Etude des mécanismes de reconnaissance du transcrit dans la terminaison de la transcription Rho-dépendante

Cédric Nadiras

UNIVERSITÉ D'ORLÉANS

ÉCOLE DOCTORALE SANTE, SCIENCES BIOLOGIQUES ET CHIMIE DU VIVANT

Centre de Biophysique Moléculaire

THÈSE présentée par :

Cédric Nadiras

soutenue le : **7 décembre 2018**

pour obtenir le grade de : **Docteur de l'université d'Orléans**

Discipline/ Spécialité : Biologie Moléculaire et Cellulaire

Etude des mécanismes de reconnaissance du transcrit dans la terminaison de la transcription Rho-dépendante

THÈSE dirigée par :

Dr. Marc Boudvillain	Directeur de Recherche, CNRS Orléans

RAPPORTEURS :

Dr. Véronique Arluison	Maitre de conférences de l'université de Paris Diderot
Dr. Bruno Sargueil	Directeur de Recherche, CNRS Paris Sud

JURY :

Dr. Bertrand Castaing	Directeur de Recherche, CNRS Orléans
Dr. Bruno Sargueil	Directeur de Recherche, CNRS Paris Sud
Dr. Hervé Le Hir	Directeur de Recherche, CNRS Paris
Dr. Isabelle Virlogeux-Payan	Directrice de Recherche, INRA Tours
Dr. Marc Boudvillain	Directeur de Recherche, CNRS Orléans
Dr. Véronique Arluison	Maitre de conférences de l'université de Paris Diderot

Remerciements

Les travaux qui ont fait l'objet d'un mémoire de thèse ont été réalisés au Centre de Biophysique Moléculaire d'Orléans sous la direction de Monsieur Marc Boudvillain.

C'est donc tout naturellement que je tiens à remercier en premier mon directeur de thèse. Je le remercie de m'avoir accueilli au sein de son équipe durant ces trois années, de m'avoir donné ma chance et d'avoir pu travailler dans les meilleures conditions possibles. Je le remercie pour l'ensemble du savoir scientifique et moral qu'il m'a permis d'acquérir. Je suis ravi d'avoir travaillé en sa compagnie, car outre son appui scientifique, il a toujours été là pour me conseiller, m'aider, et il s'est toujours impliqué dans l'orientation de mes travaux. Nos rapports ont bien évolué, et tes précieux conseils m'accompagneront dans la suite de ma vie professionnelle.

Je remercie infiniment les Docteurs Hervé Le Hir et Véronique Arluison de l'honneur qu'ils m'ont fait pour avoir jugé mon travail en tant que rapporteurs. Je remercie profondément le Docteur Bertrand Castaing d'avoir présidé le jury de ma thèse, les Docteurs Bruno Sargueil et Isabelle Virlogeux-Payant d'avoir accepté d'être dans mon jury de soutenance de thèse. J'ai été très honoré de leurs présences et je leur suis très reconnaissant qu'ils aient consacré de leurs temps pour assister à la présentation de ma thèse.

Je remercie très sincèrement Éric Eveno pour son aide précieuse et sans qui tout ce travail n'aurait pas été possible, notamment sur les aspects de bio-informatique. Nos différentes discussions de travail vont me manquer et je garderai un excellent souvenir de cette collaboration.

Je remercie également Annie Schwartz pour ses précieux conseils, son enseignement et sa bonne humeur, toutes ses valeurs m'ont accompagné en permanence durant ces trois années. Je remercie aussi Émilie Soares et Mildred Delaleau pour leurs nombreux conseils et leurs présences durant mes travaux.

Je remercie également les Docteurs Nara Figueroa-Bossi, Lionello Bossi et Emmanuel Margeat pour leurs collaborations et leurs conseils pour l'optimisation des travaux.

Je remercie l'ensemble des doctorants du CBM avec qui j'ai passé de nombreux moments très agréables, notamment au travers de l'Association des Doctorants du CBM (ADOC). Une mention spéciale pour Justine qui m'a épaulé durant ma présidence de l'association.

Je remercie aussi l'ensemble de mes amis que je me suis fait durant mes années d'études à travers la France…notamment : Tehina, Alysson, Marc, Emmanuel, Stéphanie, Damien, Romain, '*la Strasbourg Team*' : Guillaume, Julien, Jeanne, Antoine, Clémentine, Laura, tous m'ont soutenu et encouragé tout au long de mon cursus.

Je remercie mes parent Marylène et Pierre Nadiras, mon frère Mathieu et ma sœur Laura, ainsi que toute ma famille, qui ont toujours été présents à mes côtés.

Une grande pensée pour ma grand-mère Denise Aujol et pour mon grand-père Claude Nadiras, partis 6 mois avant la fin de ma thèse, ils ont toujours été à mes côtés, pour m'épauler et me soutenir, aujourd'hui j'espère mais j'en suis sûr qu'ils sont fiers du travail que j'ai accompli.

Mes derniers remerciements vont à Lucas D'Agui pour son soutien indéfectible.

Abréviations

- ^{32}P : Phosphore 32
- A : Adénine
- ADN : Acide Désoxyribonucléique
- ADNc : ADN complémentaire
- Ala : Alanine
- Alx : Alexa 488
- anti-IgG : anti- Immunoglobuline G
- ARN : Acide Ribonucléique
- ARNm : Acide Ribonucléique messager
- ARNP : ARN Polymérase
- ARNr : Acide Ribonucléique ribosomique
- ARNt : Acide Ribonucléique transfert
- Ato : ATTO 647N
- ATP : Adénosine Triphosphate
- bcm : Bicyclomycine
- BHQ2 : *Black Hole Quencher® Dyes in 2000*
- BSA : Albumine de Sérum Bovin (*Bovine Serum Albumin*)
- *B. subtilis* : *Bacillus subtilis*
- C : Cytosine
- CET : Complexe d'Elongation
- *C. glutamicum* : *Corynebacterium glutamicum*
- ChiP-array : Immunoprécipitation de la chromatine sur Chip (*Chromatin Immunoprecipitation followed by microarray identification*)
- Clip-seq : Immunoprécipitation des composés fixés couplé à un séquençage haut-débit (*Cross-Linking Immunoprecipitation coupled to high-throughput sequencing*)
- CIP : Phosphatase Alcaline (*Calf Intestine Phosphatase*)
- Cryo-EM : Cryo-microscopie électronique (*Cryo-electron microscopy*)
- CTD : Domaine Carboxy-terminal (*Carboxy-terminal Domain*)
- C-ter : Extrémité Carboxy-terminal
- CTP : Cytidine Triphosphate
- Cy : Cyanine
- DTT : Dithiothréitol
- *E. coli* : *Escherichia coli*
- EcRho : Rho d'*Escherichia coli*
- FMN : Flavin mononucleotide
- FTH : *β Fla-Tip Helix*
- G : Guanine
- EDTA : Éthylènediaminetétraacétique
- ePEC : Etat de Pause « Elémentaire » du Complexe d'élongation
- GFP : Protéine fluorescente verte (*Green Fluorescent Protein*)
- Gly : Glycine
- GTP : Guanosine Triphosphate

- iCLIP-seq : Résolution individuelle des nucléotides fixés aux UV et immunoprécipité couplé à un séquençage haut-débit (Individual-nucleotide resolution UV Crosslinking and Immunoprecipitation coupled to high-throughput sequencing)
- IF1, IF2, IF3 : Facteurs d'initiation 1, 2 et 3
- ¡RAPs : *RNA polymerase binding aptamers inhibitory* RAPs
- kB : Kilo Base
- Kd : Constante de dissociation
- *K. pneumoniae : Klebsiella pneumoniae*
- LB : Luria Broth
- Mb : Méga Base
- MBP : Sonde Moléculaire (*Molecular Beacon Probe*)
- Mg^{2+} : Ion Magnésium
- MW : Marqueur de poids moléculaire (Molecular Weight marker)
- NER : Réparation par excision de nucléotides (*Nucleotide Excision Repair*)
- NHB : *N-terminal Helix Bundle*
- NGS : Séquençage de Nouvelle Génération (*Next-Generation Sequencing*)
- Nt : Nucléotide
- N-ter : Extrémité Amine-terminale
- NTD : Domaine Amino-terminal (*Amino-terminal Domain*)
- NTP : Nucléotide Triphosphate
- *Nut :* Site de fixation de NusA (*NusA-utilization*)
- *ops* : *Operon Polarity Suppressor*
- PAGE : Électrophorèse sur gel de polyacrylamide (*Polyacrylamide Gel Electrophoresis*)
- Pb : Paire de base
- pb/s : Paire de base par seconde
- PBS : Site primaire de liaison (*Primary Binding Site*)
- PCR : Réaction en chaîne par polymérase (*Polymerase Chain Reaction*)
- Phe : Phénylalanine
- P_i : phosphate
- PIFE : Amélioration de la fluorescence induite par la protéine (*Protein Induced Fluorescence Enhancement*)
- PK : Protéinase K
- *Qut :* Site de fixation de Q (*Q-utilization*)
- QSAR : Relation quantitative de l'activité structurelle (*Quantitative Structure Activity Relationship*)
- R : Purine (A et G)
- RAPs : *RNA polymerase binding aptamers*
- RARE : Élément ARN antagoniste de Rho (*Rho-Antagonizing RNA Element*)
- SAM : S-adénosyl Méthionine

- SBS : Site secondaire de liaison (*Secondary Binding Site*)
- SDS : Dodécylsulfate de Sodium (*Sodium Dodecyl Sulfate*)
- Ser : Sérine
- sRNA : Petit ARN (*small RNA*)
- *S. Typhimurium* : *Salmonella Typhimurium*
- T : Thymine
- T4 PNK : T4 Polynucléotide Kinase
- TBS : Tampons à base saline (*Tris-buffered Saline*)
- RBS : Site de fixation du ribosome (*Ribosome binding site)*
- RBP : Protéine de liaison à l'ARN (*RNA Binding protein*)
- R_i : ribose
- Rif : Rifampicine
- RMN : Résonnance Magnétique Nucléaire
- RNAseq : Séquencage de l'ARN (*RNA sequencing)*
- rNTP : Ribonucleoside Tri-phosphate
- RP_c : Complexe de transcription fermé
- RP_o : Complexe de transcription ouvert
- Rut : Site de fixation de Rho (*Rho-utilization*)
- TCR : Réparation couplée à la transcription (*Transcription Coupled Repair*)
- TEapp : Terminaison apparente
- TL : *Trigger Loop*
- T_{tb} : Tige boucle
- TPP : Thiamine Pyrophosphate
- *tsp* : Site de terminaison (*termination stop points*)
- Tyr : Tyrosine
- U : Uracile
- UTP : Uridine Triphosphate
- UTR : Région transcrite non-traduite (*Untranslated Transcribed Region*)
- UV : Ultraviolet
- Y : Pyrimidine (C et T)

Table des matières

A. Introduction

Les bactéries sont présentes dans l'ensemble des niches écologiques car elles ont réussi à s'adapter à une très grande variété de conditions physico-chimiques et biologiques. Cette adaptabilité bactérienne est largement liée à l'émergence de mécanismes de réponse au changement du milieu qui contribuent à moduler l'expression génique, en particulier au niveau de la transcription (Campagne et al., 2014; Murakami, 2015; Peters et al., 2011; Ray-Soni et al., 2016; Roberts et al., 2008; Washburn and Gottesman, 2015). Cette modulation est gouvernée par divers facteurs spécifiques, de type ARN intervenant en *cis* (riboswitches, par exemple) ou en *trans* (petits ARN non-codants « sRNAs ») ou de type protéique (par exemple : CsrA, Hfq, Rho, ribosome) (Boudvillain et al., 2013; Holmqvist et al., 2016; Kim and Breaker, 2008).

Parmi ces facteurs de régulation, Rho (Roberts, 1969) contribue à la stabilité et à l'expression du génome bactérien grâce à sa capacité à transloquer le long de l'ARN et à induire la terminaison de la transcription (**Figure 1**) (pour revue : (Boudvillain et al., 2013; Rabhi et al., 2010a; Ray-Soni et al., 2016)).

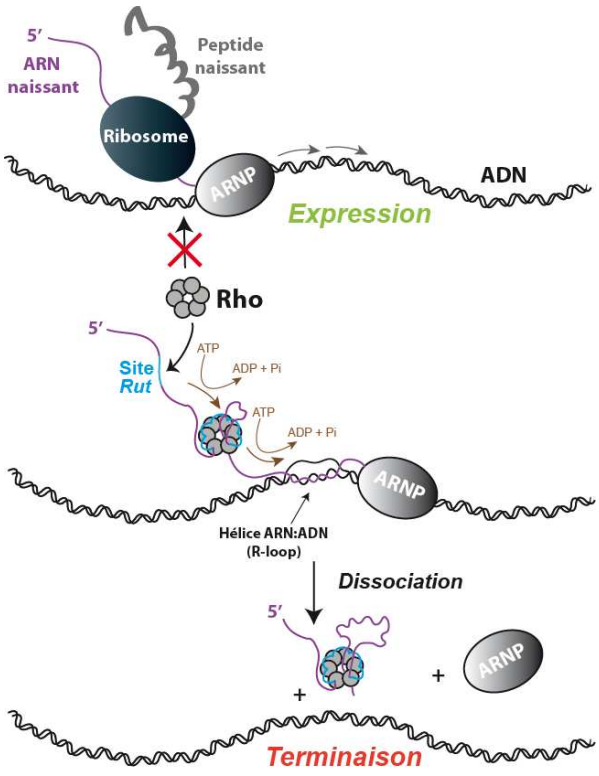


Figure 1 : Principe de la terminaison de la transcription Rho-dépendante. Le facteur Rho se fixe au transcrit naissant au niveau d'un site *Rut* (en bleu clair), si ce dernier est accesible, puis transloque de manière ATP-dépendante vers l'extrémité 3' du transcrit pour induire le démantelement du complexe d'élongation de la transcription (CET). Figure issue de (Nadiras et al., 2018a).

Son action est souvent directe (à la fin des unités transcriptionelles, par exemple) mais est parfois couplée à celle d'un autre des facteurs ARN/protéine évoqués ci-dessus (Boudvillain et al., 2013; Grylak-Mielnicka et al., 2016; Kriner et al., 2016). Dans ce cas, la liaison de Rho à l'ARN au niveau d'un site de reconnaissance *Rut*[1] (pour *Rho-utilization*) (Richardson and Richardson, 1996) peut-être sous le contrôle du second facteur, rendant ainsi « conditionnelle » la terminaison Rho-dépendante à certains *locus* du génome bactérien (Bastet et al., 2017; Bossi et al., 2012; Brandis et al., 2016; Chauvier et al., 2017; Figueroa-Bossi et al., 2014; Hollands et al., 2012; Kriner and Groisman, 2015; Sedlyarova et al., 2017; Sedlyarova et al., 2016; Takemoto et al., 2015). L'activité de Rho est aussi parfois modulée de façon conditionnelle par le second facteur à une étape ultérieure du processus de terminaison de la transcription Rho-dépendante (Kriner et al., 2016). Cette versatilité d'action, qui contribue à protéger le génome et à en adapter l'expression aux conditions du milieu (Grylak-Mielnicka et al., 2016), explique pourquoi Rho est vital pour de nombreuses bactéries (Botella et al., 2017; Bubunenko et al., 2007; Grylak-Mielnicka et al., 2016).

Une difficulté pour appréhender le spectre d'action de Rho dans son ensemble provient du fait que les sites *rut* ne présentent pas de consensus de séquence clair (Boudvillain et al., 2013; Ciampi, 2006; Ray-Soni et al., 2016). Quelques caractéristiques générales ont été définies pour ces zones d'ancrage à l'ARN (**Figure 1**), qui restent cependant insuffisantes pour prédire les sites de terminaison de la transcription Rho-dépendante (Ciampi, 2006). Ces derniers peuvent exister par centaines au sein d'un même génome. Ainsi, environ 1300 sites ont été identifiés par une analyse transcriptomique chez *Escherichia Coli* (Peters et al., 2012). Des *locus* Rho-dépendants ont été identifiés en nombre comparable chez *Mycobacterium tuberculosis* (Botella et al., 2017) et en plus faible proportion chez les firmicutes (*Bacillus subtilis* et *Staphylococcus aureus*) (Bidnenko et al., 2017; Mader et al., 2016; Nicolas et al., 2012) où Rho n'est pas indispensable. Ces listes de sites Rho-dépendants sont très probablement loin d'être exhaustives en raison de limitations techniques (seuil de détection des transcrits de faible abondance ; nature des critères de sélection et du pipeline bio-informatique) ou parce qu'une fraction du transcriptome Rho-dépendant n'est apparente que dans des conditions de culture (Bossi et al., 2012; Brandis et al., 2016; Chauvier et al., 2017; Figueroa-Bossi et al., 2014; Gall et al.,

[1] Par convention, la dénomination « *rut* » est utilisée pour la séquence ADN encodant le motif « *Rut* » au sein du transcrit ARN.

2016; Gall et al., 2018; Hollands et al., 2012; Kriner and Groisman, 2017; Sedlyarova et al., 2017; Sedlyarova et al., 2016; Sullivan and Gottesman, 1992; Takemoto et al., 2015) ou de contexte génétique (Raghunathan et al., 2018) bien spécifiques.

Pour résoudre ce problème, l'objectif principal de ma thèse était de **mieux cerner les éléments de séquence constitutifs de la terminaison Rho-dépendante** et de tenter d'utiliser ces derniers à des fins de prédiction à l'échelle génomique. Pour cela, j'ai caractérisé un grand nombre de séquences ADN à l'aide d'essais biochimiques de la terminaison Rho-dépendante. Puis, en collaboration avec un bio-informaticien dans l'équipe (Dr. Eric Eveno), j'ai défini un ensemble de descripteurs plus ou moins complexes pour ces séquences (par exemple : richesse/pauvreté en A, C, G, T). En utilisant des approches statistiques multivariées, nous avons recherché quels descripteurs étaient corrélés à l'efficacité de terminaison Rho-dépendante. Cette approche nous a permis d'identifier une combinaison de descripteurs permettant de **prédire les sites de terminaison Rho-dépendante** au sein des génomes d'*Escherichia coli* MG1655 et *Salmonella* LT2 avec un taux de succès élevé.

Au cours de cette étude, j'ai pu apprécier les limites des méthodes courantes de caractérisation biochimique de la terminaison Rho-dépendante (complexité de mise en œuvre ; utilisation de radio-isotopes ; difficulté à quantifier les signaux). Pour tenter d'y remédier, j'ai développé deux variantes d'approches existantes :

- La première permet **d'évaluer rapidement l'effet *in vitro* des petits ARN non-codants sur la terminaison Rho-dépendante.** Cette méthode repose sur un principe de marquage interne des transcrits au phosphore 32 et de détection par « *phosphorimaging* » et constitue une bonne solution d'approche de « première intention ».
- La seconde méthode permet de **changer de modalité de détection en utilisant la fluorescence** de sondes de type « *Molecular Beacon* » pour la caractérisation rapide, potentiellement quantitative, de séquences pouvant contenir des sites de terminaison Rho-dépendante.

J'ai complété mon étude de la terminaison Rho-dépendante par deux lignes de travaux exploratoires. D'une part, j'ai caractérisé *in vitro* plusieurs régions génomiques que je suspectais pouvoir participer à la régulation de gènes suivant un **mécanisme conditionnel de**

terminaison Rho-dépendante médié par un facteur protéique, un petit ARN non-codant, ou la température.

D'autre part, j'ai tenté **d'adapter l'approche CLIP-seq** (*Cross-Linking Immunoprecipitation coupled to high-throughput sequencing*) pour identifier précisément *in vivo* les sites de fixation de Rho aux transcrits chez *Salmonella*.

Dans les pages qui suivent, je présente les grandes lignes du processus d'expression génique bactérienne en insistant plus particulièrement sur les mécanismes gouvernants l'élongation de la transcription. Puis je décris précisément l'état de l'art concernant la terminaison de la transcription Rho-dépendante et discute mes résultats personnels dans ce contexte. Enfin, j'évoque les perspectives qu'offre mon travail de thèse.

B. Contexte de l’étude

I. Expression des gènes chez la bactérie

L'expression des gènes est un processus complexe constitué de plusieurs étapes interdépendantes et hautement régulées. Une caractéristique intrinsèque de l'expression génique bactérienne est que l'ensemble des étapes occupe le même compartiment spatial mais également, bien souvent temporel. Bien que les mécanismes de dégradation et de recyclage des transcrits contribuent largement à l'expression génique bactérienne, ils ne sont pas déterminants pour le sujet de cette thèse et, par soucis de concision, ne sont pas évoqués dans ce qui suit.

1) La transcription

La transcription est la première étape du processus d'expression génique. Elle est assurée par l'ARN polymérase (ARNP), une enzyme généralement composée de plusieurs sous-unités et hautement conservée dans le processus évolutif (virus mis à part), qui synthétise l'ARN à partir d'une matrice ADN double-brin (Zhang et al., 1999).

La transcription est composée de quatre étapes principales: fixation de l'ARNP à l'ADN, initiation, élongation et terminaison (**Figure 2**) (Murakami, 2015; Peters et al., 2011; Ray-Soni et al., 2016; Roberts et al., 2008; Washburn and Gottesman, 2015).

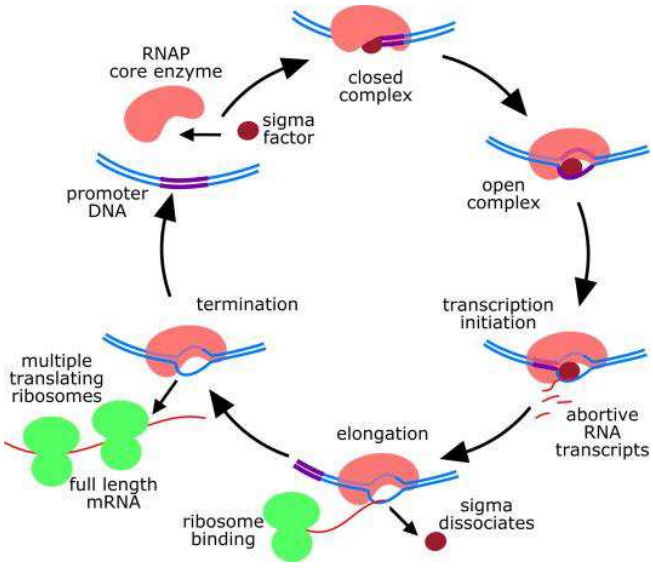


Figure 2 : Cycle de la transcription bactérienne. L'ARN polymérase (ARNP) liée au facteur d'initation Sigma, se fixe à l'ADN au niveau d'un promoteur (en violet) sous la forme d'un complexe fermé. L'isomérisation en complexe ouvert et la formation de la bulle de transcription caractérisent l'étape d'initiation. L'ARNP entre en phase d'élongation lorsque la chaine ARN atteint une taille d'environ 15 nt, ce qui stabilise le complexe de transcription et est accompagné de la dissociation du facteur sigma. Chez les bactéries, la traduction de l'ARNm (en rouge) en protéines par les ribosomes est concomitante à la synthèse de l'ARN. Au cours de l'étape de terminaison, le complexe d'élongation de la transcription (CET) est dissocié de façon irréversible. Figure issue de (Stracy and Kapanidis, 2017).

L'ARNP bactérienne est composée de cinq sous-unités (2 sous-unités α et les sous-unités β, β' et ω) formant le cœur de l'enzyme (« *core enzyme* ») (**Figure 3**). Ce dernier est capable de synthétiser *in vitro* de l'ARN à partir d'une matrice ADN de séquence non-spécifique (Burgess and Travers, 1970; Murakami, 2015).

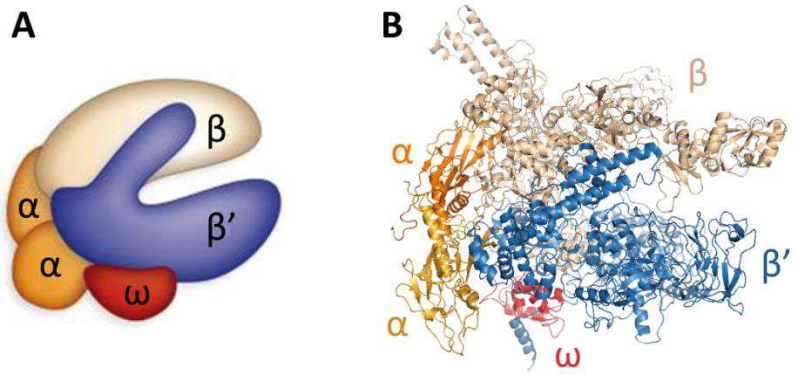


Figure 3 : Organisation générale de l'ARNP bactérienne. Le « core enzyme » adopte une conformation en forme de « pince de crabe ». Représentation **(A)** schématique issue de (Freeman, 2012) et **(B)** structurale de l'ARNP d'*E. coli* (PDB : 3LU0) (Opalka et al., 2010).

L'addition de la sous-unité σ à ce coeur permet de constituer l'holoenzyme capable de synthétiser spécifiquement, *in vitro* et *in vivo*, de l'ARN à partir d'un promoteur bactérien (**Figure 4A**). Il existe plusieurs facteurs σ assurant des fonctions biologiques différentes (**Tableau 1**) (Bae et al., 2015; Campagne et al., 2014; Campbell et al., 2002; Cook and Ussery, 2013; Darst et al., 2014; Davis et al., 2017).

Tableau 1 : Facteurs Sigma présents chez *E. coli* et leurs fonctions. Inspiré de (Cook and Ussery, 2013).

Gène	Alias	Fonctions
rpoD	σ^{70}, σ^{D}	Essentiel à la survie de la bactérie. Assure la transcription des gènes de ménage pendant la phase exponentielle de croissance.
rpoS	σ^{38}, σ^{S}	Pas essentiel à la survie et la croissance bactérienne. Régulateur principal de la réponse au stress, notamment pendant la phase stationnaire.
rpoH	σ^{32}, σ^{H}	Contrôle la transcription des gènes impliqués dans les chocs thermiques.
fliA	σ^{28}, σ^{F}	Intervient dans le contrôle des gènes liés au développement, par exemple la biosynthèse des flagelles ou lors de la sporulation.
rpoN	σ^{54}, σ^{N}	Contrôle l'expression de gènes liés à l'azote.
rpoE	σ^{24}, σ^{E}	Contrôle la transcription des gènes impliqués dans les chocs thermiques ou osmotiques.
fecI	σ^{19}	Intervient dans le contrôle des gènes liés au transport du fer.

L'holoenzyme se fixe à l'ADN au niveau de séquences promotrices adaptées à chaque facteur σ (**Figure 4A**) (Bae et al., 2015; Browning and Busby, 2016; deHaseth et al., 1998; Winkelman and Gourse, 2017; Zuo and Steitz, 2015). Le facteur σ^{70} d'*E. coli* reconnait les séquences ADN « -10 » (séquence consensus : 5'-TATAAT-3'), « -10 étendue » (5'-TGTG$_n$-3') et « -35 » (5'-TTGACA-3') des promoteurs (**Figure 4B**) (Gaal et al., 2001; Zuo and Steitz, 2015). Cette interaction initie la formation du complexe fermé (RP_c) entre l'holoenzyme et l'ADN (**Figure 4B**). Les modifications structurales de l'ARNP et de l'ADN engendrées par la formation du complexe fermé vont s'accentuer lors de la conversion de ce dernier en complexe ouvert (RP_o) (**Figures 2 & 4A**). Cette conversion est caractérisée notamment par la formation de la bulle de transcription correspondant à l'ouverture d'une région ADN qui finira par atteindre une taille de 16 à 18 paires de bases [pb] (Bae et al., 2015; Chakraborty et al., 2012; Saecker et al., 2011; Zuo and Steitz, 2015).

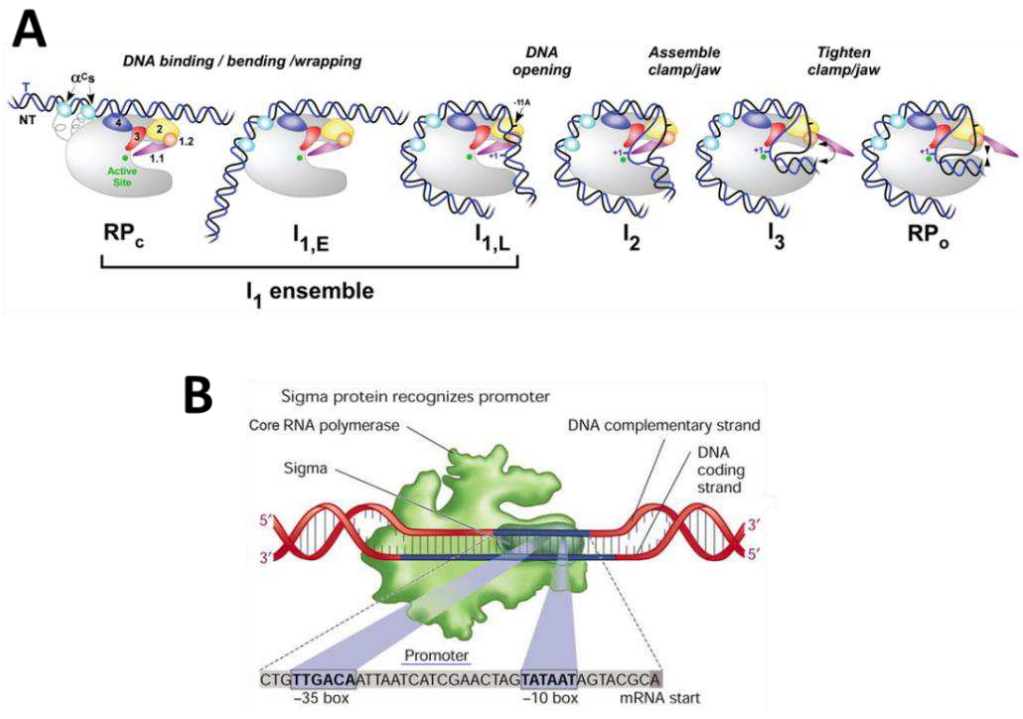


Figure 4 : Etapes de l'initiation de la transcription bactérienne via le facteur σ. (A) La reconnaissance du promoteur ADN par l' ARNP:σ, décrite dans le panneau B, forme le complexe fermé de transcription (RP_c). Puis des réarrangements (torsion de l'ADN, ouverture de la bulle de transcription, contact avec les deux pinces β de l'ARNP) aboutissent à la formation d'un complexe ouvert (RP_o). Figure issue de (Ruff et al., 2015). **(B)** Reconnaisance par le complexe ARNP:σ du promoteur et des éléments « -35 » et « -10 » sur l'ADN . Figure tirée d'internet (http://helicase.pbworks.com/).

Le complexe ouvert (RP_o) est compétent pour synthétiser un court fragment d'ARN de 9 à 11 nucléotides (nt). Durant ce processus, les contacts entre l'ARNP et le promoteur

restent intacts mais ce dernier est progressivement "tordu" alors que la bulle de transcription est agrandie (**Figure 4A**) (Kapanidis et al., 2006). Ces changements induisent un stress structural, caractérisé également par le clash stérique entre l'extrémité 3' du fragment d'ARN naissant et la région $\sigma^{3.2}$ du facteur σ insérée dans le canal de sortie de l'ARN, qui va se traduire soit par la dissociation de la chaine ARN (transcription abortive), soit par la poursuite de la synthèse d'ARN et la transition vers la phase d'élongation (**Figure 2**) (Krummel and Chamberlin, 1992; Murakami et al., 2002; Petushkov et al., 2017; Saecker et al., 2011). Dans ce dernier cas, le stress structural est brutalement réduit par la rupture de la liaison entre l'ARNP et le promoteur ADN et par l'expulsion du domaine $\sigma^{3.2}$ en dehors du canal de sortie de l'ARN induite par le transcrit naissant (ce qui va conduire à la dissociation du facteur σ). La conversion du complexe d'initiation en complexe d'élongation est considérée achevée lorsque la chaine d'ARN est suffisamment longue (≈15 nt) pour occuper pleinement le canal de sortie de l'ARN (**Figure 5**) (Bae et al., 2015; Saecker et al., 2011).

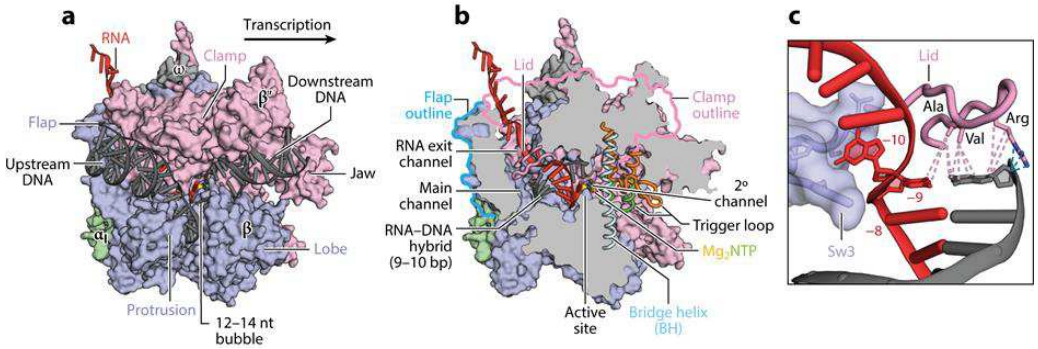


Figure 5 : Structure du complexe d'élongation de la transcription (CET) de *Thermus aquaticus*. Les éléments importants : l'ADN en gris foncé, l'ARN en rouge, les sous-unités de l'ARNP (αI en vert, β en violet, β' en rose, ω en gris) et les domaines régulateurs (« *Triger loop* » en orange, « *Clamp* » et « *Lid* » en rose, « *Flap* » en bleu et « *Bridge helix* » en bleu clair). Le domaine « *Lid* » (panneau C) de la sous-unité β' assure la séparation de l'hybride ADN:ARN. La poche Sw3 assure une interaction hydrophobe avec le premier nucléotide ARN (i_{-10}) issu de l'hybride ADN:ARN. Figure issue de (Ray-Soni et al., 2016).

La phase d'élongation est caractérisée par une grande processivité du CET (Complexe d'Elongation de la Transcription) due notamment à la stabilité apportée par le réseau d'interactions entre l'ARNP, la matrice ADN et le transcrit naissant (**Figure 5**). L'addition de chaque nucléotide à l'extrémité 3' du transcrit est un processus constitué de deux étapes (**Figure 6**). Dans la première étape, une nouvelle liaison phosphodiester entre le groupement 3'-OH du brin d'ARN et le α-phosphate du NTP entrant est formée (état pré-transloqué). Puis, le complexe transcriptionnel avance d'une paire de bases [pb] de manière à aligner la nouvelle extrémité 3'-OH du brin ARN avec le site actif (état post-transloqué) (**Figure 6**). Ce

mécanisme cyclique assure une vitesse moyenne de translocation du CET de 10 à 100 pb/s *in vivo* (Peters et al., 2011; Roberts et al., 2008).

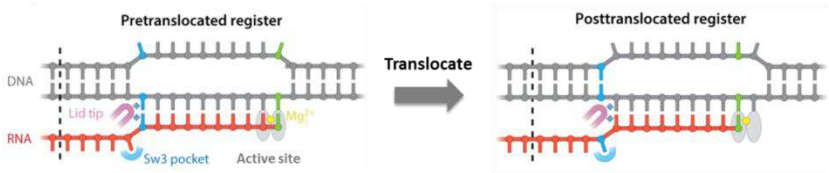


Figure 6 : Cycle de translocation du CET. Le CET alterne entre l'état « pré-transloqué » (après addition d'un nouveau NTP, en vert) et l'état « post-transloqué » (pour réaligner l'extrémité 3' du transcrit ARN dans le site actif). Figure issue de (Ray-Soni et al., 2016).

2) Le couplage transcription-traduction

La traduction des ARNm en protéines est assurée par le ribosome organisé de manière asymétrique en deux sous-unités formées d'ARNr et de protéines (**Tableau 2**).

Tableau 2 : Composition du ribosome d'*Escherichia coli*.

	Ribosome 70S	Sous-unité 50S	Sous-unité 30S
ARN ribosomaux (ARNr)	3	2 (23S et 5S)	1 (16S)
Protéines ribosomiques	55	34 (L1 → L34)	21 (S1 → S21)

Pour permettre l'initiation de la traduction bactérienne, l'ARNr 16S (via sa séquence 5'-CCUCCU-3') de la petite sous-unité 30S doit tout d'abord interagir avec une courte séquence de l'ARNm riche en purines appelée séquence *Shine-Dalgarno* ou RBS (*Ribosome-binding Site*, de séquence consensus 5'-GGAGG-3'), située 5-10 nucléotides en amont du codon initiateur (**Figure 7**) (Shine and Dalgarno, 1974; Steitz and Jakes, 1975). Cela permet le recrutement des facteurs d'initiations IF1, IF2, IF3, de l'ARNt initiateur (fMet-tRNA^{fMet}) et de la grande sous-unité 50S. La traduction commence au niveau d'un codon initiateur (5'-AUG-3' / 5'-GUG-3' / 5'-AUG-3') et se poursuit jusqu'à ce que le ribosome rencontre un codon stop (5'-UAG-3' / 5'-UGA-3' / 5'-UAA-3') dans le même cadre de lecture (**Figure 7**). Les ARNm peuvent contenir une ou plusieurs séquences codantes ; ils sont alors appelés respectivement ARNm mono- et poly-cistroniques (Huttenhofer and Noller, 1994; Steitz and Jakes, 1975).

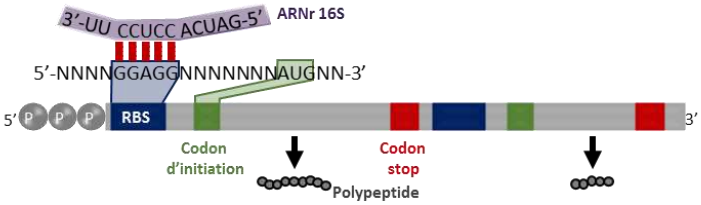


Figure 7 : Reconnaissance du site RBS par la sous-unité 30S du ribosome. Figure inspirée de (Watson et al., 2009)

Chez les eucaryotes, la traduction a lieu dans le cytoplasme et est donc physiquement séparée de la transcription qui est un processus nucléaire. Cette séparation n'existe pas chez les bactéries où transcription et traduction sont physiquement et temporellement couplées (Burmann et al., 2010; Miller et al., 1970; Saxena et al., 2018; Strauss et al., 2016). Ce couplage est assuré par le facteur protéique essentiel NusG ou, dans certains cas, par son paralogue RfaH.

NusG contient deux domaines pour relier l'ARNP et le ribosome « *leader* » du polysome : le domaine N-terminal (N-ter) interagit avec les sous-unité β/β' de l'ARNP (Kang et al., 2018; Mooney et al., 2009b; Strauss et al., 2016; Svetlov et al., 2007) alors que le domaine C-terminal (C-ter) se lie à la protéine ribosomique S10 (aussi appelée NusE) de la sous-unité 30S du ribosome (**Tableau 2 et Figure 8 et 9A**) (Burmann et al., 2010; Strauss et al., 2016).

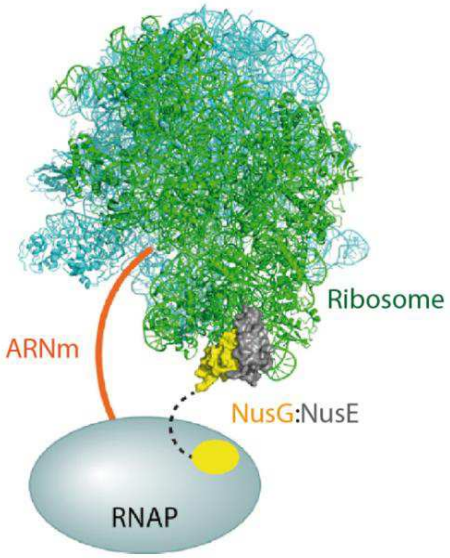


Figure 8 : Couplage entre la transcription et la traduction via l'interaction NusE:NusG. Figure issue de (Ma et al., 2016).

Le facteur paralogue RfaH se lie également à la sous-unité β' de l'ARNP (Belogurov et al., 2009) mais uniquement quand une séquence ***ops*** (*Operon Polarity Suppressor*) est contenue dans le brin ADN non codant (ADN-nt) (**Figure 9A-B**) (NandyMazumdar and Artsimovitch, 2015; Zuber et al., 2018). Tout comme pour NusG, la partie C-ter de RfaH se lie à la protéine ribosomique S10 (NusE) pour favoriser le couplage entre la transcription et la traduction (Burmann et al., 2012). RfaH active les opérons assez longs qui codent pour les antibiotiques, les capsules, les toxines et les pili en inhibant la terminaison Rho-dépendante

(NandyMazumdar and Artsimovitch, 2015). Une étude récente par cryo-microscopie électronique (Cryo-EM) démontre que les parties N-ter de NusG et de RfaH interagissent de façon très similaire avec le CET (**Figure 9A**). Une différence notable existe malgré tout : NusG reconnaît une «protubérance » de la sous-unité β (**Figure 9C**) alors que RfaH se lie à la séquence ***ops*** qui adopte une conformation en tige boucle au sein du brin ADN-nt dans la bulle de transcription (**Figure 9B-C**) (Kang et al., 2018). Cette interaction confère une plus grande stabilité au complexe formé avec RfaH qui est ainsi capable d'exclure NusG à cette étape lors de la transcription de régions *ops*-dépendantes (Hu and Artsimovitch, 2017). Par la suite, la partie RfaH-Nter rompt ces contacts avec la séquence ***ops*** pour établir de nouvelles interactions au niveau de la « protubérance « de la sous-unité β.

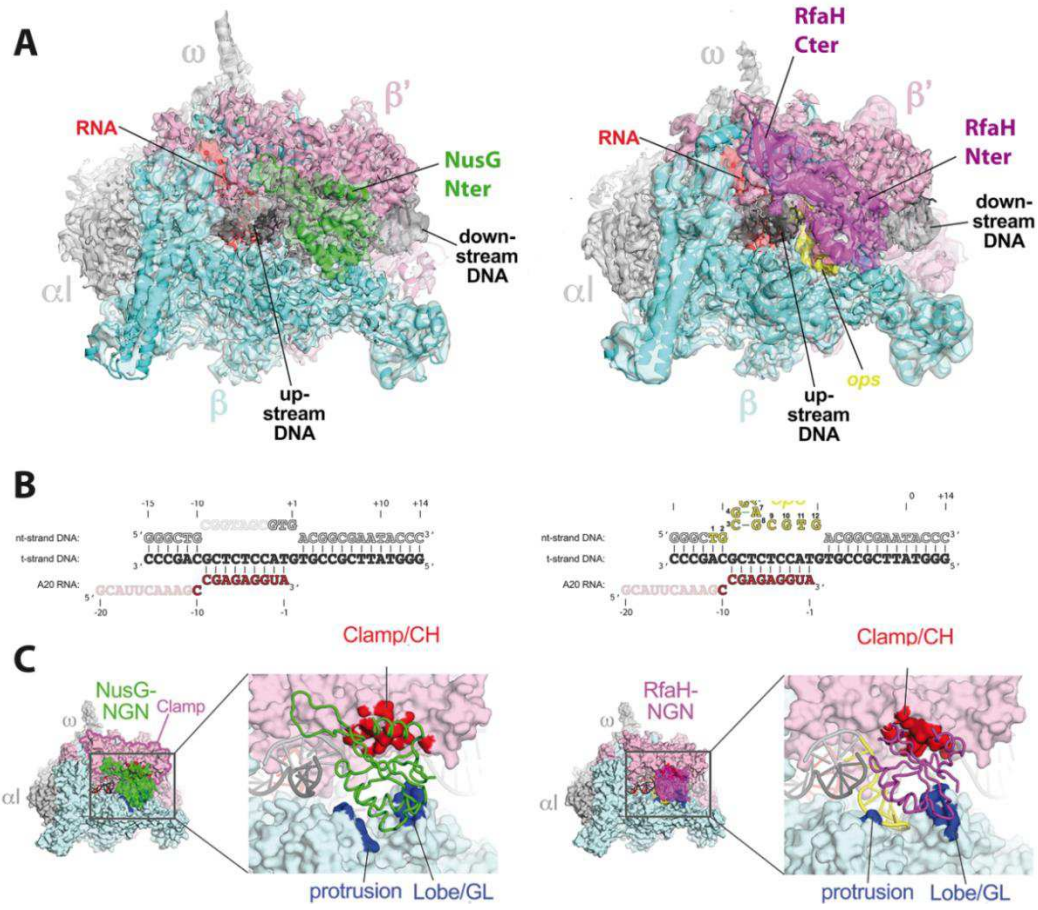


Figure 9 : Structures de NusG et RfaH en interaction avec le CET. (A) Structures par Cryo-EM (PDB [NusG] : 6C6U ; PDB [RfaH] : 6C6S). L'ADN est repésenté en noir, et l'ARN en rouge. **(B)** Sequence d'acide nucléique utilisé pour l'observation en Cryo-EM de RfaH. Cette séquence permet de bien visualiser la partie ***ops*** qui forme une petite tige boucle. **(C)** Interactions de NusG-Nter et de RfaH-Nter avec l'ARNP. Les sous-unités β et β', sont en bleu clair et rose, respectivement. Figure issue de (Kang et al., 2018)

Une autre étude récente par Cryo-EM a révélé un emboitement des appareils de transcription et de traduction qui, ensemble, semblent former une machinerie unique

appelée « *expressome* » par les auteurs (**Figure 10**) (Kohler et al., 2017). Dans ce complexe, le ribosome contacte directement le domaine C-ter de la sous-unité α de l'ARNP. Ce domaine est impliqué dans des interactions avec le facteur σ et avec le promoteur ADN lors de la phase d'initiation (**Figure 4**) et n'est donc pas disponible pour former l' « *expressome* » dès cette étape. Les auteurs indiquent que l'interaction entre ribosome et ARNP au sein de l' « *expressome* » n'est pas incompatible « stériquement » avec la liaison des facteurs d'élongation transcriptionnelle GreA, GreB, NusG ou RfaH (Kohler et al., 2017). Dans ce contexte, NusG (ou RfaH) n'aurait alors qu'un rôle de stabilisation de l' « *expressome* » bactérien.

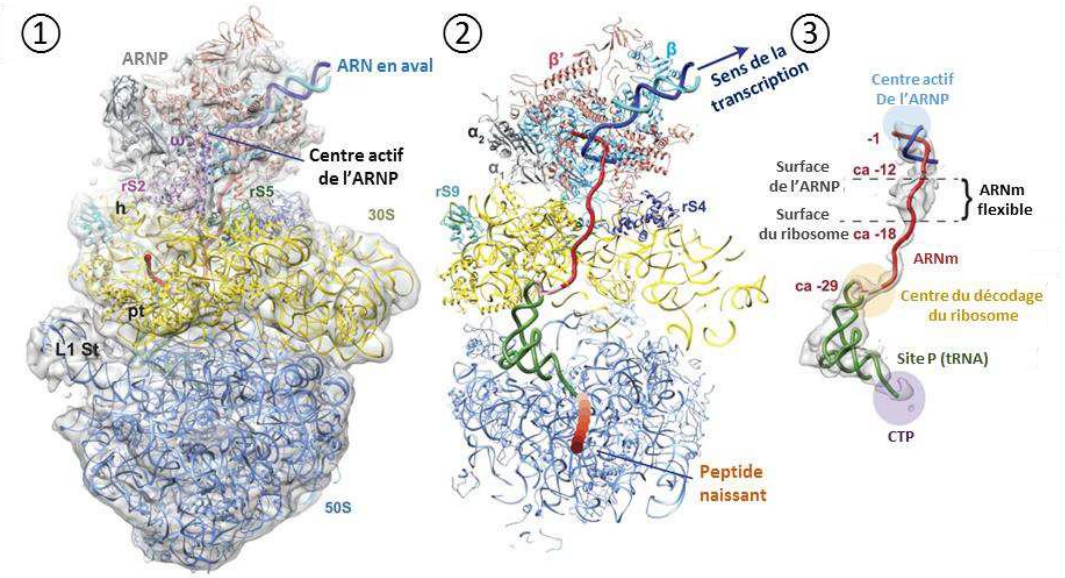


Figure 10 : Architecture de l' « *expressome* » bactérien. Le ribosome, en bleu foncé (50S) et jaune (30S), est disposé sous l'ARNP avec laquelle il « s'emboite » parfaitement. CTP : centre de transfert peptidique. Figure adaptée de (Kohler et al., 2017).

3) Régulation transcriptionelle versus régulation post-transcriptionelle

La régulation transcriptionnelle permet de contrôler le niveau d'expression des gènes en modulant directement le processus de transcription de l'ADN (Holmqvist and Vogel, 2018). La régulation post-transcriptionnelle est, comme son nom le suggère, une phase plus tardive regroupant les mécanismes de régulation qui impliquent les molécules d'ARN néoformées (Holmqvist and Vogel, 2018). A l'origine, ces deux phases étaient considérées (et étudiées) indépendamment, modulant respectivement l'une la transcription et l'autre les mécanismes

de traduction et de dégradation des ARN (chez les eucaryotes, d'autres étapes comme l'épissage et l'export nucléaire participent également à la régulation post-transcriptionnelle). Cependant, une vision plus intégrée de ces deux phases, en accord avec la notion d' « *expressome* » bactérien (partie B-I-2) (**Figure 10**) (Kohler et al., 2017) doit maintenant être sérieusement considérée. Je détaille ci-dessous deux exemples de régulation regroupant des composantes appartenant aux deux phases et qui illustrent cette vision plus intégrée.

Le premier mécanisme, dit de « riborégulation », implique une séquence d'ARNm située dans une région non traduite (5'UTR le plus souvent) qui affecte l'expression du gène ou de l'opéron correspondant. Cette séquence appelée « riboswitch » est capable de former des structures alternatives en fonction de la présence (et de l'interaction) ou de l'absence d'un ligand (petits métabolites, ions métalliques) ou d'un ARN de transfert (ARN_t) (**Figure 11**) (Dann et al., 2007; Epshtein et al., 2003; Gutierrez-Preciado et al., 2009; Hammann and Westhof, 2007; Kim and Breaker, 2008; Serganov et al., 2008; Sudarsan et al., 2006). Les riboswitches sont composés de deux domaines interconnectés : l'aptamère et la plateforme d'expression (**Figure 11**) (Corbino et al., 2005). La liaison au ligand est assurée par le domaine aptamère et modifie la structure de la plateforme d'expression. Cette modification peut induire une régulation post-transcriptionelle via l'exposition/l'occultation du RBS ou du codon initiateur et le contrôle de la traduction (**Figure 11A**) ou bien une régulation transcriptionnelle via la structuration/déstructuration d'un terminateur intrinsèque (Dann et al., 2007; Epshtein et al., 2003; Gutierrez-Preciado et al., 2009; Hammann and Westhof, 2007; Kim and Breaker, 2008; Serganov et al., 2008; Sudarsan et al., 2006) (**Figure 11B**) ou Rho-dépendant (Hollands et al., 2012; Proshkin et al., 2014; Takemoto et al., 2015). Certains riboswitches traductionnels (**Figure 11A**) induisent également la terminaison Rho-dépendante à un site situé plus en aval (du fait du découplage-transcription-traduction engendré) (Bastet et al., 2017; Chauvier et al., 2017) et, de ce fait, combinent effets transcriptionnels et post-transcriptionnels.

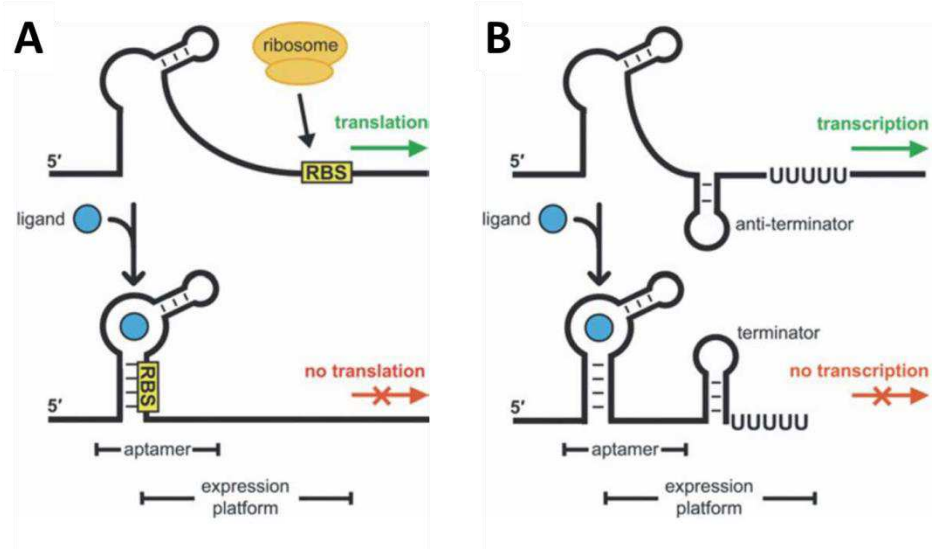


Figure 11 : Principe de riborégulation médiée par un riboswitch. Dans l'illustration, la conformation sans ligand (effecteur) du riboswitch permet l'expression du système en occultant **(A)** un site anti *Shine-Dalgarno* ou **(B)** un terminateur intrinsèque. La fixation du ligand au domaine « aptamère » du riboswitch induit un remodelage structural assurant **(A)** l'occultation du RBS (Liberman et al., 2015) ou **(B)** la formation du terminateur intrinséque. Certains riboswitches contrôlent l'accès de Rho au transcrit naissant, soit directement (Hollands et al., 2012; Proshkin et al., 2014; Takemoto et al., 2015) , soit après découplage transcription-traduction (Bastet et al., 2017; Chauvier et al., 2017). Figure inspirée de (Kim and Breaker, 2008).

Le second exemple de mécanisme régulateur mobilise des acteurs en *trans* sous la forme d'un petit ARN non-codant (sRNA ; ≈ 40 à 400 nt) généralement assisté d'une protéine chaperonne (Hfq ou ProQ), du moins chez les bactéries à gram négatif (pour revue : (Kavita et al., 2018)). Un sRNA est souvent capable de réguler plusieurs cibles en s'hybridant aux ARNm correspondants. Cette hybridation, en général caractérisée par une double hélice courte (10-25 pb) (Waters and Storz, 2009), imparfaite et localisée à proximité du RBS, inhibe la traduction de l'ARNm (**Figure 12**) et/ou entraine sa dégradation en exposant des sites de clivage pour la RNaseE (pour revue : (Kavita et al., 2018)).

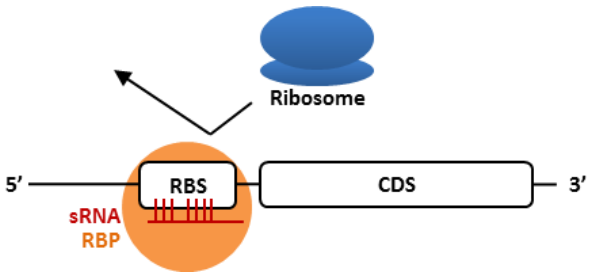


Figure 12: Exemple d'action des sRNA. L'hybridation oculte le RBS et inhibe la traduction. Figure inspirée de (Van Assche et al., 2015).

La participation d'une protéine « ARN chaperonne » comme Hfq (Arluison et al., 2007; Morita et al., 2017) ou ProQ (Holmqvist et al., 2018; Smith et al., 2004) permet à la fois de

favoriser la formation des hybrides sRNA:ARNm (Arluison et al., 2007; Bossi et al., 2012; Santiago-Frangos et al., 2016) et de protéger les sRNA de la dégradation (Holmqvist et al., 2018; Masse et al., 2003; Smirnov et al., 2016; Wagner and Romby, 2015). De nombreuses analyses transcriptomiques ont été réalisées chez *E. coli* et *Salmonella* pour identifier les cibles ARNm et sRNA de ces deux RBP (*RNA Binding protein*) (Chao et al., 2012; Holmqvist et al., 2018; Holmqvist et al., 2016; Smirnov et al., 2016; Zhang et al., 2003). Comme dans le cas des riboswitches traductionnels, certains sRNA sont également capables d'activer la terminaison Rho-dépendante et, de ce fait, mobilisent composantes transcriptionnelle et post-transcriptionnelle (Bossi et al., 2012). Cet aspect est détaillé dans le chapitre consacré à la terminaison Rho-dépendante (page 59).

II. Régulation de la transcription

Dans cette section, je décris brièvement les mécanismes qui régulent le complexe transcriptionnel bactérien sans, toutefois, aborder ce qui relève spécifiquement de la phase d'initiation. Bien que les mécanismes de contrôle de l'initiation de la transcription soient très nombreux et variés (pour revue : (Saecker et al., 2011)), aucun lien direct n'a à ce jour été établi avec la terminaison Rho-dépendante, ce qui n'est pas le cas des mécanismes décrits ci-dessous.

1) Signaux de pause et d'arrêt du complexe d'élongation de la transcription

Le CET alterne entre des phases d'élongation et des pauses provoquées par différents signaux. Ces pauses sont très importantes pour la régulation des gènes car elles permettent la synchronisation de l'action de l'ARNP avec celles des cofacteurs se fixant aux différents composants du CET (ARN, ARNP, ADN). Les signaux de pause facilitent également la formation de l' « *expressome* » bactérien (Kohler et al., 2017) et contribuent au repliement natif de l'ARN en cours de transcription (repliement co-transcriptionnel) (Larson et al., 2014; Pan and Sosnick, 2006). Des évènements de pause de l'ARNP sont également parties intégrantes des signaux d'arrêt et de terminaison décrits plus loin (Greive and von Hippel, 2005; Roberts et al., 2008).

Un premier type de pause est lié au passage du CET d'un état de « *pré-translocation* » à un état de « *post-translocation* » (**Figure 6 et 15①**). Cette transition nécessite un remodelage des liaisons entre protéines et acides nucléiques au sein du CET afin que ce dernier puisse se repositionner sur l'ADN en position i_{+1}. Ce remodelage est parfois perturbé, ce qui entraîne un état de pause « élémentaire » du complexe d'élongation (ePEC) en « pince ouverte » en raison de l'écartement transitoire du domaine « β *clamp* » (**Figure 15②**) (Zhang and Landick, 2016).

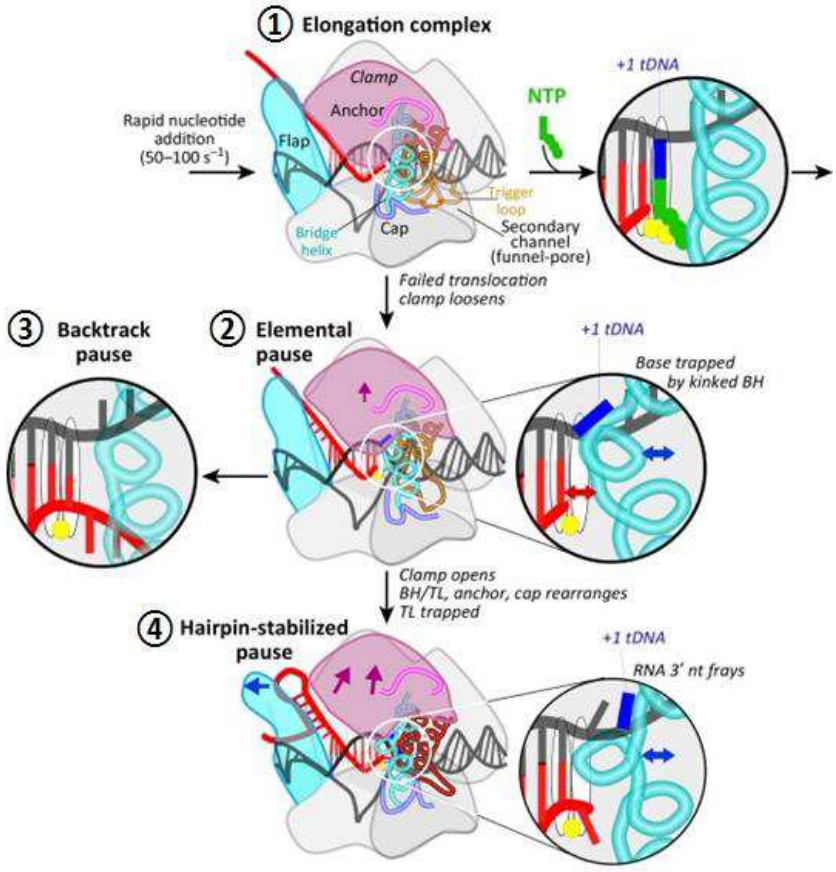


Figure 15 : Modèle de pause élémentaire des CET. ① CET dans un processus normal d'élongation durant lequel l'addition de nucléotide s'effectue suivant des oscillations de la boucle « *Trigger Loop* » (TL) près du site actif. La liaison du nucléoside triphosphate à l'ADN matrice initie le mouvement de la TL et le positionnement du NTP et du Mg^{2+} dans le site actif. ② CET dans un état de pause élémentaire avec des changements structuraux : un écartement transitoire du domaine « ***β Clamp*** », modification de la « ***Bridge helix*** », piège de la **base i_{+1}** de l'ADN. L'état de pause élémentaire peut conduire à un état de pause prolongée ou d'arrêt suite à un *backtracking* ③ ou à la formation d'une structure ARN en tige boucle dans le canal de sortie ④. La formation de la structure ARN cause l'ouverture du canal (en écartant les domaines « ***β Clamp*** » et « ***β Flap*** »), la conformation improductive de la boucle TL et l'éloignement entre les nucléotides i_{+1} de l'ARN et de l'ADN. Figure issue de (Zhang and Landick, 2016).

Cet état de pause est généralement très court avant que le CET finisse par adopter une configuration correcte en position i_{+1}. Il est néanmoins favorisé par certaines séquences. Ainsi, deux équipes ont mis en évidence un motif consensus pouvant induire une pause « élémentaire » du CET : 5'-**$G_{-10}Y_{-1}G_{+1}$**-3' (Larson et al., 2014; Vvedenskaya et al., 2014) (**Figure 16**). Les auteurs proposent que ce motif est thermodynamiquement défavorable à un repositionnement du CET en position i_{+1} pour plusieurs raisons :

- Le CET aurait du mal à ouvrir le double brin d'ADN en position i_{+1} à cause de la paire 5'-dG_{+1} /dC_{+1}-3' (comparé à une paire 5'-dA_{+1} /dT_{+1}-3').

- La dissociation de l'hybride ADN:ARN au niveau du canal de sortie en position i_{-10} serait défavorisée par la paire r**G**$_{-10}$ /d**C**$_{-10}$ (**Figure 16**).
- Enfin, l'élément 5'-r**Y**r**G**-3'/3'-d**R**-d**Y**-5' en position i_{-1} et i_{+1} défavoriserait la translocation de l'ARNP en position i_{+1} (ce qui n'es pas le cas avec un motif 5'-r**R**-r**Y**/3'-d**Y**-d**R**-5').

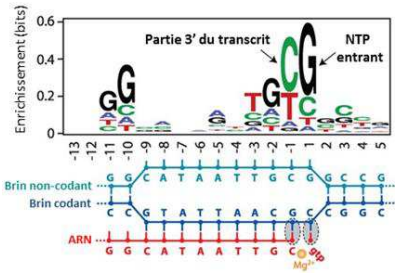


Figure 16 : Séquence consensus de pause élémentaire alignée avec la bulle de transcription. Les éléments retrouvés majoritairement sont rG$_{-10}$, rY$_{-1}$ et rG$_{+1}$. Figure issue de (Larson et al., 2014).

Ces éléments favoriseraient un état de « *pré-translocation* » et défavoriseraient l'état de « *post-translocation* ». Les auteurs ont recensé 20 000 sites possibles de pause élémentaire chez *E. coli* et démontré que ce motif consensus était conservé dans diverses lignées bactériennes (Larson et al., 2014). La séquence **G$_{-10}$Y$_{-1}$G$_{+1}$** est notamment présente près des sites RBS, ce qui permettrait de favoriser le repliement correct de l'ARN naissant et limiterait la synthèse d'ARN jusqu'à ce que la traduction débute (Larson et al., 2014). Les pauses élémentaires sont généralement très courtes mais peuvent initier une transition vers des états inactivés plus durables résultant d'un remodelage marqué du CET (Weixlbaumer et al., 2013). Nous pouvons distinguer deux catégories principales de signaux d'inactivation durable :

- Les signaux de pause (de type I ou II) à partir desquels l'ARNP peut retrouver spontanément une configuration active ; cette réactivation peut être facilitée par certains cofacteurs (NusA et NusG par exemple).
- Les signaux d'arrêt caractérisés par une inactivation irréversible de l'ARNP en absence de cofacteurs ré-activateurs (GreA et GreB par exemple).

Les signaux de pause de type I requièrent la formation d'une structure tige-boucle (*hairpin*) de 14 nt au sein du transcrit dont 5 nt peuvent se positionner dans le canal de sortie de l'ARNP. Cette structure va déplacer la base ARN$_{i-10}$ de sa poche de liaison Sw3 (**Figure 5**) et interagit avec le domaine « *flap* » (**Figure 15-④**), ce qui inhibe la fermeture de la pince et inactive transitoirement l'ARNP (**Figure 15-④ et Figure 17-①**) (Greive and von Hippel, 2005; Washburn and Gottesman, 2015).

Les signaux de pause de type II impliquent une « rétro-translocation » de l'ARNP le long de l'ADN (*backtracking*), ce qui positionne l'extrémité 3' de l'ARN dans le canal d'entrée des NTP (appelé canal secondaire) et obstrue ce dernier (**Figure 15-① et Figure 17 ②**). Il arrive parfois que la rétro-translocation piège le CET dans un état stable dont il ne peut échapper spontanément. Cet état est qualifié d'arrêt. Dans ce cas, la récupération du registre catalytique requière l'intervention de facteurs extérieurs qui vont aider l'ARNP à ré-avancer le long de l'ADN ou bien couper la portion d'ARN occupant le canal d'entrée des NTP. Le phénomène d'arrêt peut être favorisé par une incorporation erronée ou une carence en nucléotides (Nudler et al., 1997).

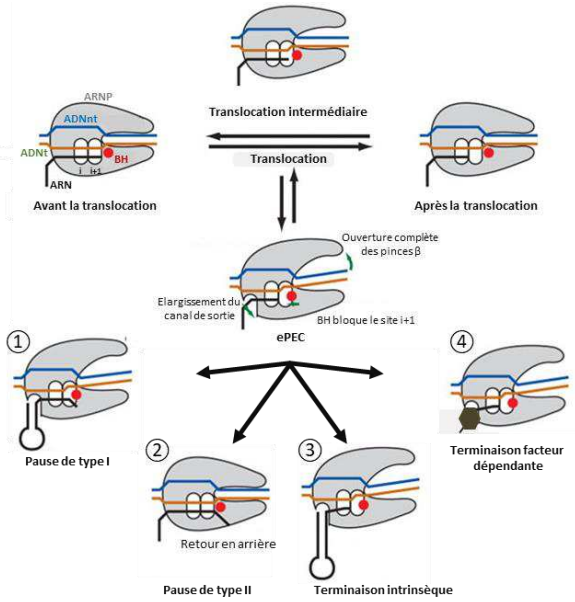


Figure 17 : Signaux de pause et de terminaison lors de l'élongation de la transcription. L'entrée dans un état de pause élémentaire transitoire (ePEC) peut conduire à une inactivation du CET plus durable voire irréversible. Différents cas sont distingués : ① pause de type I, induit par la formation d'une structure ARN en tige-boucle dans le canal de sortie de l'ARNP ; ② pause de type II liée à un recul de l'ARNP le long de l'ADN (*backtracking*) conduisant à une extrusion de l'extrémité 3' du transcrit dans le canal secondaire ; ③ terminaison intrinsèque induite par la formation d'un motif ARN en tige-boucle à proximité de l'extrémité 3' du transcrit ; ④ terminaison facteur-dépendante nécessitant l'intervention d'un facteur tel que Rho pour démanteler le CET. Figure inspirée de (Weixlbaumer et al., 2013).

Il existe un dernier type de pause dite « proximale » (ou σ-dépendante), induite par le facteur σ et localisée près des promoteurs (Mooney et al., 2005). Lors de l'initiation de la transcription, le facteur σ en interaction avec l'ARNP se disloque pour permettre le passage du complexe en phase d'élongation (voir partie B-I-1, page 9). Bien que les domaines $\sigma^{1.1}$ et $\sigma^{3.2}$ soient exclus de l'interaction avec l'ARNP à ce stade, σ reste parfois lié quelque temps au CET via son domaine σ^2 (Kapanidis et al., 2005; Mooney et al., 2005) et peut ainsi encore reconnaitre certaines séquences ADN ressemblant aux promoteurs. Cette interaction peut induire une pause du CET jusqu'au démantèlement complet du facteur σ. Il a été proposé que les signaux de pause proximale favorisent le couplage entre la transcription et la traduction (Mooney et al., 2005; Mooney et al., 2009a).

2) Régulation des signaux de pause et d'arrêt

Les signaux de pause et d'arrêt sont étroitement régulés car ils peuvent avoir des conséquences délétères. Par exemple, les CET arrêtés peuvent représenter des obstacles pour les réplisomes et entrainer une instabilité génomique (Dutta et al., 2011). Cette dernière semble liée à la formation de structures en « R-loops » (hybrides ARN:ADN) au niveau des CET retro-transloqués, suivant un mécanisme moléculaire qui reste à préciser (Dutta et al., 2011).

Différents facteurs protéiques (les principaux étant NusA, NusG, RfaH, Mfd, GreA et GreB) participent à la régulation des signaux de pause et d'arrêt. Des contraintes physiques peuvent également limiter l'impact des signaux d'arrêt ou de pause de type II. Par exemple, la présence d'un ribosome accolé à l'ARNP au sein de l' « *expressome* » (**Figure 10**) et lui-même engagé dans un mouvement de translocation directionnelle (le long de l'ARNm) limite la capacité du CET à reculer le long de l'ADN dans les régions codantes (Proshkin et al., 2010).

Le facteur NusA stimule la pause du type I et la terminaison intrinsèque du CET (voir paragraphe suivant) en favorisant la formation d'une tige boucle ARN dans le canal de sortie (Guo et al., 2018; Gusarov and Nudler, 2001; Ha et al., 2010; Strauss et al., 2016; Washburn and Gottesman, 2015; Yakhnin and Babitzke, 2010). Pour cela, NusA se lie à l'ARNP, dépourvue du facteur σ, au niveau du canal de sortie et utilise un réseau d'interactions (Guo et al., 2018; Mooney et al., 2009a) (**Figure 17**) pour :

- guider l'ARN naissant le long de la surface chargée positivement ce qui stabilise l'hélice **FTH** (*β Fla-Tip Helix*) et empêche celle-ci d'interférer avec la formation du duplex ARN dans le canal de sortie de l'ARN (**Figure 17-B**).
- stabiliser une conformation de pause de type I grâce à des interactions protéiques avec l'ARNP et la tige boucle de l'ARN (**Figure 18-B**) (Guo et al., 2018).

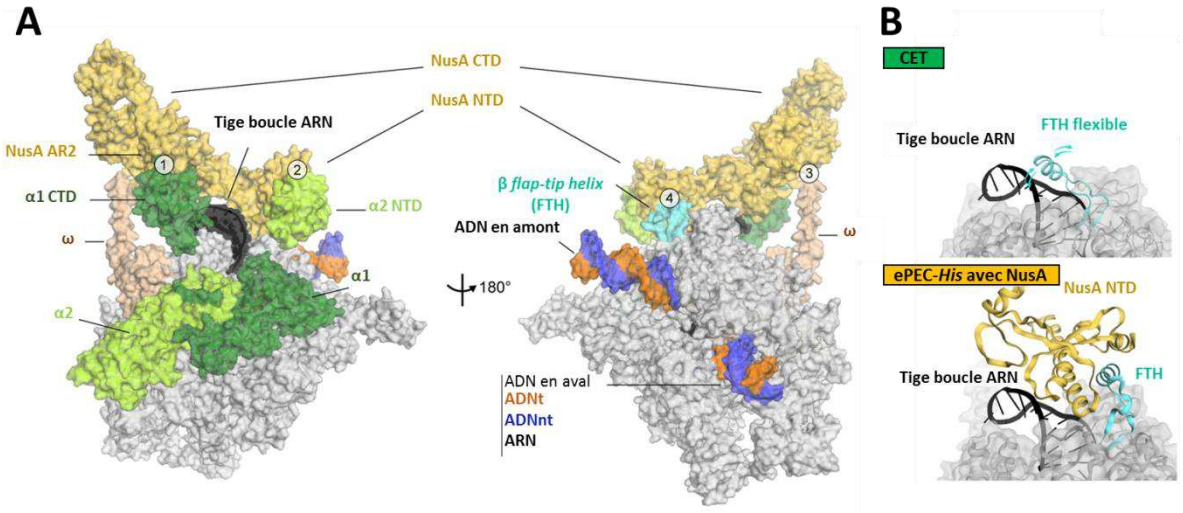


Figure 18 : Représentation de la liaison entre NusA et le CET. (A) Interaction entre NusA et le ePEC (pour l'opéron *his* d'*E. coli*, (PDB: 6FLQ). Le facteur NusA effectue quatre interactions avec le ePEC-*his* : ① ARNP-α^1-CTD:NusA-AR2, ② ARNP-α^2-CTD:NusA-NTD, ③ ARNP-ω:NusA-KH1/KH2, ④ ARNP-FTH:NusA-NTD. **(B)** Comparaison du domaine FTH (cyan) de la sous-unité β de l'ARNP dans les structures du CET et de l'ePEC lié à NusA. Dans le CET la partie FTH est en général flexible. La liaison de NusA-NTD au FTH va stabiliser celui-ci dans une position distale par rapport à la tige-boucle ARN. Figure issue de (Guo et al., 2018).

Le facteur NusG (**Figure 9**), en plus de permettre le couplage de la transcription et de la traduction (voir partie B-I-2, page 13) (**Figure 8**), favorise les mouvements du CET le long de l'ADN en stimulant le passage dans l'état de « *post-translocation* » (**Figure 6**). Ainsi, NusG-NTD diminue le taux de pause élémentaire du CET (ePEC), et par conséquent le taux de pauses prolongées (Herbert et al., 2010; Mooney et al., 2009b). Des analyses récentes ont mis en évidence que NusG stabilise à la fois l'appariement de base du duplex ADN, en amont de la bulle de transcription, pour éviter les phénomènes de « *backtracking* » (Kang et al., 2018) et aussi les interactions de l'ARNP avec l'ADN durant la phase d'élongation (Svetlov and Nudler, 2011). Tout comme NusG, le facteur paralogue RfaH (**Figure 9**) peut favoriser les mouvements productifs du CET (de façon *osp*-dépendante) en limitant les phénomènes de « *backtracking* » (Burmann et al., 2012). Les facteurs NusA, NusG et RfaH sont aussi impliqués dans la régulation de la transcription Rho-dépendante, un point que j'aborde plus loin dans la partie B-III (pages 55-56).

Les protéines paralogues GreA et GreB permettent de réactiver les CET arrêtés par *rétro-translocation* (Esyunina et al., 2016; Roberts et al., 2008; Washburn and Gottesman, 2015). Elles sont organisées en deux domaines : un domaine N-ter en hélice qui est responsable de l'activité « anti-arrêt » et un domaine C-ter qui assure l'interaction avec l'ARNP (**Figure 19A**) (Borukhov et al., 2005; Opalka et al., 2003; Stebbins et al., 1995;

Vassylyeva et al., 2007). Ces facteurs se lient à l'ARNP au niveau du canal secondaire (**Figure 19B-C**), où ils vont couper la partie 3' extrudée du transcrit ARN (par activité endonucléolytique) permettant ainsi le réalignement de l'extrémité 3' avec le site catalytique de l'ARNP et la reprise de la transcription. Le facteur GreA coupe préférentiellement de courts segments d'ARN extrudé (2-3 nt) et aurait plutôt un rôle de prévention des phases précoces de rétrotranslocation alors que GreB peut couper des segments allant jusqu'à 18 nucléotides (Borukhov et al., 2005; Rutherford et al., 2007). Ces protéines favorisent également la transition du complexe ouvert en complexe d'élongation en réduisant le taux d'initiation abortive (partie B-I-1) et participent au contrôle de la fidélité de la transcription (Erie et al., 1993; Hsu et al., 1995; Opalka et al., 2003).

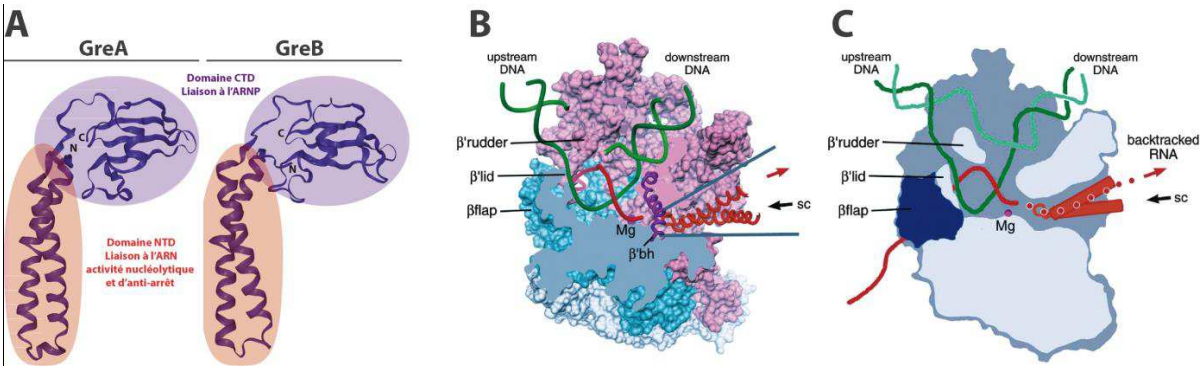


Figure 19 : Facteurs GreA et GreB. (A) Les facteurs GreA et GreB ont une organisation structurale et un fonctionnement très similaires (PDB [GreA] : 1GRJ, PDB [GreB] : 2P4V). **(B)** et **(C)** CET arrêté par rétrotranslocation avec GreB dans le canal secondaire (sc) de l'ARNP. Figure inspirée de (Opalka et al., 2003; Stebbins et al., 1995; Vassylyeva et al., 2007).

3) Terminaison de la transcription

a. Terminaison intrinsèque

La terminaison de la transcription met fin à la synthèse de l'ARN par le démantèlement irréversible du CET. Deux catégories de signaux de terminaison sont rencontrées chez la bactérie (**Figure 17 ③ & ④**). La première est la terminaison intrinsèque (Rho-indépendante) qui nécessite la présence d'une séquence ADN spécifique codant pour une structure ARN en tige-boucle localisée en amont d'une séquence de 7 à 8 nt, riches en uracile (U), à l'extrémité 3' du transcrit naissant (**Figure 17 ③ et Figure 20A**) (Gusarov and Nudler, 2001; Komissarova et al., 2002; Penno et al., 2015; Yarnell and Roberts, 1999). Les terminateurs intrinsèques sont souvent retrouvés à la fin ou entre les gènes des opérons

pour permettre la régulation de l'expression génique (pour revue : (Le et al., 2018)). La séquence U-riche du terminateur conduit à la formation d'un hybride rU/dA peu stable au sein du CET qui favorise un état de pause (Toulokhonov and Landick, 2003). Cette pause laisse le temps à la structure en tige-boucle de se former dans le canal de sortie de l'ARNP, un évènement conduisant à la déstabilisation irréversible du CET (Burmann et al., 2012; Gusarov and Nudler, 2001; Komissarova et al., 2002; Yarnell and Roberts, 1999). Différents modèles de déstabilisation ont été proposés (Peters et al., 2011; Washburn and Gottesman, 2015) qui, dans une revue bibliographique récente (Ray-Soni et al., 2016), sont déclinés en deux versions principales:

- Le modèle « *Hybrid Shearing* », propose que la formation de la structure tige-boucle dans le canal de sortie de l'ARNP « tire » sur la portion aval de la chaine ARN, détruisant ainsi l'hybride ARN:ADN et la bulle de transcription. Dans ce modèle la déstabilisation irréversible du CET ne nécessite pas de mouvement de l'ARNP le long de l'ADN (**Figure 20B-C**) (Larson et al., 2008).
- Le modèle par « *hyper-translocation* » prévoit, au contraire un mouvement de l'ARNP vers l'avant (de 2 à 4 pb) provoqué par la formation de la structure tige-boucle. Ce mouvement conjugué à l'absence d'addition de nucléotide à la chaine ARN conduit à une réduction progressive de la taille de l'hybride ARN:ADN et à la déstabilisation du CET (**Figure 20B-C**) (Larson et al., 2008; Santangelo and Roberts, 2004).

Il est important de noter que, quel que soit le modèle, l'inactivation et la déstabilisation du CET impliquent divers changements allostériques (Toulokhonov et al., 2001; Toulokhonov and Landick, 2003; Toulokhonov et al., 2007), qui peuvent eux-mêmes être sujets à régulation (Larson et al., 2008).

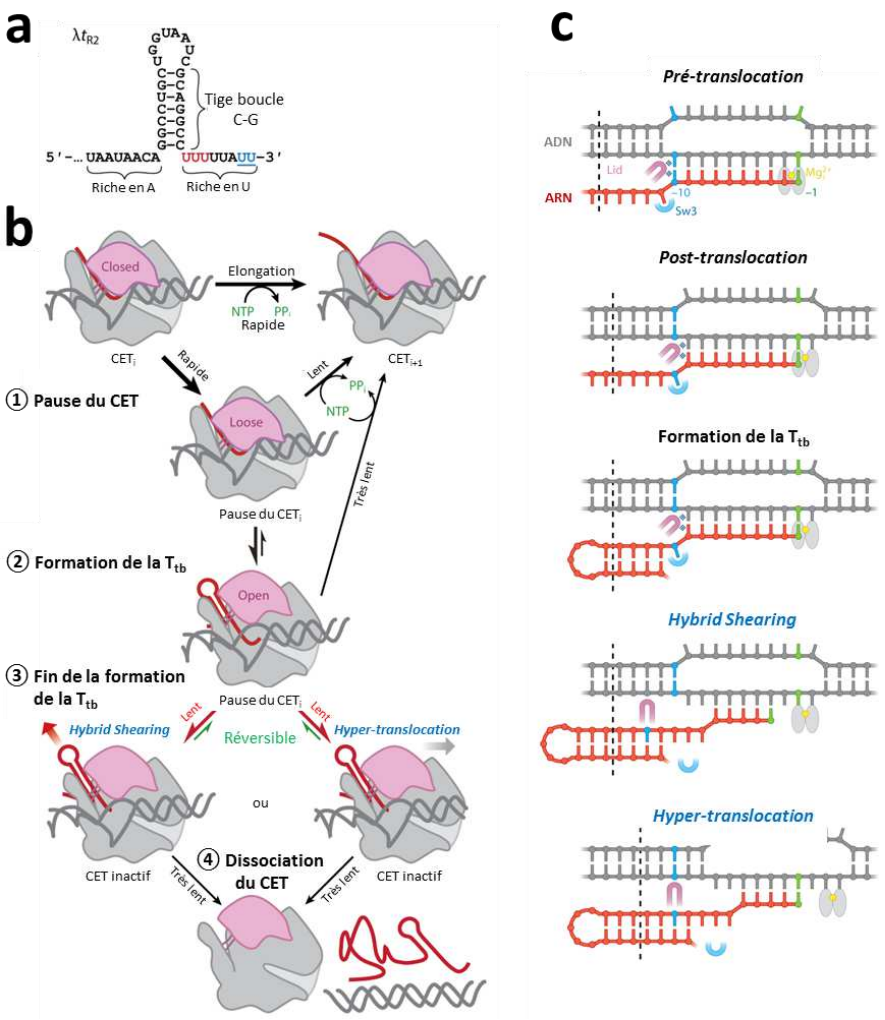


Figure 20: Terminaison intrinsèque de la transcription. (A) Exemple de terminateur intrinsèque (terminateur *λtr2*). La structure ARN en tige-boucle précède une séquence 3'-terminale riche en Uraciles. Les résidus en rouge sont ceux déshybridés en premier de l'hélice ARN:ADN tandis que ceux en bleu représentent les points préférentiels de terminaison. **(B)** Diagramme simplifié du processus de terminaison intrinsèque. ① Le CET marque une pause à cause de la séquence riche en Uraciles en partie 3' de l'ARN naissant (en rouge). Le domaine « *Clamp* » mobile de l'ARNP est représenté en rose. ② La pause du CET permet la formation de la structure tige boucle (T_{tb}) dans le canal de sortie de l'ARN et le repositionnement du domaine « *Clamp* » en configuration ouverte. Cette étape engage la fusion des paires rU:dA dans la partie amont de l'hydride ARN:ADN. ③ La fusion complète de l'hydride ARN:ADN est proposée être due à un effet de levier de l' « *hairpin* » sur le « *Clamp* » qui va soit tirer la chaine ARN dans le canal de sortie sans mouvoir l'ARNP le long de l'ADN (modèle « *Hybrid Shearing* ») soit, au contraire, pousser l'ARNP le long de l'ADN. Dans ce dernier cas, la chaine ARN ne serait pas rallongée simultanément en raison d'une configuration inadaptée du site catalytique de l'ARNP, conduisant au raccourcissement progressif de l'hydride ARN:ADN (modèle de terminaison par « *hyper-translocation* »). ④ Le processus s'achève par la dissociation du CET. **(C)** Réseau d'interactions ARN/ADN dans la bulle de transcription selon les étapes. Figure adaptée de (Ray-Soni et al., 2016).

b. Terminaison facteur-dépendante

Le second type de terminaison de la transcription nécessite la présence d'au moins un cofacteur protéique tel que Rho, Mfd, ou CodY (terminaison facteur-dépendante). Le facteur

CodY est présent uniquement chez certaines espèces Gram$^+$ comme *Bacillus subtillis* (Stenz et al., 2011). Son mécanisme d'action a été peu étudié mais semble impliquer la formation d'un complexe avec l'ADN qui fait obstacle à l'avancement du CET (Stenz et al., 2011). Les facteurs Rho et Mfd sont eux beaucoup plus conservés dans le règne bactérien (D'Heygere et al., 2013; Roberts and Park, 2004).

Le facteur protéique Mfd (130 kDa) a été décrit initialement pour son rôle dans la reconnaissance des ARNP bloquées par un dommage à l'ADN (**Figure 21**) dans le contexte de la réparation associée à la transcription (*Transcription Coupled Repair* [TCR]) (Portman and Strick, 2018). Dans ce cas, Mfd se lie simultanément à la sous-unité β de l'ARNP bloquée et à un segment d'ADN situé environ 20 pb en aval. L'action de Mfd n'est possible que si le facteur σ a été relargué après la conversion en complexe d'élongation (Park et al., 2002), ce qui n'est parfois pas complètement le cas (voir partie B-II-1, pages 9). Après fixation, Mfd dissocie le CET arrêté (Graves et al., 2015; Park and Roberts, 2006) et recrute les facteurs protéiques UvrA et UvrB de la voie NER (*Nucleotide Excision Repair*) de réparation (**Figure 21**) (pour revue : (Pani and Nudler, 2017; Portman and Strick, 2018)). Une étude *in vitro* a montré que Mfd pouvait former un complexe avec l'ARNP avançant à environ 4 pb/s sur l'ADN pour éloigner l'ARNP du site du dommage et permettre le relargage rapide du transcrit (Graves et al., 2015). Le processus de terminaison de la transcription induit par Mfd s'acheverait par la dissociation de l'ARNP (**Figure 21**).

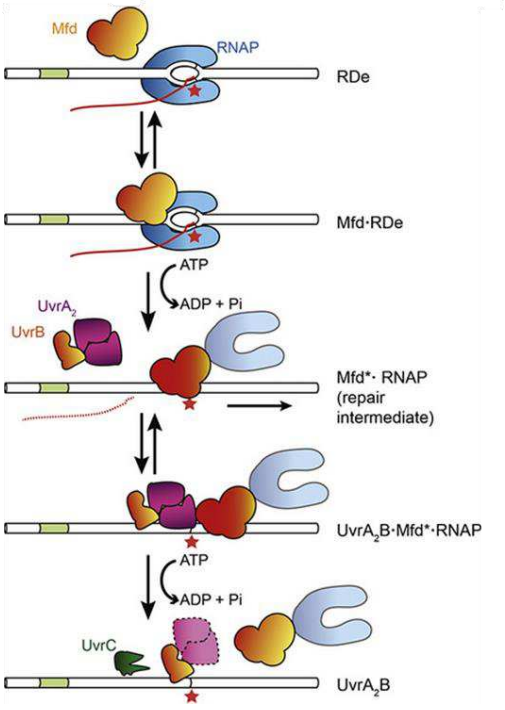


Figure 21 : Mécanisme de terminaison de la transcription induite par Mfd. Mfd est recruté au niveau d'une ARNP bloquée par une lésion de l'ADN (★). Mfd subit des réarrangements structuraux lui permettant de recruter le complexe $UvrA_2B$ et de transloquer l'ARNP vers l'avant, une activité ATPase-dépendante. Cette translocation, qui n'est pas accompagnée de l'allongement concomitant du transcrit, conduit à un effondrement de la bulle de transcription et au relargage du transcrit. La dissociation du complexe Mfd:ARNP permet au complexe $UvrA_2B$ (avec UvrC) d'initier le processus de réparation du dommage à l'ADN. Figure issue de (Portman and Strick, 2018).

Diverses études suggèrent que le rôle de Mfd n'est pas limité au TCR et s'étend à la résolution d'autres phénomènes d'arrêts transcriptionnels, en particulier ceux induits par des protéines liées à

l'ADN (Le et al., 2018) ou au transcrit naissant (Park et al., 2002). Dans ces cas, Mfd va aider le CET à passer l'obstacle (« *transcription through the roadblock* ») ou bien induire sa dissociation. Mfd a notamment été impliqué dans la résolution des conflits entre CET et réplisomes (Park et al., 2002). Chez *E. coli*, Mfd est beaucoup moins abondant que les ARNP et semble utiliser un mécanisme de reconnaissance originale des CET arrêtés pour pallier ce désavantage. En effet, des expériences récentes de nanomanipulation à l'échelle de la molécule unique ont montré que Mfd était capable de transloquer seul sur l'ADN avec une processivité d'environ 200 pb et une vitesse de 7 pb/s (Le et al., 2018). Ces paramètres, plus faibles que ceux d'un CET fonctionnant normalement (vitesse de 14 pb/s dans les mêmes conditions d'étude), permettraient à Mfd de patrouiller l'ADN en continu à la recherche de CET arrêtés suivant un mécanisme qualifié de « *release and catch-up* » (Le et al., 2018). Il est démontré dans la même étude que Mfd est capable, du moins *in vitro*, de s'opposer à la rétrotranslocation du CET ou bien de réaligner (et réactiver) des CET séverement rétrotransloqués. Ceci aiderait le CET à passer les obstacles, sauf lorsque ceux-ci sont trop résistants, auquel cas Mfd induirait la dissociation du CET (Le et al., 2018). Cette activité presente une certaine forme de redondance avec celles de GreA/B, NusG, RfaH et Rho. Les facteurs NusG, RfaH et Mfd auraient plutôt des rôles de prévention de la rétrotranslocation à l'inverse de GreA/B et Rho qui « géreraient » plus en aval les CET arrêtés et les problèmes qui en découlent (formation de structures « R-loops », par exemple).

Chez *Bacillus subtilis*, des signaux de terminaison ressemblant aux terminateurs intrinsèques (structure tige-boucle suivie d'une séquence riche en uraciles) mais n'étant pas capables, seuls, d'induire la dissociation du CET ont été identifiés (Mondal et al., 2016; Potter et al., 2011). Ces terminateurs « sous-optimaux » contiennent généralement moins de résidus uracyles dans la section ARN 3'-proximale que les terminateurs intrinsèques canoniques et requièrent la participation du cofacteur NusA (Mondal et al., 2016; Yakhnin and Babitzke, 2002). Des terminateurs intrinsèques sous-optimaux, cette fois stimulés par la présence de NusG, ont également été identifiés chez les mycobactéries (Czyz et al., 2014). Ces terminateurs sont également caractérisés par une section ARN 3'-proximale appauvrie en uraciles (Czyz et al., 2014). Ces données démontrent que la frontière entre terminateurs intrinsèques et terminateurs facteur-dépendants n'est pas toujours nette. Elles illustrent également l'importance des facteurs NusA et NusG comme régulateurs des signaux de

l'élongation de la transcription. Comme nous le verrons dans le chapitre suivant, ces facteurs sont également capables de fortement réguler la terminaison Rho-dépendante. De façon intrigante, cette régulation semble également s'exercer parfois sous la forme d'une stimulation (voire d'une activation) de terminateurs Rho-dépendants « sous-optimaux ».

III. Terminaison de la transcription Rho-dépendante

La terminaison de la transcription Rho-dépendante, sur laquelle porte ma thèse, est un mécanisme spécifique aux bactéries. Le facteur Rho est très largement retrouvé au sein du règne bactérien. Ainsi, une étude récente a montré que seulement 8% des génomes bactériens séquencés étaient dépourvus de gène(s) codant pour Rho (D'Heygere et al., 2013). Les espèces dépourvues de facteur Rho incluent certains Firmicutes (des classes : *Clostridia*, *Bacilli* et *Negativicutes*) ainsi que toutes les Cyanobacteries et Mollicutes. Chez certaines espèces, le gène *rho* est dupliqué, l'une des copies ayant souvent subi une évolution divergente de sa séquence (D'Heygere et al., 2013). C'est le cas, par exemple, de *Streptomyces* (D'Heygere et al., 2013) dont certaines espèces produisent naturellement la bicyclomycine, le seul inhibiteur connu du facteur Rho (Zwiefka et al., 1993). Il est possible que la copie divergente du gène *rho* rende ces souches résistantes à la bicyclomycine (D'Heygere et al., 2013). Le groupe de gènes responsables de la biosynthèse de la bicyclomycine a été identifié récemment (Patteson et al., 2018) et, de façon surprenante, a été retrouvé chez des espèces phylodivergentes comme *Pseudomonas aeruginosa* qui ne possèdent qu'une copie du gène *rho* (D'Heygere et al., 2013). Quel est l'intérêt pour l'hôte de ce transfert horizontal et comment se protège-t-il de la production endogène de bicyclomycine restent deux questions ouvertes à ce jour.

Bien que présent chez de nombreuses espèces bactériennes, Rho n'est pas essentiel chez toutes. Les premières études, basées sur les spectres d'action de la bicyclomycine, suggéraient que Rho est essentiel chez de nombreuses espèces à Gram-négatif et superflu chez les espèces à Gram-positif (D'Heygere et al., 2013; Rabhi et al., 2010b). Cette classification a depuis été nuancée par l'observation que Rho était essentiel chez des espèces Gram positif comme *Micrococcus luteus* (Nowatzke et al., 1997) ou *Mycobacterium tuberculosis* (Botella et al., 2017). De plus, le facteur Rho impacte profondément le transcriptome d'espèces chez qui le mutant *rho*$^-$ est viable comme *Bacillus subtilis* (Bidnenko et al., 2017) ou *Staphylococcus aureus* (Mader et al., 2016). L'importance de Rho chez *E. coli* a été attribuée à son rôle inhibiteur de l'expression du gène toxique *kil* présent dans le prophage *Rac* (Briani et al., 2000).

Les études détaillées de la terminaison Rho-dépendante ont été menées, en grande majorité, chez la bactérie modèle *E. coli*. Certaines « règles » déduites de ces études doivent donc être considérées et généralisées avec prudence. De ces études, il apparaît que la terminaison Rho-dépendante a plusieurs fonctions importantes :

- Un rôle de « ponctuation » de l'expression génique, en terminant la transcription de gènes (ou d'opérons) dépourvus de terminateurs intrinsèques dans leur partie 3'UTR (Richardson and Greenblatt, 1996).
- Un rôle de « *silencing* » de la transcription pervasive, en particulier de la transcription antisens (Peters et al., 2012).
- Un rôle de « surveillance » du couplage transcription-traduction, en induisant la terminaison des gènes incorrectement traduits (Boudvillain et al., 2013; Richardson, 1991).

Ces fonctions ont été largement confirmées par des études transcriptomiques menées chez *E. coli* (Dar and Sorek, 2018; Peters et al., 2012; Peters et al., 2009; Sedlyarova et al., 2016), *B. subtilis* (Bidnenko et al., 2017; Nicolas et al., 2012), *S. Aure*us (Mader et al., 2016), et *M. Tuberculosis* (Botella et al., 2017). L'implication de Rho dans des mécanismes de régulation conditionnels variés (riboswitches, contrôle par sRNAs, etc. ; voir également la partie III-5) a également été démontrée chez plusieurs espèces comme *E. coli*, *Salmonella* (Bastet et al., 2017; Bossi et al., 2012; Brandis et al., 2016; Chauvier et al., 2017; Figueroa-Bossi et al., 2014; Gall et al., 2016; Gall et al., 2018; Kriner and Groisman, 2015; Sedlyarova et al., 2017; Sedlyarova et al., 2016) ou *Corynebacterium glutamicum* (Takemoto et al., 2015).

Chez *E. coli*, Rho a également été impliqué dans :

- L'élimination des structures « *R-loop* » transcriptionnelles (voir page 21) et des CET arrêtés, prévenant ainsi les risques de conflit entre machineries de transcription et de réplication (Harinarayanan and Gowrishankar, 2003; Leela et al., 2013).
- La « protection » (comme un système immunitaire) contre l'expression de gènes de prophages ou d'ADN xénogénique acquis par transfert horizontal et potentiellement toxiques (Cardinale et al., 2008a; Menouni et al., 2013).

- La machinerie de dégradation de l'ARN (dégradosome) dans des conditions bien particulières de culture (Jager et al., 2004). Si cette implication est (au mieux) ponctuelle chez *E. coli*, Rho semble en revanche être toujours associé au dégradosome chez *Rhodobacter capsulatus* (Jager et al., 2001).

1) Organisation structurale et activités biochimiques de Rho

L'activité de terminaison de la transcription Rho-dépendante découle de la capacité du facteur Rho à se lier au transcrit naissant au niveau d'un site *Rut*, à transloquer « directionnellement » le long de celui-ci de façon ATP-dépendante et à dissocier le CET temporairement immobilisé à un site de pause (**Figure 1**) (Boudvillain et al., 2010a).

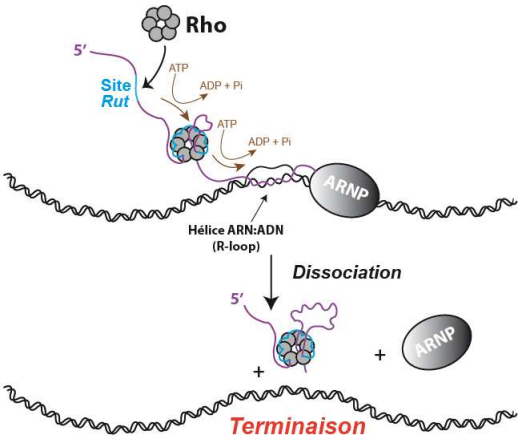


Figure 1 (rappel partiel)

Cette activité de moteur moléculaire, ATP- et ARN-dépendante, lui permet également de dissocier des doubles hélices ARN:ARN et ARN:ADN, du moins *in vitro* (Brennan et al., 1987; Brennan et al., 1990; Schwartz et al., 2007; Walmacq et al., 2004, 2006). Rho est, de ce fait, considéré comme une ARN hélicase mais adopte une organisation structurale en anneau homo-hexamèrique (Gogol et al., 1991) inhabituelle pour cette classe d'enzymes (alors qu'elle est fréquemment rencontrée pour les ADN hélicases) (Singleton et al., 2007).

Chaque monomère Rho est constitué d'une partie C-terminale (CTD) très conservée contenant les motifs essentiels pour l'oligomérisation, l'activité de moteur (motifs ATPase) et de translocation de l'ARN (via le site secondaire **SBS** : *Secondary Binding Site*) et d'une partie N-terminale (NTD), plus variable suivant les espèces (D'Heygere et al., 2013) et

contenant les motifs de reconnaissance initiale de la chaine ARN (formant le site primaire **PBS** : *Primary Binding Site*) (**Figure 22A**) (Geiselmann et al., 1992; Gogol et al., 1991; Skordalakes and Berger, 2003; Thomsen and Berger, 2009; Yu et al., 2000).

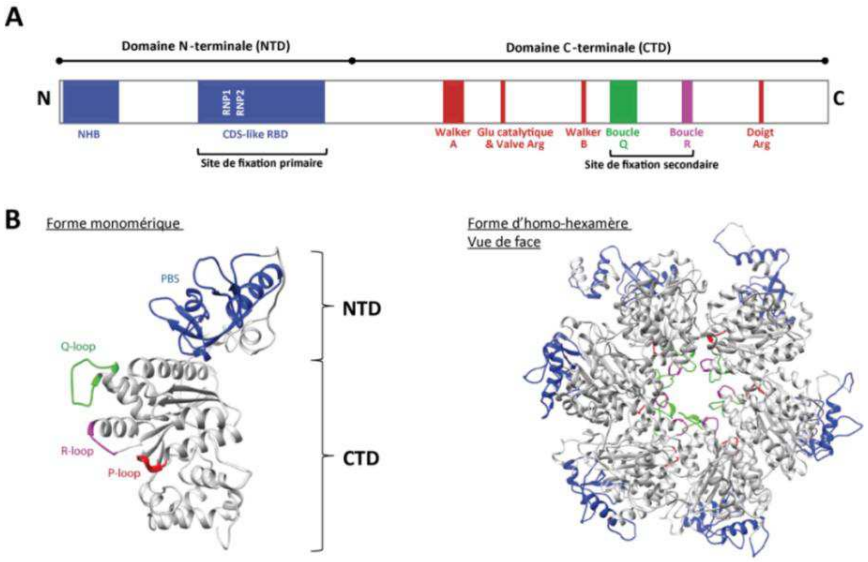


Figure 22 : Organisation structurale du facteur Rho. (A) Représentation schématique des motifs constituant chaque monomère Rho. **(B)** Structure cristalline de l'homohéxamère Rho (PDB 3ICE). Une vue transversale d'un monomère isolé est proposée sur la gauche de la figure. Figure adaptée de (Mitra et al., 2017).

Le site PBS est essentiel pour l'accrochage initial du facteur Rho au transcrit naissant. Il est constitué par six poches (une par monomère) de reconnaissance spécifique de dimères de pyrimidines 5'-CC ou 5'UC (5'YC) (Bogden et al., 1999) formant ainsi une sorte de couronne sur le dessus de l'hexamère Rho (**Figure 22B**). Le PBS inclut également les domaines **NHB** (*N-terminal Helix Bundle*) (**Figure 22**) qui pourraient être impliqués dans le recrutement non spécifique (électrostatique) de l'ARN et le guidage vers les poches de reconnaissance des dimères 5'YC (Canals et al., 2010). Cette reconnaissance n'implique pas les groupements 2'-OH de l'ARN (Bogden et al., 1999; Skordalakes and Berger, 2003) de telle sorte que le PBS accepte aussi bien des substrats ARN qu'ADN pourvu qu'ils soient « simple-brin » (ou pauvres en structures secondaires) et riches en (désoxy)cytidines (Lowery-Goldhammer and Richardson, 1974).

Le site SBS est responsable de la translocation de l'ARN au centre de l'anneau héxamérique. Il est constitué par les motifs « **Q-loop** » et « **R-loop** » qui contactent l'ARN essentiellement au niveau du squelette phosphodiester et des groupements 2'-OH (Schwartz et al., 2009; Soares et al., 2014; Thomsen and Berger, 2009), conférant ainsi à Rho sa spécificité pour l'ARN (**Figure 23**). Des expériences d'empreinte (Fe-EDTA/H_2O_2) ont en effet

montré que l'ADN simple-brin n'était pas capable d'interagir avec le SBS (Richardson, 1982; Wei and Richardson, 2001), ce qui explique également pourquoi l'ADN n'est pas capable d'activer Rho (Schwartz et al., 2009; Wang and von Hippel, 1993). Le SBS est connecté allostériquement aux sites ATPase qui sont composés des motifs **Walker A** (également appelé « *P-loop* ») et **B** et sont localisés à l'interface entre monomères (**Figure 22**). La coordination entre les six sites ATPase est assurée par les résidus conservés « **doigt arginine** », « **Glutamate catalytique** » et « **valve arginine** » (**Figure 22**) (Bogden et al., 1999; Gogol et al., 1991; Richardson, 1982; Skordalakes and Berger, 2003, 2006; Thomsen and Berger, 2009).

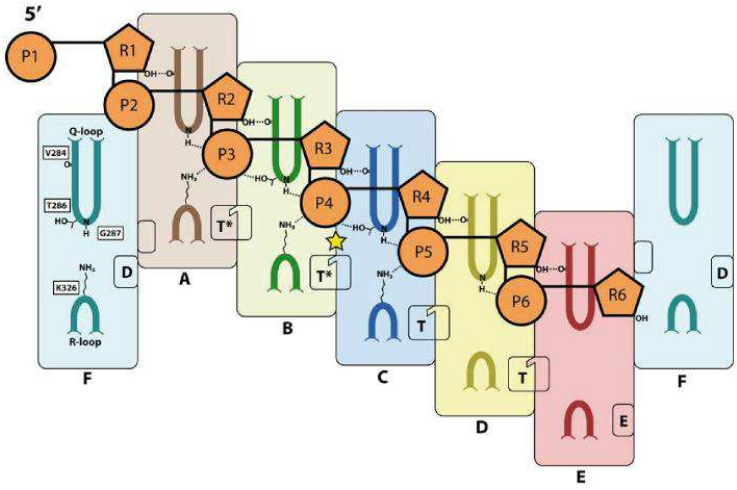


Figure 23 : Interactions ARN:SBS Les monomères de Rho sont représentés par des grands rectangles colorés où seuls les boucle Q et R sont représentés. Les poches ATPases à l'interface entre deux monomères sont représentés par des rectangles, ayant une lettre à l'intérieur qui symbole l'état du cycle ATPase (E : échange de nucléotides, T : fixation de l'ATP, T* : hydrolyse de l'ATP, D : produits d'hydrolyse). L'ARN est représenté en orange (R_i : ribose, P_i : phosphate) dont les nucléotides sont numérotés suivant leurs positions dans la chaine. Figure issue de (Thomsen and Berger, 2009).

Un mécanisme d'activation du facteur Rho impliquant la formation de l'hexamère (à partir de monomères ou d'autres formes de basse composition oligomérique) directement sur l'ARN n'est pas compatible avec les données expérimentales disponibles. Par exemple, la première structure cristalline de l'hexamère Rho montrait l'anneau dans une forme ouverte (**Figure 24**) au sein de laquelle les sites ATPase ne sont pas correctement configurés pour la catalyse (Skordalakes and Berger, 2003). Une configuration ouverte de l'anneau a également été observée lors d'expériences de microscopie électronique (Gogol et al., 1991; Yu et al., 2000) ou de SAXS (Thomsen et al., 2016) réalisées en absence d'ARN. En fait, la présence d'un ligand ARN capable de se fixer au SBS semble favoriser la forme fermée de l'anneau

hexamèrique (**Figure 24**) (Thomsen and Berger, 2009; Thomsen et al., 2016). Ces observations complétées de nombreuses expériences biochimiques (pour revue : (Ray-Soni et al., 2016)) suggèrent un mécanisme séquentiel d'activation du facteur Rho impliquant la fixation du transcrit au PBS, l'entrée de la chaine ARN au centre de l'anneau ouvert, et sa fixation au SBS provoquant la fermeture de l'anneau et une reconfiguration catalytiquement correcte des sites ATPase (**Figure 24**) (Thomsen et al., 2016). *In vitro*, ce processus complexe d'activation est lent et cinétiquement limitant (Walmacq et al., 2004). Ce processus d'activation peut être sujet à régulation, notamment par la protéine Hfq (Rabhi et al., 2011a) ou par un motif ARN présent en *cis* appelé RARE (*Rho-Antagonizing RNA Element*) (Sevostyanova and Groisman, 2015) (voir page 60).

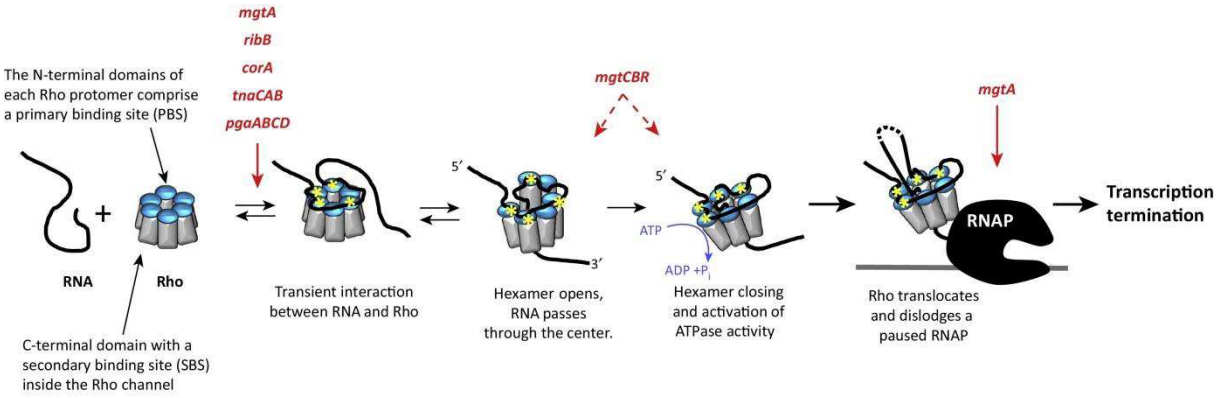


Figure 24 : Etapes principales du processus de la terminaison de la transcription Rho-dépendante. Figure issue de (Kriner et al., 2016).

2) Architecture des terminateurs Rho-dépendants

Les terminateurs Rho-dépendants sont constitués de séquences beaucoup plus longues et moins bien conservées que les terminateurs intrinsèques. On distingue traditionnellement deux régions, la région *Rut* (« *Rho utilization* ») encodant le motif de fixation de Rho au transcrit et la région *tsp* (« *termination stop points* ») plus aval et contenant les sites où le CET est dissocié par Rho (**Figure 25**) (Richardson and Richardson, 1996). Ces deux régions peuvent parfois se recouvrir partiellement.

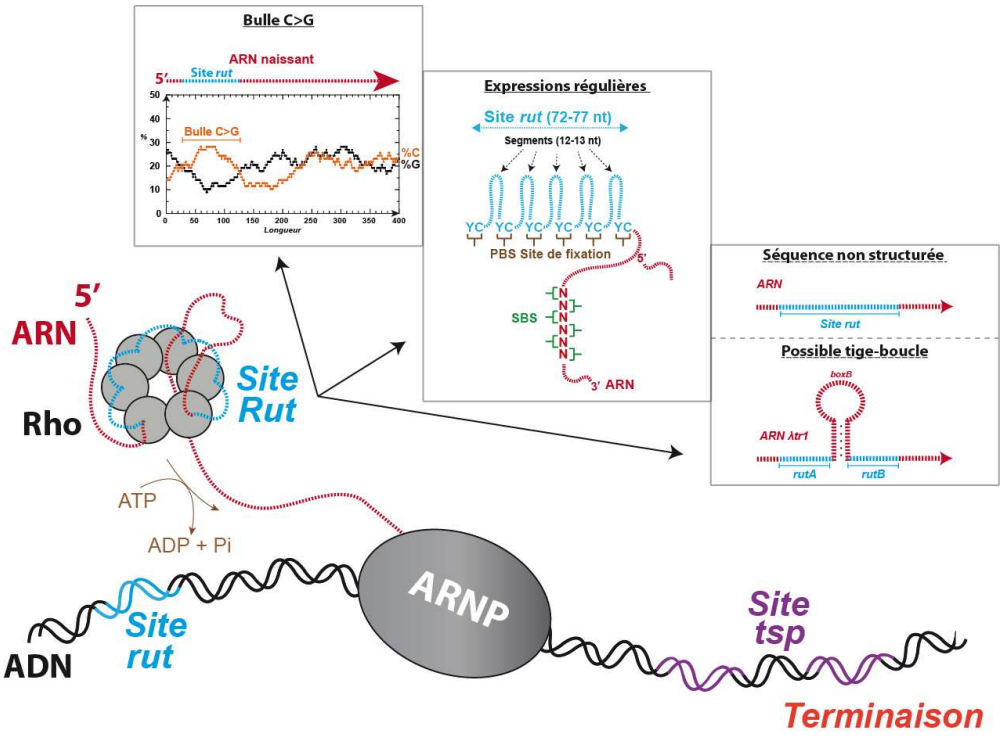


Figure 25 : Fixation du facteur Rho au transcrit naissant au niveau d'un terminateur Rho-dépendant. Le site *Rut* est généralement caractérisé par la présence d'une « bulle C>G » au sein du brin ADNn-t traduisant une richesse en dinucléotides 5'YC, non aparriés et adéquatement espacés (segment ARN simple-brin de 9-13 nt ou structure tige-boucle) pour une interaction avec le PBS.

En 1991, Alifano et ses collaborateurs suggèrent que les sites *Rut* présentent systématiquement une richesse en Cytidines (C) et une pauvreté en Guanosines (G) illustrées par la présence de « bulles C>G » lorsque les pourcentages en C et G sont calculés pour une fenêtre « glissante » de 78 pb le long du brin ADN non-matrice (**Figure 25**) (Alifano et al., 1991). Cette proposition, basée sur l'analyse des quelques terminateurs connus à l'époque, est compatible avec de nombreuses données biochimiques, biophysiques, et structurales (pour revue : (Boudvillain et al., 2010a; Ciampi, 2006; Peters et al., 2011; Rabhi et al., 2010b; Ray-Soni et al., 2016)). Par exemple, les structures cristallines de Rho ont révélé la présence de poches d'interaction spécifique 5'YC au sein du PBS (voir paragraphe précédent) (Bogden et al., 1999; Skordalakes and Berger, 2003) expliquant la préférence pour les séquences riches en C et soulignant la nécessité d'avoir des dimères 5'YC non-appariés et périodiquement espacés (**Figure 25**). Des données RMN suggèrent que d'autres contacts PBS pourraient également être formés préférentiellement avec des Cytidines en dehors des poches 5'YC (Hitchens et al., 2006) mais ces contacts ne sont pas résolus dans les structures cristallines de Rho. La distance séparant les poches 5'YC (entre deux protomères adjacents) dans les structures cristallines correspondrait à un segment d'ARN simple brin de 12 à 13 nt

(Koslover et al., 2012; Skordalakes and Berger, 2003). Ainsi, il faudrait au minimum 72 à 77 nt d'ARN simple brin pour couvrir complètement la couronne du PBS (McSwiggen et al., 1988). Des grandeurs comparables ont été déduites d'analyses biochimiques (Skordalakes and Berger, 2003) et biophysiques (Koslover et al., 2012) même si parfois la distance inter-monomère proposée est plus courte (9-10 nt). Cette distance inter-monomère (ou plus exactement inter-poches 5'YC) semble, dans certains cas, pouvoir être couverte avantageusement par une structure ARN en tige-boucle (Schwartz et al., 2007; Vieu and Rahmouni, 2004a). C'est le cas, par exemple, de la tige-boucle *boxB* qui sépare le site *Rut* du terminateur tR1 du phage lambda (*λtR1*) en deux parties, *rutA* et *rutB* (**Figure 25**) (Vieu and Rahmouni, 2004b).

Une analyse récente du transcriptome Rho-dépendant chez *E. coli* a confirmé la présence d'un biais C>G à proximité des sites de terminaison mais n'a pas révélé d'autres éléments consensus de séquence (Peters et al., 2012). Il ne peut toutefois pas être exclu que cette analyse ait été « bruitée » par une localisation peu précise des sites de terminaison due aux modifications post-transcriptionelles des transcrits par les exoribonucléases (Dar and Sorek, 2018; Peters et al., 2012).

La région *tsp* contient l'ensemble des sites où Rho est capable de dissocier le CET et, de ce fait, délimite la fenêtre d'opportunité pour la terminaison (**Figure 25**). Il est proposé que chaque site de relargage correspond à un site potentiel de pause du CET (Lau et al., 1982) et des caractéristiques rappelant les sites de pause de type I (structure en tige-boucle ; voir page 21) ont été parfois retrouvées (Lau et al., 1983; Morgan et al., 1983). Néanmoins, il ne semble pas y avoir de corrélation directe entre l'efficacité ou la durée de la pause du CET à ces sites et l'efficacité de la terminaison Rho-dépendante (Richardson and Richardson, 1996). Une hypothèse alternative est que ces sites de pause favoriseraient un réarrangement structural de l'ARNP facilitant le démantèlement du CET par le facteur Rho (Artsimovitch and Landick, 2000).

3) Mécanisme de la terminaison Rho-dépendante

En 1992, Jin et ses collaborateurs ont établi que l'efficacité de la terminaison Rho-dépendante pouvait être modulée par des mutations de Rho ou de l'ARNP susceptibles

d'affecter la vitesse de translocation de ces enzymes (Jin et al., 1992). Des résultats similaires ont été obtenus en faisant varier la concentration en nucléotides (et donc la vitesse de translocation de l'ARNP) (Jin et al., 1992). Ces données révèlent l'existence d'un couplage cinétique entre Rho et l'ARNP qui contraint la fenêtre d'opportunité pour la terminaison Rho-dépendante. Cette fenêtre est également dépendante de la processivité de Rho qui semble assez limitée, du moins *in vitro*, autour de 60 à 80 nt (Gocheva et al., 2015; Walmacq et al., 2004).

Des expériences de nanomanipulation à l'échelle de la molécule unique ont démontré que Rho était capable de transloquer sur l'ARN à une vitesse d'environ 56 nt/s, soit 2 à 5 fois plus vite que l'ARNP, le long de l'ADN dans les mêmes conditions (Gocheva et al., 2015). Ces expériences ont également confirmé que Rho transloquait l'ARN suivant un mécanisme de « *tethered tracking* » (Gocheva et al., 2015; Koslover et al., 2012). Ce mécanisme, initialement déduit d'expériences biochimiques astucieuses, stipule que Rho conserve son interaction initiale avec le site *Rut* en cours de la translocation (Steinmetz and Platt, 1994). Cela est rendu possible par la séparation entre les sites dédiés à la reconnaissance initiale (PBS) et à la translocation (SBS) du transcrit et se traduit par la formation d'une boucle ARN entre les deux sites qui grandit en cours de translocation (**Figure 26**) (Gocheva et al., 2015; Koslover et al., 2012; Soares et al., 2014; Steinmetz and Platt, 1994).

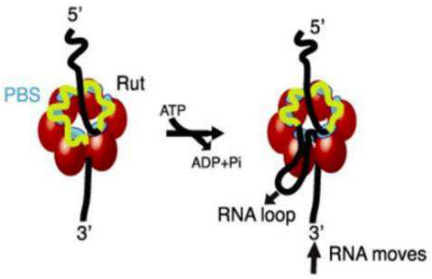


Figure 26 : Mécanisme de « *tethered tracking* ». Figure issue de (Gocheva et al., 2015).

Les structures les plus récentes de l'hexamère Rho suggèrent que le site SBS transloque l'ARN suivant un mécanisme dit « d'escorte » où chaque nucléotide est successivement pris en charge par une sous-unité qui lui fait traverser le canal central de l'hexamère grâce à des mouvements structuraux des boucles Q et R médiés par l'hydrolyse de l'ATP dans la poche ATPase adjacente (**Figures 23 et 27**) (Thomsen and Berger, 2009; Thomsen et al., 2016). Un mécanisme similaire d'escorte a été proposé pour l'ADN hélicase E1 du papillomavirus (Enemark and Joshua-Tor, 2006).

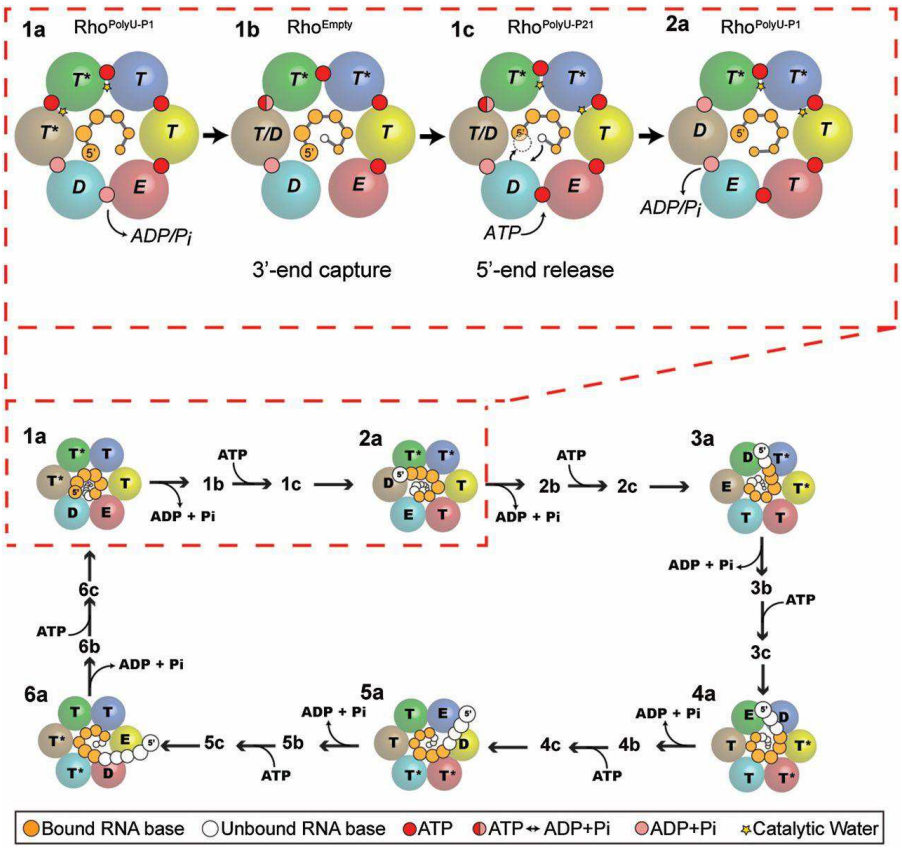


Figure 27 : Modèle de translocation par « *RNA escort* ». Représentation de la translocation de l'ARN (en orange) dans le canal central de l'hexamère suivant l'état d'hydrolyse de l'ATP (E : échange de nucléotides, T : fixation de l'ATP, T* : hydrolyse de l'ATP, D : produits d'hydrolyse). Figure issue de (Thomsen et al., 2016).

Bien que le mécanisme d'escorte stipule que chaque nucléotide est pris en charge de la même façon par le SBS, des expériences d'interférence aux sondes chimiques indiquent que la présence d'un groupement 2'OH dans la chaine ARN n'est réellement critique que tous les 7 nt en moyenne (Rabhi et al., 2011b; Schwartz et al., 2009). Cette contradiction apparente pourrait être expliquée par les contacts PBS-*Rut* persistant en cours de translocation (« *tethered tracking* ») qui provoquent une asymétrie au sein de l'anneau hexamèrique où l'une des six interfaces entre monomères n'est pas pontée par la chaine ARN (**Figure 28**). La déstabilisation induite par la perte d'un groupement 2'OH serait ainsi périodiquement plus forte lorsque cette interface moins « solide » doit prendre en charge un nouveau nucléotide (Soares et al., 2014). Cette interface, plus susceptible de s'ouvrir transitoirement et de laisser ressortir la chaine ARN de l'anneau hexamèrique (**Figure 28**) pourrait expliquer la relativement faible processivité (60-80 nt) du facteur Rho (Soares et al., 2014).

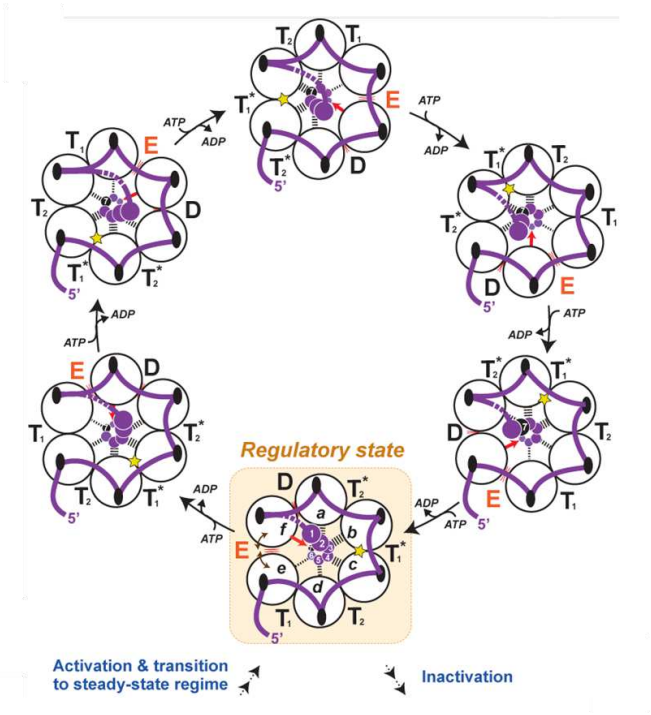


Figure 28 : Translocation de Rho expliquant l'importance d'un groupement 2'OH tous les ~7 nt dans la chaine ARN. L'ARN (en violet) se lie au PBS puis au SBS dans le canal central déclenchant la fermeture de l'anneau hexamérique et la formation d'un complexe Rho:ARN catalytiquement compétent. En raison de la translocation par « *tethered tracking* », le complexe Rho:ARN a une organisation « asymétrique » avec une interface « faible » qui n'est pas pontée par le PBS. Lorsque cette interface doit, à son tour, prendre en charge un nouveau nucléotide (pour l'escorter au travers le canal central), elle est plus facilement déstabilisée par une peturbation telle que l'absence d'un groupement 2'OH (avec lequel une des sous-unités de l'interface est censée interagir à cette étape ; interaction symbolisée par une flèche rouge). Figure adaptée de (Soares et al., 2014).

Comme nous l'avons vu plus haut, la translocation de Rho le long de l'ARN est sans doute beaucoup plus rapide que celle de l'ARNP le long de l'ADN (Gocheva et al., 2015) et il est probable que l'action de Rho soit cinétiquement limitée au niveau de sa phase initiale d'activation qui est lente *in vitro* (Walmacq et al., 2004). Une fois que Rho a rattrapé l'ARNP, il doit induire la dissociation du CET suivant un mécanisme qui reste encore aujourd'hui débattu (Peters et al., 2012; Ray-Soni et al., 2016). Les modèles actuels ressemblent fortement à ceux proposés pour la terminaison intrinsèque (page 27), chacun étant décliné en deux versions suivant que Rho forme ou non une interaction constitutive avec l'ARNP (voir paragraphe suivant) :

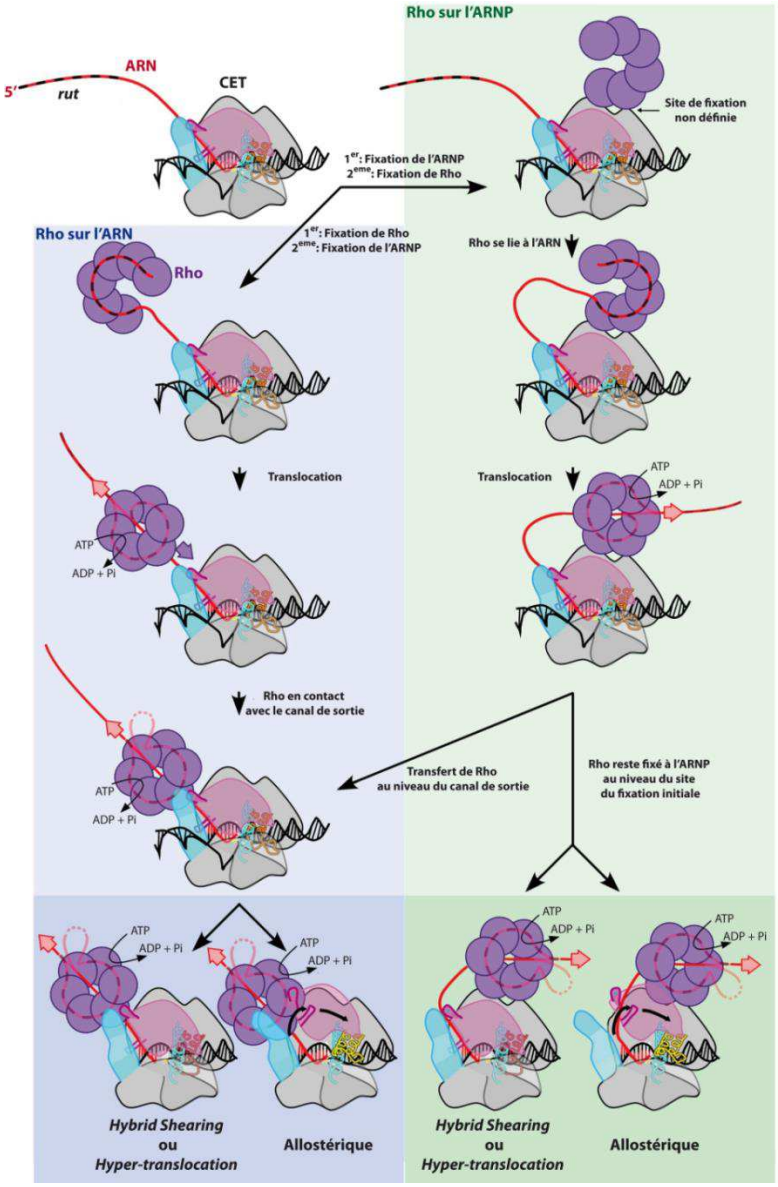


Figure 29 : Illustration des différents modèles de terminaison de la transcription Rho-dependante. Figure adaptée de (Peters et al., 2011).

- Le modèle par « *Hybrid Shearing* », propose que le facteur Rho entre en contact avec l'ARNP au niveau du canal de sortie de l'ARN où il prendrait appui pour extraire de force l'ARN et détruire l'hybride ADN:ARN (**Figure 29**) (Richardson, 2002).
- Le modèle par « *hyper-translocation* » propose que Rho pousse l'ARNP le long de l'ARN. Pour maintenir le registre correct de la bulle de transcription, l'ARNP avancerait également le long de l'ADN mais ce mouvement se ferait sans addition de nucléotide à l'extrémité 3' du transcrit. L'hybride ADN:ARN serait ainsi progressivement réduit jusqu'à provoquer l'effondrement de la bulle de transcription (**Figure 29**) (Park and Roberts, 2006).

- Le modèle « allostérique » propose que le contact entre Rho et l'ARNP provoque des modifications allostériques telles que l'ouverture des pinces β/β' qui conduisent à la déstabilisation du CET (**Figure 29**) (Dutta et al., 2011).

Des données expérimentales obtenues *in vitro* (Epshtein et al., 2010) et *in vivo* (Mooney et al., 2009a) suggèrent que Rho pourrait former une interaction constitutive avec l'ARNP, du moins dans le contexte d'un CET (**Figure 29**, panneau de droite). Par exemple, la pré-incubation de CET artificiellement immobilisés sur billes magnétiques (au sein desquels le transcrit naissant est encore trop court pour dépasser du canal secondaire) avec un mutant inactif de Rho semble suffisante pour empêcher la terminaison Rho-dépendante lorsque ces CET, après lavage des billes magnétiques (pour éliminer ce qui n'y est pas fixé), sont ensuite incubés avec des rNTPs et Rho sauvage (Epshtein et al., 2010). De plus, des expériences d'immunoprécipitation de la chromatine (ChiP-array) ont révélé des distributions similaires de Rho et de l'ARNP le long du génome d'*E. coli*, ce qui n'est pas le cas des cofacteurs NusA ou NusG et supporte l'idée d'une interaction Rho:ARNP constitutive (Mooney et al., 2009a). D'autres études, *in vitro,* plus récentes contestent cette hypothèse (Kalyani et al., 2011; Koslover et al., 2012). On peut remarquer qu'à ce jour aucune équipe (la nôtre comprise) n'a pu détecter d'interaction stable entre Rho et l'ARNP d'*E. coli*. Cette interaction, si elle existe, requière donc la configuration particulière adoptée par l'ARNP au sein du CET ou, peut-être, la participation d'un cofacteur qui reste à ce jour inconnu.

4) Les sites répertoriés de la terminaison Rho-dépendante

L'absence d'éléments de séquence consensus précis a longtemps limité l'identification des terminateurs Rho-dépendants dont seuls quelques exemples étaient connus au milieu des années 2000 (Ciampi, 2006). A l'époque, des centaines de terminateurs intrinsèques étaient connus ou prédits dans de nombreux génomes bactériens (de Hoon et al., 2005; Lesnik et al., 2001; Unniraman et al., 2002), de telle sorte que le rôle de Rho dans la régulation génique bactérienne était souvent perçu comme secondaire. Cette perception a largement évolué à partir de 2008 lorsque la première étude de l'effet de la biyclomycine sur le transcriptome et le protéome d'*E. coli* a démontré un rôle de régulateur global pour le facteur Rho (Cardinale et al., 2008b). Cette étude a également démontré un rôle important de Rho dans le « *silencing* » des prophages et de l'ADN xénogénique. En 2009, le groupe de Robert Landick

identifiait ≈200 sites de terminaison Rho-dépendante chez *E. coli* en comparant les profils ChiP-array (*Chromatin Immunoprecipitation followed by microarray identification*) de l'ARNP obtenus avant et après traitement à la bicyclomycine (Peters et al., 2009). Une partie de ces sites est localisée en aval de gènes non-codants (tRNAs, sRNAs) ou bien dans des régions antisens, soulignant, pour la première fois, le rôle de Rho dans la régulation de la transcription non-codante (Peters et al., 2009). Ce rôle fut confirmé dans une seconde étude du même groupe où des approches transcriptomiques plus sensibles (combinant Microarray et RNAseq) permirent d'identifier ≈1300 *loci*[1] Rho-dépendants chez *E. coli* (Peters et al., 2012). La très grande majorité de ces *loci* (88%) se trouve dans des régions antisens (**Figure 30**), ce qui constitue la première démonstration du rôle prévalent de Rho dans le contrôle de la transcription antisens, et plus généralement de la transcription dite envahissante (« *pervasive* ») ou illégitime (« *spurious* »).

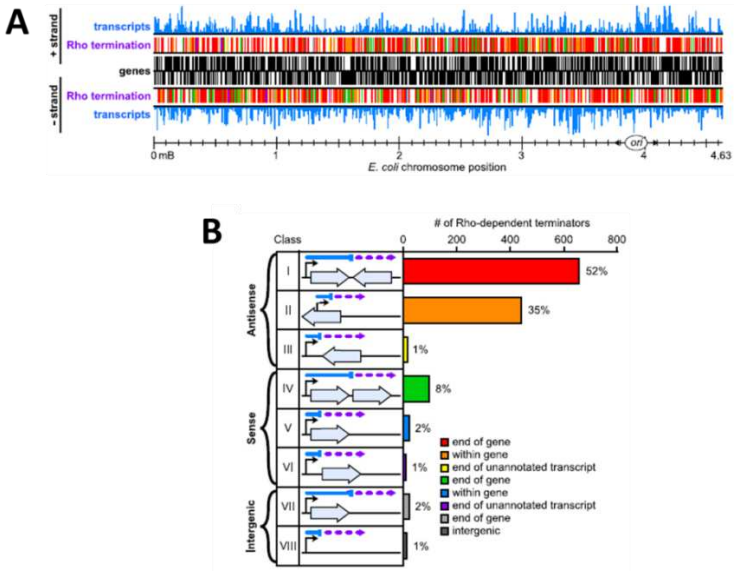


Figure 30 : Détection des *loci* Rho-dépendants dans le génome d'*E. coli*. (A) Les *loci* sont répartis dans tout le génome sans préférence pour l'un des brins ADN. **(B)** Classification des *loci* suivant leurs localisations génomique. Figure issue de (Peters et al., 2012).

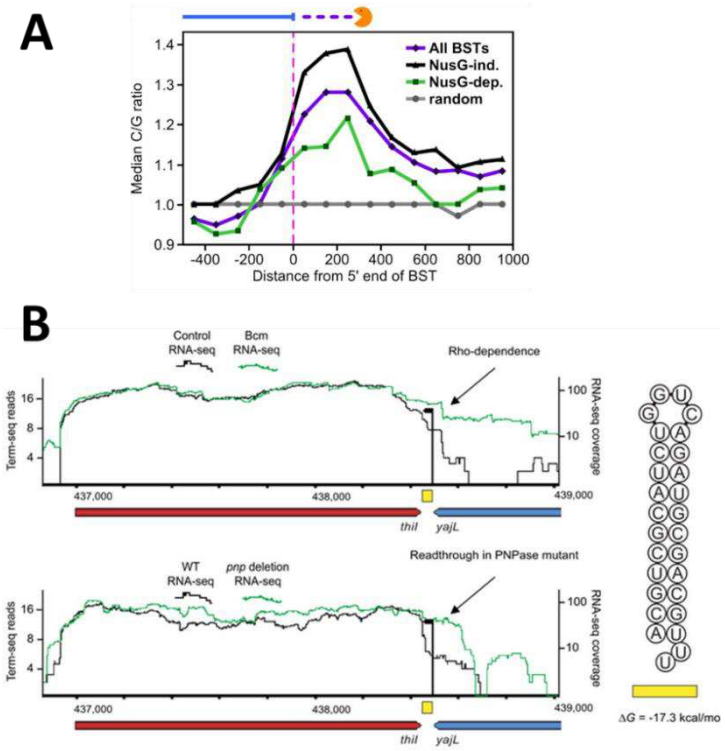
L'étude a également révélé une forte corrélation de position entre les *loci* Rho-dépendants et les sites de fixation à l'ADN de la protéine chaperonne H-NS ainsi que l'existence d'une sous-catégorie de *loci* Rho-dépendants (≈20%) requérant la participation du facteur NusG (Peters et al., 2012). Ces observations ont été globalement confirmées par d'autres groupes étudiant le transcriptome Rho-dépendant chez *E. coli* (Dar and Sorek, 2018;

[1] Le terme « *loci* » est utilisé dans ce manuscrit pour désigner les « sites » identifiés *in vivo* dont la localisation précise est incertaine en raison de l'activité post-transcriptionnelle des exoribonucléases (voir plus loin).

Raghunathan et al., 2018; Sedlyarova et al., 2016). Ces études plus récentes ont également démontré que l'identification des *loci* Rho-dépendants était fortement dépendant des conditions expérimentales et de la méthode d'analyse bioinformatique. Par exemple, de nouveaux *loci* Rho-dépendants ont été identifiés dans les régions 5'UTRs, soulignant le rôle potentiel de Rho dans la régulation conditionnelle (Sedlyarova et al., 2016) (voir aussi page 50). D'autres sites, indétectables dans des conditions normales de croissance, ont mis en évidence l'importance de Rho dans le contrôle des R-loops (Raghunathan et al., 2018) (voir aussi page 25).

Des analyses transcriptomiques similaires ont également été menées chez *B. subtilis* (Bidnenko et al., 2017; Nicolas et al., 2012)*, S. Aureus* (Mader et al., 2016) *et M. tuberculosis* (Botella et al., 2017). Ces analyses ont confirmé l'implication de Rho dans le contrôle de la transcription « pervasive ». On peut noter que le nombre de *loci* Rho-dépendants identifiés est globalement plus faible chez les firmicutes, pour lesquels Rho n'est pas essentiel, que chez *M. tuberculosis* pour lequel Rho est vital. Néanmoins, l'inactivation de Rho chez des espèces où le facteur n'est pas primordial peut avoir des conséquences physiologiques significatives. Par exemple, cette inactivation perturbe l'homogénéité et la mobilité cellulaires, la formation de biofilm et la sporulation chez *B. subtilis* (Bidnenko et al., 2017).

Comme évoqué brièvement plus haut, la localisation précise des sites de terminaison Rho-dépendante par les approches transcriptomiques est compliquée par l'activité post-transcriptionnelle des exoribonucléases (Dar and Sorek, 2018; Peters et al., 2012). Cette activité est essentielle et ne peut être totalement inhibée. Ainsi, le biais C>G caractéristique des sites *Rut* et des fameuses « bulles C>G » (Alifano et al., 1991) est, une fois moyenné, retrouvé en aval (et non en amont) de la position des sites de terminaison identifiés par ces approches (**Figure 31A**). Dans une étude récente, Dar et Sorek ont comparé les positions de *loci* Rho-dépendants identifiés par le groupe de Robert Landick (Peters et al., 2012) avec les extrémités 3' des transcrits générés dans chaque mutant (viable) des trois enzymes responsables de la dégradation exonucléolytique 3→5' chez *E. coli*: PNPase (*pnp*$^-$), RNase II (*rnb*$^-$) et Rnase R (*rnr*$^-$) (Dar and Sorek, 2018). Bien que menée sur un nombre limité de *loci* (144), cette analyse a confirmé le rôle essentiel des exonucléases dans la prise en charge des transcrits Rho-dépendants qui sont visiblement raccourcis jusqu'à ce que les exonucléases



soient bloquées par une structure en tige-boucle suffisamment stable située en amont (**Figure 31B**) (Dar and Sorek, 2018).

Figure 31 : L'activité des exoribonucléases altère l'identification des sites de terminaison Rho-dépendante. (A) Ratio C>G moyen en fonction de la position du début de signal Rho-dépendant (BST) détecté par RNAseq. Le ◗ suggère une digestion possible des transcrits par les exoribonucléases 3'→ 5'. Figure issue de (Peters et al., 2012). **(B)** Comparaison de profils RNAseq obtenus pour la souche sauvage traitée (**en vert**) ou non (**en noir**) à la BCM et pour un mutant inactivé de la PNPase. Ces données suggèrent que la PNPase dégrade le transcrit Rho-dépendant jusqu'à une structure en tige-boucle stable. Figure issue de (Dar and Sorek, 2018).

Un nombre limité de terminateurs Rho-dépendants a fait l'objet d'études plus précises menées *in vitro* et/ou *in vivo* qui ont permis de mieux en préciser la position et les caractéristiques principales (**Table 3**). Certains de ces sites fonctionnent de manière constitutive alors que d'autres ne sont opérationnels que dans des conditions bien particulières (**Table 3**) qui vont masquer/démasquer le site *Rut* ou qui vont moduler une étape plus tardive du processus de terminaison (comme l'activation catalytique du facteur Rho). Ces mécanismes de régulation conditionnelle impliquant Rho sont détaillés plus loin (page 50).

Tableau 3 : Terminateurs Rho-dépendants connus.

Localisation génomique	Organisme	Régulation	Position du terminateur	Références
galE	*E. coli*	Conditionnel (accès au site *Rut* dépendant NusG)	Intragénique	(De Crombrugghe et al., 1973; Sullivan and Gottesman, 1992)
ilv	*E. coli*	/	Intragénique (*ilvGM*)	(Wek et al., 1987)
lac	*E. coli*	Conditionnel (accès au site *rut* dépendant de la rupture du couplage transcription/traduction	Intragénique	(Ruteshouser and Richardson, 1989; Stanssens et al., 1986)
lamB	*E. coli*	/	Intragénique	(Colonna and Hofnung, 1981)
lysC	*E. coli*	Conditionnel (accès au site *Rut* dépendant d'un riboswitch)	5'UTR	(Bastet et al., 2017)
nadD	*E. coli*	Conditionnel (accès au site *Rut* dépendant de la transcription d'un iRAP)	Intragénique (*nadD*)	(Sedlyarova et al., 2017)

pgaA	*E. coli*	Conditionnel (accès au site *rut* dépendant CsrA)	5'UTR	(Figueroa-Bossi et al., 2014)
rho	*E. coli*	Conditionnel (autorégulation dépendante de [Rho])	5'UTR	(Matsumoto et al., 1986)
ribB	*E. coli*	Conditionnel (accès au site *Rut* dépendant d'un riboswitch)	5'UTR	(Bastet et al., 2017; Hollands et al., 2012)
rpoS	*E. coli*	Conditionnel (accès au site *Rut* dépendant de *dsrA; arcZ*; *rprA*)	5'UTR	(Sedlyarova et al., 2016)
tfaS	*E. coli*	Constitutif	3'UTR	(Menouni et al., 2013)
thiC	*E. coli*	Conditionnel (accès au site *Rut* dépendant d'un riboswitch)	5'UTR	(Bastet et al., 2018; Chauvier et al., 2017)
thiB	*E. coli*	Conditionnel (accès au site *Rut* dépendant d'un riboswitch)	5'UTR	(Bastet et al., 2018; Chauvier et al., 2017)
thiM	*E. coli*	Conditionnel (accès au site *Rut* dépendant d'un riboswitch)	5'UTR	(Bastet et al., 2018; Chauvier et al., 2017)
tna	*E. coli*	Conditionnel (accès au site *Rut* dépendant de la traduction d'un peptide leader)	5'UTR	(Gong and Yanofsky, 2003; Stewart et al., 1986)
trp	*E. coli*	/	Intragénique (*trpE*)	(Korn and Yanofsky, 1976)
	E. coli	Constitutif	3'UTR	(Platt, 1981; Wu et al., 1981; Zalatan et al., 1993)
tyrT	*E. coli*	Constitutif	3'UTR	(Kupper et al., 1978; Rossi et al., 1981)
chiP	*S. Typhimurium*	Conditionnel (accès au site *Rut* dépendant de *chiX*)	5'UTR	(Bossi et al., 2012)
corA	*S. Typhimurium*	Conditionnel (accès au site *Rut* dépendant de la traduction d'un peptide leader)	5'UTR	(Kriner and Groisman, 2015; Silverman, 1974)
his (Plusieurs sites)	*S. typhimurium*	Constitutifs (mais NusA-dépendant)	Intragénique (*hisG* et *hisC)*	(Alifano et al., 1991; Carlomagno and Nappo, 2003; Ciampi et al., 1989; Ciampi et al., 1982)
mgtA	*S. Typhimurium*	Conditionnel (accès au site *Rut* dépendant de la traduction d'un peptide leader)	5'UTR	(Gall et al., 2016; Hollands et al., 2012)
mgtC	*S. Typhimurium*	Conditionnel (accès au site *Rut* dépendant de la traduction d'un peptide leader)	5'UTR	(Gall et al., 2018; Sevostyanova and Groisman, 2015)
tufB	*S. Typhimurium*	Conditionnel (autorégulation dépendante de [TufB])	5'UTR	(Brandis et al., 2016)
rho	*B. subtilis*	Conditionnel (autorégulation dépendante de [Rho])	5'UTR	(Ingham et al., 1999)
trp	*B. subtilis*	Conditionnel (accès au site *Rut* dépendant de la pause du ribosome)	Intragénique (*trpE*)	(Yakhnin et al., 2001)
ribM	*C. glutamicum*	Conditionnel (accès au site *Rut* dépendant d'un riboswitch)	5'UTR	(Takemoto et al., 2015)
nifA	*K. pneumoniae*	Conditionnel (accès au site *Rut* dépendant de la rupture du couplage transcription/traduction)	Intragénique	(Govantes et al., 1996)
cro (λtR1)	*Phage λ d'E. coli*	Conditionnel (accès au site *Rut* dépendant de NusB ; NusG ; NusE/S10)	Intragénique *(cro* et *cII)*	(Chen and Richardson, 1987; Mogridge et al., 1998)
Gène IV	*Phage f1 d'E. coli*	Constitutif	3'UTR	(La Farina et al., 1990; Moses and Model, 1984)
kil	*Phage P4 d'E. coli*	Conditionnel (accès au site *Rut* dépendant de la traduction de *rac*)	Intragénique	(Briani et al., 2000; Forti et al., 1999)

5) Facteurs de régulation de la terminaison Rho-dépendante

De nombreux facteurs, agissant aussi bien en *cis* qu'en *trans*, régulent la terminaison Rho-dépendante. Dans les pages qui suivent, je présente brièvement les exemples les plus pertinents de ces facteurs.

a. Hfq, Yao et Psu

Ces trois protéines sont capables d'interagir directement avec Rho et d'en inhiber l'action suivant des mécanismes moléculaires différents.

L'activité d'anti-terminaison de la protéine endogène YaeO a été découverte par hasard par le groupe de J.P. Bouché lors de la recherche de suppresseurs de mutations affectant la division cellulaire (Pichoff et al., 1998). *In vitro*, YaeO forme avec Rho un complexe binaire relativement stable (Gutierrez et al., 2007). Une étude par résonnance magnétique nucléaire (RMN) suggère que YaeO interagit avec Rho au niveau du PBS (**Figure 32**) ce qui lui permettrait d'empêcher l'ancrage de Rho aux transcrits naissants (Gutierrez et al., 2007). Le rôle et le contexte physiologiques de cette activité d'anti-terminaison de YaeO restent à ce jour mystérieux.

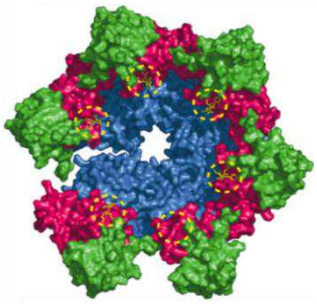


Figure 32 : Modèle du complexe Rho:YaeO. Le facteur d'anti-terminaison interagit au niveau des poches de fixation du PBS de Rho. Les zones en pointillées jaune indiquent les régions d'interférences stériques de YaeO (rouge & vert) sur Rho (bleu). Figure issue de (Gutierrez et al., 2007).

La protéine Psu de la capside du bactériophage P4 peut aussi inhiber l'activité de terminaison de Rho (Pani et al., 2006a), les deux facteurs forme un complexe binaire stable (Pani et al., 2006b; Pani et al., 2009). Bien qu'il n'existe pas de structure à haute-résolution de ce complexe, des expériences biochimiques suggèrent que Psu contacte le NTD de Rho (Ranjan et al., 2013). Les auteurs de ce travail ont également bâti un modèle moléculaire du complexe dans lequel un dimer Psu interagit avec deux interfaces entre des sous-unités de Rho se faisant face dans l'hexamère, de telle façon que Psu vient former « un couvercle » sur le canal central de Rho (**Figure 33**) (Ranjan et al., 2013). Leur proposition que Psu inhibe Rho

en empêchant l'accès au canal central, et donc au SBS, est séduisante mais n'a pas encore été testée en profondeur.

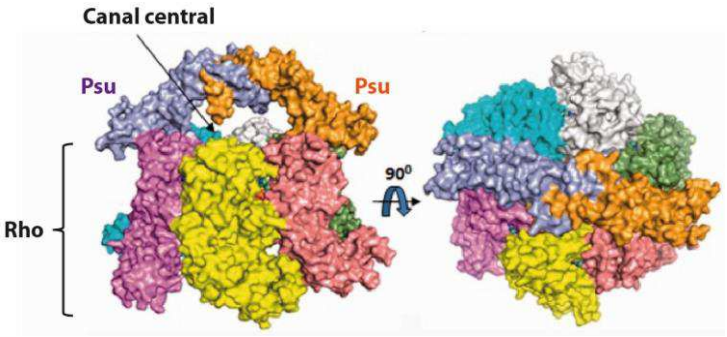


Figure 33 : Modèle du complexe Rho:Psu. Le dimère Psu (PDB : 3RX6) forme « un couvercle » sur le canal central de Rho (PDB : 3ICE). Figure inspirée de (Ranjan et al., 2013).

Enfin, il a été démontré au laboratoire que l'ARN chaperonne Hfq, un homo-héxamère en forme d'anneau (Brennan and Link, 2007; Gottesman et al., 2006) de plus petite taille que Rho était également capable *in vitro* de former un complexe binaire stable (Kd~ 40 nM) avec Rho (Rabhi et al., 2011a). Cette interaction peut conduire à l'inhibition des activités enzymatiques de Rho (ATPase, hélicase, terminaison) pourvu que Hfq contacte également l'ARN dans le complexe ternaire Rho:Hfq:ARN (Rabhi et al., 2011a). Une activité d'anti-terminaison de Hfq a pu être démontrée *in vivo* avec un terminateur Rho-dépendant modèle (*λtR1*) mais aucun système naturellement régulé de cette façon n'a encore été découvert. On peut noter que Hfq présente une homologie structurale avec le CTD de NusG et que les deux facteurs semblent interagir avec Rho de façon mutuellement exclusive, suggérant qu'il compètent pour le même site de fixation (Rabhi et al., 2011a). Enfin, Hfq ne perturbe pas l'association d'oligonucléotides avec le PBS de Rho, suggérant que son effet inhibiteur s'exerce à une étape plus tardive du processus d'activation de Rho que la simple fixation à l'ARN (Rabhi et al., 2011a).

b. Les protéines de liaison à l'ARN

Diverses protéines de liaison à l'ARN (RBP), dont Hfq, sont impliquées dans le contrôle de la terminaison Rho-dépendante. Leur liaison au transcrit naissant peut provoquer des réarrangements structuraux de l'ARN susceptibles de masquer/démasquer les sites *Rut*. Plusieurs cas ont été détectés dans les régions 5'UTR, conduisant à la régulation du gène (ou

opéron) en aval ; on parle dans ce cas d'atténuation transcriptionnelle (due à la terminaison Rho-dépendante).

La protéine CsrA (6,8 Da) est un régulateur post-transcriptionnel (Edwards et al., 2011) qui contrôle la traduction et la stabilité de divers ARNm cibles (Vakulskas et al., 2015). Pour cela, CsrA reconnaît des motifs du type AN**GGA** (ou [A/C]U**GGA** chez *Salmonella*) présents dans la partie apicale de structures en tige-boucle (Holmqvist et al., 2016) ou au niveau du RBS (A**GGA**GGUAA) ou du codon stop (AUG) où CsrA se fixe de façon compétitive avec le ribosome (Vakulskas et al., 2015). Récemment, il a été démontré au laboratoire que CsrA est aussi un régulateur transcriptionnel, au moins pour le contrôle de l'opéron *pgaABCD* (Figueroa-Bossi et al., 2014). Dans ce cas, l'accès au site *Rut* piégé dans une structure secondaire est conditionné par la fixation de CsrA à la région 5'UTR de l'ARNm *pgaA* (**Figure 34**) (Figueroa-Bossi et al., 2014).

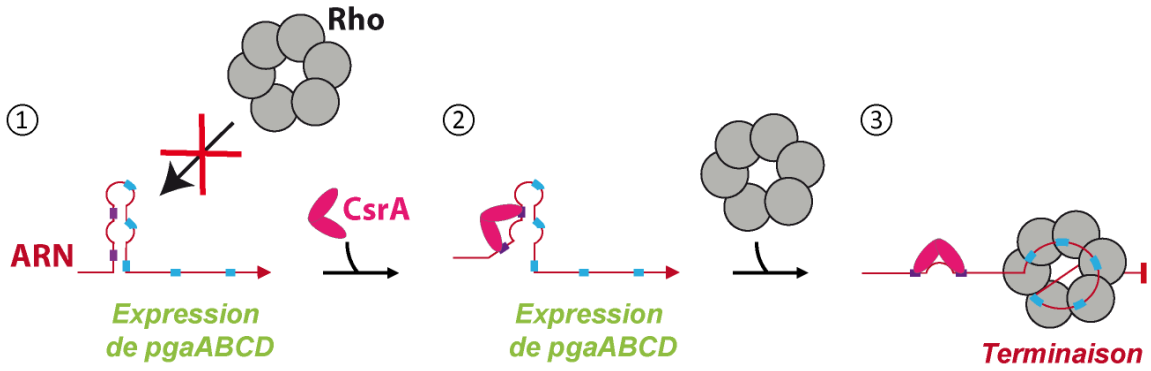


Figure 34 : Atténuation transcriptionnelle de l'opéron *pgaABCD* par CsrA et Rho. ① L'opéron *pgaABCD* est transcrit normalement, le facteur Rho ne peut pas se fixer sur l'ARN naissant. ② La protéine CsrA se fixe au niveau des sites de reconnaissance, induisant une modification de la structure secondaire de la région 5'-leader de l'ARNm *pgaA*. ③ Le facteur Rho reconnaît le site *rut* qui n'est plus piégé dans une structure secondaire, ce qui permet la terminaison de la transcription. Figure inspirée de (Figueroa-Bossi et al., 2014).

La deuxième RBP que j'évoquerais est le ribosome qui masque les sites *Rut* situés dans les régions codantes des ARNm (**Figure 1**). Le ribosome est aussi parfois impliqué dans la traduction de petites séquences situées dans les régions 5'UTR conduisant à la formation de « peptides leaders ». La synthèse de peptides leaders a été impliquée dans divers mécanismes d'atténuation transcriptionnelle basés sur la terminaison Rho-dépendante. La régulation est généralement obtenue par le masquage/démasquage d'un site *Rut* gouverné directement par la présence/absence du ribosome et/ou par le remodelage structural de la région 5'UTR de l'ARNm lors de la traduction du peptide leader (Kriner et al., 2016).

L'exemple le plus connu et le mieux étudié est sans doute celui gouvernant la synthèse de la tryptophanase (opéron *tna*) en fonction de la concentration en tryptophane (Yanofsky, 2007) (**Figure 35**).

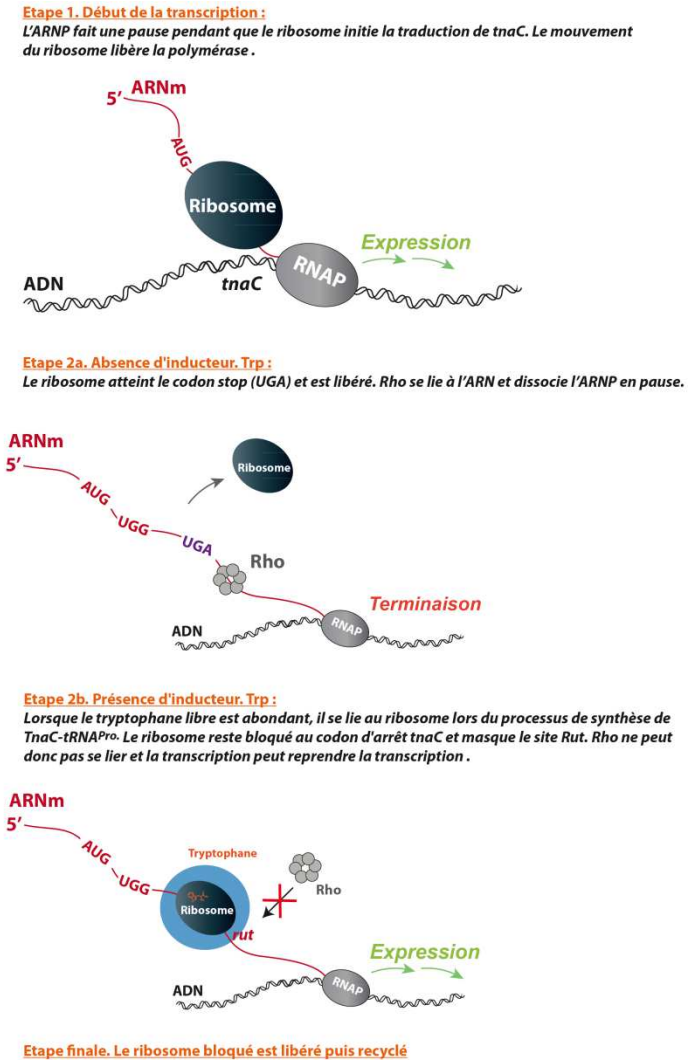


Figure 35 : Mécanismes d'expression de l'opéron *tna* conditionné par la concentration en tryptophane. Figure inspirée de (Yanofsky, 2007).

Un exemple plus récent découvert chez *Salmonella* concerne l'expression du gène *mgtA* (codant pour un transporteur du Mg^{2+}) qui participe à l'homéostasie du magnésium (Gall et al., 2016; Gall et al., 2018; Kriner and Groisman, 2015; Silverman, 1974). Dans ce cas, l'accessibilité d'un site *Rut* localisé dans la région 5'leader de l'ARNm est gouvernée par la traduction du peptide leader *mgtL* (**Figure 36**). La vitesse de traduction de *mgtL* est dépendante de la concentration en Mg^{2+} : à faible concentration, le ribosome pause au niveau de codons Proline (car la réaction de chargement de l'$ARNt^{Pro}$ requière du Mg^{2+}), ce qui favorise une conformation de la région 5'-leader masquant le site *Rut* (**Figure 36**) ; à plus

forte concentration en Mg^{2+}, le ribosome traduit normalement *mgtL* jusqu'au codon stop, l'ARN se structure d'une manière différente et permet l'accès au site *Rut* (**Figure 36**) (Gall et al., 2016; Gall et al., 2018; Kriner and Groisman, 2015; Silverman, 1974). Ironiquement, ce système a été d'abord décrit comme le premier riboswitch gouvernant la terminaison Rho-dépendante en fonction du ligand Mg^{2+} (Hollands et al., 2012), avant que l'action de remodelage du ribosome plus directe n'ait été élucidée (Gall et al., 2016).

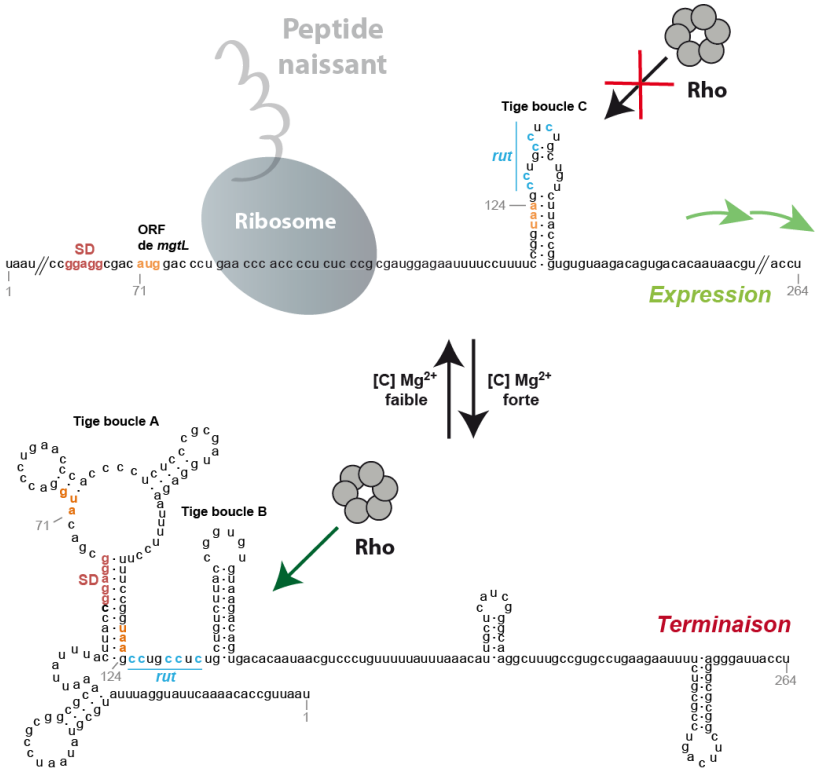


Figure 36 : Régulation encadrant la transcription de *mgtA*. Si la concentration de Mg^{2+} est faible, les ribosomes sont bloqués au niveau de *mgtL* ce qui permet la formation de la tige-boucle C dans l'ARN naissant. Cette structure séquestre le site de liaison de Rho et permet la transcription de *mgtA*. Si la concentration de Mg^{2+} est forte, la traduction de *mgtL* est rapide et complète. Ceci favorise la formation des tige-boucles A et B et expose le site *Rut*. Sur ce schéma les codons initiateur et stop de *mgtL* sont en orange et le RBS est en rouge. Figure inspirée de (Gall et al., 2016).

c. NusA

Le facteur NusA, qui est aussi une RBP, régule les pauses de type I, la terminaison intrinsèque (partie B-II-3a) et l'anti-terminaison de la transcription. Ce dernier aspect est illustré par le cas de la protéine N du bactériophage *λ* dont NusA facilite la liaison au CET au niveau d'un site *nut* (N-utilisation). Le site *nut* est composé de deux éléments distincts, *boxA* et *boxB* (une structure en tige-boucle) séparés par une séquence « *spacer* » (Friedman and Court, 1995) qui contribuent également au recrutement et à la stabilisation des autres

composants (NusB, NusE, NusG) du complexe d'antiterminaison (**Figure 37**) (Prasch et al., 2009).

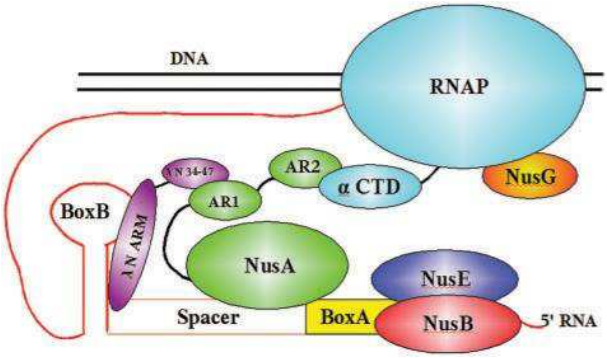


Figure 37 : Représentation schématique du complexe d'anti-terminaison chez le phage λ. Ce modele prévoit les interactions suivantes : **NusA**:***Spacer*** , **NusA-AR1**:**λN** , **NusA-AR2**:**αCTD-RNAP** , **λN**:***boxB*** et **NusE**:**NusB**:***boxA***. Figure issue de (Prasch et al., 2009).

La formation de ce complexe permet au CET d'outrepasser certains sites de terminaison présents en aval des promoteurs *λ*tL et *λ*tR. C'est, par exemple, le cas du terminateur Rho-dépendant *λ*tR1 situé dans la partie 3'UTR du gène *λcro* et qui contient un site *nut* et un site *Rut* entremêlés (Faus and Richardson, 1990; Vieu and Rahmouni, 2004b). Des motifs « *nut-like* » sont également présents au sein des opérons ribosomaux qui, n'étant pas traduits, sont protégés de l'action de Rho par la formation de complexes d'antiterminaison similaires et contenant NusA ainsi que NusG et NusE (Condon et al., 1993; Friedman and Baron, 1974). NusA assiste également la formation d'un complexe d'antiterminaison avec la protéine Q du phage *λ* qui ne nécessite pas d'autres cofacteurs mais la présence d'une séquence ADN spécifique appelée *qut* (Santangelo and Artsimovitch, 2011). Ce complexe d'anti-terminaison intervient dans la transcription à partir du promoteur *λ*tR' et, comme le complexe N-dépendant, est résistant à l'action de Rho (Shankar et al., 2007). Dans certains cas, comme celui des terminateurs *his* intragéniques de *Salmonella* (**Table 3**), NusA parait capable de stimuler la terminaison Rho-dépendante (Carlomagno and Nappo, 2003). Le mécanisme de cette stimulation reste, néanmoins, mystérieux.

d. NusG & RfaH

NusG a divers effets sur la terminaison Rho-dépendante qui semblent parfois contradictoires. Cette complexité est due à la modularité de la protéine capable d'interagir avec les sous-unités β de l'ARNP via son NTD et, de façon mutuellement exclusive, avec Rho (Sullivan and Gottesman, 1992) ou la protéine ribosomale S10 (*aka* NusE) (Mooney et al., 2009b; Strauss et al., 2016; Svetlov et al., 2007) via son CTD. L'interaction avec l'ARNP stimule l'activité et la stabilité du CET (Burova et al., 1995), ce qui est *a priori* défavorable à

la terminaison Rho-dépendante. L'interaction avec NusE garantit le couplage transcription-traduction (**Figures 8 et 38**) (Burmann et al., 2010; Strauss et al., 2016) et/ou la formation de complexes d'anti-terminaison (voir ci-dessus), des évènements également défavorables à la terminaison Rho-dépendante.

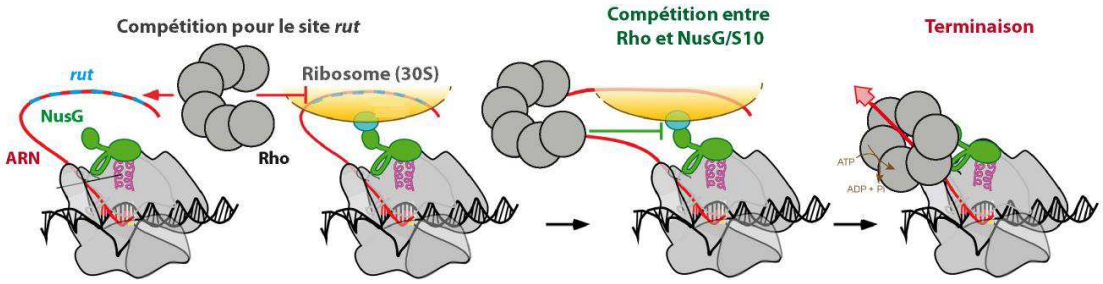


Figure 38 : Compétition entre Rho et le ribosome pour NusG-CTD. Figure issue de (Peters et al., 2011).

Rho a une affinité pour le CTD de NusG (K_d ~12 nM) qui est très supérieure à celle démontrée par NusE (K_d ~50µM) (Burmann et al., 2010). Cette différence est potentiellement compensée par les contacts directs entre ARNP et ribosome au sein de l'expressome bactérien (**Figure 10**). En cas de rupture (ou d'absence) du couplage transcription-traduction, le CTD de NusG est disponible pour interagir avec Rho (**Figure 38**). Pourtant, seule une minorité (~20 %) de *loci* Rho-dépendants parait être sensible à l'action de NusG, qui stimule la terminaison dans ce cas (Peters et al., 2012; Shashni et al., 2014). Cette minorité est caractérisée par un ratio C>G plus faible que pour les autres *loci* Rho-dépendants (**Figure 31**) (Peters et al., 2012; Shashni et al., 2014) qui pourrait refléter l'absence de sites *Rut* optimaux ou l'existence de structures secondaires délétères. *In vitro*, la stimulation de la terminaison Rho-dépendante par NusG est souvent caractérisée par un décalage des sites *tsp* vers le promoteur (Nehrke et al., 1993). D'une façon générale, le mécanisme exact par lequel NusG stimule la terminaison Rho-dépendante reste mal compris et débattu (Burns and Richardson, 1995; Valabhoju et al., 2016). Une étude très récente suggère que les boucles L1/L2 de la partie NusG-CTD réalisent une interaction avec Rho selon un modèle de « *trimer-of-dimers* » (**Figure 39A**) (Lawson et al., 2018). Cette interaction facilite l'encerclement des substrats ARN non-optimaux à Rho (**Figure 39B**) et favorise son passage en complexe fermé par un remodelage des interactions dans la partie Rho-CTD (**Figure 39C**) (Lawson et al., 2018).

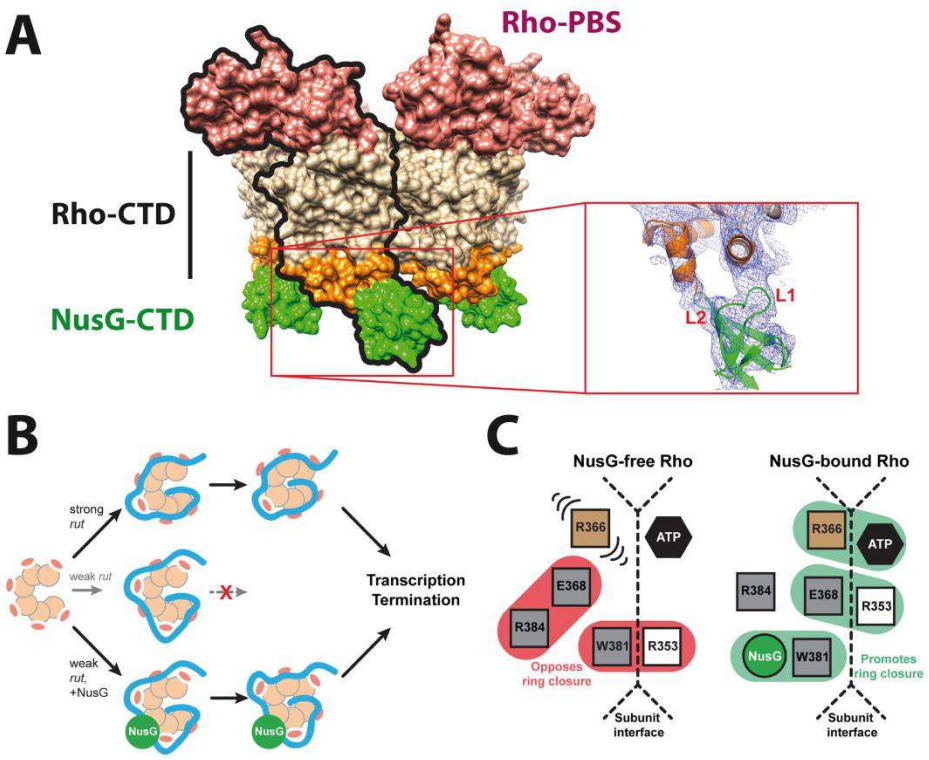


Figure 39 : Modèle de l'interaction Rho:NusG. (A) Structure crystallographique du complexe Rho:NusG-CTD, avec un zoom sur l'interaction des boucles L1 et L2 de NusG avec Rho (PDB : 6DUQ). **(B)** Influence de NusG sur la terminaison de la transcription Rho-dépendante en fonction de la constitution optimale ou non du site *Rut*. **(C)** Représentation du remodelage de la partie CTD de Rho par NusG, notamment la stabilisation de l'interaction Arg366:ATP, qui favorise le passage en complexe fermé. Figure issue de (Lawson et al., 2018).

Le facteur paralogue RfaH a des propriétés similaires à NusG mais qui sont restreintes aux régions contenant une séqence ***ops*** (**Figure 9**) (Belogurov et al., 2009; Burmann et al., 2012). Néanmoins, RfaH n'est pas capable d'interagir avec Rho et a une action inhibitrice plutôt que stimulatrice de son action (Hu and Artsimovitch, 2017).

e. sRNA

Il est bien établi que les sRNA régulent leurs cibles ARNm en formant des duplexes imparfaits avec ceux-ci, duplexes qui sont généralement stabilisés par une chaperonne, Hfq ou ProQ (Holmqvist et al., 2018; Masse et al., 2003; Santiago-Frangos et al., 2016; Wagner and Romby, 2015). Dans un grand nombre de cas, ces duplexes altèrent la traduction et/ou la dégradation des ARNm cibles (partie B-I-3) mais d'autres effets possibles ont également été découverts. Ainsi, une étude fondatrice du groupe de Lionello et Nara Bossi (CNRS, Gif/Yvette) en collaboration avec notre équipe a démontré que le sRNA *ChiX* était capable d'activer un terminateur Rho-dépendant latent (intragénique) situé dans le premier gène de

l'opéron *chiPQ* de *Salmonella* (Bossi et al., 2012). Cette capacité est en fait liée à un phénomène classique de découplage de la traduction induit par la fixation de *ChiX* à l'ARNm à proximité du RBS du gène *ChiP* (Bossi et al., 2012). Un mécanisme d'activation de la terminaison Rho-dépendante par le sRNA *Spot42* (aka spf) à la jonction *galT-galK* de l'opéron *gallETKM* a depuis été découvert (Wang et al., 2015). Les auteurs proposent que, dans ce cas, la fixation de *Spot42* à l'ARNm provoque la dissociation précoce (en amont) du ribosome, ce qui a pour effet de découvrir un site *Rut* dans la partie terminale du gène *galT* (Wang et al., 2015). Des études menées au laboratoire suggèrent que d'autres systèmes pourraient être régulés par un sRNA via une modulation de la terminaison Rho-dépendante (voir apport personnel). Récemment, un consortium d'équipes concurrentes a démontré que des sRNA pouvaient également inhiber la terminaison Rho-dépendante. Cet effet a été observé pour chacun des sRNA *rprA*, *dsrA* et *arcZ* qui, assistés de Hfq, se fixent à la région 5'UTR du gène *rpoS* d'adaptation au stress (Sedlyarova et al., 2016). Cette fixation inhibe un terminateur Rho-dépendant présent dans la région 5'UTR (Sedlyarova et al., 2016) suivant un mécanisme qu'il reste à préciser (**Figure 40**). Des observations similaires ont été faites au laboratoire avec ce système (voir Apport personnel).

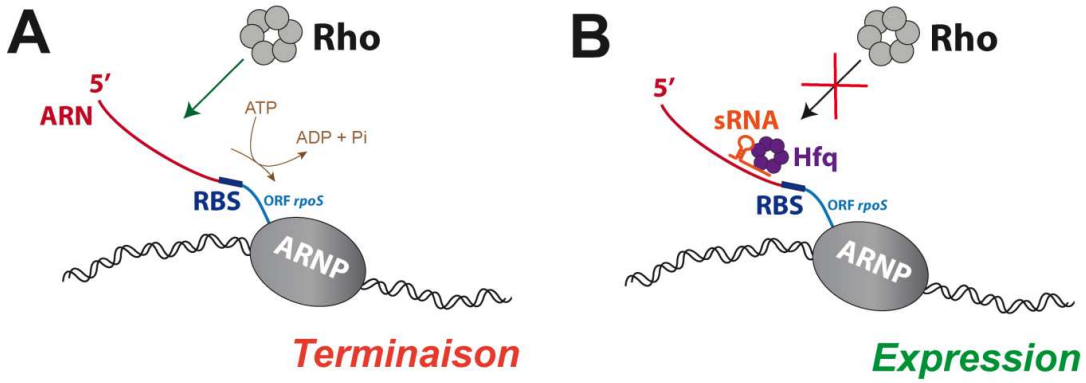


Figure 40 : Inhibition de la terminaison Rho-dépendante par un sRNA dans la région 5'UTR de *rpoS*. (A) En absence de sRNA, Rho a accès au terminateur situé dans la région 5'UTR de *rpoS*. **(B)** La présence d'un duplexe ARNm:sRNA dans le 5'UTR inactive le terminateur. Le mécanisme précis n'est pas connu et ne repose pas nécessairement sur un principe d'exclusion d'accès au transcrit comme représenté ici. Figure inspirée de (Sedlyarova et al., 2016).

f. Riboswitches

Comme évoqué plus haut, les riboswitches sont des plateformes ARN capables d'adopter deux conformations différentes en fonction de la présence ou non d'un ligand spécifique. Une étude parue en 2012 a démontré pour la première fois que ce principe était exploité par

deux riboswitches présents dans les régions 5'UTR des gènes *mgtA* de *Salmonella* (ligand : Mg^{2+}) et *ribB* d'*E. coli* (ligand : FMN) pour contrôler l'accès de Rho au transcrit (Hollands et al., 2012). Bien que le système *mgtA* soit en réalité plus complexe (voir page 54), d'autres riboswitches gouvernant la terminaison Rho-dépendante ont depuis été découverts chez diverses espèces :

- chez *E. coli : thiB*, *thiC* et *thiM* (ligand : TPP) ainsi que *lysC* (ligand : lysine) (Bastet et al., 2017; Chauvier et al., 2017).
- Chez *S. typhimurium* : *mgtC* (ligand : Mg^{2+}) (Sevostyanova and Groisman, 2015).
- Chez *C. glutamicum* : *ribM* (ligand : FMN) (Takemoto et al., 2014).

Ces riboswitches peuvent masquer/démasquer un site *Rut* directement impacté par la structure locale de la région 5'UTR ou, dans le cas de riboswitches traductionnels, contrôler l'accès de Rho à un terminateur intragénique plus éloigné en provoquant le découplage de la traduction (**Figure 41**) (pour revue : (Bastet et al., 2018)).

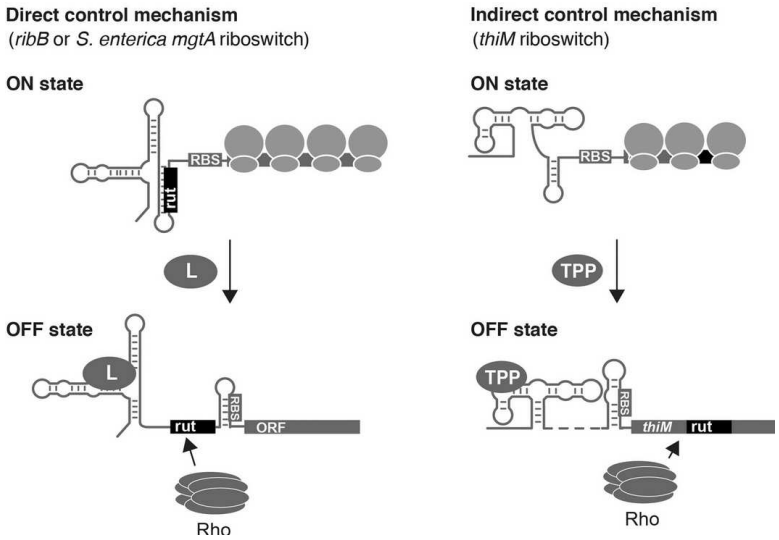


Figure 41 : Mécanismes alternatifs de fonctionnement des riboswitches Rho-dépendants. Figure issue de (Bastet et al., 2017).

g. Séquences dans le transcrit naissant (RARE, ¡RAP)

Des études récentes du riboswitch *mgtC* situé dans la région 5'UTR de l'opéron *mgtCBR* de *Salmonella* ont démontré que ce système était beaucoup plus complexe qu'anticipé. Si comme pour *mgtA* (voir page 54) la traduction d'un petit peptide leader est impliquée

(**Figure 36**), le plus surprenant est la présence d'une courte séquence inhibitrice dans la région 5'UTR de *mgtC* (Sevostyanova and Groisman, 2015). Cette séquence appelée RARE (*Rho-Antagonizing RNA Element*) est capable d'inhiber Rho fixé au transcrit et ne semble pas être en compétition directe avec l'interaction ARN:PBS canonique (Sevostyanova and Groisman, 2015). Ainsi le motif RARE piégerait le complexe Rho:ARN dans une forme non-productive. La formation des contacts inhibiteurs entre RARE et Rho serait gouvernée par la traduction du peptide leader, elle-même dépendante de la concentration en magnésium (**Figure 42**) (Gall et al., 2018; Sevostyanova and Groisman, 2015).

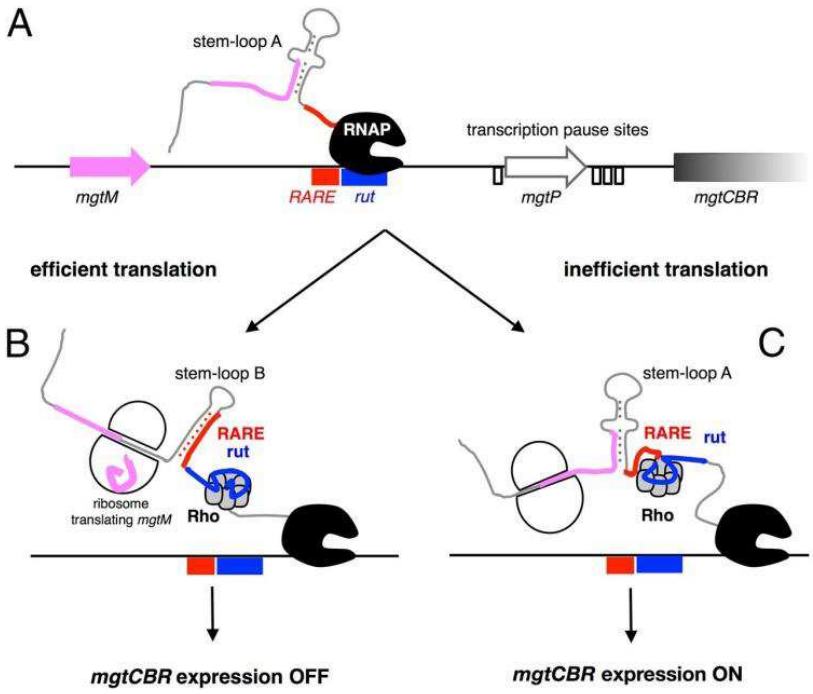


Figure 42 : Modèle de régulation de la terminaison Rho-dépendante pour l'opéron *mgtCBR de Salmonella*. (A) Représentation de la région leader de l'operon *mgtCBR*. Les sites importants sont indiquées par des rectangles colorés : **RARE**, ***Rut***, ***tsp***. **(B & C)** L'efficacité de traduction influe sur la structure secondaire de l'ARN, la libération du site **RARE** et l'expression de l'operon *mgtCBR*. Figure issue de (Sevostyanova and Groisman, 2015).

Un crible par SELEX génomique a récemment identifié ~15 000 motifs ARN (longueur moyenne de 48 nt) capables d'interagir avec l'ARNP et que les auteurs ont appelé RAPs (*RNA polymerase binding aptamers*) (Sedlyarova et al., 2017). Une fraction de ces RAPs (10-15%) est capable d'inhiber la transcription (¡RAPs pour *inhibitory* RAPs) *in vivo*. Pour au moins quelques-uns des ¡RAPs situés dans les régions codantes, l'inhibition de la transcription pourrait être due à un phénomène de découplage de la traduction et d'activation de terminateurs Rho-dépendants intragéniques situés en aval (**Figure 43**) (Sedlyarova et al., 2017).

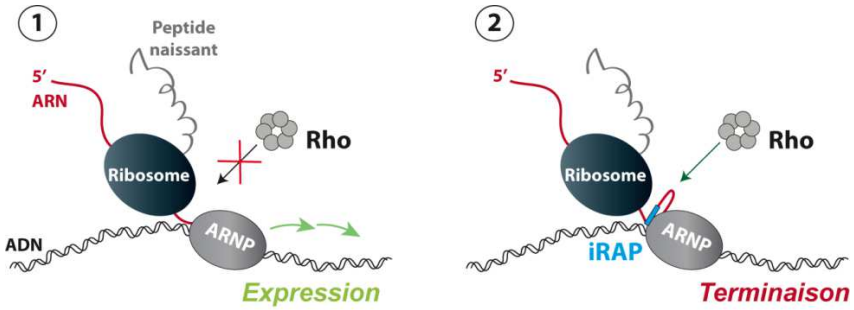


Figure 43 : Modèle d'activation d'un terminateur intragénique par un iRAP. L'interaction entre le motif iRAP et l'ARNP bloque le ribosome en amont du iRAP. La région aval de l'ARNm n'est donc plus traduite et devient accessible à Rho. Figure inspirée de (Sedlyarova et al., 2017).

C. Apport personnel

I. Introduction

Les études de la terminaison Rho-dépendante menées à l'échelle génomique ou transcriptomique (Bidnenko et al., 2017; Botella et al., 2017; Dar and Sorek, 2018; Mader et al., 2016; Nicolas et al., 2012; Peters et al., 2012; Peters et al., 2009; Raghunathan et al., 2018; Sedlyarova et al., 2016) ou au niveau de sites bien identifiés (**Tableau 4**) n'ont pas permis de dégager d'éléments consensus autres que ceux évoqués précédemment (partie B-III-2, page 38), c'est-à-dire la présence de « bulles C>G » (dans le transcrit ou le brin ARN non-matrice) (Alifano et al., 1991) et une relative pauvreté en structures secondaires dans la région comprenant le site *Rut*. Il est possible que les terminateurs Rho-dépendants soient construits à partir de règles relativement laxes permettant une intervention du facteur Rho à de nombreux points du génome, en fonction des circonstances et sous le contrôle de facteurs indirects (découplage transcription-traduction, par exemple). Toutefois, il ne peut pas être complètement exclu que des règles précises de séquence existent qui n'auraient pas été clairement identifiées, par exemple parce qu'elles sont très sophistiquées et difficilement détectables avec les approches mises en œuvre. Ainsi, de nombreuses études *in vitro* ont reposé sur l'utilisation de substrats artificiels (Ciampi, 2006; Guerin et al., 1998; Zhu and von Hippel, 1998) qui ne récapitulent pas forcément bien toutes les caractéristiques de la terminaison Rho dépendante. De plus, les études génomiques (approche ChiP-array) et transcriptomiques soufrent d'imprécision dans la localisation des sites et/ou la sensibilité de détection qui ne permettent pas d'en exploiter tout le potentiel. Dans ce cadre, il est par exemple regrettable que les jeux de *loci* Rho-dépendants identifiés chez *E. coli* varient très significativement suivant les études (Peters et al., 2012; Peters et al., 2009; Raghunathan et al., 2018; Sedlyarova et al., 2016).

Pour tenter de pallier ces défauts et offrir de nouvelles perspectives sur un mécanisme de régulation majeur, j'ai recherché de nouveaux déterminants de la terminaison Rho-dépendante en utilisant notamment des approches expérimentales inédites que j'ai contribué à développer. Au cours de ma thèse, Je me suis focalisé sur trois questions principales :

1) Est-il possible de prédire les sites de terminaison de la transcription Rho-dépendante ?

Nous avons émis l'hypothèse qu'en absence d'éléments de séquence consensus forts, une démarche de type QSAR (*Quantitative Structure Activity Relationship*) (pour revue : (Chen et al., 2015; Lo et al., 2018)) pourrait peut-être permettre d'établir une relation utile entre activité de terminaison Rho-dépendante et « contenu » de la séquence ADN transcrite. La méthode QSAR est particulièrement utilisée dans l'industrie pharmaceutique pour tenter de prédire la réponse pharmacologique de nouveaux composés à partir de relations empiriques utilisant des descripteurs moléculaires et/ou chimiques standardisés (présence de groupements fonctionnels précis, moment dipolaire, logP, etc.). Un choix adapté de descripteurs et une base de données expérimentales (à partir de composés déjà testés) suffisante pour établir la relation empirique sont des éléments essentiels au succès de cette approche chimio-métrique. En m'inspirant de celle-ci, j'ai préparé une grande banque de matrices ADN (104 séquences distinctes) dont j'ai testé *in vitro* la capacité à éliciter la terminaison Rho-dépendante afin d'établir une base de données robustes et homogènes. Puis, avec un bio-informaticien de l'équipe (Dr. Eric Eveno), nous avons développé une série de descripteurs de séquence ADN et utilisé des approches statistiques multivariées pour tenter de corréler la réponse « terminaison » des matrices ADN aux valeurs des descripteurs. Cette stratégie ressemblant à QSAR s'est avérée payante et nous a permis de bâtir le premier modèle de la terminaison Rho-dépendante. Ce modèle permet de prédire la réponse « terminaison » à l'échelle de génomes entiers avec un taux de succès de 85 % (taux évalué par validation croisée avec le jeu de matrices test). Ce travail est détaillé dans les pages suivantes sous la forme d'un article paru dans le journal « *Nucleic Acids Research* » (**Article I**, page 69).

Bien que représentant une avancée très significative, notre modèle prédictif souffre de défauts (puissance statistique modérée ; ~35 % de séquences 'hors modèle', de réponse non prédictible) liés à la taille de l'échantillon test (104 matrices ADN de ~500 bp en moyenne, soit ~60 kB total) relativement modeste comparée à celle d'un génome bactérien (~5 Mb). Bien que la méthode de caractérisation de la terminaison Rho-dépendante que j'ai utilisé soit la plus simple des approches existantes, elle reste trop couteuse et chronophage (essentiellement en raison de l'analyse des produits de réaction par électrophorèse sur gel) pour envisager d'accroitre la taille de l'échantillon test de façon significative. Dans le

deuxième volet de mon apport personnel (**Travaux non publiés I**, page 121), je présente une nouvelle méthode que j'ai développé qui repose sur un principe de détection fluorogénique de la terminaison Rho-dépendante compatible avec l'utilisation d'un lecteur de microplaques et qui devrait permettre d'améliorer sensiblement le flux de caractérisation de nouvelles matrices ADN.

2) Peut-on adapter la méthode « Clip-seq » pour étudier Rho et identifier précisément les sites *Rut* utilisés *in vivo* ?

Notre modèle prédit la réponse « terminaison » de régions génomiques (le génome étant scanné avec une fenêtre « glissante » de ~500 nt) mais ne renseigne pas sur les sites *Rut* ou *tsp* précisément utilisés par Rho au sein de ces régions (cf : **Article I**, page 69). Une identification de ces sites (ou au moins d'une partie de ceux-ci) permettrait sans doute de mieux apprécier les déterminants de la terminaison Rho-dépendante. Une étude récente suggère qu'une fraction des sites *tsp* peut être identifiée en comparant les profils RNAseq obtenus en présence ou non de bicyclomycine dans une souche sauvage avec ceux obtenus pour les mutants simples d'exoribonucléases (Dar and Sorek, 2018). Nous pensons qu'une fraction des sites *Rut* pourrait également être identifiée précisément en utilisant une approche transcriptomique appelée Clip-seq (*Cross-Linking Immunoprecipitation coupled to high-throughput sequencing*) (pour revue : (Konig et al., 2012)). Cette approche repose sur le pontage irréversible *in cellulo* (par photo-activation UV) d'une protéine d'intérêt à ses sites d'interaction ARN et sur l'identification par RNAseq de ces sites. Bien que simple dans le principe, cette approche est complexe et délicate à mette en œuvre et n'est pas adaptée à toutes les protéines liant l'ARN (Cook et al., 2015). Dans le troisième volet de mon apport personnel, je présente les travaux que j'ai mené en collaboration avec l'équipe du Dr. Nara Figueroa-Bossi (Gif-sur-Yvette) pour tenter d'adapter l'approche Clip-seq à l'étude des sites *Rut* chez *Salmonella*. Je discute les avancées mais également les difficultés rencontrées et les verrous techniques qu'il reste à lever pour étudier Rho par Clip-seq avec succès (**Travaux non publiés II**, page 137).

3) Peut-on identifier de nouveaux terminateurs Rho-dépendants conditionnels à partir des données « omiques » publiées et/ou des prédictions générées au laboratoire ?

L'avènement des méthodes « omiques » a conduit à la production de métadonnées qui restent largement sous-exploitées. Ainsi, c'est l'exploration des métadonnées générées pour Rho par le groupe de Bob Landick (Peters et al., 2012; Peters et al., 2009) qui a conduit notre équipe (en partenariat avec celle de Nara Figueroa-Bossi) à identifier un nouveau mécanisme de régulation conditionnelle dans la région 5'UTR de l'opéron *pgaABCD* d'*E. coli* (Figueroa-Bossi et al., 2014) (Voir 52 pour une description détaillée du mécanisme). Encouragés par ce succès, nous avons continué à explorer ces métadonnées (et celles générées avec notre modèle prédictif) à la recherche d'autres exemples de régulation conditionnelle mobilisant Rho, en nous focalisant particulièrement sur les régions 5'UTR d'*E. coli* de et de *Salmonella*. Les résultats de cette démarche sont présentés dans les deux dernières sections de mon apport personnel. Je décris tout d'abord l'adaptation d'essais classiques de terminaison de la transcription pour tester l'implication éventuelle de sRNA dans la régulation de l'activité de Rho. Je présente l'exemple de la région 5'UTR de *rpoS* au sein de laquelle nous avons découvert un terminateur Rho-dépendant conditionnel qui est inhibé par plusieurs sRNA s'appariant au 5'UTR de l'ARNm *rpoS*. Ces données sont présentées sous la forme d'un article paru dans *Methods in Molecular Biology* (**Article II**, page 149). Malheureusement, d'autres équipes ont fait la même découverte qu'elles ont décrit avant nous dans un journal à fort impact (Sedlyarova et al., 2016) (voir aussi page 58 de l'introduction). Dans la dernière section, je décris les autres régions génomiques que j'ai étudié à la recherche d'une régulation conditionnelle de Rho, en détaillant particulièrement les cas du gène *cspA* (**Travaux non publiés III**, page 171). Malheureusement, je n'ai pas trouvé d'éléments fortement en faveur d'une régulation conditionnelle pour ces *loci* particuliers. Néanmoins, pendant la même période, d'autres équipes ont mis en évidence de nouveaux mécanismes de régulation Rho-dépendante dans des régions 5'UTR (gènes : *mgtC*, *corA*, *tufB, thiB*, *thiC*, *thiM*, *lysC*, *nadD*) (Bastet et al., 2017; Brandis et al., 2016; Chauvier et al., 2017; Gall et al., 2018; Kriner and Groisman, 2015; Sedlyarova et al., 2017; Sevostyanova and Groisman, 2015), confirmant ainsi la validité de notre hypothèse de travail mais illustrant également la forte concurrence qui s'est maintenant établie dans ce domaine.

I. Article I:

A multivariate prediction model for Rho-dependent termination of transcription

Nucleic Acids Research (2018)

Nucleic Acids Research, 2018 **1**
doi: 10.1093/nar/gky563

A multivariate prediction model for Rho-dependent termination of transcription

Cédric Nadiras[1,2,†], Eric Eveno[1,†], Annie Schwartz[1], Nara Figueroa-Bossi[3] and Marc Boudvillain[1,*]

[1]Centre de Biophysique Moléculaire, CNRS UPR4301, rue Charles Sadron, 45071 Orléans cedex 2, France, [2]ED 549, Sciences Biologiques & Chimie du Vivant, Université d'Orléans, France and [3]Institute for Integrative Biology of the Cell (I2BC), CEA, CNRS, University of Paris-Sud, University of Paris-Saclay, Gif-sur-Yvette, France

Received April 11, 2018; Revised May 23, 2018; Editorial Decision June 07, 2018; Accepted June 08, 2018

ABSTRACT

Bacterial transcription termination proceeds via two main mechanisms triggered either by simple, well-conserved (intrinsic) nucleic acid motifs or by the motor protein Rho. Although bacterial genomes can harbor hundreds of termination signals of either type, only intrinsic terminators are reliably predicted. Computational tools to detect the more complex and diversiform Rho-dependent terminators are lacking. To tackle this issue, we devised a prediction method based on Orthogonal Projections to Latent Structures Discriminant Analysis [OPLS-DA] of a large set of *in vitro* termination data. Using previously uncharacterized genomic sequences for biochemical evaluation and OPLS-DA, we identified new Rho-dependent signals and quantitative sequence descriptors with significant predictive value. Most relevant descriptors specify features of transcript C>G skewness, secondary structure, and richness in regularly-spaced 5′CC/UC dinucleotides that are consistent with known principles for Rho-RNA interaction. Descriptors collectively warrant OPLS-DA predictions of Rho-dependent termination with a ∼85% success rate. Scanning of the *Escherichia coli* genome with the OPLS-DA model identifies significantly more termination-competent regions than anticipated from transcriptomics and predicts that regions intrinsically refractory to Rho are primarily located in open reading frames. Altogether, this work delineates features important for Rho activity and describes the first method able to predict Rho-dependent terminators in bacterial genomes.

INTRODUCTION

In bacteria, the Rho factor mediates one of the two major pathways that lead to the termination of transcription (reviewed in ref. (1–3)). Rho is a ring-shaped homohexameric enzyme with RNA-dependent ATP hydrolase activity. This ATPase activity fuels the translocation of Rho along nascent transcripts and the ensuing displacement of RNA polymerases [RNAPs] halted along the DNA template at Rho-dependent termination sites (Figure 1A). Rho-induced dissociation of transcriptional complexes contributes to the orchestration of gene expression at the genome scale but also helps maintain genome integrity by preventing conflicts between the transcription and replication machineries (3–6). Rho activity is tightly controlled *in vivo* and usually requires uncoupling of translation from transcription (or lack of translation) which exposes nascent RNA *Rut* (Rho utilization) sites to binding by Rho (Figure 1A). RNA capture is thus the primary determinant of Rho action (7–10) even though subsequent steps along the termination pathway can also be subjected to tight regulation (11,12).

Despite the importance of the initial RNA binding step, there are no known rules or consensus features that allow precisely defining (and detecting) *Rut* sites. It has been proposed that Rho-dependent termination sites lie downstream from so-called C>G 'bubbles', i.e. strand regions rich and poor in C and G residues, respectively (13) (Figure 1B). This agrees well with biochemical data showing that single-stranded C-rich-and-G-poor RNA ligands can proficiently bind and activate Rho (reviewed in (14)), thereby minimally defining the composition of *Rut* sites within nascent transcripts. The compositional bias can be better understood in the light of crystal structures where the Rho hexamer binds short C-rich oligonucleotides using a specific 5′-YC (Y being a pyrimidine) binding cleft located in the N-terminus of each Rho subunit (15,16). These structures, and a wealth of additional experimental data, support the idea that 60–80 nucleotides (nt) of single-stranded RNA are required

*To whom correspondence should be addressed. Tel: +33 238 25 55 85; Fax: +33 238 63 15 17; Email: marc.boudvillain@cnrs.fr
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

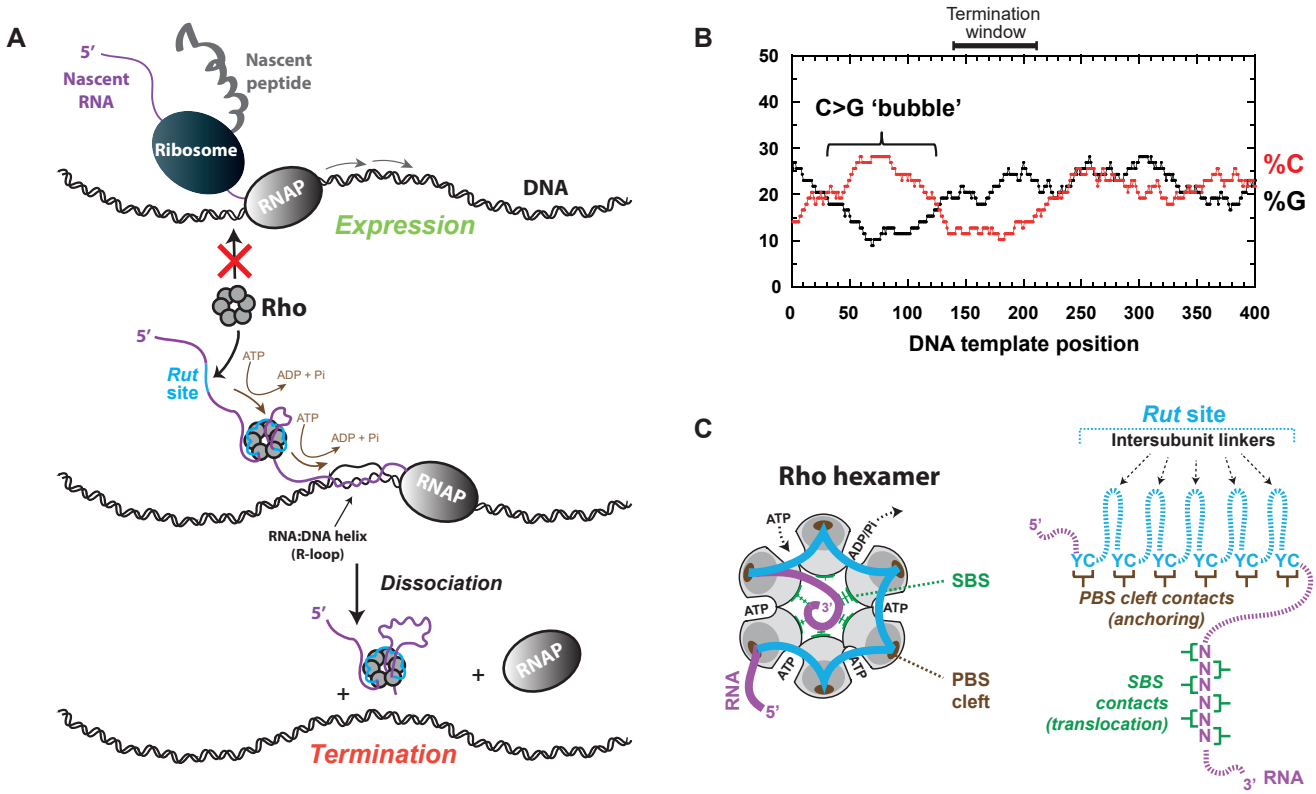


Figure 1. Rho-dependent termination of transcription. (**A**) Schematic representation of the termination process. Putative contacts between Rho and RNAP (74) are not depicted. Transcriptional R-loops, sometimes formed behind RNAPs, are also dissociated by the Rho factor (68), as depicted. (**B**) Genomic regions encoding Rho-dependent terminators usually contain a C>G bubble upstream from the termination sites (10,13,29,37). The case of the *pgaA* terminator from *E. coli* is shown as a representative example (10). (**C**) Configuration of the RNA interaction network within the Rho hexamer based on crystal structures of *E. coli*'s Rho (16,75). The tethered tracking mechanism used by Rho implies that PBS-RNA contacts are preserved while other contacts change as the RNA chain is translocated within the SBS, leading to the progressive lengthening of the PBS→SBS linker (see (76) and references therein).

to span the entire N-terminal periphery of the Rho ring, which composes the Primary Binding Site (PBS), including 9–13 nucleotide (nt)-long 'linkers' between the 5′-YC binding PBS clefts of adjacent subunits (16–19) (Figure 1C). Little else is known about *Rut* sites except that their effectiveness might be increased by the presence of pyrimidine residues in the intersubunit linkers (20), which themselves might be advantageously replaced by hairpin-like RNA motifs (21,22).

One possible reason for the lack of more precise rules to define *Rut* sites is that previous biochemical/biophysical studies have focused on a very limited set of Rho-dependent terminator sequences or have used artificial RNA ligands that may not fully recapitulate *Rut* features (1,14). The lack of reliable predictive tool is regrettable given that bacterial genomes can harbor hundreds of Rho-dependent terminators, the abundance of which requires elaborate transcriptomics/genomics strategies for discovery (23–28). For instance, ∼1300 Rho-dependent termination loci have been uncovered by transcriptomics in the genome of the *Escherichia coli* MG1655 strain (25). Even in this case, however, the precise location of the terminators (or *Rut* sites) could not be determined with certainty due to the endogenous trimming of the transcripts by essential exoribonucleases (25). Moreover, this type of approach is not ideal to detect terminators regulating low-abundant transcripts and probably misses conditional terminators that are active (and thus detectable) only under specific growth conditions (many effector-controlled Rho-dependent termination mechanisms have now been identified (9,10,28–35)). The great genomic variations among bacterial strains (notably pathogenic ones) due to horizontal transfers, mutations, and recombination events (36) are also expected to impact the Rho-dependent transcriptome (23,24,37) in proportions that would be too burdensome to systematically investigate *in vivo*. Thus, a method able to quickly outline genome-wide Rho-dependent termination in a given strain (or species) as efficiently as the algorithms developed to detect intrinsic (Rho-independent) terminators (38–43) could prove particularly useful.

In this work, we show that it is possible to develop such a method. Based on the information described above, we hypothesized that the major fraction of Rho-dependent terminators present in a given genome can be detected by seeking C>G bubble regions containing adequately spaced 5′-YC dinucleotides; only the minor fraction of Rho-dependent terminators activated by the NusG cofactor (∼20% in *E. coli* MG1655), which lack a canonical *Rut* site (25,44), may escape detection. Likewise, the rules described here may not apply to species harboring Rho factors too divergent from that of *E. coli*. Notably, caution should be exerted with the ∼35% of Rho factors (in different

species) that contain N-terminal domain (NTD) insertions (45) susceptible of altering PBS specificity (46–48). With this in mind, we focused our analysis on the genomes of two representative γ-Proteobacteria, *E. coli* MG1655 and *Salmonella enterica* LT2, whose Rho factors share 99.5% (417/419 residues) sequence identity (10). We systematically searched both genomes for the presence of C>G bubbles and $[(YC)N_{9\rightarrow13}]_{1\rightarrow6}$ sequence motifs (where N is any nucleotide). Based on this screen, we selected 104 genomic regions of diverse compositions (having a length of ∼500 base pairs [bp], on average) that we tested for the presence of Rho-dependent signals using *in vitro* transcription termination assays. The vast majority of the regions devoid of C>G bubbles (28 out of the 32 tested) showed no detectable *in vitro* termination activity, supporting the idea that a 'naked' RNA is not a sufficient condition to initiate Rho-dependent termination. Significantly, we also found that not all C>G bubbles promote Rho-dependent termination and, using multivariate statistical analysis and classification approaches, we identified explanatory variables related to the size, length, and YC-content of the C>G bubbles that have significant predictive value for Rho-dependent termination. Other major explanatory variables stemming from our analysis include the occurrence of small G-rich motifs within the non-template DNA strand as well as the template potential to encode stable RNA secondary structure. Using this information, we have built a multivariate prediction model for Rho-dependent termination based on OPLS-DA classification. The model provides reliable Rho-dependent termination prediction scores for major fractions of the MG1655 and LT2 genomes. Interestingly, sequences predicted to be refractory to Rho are overwhelmingly located within open reading frames where they may contribute to protect gene expression from unwanted transcription termination. Conversely, regions of 'Strong' termination probability are frequently located antisense of genes, consistent with Rho involvement in suppressing pervasive antisense transcription (25). Importantly, most (∼90%) of the Rho-dependent termination loci identified by transcriptomics (25) fall within regions of 'Strong' termination probability within the model-fitting portion of the MG1655 genome, thereby confirming the value of our approach. The model also predicts many termination-prone regions than were not anticipated from previous work, suggesting that additional layers of Rho-dependent regulation remain to be characterized. Taken together, the biochemical and computational data presented here identify new Rho-dependent termination signals in the *E. coli* and *Salmonella* genomes and provide the first predictive model for the automated detection of Rho-dependent terminators in bacterial genomes.

MATERIALS AND METHODS

Materials

Unless specified otherwise, chemicals and enzymes were purchased from Sigma-Aldrich and New England Biolabs, respectively. Nucleoside triphosphates and radionucleotides were purchased from GE-Healthcare and PerkinElmer, respectively. Synthetic oligonucleotides were obtained from Eurogentec. Rho protein was prepared and purified as described previously (49). Rho concentration is expressed in hexamers throughout the manuscript.

Preparation of the DNA templates

DNA templates used in *in vitro* transcription reactions were prepared by standard PCR procedures, as described previously (50). Briefly, each DNA template containing a specific genomic region downstream from the strong pT7A1 promoter was prepared in two successive PCR rounds by amplification of genomic DNA from *E. coli* MG1655 or *S. typhimurium* LT2 using synthetic DNA primers (see Supplementary Table S1 for details). Templates were purified with the GeneJET PCR purification kit (Thermo Fisher Scientific) followed by G50 size exclusion chromatography and their sizes were verified by 1.5% agarose gel electrophoresis.

Transcription termination experiments

Standard transcription termination experiments were performed as described previously (12) with minor modifications. Briefly, DNA template (0.1 pmol), *E. coli* RNAP (0.45 pmol), Rho (1.4 pmol), and Superase-In (0.5 U/μl; Ambion) were mixed in 18 μl of transcription buffer (40 mM Tris–HCl, pH 8.0, 50 mM KCl, 5 mM $MgCl_2$, 1,5 mM DTT) and incubated for 10 min at 37°C (in control reactions, Rho was omitted). Then, 2 μl of initiation mix (2 mM ATP, GTP and CTP, 0.2 mM UTP, 2.5 μCi/μl of ^{32}P-αUTP and 250 μg/ml of rifampicin in transcription buffer) were added to the reaction mixtures before further incubation for 20 min at 37°C. Transcription reactions were stopped by adding 4 μl of EDTA (0.5 M), 6 μl of tRNA (0.25 mg/ml) and 80μl of sodium acetate (0.42 M) before precipitation at –20°C with 330 μl of ethanol. Reaction pellets were dissolved in denaturing loading buffer (95% formamide, 5 mM EDTA), and analyzed by denaturing 7% polyacrylamide gel electrophoresis (PAGE) and Typhoon-9500 imaging (GE-Healthcare). Note that, due to the different internal labeling of the terminated and runoff transcripts, the efficiency of Rho-dependent termination cannot be precisely measured with this assay. Termination signals were thus categorized as 'None', 'Weak' or 'Strong' based on the visual changes in the transcription profiles induced by Rho (see legend to Figure 3 for details).

Ideally, termination efficiencies are determined from single-run transcription experiments wherein uniformly-labeled, stalled TECs are 'chased' with a mixture of unlabeled NTPs (51). However, the method is tedious and unpractical for the analysis of a large set of DNA templates and often requires template-specific modifications of the upstream region of the DNA template (to halt TECs) which may themselves alter the termination signals. On-beads transcription experiments require the same type of modifications and were thus only performed with selected DNA template controls (See results). To this effect, the sequences of the T049, T077 and T091 templates were modified to include the sequence 5′-AGATTGTATATGGTA AT between the T7A1 promoter and genomic sequences. This modification allowed preparation of TECs halted at position +26 (or +27 for the modified T091) of the templates. A 5′-biotinylated forward primer was also used dur-

ing PCR amplification of the DNA templates to allow immobilization of the TECs on streptavidin-coated beads. Preparation of the bead-immobilized TECs and transcription chase reactions, with or without Rho, were then performed as described previously (10,48). We note that, to date, single-run transcriptions with bead-affixed TECs always supported the idea that the Rho-dependent signals observed with our standard transcription termination assay stem from termination events (this work and ref. (10,48)), which attests to the robustness of this simple assay.

All standard and bead-immobilized transcription termination experiments were at least performed twice with each DNA template. For DNA templates yielding no (or very weak) termination signals in the presence of Rho, experiments were also repeated with completely distinct batches of reactants and buffers to rule out the presence of inhibitory contaminants.

Bioinformatic analysis and model building

Reference genomes used for bioinformatics analyses are U00096.3 for *E. coli* MG1655 and NC_003197.1 for *Salmonella* LT2.

Dedicated scripts for the detection of C>G bubbles and $[(YC)N_{9\rightarrow13}]_n$ motifs in genomic sequences and for the production of sequence descriptors for a given DNA template (or genomic region) were written in language Python 2.7 (the definitions of the descriptors and Python scripts are provided in Supplementary information). The C>G bubbles were identified using a 78-nt sliding window, as described previously (13). DNA regions were defined as C>G bubbles if they comply with %C ≥ %G (percentages calculated over the 78-nt window) at all positions and with %C > %G at one position at least. Regions were considered devoid of C>G bubbles if they comply with %C ≤ %G at all positions.

Minimum RNA folding free energies were determined with the Mfold software (52) using a locally installed 3.2 version and the RNA Quickfold option (standard RNA setting rules: 37°C; 1 M Na^+, 0 M Mg^{2+}; structures: 5% suboptimal, default window size, 100 folding max; no limit on maximum distance between paired bases).

The 104 DNA templates tested *in vitro* were divided into three classes ('Strong', 'Weak' or 'None') according to their capacity to elicit Rho-dependent termination (see legend to Figure 3 for details). The capacity of each sequence descriptor to discriminate between classes was evaluated with ANOVA (Analysis of Variance) and post-hoc Student–Newman–Keuls ($P \leq 0.05$) significance tests using Kaleidagraph 4.0 software (Synergy). Multivariate analyses, including Principal Component Analysis (PCA), Projection to Latent Structures Discriminant Analysis (PLS-DA), and OPLS-DA were performed with SIMCA 14.1 software (Umetrics) using a training set composed of the termination responses ('Strong', 'Weak' or 'None') measured with the 104 DNA templates used in this work (Supplementary Table S1). Explanatory variables (descriptors) were centered and auto-scaled to unit variance (UV scaling) using the default settings of SIMCA. The significance of PCA, PLS-DA and OPLS-DA components and the performance of resulting models were tested by jackknife cross validation (CV) with the training set as implemented in SIMCA (1/7th of the data held out and used as test set per CV round). Quality assessment for fit and prediction was based, respectively, on values of R^2, the fraction of the Sum of Squares (SS) explained by the component (or R^2_{cum} for the several components of a model), and Q^2, the fraction of the total variation of 'descriptors' or 'response' (i.e. dummy Y_{pred} variables in (O)PLS-DA) that can be predicted by a component (Q^2_{cum} for several components). The reliability and degree of overfitting of the (O)PLS-DA models were assessed, respectively, with CV-ANOVA of the cross-validated predictive residuals (53) and permutation plots (100 permutations) as implemented in SIMCA. Receiver Operating Characteristic (ROC) curves were calculated in SIMCA with the full training set of 104 observations and distinct moving threshold parameters for PCA-class (proprietary $P_{Mod}XPS$ probability) and (O)PLS-DA (dummy Y_{pred} response variable) models. For validated models with $Q^2_{cum} < 0.5$, we verified that the quality parameters remain stable after permutation of the rows in the dataset, as recommended (54).

Following guidelines from SIMCA's manufacturer, sequences with a 'probability of model membership' (proprietary $P_{mod}XPS^+$ parameter in SIMCA) lower than 0.05 (95% confidence level) were considered to be OPLS-DA model outliers ('out-of-model' sequences). Methods used to define the predicted 'None', 'Weak', 'Strong' and 'Out-of-model' regions for genomic predictions as well as template sequence descriptor values used for PCA and (O)PLS-DA model training are provided in Supplementary Information.

RESULTS

Genomic regions devoid of C>G bubbles do not contain significant Rho-dependent signals

Rho-dependent terminators often trigger transcript release at multiple, heterogeneous sites (1). It is usually unclear whether such a wide termination window reflects a diversity of potential Rho 'entry' points on the transcript (for instance, due to loose *Rut* site rules), a distribution of transcriptional pause sites where Rho is able to dissociate RNAP, or a combination of both. Notwithstanding, analysis of a handful of Rho-dependent terminators suggested that termination windows are invariably located downstream from so-called C>G bubble regions where the non-template DNA strand is richer in C than in G residues (13) (Figure 1B). These C>G bubble regions were identified upon scanning of the non-template DNA strand with a 78 nt-long sliding sequence window, assumed to be an ideal *Rut* size (13). We have used this 'bubble' methodology (see Supplementary Figure S1 and methods) and other sequence descriptors to probe the importance of C>G skewness in Rho-dependent termination.

If the presence of a C>G bubble is a prerequisite for Rho-dependent termination (13), then DNA sequences devoid of C>G bubbles should not contain Rho-dependent signals. Surprisingly, this prediction has never been tested thoroughly despite the fact that many genomic regions are free of C>G bubbles (hereafter named 'C>G-free' regions). For instance, the *E. coli* MG1655 genome contains 3381 'C>G-

free' regions having a length of at least 200 bp (Supplementary Figure S1) while 'C>G-free' regions ≥ 100 bp represent 25.2% of the genome (proportions are similar for the *Salmonella* LT2 genome).

To test the assumption that 'C>G-free' sequences cannot elicit Rho-dependent termination (at least in the absence of NusG), we prepared thirty-two DNA templates containing a genomic 'C>G-free' region fused to the T7A1 promoter (Figure 2A). The sizes of these genomic segments ranged from 572 to 1130 bp, representing a total of 22 083 bp (Supplementary Table S1). The templates were used in standard *in vitro* transcription experiments in the presence or absence of Rho (see methods). For twenty-eight of the DNA templates, transcription patterns were not affected by the presence of the Rho protein (representative examples are shown in Figure 2B), indicating that the corresponding sequences do not encode strong *Rut* sites or Rho-dependent terminators. For the last four DNA templates (T003, T063, T074 and T104 templates), we detected shorter-than-runoff transcripts formed in low amounts in the presence of Rho (Figure 2C, black stars), implying that these templates contain weak Rho-dependent signals. Similar profiles were observed upon using ^{32}P-αCTP (instead of ^{32}P-αUTP) to label the transcripts, ruling out effects dependent on which NTP is at a subsaturating concentration in the transcription assay (see Materials and Methods). Using RSAT oligo analysis (55), we found that the T003, T063, T074, and T104 templates are significantly enriched in AT-rich motifs in the non-template strand (e.g. TTTAAA, TTTTA, TATT, TTAT; $P < 10^{-5}$) as compared to the other 'C>G-free' templates. These motifs may facilitate productive interactions between Rho and the nascent transcript (for instance, by limiting the formation of stable RNA secondary structures) but are unlikely to represent primary determinants of Rho activation given the weakness of the associated Rho-dependent signals (see Figure 2C and below). Hence, our results support the view that genomic regions devoid of C>G bubbles do not encode canonical *Rut* sites or strong Rho-dependent terminators.

Distribution of C>G bubbles and 5′-YC dimers in the E.coli and Salmonella genomes

Next, we looked at the distribution of C>G bubbles within the genomes of the *E. coli* MG1655 and *Salmonella* LT2 strains. We found that both genomes contain a high density of C>G bubbles evenly distributed (in numbers and sizes) between both genomic strands (Supplementary Table S2). For instance, the *E. coli* genome contains 20 882 C>G bubbles having a length of at least 80 bp (∼2.2 per kilobase). This is an order of magnitude higher than the number of Rho-dependent termination loci that were identified by transcriptomics (25). This difference is all the more notable considering that more than half of the aforementioned C>G bubbles (61%) are located antisense or outside of open reading frames (ORFs) where they should not be hindered by translating ribosomes.

We also scanned the MG1655 and LT2 genomes for the presence of groups of YC dimers adequately spaced for interaction with Rho's PBS. We limited the search to $[(YC)N_{9\rightarrow13}]_n$ sequences encoding RNA motifs containing 'intersubunit linkers' (Figure 1C) of 9 to 13 nt. This criterion is a tradeoff between the proposal (based on Rho crystal structures) that a length of 12–13 nt of single-stranded RNA is required to span adjacent subunit PBS clefts (56) and biophysical observations suggesting that even shorter linkers (9–10 nt) are compatible for binding (18,19). For instance, a footprint of (57 ± 2) nt of single-stranded RNA for the full PBS has been estimated from single-molecule force extension experiments (19). Assuming that the six PBS clefts were occupied by YC dimers, this leaves (45 ± 2) nt for the five intersubunit linkers (Figure 1C), i.e. (9 ± 0.4) nt per linker on average. Of note, our $[(YC)N_{9\rightarrow13}]_n$ motif search cannot detect special cases where hairpin-forming sequences can de facto bring the distance between YC dimers inside the accepted range (21,22).

As expected, the number of genomic $[(YC)\text{-}N_{9\rightarrow13}]_n$ motifs sharply decreases as a function of the number of repeats, *n* (Supplementary Table S3). There are 32,351 $[(YC)N_{9\rightarrow13}]_6$ sequences in the MG1655 genome (∼3.5 per kilobase on average) encoding RNA segments that could theoretically span the six Rho subunits (and thus the entire PBS). This is slightly less than for the LT2 genome (∼3.9 $[(YC)N_{9\rightarrow13}]_6$ motifs per kilobase) but still much more than the ∼1300 Rho-dependent termination loci (∼0.14 per kilobase) identified by transcriptomics (25). We note, however, that three of our 'C>G-free' templates unable to elicit Rho-dependent termination (T025, T056 and T088; see above) encode $[(YC)N_{9\rightarrow13}]_6$ motifs, suggesting that the presence of such presumably ideal motifs in transcripts is not a sufficient determinant for termination.

Next, we attempted to estimate the number of C>G bubbles containing $[(YC)N_{9\rightarrow13}]_n$ motifs that are sufficiently long to encode candidate *Rut* sites. We found that known Rho-dependent terminators contain C>G bubbles of at least ∼80 bp located upstream from (or comprising) the termination sites and that all of these bubbles contain at least one $[(YC)N_{9\rightarrow13}]_4$ motif (Supplementary Table S4). The later observation suggests that the RNA chain needs to contact at least four distinct Rho subunits to make a productive interaction with the hexameric enzyme. The MG1655 and LT2 genomes harbor >16 000 C>G bubbles with these characteristics (∼80 bp or longer and containing a $[(YC)N_{9\rightarrow13}]_4$ motif) (Supplementary Table S5), a number which is still much higher than that of experimentally validated terminators. This suggests that the number of Rho-dependent terminators is largely underestimated or that C>G bubbles with the aforementioned characteristics are not sufficient (or adequate) predictors for the occurrence of termination.

Testing the Rho-dependent termination potential of genomic C>G bubble regions

To better identify features defining a productive *Rut* site, we compared the Rho-dependent transcriptional responses of a large set of DNA templates probed under the same standard *in vitro* transcription conditions. We selected 72 genomic sequences (average length ∼490 bp, representing a total of 35 171 bp) from the *E. coli* and *Salmonella* genomes (Supplementary Table S1), each bearing at least one C>G bubble starting no closer than 130 bp from the downstream

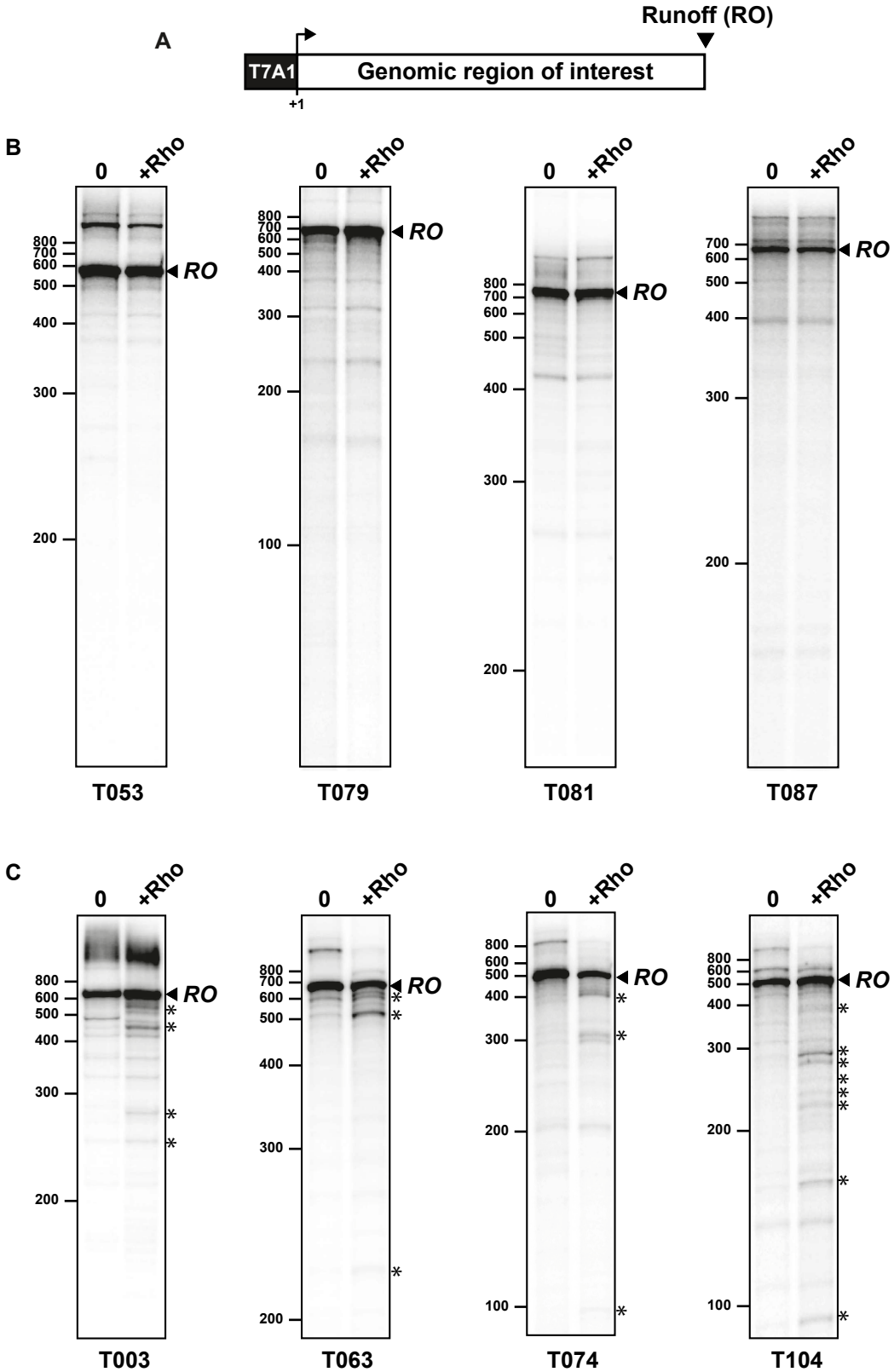


Figure 2. C>G-less genomic regions are devoid of significant Rho-dependent signals. (**A**) Schematic depiction of the DNA templates used in our standard *in vitro* transcription termination experiments. (**B**) Representative denaturing PAGE gels illustrate the absence of formation of Rho-specific truncated transcripts during transcription of C>G-less templates. The RO bands correspond to runoff transcripts. RNAs migrating more slowly than the RO bands result from template switching events (once RNAP has reached a template end; see (12) and references therein). (**C**) Representative transcription experiments with the four unusual C>G-less templates yielding low amounts of truncated transcripts (identified by stars next to the gels) in the presence of Rho.

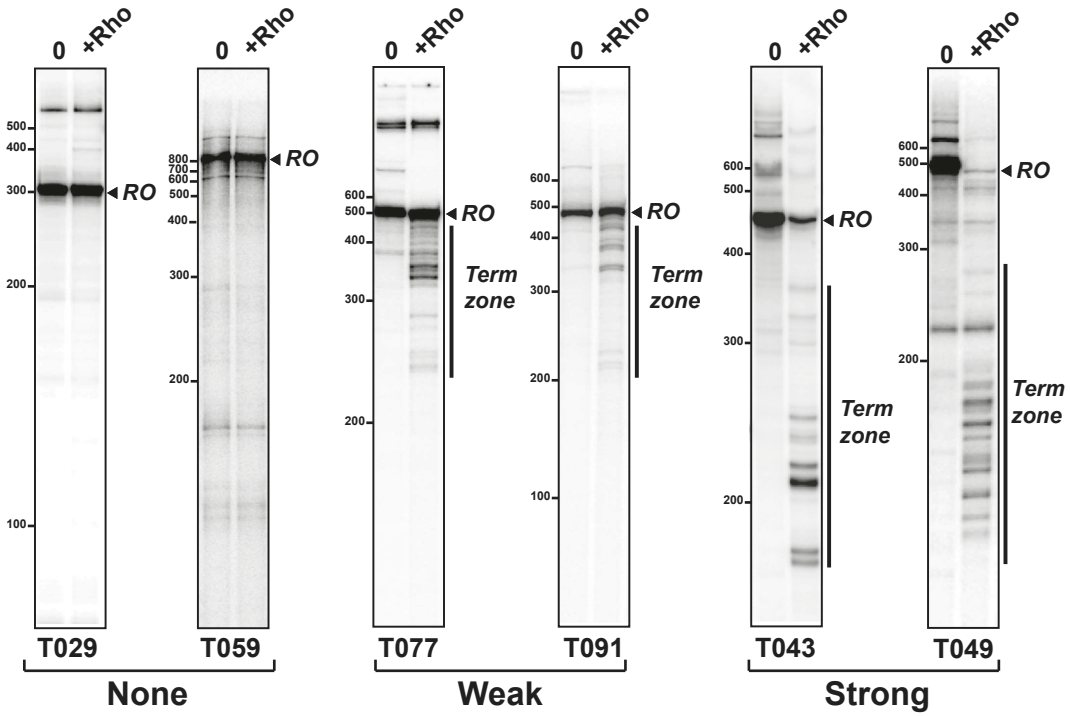


Figure 3. Rho-dependent termination signals encoded by DNA templates containing C>G bubbles. Representative denaturing PAGE gels illustrate the different classes of transcription termination signals. Rho did not change the transcription profiles for 5.6% of the tested 'C>G-plus' DNA templates ('None' category). Rho-dependent signals were considered 'Weak' when bands migrating faster than runoff transcripts appeared in the presence of Rho while the intensity of the 'runoff' band was hardly affected (47.2% of the 'C>G-plus' templates). They were considered 'Strong', when Rho elicited the appearance of fast-migrating bands and a sharp decrease of the intensity of the runoff band (47.2% of the 'C>G-plus' templates).

edge of the sequence (hereafter named 'C>G-plus' regions). These 'C>G-plus' regions were empirically selected, yet were relatively diverse in terms of C>G bubble length (ranging from 20 to 355 bp) and YC content (ranging from 0 to 18 $[(YC)N_{9\rightarrow13}]_4$ motifs per region). All 'C>G-plus' regions were fused to the T7A1 promoter (Figure 2A) and transcribed in the presence or absence of Rho as described above for the C>G-less templates. Rho had no detectable effect on the transcription of four of the 'C>G-plus' DNA templates (Figure 3, 'None' class). For the vast majority (~94%) of the templates, however, the presence of Rho induced the formation of shorter-than-runoff transcripts (representative examples are shown in Figure 3), consistent with the presence of Rho-dependent signals within the corresponding sequences. We empirically classified the signals as 'weak' or 'strong' (34 'C>G-plus' templates in each class) based on their apparent strengths estimated from transcription termination gel band profiles (Figure 3 and Supplementary Figure S2). Details on categorization of termination signals are provided in the legend to Figure 3.

We selected a few 'C>G-plus' DNA templates from the weak and strong classes to verify that the Rho-dependent signals arise from true termination events (rather than from Rho-induced transcriptional arrest). We modified the upstream sequences of the selected DNA templates in order to prepare halted transcription elongation complexes (TECs) immobilized on streptavidin-coated beads, which were then used in single-run 'chase' transcriptions (see methods). In all tested instances, we observed that the presence of Rho in the reaction mixture triggers the release of shorter-than-runoff transcripts in the supernatant (Supplementary Figure S3). This observation is consistent with the dissociation of the TECs upon transcription termination (whereas transcriptional arrest would not have disrupted the TECs).

Using RSAT oligo analysis (55), we found that the set of DNA templates unable to elicit Rho-dependent termination (T005, T029, T059, and T075) is significantly enriched in G-rich motifs in the non-template strand (e.g. CAGGG, GGGCA; $P < 10^{-4}$) as compared to the other 'C>G-plus' templates. However, many of these motifs are located in the T059 template alone and a significant enrichment is no longer detected ($P > 0.05$) if the T059 sequence is removed from the analysis. We note that the T059 template also contains the smallest C>G bubble of all 'C>G-plus' region templates (20 bp; Supplementary Figure S4) which could contribute to its 'unresponsiveness' to Rho. The other three 'unresponsive' templates, however, contain C>G bubbles that are no smaller than those found in some of the 'C>G-plus' region templates eliciting Rho-dependent termination. Notwithstanding, there appears to be some relationship between the length/area of the C>G bubble and the strength of the Rho-dependent signal to the point that all tested DNA templates containing C>G bubbles longer than 78 bp (or with an area over 700% × bp) elicit Rho-dependent termination (Supplementary Figure S4).

Seeking relevant sequence descriptors of Rho-dependent termination

To characterize the 'None', 'Weak' and 'Strong 'signals further, we systematically compared the 104 DNA templates (regardless of their C>G bubble content and discounting the T7A1 promoter region) using 111 distinct sequence explanatory variables, herein named 'descriptors' (see Supplementary methods and Supplementary Table S6). The percentages of individual monomers, dimers and trimers of nucleotides found in the non-template DNA strands (or runoff transcripts) make a first group of descriptors (larger motifs were not considered because none were evident from pairwise RSAT (55) comparisons of the termination classes of templates). A second group of descriptors provides scores for selected features of the C>G bubbles (length or area of the longest C>G bubble in template, cumulated length or area of all C>G bubbles in template, etc.). A third group of descriptors counts $[(YC)N_{9\rightarrow13}]_n$ motifs found in the C>G bubbles alone (none for 'C>G-free' templates) or in the full non-template DNA strands. The last descriptor represents the minimum free energy (per kilobase) for RNA secondary structure formation as determined for runoff transcripts with Mfold software (52).

Next, we examined how each individual descriptor is able to distinguish the three classes of templates (i.e. templates triggering 'None', 'Weak', or 'Strong' termination signals). Representative dotplots taken from this analysis are shown in Figure 4A. For 28 of the descriptors, we did not detect statistically significant differences among template classes (ANOVA, $F \leq 3$ and $P > 0.05$; Figure 4A and Supplementary Table S6). Among the 83 remaining descriptors, 27 descriptors were able to differentiate the three template classes (Student-Newman-Keuls *post hoc* test $P \leq 0.05$ for all pairwise comparisons), 49 descriptors were meaningful for only two pairwise comparisons, and 7 descriptors for only one pairwise comparison (Figure 4A and Supplementary Table S6). Among the 27 'most-differentiating' descriptors, 13 are C>G bubble descriptors (including numbers of $[(YC)N_{9\rightarrow13}]_1$ –aka YC dimers- and $[(YC)N_{9\rightarrow13}]_2$ motifs in the longest C>G bubble) while 14 are nucleotide (%C, %G), dinucleotide (%CA, %CT, %YC, %GA, %GT, %GG), and trinucleotide (%AAC, %ACA, %CTT, %CAC, %GTG, %GGT) sequence descriptors (Supplementary Table S6).

We observed that the minimal RNA folding energies calculated for the full-length transcripts are significantly higher for the 'Strong' and 'Weak' classes than for the 'None' class (Figure 4A and Supplementary Table S6). This suggests that the 'Strong' and 'Weak' transcripts are usually less structured than the 'None' transcripts, in agreement with models proposing that *Rut* sites are generally poor in RNA secondary structures (1–3). Differences between the 'Weak' and 'Strong' RNA folding energies are, however, not significant (Student–Newman–Keuls *post hoc* test $P > 0.05$). This holds true when RNA folding energies are calculated for sequences restricted to the longest C>G bubble regions (data not shown), suggesting that RNA structure content is not a discriminant between the 'Weak' and 'Strong' classes.

Overall, we found a number of sequence descriptors that vary significantly with the strength of the Rho-dependent signal (Supplementary Table S6), suggesting that Rho-dependent termination can be predicted upon scanning and detection of specific sequence features/motifs.

A predictive multivariate model for the presence of Rho-dependent termination signals

We cannot rule out that relevant descriptors have escaped detection due to insufficient statistical power or multivariate interactions. Increasing statistical power is difficult because it would require increasing sample sizes significantly, a task beyond the scope of the present work wherein an already large number (>100) of DNA templates have been characterized biochemically. By contrast, the multivariate structure of the data can be explored with statistical approaches such as Principal Component Analysis (PCA) (57). In PCA and related approaches, complex sets of variables, which are often correlated, are reduced into much smaller sets of uncorrelated variables called principal components (note that all original variables are used to build each principal component but differ in their respective 'loadings', i.e. component weights) (57,58). In this way, patterns of similarities among variables and/or observations are more easily identified and can be used to generate predictive models.

We first performed PCA with the whole set of 104 templates (i.e. observations) and 111 descriptors (i.e. variables) using the dedicated SIMCA software (see methods). The resulting PCA model contains four significant principal components which are able to separate 'None', 'Weak' and 'Strong' observations in fairly distinct populations even though some population overlap remains (Figure 4B and Supplementary Figure S5). As expected, PCA separation is most effective along principal component 1 for 'None' versus 'Weak' or 'Strong' populations (Figure 4B and Supplementary Figure S5), which is in large part due to C>G bubble and G-rich motif descriptors (Figure 4C).

The significant separation obtained with PCA (Figure 4B), where specific classes are not defined *a priori* (unsupervised approach), encouraged us to test several supervised classification approaches suited for prediction (58). Using the SIMCA software, we generated prediction models based on (i) PCA-class analysis where disjoint (one per class) PCA models are used, (ii) Projections to Latent Structures Discriminant Analysis (PLS-DA) and (iii) Orthogonal PLS-DA (OPLS-DA). The PLS-DA and OPLS-DA methods rely on similar principles to build (O)PLS components representing the best possible compromise between the description of the explanatory variables (here 'descriptors') and the prediction of the response (here, 'None', 'Weak' or 'Strong'). In OPLS-DA, orthogonal components are also produced to isolate the variation in the explanatory variables that is uncorrelated (orthogonal) to the response (58).

We found the PLS-DA method to be suboptimal, especially to predict the 'Weak' class, whereas PCA-class (Supplementary Figure S6) and OPLS-DA (Figure 5A) methods yield comparably high prediction efficiencies for the three Rho-dependent termination classes with overall success rates in the order of 85% upon jackknife cross-validation (Table 1). Since the OPLS-DA model yields slightly better ROC diagnostics than the PCA-class model (Supplementary Figure S7) and does not seem to be subjected to data

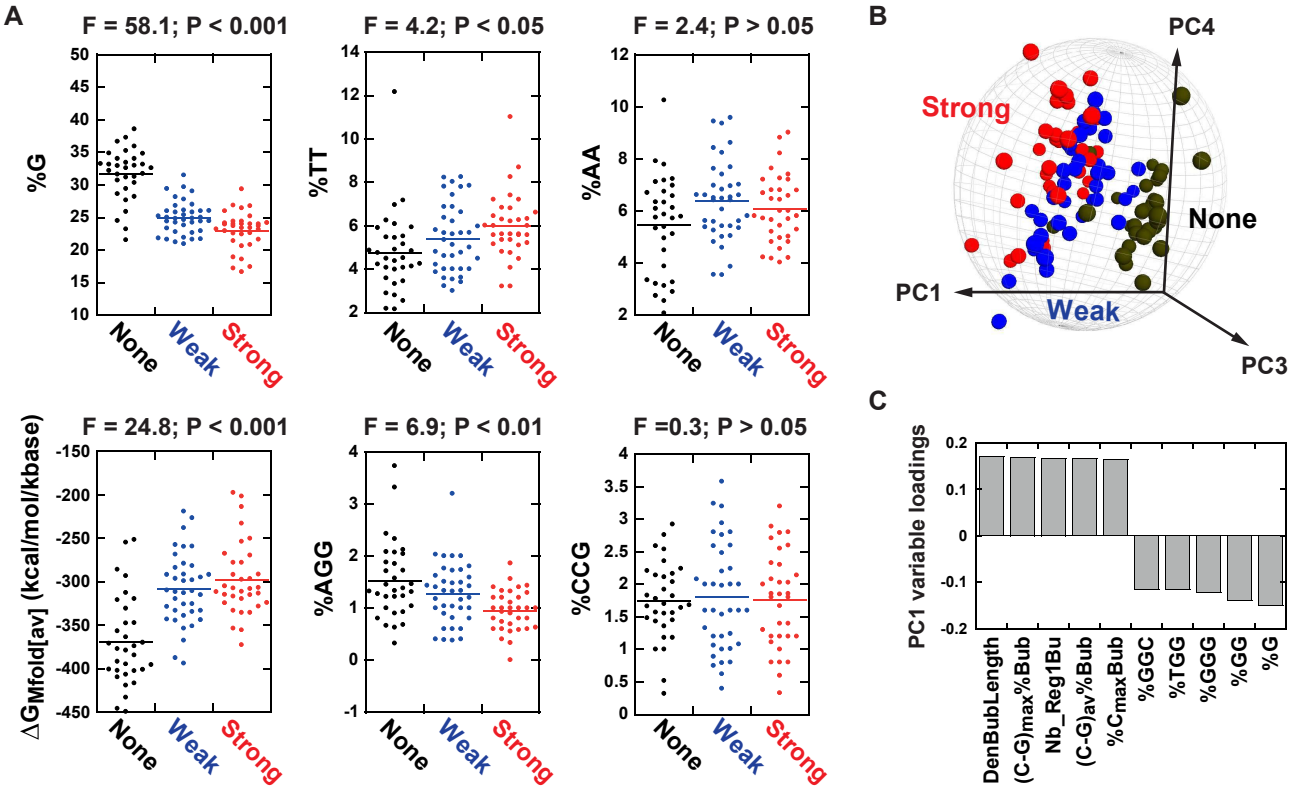


Figure 4. Statistical analysis of the Rho-dependent signals detected with the training set of DNA templates (Supplementary Table 1). (**A**) Dot plots for representative sequence descriptors. ANOVA *F*- and *P*-values are shown above each plot. (**B**) Unsupervised PCA 3D score plot obtained for the 104 DNA templates of the training set using the complete set of 111 descriptors. $Q^2_{cum} = 0.464$ and $R^2_{cum} = 0.561$ for the four PCA components. The gray sphere represents the Hotelling's $T^2 = 0.05$ limit. (**C**) The bar graph shows the best positive and negative variable loadings for the first principal component (PC1). DenBubLength: cumulated length of all C>G bubbles in non-template DNA strand relative to full strand length (density); $(C–G)_{max}$%Bub: maximal difference between %C and %G in longest C>G bubble of non-template DNA strand; Nb_Reg1Bu: Number of YC dimers in longest C>G bubble; $(C–G)_{av}$%Bub: average difference between %C and %G in longest C>G bubble; $\%C_{max}$Bub: highest %C in longest C>G bubble. %GGC, %TGG, %GGG, %GG and %G are percentages of respective motifs in the non-template DNA strand.

overfitting (Supplementary Figure S8) (54), we selected this model (OPLS-DA-111 model in Table 1 and Figure 5A) for subsequent analyses. We note that the descriptors of C>G bubbles and G-rich motifs and, to a lesser extent, RNA folding stability ($\Delta G_{Mfold[av]}$) and richness in $[(YC)N_{9\rightarrow13}]_{1\rightarrow3}$ motifs provide the largest descriptor contributions to the OPLS-DA-111 model (Supplementary Table S7). We also note that OPLS-DA models built with reduced sets of descriptors (including one with only C>G bubble descriptors) were outperformed by the OPLS-DA-111 model (higher misclassification rates upon jackknife cross-validation and poorer ROC diagnostics; see Table 1 and Supplementary Figure S7).

To further evaluate OPLS-DA modelling of Rho-dependent termination, we looked at OPLS-DA-111 predictions for a test set of known Rho-dependent terminators that were not used for model training (Supplementary Table S4). The OPLS-DA-111 model unambiguously assigned most of the test terminators (83.3%) to the 'Strong' category and none to the 'None' category (Supplementary Table S4), thereby illustrating the model predictive value. Taken together, these data support the idea that supervised multivariate (e.g. OPLS-DA) models can be built to predict the presence/absence of Rho-dependent termination signals within DNA sequences of interest.

Genome-wide OPLS-DA model prediction of Rho-dependent termination

We compiled OPLS-DA-111 probability scores as a function of genomic position for the complete *E. coli* MG1655 genome. To facilitate interpretation of the data, we divided both genome strands into successions of distinct regions wherein probability scores for either 'Out-of-model', 'None', 'Weak', or 'Strong' termination prevailed at all positions (see methods). Using this strategy, we delineated 8464 regions of interest (i.e. predicted 'Strong', 'Weak', and 'None' regions) representing ∼65% of the genomic positions (Supplementary Table S8). The predicted 'None' regions (i.e. regions where probability for 'no termination' is highest throughout) were markedly fewer and smaller than the predicted 'Weak' or 'Strong' termination regions (Figure 5B). This suggests that the portion of the MG1655 transcriptome that is refractory to Rho-dependent termination is limited. This 'refractory' portion likely corresponds to genomic regions devoid of extensive C>G bubbles. In line with this proposal, we observed that the predicted 'None' regions are characterized by a much smaller C>G bubble

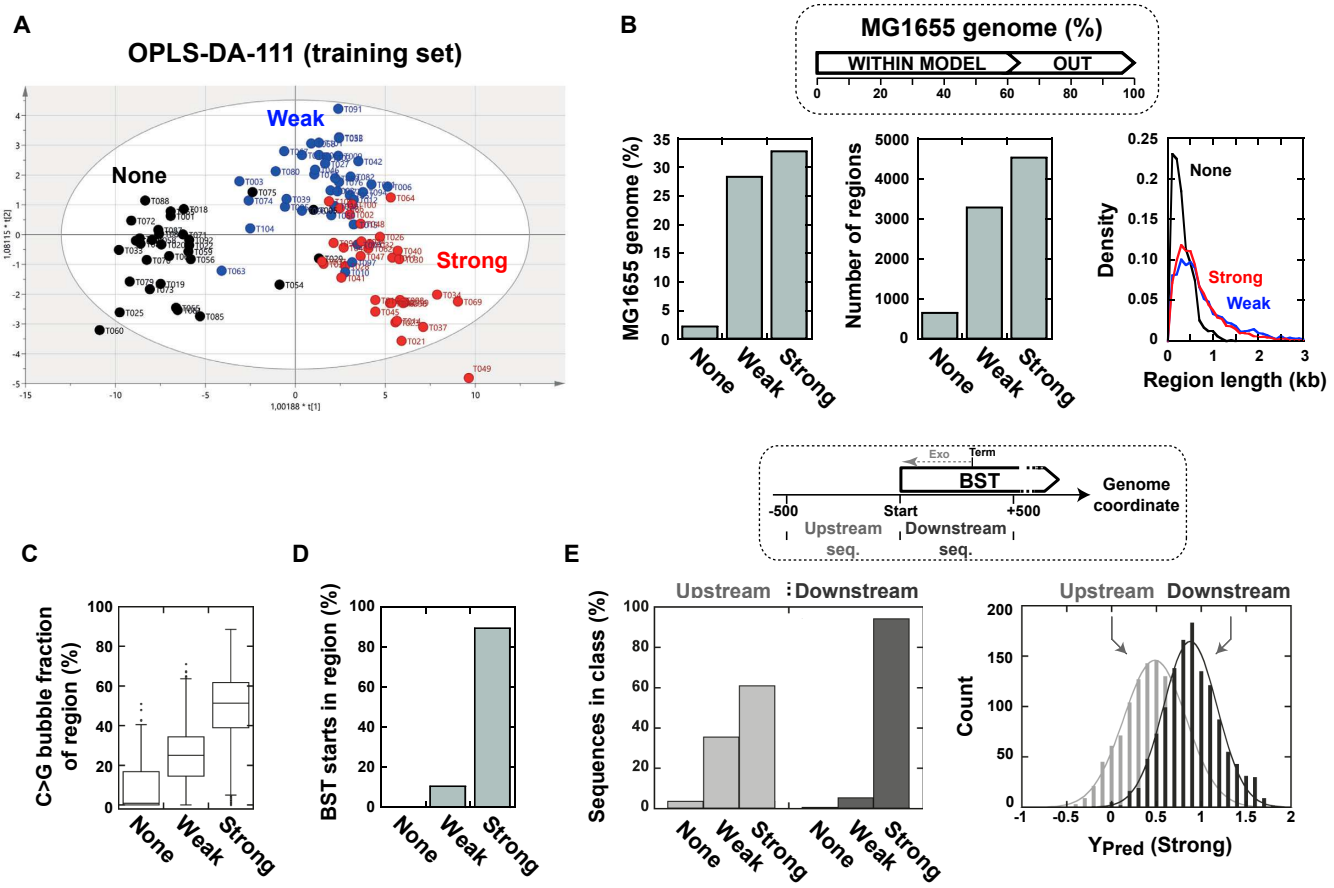


Figure 5. OPLS-DA modelling of Rho-dependent termination of transcription. (**A**) Standard 2D score plot obtained for the OPLS-DA-111 model with the training set of DNA templates. (**B**) Analysis of the *E. coli* MG1655 genome with the OPLS-DA-111 model predicts that regions refractory to Rho-dependent termination (Predicted 'None' regions) are fewer and smaller than regions eliciting termination (see Supplementary methods for definition of predicted regions). The proportions of genomic positions for which model predictions are reliable (in-model), or not (out-of-model), according to the $P_{mod}XPS^{+} = 0.05$ threshold are shown inset. (**C**) The C>G bubble content strongly correlates with the predicted termination strength of the regions (ANOVA $P < 0.001$; $F = 2713$). Regions as in panel B. (**D**) The Bicyclomycin Sensitive Transcript (BST) start points (25) fall overwhelmingly within predicted 'Strong' regions. (**E**) Comparison of the termination classes predicted with the OPLS-DA-111 model for 500 nt-long sequences located either upstream or downstream from BST start points (see diagram inset). The distributions of the predicted response scores (Y_{pred}) from SIMCA for the 'Strong' class are also shown.

content than the other predicted regions (Figure 5C). Similar observations were made upon analysis of the *Salmonella* LT2 genome with the OPLS-DA-111 model (Supplementary Figure S9 and Supplementary Table S8).

To assess the value of the OPLS-DA-111 predictions at the genome scale, we compared the map of predicted regions for the MG1655 genome (Supplementary Table S8) with that of Bicyclomycin Sensitive Transcripts (BSTs) identified by transcriptomics (25). BSTs arise nearby Rho-dependent termination loci owing to the inhibition of Rho activity by Bicyclomycin (25). About 60% of these BSTs can be used for comparison purposes as they begin within model-compatible regions ($P_{mod}XPS^{+} > 0.05$) of the *E. coli* genome (Supplementary Table S8). Remarkably, most of these BSTs (90%) begin within a 'Strong' termination region and none within a 'None' region (Figure 5D). This distribution differs significantly from a random distribution of BST start points among the predicted regions ($P < 10^{-5}$; Fisher's exact test).

Since BST start points are expected to be shifted upstream from the *bona fide* Rho-dependent termination sites due to posttranscriptional [3′→5′]-exonucleolytic trimming of the transcripts (25), we also compared OPLS-DA-111 prediction scores calculated for the 500 nt-long sequences located either upstream or downstream from the BST starts (see Figure 5E, inset). We found that the downstream sequences are more frequently and more strongly associated with the 'Strong' termination class than the upstream sequences (Figure 5D), which is consistent with 3′-exonucleolytic trimming of the Rho-dependent transcripts. Taken together, these data strongly suggest that a 'Strong' region status (as determined with the OPLS-DA-111 model) is a good prognostic for *in vivo* Rho-dependent termination. This status likely reflects a sequence composition that is intrinsically favorable to Rho activity but does not provide information on the potential presence of other regulatory signals, such as intrinsic terminators, within the same region. For instance, 2.5% of the 'Strong' regions predicted for the MG1655 genome also contain intrinsic terminators listed in regulonDB (59), accounting for ∼40% of the terminators in this database (Supplementary Table S8). Moreover, given their relatively large sizes (Figure 5B), some predicted

Table 1. Main features of the multivariate predictive models of Rho-dependent termination

						CV-ANOVA[b]		Classification of observations into known classes[c]					ROC area under curve (AUC)[e]			
Model		PCs[a]	R^2_{Xcum}	R^2_{Ycum}	Q^2_{cum}	*F*	*P*	Weak	Strong	None	Total	*P*[d]	Weak	Strong	None	Remark
PCA-class-111	N	3	0.546	n.a.	0.277	n.a.	n.a.	89.5%	79.4%	90.6%	86.5%	1.2×10^{-6}	0.861	0.928	0.956	One PCA per class with all 111 descriptors
	W	3	0.480	n.a.	0.302											
	S	5	0.580	n.a.	0.233											
PLS-DA-111		1	0.271	0.359	0.345	11.5	2.1×10^{-8}	**21.0%**	91.2%	90.6%	**65.4%**	3.0×10^{-7}	**0.578**	0.876	0.973	PLS-DA with all 111 descriptors
OPLS-DA-111		2 (2)[f]	0.513	0.639	0.410	4.4	2.1×10^{-7}	86.8%	82.3%	87.5%	85.6%	1.1×10^{-6}	0.927	0.968	0.989	OPLS-DA with all 111 descriptors
OPLS-DA-40		2 (1)	0.709	0.536	0.454	7.9	4.3×10^{-12}	71%	76.5%	84.4%	76.9%	9.6×10^{-7}	0.861	0.942	0.976	VIPs < 1 in OPLS-DA-111 removed[g]
OPLS-DA-83		2 (2)	0.572	0.627	0.453	5.9	2.4×10^{-10}	81.6%	82.4%	90.6%	84.6%	1.1×10^{-6}	0.918	0.965	0.991	OPLS-DA with only the 83 descriptors that pass ANOVA[g]
OPLS-DA-6		2 (0)	0.962	0,398	0,372	8.6	4.4×10^{-10}	44.7%	67.7%	96.9%	68.3%	8.7×10^{-7}	0.734	0.904	0.961	OPLS-DA with C>G Bubble descriptors only[h]

[a]Orthogonal Principal Components (PCs) are in parentheses when relevant. R^2_{Xcum} values do not include contributions from orthogonal PCs.
[b]ANOVA of the cross-validated predictive residuals as implemented in SIMCA software.
[c]Classification based on jackknife cross-validation (see methods).
[d]Fisher's probability of the classification occurring by chance.
[e]The AUC of the ROC curve varies from 0.5 (random prediction) to 1.0 (perfect prediction).
[f]Sources of orthogonal variance in OPLS-DA-111 uncorrelated to termination are discussed in Supplementary information.
[g]Although excluding low ranking descriptors based on ANOVA or VIP (variable Importance in the Projection) scores sometimes improves OPLS-DA predictions, this strategy was detrimental in our case (see also Supplementary Figure S5).
[h]Descriptors used were SumBubLength, SumBubSurf, DenBubLength, DenBubSurf, BublongLength and BublongSurf (see Supplementary information for details).

'Strong' regions are likely to bear several Rho-dependent signals. We thus cannot exclude that some Rho-dependent signals are silent unless transcription read through their upstream (intrinsic or Rho-dependent) terminator partner(s) occurs.

Interestingly, predicted 'Strong' regions are frequently found antisense to genes (Figure 6), which is consistent with a prevalent role of Rho-dependent termination in silencing antisense transcription (25). By contrast, 'None' and 'Weak' regions are primarily associated to intragenic sense sequences (Figure 6), most of which (~95%) correspond to open reading frames (not shown). Thus, the absence of strong Rho-dependent signals may contribute to protect the expression of some protein-coding genes, possibly to compensate for poor mRNA translation (which would limit mRNA shielding by ribosomes). In agreement with this proposal, we note that both Rho-dependent termination (Figure 4A) and mRNA translation (60) efficiencies are inversely correlated with mRNA folding stability.

DISCUSSION

Elucidation of sequence features governing Rho-dependent termination is notoriously difficult (1–3). Early attempts based on the meticulous dissection of known terminators uncovered very few consensus features (reviewed in ref. (1)). These features are: (i) a minimal transcript length of ~100 nt, (ii) the presence of a *Rut* site, which is at least ~80 nt-long and usually rich in C residues and poor in G residues, upstream from the termination endpoint(s), (iii) the paucity of RNA secondary structure in the *Rut* site region, and (iv) the requirement for signals (of highly variable compositions) triggering RNA pausing at the termination sites. Recent genomics and transcriptomics analyses of Rho-dependent termination in *E. coli* (23–25,44) essentially confirmed these observations. They also showed that the subset of NusG-dependent terminators (~20% in *E. coli* MG1655) are, on average, less C>G-skewed (25) and less dependent on *Rut*-PBS interactions (44) than the other Rho-dependent terminators but, disappointingly, did not unveil additional termination features of significant predictive value. Although this lack of strong consensus features likely reflects the flexibility necessary to accommodate Rho-dependent terminators within a variety of coding (and regulatory signal) regions (61), it complicates the development of computational models to predict Rho-dependent termination.

To get around this difficulty, we surmised that a quantitative rather than qualitative description of Rho-dependent termination could prove beneficial. We thus defined quantitative descriptors of DNA/RNA composition and examined univariate and multivariate relationships between these descriptors and the *in vitro* transcription termination responses of a training set of DNA templates. With this approach, we identified descriptors of significant predictive value (Figure 4 and Supplementary Table S7) and

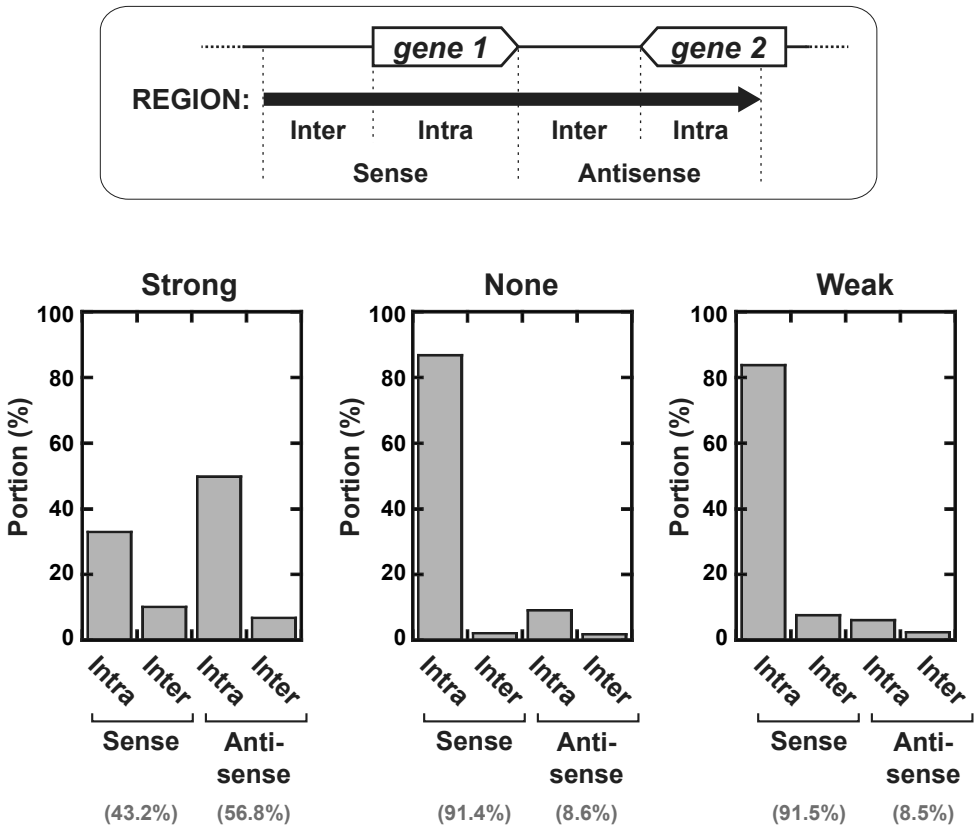


Figure 6. Distribution of the predicted 'Strong', 'None', and 'Weak' regions as a function of gene location in the MG1655 genome. Four distinct categories (Sense-Intragenic, Sense-Intergenic, Antisense-Intragenic and Antisense-Intergenic) were defined with respect to the arrangement of regions and genes (key is inset). Categorization of the intergenic regions as sense/antisense was done with respect to the next downstream gene ('sense' if the region and gene are in same strand orientation, 'antisense' otherwise). Contributions to these categories were calculated for each predicted region and then summed up for all regions of the same class, as shown in the diagrams.

were able to build a multivariate prediction model of Rho-dependent termination using OPLS-DA (Figure 5A and Table 1, OPLS-DA-111 model), a supervised classification approach widely used in analytical chemistry and 'omics' (54,58). The most relevant descriptors deduced from our analyses gauge traits (e.g. length, area) of the C>G bubbles present in the DNA templates (Supplementary Table S7 and Supplementary Figure S10), thereby lending quantitative support and validation to the C>G bubble concept (13) (Supplementary Figure S1). This concept is important because it stipulates that Rho-dependent termination requires non-template DNA strand (or transcript) regions of uninterrupted (rather than diffuse) C>G skewness. This condition likely reflects the requirement for Rho binding to a largely unstructured and YC-rich *Rut* site (2,3,14). We note, however, that not only descriptors of the longest C>G bubble but also descriptors of the sum of all C>G bubbles dispersed in the non-template DNA strand stand are among the most relevant descriptors (Supplementary Table S7). This suggests that Rho-dependent termination is a probabilistic event that not only depends on the quality but also on the density of potential *Rut* sites within the transcribed DNA region. The nature of other significant descriptors supports this view. For instance, richness in $[(YC)N_{9\rightarrow13}]_{1\rightarrow3}$ motifs (which encode 5′-YC dimers suitably spaced for collective interactions with Rho's PBS (16)), not only in the longest C>G bubble but also in the whole DNA template sequence, yields high ANOVA and OPLS-DA ranking descriptors (Supplementary Table S7). The RNA secondary structure potential and density of G-rich motifs (e.g. %G, %GG, %GGG, %GGC, %CGG, %GCG, %GGA, %TGG, %GGT) encoded by the full-length DNA templates also make significant descriptors (Supplementary Table S7), which are inversely correlated with termination (Supplementary Figure S10). This agrees well with transcription experiments showing that aberrant Rho-dependent termination sites can be activated artificially by reducing the secondary structure of transcripts upon incorporation of inosine instead of guanosine (62,63).

We note that a high frequency of pyrimidines (or YC dimers) may not only favor *Rut* recognition by Rho but may also facilitate subsequent, sequence-sensitive steps such as RNA interaction with the SBS and allosteric closure of the Rho ring required for catalytic activation ((64) and references therein). Moreover, sequences promoting RNAP pausing at the points of Rho-mediated TEC release are usually considered necessary (1–3). The negative correlation observed between the density of G-rich motifs (%G, %GG, %CG or %TG) and termination (Figure 4B and Supplementary Figure S10) argues against a favorable role of elemental pause site motifs (GN_8YG or GGN_8YG consensus) (65,66). In fact, the frequency of GN_8YG (or GGN_8YG)

motif occurrence is also negatively correlated with termination (Supplementary Figure S11). Thus, RNAP pause sites favoring Rho-dependent termination (1–3) likely follow other sequence rules that are not readily apparent from our analysis.

The mechanistic basis of other significant descriptors is more enigmatic. For instance, positively correlated descriptors %CA, %CAC, %AC, %ACA, %AAC, and %CAA (Supplementary Table S7 and Supplementary Figure S10) may reflect the presence of CA-rich auxiliary motifs in some terminators (1). However, these motifs, which are 6 to 7 nt-long (1), were not detected by RSAT oligo analysis (data not shown). Such motifs may not be frequent or conserved enough to allow detection with our training set of DNA templates which, despite its reasonable size (104 distinct templates having an average length of ∼500 bp), represents less than 0.4% of the MG1655 or LT2 sequence. We anticipate that additional relevant descriptors, of possibly higher sequence complexity, will be found with a larger training set. Increasing the number of observations would also lend higher statistical power to OPLS-DA modelling of Rho-dependent termination. In this respect, sampling genomic sequences that are not reliably modeled by the current OPLS-DA-111 model (Figure 5B, inset) could prove particularly effective. Using new observations to build a 'naïve' set of data sufficiently large and representative of each termination class to better evaluate the false positive rate of the model could also prove useful. However, a major hindrance to the significant expansion of the training set or assembly of a naïve dataset (or to the deployment of our method in species harboring potentially divergent Rho-dependent machineries (6,45,48)) rests with the relatively high cost and low-throughput of current transcription termination assays. It will thus be necessary to develop cheaper and higher throughput quantitative termination assays, directly *in vivo* if possible, to better assess the contribution of accessory factors and to achieve more comprehensive predictions of Rho-dependent termination.

Despite the abovementioned shortcomings, our OPLS-DA-111 model constitutes the first and only computational tool to probe Rho-dependent termination at the genome scale. Most notably, the model predicts that the portion of the MG1655 (or LT2) genome that is intrinsically refractory to Rho action is limited (Figure 5B and Supplementary Figure S9, 'None' regions). This suggests that Rho-dependent termination is more constrained by indirect factors such as transcript shielding by ribosomes or formation of antitermination complexes (3,67) than by a lack of termination signals. A wealth of termination signals throughout the genome would best suit Rho acting as a global curator of unwanted events such as transcription-translation uncoupling (24,35) or formation of transcriptional R-loops (68). Notwithstanding, we cannot formally exclude that termination-resistant sequences are underrepresented in OPLS-DA-111 predictions because such sequences could be predominantly located in out-of-model regions (Figure 5B, inset) and thus remain undetected. We also cannot exclude that a fraction of the 'Weak' termination signals predicted by the OPLS-DA-111 model are actually too weak to trigger significant effects *in vivo*. In line with this proposal is the observation that a predicted 'Strong' region status is the best prognostic for the Rho-dependent termination loci (Figure 5D and E) identified in *E. coli* by BST transcriptomics (25). We note, however, that there are 3.6-fold more 'Strong' regions than BSTs in the MG1655 genome and that many 'Strong' regions are long enough (Figure 5B) to contain several Rho-dependent terminators. Moreover, a number of the Rho-dependent terminators characterized in this work (Supplementary Table S1) or in other studies (28,35,69) were not detected by BST transcriptomics (25). Taken together, these observations strongly suggest that the MG1655 genome contains significantly more Rho-dependent terminators than envisioned from previous work. Such terminators may be latent in the growth conditions used for BST transcriptomics (25) or reside in poorly transcribed regions complicating experimental detection. Rho-dependent signals arranged in tandem with upstream (intrinsic or Rho-dependent) terminators are also likely to be quiescent or undetectable, yet may constitute failsafe signals against upstream terminator bypass. Considering the pervasiveness of transcription in *E. coli* and *Salmonella* (25,70,71), we thus propose that additional, unexplored layers of Rho-dependent regulation likely take effect under specific environmental/genetic backgrounds. In support of this proposal is the recent discovery of additional Rho-dependent loci in transcriptomic analyses relying on new identification (28) or genetic perturbation (72) strategies.

Among the tested DNA sequences that yield 'Strong' Rho-dependent termination signals *in vitro* (Supplementary Table S1), several are located in (or comprise) the 5′-untranslated regions (5′UTRs) of genes where they may be involved in specific attenuation mechanisms. This conjecture has already been proven true for a couple of specimens (T031 and T043 templates in Supplementary Table S1) participating in the complex regulation of the *rpoS* and *corA* genes, as revealed by studies published since we started the present work (28,31). Specimens that may also be worth investigating include sequences upstream from other highly regulated genes such as *moaA* (T064), for which there is also in *vivo* evidence of Rho-dependent termination (28), as well as *hcp* (T008), *gapA* (T023), *hmp* (T028), or *acs* (T049) (Supplementary Figure S12). More intriguing is the case of the riboswitch-regulated *btuB* gene, which bears an *in vitro* termination signal in the beginning of its coding region (T047 template; Supplementary Figure S12 and data not shown). Rho-dependent terminators at similar locations contribute to the *in vivo* regulation of other riboswitch-dependent genes but could not be detected in the case of *btuB* (69). We suspect that detection of the *btuB* terminator was hindered by an experimental limitation such as adenosyl-cobalamin contamination of the growth medium (69) or use of a terminator-selective Rho mutant (44). Indeed, Rho-dependent regulation in the well-conserved 5′UTR of *mgtA* (9,73) was also not detected with this system (69). Thus, it may be worth looking deeper into the potential involvement of Rho in *btuB* regulation.

In summary, we have built the largest collection of DNA templates for *in vitro* characterization of Rho-dependent termination and identified new terminators that could play an important role in the regulation of *E. coli, Salmonella*, and related species. The resulting dataset was used to

develop sequence descriptors and a multivariate OPLS-DA model that fits experimental data (including transcriptomics) reasonably well. This work thus establishes the proof of principle and provides a solid groundwork for computational modelling of Rho-dependent termination (the strategy might even be tested with the NusG-dependent fraction of terminators once a relevant dataset is available for model training). The interest of alternative supervised classification approaches, such as decision tree-based methods (random forest, gradient boosted trees, etc.), may also be assessed in the future. At present, our OPLS-DA-111 model may be used directly to predict Rho-dependent termination in other *E. coli* or *Salmonella* strains as well as in species having similar Rho and transcription machineries (computer scripts for descriptors and training set descriptor values necessary for OPLS-DA-111 parametrization are provided in Supplementary Information). Moreover, our method should be easily tuned to *in vivo* data training provided that a sufficiently large and accurate (i.e. not plagued by posttranscriptional processing) set of transcript endpoints can be constituted. Future efforts should also aim at minimizing the fraction of out-of-model predictions, possibly by developing additional descriptors, and at extending predictions to other, potentially divergent bacterial species.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We warmly thank Lionello Bossi for helpful discussions and critical reading of the manuscript, Yannick Berteaux for help with computer software/hardware installation and maintenance, Norbert Garnier for access to the CBM computer cluster for computationally intensive calculations, and Marylène Bertrand for help with statistical analyses.

FUNDING

French Agence Nationale de la Recherche [ANR-15-CE11-0024-02 to M.B., ANR-3-BSV3-0005 to N.F.B.] (in part); PhD fellowship from Région Centre-Val de Loire (to C.N.). Funding for open access charge: Agence Nationale de la Recherche.
Conflict of interest statement. None declared.

REFERENCES

1. Ciampi,M.S. (2006) Rho-dependent terminators and transcription termination. *Microbiology*, **152**, 2515–2528.
2. Boudvillain,M., Figueroa-Bossi,N. and Bossi,L. (2013) Terminator still moving forward: expanding roles for Rho factor. *Curr. Opin. Microbiol.*, **16**, 118–124.
3. Ray-Soni,A., Bellecourt,M.J. and Landick,R. (2016) Mechanisms of bacterial transcription Termination: All good things must end. *Annu. Rev. Biochem.*, **85**, 319–347.
4. Nudler,E. (2012) RNA polymerase backtracking in gene regulation and genome instability. *Cell*, **149**, 1438–1445.
5. Washburn,R.S. and Gottesman,M.E. (2015) Regulation of transcription elongation and termination. *Biomolecules*, **5**, 1063–1078.
6. Grylak-Mielnicka,A., Bidnenko,V., Bardowski,J. and Bidnenko,E. (2016) Transcription termination factor Rho: a hub linking diverse physiological processes in bacteria. *Microbiology*, **162**, 433–447.
7. Richardson,L.V. and Richardson,J.P. (1996) Rho-dependent termination of transcription is governed primarily by the upstream Rho utilization (rut) sequences of a terminator. *J. Biol. Chem.*, **271**, 21597–21603.
8. Guerin,M., Robichon,N., Geiselmann,J. and Rahmouni,A.R. (1998) A simple polypyrimidine repeat acts as an artificial Rho-dependent terminator in vivo and in vitro. *Nucleic Acids Res.*, **26**, 4895–4900.
9. Hollands,K., Proshkin,S., Sklyarova,S., Epshtein,V., Mironov,A., Nudler,E. and Groisman,E.A. (2012) Riboswitch control of Rho-dependent transcription termination. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 5376–5381.
10. Figueroa-Bossi,N., Schwartz,A., Guillemardet,B., D'Heygere,F., Bossi,L. and Boudvillain,M. (2014) RNA remodeling by bacterial global regulator CsrA promotes Rho-dependent transcription termination. *Genes Dev.*, **28**, 1239–1251.
11. Kriner,M.A., Sevostyanova,A. and Groisman,E.A. (2016) Learning from the Leaders: Gene regulation by the transcription termination factor rho. *Trends Biochem. Sci.*, **41**, 690–699.
12. Rabhi,M., Espeli,O., Schwartz,A., Cayrol,B., Rahmouni,A.R., Arluison,V. and Boudvillain,M. (2011) The Sm-like RNA chaperone Hfq mediates transcription antitermination at Rho-dependent terminators. *EMBO J.*, **30**, 2805–2816.
13. Alifano,P., Rivellini,F., Limauro,D., Bruni,C.B. and Carlomagno,M.S. (1991) A consensus motif common to all Rho-dependent prokaryotic transcription terminators. *Cell*, **64**, 553–563.
14. Rabhi,M., Rahmouni,A.R. and Boudvillain,M. (2010) In: Jankowsky,E. (ed). *RNA Helicases*. RSC Publishing, Cambridge, Vol. **19**, pp. 243–271.
15. Bogden,C.E., Fass,D., Bergman,N., Nichols,M.D. and Berger,J.M. (1999) The structural basis for terminator recognition by the Rho transcription termination factor. *Mol Cell*, **3**, 487–493.
16. Skordalakes,E. and Berger,J.M. (2003) Structure of the Rho transcription terminator: mechanism of mRNA recognition and helicase loading. *Cell*, **114**, 135–146.
17. McSwiggen,J.A., Bear,D.G. and von Hippel,P.H. (1988) Interactions of Escherichia coli transcription termination factor rho with RNA. I. Binding stoichiometries and free energies. *J Mol Biol*, **199**, 609–622.
18. Geiselmann,J., Yager,T.D. and von Hippel,P.H. (1992) Functional interactions of ligand cofactors with Escherichia coli transcription termination factor rho. II. Binding of RNA. *Protein Sci.*, **1**, 861–873.
19. Koslover,D.J., Fazal,F.M., Mooney,R.A., Landick,R. and Block,S.M. (2012) Binding and translocation of termination factor rho studied at the single-molecule level. *J Mol Biol*, **423**, 664–676.
20. Hitchens,T.K., Zhan,Y., Richardson,L.V., Richardson,J.P. and Rule,G.S. (2006) Sequence-specific interactions in the RNA-binding domain of escherichia coli transcription termination factor rho. *J. Biol. Chem.*, **281**, 33697–33703.
21. Vieu,E. and Rahmouni,A.R. (2004) Dual role of boxB RNA motif in the mechanisms of termination/antitermination at the lambda tR1 terminator revealed in vivo. *J. Mol. Biol.*, **339**, 1077–1087.
22. Schwartz,A., Walmacq,C., Rahmouni,A.R. and Boudvillain,M. (2007) Noncanonical interactions in the management of RNA structural blocks by the transcription termination rho helicase. *Biochemistry*, **46**, 9366–9379.
23. Cardinale,C.J., Washburn,R.S., Tadigotla,V.R., Brown,L.M., Gottesman,M.E. and Nudler,E. (2008) Termination factor rho and its cofactors NusA and NusG silence foreign DNA in E. coli. *Science*, **320**, 935–938.
24. Peters,J.M., Mooney,R.A., Kuan,P.F., Rowland,J.L., Keles,S. and Landick,R. (2009) Rho directs widespread termination of intragenic and stable RNA transcription. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 15406–15411.
25. Peters,J.M., Mooney,R.A., Grass,J.A., Jessen,E.D., Tran,F. and Landick,R. (2012) Rho and NusG suppress pervasive antisense transcription in Escherichia coli. *Genes Dev.*, **26**, 2621–2633.
26. Nicolas,P., Mader,U., Dervyn,E., Rochat,T., Leduc,A., Pigeonneau,N., Bidnenko,E., Marchadier,E., Hoebeke,M., Aymerich,S. *et al.* (2012) Condition-dependent transcriptome reveals high-level regulatory architecture in Bacillus subtilis. *Science*, **335**, 1103–1106.
27. Mader,U., Nicolas,P., Depke,M., Pane-Farre,J., Debarbouille,M., van der Kooi-Pol,M.M., Guerin,C., Derozier,S., Hiron,A., Jarmer,H. *et al.* (2016) Staphylococcus aureus transcriptome architecture: from

laboratory to infection-mimicking conditions. *PLoS Genet.*, **12**, e1005962.
28. Sedlyarova,N., Shamovsky,I., Bharati,B.K., Epshtein,V., Chen,J., Gottesman,S., Schroeder,R. and Nudler,E. (2016) sRNA-Mediated control of transcription termination in E. coli. *Cell*, **167**, 111–121.
29. Bossi,L., Schwartz,A., Guillemardet,B., Boudvillain,M. and Figueroa-Bossi,N. (2012) A role for Rho-dependent polarity in gene regulation by a noncoding small RNA. *Genes Dev.*, **26**, 1864–1873.
30. Brandis,G., Bergman,J.M. and Hughes,D. (2016) Autoregulation of the tufB operon in Salmonella. *Mol .Microbiol.*, **100**, 1004–1016.
31. Kriner,M.A. and Groisman,E.A. (2015) The bacterial transcription termination factor Rho coordinates Mg homeostasis with translational signals. *J. Mol. Biol.*, **427**, 3834–3849.
32. Sevostyanova,A. and Groisman,E.A. (2015) An RNA motif advances transcription by preventing Rho-dependent termination. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, E6835–E6843.
33. Takemoto,N., Tanaka,Y. and Inui,M. (2015) Rho and RNase play a central role in FMN riboswitch regulation in Corynebacterium glutamicum. *Nucleic Acids Res.*, **43**, 520–529.
34. Wang,X., Ji,S.C., Jeon,H.J., Lee,Y. and Lim,H.M. (2015) Two-level inhibition of galK expression by Spot 42: Degradation of mRNA mK2 and enhanced transcription termination before the galK gene. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, 7581–7586.
35. Sedlyarova,N., Rescheneder,P., Magan,A., Popitsch,N., Rziha,N., Bilusic,I., Epshtein,V., Zimmermann,B., Lybecker,M., Sedlyarov,V. *et al.* (2017) Natural RNA polymerase aptamers regulate transcription in E. coli. *Mol. Cell*, **67**, 30–43.
36. Touchon,M., Hoede,C., Tenaillon,O., Barbe,V., Baeriswyl,S., Bidet,P., Bingen,E., Bonacorsi,S., Bouchier,C., Bouvet,O. *et al.* (2009) Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths. *PLoS Genet.*, **5**, e1000344.
37. Menouni,R., Champ,S., Espinosa,L., Boudvillain,M. and Ansaldi,M. (2013) Transcription termination controls prophage maintenance in Escherichia coli genomes. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 14414–14419.
38. Lesnik,E.A., Sampath,R., Levene,H.B., Henderson,T.J., McNeil,J.A. and Ecker,D.J. (2001) Prediction of rho-independent transcriptional terminators in Escherichia coli. *Nucleic Acids Res.*, **29**, 3583–3594.
39. de Hoon,M.J., Makita,Y., Nakai,K. and Miyano,S. (2005) Prediction of transcriptional terminators in bacillus subtilis and related species. *PLoS Comput. Biol.*, **1**, e25.
40. Kingsford,C.L., Ayanbule,K. and Salzberg,S.L. (2007) Rapid, accurate, computational discovery of Rho-independent transcription terminators illuminates their relationship to DNA uptake. *Genome Biol.*, **8**, R22.
41. Gardner,P.P., Barquist,L., Bateman,A., Nawrocki,E.P. and Weinberg,Z. (2011) RNIE: genome-wide prediction of bacterial intrinsic terminators. *Nucleic Acids Res.*, **39**, 5845–5852.
42. Naville,M., Ghuillot-Gaudeffroy,A., Marchais,A. and Gautheret,D. (2011) ARNold: a web tool for the prediction of Rho-independent transcription terminators. *RNA Biol.*, **8**, 11–13.
43. Unniraman,S., Prakash,R. and Nagaraja,V. (2002) Conserved economics of transcription termination in eubacteria. *Nucleic Acids Res.*, **30**, 675–684.
44. Shashni,R., Qayyum,M.Z., Vishalini,V., Dey,D. and Sen,R. (2014) Redundancy of primary RNA-binding functions of the bacterial transcription terminator Rho. *Nucleic Acids Res.*, **42**, 9677–9690.
45. D'Heygere,F., Rabhi,M. and Boudvillain,M. (2013) Phyletic distribution and conservation of the bacterial transcription termination factor Rho. *Microbiology*, **159**, 1423–1436.
46. Nowatzke,W.L., Burns,C.M. and Richardson,J.P. (1997) Function of the novel subdomain in the RNA binding domain of transcription termination factor Rho from Micrococcus luteus. *J. Biol. Chem.*, **272**, 2207–2211.
47. Mitra,A., Misquitta,R. and Nagaraja,V. (2014) Mycobacterium tuberculosis Rho Is an NTPase with distinct kinetic properties and a novel RNA-Binding subdomain. *PLoS One*, **9**, e107474.
48. D'Heygere,F., Schwartz,A., Coste,F., Castaing,B. and Boudvillain,M. (2015) ATP-dependent motor activity of the transcription termination factor Rho from Mycobacterium tuberculosis. *Nucleic Acids Res.*, **43**, 6099–6111.
49. Boudvillain,M., Walmacq,C., Schwartz,A. and Jacquinot,F. (2010) Simple enzymatic assays for the in vitro motor activity of transcription termination factor Rho from Escherichia coli. *Methods Mol. Biol.*, **587**, 137–154.
50. Rabhi,M., Gocheva,V., Jacquinot,F., Lee,A., Margeat,E. and Boudvillain,M. (2011) Mutagenesis-based evidence for an asymmetric configuration of the ring-shaped transcription termination factor Rho. *J. Mol. Biol.*, **405**, 497–518.
51. Artsimovitch,I. and Henkin,T.M. (2009) In vitro approaches to analysis of transcription termination. *Methods*, **47**, 37–43.
52. Zuker,M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, **31**, 3406–3415.
53. Lennart,E., Johan,T. and Svante,W. (2008) CV-ANOVA for significance testing of PLS and OPLS® models. *J. Chemometrics*, **22**, 594–600.
54. Triba,M.N., Le Moyec,L., Amathieu,R., Goossens,C., Bouchemal,N., Nahon,P., Rutledge,D.N. and Savarin,P. (2015) PLS/OPLS models in metabolomics: the impact of permutation of dataset rows on the K-fold cross-validation quality parameters. *Mol. bioSyst.*, **11**, 13–19.
55. Medina-Rivera,A., Defrance,M., Sand,O., Herrmann,C., Castro-Mondragon,J.A., Delerce,J., Jaeger,S., Blanchet,C., Vincens,P., Caron,C. *et al.* (2015) RSAT 2015: Regulatory sequence analysis tools. *Nucleic Acids Res.*, **43**, W50–W56.
56. Skordalakes,E. and Berger,J.M. (2006) Structural Insights into RNA-Dependent ring closure and ATPase activation by the rho termination factor. *Cell*, **127**, 553–564.
57. Abdi,H. and Williams,L.J. (2010) Principal component analysis. *Wiley Interdiscipl. Rev.: Comput. Stat.*, **2**, 433–459.
58. Bylesjö,M., Rantalainen,M., Cloarec,O., Nicholson,J.K., Holmes,E. and Trygg,J. (2006) OPLS discriminant analysis: combining the strengths of PLS-DA and SIMCA classification. *J. Chemometrics*, **20**, 341–351.
59. Gama-Castro,S., Salgado,H., Santos-Zavaleta,A., Ledezma-Tejeida,D., Muniz-Rascado,L., Garcia-Sotelo,J.S., Alquicira-Hernandez,K., Martinez-Flores,I., Pannier,L., Castro-Mondragon,J.A. *et al.* (2016) RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. *Nucleic Acids Res.*, **44**, D133–D143.
60. Kudla,G., Murray,A.W., Tollervey,D. and Plotkin,J.B. (2009) Coding-sequence determinants of gene expression in Escherichia coli. *Science*, **324**, 255–258.
61. Richardson,J.P. (1990) Rho-dependent transcription termination. *Biochim. Biophys. Acta*, **1048**, 127–138.
62. Morgan,W.D., Bear,D.G. and von Hippel,P.H. (1983) Rho-dependent termination of transcription. I. Identification and characterization of termination sites for transcription from the bacteriophage lambda PR promoter. *J. Biol. Chem.*, **258**, 9553–9564.
63. Zhu,A.Q. and von Hippel,P.H. (1998) Rho-dependent termination within the trp t' terminator. I. Effects of rho loading and template sequence. *Biochemistry*, **37**, 11202–11214.
64. Lawson,M.R., Dyer,K. and Berger,J.M. (2016) Ligand-induced and small-molecule control of substrate loading in a hexameric helicase. *Proc. Natl. Acad. Sci. U.S.A.*, **113**, 13714–13719.
65. Vvedenskaya,I.O., Vahedian-Movahed,H., Bird,J.G., Knoblauch,J.G., Goldman,S.R., Zhang,Y., Ebright,R.H. and Nickels,B.E. (2014) Interactions between RNA polymerase and the "core recognition element" counteract pausing. *Science*, **344**, 1285–1289.
66. Larson,M.H., Mooney,R.A., Peters,J.M., Windgassen,T., Nayak,D., Gross,C.A., Block,S.M., Greenleaf,W.J., Landick,R., Weissman,J.S. *et al.* (2014) A pause sequence enriched at translation start sites drives transcription dynamics in vivo. *Science*, **344**, 1042–1047.
67. Santangelo,T.J. and Artsimovitch,I. (2011) Termination and antitermination: RNA polymerase runs a stop sign. *Nat. Rev. Microbiol.*, **9**, 319–329.
68. Leela,J.K., Syeda,A.H., Anupama,K. and Gowrishankar,J. (2013) Rho-dependent transcription termination is essential to prevent excessive genome-wide R-loops in Escherichia coli. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 258–263.
69. Bastet,L., Chauvier,A., Singh,N., Lussier,A., Lamontagne,A.M., Prevost,K., Masse,E., Wade,J.T. and Lafontaine,D.A. (2017) Translational control and Rho-dependent transcription termination are intimately linked in riboswitch regulation. *Nucleic Acids Res.*, **45**, 7474–7486.

70. Raghavan,R., Sloan,D.B. and Ochman,H. (2012) Antisense transcription is pervasive but rarely conserved in enteric bacteria. *mBio*, **3**, e00156-12.
71. Thomason,M.K., Bischler,T., Eisenbart,S.K., Forstner,K.U., Zhang,A., Herbig,A., Nieselt,K., Sharma,C.M. and Storz,G. (2015) Global transcriptional start site mapping using differential RNA sequencing reveals novel antisense RNAs in Escherichia coli. *J. Bacteriol.*, **197**, 18–28.
72. Raghunathan,N., Kapshikar,R.M., Leela,J.K., Mallikarjun,J., Bouloc,P. and Gowrishankar,J. (2018) Genome-wide relationship between R-loop formation and antisense transcription in Escherichia coli. *Nucleic Acids Res.*, **46**, 3400–3411.
73. Hollands,K., Sevostiyanova,A. and Groisman,E.A. (2014) Unusually long-lived pause required for regulation of a Rho-dependent transcription terminator. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, E1999–E2007.
74. Epshtein,V., Dutta,D., Wade,J. and Nudler,E. (2010) An allosteric mechanism of Rho-dependent transcription termination. *Nature*, **463**, 245–249.
75. Thomsen,N.D. and Berger,J.M. (2009) Running in reverse: the structural basis for translocation polarity in hexameric helicases. *Cell*, **139**, 523–534.
76. Gocheva,V., Le Gall,A., Boudvillain,M., Margeat,E. and Nollmann,M. (2015) Direct observation of the translocation mechanism of transcription termination factor Rho. *Nucleic Acids Res.*, **43**, 2367–2377.

A multivariate prediction model for Rho-dependent termination of transcription

Cédric Nadiras, Eric Eveno, Annie Schwartz, Nara Figueroa-Bossi, and Marc Boudvillain

(Supplementary material)

This document includes Supplementary methods, a brief discussion of OPLS-DA-11 orthogonal variance, Figures S1 to S12 and Tables S1 to S7. The following information is provided as separate files:

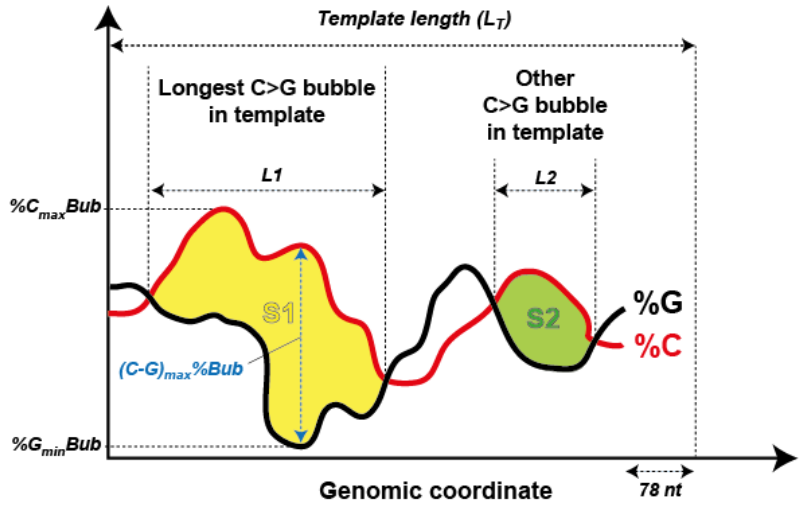
- Python scripts for descriptors and predicted regions (with ReadMe pdf file)
- Predicted termination regions within MG1655 & LT2 genomes (Table S8)
- Descriptor values for the full training set of 104 DNA templates (Table S9)

Supplementary methods

C>G bubbles descriptors

A Python script for detection of C>G bubbles and calculation of sequence descriptors is provided in supplementary information as a separate file (MakeDescriptors.py).

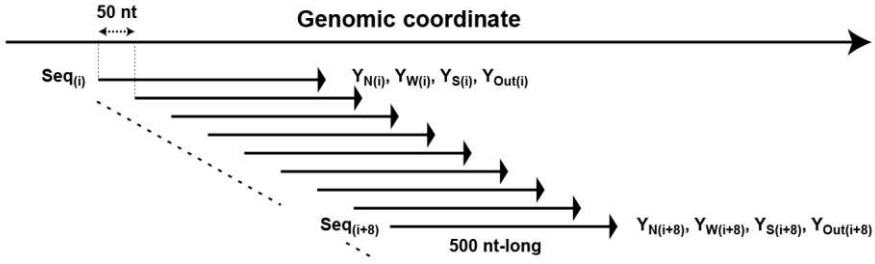
DNA template sequences (from training set) or 500 bp-long genomic sequences (upon analysis of genomes; see below) were individually scanned in search of C>G bubbles using a 78-nt sliding window (1). The C>G bubble descriptors were defined for the longest C>G bubble found in the analyzed sequence or for the sum of all C>G bubbles found, as depicted below. Names and definitions of all sequence descriptors are provided in supplementary Table 6.



For each C>G bubble, the difference %C-%G was determined at each nucleotide position. The bubble area (S_i in diagram) was determined by summing up %C-%G values over the length of the bubble (area are thus expressed in % x bp).

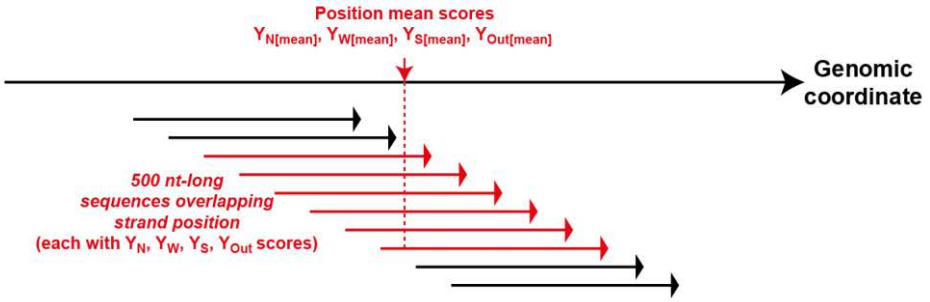
Definition of None', 'Weak', 'Strong', and 'out-of-model' regions for genomic predictions.

For MG1655 and LT2 genomic predictions, 500-nt long sequences were defined along each strand of the genome in 50 nt increments:



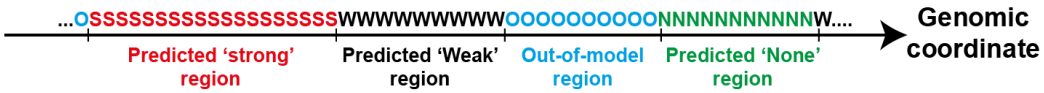
Descriptor values were determined for each 500 nt-long genomic sequence. The sequences were then analyzed with the OPL-DA-111 model in order to obtain per-class predicted response scores (YPredPS vector scores in SIMCA for the predicted dummy Y_{pred} variable responses; respectively named Y_N, Y_W, Y_S for the 'None', 'Weak' and 'Strong' classes in above diagram)

for each of them. We arbitrarily defined a fourth score variable (Y_{Out} in diagram) to account for 'out-of-model' sequences based on the proprietary SIMCA $P_{mod}XPS^{+}$ parameter (probability of model membership). Hence, for sequences with $P_{mod}XPS^{+} < 0.05$ (model outliers at 95% confidence level), the Y_{Out} score was set to 1 while the Y_N, Y_W, and Y_S scores were conservatively changed to zero. For other sequences ($P_{mod}XPS^{+} \geq 0.05$), the Y_{Out} score was set to zero while the Y_N, Y_W, and Y_S scores were kept at their original YPredPS values. Then, for each genomic strand position, mean scores $Y_{N[mean]}$, $Y_{W[mean]}$, $Y_{S[mean]}$, and $Y_{Out[mean]}$ were calculated, respectively, as the averages of the Y_N, Y_W, Y_S and Y_{Out} scores obtained for all of the 500 nt-long, same-strand sequences overlapping this position:



A winner-takes-it-all strategy was then used to assign each strand position to a single category (i.e. the one with the highest mean score). For instance, strand positions for which $Y_{N[mean]} > Y_{W[mean]}$, $Y_{S[mean]}$, or $Y_{Out[mean]}$ were assigned to the 'None' category. The Python script used for $Y_{N[mean]}$, $Y_{W[mean]}$, $Y_{S[mean]}$, and $Y_{Out[mean]}$ calculation and category assignment is provided as a separate file (MakePredPerBase.py)

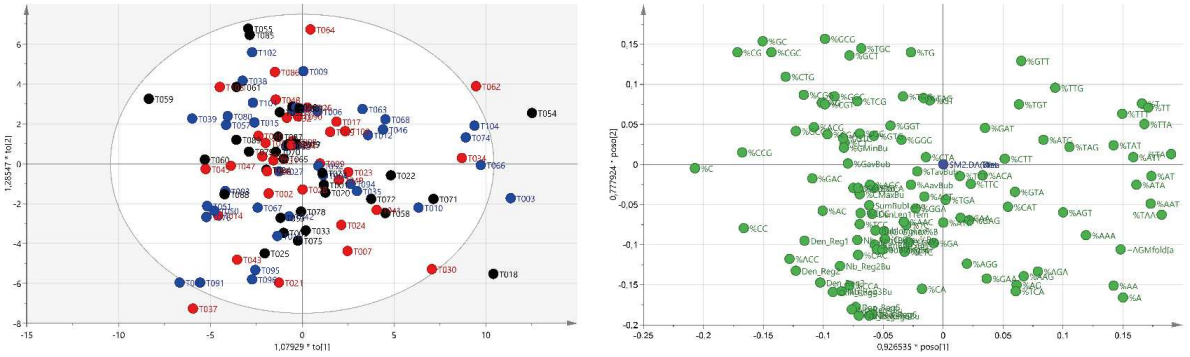
The predicted None', 'Weak', 'Strong', and 'out-of-model' regions were then defined as continuous successions of strand positions falling into the corresponding category:



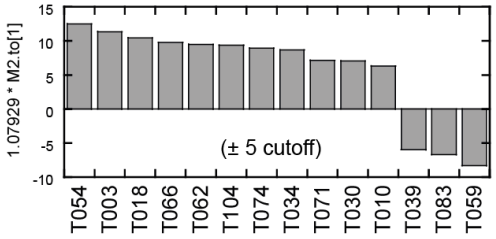
The python script used for this last step is also provided as a separate file (MakePredRegions.py).

Sources of orthogonal variance in the OPLS-DA-111 model

Orthogonal variance likely stems from a combination of experimental noise, variations in observations not adequately explained by the descriptors, and variations in some descriptors uncorrelated to the termination categorization. These sources are not easily untangled from the scattering of observations/variables along the two orthogonal PCs:

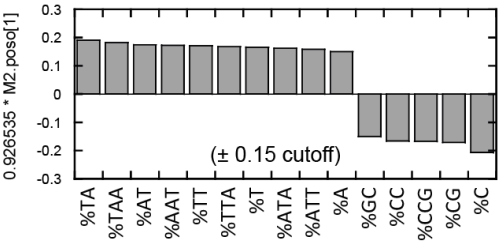


Using coordinates along the orthogonal PC1 highlights contributions from specific DNA templates:



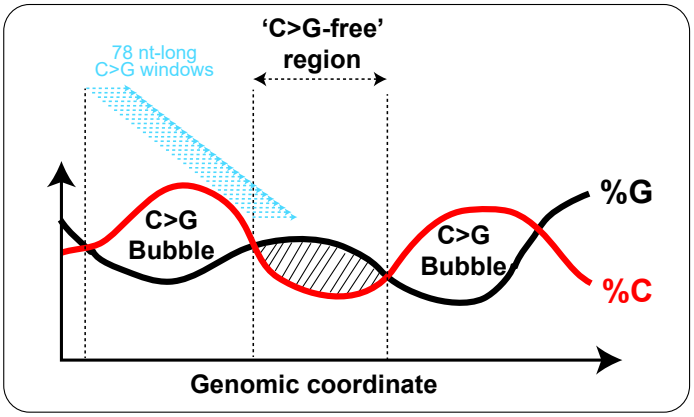
These templates include three of the four C>G-free templates displaying weak termination signals (T003, T074, T104) and one out of the four C>G-plus templates (T059) unable to elicit termination. Other templates in the bar graph belong to the 'None' (3), 'Weak' (5), and 'Strong' (2) categories.

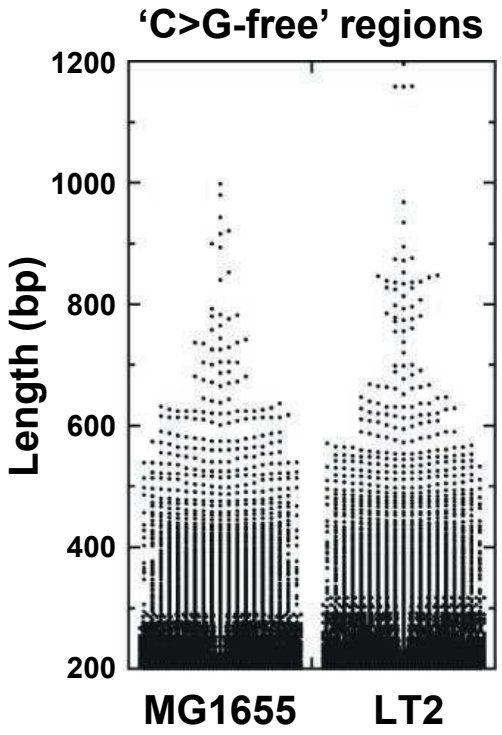
A similar analysis of descriptors along the orthogonal PC1 highlights descriptors of the frequency of A/T- and C-rich motifs within the non-template DNA strand:



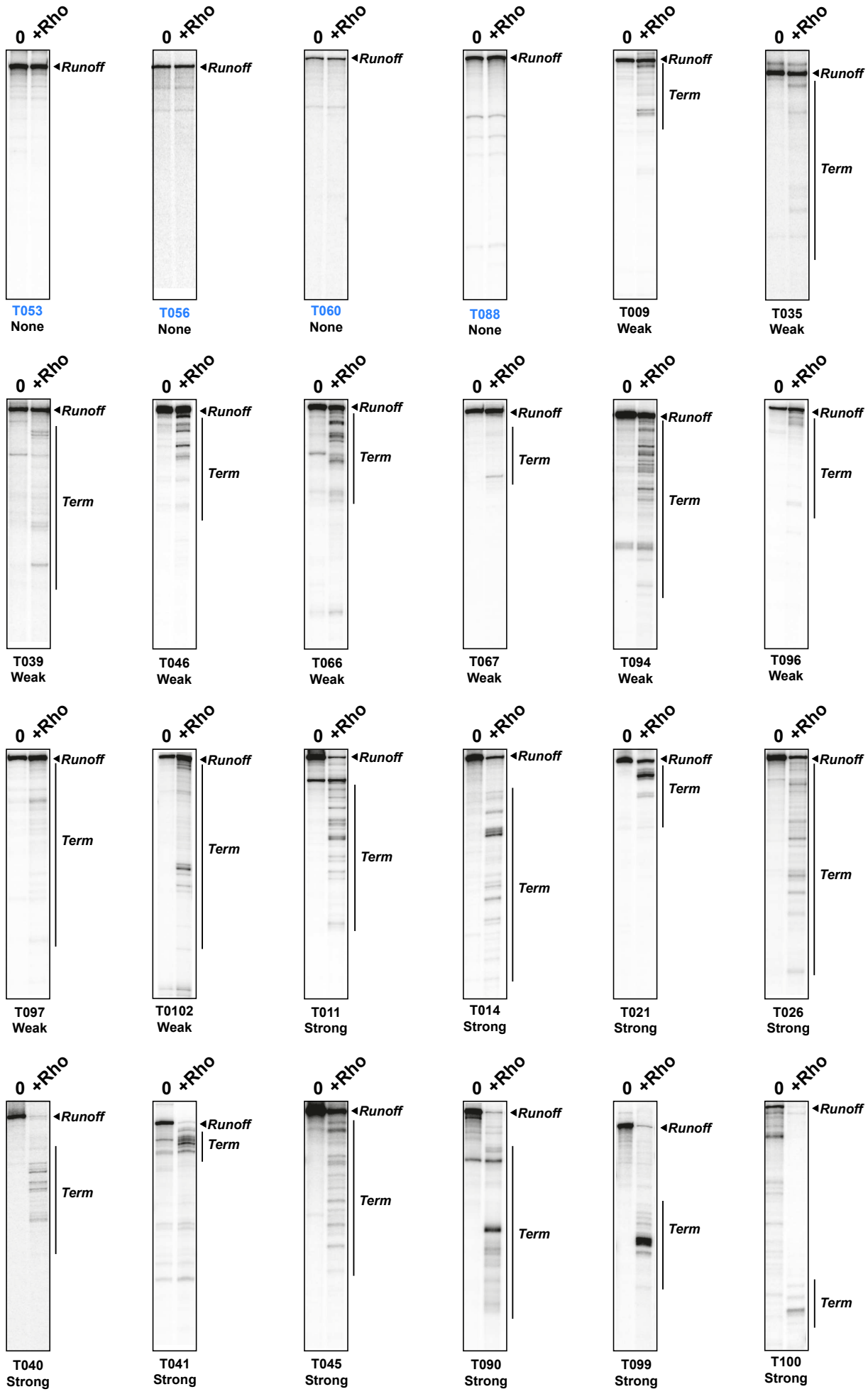
REFERENCES

1. Alifano, P., Rivellini, F., Limauro, D., Bruni, C.B. and Carlomagno, M.S. (1991) A consensus motif common to all Rho-dependent prokaryotic transcription terminators. *Cell*, **64**, 553-563.

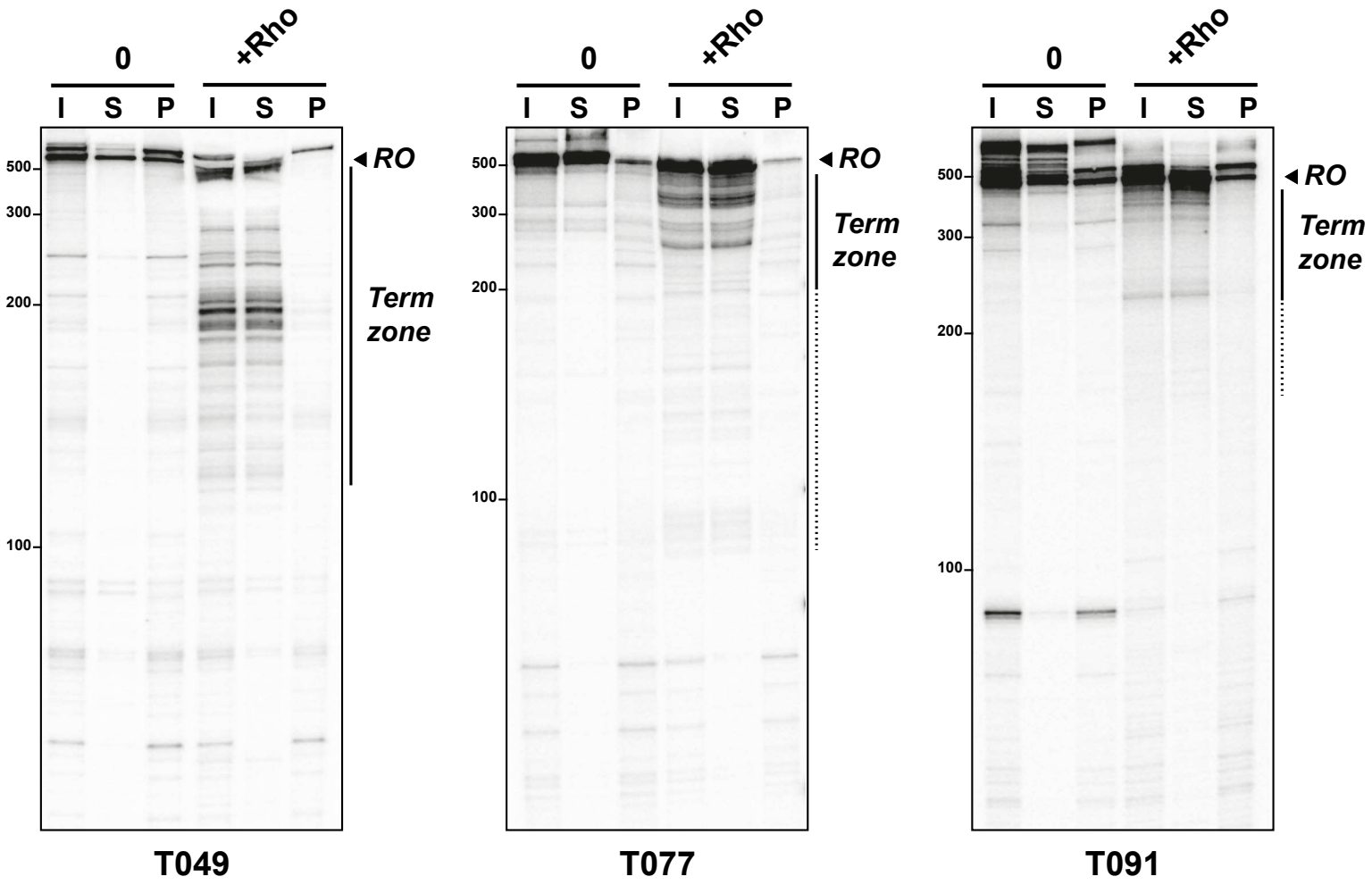




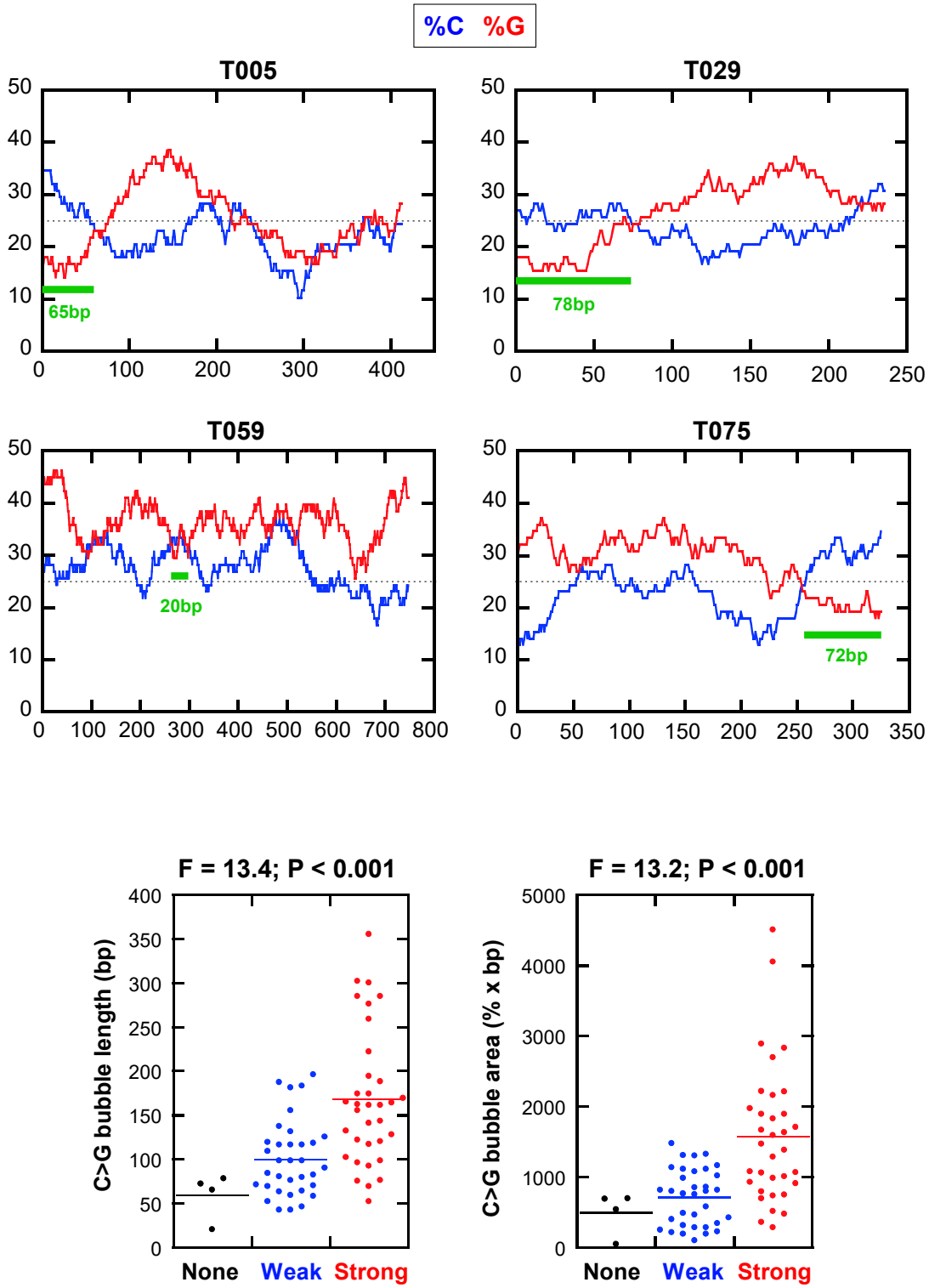
Supplementary Figure 1: Distributions of the lengths of 'C>G-free' regions (cutoff 200 bp) in the genomes of *Escherichia coli* MG1655 and *Salmonella* LT2. As depicted on the top diagram, the 'C>G-free' regions were defined as the regions between consecutive C>G bubbles and thus include the last 78 nt sliding window of the upstream C>G bubble.



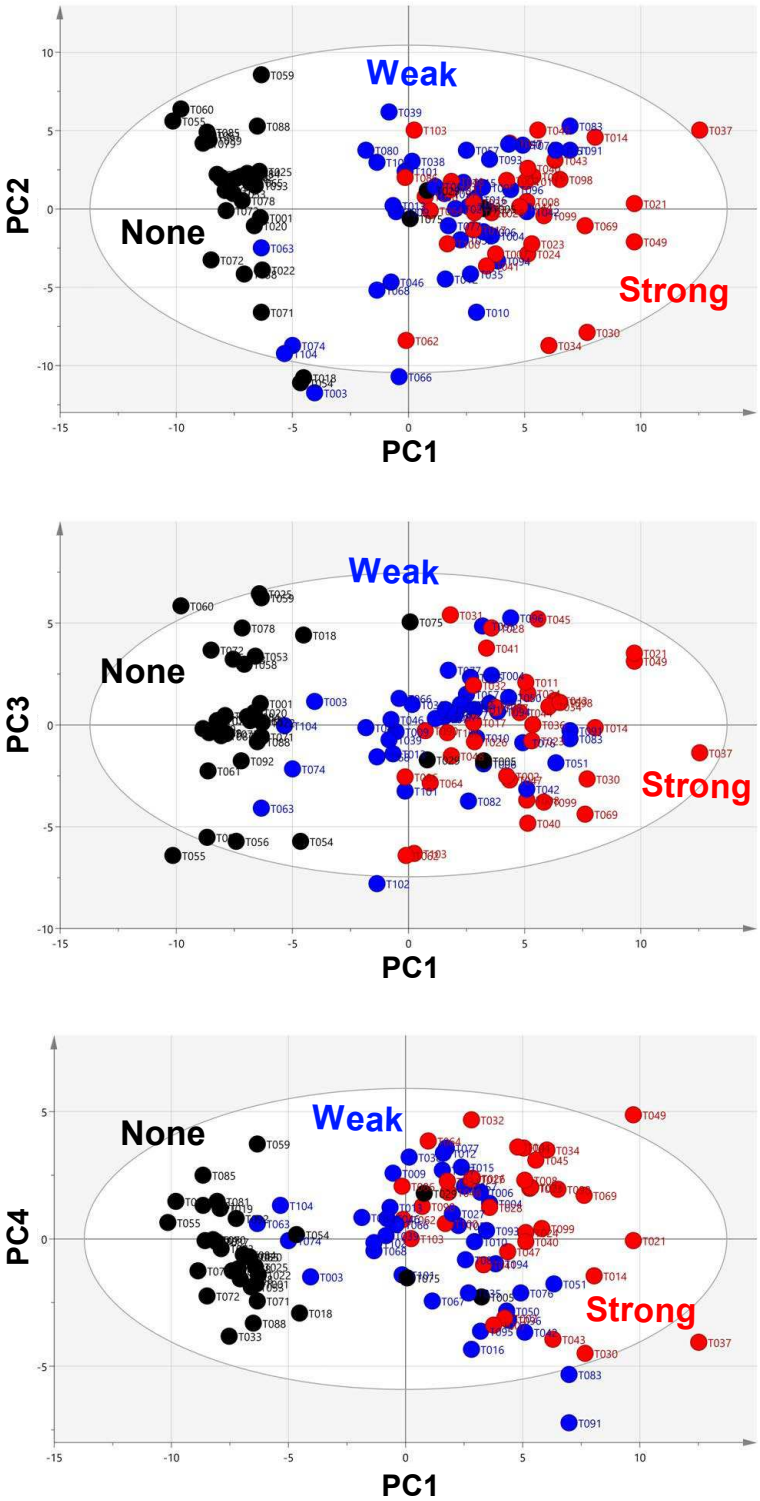
Supplementary Figure 2: Representative denaturing PAGE gels illustrating the three different classes of Rho-dependent transcription termination signals ('None', 'Weak', 'Strong') obtained with the DNA templates listed in Supplementary Table 1. The names of C>G-free templates are in blue.



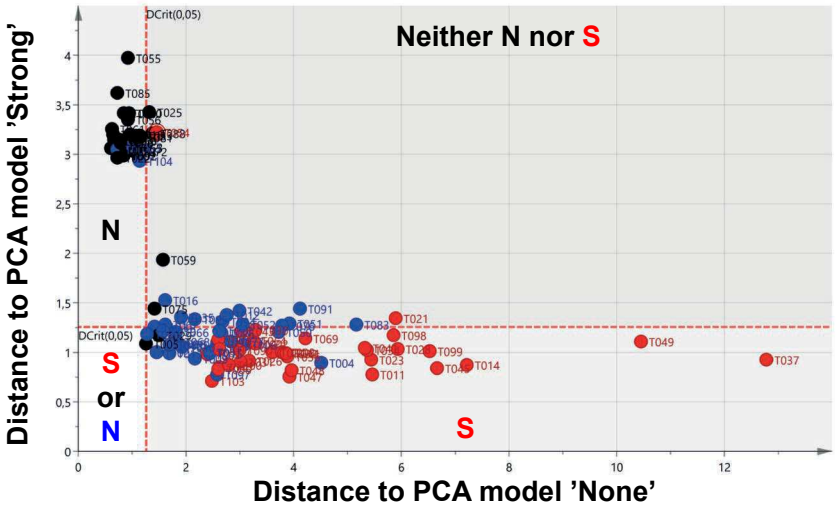
Supplementary Figure 3: Single-round transcriptions with bead-affixed transcription complexes containing biotinylated DNA templates (biotin located at the 5'-end of the non-template strand). The ^{32}P-labeled complexes, halted at position +26 or +27 (depending on the DNA template) by privation of CTP from the initiation mixture, were prepared and immobilized on streptavidin-coated magnetic beads before being 'chased' with 75µM rNTPs in the presence (+Rho lanes) or absence (0 lanes) of Rho. The presence of Rho-dependent transcripts in the supernatants (S lanes) rather than with the bead pellets (P lanes) confirm that they stem from 'true' termination events (main termination zones are indicated on left sides of gels). Reaction imputs were loaded in I lanes.

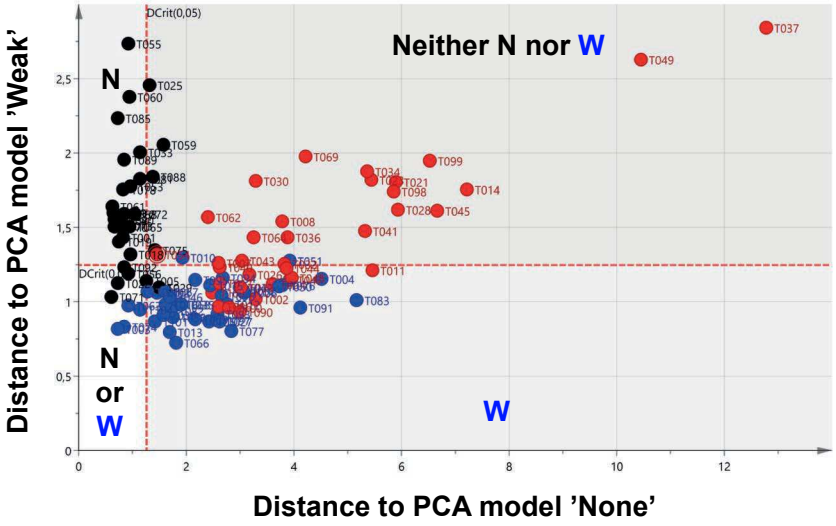


Supplementary Figure 4: Compositional profiles of the four C>G-plus templates that do not elicit Rho-dependent termination (top). Dotplots (bottom) compare the lengths and surfaces of the C>G bubbles found in these DNA templates with the ones of the C>G-plus templates eliciting weak or strong Rho-dependent signals. ANOVA's F-values and p-values are shown above graphs. Note that C>G bubble areas are calculated from the (%C - %G) values summed up over the length of the bubble and are thus expressed in percentage x DNA length (% x bp).

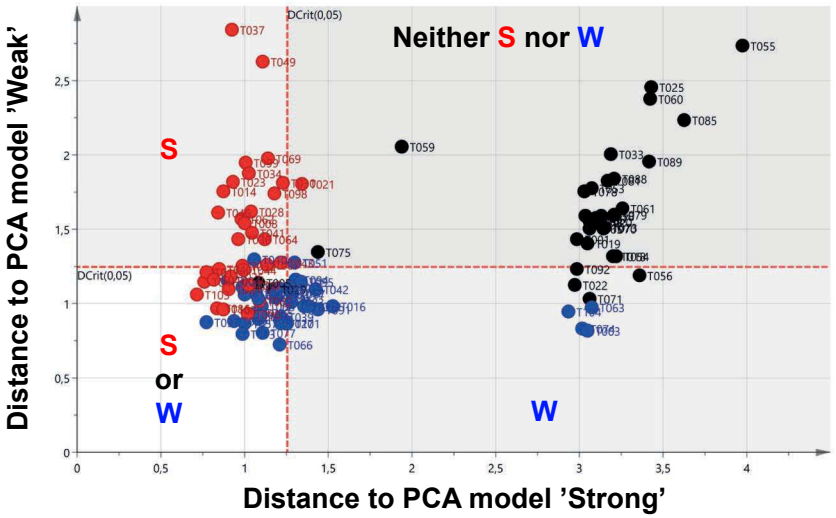


Supplementary Figure 5: Standard biplots for PCA performed with the 104 DNA template transcription termination responses and the 111 sequence descriptors.

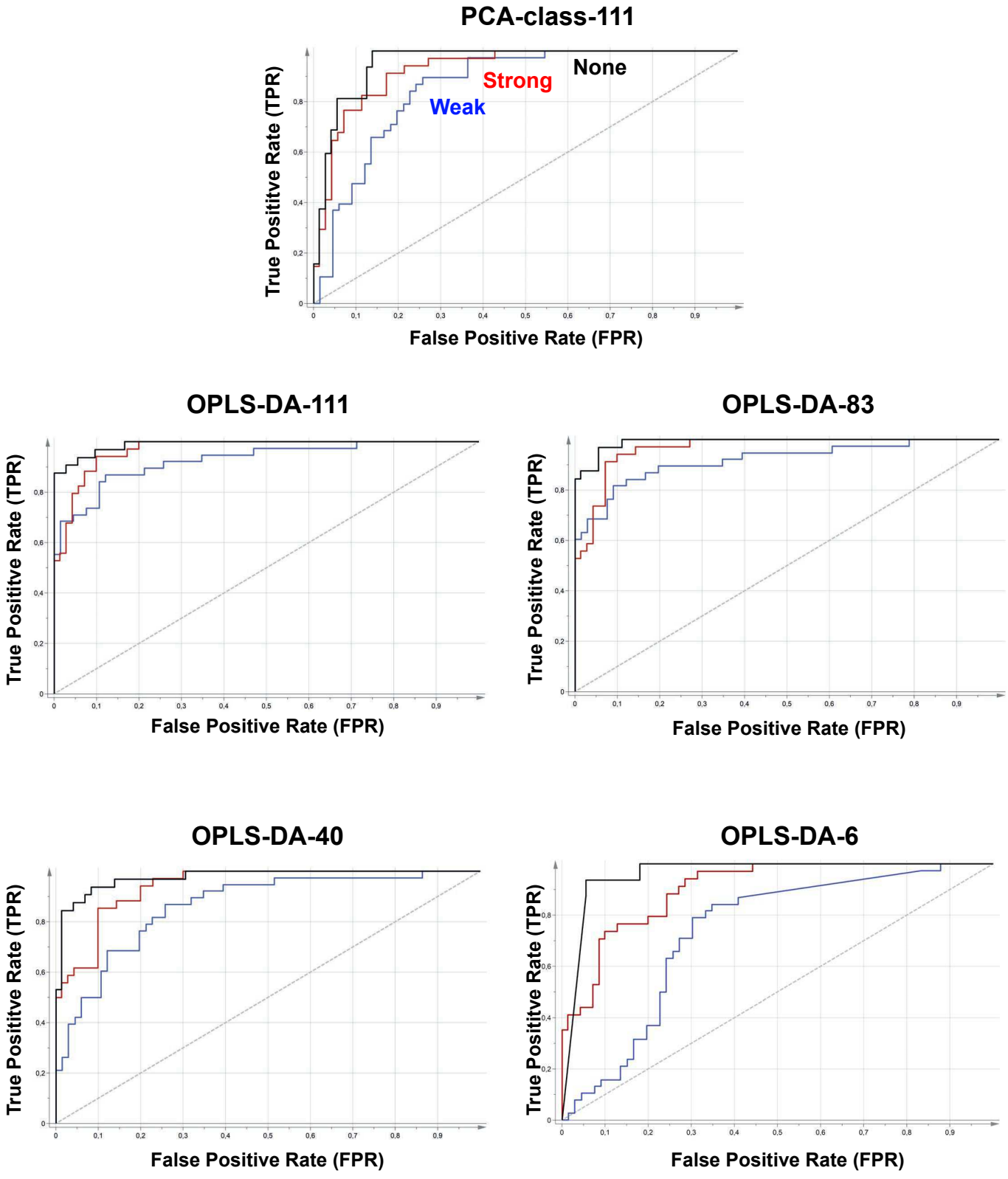




Template class:
- 'None'
- 'Weak'
- 'Strong'

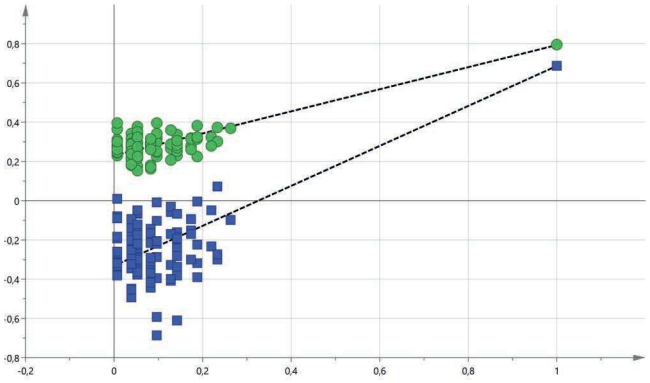


Supplementary Figure 6: Coomans plots of the residual standard deviations for predictions of Rho-dependent termination. Classification is based on disjoint PCA models for the 'None' (N), 'Weak' (W), and 'Strong' (S) classes (PCA-class-111 model in Table 1). Dashed red lines represent the distance limits to models at a significance level of 0.05.



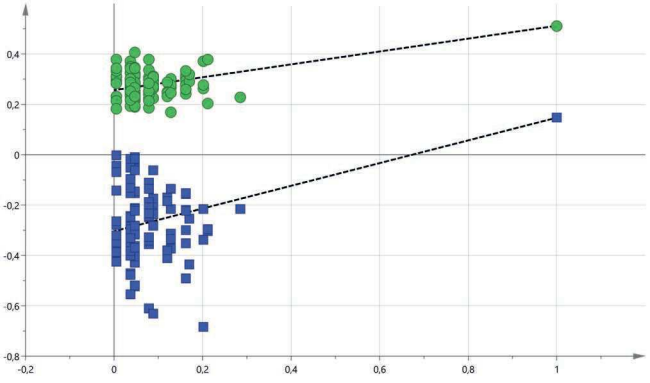
Supplementary Figure 7: Receiver Operating Characteristic (ROC) diagrams for the PCA-class-111 and OPLS-DA models presented in Table 1. As shown on the PCA-class-111 diagram, ROC plots for the 'None', 'Weak' , and 'Strong' classes are shown in black, blue, and red, respectively.

Intercepts: R^2 = (0; 0.230), Q^2 = (0; -0.332)



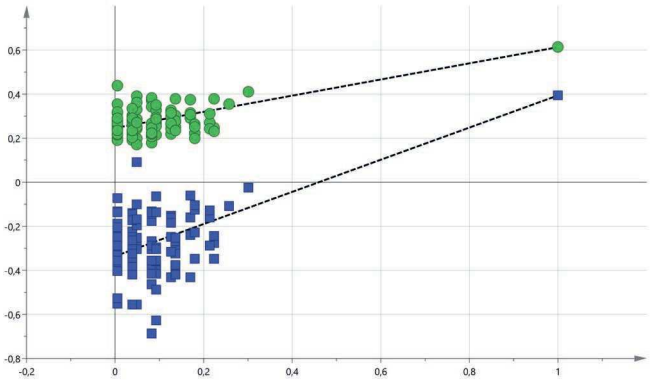
'None'

Intercepts: R^2 = (0; 0.256), Q^2 = (0; -0.304)



'Weak'

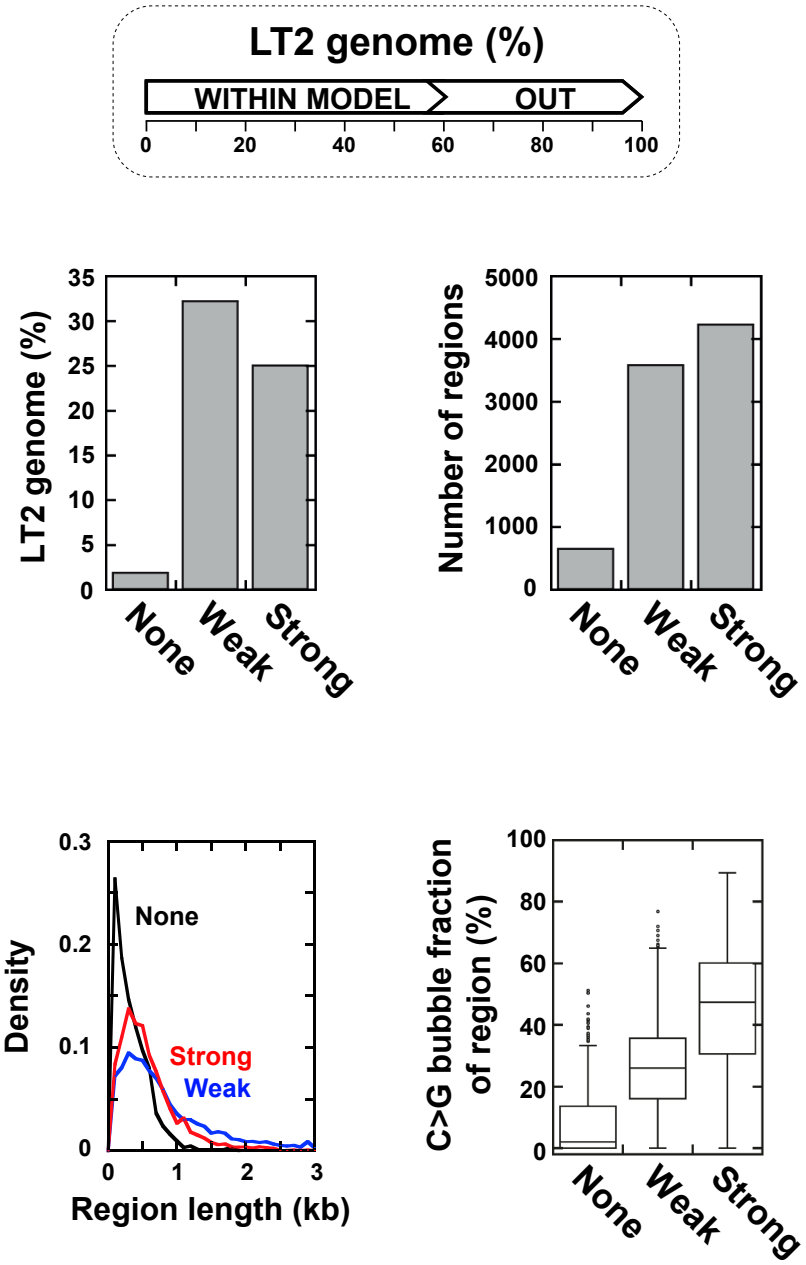
Intercepts: R^2 = (0; 0.246), Q^2 = (0; -0.336)



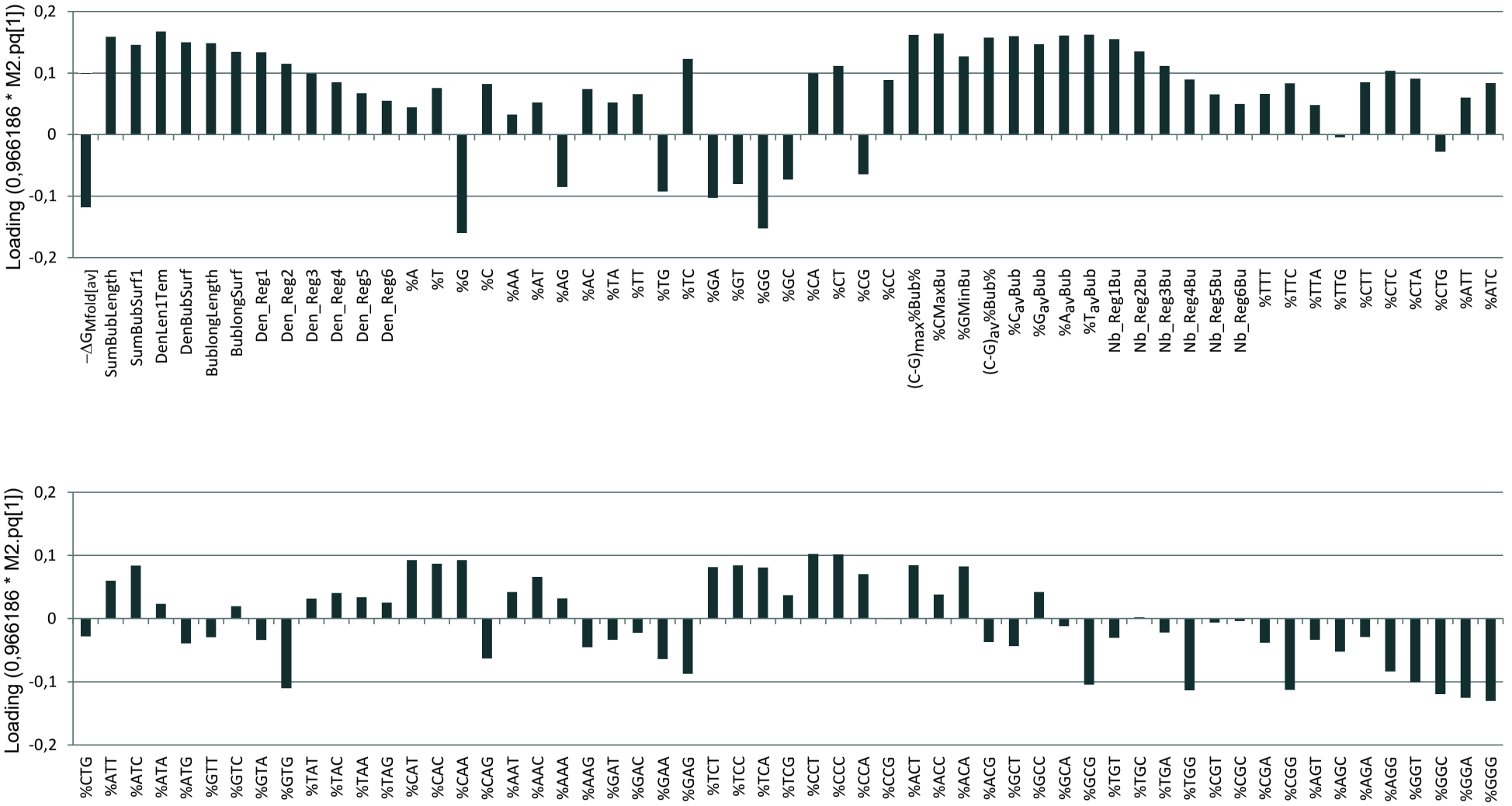
'Strong'

OPLS-DA-111 row permutation test					
	R^2_{Xcum}	R^2_{Ycum}	Q^2_{cum}	F-value	P-value
Mean	0.506	0.628	0.437	5.6	3.7×10^{-8}
SD	0.037	0.024	0.024	1.2	7.8×10^{-8}

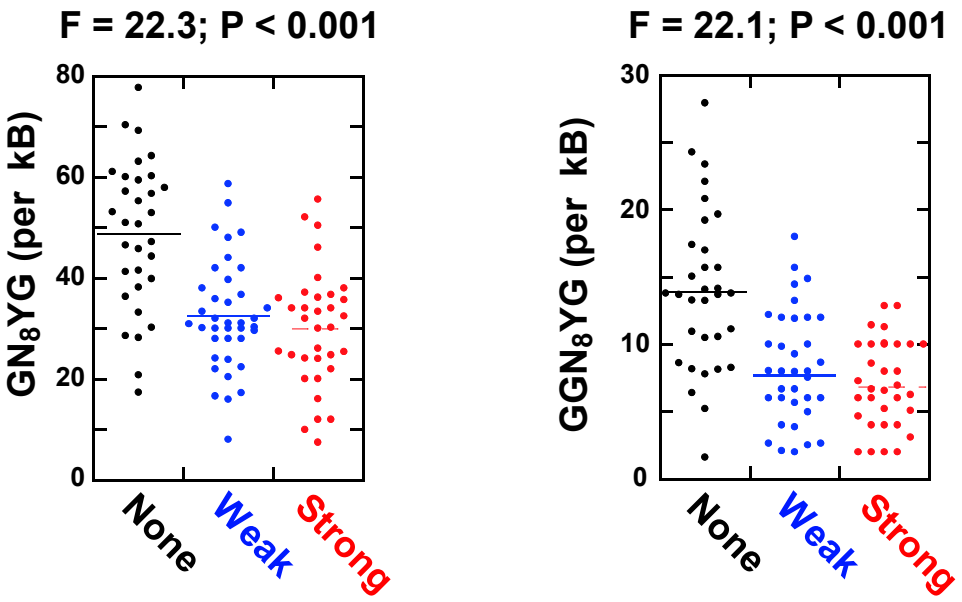
Supplementary Figure 8: SIMCA-P permutation plots (top) for the PLS-DA-111 model (100 permutations, 2 components). R^2 and Q^2 values are in green and blue, respectively. Permuted R^2 and Q^2 values cluster on the left of the plots while original R^2/Q^2 are on the right. The plots support that the performance of the OPLS-DA-111 model is not due to data overfitting since all permuted R^2 and Q^2 values are lower than the original R^2/Q^2 values while Q^2 regression curves have negative intercepts. The table (bottom) summarizes results of the distinct row permutation (n = 7) test recommended byTriba et al., 2015. The lack of strong variations imparted by permutations of the order of observations further supports the significance of the OPLS-DA-111 model.



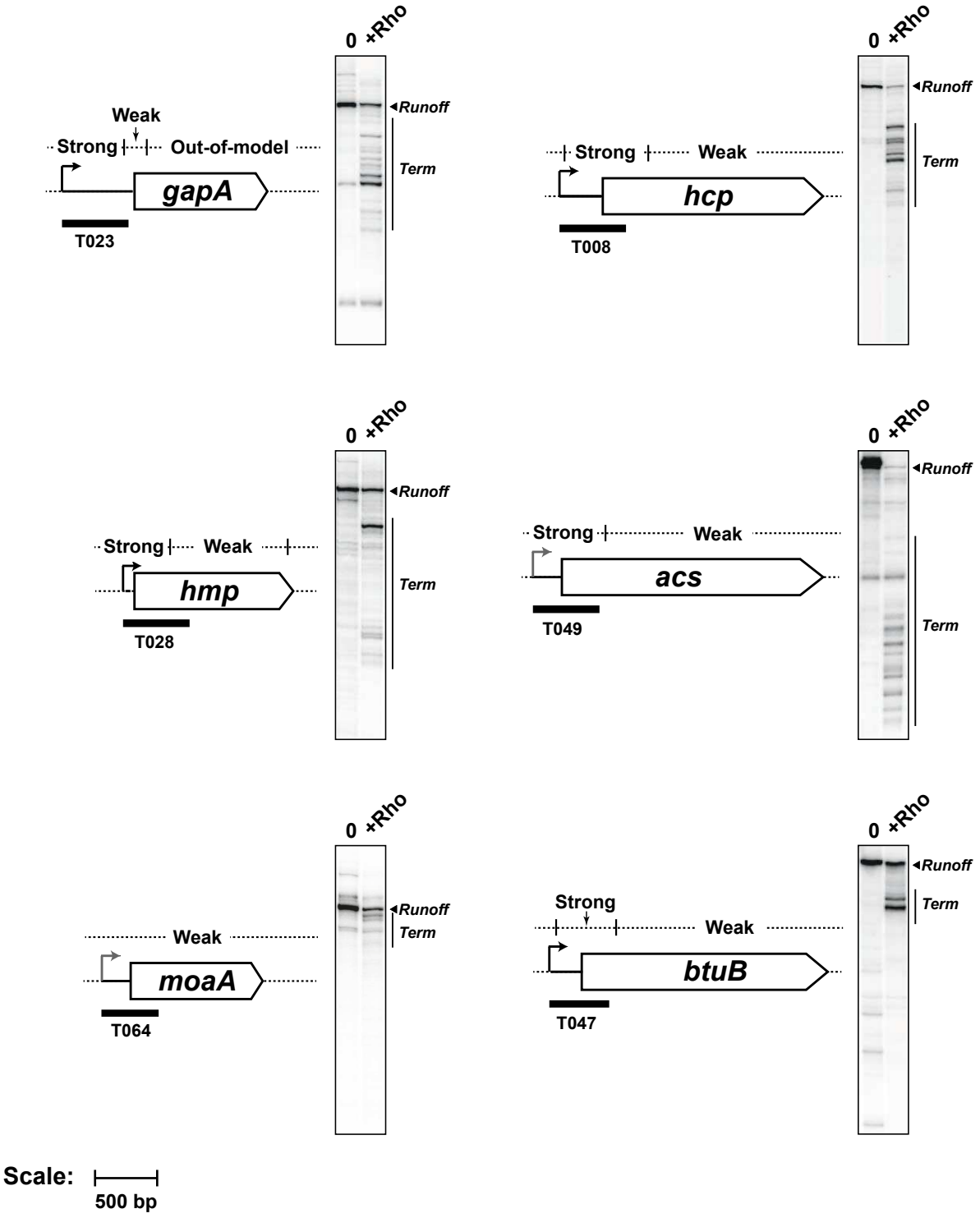
Supplementary Figure 9: Analysis of the *Salmonella* LT2 genome with the OPLS-DA-111 model. As for the MG1655 genonme, the model predicts that regions refractory to Rho-dependent termination ('None') are fewer and smaller than regions eliciting termination. The proportions of genomic positions for which model predictions are reliable (in-model), or not (out-of-model), according to the $P_{mod}XPS^{+}$ = 0.05 threshold are shown inset. The C>G bubble content strongly correlates with the predicted termination strength of the regions (ANOVA P < 0.001; F = 1784).



Supplementary Figure 10: Variable loadings along Principal Component 1 of the OPLS-DA-111 model.



Supplementary Figure 11: The frequency of consensus RNAP elemental pause motifs is inversely correlated with the strength of Rho-dependent termination. ANOVA F- and P-values are shown above each plot.



Supplementary Figure 12: Examples of Rho-dependent termination signals detected in 5'UTRs or in upstream sections of the open reading frames of highly regulated genes. The location of termination-class regions predicted with the OPLS-DA-111 model are indicated at the top of the scaled gene diagrams. DNA templates used in *in vitro* transcription termination experiments are depicted by thick black lines. Grey arrows correspond to promoters for which direct experimental evidence is lacking.

Supplementary Table 1 : DNA templates[a]								
Template id	**Strain[b]**	**Strand**	**Genomic coordinates[c]**		**Template length (bp)**	**Termination type**	**Forward primer 1[d]**	**Reverse primer**
			start	**end**				
T001	MG1655	Plus	102182	102784	603	NONE	GTCTAACCTATAGGATACTTACAGCCGTTCTTTAGCTGAAATCAAACTGAAG	ACTTAGCCATTTTTCAATCAATAC
T002	MG1655	Plus	441467	441946	480	STRONG	GTCTAACCTATAGGATACTTACAGCCGCCTGGATGCAGCCGCAGGTGTGGGCTG	TTTAAACCGCCGCCGACGTAACGTTCCAC
T003	MG1655	Plus	569438	570190	753	WEAK	GTCTAACCTATAGGATACTTACAGCCCTTAGATGACAAAAGAACGGCTTTTG	ACACCTTTACTCTTCAAAGTTTTC
T004	MG1655	Plus	638614	639302	689	WEAK	GTCTAACCTATAGGATACTTACAGCCGCATAACCTATCACTGTCATAGG	GGTCAGCAACGTCACCCAGTTCGGTCGGGC
T005	MG1655	Minus	702686	703175	490	NONE	GTCTAACCTATAGGATACTTACAGCCTTAACGAACCGGCGTCTTCTCTGG	AACAACCGCGTGGTCATCAAGAAATTC
T006	MG1655	Minus	791678	792177	500	WEAK	GTCTAACCTATAGGATACTTACAGCCATCGCGCATAAAAAACGGCTAAATTCTTG	GGTGGCGGAGGAGCTAAAAATAAAG
T007	MG1655	Plus	813151	813650	500	STRONG	GTCTAACCTATAGGATACTTACAGCCAACACGAGGCAAGCGAGAGAATACGCGG	CCAGTCACGCCAAGTAACGTCTGGTGCGCC
T008	MG1655	Minus	913630	914129	500	STRONG	GTCTAACCTATAGGATACTTACAGCCGATTCGCCGCAGCCGCTCTACTGCACTGGG	ACATCGTGGTTGATGATGCCGTATTCAC
T009	MG1655	Plus	940823	941322	500	WEAK	GTCTAACCTATAGGATACTTACAGCCGAGTGAAAATCTACCTATCTCTTTG	ATTTCAGGCGGTCCGGATTGTAGACAC
T010	MG1655	Minus	1092122	1092523	402	WEAK	GTCTAACCTATAGGATACTTACAGCCAGGCATTGGGATTTATGCCGTATTCCTGAAG	CCATGACAAAGCTGGCTGAGTATTACCC
T011	MG1655	Minus	1123604	1124103	500	STRONG	GTCTAACCTATAGGATACTTACAGCCCTGACTCCTTCATACTGAAAGTGAACTG	AATCATCGCTGATATTCTTAATCAGAC
T012	MG1655	Plus	1157626	1157925	300	WEAK	GTCTAACCTATAGGATACTTACAGCCGCCGAAAATTGGGCGGTGAATAACCACG	CGGCAGCATCAGCGATTTACCGACCTTTTGC
T013	MG1655	Minus	1189402	1189931	530	WEAK	GTCTAACCTATAGGATACTTACAGCCTCGTTAATGCTCCAACTCTATCCGG	ATCCGCTTCAATTTCATGAAAACCATTC
T014	MG1655	Minus	1263057	1263556	500	STRONG	GTCTAACCTATAGGATACTTACAGCCGTAACGCTAGCATTAAGGGTTATAACTG	GTCGGTTGCATCACCCGGTAAACC
T015	MG1655	Plus	1263614	1263913	300	WEAK	GTCTAACCTATAGGATACTTACAGCCCAA	TCTTGCAGGTTGTCCTGCTC

							TTACTCCAAAAGGGGGCG	TTCAACGC
T016	MG1655	Plus	1330788	1331287	500	WEAK	GTCTAACCTATAGGATACTTACAGCCCGTATCGGATTTTATCAGGTACAGTGTG	CTCCCAATTGTGCCACGGGTCAACCCCC
T017	MG1655	Minus	1494635	1495133	499	STRONG	GTCTAACCTATAGGATACTTACAGCCCAAGTTATCATAATCAAACAACTTCACTTG	AAACCCGCCAGAATCATAGGGTTACTC
T018	MG1655	Plus	1531043	1531667	625	NONE	GTCTAACCTATAGGATACTTACAGCCCAGCCATTCCAGAGTTGCTTAACCTACTG	TAACGCAGCCTTCTTTCTTTTCCTC
T019	MG1655	Minus	1613677	1614256	580	NONE	GTCTAACCTATAGGATACTTACAGCCCATCATTCTTGCAGGTAACGG	CCCCAGCCATCTTTTCCCCGCCC
T020	MG1655	Minus	1744961	1745859	899	NONE	GTCTAACCTATAGGATACTTACAGCCTTAGTCTCTCTACGCTCGCTACGG	CTCAACGTGTATATCCCCGGTAACTTCCCC
T021	MG1655	Plus	1755466	1755735	270	STRONG	GTCTAACCTATAGGATACTTACAGCCGCCAATTGACTCTTGAATGGTTTCAG	TTCGGTCCGATGGTGCAAACAATTTTGGTC
T022	MG1655	Minus	1844707	1845310	604	NONE	GTCTAACCTATAGGATACTTACAGCCACCGACGCCGGAAAAGCGCTATTCCATTCG	CACAAACATTACAGCGACAACAC
T023	MG1655	Plus	1862240	1862739	500	STRONG	GTCTAACCTATAGGATACTTACAGCCAGCCGGAATCATACTTGGTTTGGG	AGGTTGCCTGTAAAATTACAAAAACCTTAC
T024	MG1655	Plus	1986775	1987274	500	STRONG	GTCTAACCTATAGGATACTTACAGCCGTTCAGACGTTATTCTTATTTCAG	ACGGTTGAATCATCATTCAGTTCTTTC
T025	MG1655	Plus	2071731	2072430	700	NONE	GTCTAACCTATAGGATACTTACAGCCGGCGGAACACTGGCAAATCATG	CCACCTGCAGTCTGCCCTTCTGATTAAC
T026	MG1655	Plus	2233989	2234488	500	STRONG	GTCTAACCTATAGGATACTTACAGCCACTAAACGCTCGCCTTAATTACCTATAG	AATCGTGACGTCATAACCCTGGTTTGTC
T027	MG1655	Plus	2408831	2409330	500	WEAK	GTCTAACCTATAGGATACTTACAGCCACCTGATGATGTCATCATACGTAAGG	AGAATTCATTATTTCCGGCCGCGAGTTC
T028	MG1655	Plus	2685797	2686296	500	STRONG	GTCTAACCTATAGGATACTTACAGCCGAGATACATCAATTAAGATGCAAAAAAAGG	ATCGCGAGTACCTTCCCAACCACC
T029	MG1655	Plus	2800656	2800968	313	NONE	GTCTAACCTATAGGATACTTACAGCCATCAACAGAGAGACAACCCGACGCG	CATGCACTGGCCGCGTGTGGCGCTGGATGC
T030	MG1655	Plus	2804577	2804880	304	STRONG	GTCTAACCTATAGGATACTTACAGCCGCCTGAAACCACAATATTCAGGCGTTTTTTCG	GAACGCTCGCTGTGGATGCTCGCCAAAT
T031	MG1655	Minus	2867530	2868119	590	STRONG	GTCTAACCTATAGGATACTTACAGCCCTTCGGGTGAACAGAGTGCTAACAAAATG	CTTTCAGCGTATTCTGACTCATAAGGTGGC

T032	MG1655	Plus	3201040	3201339	300	STRONG	GTCTAACCTATAGGATACTTACAGCCTAGTATTTTGCGCCAAATTGCCATG	GATAATGATCTCCCGGACCGCTGCGGACCC
T033	MG1655	Minus	3425268	3425840	573	NONE	GTCTAACCTATAGGATACTTACAGCCCACACGGCGGGTGCTAACGTCCGTCG	CTCAGCCTTGATTTTCCGGATTTGCC
T034	MG1655	Plus	3455482	3455981	500	STRONG	GTCTAACCTATAGGATACTTACAGCCAAATAAATTTTGCTTGATTCATGCAAGCGG	ATTTTTCTAATAAACGTTCCCGGCGC
T035	MG1655	Plus	3532678	3532927	250	WEAK	GTCTAACCTATAGGATACTTACAGCCCCGTTTCGTGACAGGAATCACGGAG	AGCTCTTCCTGATACAGCAGGTCGTAGC
T036	MG1655	Plus	3544667	3544966	300	STRONG	GTCTAACCTATAGGATACTTACAGCCACAGCGCACCAAATCCCCGGCTACGCCCG	CGCGTGAACAGACCGAACAAATCCCCC
T037	MG1655	Plus	3613438	3613987	550	STRONG	GTCTAACCTATAGGATACTTACAGCCATCACTCCTCGCTGTCAGTCAGCCATTG	TGCCGCCGCCTCGGCATCGAACGGTTCACC
T038	MG1655	Plus	3739674	3740173	500	WEAK	GTCTAACCTATAGGATACTTACAGCCCCTTGCTACAGAGTTCGACAGATATCCCG	TATTCGGACGCGCAGAGACAAACAGATC
T039	MG1655	Minus	3760806	3761297	492	WEAK	GTCTAACCTATAGGATACTTACAGCCTTAAACGCCCTTCTCCGTGTGAGAGGG	ACAGGCATCTTCCGCCCCCGTAATACGGC
T040	MG1655	Plus	3777289	3777788	500	STRONG	GTCTAACCTATAGGATACTTACAGCCGTCATTATCCCTACACAACACAATTGG	TTCAAGGAACGCGCCGAAACAGAAACC
T041	MG1655	Minus	3835899	3836665	767	STRONG	GTCTAACCTATAGGATACTTACAGCCTGGTGAAGTTTATAGTCAGTTTTTTTCG	TCTCTTTAACGGACAACACTTCACTCTTC
T042	MG1655	Plus	3884071	3884423	353	WEAK	GTCTAACCTATAGGATACTTACAGCCGGACGATCCTTGCGCTTTACCCATCAGCCCG	CCTGACGACCATTTTTAGTAGCCATAC
T043	MG1655	Plus	4001191	4001633	443	STRONG	GTCTAACCTATAGGATACTTACAGCCCACGATCACTCTAAGAGGACATTCGCC	CGTCGTCTTCAAAGAAACGTGCCGATGC
T044	MG1655	Minus	4015844	4016343	500	STRONG	GTCTAACCTATAGGATACTTACAGCCCCATGACGCAAATGCGTTGCATAA	CTAAGCAACGTGGGGATATCGGCAAAATC
T045	MG1655	Plus	4100759	4101258	500	STRONG	GTCTAACCTATAGGATACTTACAGCCACTGCTTACGCGGCATTAACAATCGGCCG	GGAGAATCCTGGTTAGCAGTAGAAACC
T046	MG1655	Plus	4112869	4113266	398	WEAK	GTCTAACCTATAGGATACTTACAGCCATGGTAATCCATAAGATCATTACTTG	TCAAGCTAACGAACGCAGTGATAACTCAC
T047	MG1655	Plus	4163399	4163848	450	STRONG	GTCTAACCTATAGGATACTTACAGCCGCCGGTCCTGTGAGTTAATAGGG	GCCAAGACGGCGCAGCACATCATTGACC
T048	MG1655	Minus	4256003	4256502	500	STRONG	GTCTAACCTATAGGATACTTACAGCCGCA	GCGAAAAATTTCTGTACGCC

							TTTGCATTATTAACCAGAGG	GTTAAGC
T049	MG1655	Minus	4287096	4287595	500	STRONG	GTCTAACCTATAGGATACTTACAGCCAGAGTTAGTCAGTATCTTCCTCTTTTTCAACAG	ACGATCGCCGTTTTCTTGCAGATGGC
T050	MG1655	Minus	4579499	4579917	419	WEAK	GTCTAACCTATAGGATACTTACAGCCGTATTTAGCGCGGTGCGGATGTGCG	TTACGCGACTTTCTGTTTACCGGCAATCAC
T051	MG1655	Minus	4605195	4605694	500	WEAK	GTCTAACCTATAGGATACTTACAGCCAATCATTATCATTTAGATTACTATCCCG	TTCGATGCTGCGGCGAATGTGGTGTTGC
T052	MG1655	Plus	4611153	4611671	519	WEAK	GTCTAACCTATAGGATACTTACAGCCGTGATGACATTTCTGACGGCGTTAAATACCG	GAAACCGCTCAGGGTCACGACTTTTTGATC
T053	MG1655	Plus	4036001	4036589	589	NONE	GTCTAACCTATAGGATACTTACAGCCTTTGCTCATTGACGTTACCCGCAGAAG	GCCATGCAGCACCTGTCTCACGGTTCCC
T054	MG1655	Plus	4479588	4480163	576	NONE	GTCTAACCTATAGGATACTTACAGCCTCTTCTTACGTACGCAAGCGACTTATAAAG	ACCCACCAGAACCCTTTCATAGAGC
T055	LT2	Plus	99280	100060	781	NONE	GTCTAACCTATAGGATACTTACAGCCTGCTGGCCGCCAGCGGCGCATCCACCACG	AGCACCACCATCTTCACCCCGCTCGACAGC
T056	LT2	Plus	149442	150307	866	NONE	GTCTAACCTATAGGATACTTACAGCCCGTATTCGTCGCGTTACTTCGTTCTGG	CAATGACCATGCGGCAAGACCGCCGGGGCC
T057	LT2	Plus	185749	186248	500	WEAK	GTCTAACCTATAGGATACTTACAGCCGCGGTACCGGGCATTTACCCTACTAACTACTG	TCGTAGAAGTTATCGAACATCAGCAGGGTGTGGGAC
T058	LT2	Plus	328871	329601	731	NONE	GTCTAACCTATAGGATACTTACAGCCTAACCAGTCTGTACGGGATCAGAAG	TAATAACCAGCACACCGCCAGTCATATAC
T059	LT2	Plus	334870	335693	824	NONE	GTCTAACCTATAGGATACTTACAGCCAGCGGTTATCTGACGGCGATACAGACCACG	GTGGTCAGCTCGCCCCACTCGTTATAGCC
T060	LT2	Plus	335392	336215	824	NONE	GTCTAACCTATAGGATACTTACAGCCCCTGTTCTTCCTGGCCAGCGCCCGGCGCGG	ACAAACCGTCCACACTCCGGTGCATAATATC
T061	LT2	Plus	659505	660301	797	NONE	GTCTAACCTATAGGATACTTACAGCCATTACCGGCGTGTATGCGCATATTGGCTG	CCCTCCCTGCTGACGGCGAAACTGCGC
T062	LT2	Plus	687775	688274	500	STRONG	GTCTAACCTATAGGATACTTACAGCCACGTATACATCGTAAGGAGTATTTTGG	GCCAGGCGTTTATAGTCCCGGTTTTTTAC
T063	LT2	Plus	713480	714219	740	WEAK	GTCTAACCTATAGGATACTTACAGCCAATCGGGCAGCACAAGCTGTCCGG	TCCGCCAATCGCCACCGTCATGAGTCCC
T064	LT2	Plus	870420	870849	430	STRONG	GTCTAACCTATAGGATACTTACAGCCTAAATAGCACGATCATATCGCTATGTATATG	CCGGTAAGACGCACCTTCTCCGTGCCAAGA

T065	LT2	Plus	1023667	1024397	731	NONE	GTCTAACCTATAGGATACTTACAGCCTATCAGGAATTCAGCAATATTTTTG	TCGATCTCATCCAGCAACAGCACCGC
T066	LT2	Plus	1053980	1054460	481	WEAK	GTCTAACCTATAGGATACTTACAGCCTAAATAGAGTGTGGTTTTAATCAAAAAATG	AGTTTTCTGAAACATTGAGGAGACGATG
T067	LT2	Minus	1318898	1319397	500	WEAK	GTCTAACCTATAGGATACTTACAGCCAGACTGTTCTTATTGTTAACACAAGGGAG	ATTCGAACGCCGTGAGTTTGATGACCTC
T068	LT2	Plus	1391702	1392200	499	WEAK	GTCTAACCTATAGGATACTTACAGCCACCTAAAAAAGTAAATCCGGTTACGCAAG	CCCTTAAATACCGCAATAAAATATTGC
T069	LT2	Minus	1444798	1445120	323	STRONG	GTCTAACCTATAGGATACTTACAGCCTCCCGTCGGGCGGCATTGCCG	AGCTGGCTTATGTAACAATGCGGC
T070	LT2	Plus	1472804	1473655	852	NONE	GTCTAACCTATAGGATACTTACAGCCATTCTTAATGCGGCTCAGGCGGGCG	TCCGCCAGGCTACCGGCCTGCATTTTTTCC
T071	LT2	Plus	1481823	1482584	762	NONE	GTCTAACCTATAGGATACTTACAGCCTCAAGTACATAGTGTAATCAGG	TTAACACGCTCCCCGGCCCTTCGCTGGATAC
T072	LT2	Plus	1485301	1486071	771	NONE	GTCTAACCTATAGGATACTTACAGCCTCAACGTCTTTGGTAATAACG	GCACACATTTTTCCAGGGCTTCTTTTGCC
T073	MG16655	Minus	1967087	1967766	680	NONE	GTCTAACCTATAGGATACTTACAGCCAAAATTCACGATGGCCCGGCGGTTAACCG	CTCCAGCGTCGCGGCGGTAAATGGCTTCAC
T074	LT2	Minus	2164234	2165017	784	WEAK	GTCTAACCTATAGGATACTTACAGCCCGCTCTTGCCGCGTTTTGTGCG	CCTATCATTCCGTTTTAAATAATC
T075	LT2	Minus	2213844	2214245	402	NONE	GTCTAACCTATAGGATACTTACAGCCTTGCTTCCTCTATTGTAGTCAGC	GCGCCTCGCAAGCCCAGCGCTGATGGATTG
T076	LT2	Minus	2290444	2290943	500	WEAK	GTCTAACCTATAGGATACTTACAGCCATGACAGTCTGATGGAAACGCCGCATCG	TCATGGATGATGGCGAACGGCACGC
T077	LT2	Minus	2290809	2291308	500	WEAK	GTCTAACCTATAGGATACTTACAGCCACAAATATGCCCTTTGTCCTG	CGGTGACCATTTCATCCACCTTC
T078	LT2	Minus	2340733	2341545	813	NONE	GTCTAACCTATAGGATACTTACAGCCCTTTGCTATGGCAACAGAGGCAACCTCAAG	CGCTCACGATGTTCAACAACACCCAGC
T079	LT2	Minus	2547320	2548129	810	NONE	GTCTAACCTATAGGATACTTACAGCCACCACCGGTTACATGGGCAGCGCAG	ACAACTGCGAGGTCACCGGGCGAAAGCAC
T080	LT2	Minus	2975443	2975971	529	WEAK	GTCTAACCTATAGGATACTTACAGCCTCAACAGTACGAATTCACTATCCGG	CCTTCGATTTCCGCTTTCGGCGTTAGCGCC
T081	LT2	Minus	3069329	3070086	758	NONE	GTCTAACCTATAGGATACTTACAGCCCTG	TCCCACCAGTTCCAGCTTAA

							GCGAAATTCCTGAAAATTCACG	TTGTTGC
T082	LT2	Minus	3074966	3075449	484	WEAK	GTCTAACCTATAGGATACTTACAGCCCGT TAAGCATTTTTTGATATTGGTTTTTGG	CGGTGTGTTTAGCGCTACCG TGAACGAAC
T083	LT2	Minus	3099974	3100513	540	WEAK	GTCTAACCTATAGGATACTTACAGCCCTCC CTGTGTGAATGTCAGCCTGACGAAG	CCACATCGTGGCCTTCGCCA CCTTCCAGCAC
T084	LT2	Minus	3109944	3110792	849	NONE	GTCTAACCTATAGGATACTTACAGCCCCG CGATCGATACCCACTGGATCTGGCAGG	CTGGCGAGTGTTTCCAGCCC AATGCCGCC
T085	LT2	Minus	3166660	3167790	1131	NONE	GTCTAACCTATAGGATACTTACAGCCTAA TGCGCTGCTGTTCGCGACGCTGG	CCGCCCGGCATATTCACTCA GTATGATGCC
T086	LT2	Minus	3250543	3251242	700	STRONG	GTCTAACCTATAGGATACTTACAGCCTCT GGTGCGCGCATGTCCCCCGCAATCTG	AGTAGTCGCCGTTATAACCG TAAGATTCAC
T087	LT2	Minus	3383394	3384118	725	NONE	GTCTAACCTATAGGATACTTACAGCCATA ATCTGGTGCGTCATCCGG	TTCCTGAACATTCTCCACGG TCTGAC
T088	LT2	Minus	3454134	3454956	823	NONE	GTCTAACCTATAGGATACTTACAGCCCTG GACACCCCGGGTCACGCCGCG	TCACTTCGTCGGTAGACAGT TTCAGC
T089	LT2	Minus	3462421	3463183	763	NONE	GTCTAACCTATAGGATACTTACAGCCCAT ATCGCGCCGAATGTGCTGCG	CCGCATCCGCGATACGATG GGCGAATGCC
T090	LT2	Plus	3494109	3494808	700	STRONG	GTCTAACCTATAGGATACTTACAGCCAAA AAGCAAGGTCAGATAATGCCTGGCGAG	CGGCCAGTTCAGGATCTTTA TTCAAAAATATC
T091	LT2	Plus	3623495	3623957	463	WEAK	GTCTAACCTATAGGATACTTACAGCCGAA AAGAAATCGAGGCAAAAATGAGCAAAG	CCGCGCTTGCTGCGACGGG CGCAGGCTTC
T092	LT2	Minus	3654621	3655358	738	NONE	GTCTAACCTATAGGATACTTACAGCCCCCT ATCCGCGCGTACTACGCGG	TTTCTGCGTCCATTTCCCCTC TTTTAAGC
T093	LT2	Minus	3815011	3815510	500	WEAK	GTCTAACCTATAGGATACTTACAGCCTTCC TTGCCGATTTAGCCATGGACTTTG	GGCGTTGCGGCGAAAAACC TGATCCCGCC
T094	LT2	Plus	3836275	3836651	377	WEAK	GTCTAACCTATAGGATACTTACAGCCAAC GGTTTGACGTACAGACCATTAAAGCAG	TTACAGGCTGGTTACGTTGC CAGC
T095	LT2	Minus	3815611	3816110	500	WEAK	GTCTAACCTATAGGATACTTACAGCCCCA GGTAAAAAACAAAAAAGGCCGGGCG	TAGCTGCCGCCAGACACTTT ATGGTAC
T096	LT2	Minus	3815692	3816191	500	WEAK	GTCTAACCTATAGGATACTTACAGCCTGA GGGGCATTTTATGGAGAATCCG	AGCTCGCGCGTGGGTTTGA AATCC
T097	LT2	Plus	4169555	4170054	500	WEAK	GTCTAACCTATAGGATACTTACAGCCATA TACAGTACCTTTACATTATGGATGTG	CGGAACAGAGTGTCGATGT CCACGGAACC

T098	LT2	Minus	4342947	4343378	432	STRONG	GTCTAACCTATAGGATACTTACAGCCGTTGGGATTAAACCAATATGCAGCGGGCCG	CGGTGTCTTCAAGGTTAAACGAGAAACCGCTCCGTT
T099	LT2	Minus	4548095	4548794	700	STRONG	GTCTAACCTATAGGATACTTACAGCCTGGCGTGACAATGATCACATAAGTCACATG	TAGATAATCGGCAAAATGGTGTAGACCAC
T100	LT2	Plus	4610454	4611153	700	STRONG	GTCTAACCTATAGGATACTTACAGCCTCGGGTAAGCCATTACGCTATCCGACACAG	TTTCGTTAAATGTTCGAGGATAAATTCCC
T101	LT2	Plus	4634763	4635270	508	WEAK	GTCTAACCTATAGGATACTTACAGCCCGATGAGTCCTGATAACAGGATCGTCG	CGCCAGCGAGGCGACCCAATCGGAAAGGCC
T102	LT2	Plus	4635163	4635672	510	WEAK	GTCTAACCTATAGGATACTTACAGCCACGTACTGGT	GGAACTCCACGGCGTCACCGACATAATGAC
T103	LT2	Plus	4635563	4636257	695	STRONG	GTCTAACCTATAGGATACTTACAGCCAAGATCCGGAAAAATCGCTGCCGCGCGCG	TTAACGCATTCCCGCCATCCGTTTCTTACC
T104	LT2	Plus	4710707	4711461	755	WEAK	GTCTAACCTATAGGATACTTACAGCCCTTACTTAATATATGAACCGGG	CCATCCCTTCCTGACCTTGTAGCACAATC

[a]: DNA templates devoid of C>G bubbles are listed in blue.
[b]: *Escherichia coli* MG1655 or *Salmonella typhimirium* LT2.
[c]: from NCBI reference sequences U00096.3 (MG1655) and NC_003197.1 (LT2).
[d]: primer used in a first round of PCR to amplify the region of interest (partial T7A1 promoter sequence is underlined); the complete T7A1 sequence was introduced in a second PCR round by using a distinct forward primer: 5'-TTATCAAAAAGAGTATTGACTTAAAGTCTAACCTATAGGATACTTACAGCC-3'.

Supplementary Table 2: Number of C>G bubbles						
Minimal length (nt)	***E.coli* MG1655[a]**			***Salmonella* LT2[a]**		
	Minus strand	**Plus strand**	**Total**	**Minus strand**	**Plus strand**	**Total**
20	23820	23770	47590	24726	24660	49386
50	14372	14316	28688	14849	14894	29743
80	10419	10463	20882	10829	10837	21666
100	7365	7438	14803	7586	7633	15219
150	3639	3783	7422	3884	3868	7752
200	1976	2020	3996	2126	2150	4276
250	1074	1146	2220	1201	1220	2421
300	598	679	1277	716	709	1425
400	225	255	480	267	271	538
500	88	85	173	92	108	200
600	38	35	73	41	42	83
700	12	17	29	20	19	39
800	4	5	9	12	12	24
900	3	3	6	2	4	6
[a] Genome sizes: 4,641,652 bp (MG1655) and 4,857,432 bp (LT2)						

Supplementary Table 3: Number of $[(YC)N_{9\rightarrow13}]_n$ motifs in genome						
n	***E.coli* MG1655**			***Salmonella* LT2**		
	Minus strand	**Plus strand**	**Total**	**Minus strand**	**Plus strand**	**Total**
1	537636	539215	1076851	577518	577509	1155027
2	242346	243973	486319	264686	264441	529127
3	117984	119204	237188	130828	131249	262077
4	59426	60242	119668	66750	67449	134199
5	30662	31031	61693	34936	35435	70371
6	16148	16203	32351	18672	19040	37712
7	8752	8637	17389	10125	10367	20492
8	4672	4700	9372	5497	5680	11177
9	2566	2648	5214	3106	3202	6308
10	1345	1463	2808	1814	1799	3613
11	743	838	1581	1060	1025	2085
12	397	479	876	649	587	1236
13	226	272	498	419	329	748
14	115	138	253	267	174	441
15	73	79	152	185	91	276
16	45	46	91	132	52	184
17	27	30	57	101	30	131
18	14	21	35	78	14	92
19	7	13	20	64	8	72
20	5	10	15	55	6	61
21	3	8	11	50	4	54
22	1	6	7	43	2	45
23	0	5	5	38	1	39
24	0	3	3	34	0	34
25	0	1	1	31	0	31
26	0	0	0	26	0	26
27	0	0	0	22	0	22
28	0	0	0	19	0	19
29	0	0	0	14	0	14
30	0	0	0	11	0	11
31	0	0	0	8	0	8
32	0	0	0	6	0	6
33	0	0	0	3	0	3
34	0	0	0	1	0	1
35	0	0	0	0	0	0

Supplementary Table 4: Sequence features of representative Rho-dependent terminators

Terminator	strand	C>G bubble[d]				Number of $[(YC)N_{9\rightarrow13}]_n$ motifs (n)				Pred[e]	Ref
		start[a]	end[a]	length	area	3	4	5	6		
chiP	+	708428 (708427)	708543 (708677)	116	1098,7	8	3	0	0	Weak	(1)
tfaS	+	2471085 (2471028)	2471238 (2471428)	154	1676,9	8	3	2	1	Strong	(2)
tnA	+	3888423 (3888420)	3888561 (3888650)	139	1044,9	9	5	3	0	Strong	(3)
ilv	+	3951910 (3951900)	3952089 (3952180)	180	1616,7	10	5	2	0	Strong	(4)
lacZ	-	366164 (366050)	366314 (366320)	151	1484,6	10	6	4	2	Strong	(5)
		365982 (365876)	366087 (366192)	106	1278,2	11	6	4	1	Weak	
pgaA	-	1092398 (1092287)	1092495 (1092523)	98	978,2	7	4	1	0	Strong	(6)
tyrT	-	1287050 (1286960)	1287152 (1287155)	103	1423,1	6	3	0	0	Strong	(7)
trpA	-	1316095 (1316010)	1316475 (1316480)	381	4174,4	11	5	1	0	Strong	(8)
λtR1[b]	+	38170[b] (38150)	38263[b] (38340)	94	1230,8	3	1	0	0	Strong	(9)
hisG[c]	+	2149821[c] (2149800)	2149903[c] (2149985)	83	443,6	2	1	0	0	Weak	(10)
		2149923[c] (2149915)	2150021[c] (2150110)	99	1105,1	13	10	5	1	Weak	(10)
mgtA[c]	+	4699628[c] (4699450)	4699766[c] (4699951)	139	1124,4	10	8	7	5	Strong	(11)

[a] : Reference sequence for *E. coli* MG1655 genome: U00096.3. Genomic coordinates in parentheses correspond to the DNA sequences encompassing the C>bubbles which were used for prediction with the OPLS-DA-111 model.
[b] : Reference sequence for phage lambda genome: NC_001416.1
[c] : Reference sequence for *Salmonella* LT2 genome: NC_003197.1
[d] : The length in bp and and area in %xbp are calculated as for Table S6.
[e] : Prediction of termination class with the OPLS-DA-111 model developed in the present work.

REFERENCES

1. Bossi, L., Schwartz, A., Guillemardet, B., Boudvillain, M. and Figueroa-Bossi, N. (2012) A role for Rho-dependent polarity in gene regulation by a noncoding small RNA. *Genes Dev*, **26**, 1864-1873.
2. Menouni, R., Champ, S., Espinosa, L., Boudvillain, M. and Ansaldi, M. (2013) Transcription termination controls prophage maintenance in Escherichia coli genomes. *Proc Natl Acad Sci U S A*, **110**, 14414-14419.
3. Stewart, V., Landick, R. and Yanofsky, C. (1986) Rho-dependent transcription termination in the tryptophanase operon leader region of Escherichia coli K-12. *J Bacteriol*, **166**, 217-223.
4. Wek, R.C., Sameshima, J.H. and Hatfield, G.W. (1987) Rho-dependent transcriptional polarity in the ilvGMEDA operon of wild-type Escherichia coli K12. *J Biol Chem*, **262**, 15256-15261.
5. Ruteshouser, E.C. and Richardson, J.P. (1989) Identification and characterization of transcription termination sites in the Escherichia coli lacZ gene. *J Mol Biol*, **208**, 23-43.

6. Figueroa-Bossi, N., Schwartz, A., Guillemardet, B., D'Heygere, F., Bossi, L. and Boudvillain, M. (2014) RNA remodeling by bacterial global regulator CsrA promotes Rho-dependent transcription termination. *Genes Dev*, **28**, 1239-1251.
7. Kupper, H., Sekiya, T., Rosenberg, M., Egan, J. and Landy, A. (1978) A rho-dependent termination site in the gene coding for tyrosine tRNA su3 of Escherichia coli. *Nature*, **272**, 423-428.
8. Zalatan, F., Galloway-Salvo, J. and Platt, T. (1993) Deletion analysis of the Escherichia coli rho-dependent transcription terminator trp t'. *J. Biol. Chem.*, **268**, 17051-17056.
9. Morgan, W.D., Bear, D.G. and von Hippel, P.H. (1983) Rho-dependent termination of transcription. I. Identification and characterization of termination sites for transcription from the bacteriophage lambda PR promoter. *J. Biol. Chem.*, **258**, 9553-9564.
10. Alifano, P., Rivellini, F., Limauro, D., Bruni, C.B. and Carlomagno, M.S. (1991) A consensus motif common to all Rho-dependent prokaryotic transcription terminators. *Cell*, **64**, 553-563.
11. Hollands, K., Proshkin, S., Sklyarova, S., Epshtein, V., Mironov, A., Nudler, E. and Groisman, E.A. (2012) Riboswitch control of Rho-dependent transcription termination. *Proc Natl Acad Sci U S A*, **109**, 5376-5381.

Supplementary Table 5: Total numbers of C>G bubbles containing $[(YC)N_{9\rightarrow13}]_n$ motifs [a]								
Bubble length (nt)	***E.coli* MG1655[b]**				***Salmonella* LT2[b]**			
	n = 3	**n = 4**	**n = 5**	**n = 6**	**n = 3**	**n = 4**	**n = 5**	**n = 6**
20	39756	26402	15489	8661	41619	27992	16602	9582
50	26495	20148	13074	7754	27497	21172	13956	8520
80	19950	16093	11010	6833	20698	16932	11781	7527
100	14406	12222	8844	5711	14841	12812	9394	6277
150	7348	6683	5249	3604	7677	7077	5660	4081
200	3974	3749	3116	2294	4258	4071	3456	2648
250	2215	2124	1841	1441	2415	2343	2088	1686
300	1275	1252	1129	899	1424	1401	1290	1078
400	480	476	453	388	538	536	520	463
500	173	170	164	148	200	200	197	184
600	73	73	73	68	83	83	82	77
700	29	29	29	25	39	39	39	36
800	9	9	9	9	24	24	24	24
900	6	6	6	6	6	6	6	6

[a]: Note that the motifs can end up to 78 nt farther from the downstream edge of the C>G bubble to account for the length of the last sliding window defining the bubble.
[b]: Reference genome sequences: U00096.3 (MG1655) and NC_003197.1 (LT2)

Supplementary Table 6: list and statistical significance of template sequence descriptors.						
Descriptor[b]	**Name**	**ANOVA[a]**				
		F	**P**	**Post hoc SNK test P**		
				Strong vs None	**Weak vs None**	**Strong vs Weak**
%A	%A	4.3	*	*	*	ns
%T	%T	4.9	**	**	*	ns
%G	%G	58.1	***	***	***	**
%C	%C	8.9	***	**	*	*
%AA	%AA	2.4	ns	-	-	-
%AT	%AT	3.3	*	ns	*	ns
%AG	%AG	5.9	**	**	ns	ns
%AC	%AC	8.8	***	***	ns	*
%TA	%TA	4.1	*	*	*	ns
%TT	%TT	4.2	*	*	ns	ns
%TG	%TG	15.2	***	***	***	ns
%TC	%TC	19.7	***	***	***	ns
%GA	%GA	12.9	***	***	*	**
%GT	%GT	10.6	***	***	*	*
%GG	%GG	42.1	***	***	***	*
%GC	%GC	8.5	***	***	**	ns
%CA	%CA	17.2	***	***	**	**
%CT	%CT	16.8	***	***	**	*
%CG	%CG	6.0	**	**	*	ns
%CC	%CC	11.0	***	***	**	ns
%ATT	%ATT	3.8	*	*	*	ns
%ATC	%ATC	7.6	***	**	**	ns
%ATA	%ATA	1.3	ns	-	-	-
%ATG	%ATG	4.8	**	*	ns	*
%AAT	%AAT	1.9	ns	-	-	-
%AAC	%AAC	6.6	**	**	**	*
%AAA	%AAA	2.7	ns	-	-	-
%AAG	%AAG	3.5	*	ns	*	ns
%ACT	%ACT	8.3	***	***	**	ns
%ACC	%ACC	2.6	ns	-	-	-
%ACA	%ACA	12.9	***	***	**	*
%ACG	%ACG	3.0	ns	-	-	-
%AGT	%AGT	1.1	ns	-	-	-
%AGC	%AGC	4.2	*	*	*	ns
%AGA	%AGA	1.4	ns	-	-	-
%AGG	%AGG	6.9	**	**	ns	**
%TTT	%TTT	5.1	**	**	ns	ns
%TTC	%TTC	7.8	***	***	*	ns
%TTA	%TTA	1.7	ns	-	-	-
%TTG	%TTG	0.1	ns	-	-	-
%TAT	%TAT	4.6	*	ns	**	ns

%TAC	%TAC	1.1	ns	-	-	-
%TAA	%TAA	1.4	ns	-	-	-
%TAG	%TAG	0.8	ns	-	-	-
%TCT	%TCT	6.9	**	**	ns	*
%TCC	%TCC	9.1	***	**	***	ns
%TCA	%TCA	8.2	***	***	**	ns
%TCG	%TCG	1.3	ns	-	-	-
%TGT	%TGT	1.7	ns	-	-	-
%TGC	%TGC	1.2	ns	-	-	-
%TGA	%TGA	3.8	*	ns	ns	*
%TGG	%TGG	17.9	***	***	***	ns
%CTT	%CTT	10.7	***	***	*	*
%CTC	%CTC	13.9	***	***	***	ns
%CTA	%CTA	14.3	***	***	***	ns
%CTG	%CTG	2.8	ns	-	-	-
%CAT	%CAT	7.7	***	***	*	ns
%CAC	%CAC	11.6	***	***	**	*
%CAA	%CAA	13.0	***	***	***	ns
%CAG	%CAG	3.4	*	ns	*	ns
%CCT	%CCT	15.9	***	***	***	ns
%CCC	%CCC	16.5	***	***	***	ns
%CCA	%CCA	6.4	**	**	ns	*
%CCG	%CCG	0.3	ns	-	-	-
%CGT	%CGT	0.4	ns	-	-	-
%CGC	%CGC	0.1	ns	-	-	-
%CGA	%CGA	1.5	ns	-	-	-
%CGG	%CGG	21.9	***	***	***	ns
%GTT	%GTT	2.2	ns	-	-	-
%GTC	%GTC	0.4	ns	-	-	-
%GTA	%GTA	5.3	**	*	ns	**
%GTG	%GTG	19.3	***	***	***	*
%GAT	%GAT	1.7	ns	-	-	-
%GAC	%GAC	0.3	ns	-	-	-
%GAA	%GAA	6.3	**	**	ns	**
%GAG	%GAG	6.1	**	**	*	ns
%GCT	%GCT	2.3	ns	-	-	-
%GCC	%GCC	2.7	ns	-	-	-
%GCA	%GCA	0.3	ns	-	-	-
%GCG	%GCG	19.7	***	***	***	ns
%GGT	%GGT	16.2	***	***	*	***
%GGC	%GGC	23.0	***	***	***	ns
%GGA	%GGA	20.2	***	***	***	ns
%GGG	%GGG	37.0	***	***	***	ns
Cumulated length of all C>G bubbles	SumBubLength	107.3	***	***	***	***
Cumulated area of all C>G bubbles	SumBubSurf	63.4	***	***	***	***
Density of C>G bubbles (in length)	DenBubLength	109.8	***	***	***	***

Density of C>G bubbles (in area)	DenBubSurf	60.9	***	***	***	***
Length of the longest C>G bubble	Bub_{long}Length	70.6	***	***	***	***
Area of the longest C>G bubble	Bub_{long}Surf	48.8	***	***	***	***
Average %C-%G in longest C>G bubble	$(C\text{-}G)_{av}$%Bub	64.6	***	***	***	***
maximal %C-%G in longest C>G bubble	$(C\text{-}G)_{max}$%Bub	78.6	***	***	***	***
Maximal %C in longest C>G bubble	$\%C_{max}$Bub	99.9	***	***	***	*
Minimal %G in longest C>G bubble	$\%G_{min}$Bub	47.8	***	***	***	ns
Average %C in longest C>G bubble	$\%C_{av}$Bub	91.8	***	***	***	ns
Average %G in longest C>G bubble	$\%G_{av}$Bub	73.4	***	***	***	ns
Average %A in longest C>G bubble	$\%A_{av}$Bub	86.7	***	***	***	ns
Average %T in longest C>G bubble	$\%T_{av}$Bub	92.2	***	***	***	*
Number of $[(YC)N_{9\rightarrow13}]_1$ motifs (aka YC dimers) in longest C>G bubble	Nb_Reg1Bu	77.8	***	***	***	***
Number of $[(YC)N_{9\rightarrow13}]_2$ motifs in longest C>G bubble	Nb_Reg2Bu	41.5	***	***	***	***
Number of $[(YC)N_{9\rightarrow13}]_3$ motifs in longest C>G bubble	Nb_Reg3Bu	20.6	***	***	***	*
Number of $[(YC)N_{9\rightarrow13}]_4$ motifs in longest C>G bubble	Nb_Reg4Bu	12.0	***	***	***	ns
Number of $[(YC)N_{9\rightarrow13}]_5$ motifs in longest C>G bubble	Nb_Reg5Bu	5.9	***	***	***	ns
Number of $[(YC)N_{9\rightarrow13}]_6$ motifs in longest C>G bubble	Nb_Reg6Bu	3.5	*	*	*	ns
Density of $[(YC)N_{9\rightarrow13}]_1$ motifs (aka %YC)	Den_Reg1	29.6	***	***	***	*
Density of $[(YC)N_{9\rightarrow13}]_2$ motifs	Den_Reg2	21.4	***	***	***	ns
Density of $[(YC)N_{9\rightarrow13}]_3$ motifs	Den_Reg3	16.2	***	***	***	ns
Density of $[(YC)N_{9\rightarrow13}]_4$ motifs	Den_Reg4	11.6	***	***	***	ns
Density of $[(YC)N_{9\rightarrow13}]_5$ motifs	Den_Reg5	6.9	**	**	**	ns
Density of $[(YC)N_{9\rightarrow13}]_6$ motifs	Den_Reg6	4.3	*	*	*	ns
Minimal RNA folding energy (averaged per kilobase) for the full-length sequence	$\Delta G_{Mfold[av]}$	24.8	***	***	***	ns

[a]One-way ANOVA analyses were performed with Kaleidagraph 4.0 using a *post hoc* Student-Newman-Keuls test when relevant. ns: $P > 0.05$; *: $P \leq 0.05$; **: $P \leq 0.01$; ***: $P \leq 0.001$.
[b]Descriptors are for the non-template DNA strand.

Supplementary Table 7: VIP (Variable Importance in the Projection) score rankings for the OPLS-DA models listed in Table 1.

OPLS-DA-111			OPLS-DA-40			OPLS-DA-83			ANOVA[a]	
Descriptor[b]	VIP	$_{cv}$SE[c]	Descriptor[b]	VIP	$_{cv}$SE[c]	Descriptor[b]	VIP	$_{cv}$SE[c]	Descriptor[b]	F
BublongSurf	1.61	0.44	BublongSurf	1.27	0.23	$\%G_{min}Bub$	1.52	0.55	DenBubLength	109.8
$\%G_{min}Bub$	1.57	0.64	SumBubSurf	1.21	0.18	BublongSurf	1.48	0.38	SumBubLength	107.3
DenBubSurf	1.53	0.32	DenBubSurf	1.20	0.18	$\%G_{av}Bub$	1.44	0.38	$\%C_{max}Bub$	99.9
$\%G_{av}Bub$	1.49	0.45	BublongLength	1.19	0.17	DenBubSurf	1.41	0.33	$\%T_{av}Bub$	92.2
SumBubSurf	1.48	0.34	Nb_Reg1Bu	1.14	0.15	SumBubSurf	1.37	0.30	$\%C_{av}Bub$	91.8
BublongLength	1.44	0.32	DenBubLength	1.13	0.10	$\%C_{av}Bub$	1.36	0.25	$\%A_{av}Bub$	86.7
$\%C_{av}Bub$	1.43	0.32	%G	1.12	0.12	$\%C_{max}Bub$	1.33	0.20	$(C\text{-}G)_{max}\%Bub$	78.6
$\%C_{max}Bub$	1.41	0.26	$(C\text{-}G)_{max}\%Bub$	1.10	0.11	BublongLength	1.33	0.27	Nb_Reg1Bu	77.8
$\%A_{av}Bub$	1.40	0.33	$\%C_{max}Bub$	1.09	0.16	$\%A_{av}Bub$	1.32	0.25	$\%G_{av}Bub$	73.4
DenBubLength	1.39	0.16	SumBubLength	1.09	0.07	$\%T_{av}Bub$	1.29	0.17	BublongLength	70.6
Nb_Reg1Bu	1.39	0.25	$\%C_{av}Bub$	1.08	0.19	DenBubLength	1.28	0.13	$(C\text{-}G)_{av}\%Bub$	64.6
$\%T_{av}Bub$	1.38	0.17	$(C\text{-}G)_{av}\%Bub$	1.08	0.15	Nb_Reg1Bu	1.27	0.22	SumBubSurf	63.4
%G	1.38	0.17	%GG	1.08	0.14	%G	1.26	0.16	DenBubSurf	60.9
$(C\text{-}G)_{max}\%Bub$	1.37	0.18	$\%A_{av}Bub$	1.07	0.23	$(C\text{-}G)_{max}\%Bub$	1.25	0.15	%G	58.1
$(C\text{-}G)_{av}\%Bub$	1.34	0.20	$\%G_{av}Bub$	1.06	0.30	$(C\text{-}G)_{av}\%Bub$	1.23	0.18	BublongSurf	48.8
SumBubLength	1.32	0.12	%GGC	1.06	0.20	SumBubLength	1.22	0.09	$\%G_{min}Bub$	47.8
%GG	1.31	0.24	Nb_Reg2Bu	1.06	0.23	%GG	1.21	0.22	%GG	42.1
%GGG	1.31	0.59	$\%G_{min}Bub$	1.06	0.44	%GGC	1.19	0.24	Nb_Reg2Bu	41.5
%CA	1.28	0.63	$\%T_{av}Bub$	1.05	0.19	%GGG	1.18	0.49	%GGG	37
Nb_Reg2Bu	1.28	0.37	$\Delta G_{Mfold[av]}$	0.99	0.25	Nb_Reg2Bu	1.17	0.33	Den_Reg1	29.6
%GGC	1.27	0.25	Den_Reg1	0.99	0.19	%CA	1.14	0.57	$\Delta G_{Mfold[av]}$	24.8
Den_Reg1	1.23	0.26	%GGG	0.98	0.31	Den_Reg1	1.12	0.20	%GGC	23
$\Delta G_{Mfold[av]}$	1.19	0.24	Den_Reg2	0.96	0.15	$\Delta G_{Mfold[av]}$	1.11	0.25	%CGG	21.9
Den_Reg2	1.18	0.24	%CGG	0.94	0.34	%C	1.08	0.23	Den_Reg2	21.4
%C	1.14	0.22	%GCG	0.92	0.33	Den_Reg2	1.07	0.20	Nb_Reg3Bu	20.6
%CGG	1.13	0.43	Nb_Reg3Bu	0.91	0.23	%CGG	1.05	0.43	%GGA	20.2
%GC	1.11	0.47	Den_Reg3	0.90	0.23	%GC	1.04	0.42	%TC	19.7
%GAA	1.11	0.55	%AAG	0.89	0.38	%AAG	1.03	0.89	%GCG	19.7
%TC	1.11	0.50	%GC	0.89	0.46	Nb_Reg3Bu	1.02	0.33	%GTG	19.3
Den_Reg3	1.10	0.32	%C	0.89	0.41	%CTA	1.02	0.26	%TGG	17.9
Nb_Reg3Bu	1.10	0.39	%TGG	0.87	0.23	%GCG	1.02	0.43	%CA	17.2
%GCG	1.08	0.41	%CA	0.85	0.35	%TA	1.01	0.44	%CT	16.8
%AAG	1.08	0.91	%CC	0.85	0.25	Den_Reg3	1.00	0.25	%CCC	16.5
%GA	1.07	0.67	%GAA	0.85	0.38	%GAA	0.97	0.50	%GGT	16.2
%CTA	1.06	0.24	%TC	0.85	0.43	%TC	0.96	0.42	Den_Reg3	16.2
%TA	1.05	0.50	Den_Reg4	0.82	0.31	%CC	0.96	0.15	%CCT	15.9
%GGA	1.05	0.24	%TA	0.82	0.20	%TGG	0.95	0.37	%TG	15.2
%CC	1.02	0.15	%GA	0.82	0.43	%GGT	0.94	0.46	%CTA	14.3
%TGG	1.01	0.40	%GGA	0.78	0.29	%GGA	0.94	0.19	%CTC	13.9
Den_Reg4	1.00	0.41	%CTA	0.75	0.25	Den_Reg4	0.93	0.32	%CAA	13
%GGT	1.00	0.46				%TAT	0.92	0.59	%GA	12.9

%CT	0.98	0.32				%GA	0.92	0.61	%ACA	12.9
%CTT	0.98	1.09				Nb_Reg4Bu	0.91	0.25	Nb_Reg4Bu	12
%CG	0.98	0.12				%CAC	0.89	0.51	%CAC	11.6
%CAC	0.96	0.51				%CT	0.89	0.28	Den_Reg4	11.6
Nb_Reg4Bu	0.96	0.30				Den_Reg5	0.88	0.48	%CC	11
%TT	0.96	0.43				%GT	0.88	0.81	%CTT	10.7
%A	0.95	0.84				%GTG	0.88	0.38	%GT	10.6
%GTG	0.95	0.43				%CTT	0.88	1.06	%TCC	9.1
%TAT	0.95	0.70				%T	0.87	0.20	%C	8.9
%T	0.95	0.17				%TT	0.87	0.22	%AC	8.8
%TCC	0.94	0.77				%TCC	0.87	0.69	%GC	8.5
%CAG	0.94	0.73				%CG	0.86	0.20	%ACT	8.3
%TGA	0.94	0.92				%CAT	0.86	0.49	%TCA	8.2
%CTC	0.93	0.19				%AT	0.86	0.36	%TTC	7.8
%GTA	0.93	0.81				%AC	0.86	0.78	%CAT	7.7
%GCA	0.93	0.62				%CTC	0.85	0.16	%ATC	7.6
Den_Reg5	0.93	0.60				%GTA	0.84	0.70	%AGG	6.9
%GT	0.92	0.84				%A	0.84	0.82	%TCT	6.9
%AT	0.92	0.36				%CCT	0.84	0.28	Den_Reg5	6.9
%AA	0.91	0.89				%CCC	0.82	0.26	%AAC	6.6
%AG	0.91	0.67				Den_Reg6	0.82	0.49	%CCA	6.4
%AC	0.91	0.93				%TGA	0.82	0.69	%GAA	6.3
%CCC	0.89	0.28				%CCA	0.81	0.29	%GAG	6.1
%CCA	0.89	0.27				Nb_Reg5Bu	0.80	0.21	%CG	6
%CCT	0.88	0.30				%AG	0.80	0.79	%AG	5.9
%CAT	0.87	0.42				%TTT	0.80	0.26	Nb_Reg5Bu	5.9
%TG	0.87	0.22				%ATT	0.79	0.12	%GTA	5.3
%ACA	0.86	0.82				%TCA	0.79	0.15	%TTT	5.1
%TCA	0.86	0.24				%ACA	0.78	0.63	%T	4.9
Den_Reg6	0.85	0.63				%TG	0.77	0.14	%ATG	4.8
%TTT	0.85	0.45				%AGG	0.77	0.62	%TAT	4.6
%AGG	0.84	0.52				Nb_Reg6Bu	0.76	0.37	%A	4.3
%AAT	0.84	0.46				%CAG	0.75	0.66	Den_Reg6	4.3
%ATC	0.84	0.76				%AAC	0.74	0.60	%TT	4.2
%CGT	0.83	0.67				%CAA	0.73	0.33	%AGC	4.2
%TAA	0.83	0.27				%ATC	0.72	0.62	%TA	4.1
%TCG	0.83	1.07				%GAG	0.68	0.84	%ATT	3.8
Nb_Reg5Bu	0.82	0.17				%ACT	0.67	0.41	%TGA	3.8
%AAC	0.81	0.73				%TTC	0.66	0.40	%AAG	3.5
%ATT	0.81	0.12				%TCT	0.63	0.27	Nb_Reg6Bu	3.5
%CAA	0.81	0.45				%ATG	0.59	0.64	%CAG	3.4
%TGC	0.81	1.01				%AGC	0.45	0.28	%AT	3.3
%TTA	0.81	0.10							%ACG	3
%CCG	0.80	0.59							%CTG	2.8
%CTG	0.78	0.64							%AAA	2.7

%AAA	0.78	0.62							%GCC	2.7
%GAG	0.77	0.80							%ACC	2.6
Nb_Reg6Bu	0.76	0.40							%AA	2.4
%ACT	0.73	0.43							%GCT	2.3
%CGC	0.72	0.17							%GTT	2.2
%ACC	0.72	0.27							%AAT	1.9
%ATA	0.71	0.24							%TTA	1.7
%TTC	0.69	0.40							%TGT	1.7
%AGA	0.68	0.82							%GAT	1.7
%TCT	0.68	0.29							%CGA	1.5
%GCC	0.67	0.47							%AGA	1.4
%GCT	0.63	0.30							%TAA	1.4
%TTG	0.63	0.92							%ATA	1.3
%ATG	0.63	0.89							%TCG	1.3
%ACG	0.60	0.89							%TGC	1.2
%GAC	0.57	0.61							%AGT	1.1
%GTC	0.56	0.68							%TAC	1.1
%TAG	0.55	0.53							%TAG	0.8
%CGA	0.55	0.29							%CGT	0.4
%GTT	0.53	0.28							%GTC	0.4
%AGT	0.53	0.25							%CCG	0.3
%TAC	0.51	0.55							%GAC	0.3
%AGC	0.50	0.27							%GCA	0.3
%GAT	0.48	1.16							%TTG	0.1
%TGT	0.43	0.27							%CGC	0.1

[a]Descriptors are ranked according to F-values (descriptors with ANOVA P > 0.05 are shown in red).
[b]Definitions of the descriptors are provided in Supplementary Table 6.
[c]jack-knife cross-validated standard error as implemented in SIMCA-P.

II. Travaux non publiés I :

Un essai fluorogènique pour suivre la terminaison de la transcription Rho-dépendante

Les travaux décrits dans cette section forment un ensemble homogène et finalisé qui devrait déboucher rapidement sur une publication dans un journal à comité de lecture. Je présente ci-dessous une version adaptée en Français du manuscrit qui est en cours de préparation.

1) Détection de la terminaison de la transcription Rho-dépendante à l'aide de sondes fluorescentes de type « Molecular Beacon »

Pour détecter la terminaison de la transcription Rho-dépendante en utilisant la fluorescence, j'ai adapté un essai fluorogénique développé par Lafontaine et ses collaborateurs pour surveiller la régulation de la transcription par un riboswitch (Chinnappan et al., 2013). Cet essai repose sur l'utilisation simultanée de deux sondes oligonucléotidiques dont la fluorescence est exaltée lors de l'appariement avec une séquence cible (*Molecular Beacon probe* [MBP]). Une première sonde MBP s'hybride au transcrit naissant en amont du riboswitch (pour détecter l'activité de transcription du promoteur), tandis que la seconde sonde MBP se lie en aval du motif régulateur et ne détecte que les transcrits de plus grande longueur. Cette configuration permet une surveillance simultanée de l'efficacité de la transcription et de sa régulation par le riboswitch (Chinnappan et al., 2013). Suivant un principe similaire, j'ai conçu une matrice ADN codant pour un promoteur fort (T7A1) en amont d'une cassette « *reporter* » contenant deux séquences cibles distinctes (séquences antiMBP$_C$ et antiMBP$_T$) situées de part et d'autre du terminateur Rho-dépendant *λtR1* (**Figure 44**).

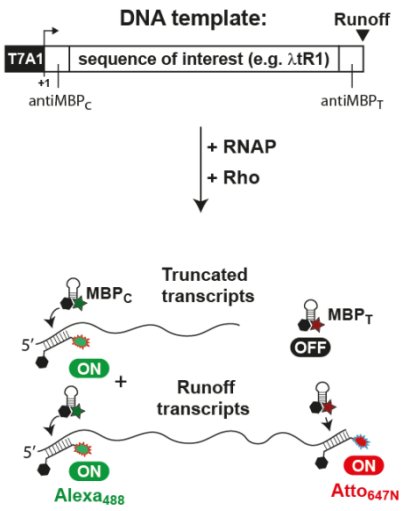


Figure 44 : Principe de l'essai fluorogénique pour le suivi de la terminaison Rho dépendante. La matrice *λtR1$_{C/T}$* est schématisée en haut de la figure.

J'ai utilisé les séquences MPB (et antiMBP) optimisées pour l'étude des riboswitches (Chinnappan et al., 2013 1912) mais portant des fluorochromes plus compatibles avec l'activité de la protéine Rho que les originaux Cy3 et Cy5 (voir ci-dessous, page 130). Ainsi, la sonde « amont » MBP$_C$ porte le fluorochrome Alexa488 et le *quencher* Dabcyl tandis que la sonde « aval » MBP$_T$ porte le fluorochrome Atto647N et le *quencher* BHQ2 (**Figure 44**). Pour suivre les réactions

de transcription, j'ai utilisé des microplaques 384-puits et un lecteur de microplaques car cette configuration économise les réactifs et convient mieux aux applications éventuelles de criblage haut-débit que le format standard développé pour la caractérisation des riboswitches (Chinnappan et al., 2013).

J'ai d'abord réalisé des expériences de transcription avec la matrice ADN contenant *λtR1*, les sondes MBP_C et MBP_T (**Figure 44**) et l'ARNP d'*Escherichia coli* (voir méthodes, page 132). J'ai observé une augmentation de la fluorescence des deux sondes en fonction du temps (**Figure 45A**), ce qui correspond au résultat attendu pour la production en continu de transcrits « *runoff* » de longueur totale (**Figure 44**).

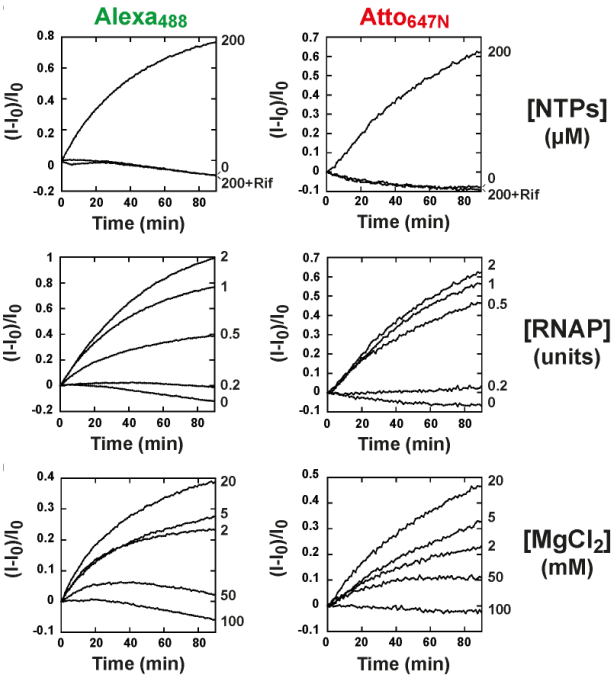


Figure 45 : Détection de la transcription de la matrice *λtR1*$_{C/T}$ à l'aide des sondes MBP_C et MBP_T suivant le principe décrit figure 38. Sauf indiqué différemment, les réactions contenaient 1U d'ARNP (28 nM), 20 mM de $MgCl_2$ et 200 µM de chaque rNTP.

Des expériences de contrôle ont confirmé que les réponses fluorescentes des sondes MBP_T et MBP_C reflètent de manière adéquate et quantitative la production de transcrits « *runoff* ». Par exemple, la présence de rifampicine (Rif), un inhibiteur de l'ARNP, supprime complétement la réponse des sondes MBP_T et MBP_C (**Figure 45A**). De plus, il n'y a pas de changements détectables de fluorescence de la MBP lors de la transcription d'une matrice ADN dépourvue des séquences antiMBP$_C$ et antiMBP$_T$. (Matrice *λtR1*$_0$, **Figure 47**). Enfin, une corrélation entre la magnitude des réponses de fluorescence MBP_T et MBP_C et la concentration de l'ARNP (**Figure 45B**) confirme également le suivi fidèle de la réaction de transcription en condition de cycles multiples.

La réponse des sondes MBP_T et MBP_C est affectée par la concentration des ions Mg^{2+} présents dans la réaction, avec une stimulation maximale observée pour les deux sondes à 20 mM de $MgCl_2$ (**Figure 45C**). Cette concentration reflète probablement un bon compromis

entre les effets antagonistes des ions Mg^{2+} sur l'efficacité de la transcription (Artsimovitch and Henkin, 2009) et sur la stabilité structurale des sondes MBP (en tige-boucle lorsqu'elles ne sont pas appariées) et des doubles hélices qu'elles forment avec les transcrits.

J'ai ensuite comparé les réponses des sondes MBP_T et MBP_C lors de réactions de transcription effectuées avec ou sans Rho. J'ai observé que la réponse de la sonde amont MBP_C (Alexa488) n'est pas affectée par la présence de Rho alors que celle de la sonde aval MBP_T (Atto647N) diminue significativement (**Figure 46A**, comparer courbes noires et rouges). Cet effet, qui peut être aboli par la présence de bicyclomycine [BCM] (**Figure 46A**, courbes bleues), un inhibiteur de Rho, concorde avec la formation de transcrits « *runoff* » en plus faibles proportions lorsque le facteur de terminaison est présent dans la réaction.

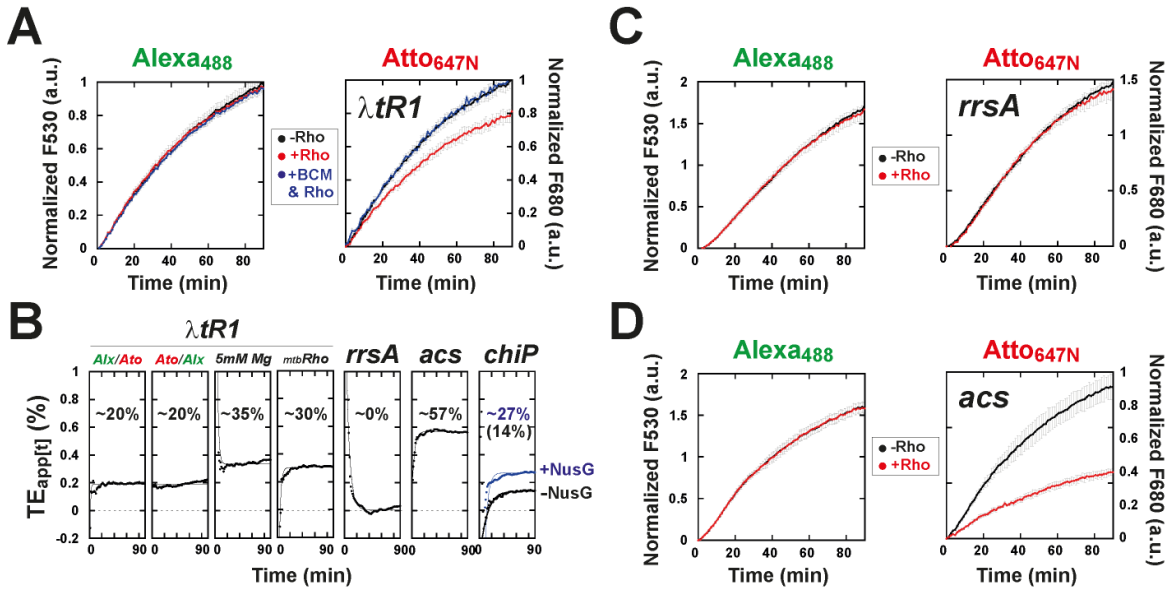


Figure 46 : Réponses des sondes MBP_C et MBP_T en fonction de la présence de Rho dans la réaction de transcription (schéma de principe décrit Figure 44). Les traces représentent les moyennes de triplicats expérimentaux avec les SD indiquées. **(A)** Intensités de fluorescence normalisées des sondes MBP_C (Alexa488) et MBP_T (Atto647N) mesurées lors d'une expérience de transcription de la matrice $\lambda tR1_{C/T}$ (0 ou 20 nM Rho). **(B)** Courbes d'efficacité apparente de la terminaison de la transcription en fonction du temps. Ato/Alx : expérience avec la matrice $\lambda tR1_{T/C}$ dans laquelle les positions des séquences antiMBP$_C$ et antiMBP$_T$ sont permutées ; *mtbRho* : expérience avec le facteur Rho de *M. tuberculosis*. Les traces de fluorescence MBP normalisées mesurées pour les matrices $rrsA_{C/T}$ **(C)** et $acs_{C/T}$ **(D)** sont également présentées.

L'efficacité de terminaison apparente (TE_{app}) déduite des réponses des sondes MBP (voir méthodes) avec la matrice $\lambda tR1_{C/T}$ atteint un plateau d'environ 20 % après quelques minutes d'incubation (**Figure 46B**, courbe « Alx/Ato »). Cette valeur correspond probablement à l'efficacité de terminaison à l'état stationnaire dans les conditions de transcription multi-cycles du test. Une TE_{app} similaire a également été mesurée avec une

matrice ADN dérivée de *λtR1* dans laquelle les positions des séquences antiMBP$_C$ et antiMPB$_T$ ont été permutées (**Figure 46B**, courbe « Ato/Alx »). Cette TE$_{app}$ est inférieure à l'efficacité de ≈65 % déterminée pour le terminateur *λtR1* avec un essai de terminaison classique (D'Heygere et al., 2015). Cette différence est probablement due à des conditions réactionnelles différentes, notamment des concentrations plus faibles de rNTP et de $MgCl_2$, ainsi qu'un régime réactionnel en cycle unique (« *single-run* ») pour l'essai classique (D'Heygere et al., 2015). En accord avec cette hypothèse, j'ai observé que la diminution de la concentration de $MgCl_2$ de 20 à 5 mM dans l'essai fluorogénique permettait de quasiment doubler la TE$_{app}$ (**Figure 46B)**. De plus, des TE$_{app}$ comparables ont été obtenues avec le facteur Rho de *M. tuberculosis* (**Figure 46B**), ce qui est cohérent avec les données publiées (D'Heygere et al., 2015) et suggère que l'essai fluorogénique peut être utilisé pour sonder l'activité de facteurs Rho distincts.

Pour tester davantage la robustesse de l'essai fluorogénique, j'ai effectué des expériences avec deux matrices ADN dans lesquelles la séquence originale *λtR1* (**Figure 44**) a été remplacée par une autre séquence capable d'éliciter une terminaison forte (provenant de la région 5'UTR du gène *acs* d'*E. coli*) ou, au contraire, par une séquence résistant à la terminaison (issue du gène ribosomal *rrsA* d'*E. Coli*) (Nadiras et al., 2018a). Les réponses de fluorescence MBP normalisées enregistrées avec et sans Rho n'ont révélé aucune différence significative pour la matrice ADN contenant la séquence *rrsA* sans terminateur mais un fort effet Rho-dépendant pour la matrice contenant le terminateur *acs* (**Figure 46B-C &D**). Ces observations sont conformes aux analyses de terminaison menées pour les séquences *rrsA* et *acs* avec un essai classique (Nadiras et al., 2018a). Nous avons également testé une matrice ADN contenant un terminateur Rho-dépendant faible issu du gène *chiP* de *Salmonella* (Bossi et al., 2012). Ce terminateur est caractérisé par une faible TE$_{app}$ (14 %) qui, cependant, double (≈27 %) en présence du cofacteur NusG. Ce résultat est également cohérent avec les données publiées montrant que le terminateur *chiP* (Bossi et al., 2012) et d'autres signaux Rho-dépendants sous-optimaux (Peters et al., 2012; Shashni et al., 2014) sont stimulés par NusG. Prises ensemble, mes données démontrent l'intérêt de l'essai fluorogénique pour suivre la terminaison de la transcription Rho-dépendante.

2) Optimisation de l'essai par utilisation simultanée d'une matrice « témoin »

Une limitation de notre essai fluorogénique est sa gamme dynamique relativement modeste dans des conditions qui seraient à la fois économiques (faible(s) concentration(s) en enzyme(s), par exemple) et optimales pour la terminaison (5 mM $MgCl_2$, par exemple) (**Figure 45C**). Cette limitation découle, au moins en partie, de la présence de la séquence antiMBP$_C$ en amont de la séquence testée (par exemple *λtR1*) dans la matrice d'ADN qui réduit significativement le rendement de transcription (**Figure 47** ; comparer les matrices *λtR1*$_{C/T}$ et *λtR1*$_T$). Ce rendement peut être amélioré par la permutation des séquences antiMBP$_C$ et antiMBP$_T$ dans la matrice ADN (**Figure 47**, matrice *λtR1*$_{T/C}$), probablement parce qu'aucune des deux séquences antiMBP n'est idéale pour initier la transcription. On peut noter également que la présence d'une séquence antiMBP dans la partie amont des transcrits risque de perturber artificiellement le repliement co-transcriptionnel de la séquence « testée » et/ou sa reconnaissance par Rho.

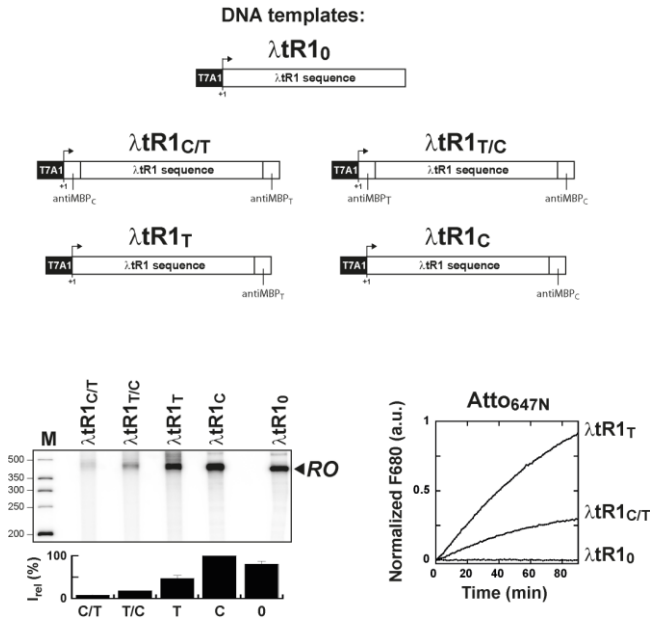


Figure 47 : Effet de la séquence antiMBP$_C$ en amont sur l'efficacité de transcription. Les expériences ont été analysées par PAGE (essai classique, gel à gauche) ou par lecteur de microplaques (essai fluorogénique, à droite) en présence de la sonde MBP$_T$.

Pour pallier à ces limitations, j'ai développé une variante de l'essai fluorogénique dans lequel la matrice ADN « testée » ne contient que la séquence antiMBP$_T$ en aval (matrice « test ») tandis qu'une seconde matrice ADN dépourvue de terminateur Rho-dépendant (« matrice témoin ») et contenant la séquence antiMBP$_C$ à son extrémité « aval » est utilisée pour fournir une référence co-transcriptionnelle (**Figure 48A**). Avec cette configuration, la séquence antiMBP$_T$ (ou antiMBP$_C$) est transcrite en dernier (**Figure 48A**) et, par conséquent, ne devrait pas affecter la synthèse ou le repliement de la séquence d'ARN testée (ou de référence) et sa sensibilité à Rho. Dans cette configuration de transcription en

« double matrice », la sonde MBP_T est utilisée pour quantifier la transcription de la séquence « test » tandis que la sonde MBP_C fournit un signal « témoin » pour la normalisation (**Figure 48A**).

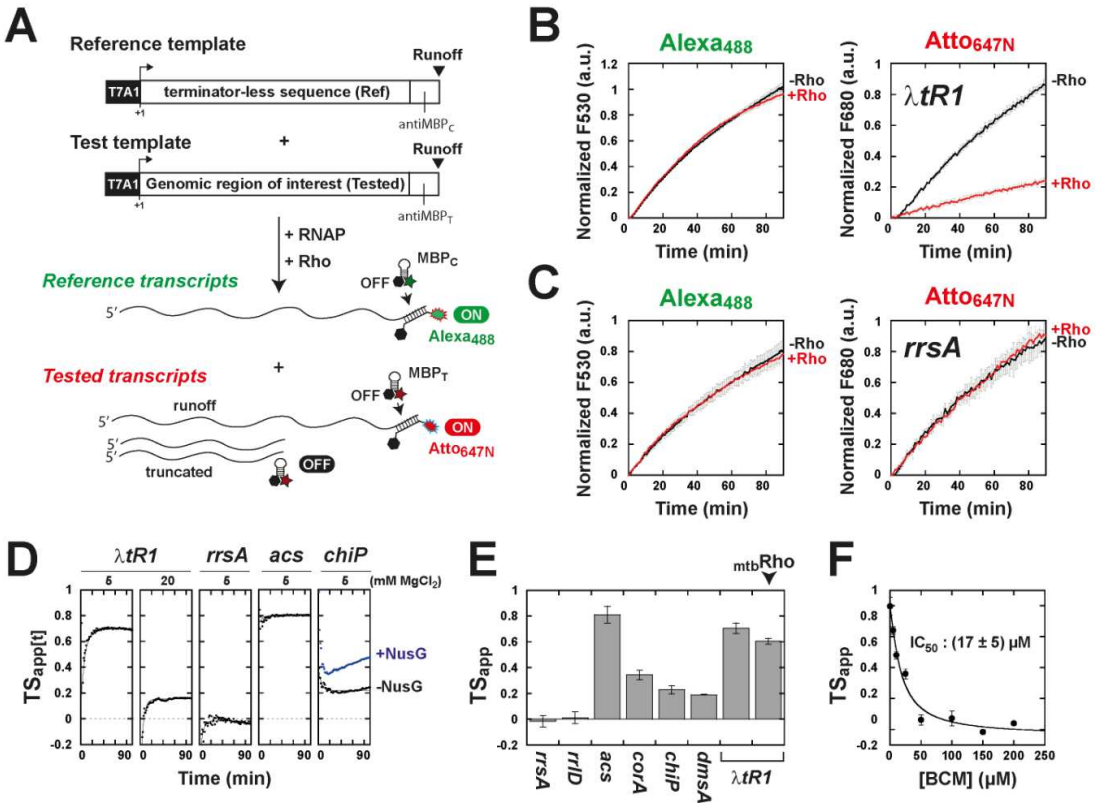


Figure 48 : Essai fluorogénique « à deux matrices » de la terminaison de la transcription Rho-dépendante. Les réactions contenaient 200 µM de rNTP, 1 U d'ARNP, 5 mM de $MgCl_2$ et 0 ou 70 nM de Rho. **(A)** Principe de l'essai. La matrice $rrsA_C$ contenant la séquence *rrsA* résistante à la terminaison (Nadiras et al., 2018a) est utilisée comme référence. Intensités de fluorescence normalisées des sondes MBP_C (Alexa488) et MBP_T (Atto647N) mesurées avec la matrice « test » $\lambda tR1_T$ **(B)** ou $rrsA_T$ **(C).** Les traces représentent les moyennes des triplicats expérimentaux avec les SD indiquées. **(D-E)** Courbes et valeurs de la « force » de terminaison apparente (TS_{app}) mesurée pour différentes matrices « test » et différentes conditions expérimentales. Leurs plateau ajustées (± barres SD) ont été obtenues pour divers matrices test. **(F)** Les valeurs de TS_{app} sont relativement bien corrélées à l'efficacité de terminaison mesurée avec un essai classique de référence ($TE_{classic}$).

J'ai d'abord testé cette configuration alternative avec une matrice « témoin » contenant la séquence résistante à la terminaison *rrsA* (matrice $rrsA_C$) et une matrice « test » contenant la séquence de terminaison *λtR1* (matrice $\lambda tR1_T$). Les deux matrices ont été mélangées et transcrites ensemble en présence des sondes MBP_C et MBP_T, dans un tampon contenant 5 mM de $MgCl_2$ avec ou sans Rho. J'ai observé une augmentation de la fluorescence Alexa488 (sonde MBP_C) en fonction du temps qui n'est quasiment pas affectée par Rho (**Figure 48B**), ce qui est cohérent avec la formation de transcrits $rrsA_C$ résistants à la terminaison (Nadiras et al., 2018a). En revanche, l'augmentation en fonction du temps de la

fluorescence Atto647N (sonde MBP_T), qui mesure la formation de transcrits « *runoff* » $\lambda tR1_T$ (**Figure 48A**), est beaucoup plus faible en présence qu'en absence de Rho (**Figure 48B**). L'ampleur de cet effet Rho-dépendant varie avec la concentration de $MgCl_2$ (**Figure 48D**) dans des proportions apparemment plus élevées qu'avec le test fluorogénique original (**Figure 45C**). On peut noter qu'avec la séquence *rrsA* à la place de *λtR1* (*rrsA* est alors présente dans les deux matrices « témoin » et « test »), il n'y a pas d'effet de Rho sur la réponse de fluorescence Atto647N (**Figure 48C**).

Une estimation directe de l'efficacité de la terminaison n'est pas possible avec la configuration « à deux matrices » car les efficacités d'initiation de la transcription pour chaque matrice ne sont pas directement mesurables (les matrices ne contenant pas de séquences antiMBP en amont) (**Figure 48A**). Nous avons donc défini une grandeur mesurable alternative, la différence normalisée entre les signaux Atto647N mesurés sans et avec Rho, ci-après dénommée TS_{app} (voir méthodes), pour comparer les forces des différents signaux de terminaison. Cette variable devient à peu près constante avec le temps, une fois que l'état stationnaire de la réaction multi-cycles est atteint (**Figure 48D**, courbes noires). Il existe toutefois une exception notable (TS_{app} augmentant avec le temps) lorsque NusG est également présent dans la réaction (**Figure 48D**, courbe bleue). Ainsi, bien que la stimulation du terminateur *ChiP* par NusG soit correctement détectée avec l'essai (**Figure 48D**), l'amplitude de cette stimulation ne peut pas être estimée avec confiance.

J'ai utilisé l'essai « à deux matrices » pour évaluer les capacités de diverses séquences génomiques à provoquer une terminaison Rho-dépendante. Les valeurs de TS_{app} sont élevées pour les séquences *λtR1* et *acs* (0.7-0.8), intermédiaires pour les séquences *dmsA*, *chiP* et *corA* (0.2-0.4) et proches de zéro pour les séquences *rrsA* et *rrlD* résistantes à la terminaison (**Figure 48E**). De façon remarquable, cet ordre d'efficacité TS_{app} varie linéairement avec l'efficacité de terminaison réelle ($TE_{classic}$) mesurée à l'aide d'un essai classique de transcription « *single-run* » **Figure 48F**), suggérant que l'essai « à deux matrices » jauge quantitativement la terminaison Rho-dépendante. Pour tester cette hypothèse, j'ai effectué des expériences avec la matrice *acs* qui contient un terminateur Rho-dépendant fort (Nadiras et al., 2018a) (**Figure 49**), en présence de quantités croissantes de bicyclomycine [BCM] qui inhibe Rho sans affecter la machinerie de transcription. J'ai observé

une diminution de la valeur de TS_{app} en fonction de la concentration en BCM, résultant en une IC_{50} apparente de 17 (±5) µM (**Figure 49**).

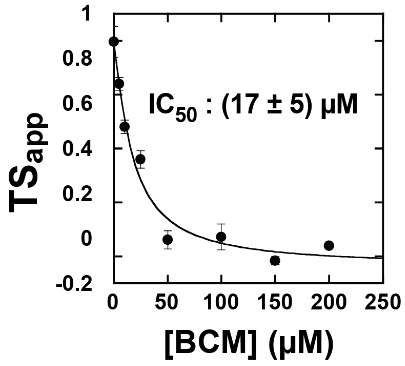


Figure 49 : Effet de la bicyclomycine [BCM] sur la force apparente de terminaison (TS_{app}) mesurée avec l'essai fluorogénique « à deux matrices ».

Cette valeur s'accorde bien avec les mesures antérieures du potentiel inhibiteur de la bicyclomycine (IC_{50} ≈25 µM) réalisées par des essais distincts d'ATPase (Kalarickal et al., 2010 717) ou de terminaison (D'Heygere et al., 2015). Prises ensemble, ces données confirment que la configuration « à deux matrices » suit la terminaison Rho-dépendante de manière appropriée et peut s'avérer utile pour la caractérisation de nouveaux terminateurs Rho-dépendants ou de nouveaux inhibiteurs de Rho.

3) Importance des sondes fluorescentes

Des expériences préliminaires avaient été menées avec les sondes MBP développées par Lafontaine et ses collaborateurs (Chinnappan et al., 2013 1912) dans lesquelles les fluorochromes sont, respectivement, Cy3 et Cy5 au lieu d'Alexa488 et d'Atto647N. Cependant, les réponses de ces sondes n'étaient pas toujours compatibles avec les effets Rho-dépendants observés avec un test classique de terminaison de la transcription. Par exemple, des expériences menées avec une matrice « test » portant la séquence *dmsA* et la matrice « témoin » *rrsA* (essai fluorogénique « à deux matrices ») ont révélé une augmentation de la fluorescence Cy5 en présence de Rho au lieu de la diminution attendue (**Figure 50A-B**). Cet effet stimulateur n'est pas compatible avec la formation de produits de terminaison, plus courts, induite par Rho au détriment de la formation de transcrits « *runoff* » (**Figure 50C**). Nous pensons que ce résultat inattendu pourrait être lié à un effet PIFE (*Protein induced fluorescence enhancement*) de Rho. L'effet PIFE est lié à la photo-isomérisation *cis-trans* des chromophores à architecture « cyanine » qui peut être perturbé par la présence de protéines associées à proximité du chromophore et conduire à une

augmentation de l'intensité de fluorescence de ce dernier ((Hwang and Myong, 2014) et références incluses).

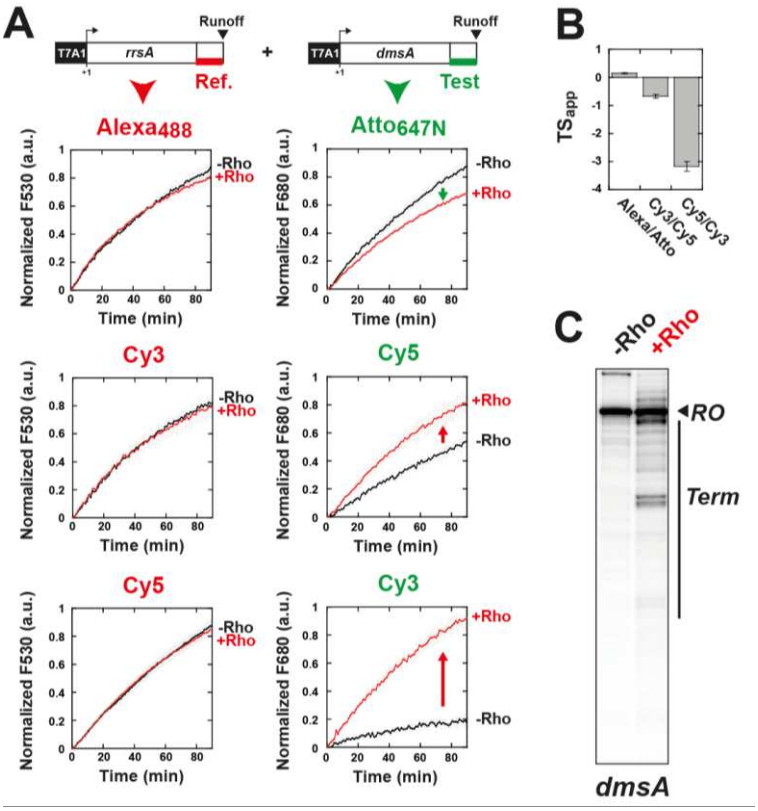


Figure 50 : Importance des fluorophores portés par les sondes moléculaires. Une estimation incorrecte de la terminaison Rho-dépendante avec les colorants Cy3 et Cy5 est probablement due à un effet PIFE (un effet fluorescent induit par une protéine) indésirable déclenché par Rho. Le PIFE est un problème commun et documenté avec les fluorophores cyanine.

4) Conclusion

Les travaux présentés ci-dessus démontrent pour la première fois qu'il est possible d'utiliser une détection fluorogénique pour suivre et caractériser la terminaison Rho-dépendante. Cette modalité de détection permet un suivi en temps réel et l'utilisation de quantités réduites de réactifs sans recours au radioisotope ^{32}P présent (et nécessaire pour la détection) dans les essais classiques (Artsimovitch and Henkin, 2009). Elle offre donc une alternative plus facile (et moins dangereuse) à mettre en œuvre et potentiellement automatisable, compatible avec un crible de moyen ou haut débit. Un tel crible pourrait être utilisé pour caractériser plus rapidement la réponse « terminaison » de nombreuses matrices ADN afin d'optimiser notre modèle prédictif de la terminaison Rho-dépendante (cf : **Article I**,

page 69) ou bien pour rechercher de nouveaux inhibiteurs de Rho (et donc de nouveaux antibiotiques) au sein de chimiothèques.

Nos essais fluorogéniques à une ou deux matrices offrent de nouvelles possibilités de caractérisation simple et rapide de la terminaison Rho-dépendante mais présentent également quelques limitations. Tout d'abord, ils ne semblent pas toujours répondre de façon adéquate à une complexification du milieu réactionnel, (due à l'ajout de cofacteurs comme NusG, par exemple). De plus, ils ne permettent pas d'évaluer la position exacte des sites de terminaison (sites *tsp*) au sein des matrices ADN qui requière une caractérisation de la longueur des transcrits produits par la terminaison.

5) Matériels et Méthodes

Matériels. Sauf indication contraire, les produits chimiques et les enzymes proviennent, respectivement, de Sigma-Aldrich et New England Biolabs. La bicyclomycine a été obtenue auprès de Santa Cruz Biotechnology. Les nucléosides triphosphates et les radionucléotides ont été achetés respectivement auprès de GE-Healthcare et PerkinElmer. Les oligonucléotides, y compris les sondes MBP, ont été achetés chez Eurogentec. Les facteurs Rho d'*E. coli* (*EcRho*) et de *M. tuberculosis* (*MtbRho*) et la protéine NusG d'*E. Coli* ont été préparés et purifiés comme décrit précédemment (Artsimovitch and Landick, 2000; Boudvillain et al., 2010b; D'Heygere et al., 2015). Les concentrations d'*EcRho* et de *MtbRho* sont exprimées en hexamères. Les matrices d'ADN utilisées dans les réactions de transcription *in vitro* ont été préparées par des procédures de PCR standard, comme décrit précédemment (Rabhi et al., 2011a). Les matrices d'ADN ont été purifiées avec le kit de purification GeneJET (Thermo Fisher Scientific). Leur composition est détaillée dans le **tableau S1** présenté en annexe (page 187).

Réactions « standard » de terminaison de la transcription in vitro. Les expériences ont été menées suivant un protocole décrit précédemment (Rabhi et al., 2011a). Brièvement, la matrice d'ADN (0,1 pmol), l'ARNP d'*E. Coli* (0,45 pmol), le facteur Rho (1,4 pmol) et de la SUPERase-In (0,5 U/µl ; Ambion) ont été mélangés dans 18 µl de tampon de transcription (40 mM Tris-HCl, pH 8,0, 50 mM KCl, 5 mM $MgCl_2$, 1,5 mM DTT) et incubés pendant 10 min à 37°C. Puis, 2 µl de mélange d'initiation (2 mM d'ATP, GTP et CTP, 0,2 mM UTP, 2,5 µCi/µL de

^{32}P-αUTP et 0 [*multi-run*] ou 250 [*single-run*] µg/mL de rifampicine dans le tampon de transcription) ont été ajoutés aux mélanges réactionnels qui ont été incubés pendant 20 min à 37°C. Les réactions de transcription ont été stoppées par ajout de 4 µL d'EDTA (0,5 M), 6 µL d'ARNt (0,25 mg/mL) et 80 µL d'acétate de sodium (0,42 M) avant précipitation à l'éthanol. Les culots de précipitation ont été dissous dans du tampon de charge dénaturant (95% formamide, 5 mM EDTA) et analysés par électrophorèse sur gel de polyacrylamide (7 %) dénaturant (PAGE).

Réactions de terminaison de la transcription in vitro avec marquage homogène des transcrits (« single-run chase transcription »). Pour ces expériences, les matrices ADN contiennent, directement en aval du promoteur, une séquence de 15-20 nt commençant par le dinucléotide AU (ou AG) et dépourvue de Cytosine (ou d'un des 3 autres nucléotides) dans le brin ADNnt (**Figure 51 et tableau S1** en annexe) (Artsimovitch and Henkin, 2009).

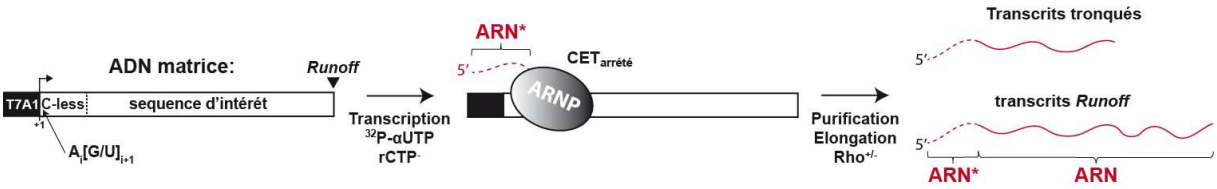


Figure 51 : Principe de l'essai « single-run chase transcription ».

La matrice d'ADN (0,6 pmol), l'ARNP d'E. Coli (0,51 pmol), et de la SUPERase-In (0,5 U/µl ; Ambion) ont été mélangés dans 6 µl de tampon de transcription (40 mM Tris-HCl, pH 8,0, 100 mM KCl, 5 mM MgCl2, 1 mM DTT, 0,05 mg/ml BSA) et pré-incubés pendant 5 min à 37°C. Puis, 1,5 µl de mélange d'initiation (5 µM rATP, GTP, 0,8 µM UTP, 10 µM de ApU (ou de ApG) (Iba) et 0,5 µCi/µL de ^{32}P-αUTP) ont été ajoutés aux mélanges réactionnels qui ont été incubés pendant 10 min à 37°C (**Figure 51**) (Artsimovitch and Henkin, 2009). L'ARNP peut ainsi transcrire la courte région « *C-less* » en incorporant le nucléotide marqué au ^{32}P avant d'être stoppé juste après la transition vers la phase d'élongation par le manque de CTP dans le milieu réactionnel. Le complexe d'élongation ainsi artificiellement arrêté (CET$_{arrêté}$) est ensuite purifié (pour éliminer l'excès de ^{32}P-αUTP) par gel-filtration sur une micro-colonne de G50 (équilibrée dans un tampon composé de 40 mM Tris-HCl, pH 8,0, 50 mM KCl, 5 mM MgCl2, 1 mM DTT, 0,025 mg/ml BSA). Pour chaque échantillon, une fraction du CET$_{arrêté}$ (≈0,24 pmol) a été mélangée avec 0 ou 1,4 pmol de facteur Rho et 0 ou 2,8 pmol de facteur NusG dans 6 µl de tampon de transcription (40 mM Tris-HCl, pH 8,0, 50 mM KCl, 5 mM

MgCl2, 1 mM DTT, 0,1 mg/ml BSA) avant d'être incubée pendant 10 min à 37°C. Puis, 4 µl du mélange de « chasse » (200 µM de chaque ribonucleotide rATP, rCTP, rGTP , rUTP et 25 µg/ml de rifampicine dans le tampon de transcription) ont été ajoutés avant incubation pendant 10 min à 37°C. Les réactions ont été stoppées par ajout de 4 µL d'EDTA (0,5 M), 6 µL d'ARNt (0,25 mg/mL) et 80 µL d'acétate de sodium (0,42 M) avant précipitation à l'éthanol. Les culots de précipitation ont été dissous dans du tampon de charge dénaturant (95 % formamide, 5mM EDTA) et analysés par électrophorèse sur gel de polyacrylamide (7 %) dénaturant.

Détection fluorogénique de la terminaison Rho-dépendante. Pour détecter simultanément l'initiation de la transcription et la formation de transcrits « *runoff* » (de taille maximale), des matrices ADN ont été conçues pour coder la séquence d'intérêt entourée des séquences antiMBP$_C$ et antiMBP$_T$ et située en aval du promoteur T7A1 fort (**Figure 44**). La séquence antiMBP$_C$ est complémentaire de la sonde MBP$_C$ (5'-Alexa488-CGCUUUUUUUUUUUGCG-Dabcyl-3') tandis que la séquence antiMBP$_T$ est complémentaire de la sonde MBP$_T$ (5'-Atto647N-CGCUUGUUAAUUUUUGCG-BHQ2-3'). Les deux sondes MBP$_C$ et MBP$_T$ sont constituées de résidus 2'-O-méthyle. Les réactions de transcription ont été réalisées dans des microplaques NBS-384 puits (Corning) en mélangeant la matrice d'ADN (0,2 pmol), l'ARNP d'*E. Coli* (1,8 pmol [1 U] sauf indication contraire), de la SUPERase-In (1 U/µL; Ambion), les sondes MBP$_T$ et MBP$_c$ (4 pmol chacune) et Rho (0 ou 2,8 pmol) dans 35 µL de tampon de transcription standard (40 mM Tris-HCl, pH 8,0, 50 mM KCl, 0,1 mM DTT) additionné de la quantité de $MgCl_2$ indiquée. Le suivi des réactions a été effectué avec un lecteur de microplaques « Synergy H1FD » (Biotek) équipé de filtres pour la détection simultanée des chromophores Alexa488 (ex : 485/20 nm, em : 530/25 nm) et Atto647N (ex : 620/40 nm, em : 680/30 nm). Après incubation des microplaques dans l'instrument (37°C, agitation à 280 rpm) pendant 10 min, les micro-injecteurs de l'instrument ont été utilisés pour injecter 5 µL de rNTP (à une dilution de 2 mM chacun dans le tampon de transcription) ou 5 µL de tampon de transcription seul (réactions de contrôle). La fluorescence des sondes Alexa488 et Atto647N a été suivie pendant 90 minutes à 37°C. Chaque expérience comprenait des triplicats d'échantillons avec NTP et Rho (+Rho), avec uniquement des NTP (-Rho) et sans NTP (-NTP). Les traces F_{530} et F_{680} normalisées ont été obtenues en soustrayant le bruit de fond « $(I-I_0)/I_0$ » (-NTP) des traces « $(I-I_0)/I_0$ » obtenues en présence de NTP (où I_0

et I représentent les intensités de fluorescence brute mesurées respectivement à 0 min et à l'instant t). Les efficacités de terminaison apparentes à l'instant t ont été déterminées avec l'équation suivante :

$$TE_{app[t]} = \left(1 - \frac{\left[\frac{F_{680+}}{F_{530+}}\right]}{\left[\frac{F_{680-}}{F_{530-}}\right]}\right) \times 100$$

Où $[F_{680+}/F_{530+}]$ et $[F_{680-}/F_{530-}]$ sont les rapports F_{680}/F_{530} mesurés en présence et en absence de Rho, respectivement. Les valeurs TE_{app} (en %) sont les valeurs de plateau obtenues par le « fit » des données $TE_{app[t]}$ avec une équation décrivant une croissance (ou décroissance, selon ce qui convient le mieux) exponentielle.

Dans la configuration expérimentale alternative « à deux matrices », les séquences $antiMBP_T$ et $antiMBP_C$ ont été introduites dans des matrices ADN distinctes « test » et « témoin » qui sont transcrites ensemble (**Figure 42A**). Dans ce cas, la matrice « test » ne contient que la séquence $antiMBP_T$ en aval de la séquence d'intérêt tandis que la matrice « témoin » ne contient que la séquence $antiMBP_C$ en aval d'une séquence de 500 pb résistant à la terminaison (provenant du gène *rrsA* d'*E.coli*) (Nadiras et al., 2018a). La transcription de la matrice « témoin » (quantifiée avec la sonde MBP_C) est utilisée pour normaliser la production de transcrits « *runoff* » issus de la matrice « test » (quantifiés avec la sonde MBP_T). Des stocks concentrés du mélange équimolaire des matrices « test » et « témoin » ont été préparés au préalable pour garantir que tous les réplicats expérimentaux contiennent exactement le même rapport d'ADN « test »/« temoin ». Les réactions de transcription ont été assemblées dans une microplaque NBS-384 puits comme décrit ci-dessus, en utilisant le mélange des matrices « test » et « témoin » (à 0,1 pmol chacune) au lieu d'une seule matrice d'ADN. Toutes les étapes ultérieures ont été effectuées comme décrit ci-dessus. Avec cette configuration « à double matrice », nous avons estimé la force des terminateurs à partir des variations d'intensité de fluorescence des MBP mesurées avec et sans Rho en utilisant l'équation suivante :

$$TS_{app[t]} = \frac{\left(\left[\frac{F_{680-}}{F_{530-}}\right] - \left[\frac{F_{680+}}{F_{530+}}\right]\right)}{\left[\frac{F_{680-}}{F_{530-}}\right]}$$

Où F_{530+}, F_{680+}, F_{530-} et F_{680-} sont les intensités de fluorescence normalisées des sondes Alexa488 (F_{530}) et Atto647N (F_{680}) mesurés en présence (+) ou en absence (-) de Rho. Les valeurs TS_{app} indiquées dans le texte sont les valeurs de plateau obtenues par le « *fit* » des points $TS_{app[t]}$ avec une équation de croissance/décroissance exponentielle.

III. Travaux non publiés II : Vers une utilisation de l'approche « CLIP-seq » pour la détection des sites *rut*

1) Formation de photo-pontages entre Rho et l'ARN *in vitro*

La stratégie Clip-seq repose sur l'analyse RNAseq des fragments du transcriptome qu'il est possible de photo-ponter *in vivo* par les UV-C à la protéine d'intérêt (**Figure 52**) :

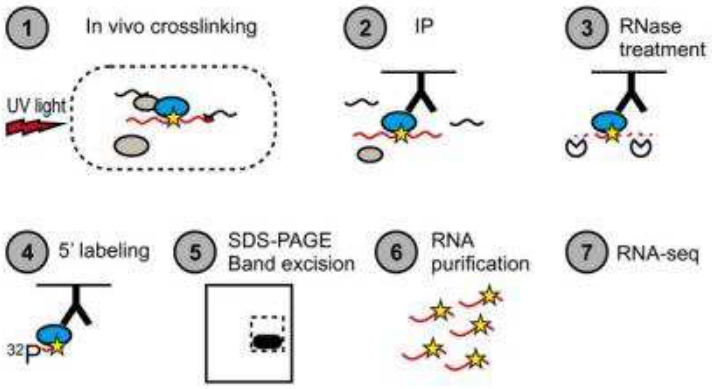


Figure 52 : principe général de l'approche CLIP-seq. Figure issue de (Holmqvist et al., 2016).

La formation d'un pontage covalent dépend généralement de la présence d'un acide aminé aromatique (Phénylalanine, Tyrosine, Tryptophane) dans la poche de reconnaissance ARN de la protéine, de telle sorte que toutes les protéines se liant à l'ARN ne sont pas forcément étudiables par Clip-seq (Cook et al., 2015). Par chance, la poche PBS de Rho contient deux résidus aromatiques (Phe64 et Tyr80) relativement bien conservés à proximité du dimère 5'YC associé (Skordalakes and Berger, 2003) (**Figure 53).**

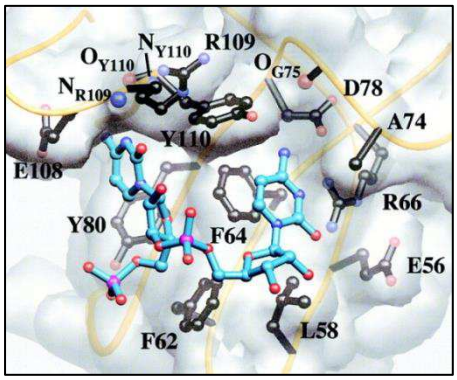


Figure 53 : Poche d'interaction PBS contenant le dinucléotide 5'-CC (en bleu). Figure issue de (Bogden et al., 1999).

Ceci suggère que les interactions entre Rho et les séquences *Rut* des transcrits devraient pouvoir être étudiées par Clip-seq. Des expériences préliminaires menées par Annie Schwartz au laboratoire ont confirmé qu'il était possible de photo-ponter Rho à un transcrit ARN contenant un terminateur (et donc un site *Rut*) connu (**Figure 54**). La quantité de transcrits photo-pontés *in vitro* dépend du facteur Rho (EcRho ou MtbRho) et de la dose d'irradiation (**Figure 54**).

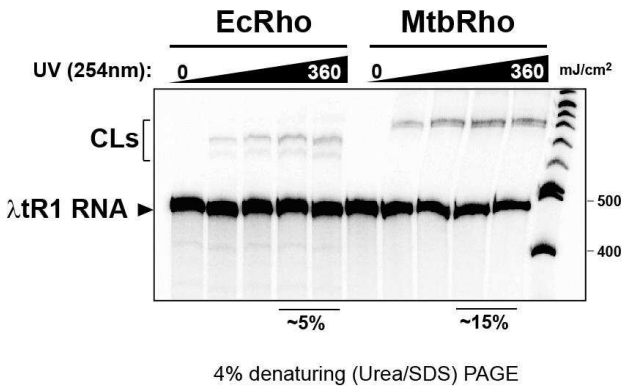


Figure 54 : Photo-pontage induit aux UV-C entre le facteur Rho d'*E. coli* (*EcRho*) ou de *M. tuberculosis* (*MtbRho*) et un ARN témoin contenant le terminateur *λtr1* (Annie Schwartz & Marc Boudvillain, résultats non publiés).

Cette quantité reste néanmoins modeste (~5% au maximum pour *EcRho*), en accord avec un mécanisme de photo-pontage protéine:ARN généralement peu efficace (Hockensmith et al., 1986; Wheeler et al., 2018). Ces expériences préliminaires nous ont encouragés à tenter d'adapter l'approche Clip-seq à l'étude de Rho *in vivo*.

2) Implémentation du Clip-seq adapté à Rho chez *Salmonella*

Un autre élément clé de l'approche Clip-seq est la purification des complexes protéine-ARN avant et après digestion (avec une endonucléase) des segments ARN non protégés par l'interaction avec la protéine (**Figure 52**, étapes 2 & 3). L'avantage procuré par le photo-pontage est que cette purification peut être faite par immunoprécipitation (IP) dans des conditions relativement stringentes qui permettent d'éliminer une grande partie des complexes protéine:ARN non spécifiques et/ou formés après lyse (Huppertz et al., 2014). Bien que nous disposons au laboratoire d'anticorps anti-Rho « à façon » (polyclonaux), ceux-ci ne sont pas suffisamment sélectifs pour une utilisation dans le Clip-seq (L. Bossi & N. Figueroa-Bossi, communication personnelle). Nous avons donc opté pour l'utilisation d'une étiquette peptidique « 3XFLAG » (DYKDHD-G-DYKDHD-I-DYKDDDDK) qui est souvent utilisée pour le Clip-seq (Holmqvist et al., 2016; Sundararaman et al., 2016) et pour laquelle il existe des anticorps commerciaux de bonne qualité.

L'introduction d'étiquettes en N-ter ou en C-ter de Rho n'est pas viable chez *E. coli* ou *Salmonella* mais notre consortium a identifié une position interne où cette introduction est tolérée (M Boudvillain, L. Bossi & N. Figueroa-Bossi, résultats non publiés). Nos

collaborateurs ont donc préparé une souche de *Salmonella* contenant un facteur Rho étiqueté « 3XFLAG » de cette façon (Rho_{3XFLAG}) tandis qu'une souche portant un facteur Rho modifié à la même position par une étiquette GFP (Rho_{GFP}) est également disponible comme « contrôle ».

Dans un premier temps, j'ai vérifié par Western Blot qu'il était possible de détecter la présence de Rho_{3XFLAG} dans un lysat cellulaire après immunoprécipitation avec l'anticorps anti-FLAG (**Figure 48B**). J'ai utilisé des conditions de lyse mécanique (Mellin et al., 2014) et une étape de dénaturation à l'urée (avant immunoprécipitation) pour diminuer les contaminations potentielles (voir méthodes). J'ai déterminé qu'un temps long d'incubation avec l'anticorps (12h à 4°C sous agitation) était nécessaire pour détecter un signal suffisant (**Figure 55A**).

Puis, j'ai évalué l'effet de la dose d'irradiation d'UV-C sur la quantité d'ARN photo-ponté qui est globalement récupéré après immunoprécipitation. Pour cela, j'ai déphosphorylé puis marqué au P^{32} les transcrits photo-pontés en réalisant ces traitements directement sur les complexes immunoprécipités (immobilisés sur billes magnétiques). Puis j'ai traité les complexes à la protéinase K pour libérer les transcrits que j'ai analysé par PAGE dénaturant. J'ai observé un effet dose-dépendant de la quantité d'ARN récupéré (**Figure 55B**), ce qui est consistant avec une purification adéquate des transcrits photo-pontés. Des doses de 400 à 800 mJ/cm², qui représentent un bon compromis entre quantités d'ARN récupérées et mort cellulaire engendrée (un paramètre qui a été évalué indépendamment par notre collaborateur L. Bossi), ont été utilisées pour les expériences ultérieures.

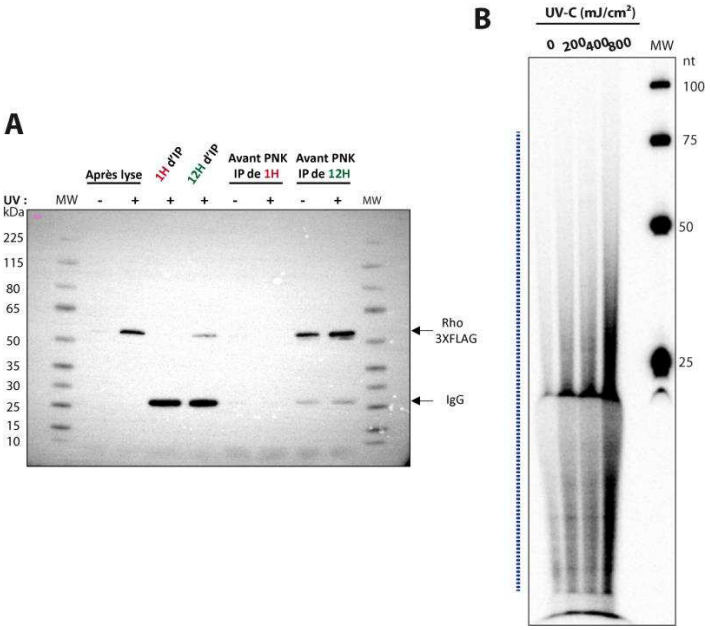


Figure 55 : Expériences d'optimisation préliminaires. (A) Importance de la durée d'immunoprécipitation. Western-Blot anti-FLAG (voir méthodes) des échantillons après lyse, immunoprécipitation, et déphosphorylation des transcrits au sein des complexes immunoprécipités. **(B)** Effet de la dose d'UV-C sur la quantité d'ARN photo-pontés détectés (10 % PAGE dénaturant). MW : marqueur de poids moléculaire.

Lors des expériences préliminaires, j'ai observé que les ARN récupérés étaient relativement de petites tailles (20-30 nt, pour la majorité) (**Figure 55B**) sans qu'il ne soit nécessaire de traiter les échantillons avec une nucléase (**Figure 52**, étape 3). On peut noter que le protocole Clip-seq a été développé initialement pour étudier des protéines eucaryotes ; il est probable que l'activité RNase endogène eucaryote n'est pas suffisante pour digérer les ARN photo-pontés correctement. Néanmoins, un traitement RNase exogène semble avoir été également nécessaire pour la préparation des échantillons dédiés à l'étude Clip-seq des protéines ProQ et Hfq chez *Salmonella* (Holmqvist et al., 2016). Nous nous sommes très largement inspirés de cette étude pour développer notre protocole mais il n'est toujours pas clair pourquoi l'activité RNase endogène est suffisante pour digérer les transcrits dans notre cas.

L'obtention de fragments ARN de relativement petites tailles (**Figure 55B**) suggère qu'ils n'ont été protégés que par un seul monomère de Rho, ce qui semble cohérent avec un faible rendement de photo-pontage (voir plus haut) mais également avec une dissociation de l'hexamère Rho avant que l'activité RNase n'ait été éliminée (ce qui devrait être le cas après les nombreuses étapes de lavage des complexes immunoprécipités sur billes magnétiques ; voir méthodes). Cette distribution de petites tailles peut éventuellement compliquer le « *mapping* » des fragments RNAseq sur le génome de référence (Murigneux et al., 2013), bien que ce facteur soit moins limitant pour les génomes bactériens que pour les « gros » génomes eucaryotes.

J'ai ensuite préparé des échantillons d'ARN photo-pontés à Rho issus de cellules irradiées à 400 ou 800 mJ/cm² et contenant Rho_{3XFLAG} (test) ou Rho_{GFP} (contrôle). Pour ce faire, j'ai radiomarqué les complexes au P^{32} (une fraction de chaque échantillon est en fait marquée à des fins de détection puis re-mélangé au reste de l'échantillon « froid ») et j'ai purifié ces derniers par migration sur un gel NuPAGE 4-12%. Ce système électrophorétique préserve l'intégrité des ARN, ce qui n'est pas le cas des gels SDS-PAGE classiques (Huppertz et al., 2014). Les complexes sont ensuite électro-transférés sur membrane de nitrocellulose ce qui constitue une étape essentielle de purification dans l'approche Clip-seq (Holmqvist et al., 2016). Dans notre cas, on peut observer sur la membrane une bande de taille apparente autour de 60 kDa qui est présente dans les échantillons Rho_{3XFLAG} mais absente dans le contrôle Rho_{GFP} (**Figure 56A-B**), suggérant qu'elle représente les complexes Rho:ARN photo-

pontés spécifiques. La population d'ARN correspondante a été récupérée par incubation de la région de la membrane correspondante avec de la protéinase K. L'analyse de cette population par PAGE dénaturant est cohérente avec les résultats présentés plus haut (ARN de petites tailles) (**Figure 56C**).

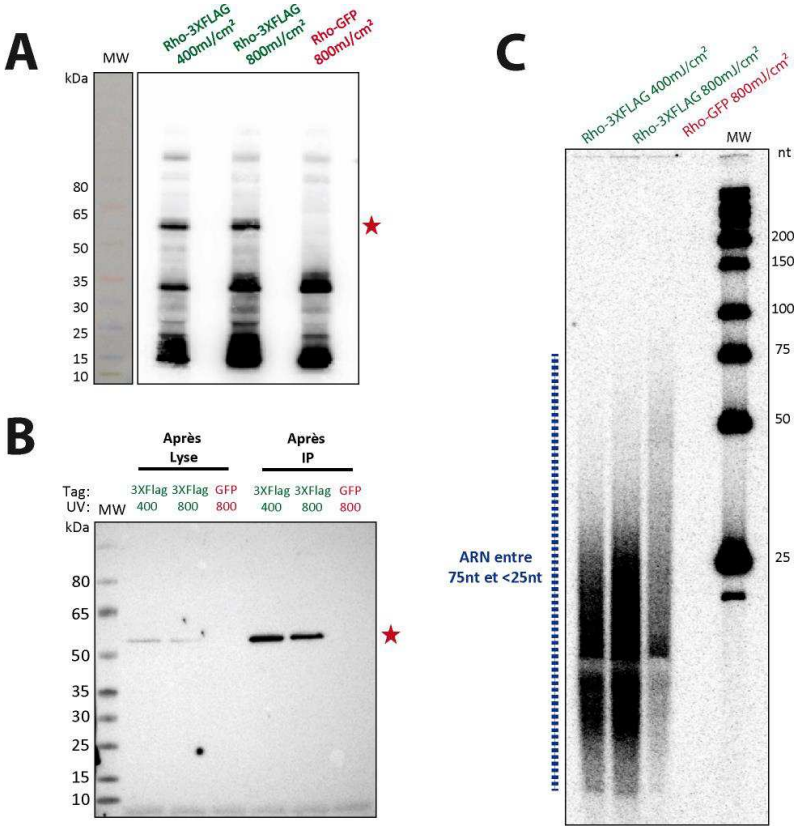


Figure 56 : Complexes Rho:ARN photo-pontés *in vivo*. (A) Profils obtenus (par phosphorimaging) après migration des échantillons sur gel NuPAGE 4-12% et transfert sur membrane de nitrocellulose. L'étoile rouge indique la population d'intérêt. **(B)** Western-Blot anti-FLAG des échantillons ARN:Rho$_{(3XFLAG\ ou\ GFP)}$ après lyse bactérienne et immunoprécipitation (IP). **(C)** Analyse par PAGE dénaturant (10%) des ARN récupérés après traitement à la protéinase K des complexes ARN:Rho$_{(3XFLAG\ ou\ GFP)}$ purifiés par NuPAGE et transfert sur membrane. MW : marqueur de poids moléculaire.

J'ai utilisé ce protocole pour préparer trois échantillons destinés au séquençage haut-débit (NGS) sur la plateforme IMAGIF (CNRs, Gif/Yvette). Bien que les quantités finales d'ARN que j'ai obtenues aient été très faibles (1-3 ng par échantillon) et sous-optimales pour le NGS, l'équipe d'IMAGIF a accepté de tenter de préparer les banques d'ADNc et de les séquencer. Malheureusement, le résultat n'a pas été concluant, la présence de contaminants vraisemblablement d'origine « microbiome » ayant été détectée en grande quantité (**Figure 57**). La quantité mais peut-être aussi la qualité d'ARN contenu dans mes

échantillons n'était vraisemblablement pas suffisante. Je propose quelques pistes pour résoudre ce problème dans la partie conclusions et perspectives.

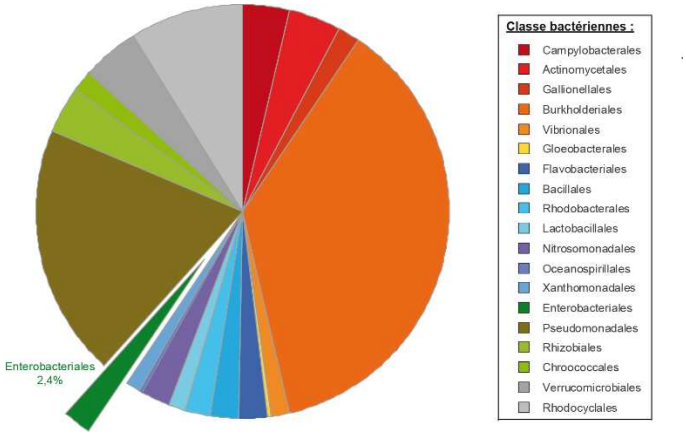


Figure 57 : Analyse métagénomique des données NGS obtenues à partir des échantillons CLIP-seq. La proportion de matériel provenant d'entérobactéries (dont *Salmonella*) est très faible (2,4%) indiquant une contamination importante des échantillons.

3) Matériels et Méthodes

Matériels. Sauf indication contraire, les produits chimiques et les enzymes ont été achetés, respectivement, chez Sigma-Aldrich et New England Biolabs. Les radionucléotides ont été achetés chez PerkinElmer. Les produits dédiés aux expériences d'électrophorèse et de transfert sur membrane ont été acheté chez Invitrogen. Les travaux ont été menés avec les souches *Salmonella typhimurium* SL1344 rho:3XFLAG ou rho:gfp en étroite collaboration avec l'équipe de Nara Figueroa-Bossi à l'I2BC (CNRS, Gif/Yvette) où j'ai réalisé une partie des expériences.

Préparation des échantillons. Les deux souches bactériennes rho:3XFLAG et rho:gfp ont été mises en culture dans du LB (*Luria Broth*) pendant une nuit à 37°C sous agitation. Puis 150 µl de chaque culture ont été étalés sur une boîte de Petri LB agar et incubés à 37°C durant 4h45. Le photo pontage aux UV-C est directement réalisé sur les tapis bactériens dans une enceinte Stratalinker réglée pour des irradiations de 0, 200, 400 ou 800 mJ/cm². Les bactéries traitées sont récupérées dans 3,5 ml de LB (pour avoir plus de matériel, on regroupe les bactéries de deux boites identiques). Après centrifugation à 13.000 g durant 15 min à 4°C, les culots sont stockés à -80°C.

Les culots bactériens congelés sont remis en suspension dans 500 µl de tampon de préparation froid (50 mM Tris-HCl, pH 7,4, 100 mM NaCl, 0,05 % Tween-20) contenant 25 µl d'un cocktail anti protéase (Solution stock : 1 tablette dissout dans 2 ml de tampon [10 mM Tris-HCl, pH 7,4] ;cOmplete-Roche), et 5 µl de SUPERase-In (Invitrogen, 100 U). Les suspensions bactériennes sont ensuite transférées dans des tubes à vis (Axygen, 2 ml) contenant 0.5 g de billes de verre (Sigma-Aldrich ; 0,1 mm de rayon). La lyse des bactéries est effectuée dans un instrument « Fast Prep Apparus » (vitesse : 6,5) en 7 cycles d'une minute interrompus par des périodes d'incubation sur glace d'une minute (Mellin et al., 2014). Après ajout de 300 µl de tampon de préparation froid, les échantillons sont centrifugés deux fois (15 min, à 13.000 g et 4° C) pour récupérer les surnageants dépourvus de billes de verre ou de débris cellulaires.

Les lysats clarifiés sont mélangés avec 1 volume de tampon de préparation contenant 8 M d'urée. Cette étape permet de limiter l'immunoprécipitation des protéines liées à la protéine d'intérêt (Huppertz et al., 2014). Après 5 minutes d'incubation à 65°C sous agitation (900 rpm dans un ThermoMixer Eppendorf), les échantillons sont dilués avec 9 équivalents-volume de tampon de préparation froid (contenant 25 µl de cocktail cOmplete-Roche, et 5 µl de SUPERase-In). Les billes magnétiques anti-FLAG® M2 (Sigma-Aldrich), préalablement rincées avec du tampon de préparation froid, sont ajoutées aux échantillons (30 µl de suspension à 50% de billes par échantillon). L'immunoprécipitation est obtenue par incubation dans un incubateur rotatif durant 12h à 4°C suivie d'une courte centrifugation (5 min à 800 g et 4°C). Un portoir magnétique est utilisé pour éliminer le surnageant et conserver les billes magnétiques qui sont resuspendues dans 900 µl de tampon de préparation froid. Pour limiter les interactions aspécifiques, les billes sont rincées quatre fois avec 900 µl de tampon riche en sel (50 mM Tris-HCl, pH 7,4, 1 M NaCl, 1 mM EDTA, 0,05% Tween20) puis deux fois avec un tampon CIP 1X (50 M Tris-HCl, 0,1 mM EDTA, pH 8,5). Les ARN retenus sur les billes sont déphosphorylés par incubation des billes dans 100 µL de tampon CIP 1X en présence de 10 U de phosphatase alcaline d'intestin de veau (CIP, Roche) et de 0,5 µl de SUPERase-In durant 30 min à 37°C et 800 rpm. Puis les billes sont rincées deux fois avec du tampon riche en sel pour éliminer la CIP. Une fraction des billes (~10%) est rincée avec du tampon PNK 1X (70 mM Tris-HCl, 10 mM MgCl2, 5 mM DTT, pH 7,6) puis re-suspendue dans 40 µL de tampon PNK 1X contenant 0.4 µl de phosphatase alcaline (T4 PNK)

et 0.4 µCi de γ^{32}P-ATP et incubée pendant 30 min à 1100 rpm afin de radio-marquer les ARN correspondants. Les billes sont ensuite rincées deux fois avec du tampon fort en sel puis quatre fois avec du tampon de préparation pour éliminer le γ^{32}P-ATP libre.

Purification des populations d'intérêt de complexes Rho:ARN. Les différents échantillons (marqués au ^{32}P ou non) sont repris dans un mélange de 16 µl de tampon de charge LDS NuPAGE 1X (26,5 mM Tris-HCl, 35,25 mM Tris Base, 0,5% LDS, 2,5% Glycerol, 127 µM EDTA, 55 µM SERVA Blue G250, 44 µM Phenol Red, pH 8,5) et de 2 µl d'antioxydant NuPAGE (10X) puis incubés 10 min à 70°C. Les surnageants dépourvu de billes sont collectés sur portoir magnétique puis analysés par électrophorèse sur un gel NuPAGE 4-12% (migration à 180 V durant 1h dans du tampon MOPS 1X ([50 mM MOPS, 50 mM Tris base, 3,5 mM EDTA, 1 mM EDTA, pH 7,7] contenant 1:2000 [v:v] d'antioxydant NuPAGE) dans un système XCell *SureLock*® Mini-Cell (Thermo Fisher Scientific). Les fractions « marquée » (10 %) et « non-marquée » (90 %) des échantillons sont déposés côte à côte sur le gel (**Figure 58**), un aliquot de marqueur de taille coloré (« *Spectra Multicolor Broad Range Protein* », Thermo Scientific) étant également déposé. Après migration, les espèces protéiques sont transférées sur une membrane de nitrocellulose, soit pour purifier les ARN d'intérêt, soit pour leur immunodetection (**Figure 56**, page 147).

Dans le premier cas, des feuilles de papiers filtre et la membrane de nitrocellulose sont baignées dans 100 ml de tampon A (89,5 mL de tampon de transfert NuPAGE 1X, 500 µl NuPAGE antioxydant, 10 ml de méthanol) tandis que le gel NuPAGE est baigné dans 100 ml de tampon B (99,5 mL tampon de transfert NuPAGE 1X [25 mM Bicine, 25 mM Bis-Tris, 1 mM EDTA, pH 7,2], 500 µl NuPAGE antioxydant) durant 10 min. L'ensemble est assemblé à l'horizontale dans un instrument d'électro-transfert « Trans-Blot Semi-Dry » (Bio-Rad) en superposant deux papiers filtres, la membrane, le gel et deux papiers filtres. Après électro-transfert à 15 V pendant 30 min, la membrane est séchée à l'air libre. Des gouttes d'une solution diluée de γ^{32}P-ATP sont déposées sur la membrane aux niveaux du marqueur de taille coloré pour servir de repères. Puis, la membrane est enveloppée dans du Saran et placée dans une cassette contenant un écran de phosphorimaging pendant une nuit. Après phosphorimaging de l'écran avec un Typhon FLA9500 (GE-Healthcare), une impression à taille réelle est utilisée comme guide pour découper les régions d'intérêt de la membrane (**Figure 57**).

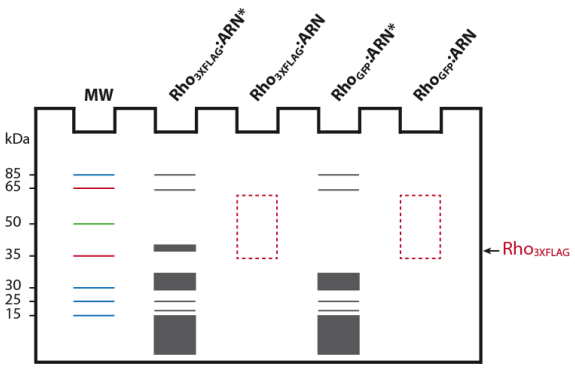


Figure 58 : Schéma de principe des dépôts d'échantillons sur gel NuPAGE. Les zones en pointillés rouges sont celles à découper (après transfert sur membrane) pour récupérer les ARN d'intérêts.

Les fragments découpés sont incubés dans 160 µl de tampon PK (100 mM Tris-HCl, pH 7,4, 50 mM NaCl, 10 mM EDTA) supplémentés de 0,5 µl de SUPERase-In (10 U) et de 40 µl de protéinase K (Thermo Scientific, 20 µg/µl) pendant 30 min à 37°C et sous agitation (1100 rpm). Puis, 200 µl de tampon PK contenant 9 M d'urée sont ajoutés avant une nouvelle incubation de 30 min dans les mêmes conditions. Les fragments de membrane sont éliminés avant ajout de 40 µl d'une solution d'acétate de sodium (3 M, pH 5,5). Les protéines de chaque échantillon sont éliminées par extraction avec 1 équivalent-volume d'un mélange phénol:chloroforme:alcool isoamylique [25 :24 :1] (pH 6,7) suivie de deux extractions avec 1 équivalent-volume de chloroforme. Les phases aqueuses sont transférées dans des tubes LoBind (Eppendorf) puis mélangées avec 1 µl d'entraineur GlycoBlue (Invitrogen ; 20 µg/µl de glycogène) et 3 équivalents-volume d'éthanol. Après précipitation pendant une nuit à 20°C, les culots sont repris dans 10 µl d'$H2O_{miliQ}$ et leurs concentrations en ARN sont estimées par fluorescence avec un kit dédié (Quant-IT RiboGreen) et un spectrofluorimètre NanoDrop 3300 (ThermoFisher Scientific). Les échantillons non-marqués au ^{32}P ont été envoyés à la plateforme IMAGIF (I2BC ; CNRS, Gif/Yvette) pour séquençage à Haut Débit de type « 75 paired end » avec un instrument Illumina NextSeq 500 (les banques d'ADNc ont été préparées par le personnel de la plateforme suivant un protocole qui ne nous a pas été communiqué). L'analyse initiale des résultats a été effectuée par Eric Eveno (CBM Orléans) avec l'aide de Daniel Gautheret (I2BC, Gif/Yvette).

Expériences de Western-Blot. Après séparation par électrophorèse sur gel NuPAGE et électro-transfert sur membrane de nitrocellulose, la présence de la protéine d'intérêt (Rho_{3XFLAG}) peut être contrôlée par immunodétection. Pour cela, la membrane est incubée pendant une nuit et sous agitation dans un bain de tampon TBS 1X (20 mM Tris-HCl, pH 7,5, 150 mM NaCl) contenant 5% (p/v) de lait écrémé (pour saturer les sites aspécifiques). La membrane est ensuite lavée trois fois par agitation pendant 5 minutes à température ambiante dans 30 ml de TBS 1X. Puis, elle est incubée dans les mêmes conditions pendant 2h

dans une solution contenant l'anticorps primaire anti-FLAG (dilué au 1/10 000 dans du tampon TBS 1X contenant 1% de lait écrémé et 0.1 % de Tween-20). Après trois rinçages de 5 min dans des bains de tampon TBS 1X contenant 0.1 % de Tween-20, la membrane est incubée pendant 1h avec l'anticorps secondaire couplé à la peroxydase dirigée contre les immunoglobines de l'anticorps primaire (anti-IgG de souris ; anticorps secondaire dilué au 1/10 000 dans du tampon TBS 1X contenant 1 % de lait écrémé et 0.1 % de Tween-20). La membrane est de nouveau rincée trois fois dans des bains de tampon TBS 1X contenant 0.1 % de Tween-20. L'activité de la peroxydase couplée à l'anticorps secondaire est révélée par chimioluminescence après incubation de la membrane avec un substrat chromogène spécifique (SuperSignal West Dura Extended Duration Substrate) suivant les recommandations du fournisseur (Thermo Fisher Scientific). La chimioluminescence est détectée avec un imageur PXi (Singene).

IV. Article II:

Evaluating the Effect of Small RNAs and Associated Chaperones on Rho-Dependent Termination of Transcription In Vitro

Methods in Molecular Biology (2018)

Chapter 7

Evaluating the Effect of Small RNAs and Associated Chaperones on Rho-Dependent Termination of Transcription In Vitro

Cédric Nadiras, Annie Schwartz, Mildred Delaleau, and Marc Boudvillain

Abstract

Besides their well-known posttranscriptional effects on mRNA translation and decay, sRNAs and associated RNA chaperones (e.g., Hfq, CsrA) sometimes regulate gene expression at the transcriptional level. In this case, the sRNA-dependent machinery modulates the activity of the transcription termination factor Rho, a ring-shaped RNA translocase/helicase that dissociates transcription elongation complexes at specific loci of the bacterial genome. Here, we describe biochemical assays to detect Rho-dependent termination signals in genomic regions of interest and to assess the effects of sRNAs and/or associated RNA chaperones on such signals.

Key words Rho, Transcription, Termination, Ring-shaped, Helicase, Hexamer, sRNA, Chaperone

1 Introduction

In bacteria, transcription termination is triggered by two main types of signals having distinct nucleic acid (NA) and protein cofactor requirements (reviewed in: [1, 2]). Intrinsic (Rho-independent) terminators are encoded by a NA signal that is usually sufficient to destabilize the transcription elongation complex. Intrinsic terminators are relatively easy to detect in bacterial genomes as they most often involve the formation of a GC-rich hairpin followed by a run of U residues at the 3′ end of the nascent RNA transcript. By contrast, Rho-dependent terminators are characterized by a strict requirement for the protein factor Rho (Fig. 1) but have relaxed sequence determinants rendering them difficult to detect in genome sequences [1].

Transcription termination signals are sometimes found in the 5′-untranslated regions (5′-UTRs) of genes or operons where they contribute to the regulation of the gene/operon through conditional attenuation of transcription [3, 4]. Attenuation mechanisms involving intrinsic terminators are diverse and wide-

Véronique Arluison and Claudio Valverde (eds.), *Bacterial Regulatory RNA: Methods and Protocols*, Methods in Molecular Biology, vol. 1737, https://doi.org/10.1007/978-1-4939-7634-8_7,

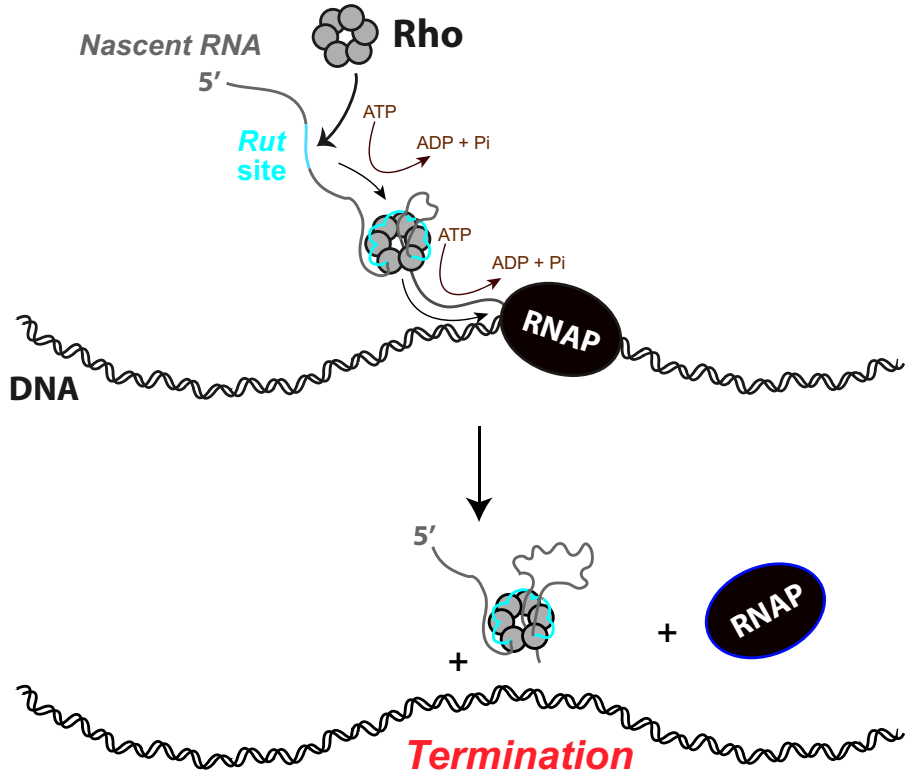


Fig. 1 Rho-dependent termination of transcription. The Rho hexamer binds a *Rut* site within a naked (i.e., untranslated) portion of the nascent transcript. Rho anchoring to the transcript activates the ATP-dependent translocation of RNA within the hexamer central channel (note that the initial interaction with the *Rut* site is maintained throughout) [27]. Then, Rho triggers dissociation of the transcription elongation complex upon catching up with the RNA polymerase (RNAP). Various mechanisms whereby sRNAs and RNA chaperones regulate Rho-dependent termination of transcription have been described (not shown on the figure). These mechanisms include formation of an Hfq-dependent antitermination complex [17], modulation of Rho access to a *Rut* site through inhibition of translation [14] or structural remodeling of the nascent RNA by CsrA [11] as well as less clear modes of structural interference by sRNA:mRNA complexes [10, 15]

spread while, until recently, known instances of comparable Rho-dependent mechanisms were few [5–7]. This situation is, however, quickly evolving with the discovery of riboswitches governing Rho-dependent termination [8, 9] and, more to the point of the present chapter, of conditional Rho-dependent mechanisms governed by sRNAs and/or RNA chaperones. For instance, a Rho-dependent terminator has been identified recently in the 5′-UTR of the *rpoS* gene of *Escherichia coli* where it is regulated negatively by the binding of sRNA DsrA, ArcZ, or RprA to the *rpoS* mRNA leader [10]. A conditional Rho-dependent terminator operating by a distinct mechanism has also been identified recently in the 5′-UTR of the *pgaABCD* operon of *E. coli* [11]. In this case, remodeling of the mRNA leader by the CsrA chaperone allows Rho access to a binding *Rut* (*R*ho *ut*ilization) site that is otherwise

sequestered in a RNA secondary structure [11]. This mechanism contributes to the multilayered regulation of *pgaABCD* by CsrA, which notably includes scavenging of CsrA by CsrB, CsrC, and McaS acting as sRNA sponges [12, 13].

Conditional Rho-dependent termination can also be triggered by the binding of a sRNA to its mRNA target. In the first case described, binding of the sRNA ChiX to the 5′-leader of the *chiPQ* operon of *Salmonella* prevents *chiP* translation [14]. This in turn exposes an intragenic *Rut* site within the *chiP* mRNA that is otherwise hidden by translating ribosomes, and triggers Rho-dependent transcriptional polarity within the operon [14]. Similarly, the sRNA Spf (aka Spot 42) triggers Rho-dependent termination at the end of the *galT* gene within the *galETKM* operon of *E. coli* [15]. These two known cases suggest that the activation of Rho-dependent transcriptional polarity by sRNAs could be a general mechanism contributing to gene/operon silencing. When exploring this possibility, however, one needs to consider the potential contribution of accessory factors. For instance, the Hfq chaperone can inhibit Rho-dependent termination in a manner that is antagonized by NusG—an essential, multi-role transcription factor [16] or by NA ligands that alter Hfq interaction with Rho and RNA [17]. Moreover, some Rho-dependent terminators are effective only in the presence of NusG, probably because their *Rut* sites are too weak to recruit or activate Rho by themselves [14, 18, 19].

Here, we describe in vitro methods to probe whether transcription termination factor Rho is involved in sRNA-mediated regulation of a given bacterial gene or operon. Together with in vivo probing, these methods should prove useful to unravel the potentially complex interplays existing between Rho, sRNAs, and accessory factors such as RNA chaperones.

2 Materials

2.1 Preparation of DNA Templates for Transcription Termination Assays

1. 100 μM stock solutions of DNA primers FWD, REV, and T7A1 (Table 1).
2. Genomic DNA or frozen stock of bacterial strain of interest (*see* **Note 1**).
3. Stock solution of dNTPs (10 mM each).
4. 2 U/μLVent DNA polymerase (*see* **Note 2**).
5. 10× Vent reaction buffer: 100 mM $(NH_4)_2SO_4$, 100 mM KCl, 20 mM $MgSO_4$, 1% Triton X-100, 200 mM Tris–HCl, pH 8.8.
6. PCR thermocycler equipped with a heated lid and a thermal block for 0.5 mL microtubes.
7. PCR purification kit (*see* **Note 3**).
8. Benchtop centrifuge.

Table 1
Sequences of the oligonucleotide primers[a]

Template	Primer	Sequence[b]
rpoS[c]	$FWD_{(rpoS)}$	5′-GTCTAACCTATAGGATACTTACAGCCCTTCGGGTGAACAGAGTGCTAACAAAATG
	$REV_{(rpoS)}$	5′-ATAACAGCTCTTCTTCAGCCAGGTCGTTATCAC
	T7A1[d]	5′-TTATCAAAAAGAGTATTGACTTAAAGTCTAACCTATAGGATACTTACAGCC
DsrA[e]	$TOP_{(DsrA)}$	5′-TAATACGACTCACTATAGGAACACATCAGATTTCCTGGTGTAACG
	$BOTT_{(DsrA)}$	5′-AAATCCCGACCCTGAGGGGGTCGGGATGAAAC
ArcZ[e]	$TOP_{(ArcZ)}$	5′-TAATACGACTCACTATAGGGGTGCGGCCTG AAAAACAGTG CTGTGCCCTT G
	$BOTT_{(ArcZ)}$	5′-AAAAAATGACCCCGGCTAGACCGGGGTGCGC
FnrS[e]	$TOP_{(FnrS)}$	5′-TAATACGACTCACTATAGGCAGGTGAATGCAACGTCAAGCGATGGGC
	$BOTT_{(FnrS)}$	5′-AAAAAGCCGACTCATCAAAGTCGGCGTCGTACGAATCAATTGTGCTATG
Spot 42[e] (Spf)	$TOP_{(Spf)}$	5′-TAATACGACTCACTATAGGTAGGGTACAGAGGTAAGATGTTCTATCTTTC
	$BOTT_{(Spf)}$	5′-TAAAAAACGCCCCAGTCATTACTGACTGGGGCGG

[a]Primers used to study sRNA effects on Rho-dependent termination in the 5′-leader region of *rpoS*
[b]Promoter sequences are underlined. Other oligonucleotide regions are chosen based on which genomic region needs to be amplified (Fig. 2a)
[c]For transcription termination assays with RNA polymerase from *E. coli*
[d]T7A1 is a universal primer used only in the second round of PCR amplification
[e]For preparation of indicated sRNA by transcription with T7 RNA polymerase

9. 1× $T_{10}E_1$ buffer: 10 mM Tris–HCl, 1 mM EDTA, pH 7.5 (*see* **Note 4**).
10. Agarose.
11. Commercial DNA ladder (e.g., #N3233S ladder from New England Biolabs).
12. 6× agarose gel loading buffer: 15% Ficoll-400, 0.1% SDS, 0.1% bromophenol blue, 20 mM Tris–HCl, 66 mM EDTA, pH 8.0.
13. 20× TAE buffer: dissolve 98 g of Tris base in approximately 800 mL of water, then add 22.8 mL of glacial acetic acid and 40 mL of 0.5 M EDTA pH 8.0, and add water to obtain a final volume of 1 L.
14. 1× TAE buffer: obtained by dilution of the 20× TAE stock buffer.
15. Horizontal gel electrophoresis system and power supply.
16. 10 mg/mL ethidium bromide stock solution (*see* **Note 5**).
17. Microwave oven.
18. UV transilluminator or dedicated gel documentation system.
19. Sephadex G-50 spin columns (e.g., Microspin columns from GE Healthcare).
20. UV spectrophotometer suitable for micro-volumes measurements (e.g., Nanodrop 2000c from Thermo Scientific).

2.2 Preparation of sRNAs

1. Items 2–20 of Subheading 2.1.
2. 100 µM stock solutions of DNA primers TOP and BOTT (Table 1).
3. 5× Transcription buffer: 0.12 M $MgCl_2$, 0.4 M HEPES, pH 7.5, 0.1 M DTT, 0.05% Triton X-100, and 5 mM Spermidine (*see* **Note 4**).
4. 50 U/µL T7 RNA polymerase.
5. 1 U/µL RQ1 DNase (Promega).
6. Dry bath incubator with shaking capability.
7. 20 U/µL SUPERase-IN™ (Thermo Fisher Scientific).
8. Deionized RNA-grade water (*see* **Note 4**).
9. Set of high-grade rNTPs (100 mM each).
10. 0.5 M EDTA solution, adjusted to pH 7.5 with NaOH (*see* **Note 4**).
11. Phenol:Chloroform:Isoamyl alcohol (25:24:1) mix, pH 6.7.
12. Diethyl ether.

13. 3 M sodium acetate (NaAc), adjusted to pH 6.5 with acetic acid (*see* **Note 4**).
14. 20× TBE buffer: dissolve 216 g of Tris base, 110 g of boric acid, 14.9 g EDTA in 1 L of water. Filter on Whatman paper and store at room temperature.
15. 1× TBE buffer: obtained by dilution of the 20× TBE stock.
16. 40% acrylamide:bis-acrylamide [29:1 ratio] commercial stock solution.
17. Denaturing acrylamide solution for sRNA preparation (6% acrylamide:bis-acrylamide [29:1 ratio] and 7 M urea in 1× TBE buffer). Mix 12.6 g of urea, 4.5 mL of 40% acrylamide:bis-acrylamide [29:1 ratio] solution, 1.5 mL 20× TBE, and 10 mL of deionized water. Heat the solution to dissolve urea completely. Adjust volume to 30 mL with deionized water and cool down to room temperature. Prepare fresh solution before use.
18. N,N,N,N′-tetramethyl-ethylenediamine (TEMED).
19. 25% (w/v) ammonium persulfate (APS) in water.
20. Vertical electrophoresis system and power supply for DNA sequencing (e.g., adjustable 20 × 42 cm sequencing kit from CBS scientific).
21. Denaturing loading buffer: 95% formamide, 5 mM EDTA, 0.01% (w/v) bromophenol blue.
22. X-ray intensifying screen or a fluor-coated TLC plate.
23. Hand-held 254 nm UV lamp.
24. 1× Elution buffer: 0.3 M NaAc, 10 mM MOPS, 1 mM EDTA, pH 6.0.
25. 1× $M_{10}E_1$ buffer: 10 mM MOPS, 1 mM EDTA, pH 6.0 (*see* **Note 4**).

2.3 Transcription Termination Assay

1. Items 6–21 of Subheading 2.2.
2. “Protein low binding” and “DNA low binding” 1.5 mL microtubes.
3. Commercial DNA ladder. Select a ladder made of DNA fragments that are not phosphorylated at 5′-ends (e.g., #N3233S ladder from New England Biolabs).
4. 10 U/μL T4 polynucleotide kinase.
5. 10× PNK buffer: 100 mM $MgCl_2$, 50 mM DTT, and 700 mM Tris–HCl, pH 7.6.
6. γ-^{32}P ATP at 3000 Ci/mmol [10 mCi/mL].

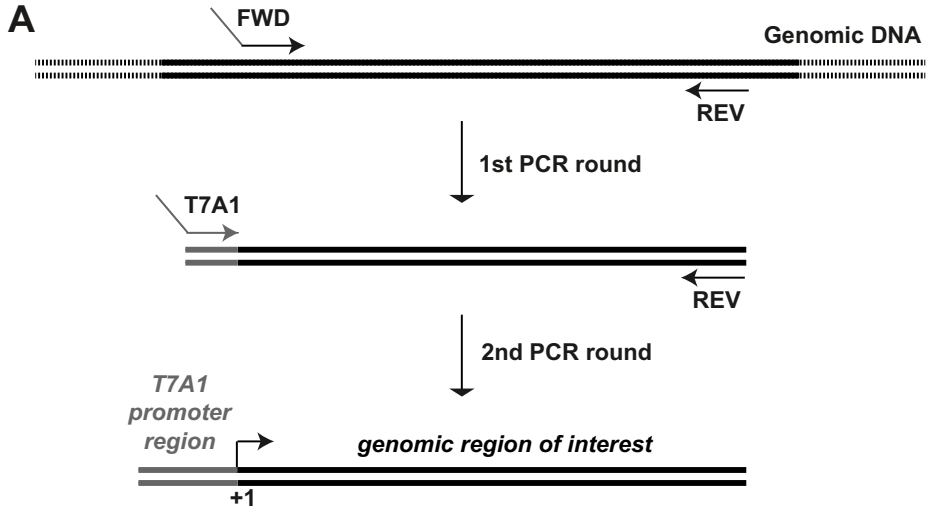
7. Sephadex G-50 spin columns (e.g., Microspin columns from GE Healthcare).
8. 1.4 μM Rho stock solution (*see* **Note 6**) in Rho storage buffer (50% glycerol, 100 mM KCl, 0.1 mM EDTA, 0.1 mM DTT, 10 mM Tris–HCl, pH 7.9). Preparation of the Rho protein from *E. coli* is detailed in the volume 587 of *Methods in Molecular Biology* [20] (*see* **Note 7**).
9. 2 μM NusG stock solution (*see* **Note 8**).
10. 2 μM stock solution of CsrA or Hfq chaperone (optional) (*see* **Note 8**).
11. 1 U/μL *E. coli* RNA Polymerase, Holoenzyme (New England Biolabs).
12. 5 μM stocks of sRNAs (as prepared in Subheading 3.2). Stocks should include the sRNAs under investigation (i.e., the ones expected to pair with the mRNA target) and at least one negative control (sRNA not expected to bind to the mRNA target). We also like to include a shorter, synthetic oligoribonucleotide that is fully complementary to the mRNA sequence targeted by the sRNA(s) and serves as positive control.
13. 5× transcription termination buffer: 250 mM KCl, 25 mM $MgCl_2$, 7.5 mM DTT, 0.25 mg/mL bovine serum albumin, and 200 mM Tris–HCl, pH 8.0 (*see* **Note 4**).
14. 10× initiation mixture: 2 mM ATP, 2 mM GTP, 2 mM CTP, 0.2 mM UTP, 250 μg/mL rifampicin, and 2 μCi/μL ^{32}P-αUTP in 1× transcription termination buffer. Prepare right before use (*see* **Note 9**).
15. 0.25 mg/mL tRNA stock.
16. 1× resuspension buffer: mix 460 μL of $M_{10}E_1$ buffer with 40 μL of 0.5 M EDTA, pH 7.5.
17. Denaturing acrylamide solution (termination assay): 7% acrylamide:bis-acrylamide [19:1 ratio] and 7 M urea in 1× TBE buffer (*see* Subheading 2.2, item 17 for preparation from stocks, using a 19:1 rather than 29:1 acrylamide:bis-acrylamide commercial mixture).
18. Vacuum drying system for electrophoresis gels.
19. Phosphor imager equipped with 35 × 43 cm phosphor imaging plates and dedicated analysis software (e.g., Typhoon Trio imager and ImageQuant TL software from GE-Healthcare).

3 Methods

3.1 Preparation of DNA Templates for Transcription Termination Assays

We prefer working with DNA templates that do not exceed 1000 bp, as longer templates may yield transcription termination products that are not easily separated from runoff transcripts by denaturing polyacrylamide gel electrophoresis. The DNA templates contain the genomic region of interest downstream from the T7A1 promoter (which is efficiently used by the RNA polymerase from *E. coli*) and are prepared in two successive rounds of PCR amplification (Fig. 2a) as follows:

1. Using a sterile loop, scrap and transfer a tiny piece of frozen bacteria (*see* **Note 1**) into a 1.5 mL microtube. Add 100 µL of water, incubate for 5 min at 95 °C, and then keep on ice (this solution can be stored at −20 °C for further use). A diluted solution of purified genomic DNA may be used instead.
2. Transfer 3 µL of above solution into a 0.5 mL PCR microtube kept on ice. Add 39 µL of water, 5 µL of 10× Vent reaction buffer, 0.5 µL of each 100 µM solution of primers FWD and REV (Table 1), 1 µL of the stock mixture of dNTPs (10 mM each), and 1 µL of Vent polymerase.
3. Transfer mixture into a thermocycler and heat it for 3 min at 94 °C. Then, perform 30 cycles of PCR amplification using the following cycle parameters: 94 °C for 1 min, 50 °C for 30 s, 72 °C for 1 min. Incubate for 10 min at 72 °C at the end of the program.
4. Transfer 0.6 µL of the PCR mixture into a new 0.5 mL PCR microtube kept on ice. Add 41.4 µL of water, 5 µL of 10× Vent reaction buffer, 0.5 µL of each 100 µM solution of primers T7A1 and REV (Table 1), 1 µL of the stock mixture of dNTPs (10 mM each), and 1 µL of Vent polymerase.
5. Perform the second PCR round as in **step 3**. At the end of amplification, add 2 µL of 0.5 M EDTA solution and store on ice.
6. Prepare agarose gel by dissolving 1.5 g of agarose into 100 mL of 1× TAE buffer (perform short heating bursts in microwave oven). Add 1 µL of ethidium bromide stock solution (*see* **Note 5**) and pour into gel tray. Once the gel has solidified, install it into a horizontal electrophoresis unit filled with 1× TAE buffer.
7. Mix 1 µL of PCR reaction mixture with 7.3 µL of 1× $T_{10}E_1$ buffer, and 1.7 µL of 6× agarose gel loading buffer. Similarly prepare a sample containing ~0.5 µg of commercial DNA ladder.
8. Load samples on agarose gel and run the gel for 1 h at 100 Volts.
9. Visualize gel with a transilluminator or dedicated gel documentation system. Only one band migrating at the expected rate for the correct DNA fragment should be visible (Fig. 2c).



B **DNA template for *in vitro* transcription**

5′.TTATCAAAAAGAGTATTGACTTAAAGTCTAACCTATAGGATACTTACAGCCCTTCGGGTGAACAGAGTGCTA
ACAAAATGTTGCCGAACAACAAGCCAACTGCGACCACGGTCACAGCGCCTGTAACGGTACCAACAGCAAGCACAA
CCGAGCCGACTGTCAGCAGTACATCAACCAGTACGCCTATCTCCACCTGGCGCTGGCCGACTGAGGGCAAAGTGA
TCGAAACCTTTGGCGCTTCTGAGGGGGGCAACAAGGGGATTGATATCGCAGGCAGCAAAGGACAGGCAATTATCG
CGACCGCAGATGGCCGCGTTGTTTATGCTGGTAACGCGCTGCGCGGCTACGGTAATCTGATTATCATCAAACATA
ATGATGATTACCTGAGTGCCTACGCCCATAACGACACAATGCTGGTCCGGGAACAACAAGAAGTTAAGGCGGGGC
AAAAAATAGCGACCATGGGTAGCACCGGAACCAGTTCAACACGCTTGCATTT***TGAAATTCGTTACAAGGGGAAAT***
CCGTAAACCCGCTGCGTTATTTGCCGCAGCGATAAATCGGCGGAACCAGGCTTTTGCTTGAATGTTCCGTCAAGG
GATCACGGGTAGGAGCCACCTT**atg**AGTCAGAATACGCTGAAAGTTCATGATTTAAATGAAGATGCGGAATTTGA
TGAGAACGGAGTTGAGGTTTTTGACGAAAAGGCCTTAGTAGAACAGGAACCCAGTGATAACGATTTGGCCGAAGA
GGAACTGTTAT.3′

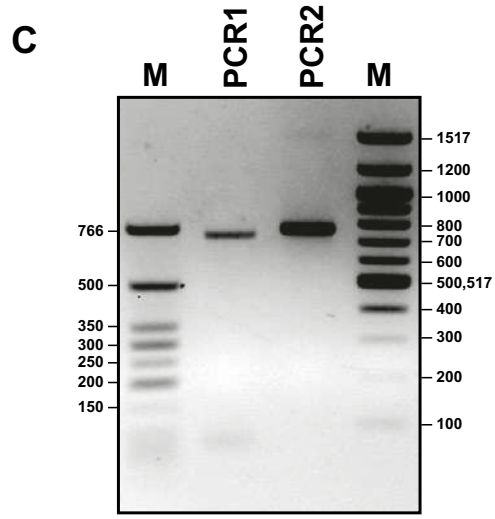


Fig. 2 Preparation of DNA templates for in vitro transcription. (**a**) Outline of the method. (**b**) Sequence of the *rpoS* template used as example in the present chapter. The template contains the 5′UTR as well as the first 139 nucleotides of the *rpoS* coding region. The sequence of the T7A1 promoter is underlined while the region recognized by sRNAs ArcZ and DsrA [28] and the ATG start codon are in bold. (**c**) Representative 1.5% agarose gel of the DNA fragments obtained after rounds 1 and 2 of PCR amplification of the *rpoS* template using primers FWD and REV (*see* Table 1)

10. Purify the PCR reaction mixture in two steps, first with a commercial silica-based purification kit (*see* **Note 3**) and then with a commercial G-50 spin column following manufacturer's instructions.
11. Use a μL spectrophotometer to determine the molar concentration of the DNA template from the absorbance of the solution at 260 nm, assuming ε_{260} ~ [13,200 × number of base pairs] L/mol/cm. Typical yields for a 1000 bp DNA fragment are around 10 pmole.
12. Adjust DNA template concentration to 100 nM with $T_{10}E_1$ buffer and store at −20 °C.

3.2 Preparation of sRNAs

Most sRNAs exceed the size limit of commercial synthetic oligoribonucleotides (~80 nts) and need to be prepared by in vitro transcription with a phage RNA polymerase using either a dedicated commercial kit or the following protocol:

1. Follow instructions in Subheading 3.1 to prepare the DNA template encoding the sRNA of interest. To introduce a promoter for T7 RNA polymerase (instead of the T7A1 promoter recognized by the RNA polymerase from *E. coli*), use oligonucleotides TOP and BOTT instead of primers FWD (or T7A1) and REV in both rounds of PCR.
2. Gently thaw transcription buffer, DNA template, and rNTP stocks on ice. Homogenize each solution by brief vortexing and centrifugation.
3. In a microtube, assemble on ice a mixture of 119 μL of water, 50 μL of 5× transcription buffer, 12.5 μL of each 100 mM rNTP stock, 1 μL of SUPERase-IN™, and 20 μL of the 100 nM DNA template solution.
4. Add 10 μL of T7 RNA polymerase and incubate mixture for 2 h at 37 °C.
5. Add 5 μL of RQ1 DNase to digest the DNA template and incubate for 20 min at 37 °C.
6. Add 12 μL of 0.5 M EDTA and 28 μL of 3 M NaAc.
7. Extract with one volume of Phenol:Chloroform:Isoamyl alcohol mix. Vortex and centrifuge briefly to separate phases. Transfer the aqueous (top) phase to a new tube and extract twice with one volume of ether, discarding the top (ether) phase in each case.
8. Add 900 μL of ice-cold ethanol. Incubate overnight at −20 °C.
9. Centrifuge for 30 min at 20,000 × *g* in a refrigerated centrifuge and discard supernatant. Wash the pellet with 150 μL of ice-cold ethanol and centrifuge for 15 min at 20,000 × *g*.

10. Remove the ethanol and leave the microtube open for ~20 min at room temperature to dry the RNA pellet (the process can be sped up by incubating at 37 °C in a dry bath) (*see* **Note 10**).
11. Dissolve pellet in a mix of 20 µL of $M_{10}E_1$ buffer and 20 µL of denaturing loading buffer. Leave sample on ice while preparing the denaturing polyacrylamide gel for purification.
12. Assemble gel plates and spacers according to manufacturer instructions. We use custom-made 20 × 20 cm gel plates equipped with 0.8 mm spacers, a 10-teeth comb, and a bottom tape seal.
13. Mix 30 mL of 6% denaturing (29:1) polyacrylamide gel solution with 90 µL of APS and 45 µL of TEMED (volumes may need to be adjusted for commercial sets of plates and spacers). Quickly pour the mixture between the gel plates, and insert comb.
14. Once the gel has polymerized (~30 min), remove the comb and wash the gel wells with 1× TBE using a 5 mL syringe. Install the gel into an electrophoresis unit and fill the top and bottom tanks with 1× TBE.
15. Set the power supply at 20 W and perform a pre-electrophoresis for 20 min.
16. Heat-denature the RNA sample (from **step 11**) for 2 min at 95 °C.
17. Turn off the power supply and flush diffusing urea from gel wells using a syringe containing 1× TBE. Distribute sample into two wells using a flat gel loading tip.
18. Run the gel at 20 W until the band corresponding to bromophenol blue reaches the bottom of the gel.
19. Carefully remove the glass plates and wrap the gel in saran sheets.
20. Place the gel on an X-ray intensifying screen (or a fluor-coated TLC plate) and visualize the band corresponding to the transcript by UV shadowing in a dark room with a hand-held 254 nm lamp (*see* **Note 11**).
21. Cut the band with a clean scalpel and crush it by passage through a 1 mL syringe. Soak the gel pieces in 3 mL of 1× elution buffer in a sterile 14 mL culture tube. Shack the tube overnight at 4 °C.
22. Pass the gel slurry through a 5-mL syringe equipped with a glass wool or cotton plug (to retain gel particles) and measure the volume of the resulting solution.
23. Extract the filtered solution with one volume of Phenol:Chloroform:Isoamyl alcohol mix. Vortex and centrifuge briefly to separate phases. Transfer the aqueous (top) phase to a new tube and extract twice with one volume of ether.

24. Add three volumes of ice-cold ethanol and incubate overnight at −20 °C.
25. After centrifugation for 30 min at 20,000 × *g* in a refrigerated benchtop centrifuge, discard supernatant and wash carefully the RNA pellet with 300 µL of 70% ice-cold ethanol. Centrifuge again for 10 min at 20,000 × *g* and discard the ethanol wash.
26. Leave the microtube open for ~20 min at room temperature to dry the RNA pellet (*see* **Note 10**) and dissolve it in 50–100 µL of $M_{10}E_1$ buffer.
27. Use a µL spectrophotometer to determine RNA concentration from the absorbance of the solution at 260 nm, assuming ε_{260} ~ [10^4 × number of nucleotides] L/mol/cm. Typical yields range between 1 and 4 nmoles of purified sRNA for a 250 µL transcription.
28. Store sRNA stock solution at −20 °C.

3.3 Transcription Termination Assay

3.3.1 Detection of Rho-Dependent Signals

To probe the effect of a sRNA or chaperone on a known, well-characterized Rho-dependent termination signal, one may skip this section and proceed directly to Subheading 3.3.2.

If the evidence connecting the effects of Rho and a sRNA (or RNA chaperone) on a given gene or operon is vague or indirect, the first step is to determine if this gene or operon contains a Rho-dependent signal. We advise to also check for the potential effect of NusG as this factor can dramatically stimulate Rho-dependent termination at suboptimal *Rut* sites (as was observed for the ChiX-regulated *chiP* terminator of *Salmonella*) [14]. To undertake these tasks, proceed as follows:

1. To a regular 1.5 mL microtube, add in the following order, 3 µL of deionized water, 1 µL of 10× PNK buffer, 2 µL of DNA ladder (1 µg/µL), 3 µL of γ-^{32}P-ATP (*see* **Note 9**), and 1 µL of T4 polynucleotide kinase. Incubate 40 min at 37 °C.
2. Add 2 µL of 0.5 M EDTA and 78 µL of $T_{10}E_1$ buffer. Pass through a G-50 spin column, following manufacturer's instructions. Discard the G-50 column and store the eluate containing the ^{32}P-labeled DNA ladder at −20 °C (*see* **Note 9**). The ^{32}P-labeled ladder can be used for up to ~2 months.
3. Prepare three 1.5 mL microtubes ("Protein low binding" grade) labeled "T," "Rho," and "NusG," respectively. Place tubes on ice.
4. In a "Protein low binding" microtube prepare a master mix containing 35.2 µL of deionized water H_2O, 10 µL of 5× transcription termination buffer, 3.2 µL of the 100 nM stock solution of DNA template, 0.6 µL of SUPERase-IN™, and 0.9 µL of *E. coli* RNA polymerase.

5. Dispatch 15.6 µL of the master mix in tubes "T," "Rho," and "NusG."
6. Add 2.4 µL of 1× transcription termination buffer to tube "T." To tube "Rho," add 1.4 µL of 1× transcription termination buffer and 1 µL of the 1.4 µM Rho stock. Add 1 µL of the 1.4 µM Rho stock and 1.4 µL of the 2 µM NusG stock to tube "NusG." Vortex gently and centrifuge briefly to homogenize each tube solution.
7. Incubate reaction tubes for 10 min at 37 °C.
8. Add 2 µL of 10× initiation mixture (*see* **Note 12**). From this step to the end of the procedure, samples will contain substantial amounts of radioactive material and should be handled accordingly (*see* **Note 9**).
9. Incubate reactions for 20 min at 37 °C. Then, stop reactions by adding, to each tube, 4 µL of EDTA (0.5 M), 2 µL of tRNA (0.25 mg/mL), 64 µL of deionized water, and 10 µL of 3 M NaAc pH 6.5.
10. Extract each tube solution with one volume of Phenol:Chloroform:Isoamyl Alcohol mix. Vortex and centrifuge briefly to separate phases. Transfer the aqueous (top) phases to new tubes (use only "DNA low binding" tubes from this step) and extract twice with one volume of ether.
11. Add three volumes of ethanol (stored at room temperature) to each tube and incubate for 35 min on ice. Under these conditions, most of the free ^{32}P-αUTP will not precipitate.
12. Centrifuge for 20 min at 20,000 × *g* in a bench centrifuge at room temperature. Discard supernatants and wash carefully the pellets with 150 µL of 100% ethanol (stored at room temperature). Centrifuge again for 10 min at 20,000 × *g* and discard the ethanol washes. Open the tubes and dry the pellets for ~5 min at room temperature (*see* **Note 10**).
13. Following instructions in **steps 12** and **13** of Subheading 3.2 but using 20 × 40 cm (w × l) gel plates and 0.4 mm spacers, prepare a 7% denaturing (19.1) polyacrylamide gel and install it in a vertical electrophoresis unit.
14. Perform a pre-electrophoresis for 45 min at 45 W (gel plates should become warm to the touch).
15. In a 1.5 mL microtube, mix 9 µL of denaturing loading buffer with 1 µL of ^{32}P-labeled ladder from **step 2** (increase volume of ladder preparation if older than a few days).
16. Dissolve pellets from **step 12** in 6 µL of 1× resuspension buffer (incubate tubes at 30 °C for a few minutes to help dissolution) and add 7 µL of denaturing loading buffer.

17. Heat sample and ladder tubes for 2 min at 90 °C. Then, flush diffusing urea from gel wells using a syringe containing 1× TBE and quickly load samples and ladder in the gel wells.
18. Run the gel at 45 W until the band corresponding to xylene cyanol is ~25 cm from the bottom of gel wells.
19. Carefully remove one of the glass plates and replace it with a sheet of Whatman paper. Remove the second glass plate and replace it with saran wrap.
20. Dry the gel in a vacuum gel dryer and expose it overnight to a phosphorimager screen in an exposure cassette.
21. Scan the phosphorimager screen with a dedicated system. The presence of a Rho-dependent signal in the DNA template of interest (Fig. 3a) is made apparent by the apparition of fast migrating bands in the samples containing Rho and by the concomitant decrease in the intensity of the band corresponding to the formation of runoff transcripts (Fig. 3b) (*see* **Note 13**). This trend is usually accentuated in samples containing both Rho and NusG (Fig. 3b) (*see* **Note 14**).

3.3.2 Effects of sRNAs and/or RNA Chaperones

Below we describe our procedure to probe the effects of sRNAs on Rho-dependent termination. The protocol can be easily adjusted to probe the effects of RNA chaperones such as Hfq [17] or CsrA [11]. Because Hfq alone can inhibit Rho [17], we recommend testing Hfq and sRNAs separately, and using control oligoribonucleotides that form perfect duplexes with the mRNA regions targeted by the sRNAs. To find transcription conditions optimal for sRNA-mRNA pairing, we also recommend to run a few exploratory experiments where the concentration of the tested sRNA or the concentration of KCl is varied. In the case of the *rpoS* template, for instance, we observed that the effect of the sRNA DsrA is optimal in the presence of 75 mM KCl (Fig. 3c) and we adjusted the composition of the transcription buffer accordingly.

1. Prepare and label one 1.5 mL "Protein low binding" microtube per planned sample (including "minus Rho" and "minus sRNA" controls) and place tubes on ice.
2. For the n planned samples, prepare a master mix containing $9.4 \times (n + 1)$ µL of deionized water, $2.7 \times (n + 1)$ µL of 5× transcription termination buffer (containing 375 mM KCl in the case of the *rpoS* template), $(n + 1)$ µL of the 100 nM stock solution of DNA template, $0.2 \times (n + 1)$ µL of SUPERase-IN™, and $0.27 \times (n + 1)$ µL of *E. coli* RNA polymerase.
3. Dispatch 13.6 µL of the master mix into each sample tube.
4. Add 2.4 µL of 1× transcription termination buffer (containing 75 mM KCl in the case of the *rpoS* template) to tube "minus Rho."

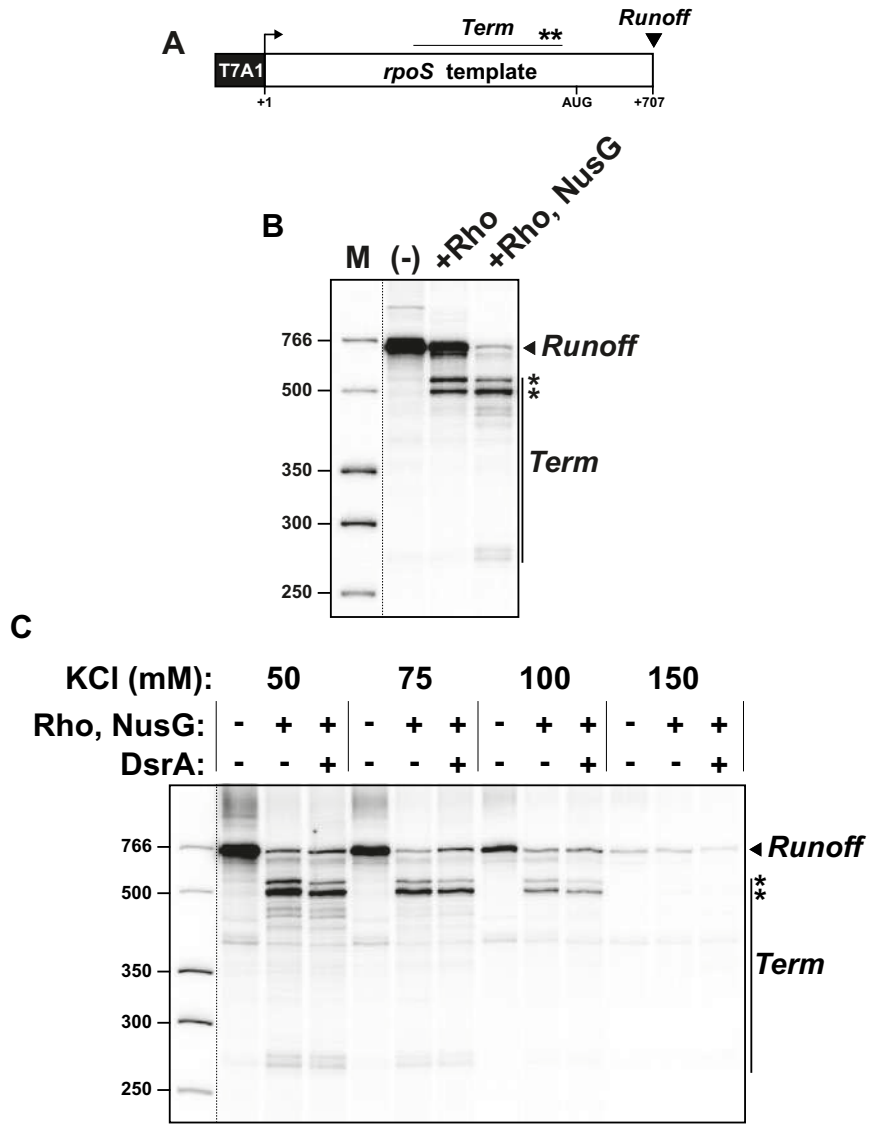


Fig. 3 Representative exploratory transcription termination experiments performed with the *rpoS* template. (**a**) Schematic of the *rpoS* template and transcription termination features. (**b**) Initial detection of Rho-dependent signals within the *rpoS* template. (**c**) Analysis of the effect of KCl. In the example shown, 75 mM KCl represents the best compromise between the effects of the salt on transcription initiation, Rho-dependent termination, and DsrA efficiency. The KCl concentration was increased right from the start of the reaction assay but one may change it at a later stage (to limit the inhibitory salt effects on transcription initiation such as shown on the present gel) by adjusting the composition of the 10× initiation mixture rather than the 5× transcription termination buffer. Note that the contrast of the DNA ladder lane has been adjusted separately as indicated by the dotted line separating it from other gel lanes

To each other tube (including "minus sRNA" tube), add 1 µL of the 1.4 µM Rho stock and 1.4 µL of the 2 µM NusG stocks.

5. Mix reactants by gently pipetting up and down and incubate for 10 min at 37 °C.
6. Add 2 µL of 1× transcription termination buffer to tubes "minus sRNA." To each other tube, add 2 µL of the 5 µM stock of appropriate sRNA or control oligoribonucleotide (*see* **Note 15**). Vortex gently and centrifuge briefly to homogenize each tube solution.
7. Proceed as described in **steps 8–21** of Subheading 3.3.1
8. Determine the apparent percentage of runoff product (*see* **Note 16**) using the phosphorimager analysis software. For instance, one may use the analysis toolbox of ImageQuant TL software (GE Healthcare) to box the runoff and termination product bands (rectangle area function; *see* Fig. 4, inset), select the "local average" option for background and "percent" option for view, and obtain percentages of band intensities directly and in an Excel-compatible format.
9. Evaluate sRNA- or chaperone-dependent effects through the comparison of the apparent percentages of runoff product (*see* **Note 16**) obtained for the different gel lanes (Fig. 4, diagram). We recommend using average percentage values obtained from at least three independent experiments to mitigate potential RNA degradation effects and the experimental variability inherent to such complex biochemical setups (*see* **Note 17**).

4 Notes

1. For the case chosen to illustrate the present chapter (analysis of the *rpoS* region of *E. coli*), we used a 20% glycerol stock of the *E. coli* reference strain MG1655 stored at −80 °C. The strain (or genomic DNA) can be obtained from biological repositories such as DSMZ (www.dsmz.de).
2. The Vent DNA polymerase works well for the preparation of most DNA templates having lengths in the 300–1000 bp range. Other high-fidelity DNA polymerases may be used or better when preparing DNA templates longer than 1 kB.
3. In our hands, the GeneJET PCR purification kit (ThermoFisher Scientific) works optimally when following manufacturer's instructions.
4. To eliminate bacteria, the major source of RNase contamination, stock solutions and buffers for transcription assays should be prepared with RNA-grade chemicals and water in small

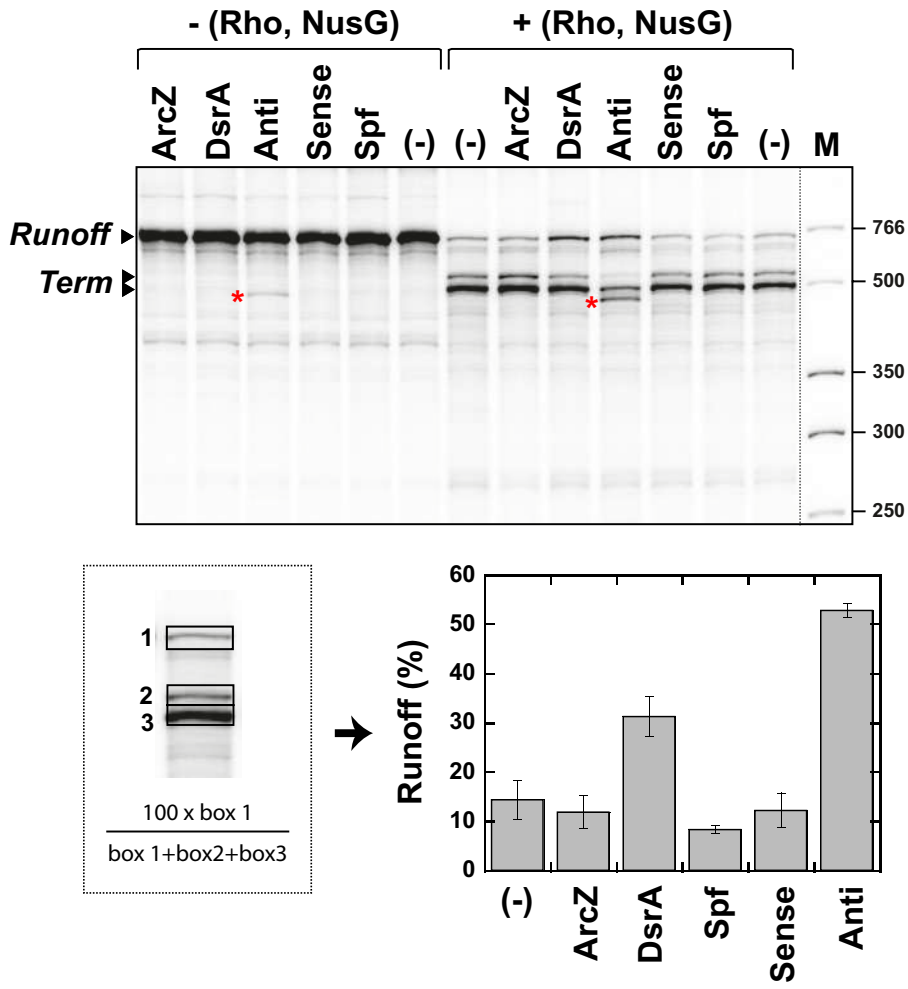


Fig. 4 Effects of sRNAs and control oligoribonucleotides on the *rpoS* terminator. Reactions were performed in the presence of 75 mM KCl and 500 nM of sRNA or control oligonucleotide "Sense" (5′ACUUUAAGCAAUGUUCCCCUUUAGGC) or "Anti" (5′CGGAUUUCCCCUUGUAACGAAUUUCA). The bands identified by asterisks correspond to a Rho-independent product that is formed only in the presence of the "Anti" strand, possibly because pairing of the oligonucleotide to the mRNA triggers intrinsic termination [29]. Runoff percentages shown in the diagram are average values obtained from 3 to 6 independent experiments (*see* **Note 16**). The ArcZ sRNA has no significant effect under the present conditions whereas it inhibited Rho-dependent termination at a higher concentration (1 μM) in a slightly different experimental setup [10]. This is consistent with the observation that ArcZ forms a less stable hybrid with the *rpoS* mRNA than does DsrA [28]

amounts (<50 mL) and sterilized with a 0.22 μm filter unit. We routinely obtain RNA-grade water by filtering ultrapure MilliQ (Millipore) water with 0.22 μm bottle-top sterile filter units. We usually avoid DEPC-treated water because harmful contaminants, such as rust particles, are often introduced during the autoclaving step required to remove excess DEPC.

5. Ethidium bromide is toxic. Solutions and waste (including gels) should be handled with great care following current safety regulation. Potentially less toxic DNA stains are available from various commercial sources and may be preferable for inexperienced users.
6. Concentrations of the Rho factor are expressed in hexamers throughout the chapter.
7. It is important to assess the enzymatic activity of every fresh Rho preparation and to verify that it does not decay upon storage at −20 °C for extended period of times. The simplest test for Rho activity is the determination of the rate of steady-state ATPase turnover as described in volume 1259 of the series [21].
8. NusG, CsrA, and Hfq proteins are now available from commercial sources (e.g., MyBioSource) which we have not tested. We prepare and use His-tagged versions of NusG [22] and CsrA [11] and obtained purified Hfq from V. Arluison [17].
9. Manipulation of ^{32}P-containing materials should be performed exclusively by individuals who have received proper training and authorization from a radiation safety officer.
10. Overdried RNA pellets are difficult to dissolve in aqueous buffers.
11. Full-length RNA should be seen clearly as one strong UV-shadow. If several shadows of comparable intensities are visible, one should assume that RNA degradation and/or formation of abortive transcription products have occurred in large proportion. It is then advisable to repeat the transcription/purification procedure with fresh solutions and reactants.
12. Rifampicin is present in the initiation mix to block subsequent rounds of transcription. Heparin cannot be used instead of rifampicin because it will trap Rho [23] in addition to free RNA polymerase.
13. Degradation of the transcripts (e.g., upon contamination of the samples by RNases) may yield similar band patterns. To ensure that the apparition of shorter-than-runoff termination products is Rho-dependent, it is advisable to repeat the experiment several times. Moreover, to ensure that neither the Rho nor NusG stock is contaminated by RNases, one may prepare control samples containing 150 μM of the Rho inhibitor bicyclomycin (supplemented in the initial mix of reactants). Rho-dependent bands will disappear at the profit of the runoff band whereas degradation product band patterns will not be affected by the presence of bicyclomycin.
14. Strictly speaking, the assay does not distinguish between truncated transcripts resulting from Rho-induced dissociation of transcription elongation complexes (termination) or from

transcription complexes that would be stably paused (or arrested) by Rho along the DNA template. To the best of our knowledge, however, the second scenario (Rho inducing transcriptional pausing or arrest) is purely hypothetical, without experimental support to date. We thus feel confident that the more complicated transcription assays required to distinguish between termination, pausing, and arrest [24] are not necessary in most cases.

15. When testing the effect of an RNA chaperone, the chaperone is added at this step in the place of the sRNA(s).
16. The apparent percentage of runoff product determined under these conditions should be used only to detect sRNA-dependent effects through the comparison of gel lanes (samples), as shown in the diagram of Fig. 4. This is because the apparent percentage of runoff product is not an accurate measure of terminator read-through as transcripts of different lengths contain different numbers of ^{32}P labels upon their internal labeling with ^{32}P-αUTP during transcription. Normalization of the intensities of gel bands for the uracil content of the corresponding transcripts is possible only when the sequences (lengths) of the various transcript species are known [25].
17. Because Rho is a nucleic acid-binding protein displaying some level of sequence specificity [26], indirect "sequestration" effects due to sRNA binding to Rho cannot be excluded beforehand, especially when considering the respective concentrations of Rho (70 nM) and sRNAs (500 nM) used in the assay. We strongly recommend testing this possibility by performing control transcription termination experiments with DNA template(s) encoding Rho-dependent terminator(s) unrelated to the original mRNA target sequence.

Acknowledgments

This work was supported by a PhD scholarship from Région Centre Val-de-Loire to C.N. and by a grant from Agence Nationale de la Recherche (ANR-15-CE11-0024-02) and CNRS core funding to M.B.

References

1. Porrua O, Boudvillain M, Libri D (2016) Transcription termination: variations on common themes. Trends Genet 32:508–522
2. Ray-Soni A, Bellecourt MJ, Landick R (2016) Mechanisms of bacterial transcription termination: all good things must end. Annu Rev Biochem 85:319–347
3. Santangelo TJ, Artsimovitch I (2011) Termination and antitermination: RNA polymerase runs a stop sign. Nat Rev Microbiol 9:319–329

4. Kriner MA, Sevostyanova A, Groisman EA (2016) Learning from the leaders: gene regulation by the transcription termination factor Rho. Trends Biochem Sci 41:690–699
5. Gish K, Yanofsky C (1995) Evidence suggesting cis action by the TnaC leader peptide in regulating transcription attenuation in the tryptophanase operon of Escherichia coli. J Bacteriol 177:7245–7254
6. Matsumoto Y, Shigesada K, Hirano M et al (1986) Autogenous regulation of the gene for transcription termination factor rho in Escherichia coli: localization and function of its attenuators. J Bacteriol 166:945–958
7. Yakhnin H, Babiarz JE, Yakhnin AV et al (2001) Expression of the Bacillus subtilis trpEDCFBA operon is influenced by translational coupling and Rho termination factor. J Bacteriol 183:5918–5926
8. Hollands K, Proshkin S, Sklyarova S et al (2012) Riboswitch control of Rho-dependent transcription termination. Proc Natl Acad Sci U S A 109:5376–5381
9. Takemoto N, Tanaka Y, Inui M (2015) Rho and RNase play a central role in FMN riboswitch regulation in Corynebacterium glutamicum. Nucleic Acids Res 43:520–529
10. Sedlyarova N, Shamovsky I, Bharati BK et al (2016) sRNA-mediated control of transcription termination in E. coli. Cell 167:111–121.e113
11. Figueroa-Bossi N, Schwartz A, Guillemardet B et al (2014) RNA remodeling by bacterial global regulator CsrA promotes Rho-dependent transcription termination. Genes Dev 28:1239–1251
12. Jorgensen MG, Thomason MK, Havelund J et al (2013) Dual function of the McaS small RNA in controlling biofilm formation. Genes Dev 27:1132–1145
13. Wang X, Dubey AK, Suzuki K et al (2005) CsrA post-transcriptionally represses pgaABCD, responsible for synthesis of a biofilm polysaccharide adhesin of Escherichia coli. Mol Microbiol 56:1648–1663
14. Bossi L, Schwartz A, Guillemardet B et al (2012) A role for Rho-dependent polarity in gene regulation by a noncoding small RNA. Genes Dev 26:1864–1873
15. Wang X, Ji SC, Jeon HJ et al (2015) Two-level inhibition of galK expression by spot 42: degradation of mRNA mK2 and enhanced transcription termination before the galK gene. Proc Natl Acad Sci U S A 112:7581–7586
16. Yakhnin AV, Babitzke P (2014) NusG/Spt5: are there common functions of this ubiquitous transcription elongation factor? Curr Opin Microbiol 18:68–71
17. Rabhi M, Espeli O, Schwartz A et al (2011) The Sm-like RNA chaperone Hfq mediates transcription antitermination at Rho-dependent terminators. EMBO J 30:2805–2816
18. Peters JM, Mooney RA, Grass JA et al (2012) Rho and NusG suppress pervasive antisense transcription in Escherichia coli. Genes Dev 26:2621–2633
19. Shashni R, Qayyum MZ, Vishalini V et al (2014) Redundancy of primary RNA-binding functions of the bacterial transcription terminator Rho. Nucleic Acids Res 42:9677–9690
20. Boudvillain M, Walmacq C, Schwartz A et al (2010) Simple enzymatic assays for the in vitro motor activity of transcription termination factor Rho from Escherichia coli. Methods Mol Biol 587:137–154
21. D'Heygere F, Schwartz A, Coste F et al (2015) Monitoring RNA unwinding by the transcription termination factor rho from Mycobacterium tuberculosis. Methods Mol Biol 1259:293–311
22. Artsimovitch I, Landick R (2000) Pausing by bacterial RNA polymerase is mediated by mechanistically distinct classes of signals. Proc Natl Acad Sci U S A 97:7090–7095
23. Nowatzke W, Richardson L, Richardson JP (1996) Purification of transcription termination factor Rho from Escherichia coli and Micrococcus luteus. Methods Enzymol 274:353–363
24. Kashlev M, Nudler E, Severinov K et al (1996) Histidine-tagged RNA polymerase of Escherichia coli and transcription in solid phase. Methods Enzymol 274:326–334
25. Rabhi M, Gocheva V, Jacquinot F et al (2011) Mutagenesis-based evidence for an asymmetric configuration of the ring-shaped transcription termination factor Rho. J Mol Biol 405:497–518
26. Rabhi M, Rahmouni AR, Boudvillain M (2010) Transcription termination factor Rho: a ring-shaped RNA helicase from bacteria. In: Jankowsky E (ed) RNA helicases, vol 19. RSC Publishing, Cambridge, pp 243–271
27. Gocheva V, Le Gall A, Boudvillain M et al (2015) Direct observation of the translocation mechanism of transcription termination factor Rho. Nucleic Acids Res 43:2367–2377
28. Soper T, Mandin P, Majdalani N et al (2010) Positive regulation by small RNAs and the role of Hfq. Proc Natl Acad Sci U S A 107:9602–9607
29. Yarnell WS, Roberts JW (1999) Mechanism of intrinsic transcription termination and antitermination. Science 284:611–615

V. Travaux non publiés III :

Etude de sites potentiellement soumis à une régulation Rho-dépendante conditionnelle

L'exploration des données 'omiques' publiées (Peters et al., 2012; Peters et al., 2009; Raghunathan et al., 2018; Sedlyarova et al., 2016) et de de nos prédictions (**Article I**, page 69) m'a amené à considérer un certain nombre de gènes ou d'opérons comme des candidats potentiels à une régulation conditionnelle mobilisant la terminaison Rho-dépendante (**Tableau 4**). Pour ces candidats, j'ai testé la présence de terminateur(s) Rho-dépendants dans les régions 5'UTR et l'influence d'inducteurs/régulateurs potentiels à l'aide d'essais *in vitro* de terminaison de la transcription (voir chapitres précédents pour les méthodes).

Tableau 4 : *Loci* suspectés d'être sujets à régulation Rho-dépendante conditionnelle et testés *in vitro*.

Gène/opéron	Cible potentielle de la régulation	Matrice[1]	Bactérie[2]	Terminaison Rho-dépendante observée	Inducteur/régulateur suspecté	Effet de l'inducteur observé
thiI	Synthèse thiamine	T002	MG16655	**Forte**	SAM (*S-adenosyl methionine*)	Aucun
rpoS[3]	Facteur sigma σ^S	T031	MG16655	**Forte**	NusG	Stimulation
					sRNA (*DsrA*)	inhibition
					sRNA (*RprA*)	inhibition
					sRNA (*ArcZ*)	inhibition
sodA	Superoxyde dismutase	T045	MG16655	**Forte**	NusG	Stimulation
					sRNA (*FnrS*)	Aucun
					CsrA	Aucun
					sRNA (*RyhB*)	Non testé
acs	Biosynthèse de l'acétyl-CoA	T049	MG16655	**Forte**	CsrA	Aucun
acnB	Aconitase	T057	LT2	**Faible**	NusG	Stimulation
					AcnB	Non testé
					sRNA (*RyhB*)	Non testé
sopD2	Protéines effectrices	T066	LT2	**Faible**	NusG	Stimulation
					CsrA	Aucun
phoP	Régulateur du taux de Mg^{2+} cytosolique	T067	LT2	**Faible**	NusG	Stimulation
					sRNA (*GcvB*)	Non testé
					sRNA (*MicA*)	Non testé
folE	GTP cyclohydrolase	T077	LT2	**XX**	NusG	Stimulation
					sRNA (*FnrS*)	Aucun
					sRNA (*SgrS*)	Non testé
nirB	Nitrite réductase	T091	LT2	**Faible**	NusG	Stimulation
					sRNA (*RyhB*)	Non testé
cspA	Protéine d'adaptation au choc froid	T094	LT2	**Faible**	NusG 37°C	Stimulation
					NusG 20°C	Stimulation
					CspA 20°C	Aucun
metE	Biosynthèse de la méthionine	T097	LT2	**Faible**	NusG	Stimulation
					sRNA (*FnrS*)	Non testé
cycA	Transporteur de Gly, Ala et Ser	T101	LT2	**Faible**	NusG	Stimulation
					sRNA (*GcvB*)	Non testé

[1] **:** Matrice ADN référencée *dans* (Nadiras et al., 2018a).
[2] : Génomes de référence : *E. coli* MG1655 (U00096.3), *Salmonella typhimirium* LT2 (NC_003197.1).
Les termes « effet positif » ou « effet négatif » veut dire que l'élément testé respectivement augmente ou diminue l'efficacité de terminaison Rho-dépendante.
[3] : Ce régulon a fait l'objet de publications récentes (Nadiras et al., 2018b; Sedlyarova et al., 2016).

Je me suis d'abord intéressé à la région 5'UTR du gène *thiI* (impliqué dans la synthèse de la thiamine) qui contient un terminateur Rho-dépendant facilement détectable *in vitro*

(**Figure 59A**) (Nadiras et al., 2018a) et qui pourrait également abriter un riboswitch induit par la S-adénosyl méthionine [SAM] (Raghavan et al., 2011). J'ai donc testé si ce riboswitch putatif pouvait contribuer à moduler l'activité du terminateur Rho-dépendant en réalisant des expériences de terminaison de la transcription *in vitro* en présence ou absence de SAM. Je n'ai pas observé de différences significatives (**Figure 59A**), suggérant que le riboswitch, s'il est fonctionnel (ce qui n'est pas complètement clair au vu de la littérature), ne conditionne pas l'accès au terminateur Rho-dépendant.

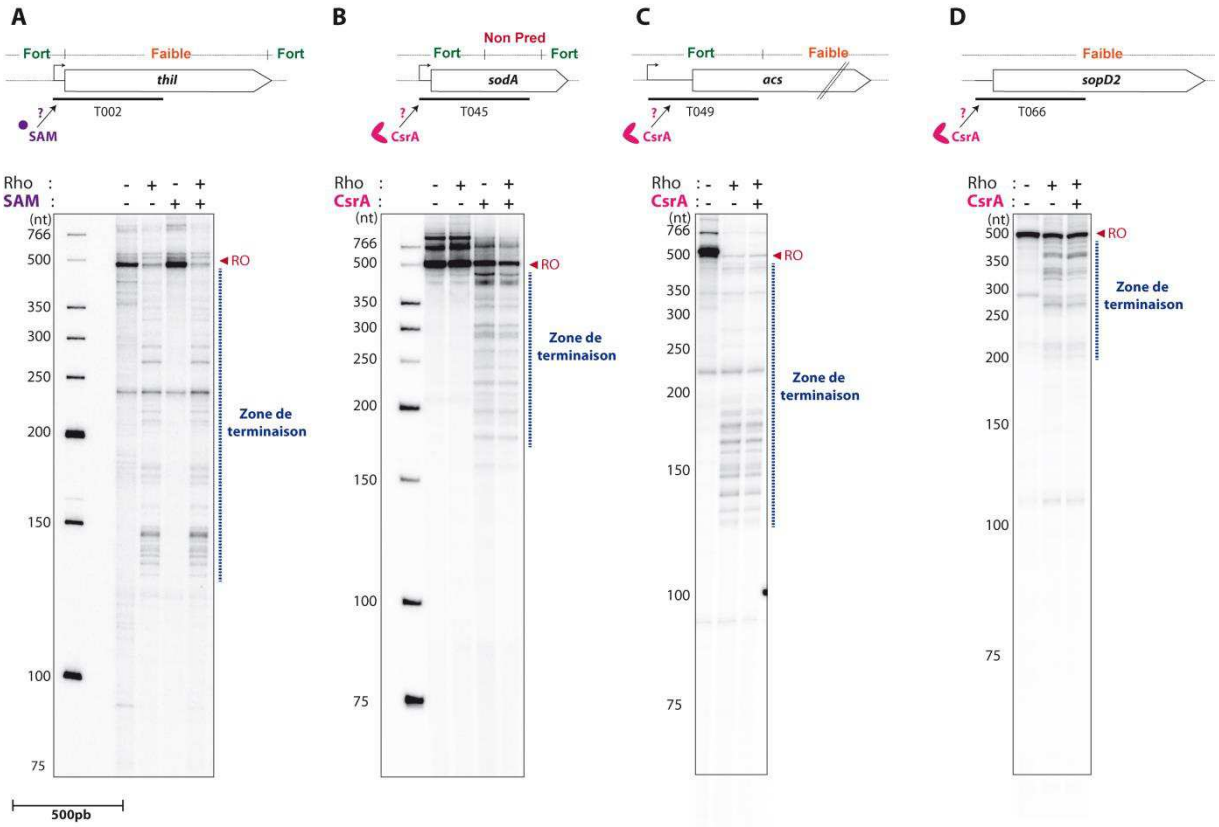


Figure 59 : Effet d'un riboswitch ou de CsrA sur certains terminateur Rho-dépendant. Impact du ligand [SAM] sur le terminateur Rho-dépendant du gène *thiI* **(A)** et de la proteine CsrA sur les terminateurs des gènes *sodA* **(B)**, *acs* **(C)** et *sopD2* **(D)**. Pour chaque cas, on retrouve une représentation, à l'échelle, du gène avec l'emplacement de la matrice ADN testé (en noir) et la prédiction de la présence de terminateur Rho-dépendant (Nadiras et al., 2018a). RO : *Runoff*. Les matrice sont analysées suivant la méthode d'écrite dans l'article méthode en présence de Rho, de NusG et de l'élément testé (partie 146) (Nadiras et al., 2018b).

J'ai également recherché de nouveaux exemples de régulation de la terminaison Rho-dépendante qui impliqueraient la protéine CsrA. Cette dernière conditionne l'accès à un site *Rut* localisé dans la région 5'leader de l'opéron *pgaABCD* d'*E. coli* (**Figure 34**) (Figueroa-Bossi et al., 2014) . Des études récentes montrent que CsrA contribue à la régulation des gènes *sopD2* (Holmqvist et al., 2016), *acs* (Wei et al., 2000), et *sodA* (Holmqvist et al., 2016). J'ai donc testé pour ces gènes si l'activité de CsrA repose, ou non, sur un mécanisme de

terminaison Rho-dépendante (comme c'est le cas pour *pgaABCD*). Bien que j'ai détecté des signaux de terminaison Rho-dépendante dans les régions analysées des gènes *sodA*, *acs* et *sopD2* (Nadiras et al., 2018a), je n'ai pas pu mettre en évidence de régulation qui dépendrait de la présence/absence de CsrA (**Figure 59B-C-D**).

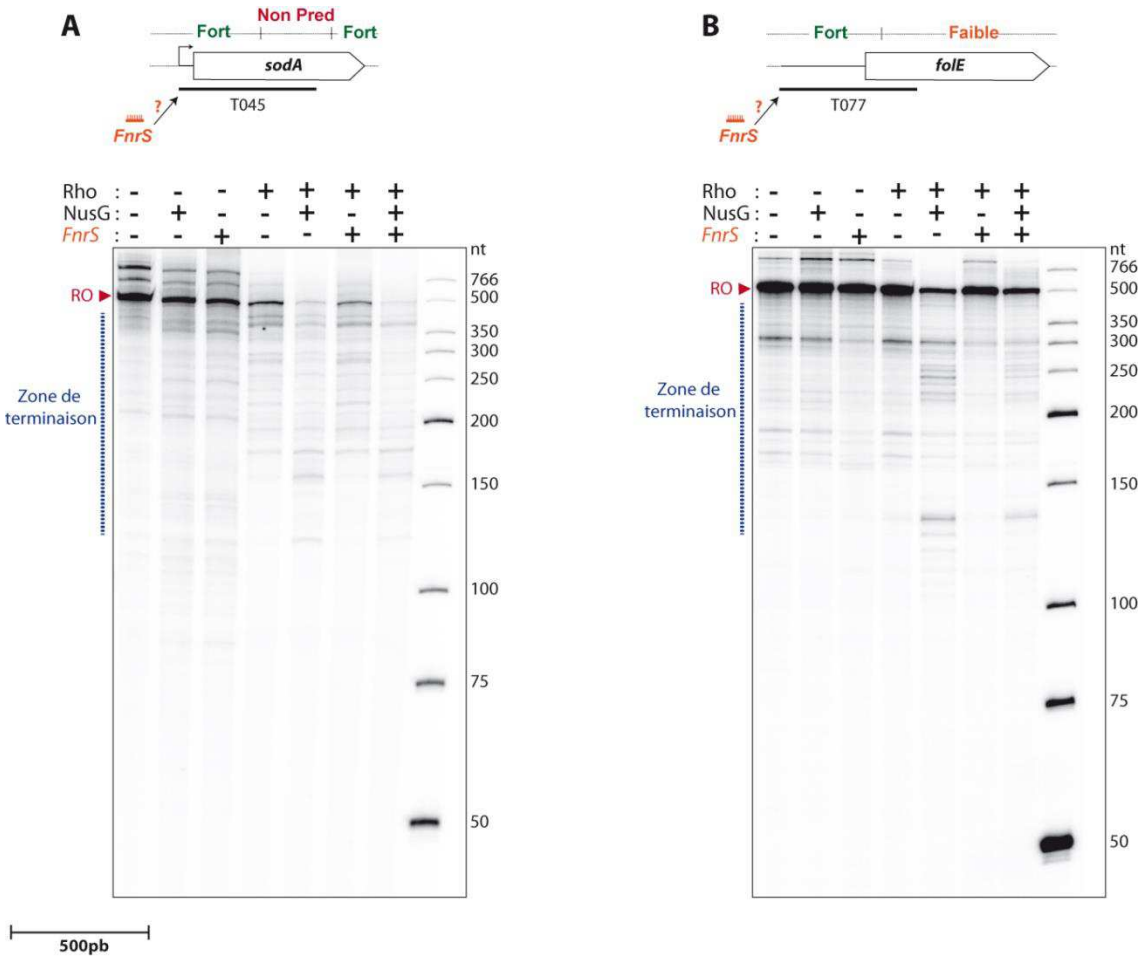


Figure 60 : Effet de *FnrS* sur les terminateurs Rho-dépendants des gènes *sodA* (A) et *acs* (B). Le sRNA est présent à une concentration de 0 ou 500 nM suivant les échantillons.

Je me suis aussi penché sur le cas des gènes *sodA* et *folE* dont la traduction est régulée par le sRNA *FnrS* (Boysen et al., 2010) et qui contiennent au moins un terminateur Rho-dépendant dans leur partie amont ou dans la région 5'UTR (**Figure 60**) (Nadiras et al., 2018b). Les expériences préliminaires que j'ai réalisées suggèrent que la présence de *FnrS* n'affecte pas la terminaison Rho-dépendante dans les deux cas (**Figure 60**). J'ai également détecté des signaux de terminaison Rho-dépendante pour d'autres gènes (*acnB*, *metE*, *phoP*, *nirB*, *cycA*) (**Figure 61**) connus pour être régulés par des sRNA (*GcvB*, *MicA*, *FnrS*, *RyhB*, *SgrS*) ou des protéines (AcnB) (**Tableau 4**), mais je n'ai pas eu le temps de tester si cette régulation et l'activité de Rho était liée dans ces cas.

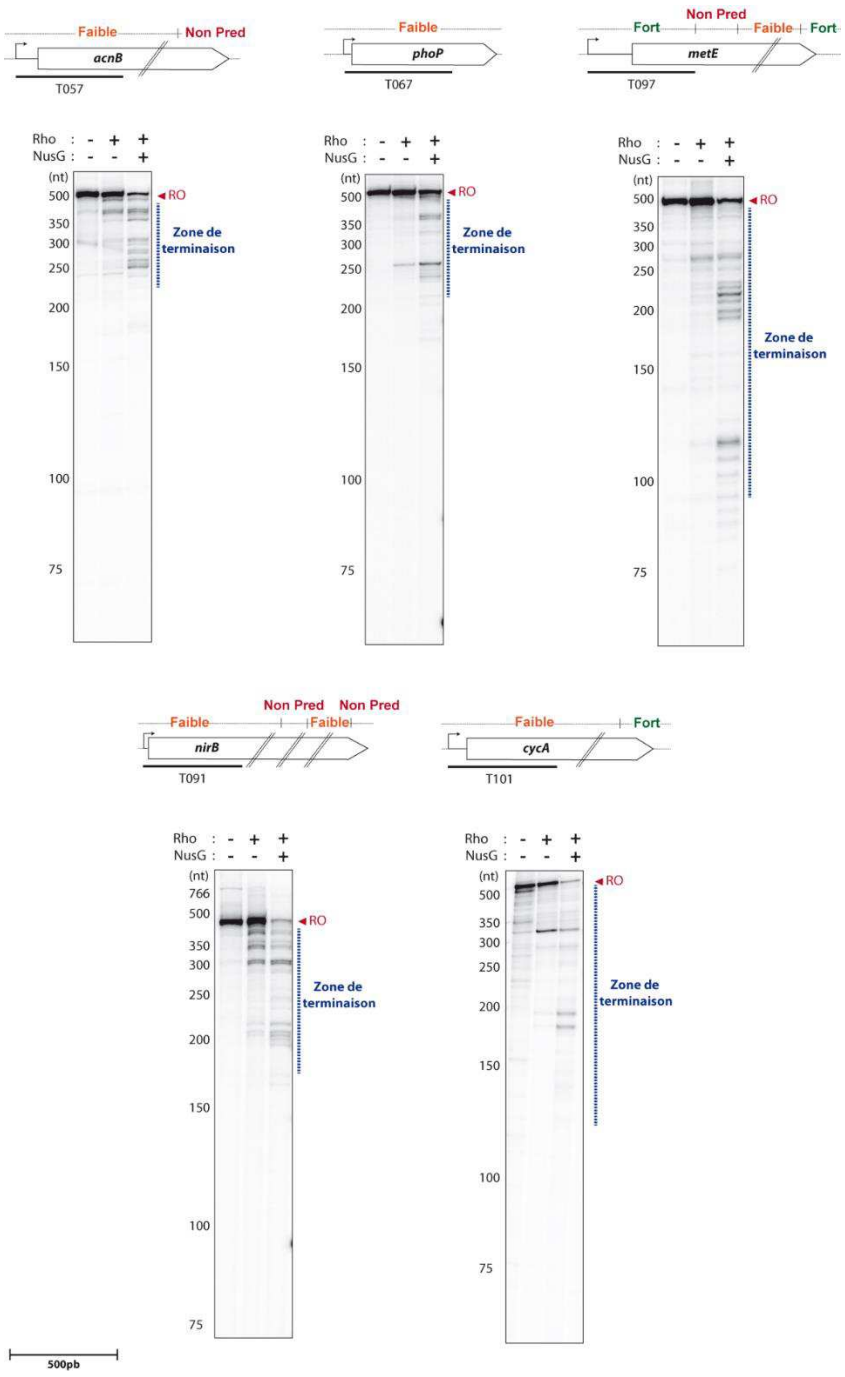


Figure 61 : Identification de signaux de terminaison Rho-dépendante pour les genes *acnB*, *phoP*, *metE*, *nirB* et *cycA*.

Enfin, je me suis intéressé au gène *cspA* qui code pour une ARN chaperonne (CspA) contribuant à l'adaptation thermique (choc froid 37°C → 10°C par exemple) (Goldstein et al., 1990). Les niveaux d'expression de *cspA* varient fortement avec la température (Giuliodori et al., 2010). Une étude a montré que la région 5'leader de l'ARNm *cspA* adopte des structures distinctes à faible (18°C) et haute (37°C) températures et qu'à la manière d'un riboswitch thermosensible, la structure « basse température » favorise l'accès au RBS et la traduction

(**Figure 62A**) (Giuliodori et al., 2010). Il apparaît aussi que CspA autorégule son expression par atténuation de la transcription en se fixant à l'ARNm au niveau d'une séquence « *cold box* » (Bae et al., 1997). L'hypothèse que j'ai exploré est un lien possible entre ces phénomènes avérés (changement de structure secondaire, interaction avec CspA) et la présence d'au moins un signal Rho-dépendant putatif dans la même région (ce signal étant prédit par notre modèle OPLS-DA ; cf. article I, page 69) (**Figure 62B**, diagramme).

J'ai tout d'abord vérifié par transcription in vitro standard à 37°C que ce signal existait bien, même s'il est relativement faible et est largement stimulé par NusG (**Figure 62B**).

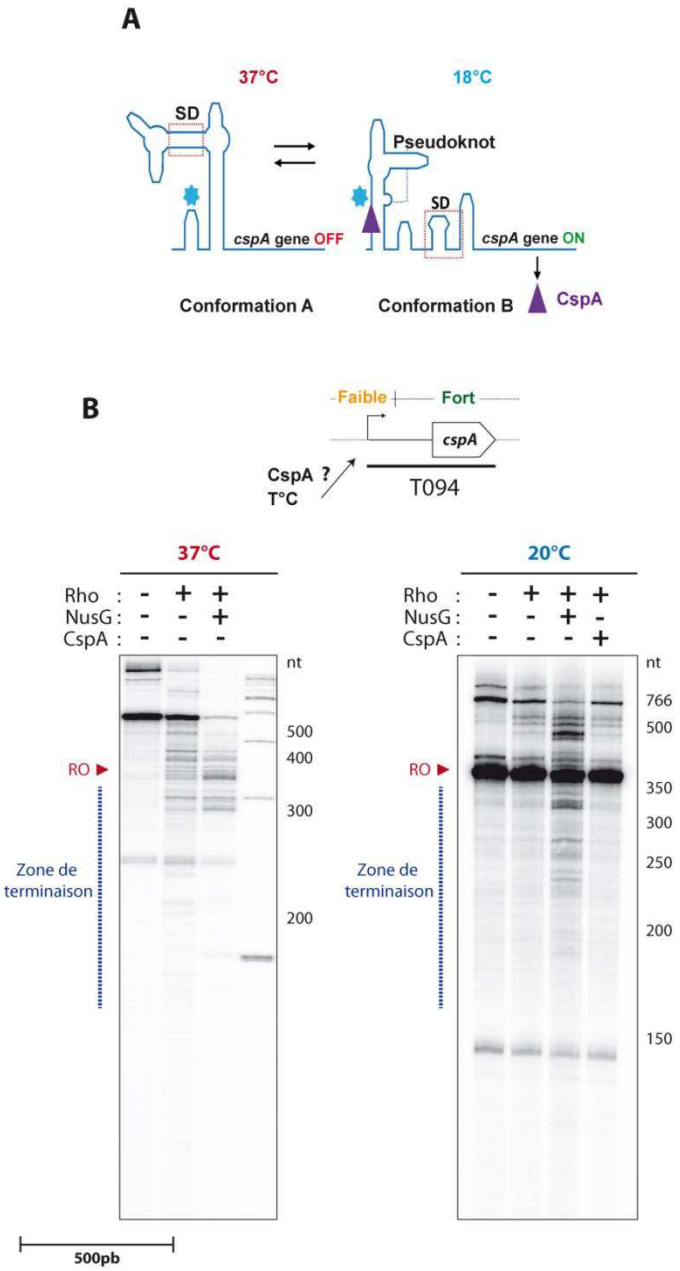


Figure 62 : Signal Rho-dépendant détecté *in vitro* pour le gène *cspA d'E. coli*. (A) La région 5'leader de *cspA* adopte des structures secondaires alternatives suivant la température. Le motif « *cold box* » est signalé par l'étoile bleue (✸) et la protéine CspA par un triangle violet (▲). Figure issue de (Guijarro et al., 2015). **(B)** Signaux Rho-dépendants prédits par notre modèle (diagramme) et analysés par transcription *in vitro*. CspA est présente à une concentration de 0 ou 250 nM suivant les échantillons.

Ce signal Rho-dépendant est d'ailleurs encore plus faible lorsque l'expérience est conduite à 20°C (**Figure 62B**), ce qui pourrait contribuer à la surexpression de CspA à cette température (Goldenberg et al., 1996). Cet effet de la température sur la terminaison Rho-dépendante n'est pas observé avec le terminateur témoin *λtr1* (données non présentées), suggérant qu'il est bel et bien lié à une restructuration de la région 5'-leader de *cspA*. Je n'ai pas détecté d'effet direct de la protéine CspA dans ce contexte (**Figure 62B**) même si des

expériences de contrôle additionnelles seraient nécessaires pour confirmer ce point. En résumé, mes données préliminaires suggèrent que la terminaison Rho-dépendante pourrait contribuer à réguler l'expression de *cspA* en fonction de la température. Cette contribution semble néanmoins modeste *in vitro* et mériterait d'être confirmée par des expériences menées *in vivo* dans des conditions de choc froid.

D. Conclusions & Perspectives

Mes travaux de thèse ont été consacrés à l'étude de la terminaison de la transcription Rho-dépendante, un mécanisme de régulation majeur et spécifique aux bactéries. J'ai utilisé essentiellement des approches biochimiques, complétées d'analyses bio-informatiques réalisées dans l'équipe par le Dr. Eric Eveno, pour tenter de mieux cerner les éléments de séquence ADN (ou ARN) qui, au sein du génome (ou transcriptome) gouvernent l'action du facteur Rho. Nos résultats confirment qu'il n'existe probablement pas de séquence ADN (ou ARN) consensus « simple » caractérisant les terminateurs Rho-dépendants. Néanmoins, nous avons pu mettre en évidence un ensemble de « descripteurs » de séquence qui, pris collectivement, permettent de prédire la propension d'une séquence ADN à induire la terminaison Rho-dépendante. Cette capacité prédictive du groupe de descripteurs a été mise à l'épreuve à grande échelle pour rechercher les zones potentiellement favorables à la terminaison Rho-dépendante au sein des génomes d'*E. coli* MG1655 et de *Salmonella* LT2. Les résultats obtenus mettent en lumière une grande proportion de ces zones favorables à l'action de Rho à l'échelle génomique (**Figure 63**). Ainsi, nos données confortent l'idée que les terminateurs Rho-dépendants sont construits à partir de règles suffisamment laxes pour autoriser l'intervention de Rho à de très nombreux points du génome, probablement sous le contrôle d'autres facteurs (découplage transcription-traduction, CET en mode « arrêt », formation d'une structure « R-loop », repliements alternatifs du transcrit, etc.).

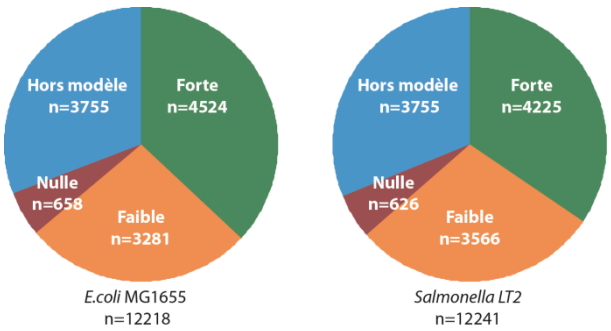


Figure 63 : Répartition des prédictions de terminaison Rho-dépendante par catégories. Les quantités respectives de régions pour lesquelles la terminaison est prédite « forte », « faible », « nulle » ou ne peut être prédite correctement (« hors modèle ») (Nadiras et al., 2018a) sont indiquées.

Notre modèle fournit des prédictions crédibles pour environ 65 % du génome d'*E. coli* ou de *Salmonella* mais n'est pas compatible avec les séquences contenues dans la fraction restante (35 %) (Nadiras et al., 2018a). Cette incompatibilité est probablement liée à la taille de l'échantillon de matrices ADN caractérisées expérimentalement et utilisées pour bâtir le modèle (104 « *training* » matrices) qui reste trop réduite pour représenter exhaustivement toutes les séquences génomiques possibles. Cette faiblesse du modèle pourrait sans doute être corrigée par la caractérisation et l'inclusion de nouvelles séquences ADN choisies au sein des régions « hors-modèle »

(**Figure 63**). Cette tâche pourrait être facilitée et accélérée par l'utilisation de l'essai fluorogénique de terminaison que j'ai développé (Section 121). De cette manière, on peut espérer augmenter le pouvoir de prédiction du modèle et offrir une cartographie plus complète pour des études d'ontologie. Ce genre d'étude pourrait, à son tour, permettre de déterminer si certaines catégories de fonctions (ou niveaux d'expression de gènes/opérons) sont plus particulièrement associées (ou non) à la terminaison Rho-dépendante.

Une analyse ontologique des régions prédites par notre modèle (**Figure 63**) pourrait peut-être également faciliter l'identification de nouveaux mécanismes/circuits de régulation conditionnelle impliquant le facteur Rho. Au cours de ma thèse, je me suis concentré sur les régions 5'UTR pour rechercher et tester des candidats sujets à une telle régulation (**Tableau 4**), une démarche qui a également été suivie par d'autres équipes (Bastet et al., 2017; Brandis et al., 2016; Chauvier et al., 2017; Gall et al., 2018; Kriner and Groisman, 2015; Sedlyarova et al., 2017; Sevostyanova and Groisman, 2015). Néanmoins, une étude récente a identifié un mécanisme de ce type dans une région intergénique de l'opéron *galIETKM* (Wang et al., 2015), suggérant que la régulation conditionnelle Rho-dépendante reste largement sous-estimée. Dans ce contexte, notre nouvel essai fluorogénique pourrait également se montrer utile comme outil de première intention pour sonder de nouvelles séquences génomiques.

Parmi les candidats sujets à régulation conditionnelle que j'ai testés au cours de ma thèse (**Tableau 4**), le cas de la région 5'UTR du gène *rpoS* est sans doute le plus emblématique. Le gène *rpoS* encode le facteur sigma alternatif σ^S, un facteur général de réponse au stress soumis à une régulation complexe et multifactorielle (**Figure 64**).

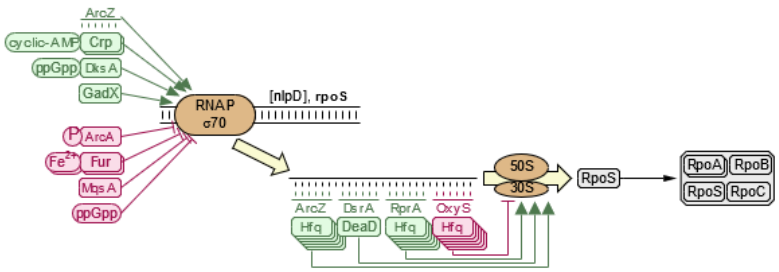


Figure 63 : Diagramme résumant la régulation complexe du gène *rpoS* d'*E. coli*. Extrait de la base de données EcoCyc (https://ecocyc.org).

En découvrant un terminateur Rho-dépendant dans la région 5'UTR de *rpoS*, qui peut être inactivé par les sRNA *ArcZ*, *DsrA*, *RprA*, une équipe concurrente (Sedlyarova et al., 2016) et la nôtre (Nadiras et al., 2018b) ont révélé une couche supplémentaire de complexité pour ce système hautement régulé. L'analyse des autres candidats potentiels (**Tableau 4**) n'a malheureusement pas révélé d'éléments de régulation conditionnelle aussi marquants. Le cas de la région 5'UTR de *cspA* reste néanmoins intriguant. Mes données suggèrent que cette région contient un terminateur Rho-dépendant faible qui peut être stimulé par NusG (**Figure 62** et texte page 177). L'activité de ce terminateur est diminuée *in vitro* lorsque la température d'incubation est abaissée à 20°C (**Figure 62**). Bien que de nombreux tests de vérification *in vitro* et *in vivo* restent à faire, ces données préliminaires suggèrent que Rho pourrait être impliqué dans le mécanisme de régulation de l'expression de *cspA* en réponse à un « choc froid » (**Figure 65**).

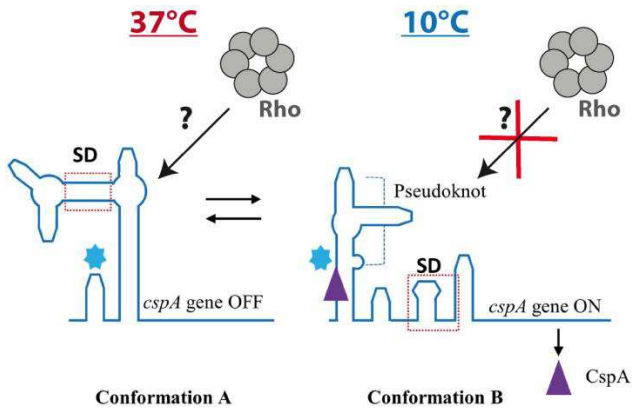


Figure 65 : Mécanisme possible impliquant Rho dans la régulation de *cspA* en fonction de la température. La séquence « *cold box* » est indiquée par l'étoile bleue (✸) et la protéine CspA représentée par un triangle violet (▲). Figure inspirée de (Guijarro et al., 2015).

L'adaptation de la stratégie Clip-seq à l'étude de Rho pourrait peut-être également permettre d'identifier de nouveaux terminateurs Rho-dépendants et sites potentiels de régulation conditionnelle. De plus, cette méthode permettrait de mieux localiser les terminateurs et sites *Rut* au sein des génomes car, à l'inverse de l'approche RNAseq utilisée jusqu'à présent pour caractériser les transcriptomes Rho-dépendants (Holmqvist et al., 2016) elle n'est pas sensible aux modifications post-transcriptionnelles des transcrits par les exonucléases. Bien qu'inspirées de protocoles ayant fait leur preuve (Holmqvist et al., 2018; Holmqvist et al., 2016; Potts et al., 2017), mes tentatives d'adaptation de l'approche Clip-seq à l'étude de Rho chez *Salmonella* (page 140) se sont heurtées à deux problèmes majeurs que

je n'ai pu résoudre dans la durée de ma thèse : D'une part, les ARN récupérés étaient relativement de petites tailles (≈25 nt) (**Figure 56C**), suggérant une activité « nucléase » anormalement élevée dans les souches utilisées et/ou mal contrôlée lors des étapes de lyse et d'immunoprécipitation. D'autre part, la quantité d'ARN récupérée après immunoprécipitation et purification par PAGE et transfert sur membrane s'est avérée beaucoup plus faible (~3 ng/échantillon) que recommandée pour la préparation des banques d'ADNc pour le séquençage NGS (> 20 ng). Ceci contribue vraisemblablement à expliquer pourquoi les résultats du séquençage indiquent une contamination de type « microbiome » des échantillons. Ces éléments démontrent que de nombreux efforts d'optimisation restent à faire pour adapter convenablement l'approche Clip-seq à l'étude de Rho. Une piste possible est l'utilisation d'une variante sophistiquée appelée iCLIP-seq (*Individual-nucleotide resolution UV crosslinking and immunoprecipitation coupled to high-throughput sequencing*) (Huppertz et al., 2014) qui limite la perte de matériel lors de la préparation des banques d'ADNc. Cette perte est notamment due à la faible efficacité des transcriptases inverses à synthétiser le brin ADNc au-delà du site de photo-pontage. Ce problème est résolu en partie dans la variante iCLIP-seq par l'introduction du segment « adapteur » amont requis pour le séquençage après l'étape de transcription inverse (plutôt qu'avant cette étape dans le protocole classique) grâce à une étape de circularisation/clivage (**Figure 66**).

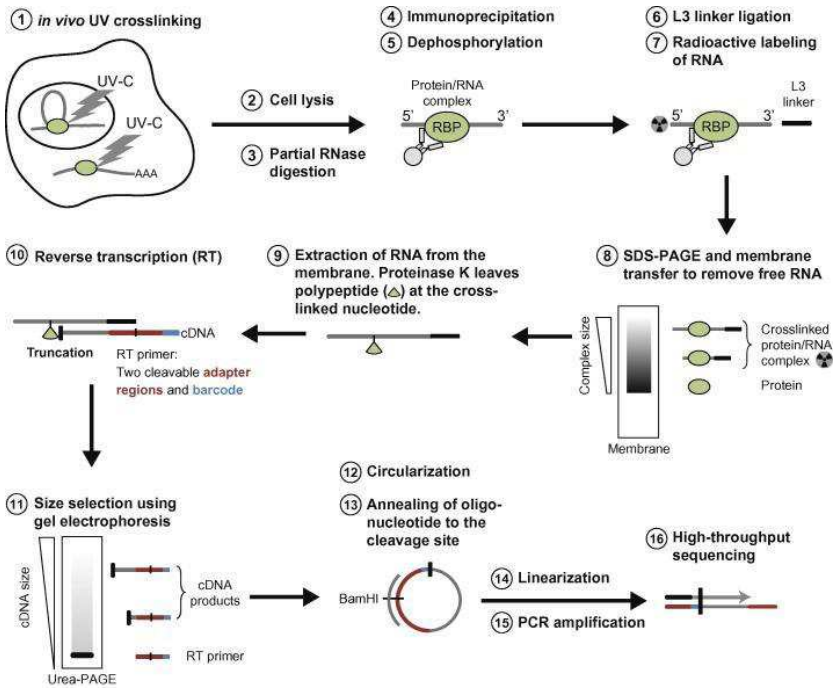


Figure 66 : Schéma de principe du protocole de ¡CLIP-seq. Figure issue de (Huppertz et al., 2014).

Pour finir, je voudrais souligner que mes travaux ont porté sur l'étude de la terminaison Rho-dépendante chez *E. coli* et *Salmonella* et que leur transposition au reste du règne bactérien devrait être menée avec prudence. Ainsi, si notre modèle de prédiction est probablement utilisable pour explorer le génome de bactéries proches (γ-Protéobactéries) et/ou contenant une machinerie de transcription et un facteur Rho similaires (D'Heygere et al., 2013), il n'est probablement pas valide dans les autres cas. Néanmoins, il est probable qu'une stratégie similaire de modélisation de la terminaison Rho-dépendante puisse être menée chez ces espèces philodivergentes, à condition que leur machinerie de transcription (Rho compris) soit purifiable et utilisable *in vitro* pour bâtir une base de données expérimentale fiable. Ce travail de longue haleine permettrait d'apprécier les différentes formes et fonctions assumées par la régulation Rho-dépendante au travers de l'ensemble du règne bactérien.

E. Annexe

Tableau supplémentaire 1: Matrice ADN

[a]: Séquence insérée entre les séquences cibles MBP dans la cassette reportrice (voir le texte principal).
[b]: Séquences génomiques de référence U00096.3 pour *Escherichia coli* MG1655, NC_003197.1 pour *Salmonella* LT2 et NC_001416.1 pour le phage lambda.
[c]: coordonnées du génome de la séquence insérée.
[d]: amorce utilisée dans un premier cycle de PCR pour amplifier la région d'intérêt avec une séquence promotrice **partielle** du T7A1; la séquence T7A1 complète a été introduite dans un second cycle de PCR en utilisant l'amorce: 5'-TTATCAAAAAGAGTATTGACTTAAAGTCTAACCTATAGGATACTTACAGCC-3' pour les séquences cibles MBP référencé «aucune» en amont, et 5'-TTATCAAAAAGAGTATTGACTTAAAGTCTAACCTATAGGATACTTACAGCCCGCAAAAA-3' pour les séquences cibles référencé « antiMBPT » ou « antiMBPC » en amont.

Nom de la matrice	Séquence génomique [a]				Séquence cible MBP		Longueur de la matrice (bp)	Oligonucléotide « forward »1[d]	Oligonucléotide « reverse »
	Génome [b]	Sens	Début [c]	Fin [c]	En amont	En aval			
$\lambda tR1_0$	Phage λ	+	38023	38406	Aucune	Aucune	435	GATACTTACAGCCATGTACTAAGGAGGTTGTATGG	GCACTCTCGATTCGTAGAGCC
$\lambda tR1_C$	Phage λ	+	38023	38406	Aucune	antiMBP$_C$	453	GATACTTACAGCCATGTACTAAGGAGGTTGTATGG	CGCTTTTTTTTTTTGCGGCACTCTCGATTCGTAGAGCCTCGTTGC
$\lambda tR1_T$	Phage λ	+	38023	38406	Aucune	antiMBP$_T$	453	GATACTTACAGCCATGTACTAAGGAGGTTGTATGG	CGCTTGTTAATTTTTGCGGCACTCTCGATTCGTAGAGCCTCGTTGC
$\lambda tR1_{C/T}$	phage λ	+	38023	38406	antiMBP$_C$	antiMBP$_T$	471	GATACTTACAGCCCGCAAAAAAAAAAAGCGATGTACTAAGGAGGTTGTATGG	CGCTTGTTAATTTTTGCGGCACTCTCGATTCGTAGAGCCTCGTTGC
$\lambda tR1_{T/C}$	Phage λ	+	38023	38406	antiMBP$_T$	antiMBP$_C$	471	GATACTTACAGCCCGCAAAATTAACAAGCGATGTACTAAGGAGGTTGTATGG	CGCTTTTTTTTTTTGCGGCACTCTCGATTCGTAGAGCCTCGTTGC
$rrsA_0$	MG1655	+	4036001	4036500	Aucune	Aucune	551	GTCTAACCTATAGGATACTTACAGCCTTTGCTCATTGACGTTACCCGCAGAAG	GTTGCATCGAATTAAACCACATGCTCC
$rrsA_C$ (matrice R)	MG1655	+	4036001	4036500	Aucune	antiMBP$_C$	569	GTCTAACCTATAGGATACTTACAGCCTTTGCTCATTGACGTTACCCGCAGAAG	CGCTTTTTTTTTTTGCGGTTGCATCGAATTAAACCACATGCTCC
$rrsA_T$	MG1655	+	4036001	4036500	Aucune	antiMBP$_T$	569	GTCTAACCTATAGGATACTTACAGCCTTTGCTCATTGACGTTACCCGCAGAAG	CGCTTGTTAATTTTTGCGGTTGCATCGAATTAAACCACATGCTCC
$rrsA_{C/T}$	MG1655	+	4036001	4036500	antiMBP$_C$	antiMBP$_T$	587	GATACTTACAGCCCGCAAAAAAAAAAAGCGTTTGCTCATTGACGTTACCCGC	CGCTTGTTAATTTTTGCGGTTGCATCGAATTAAACCACATGCTCC
acs_0	MG1655	-	4287096	4287595	Aucune	Aucune	551	GTCTAACCTATAGGATACTTACAGCCAGAGTTAGTCAGTATCTTCCTCTTTTTCAACAG	ACGATCGCCGTTTTCTTGCAGATGGC
acs_{Cless}	MG1655	-	4287096	4287595	Aucune	Aucune	551	GTCTAACCTATAGGATACTTACAGCCATAGTTAGTTAGTATTTTCCTCTTTTTCAACAG	ACGATCGCCGTTTTCTTGCAGATGGC
acs_T	MG1655	-	4287096	4287595	Aucune	antiMBP$_T$	569	GTCTAACCTATAGGATACTTACAGCCAGAGTTAGTCAGTATCTTCCTCTTTTTCAACAG	CGCTTGTTAATTTTTGCGACGATCGCCGTTTTCTTGCAGATGGC
$acs_{C/T}$	MG1655	-	4287096	4287595	antiMBP$_C$	antiMBP$_T$	587	GATACTTACAGCCCGCAAAAAAAAAAAGCGAGAGTTAGTCAGTATCTTCCTCTTTTTC	CGCTTGTTAATTTTTGCGACGATCGCCGTTTTCTTGCAGATGGC
$chiP_0$	LT2	+	749445	750006	Aucune	Aucune	611	GTCTAACCTATAGGATACTTACAGCCGTAGTAAATGAGGTTTCTCAGCGCTAC	GTCTGACCGGTAGGCTGAAT
$chiP_{Cless}$	LT2	+	749445	750006	Aucune	Aucune	611	GTCTAACCTATAGGATACTTACAGCCATAGTAAATGAGGTTTTTCAGCGCTAC	GTCTGACCGGTAGGCTGAAT
$chiP_T$	LT2	+	749445	750006	Aucune	antiMBP$_T$	629	GTCTAACCTATAGGATACTTACAGCCGTAGTAAATGAGGTTTCTCAGCGCTAC	CGCTTGTTAATTTTTGCGGTCTGACCGGTAGGCTGAATATAACCGGCG
$chiP_{C/T}$	LT2	+	749445	750006	antiMBP$_C$	antiMBP$_T$	647	GATACTTACAGCCCGCAAAAAAAAAAAGCGGTAGTAAATGAGGTTTCTCAGCGCTAC	CGCTTGTTAATTTTTGCGGTCTGACCGGTAGGCTGAATATAACCGGCG
$corA_0$	LT2	+	4157936	4158382	Aucune	Aucune	498	GTCTAACCTATAGGATACTTACAGCCCACGATTACGCTAAGAGGACATTCGCC	CTTCGTCTTCAAAGAAGCGCGCGGATGC
$corA_{Cless}$	LT2	+	4157936	4158382	Aucune	Aucune	498	GTCTAACCTATAGGATACTTACAGCCATAGATTATGTTAAGAGGACATTCGCCTTGG	CTTCGTCTTCAAAGAAGCGCGCGGATGC
$corA_T$	LT2	+	4157936	4158382	Aucune	antiMBP$_T$	516	GTCTAACCTATAGGATACTTACAGCCCACGATTACGCTAAGAGGACATTCGCC	CGCTTGTTAATTTTTGCGCTTCGTCTTCAAAGAAGCGCGCGGATGC
$dmsA_0$	MG1655	+	940823	941322	Aucune	Aucune	551	GTCTAACCTATAGGATACTTACAGCCGAGTGAAAATCTACCTATCTCTTTG	ATTTCAGGCGGTCCGGATTGTAGACAC

***dmsA*$_{Cless}$**	MG1655	+	940823	941322	Aucune	Aucune	551	GTCTAACCTATAGGATA CTTACAGCCATGTGAAA ATGTATTTATGTCTTTG	ATTTCAGGCGGTCCGGA TTGTAGACAC
***dmsA*$_{C}$**	MG1655	+	940823	941322	Aucune	antiMBP$_{C}$	569	GTCTAACCTATAGGATA CTTACAGCCGAGTGAAA ATCTACCTATCTCTTTG	CGCTTTTTTTTTTTGCG ATTTCAGGCGGTCCGGA TTGTAGACAC
***dmsA*$_{T}$**	MG1655	+	940823	941322	Aucune	antiMBP$_{T}$	569	GTCTAACCTATAGGATA CTTACAGCCGAGTGAAA ATCTACCTATCTCTTTG	CGCTTGTTAATTTTTGCG ATTTCAGGCGGTCCGGA TTGTAGACAC
***rrlD*$_{0}$**	MG1655	-	3425268	3425840	Aucune	Aucune	624	GTCTAACCTATAGGATA CTTACAGCCCACACGGC GGGTGCTAACGTCCGTC G	CTCAGCCTTGATTTTCCG GATTTGCC
***rrlD*$_{T}$**	MG1655	-	3425268	3425840	Aucune	antiMBP$_{T}$	642	GTCTAACCTATAGGATA CTTACAGCCCACACGGC GGGTGCTAACGTCCGTC G	CGCTTGTTAATTTTTGCG CTCAGCCTTGATTTTCCG GATTTGCC

F. Bibliographie

Alifano, P., Rivellini, F., Limauro, D., Bruni, C.B., and Carlomagno, M.S. (1991). A consensus motif common to all Rho-dependent prokaryotic transcription terminators. Cell *64*, 553-563.

Arluison, V., Hohng, S., Roy, R., Pellegrini, O., Regnier, P., and Ha, T. (2007). Spectroscopic observation of RNA chaperone activities of Hfq in post-transcriptional regulation by a small non-coding RNA. Nucleic Acids Res *35*, 999-1006.

Artsimovitch, I., and Henkin, T.M. (2009). In vitro approaches to analysis of transcription termination. Methods *47*, 37-43.

Artsimovitch, I., and Landick, R. (2000). Pausing by bacterial RNA polymerase is mediated by mechanistically distinct classes of signals. Proc Natl Acad Sci USA *97*, 7090-7095.

Bae, B., Feklistov, A., Lass-Napiorkowska, A., Landick, R., and Darst, S.A. (2015). Structure of a bacterial RNA polymerase holoenzyme open promoter complex. eLife *4*.

Bae, W., Jones, P.G., and Inouye, M. (1997). CspA, the major cold shock protein of Escherichia coli, negatively regulates its own gene expression. J Bacteriol *179*, 7081-7088.

Bastet, L., Chauvier, A., Singh, N., Lussier, A., Lamontagne, A.M., Prevost, K., Masse, E., Wade, J.T., and Lafontaine, D.A. (2017). Translational control and Rho-dependent transcription termination are intimately linked in riboswitch regulation. Nucleic Acids Res *45*, 7474-7486.

Bastet, L., Turcotte, P., Wade, J.T., and Lafontaine, D.A. (2018). Maestro of regulation: Riboswitches orchestrate gene expression at the levels of translation, transcription and mRNA decay. RNA biology, 1-4.

Belogurov, G.A., Mooney, R.A., Svetlov, V., Landick, R., and Artsimovitch, I. (2009). Functional specialization of transcription elongation factors. EMBO J *28*, 112-122.

Bidnenko, V., Nicolas, P., Grylak-Mielnicka, A., Delumeau, O., Auger, S., Aucouturier, A., Guerin, C., Repoila, F., Bardowski, J., Aymerich, S., *et al.* (2017). Termination factor Rho: From the control of pervasive transcription to cell fate determination in Bacillus subtilis. PLoS Genet *13*, e1006909.

Bogden, C.E., Fass, D., Bergman, N., Nichols, M.D., and Berger, J.M. (1999). The structural basis for terminator recognition by the Rho transcription termination factor. Mol Cell *3*, 487-493.

Borukhov, S., Lee, J., and Laptenko, O. (2005). Bacterial transcription elongation factors: new insights into molecular mechanism of action. Mol Microbiol *55*, 1315-1324.

Bossi, L., Schwartz, A., Guillemardet, B., Boudvillain, M., and Figueroa-Bossi, N. (2012). A role for Rho-dependent polarity in gene regulation by a noncoding small RNA. Genes Dev *26*, 1864-1873.

Botella, L., Vaubourgeix, J., Livny, J., and Schnappinger, D. (2017). Depleting Mycobacterium tuberculosis of the transcription termination factor Rho causes pervasive transcription and rapid death. Nature communications *8*, 14731.

Boudvillain, M., Figueroa-Bossi, N., and Bossi, L. (2013). Terminator still moving forward: expanding roles for Rho factor. Curr Opin Microbiol *16*, 118-124.

Boudvillain, M., Nollmann, M., and Margeat, E. (2010a). Keeping up to speed with the transcription termination factor Rho motor. Transcription *1*, 70-75.

Boudvillain, M., Walmacq, C., Schwartz, A., and Jacquinot, F. (2010b). Simple enzymatic assays for the in vitro motor activity of transcription termination factor Rho from Escherichia coli. Methods Mol Biol *587*, 137-154.

Boysen, A., Moller-Jensen, J., Kallipolitis, B., Valentin-Hansen, P., and Overgaard, M. (2010). Translational regulation of gene expression by an anaerobically induced small non-coding RNA in Escherichia coli. J Biol Chem *285*, 10690-10702.

Brandis, G., Bergman, J.M., and Hughes, D. (2016). Autoregulation of the tufB operon in Salmonella. Mol Microbiol *100*, 1004-1016.

Brennan, C.A., Dombroski, A.J., and Platt, T. (1987). Transcription termination factor rho is an RNA-DNA helicase. Cell *48*, 945-952.

Brennan, C.A., Steinmetz, E.J., Spear, P., and Platt, T. (1990). Specificity and efficiency of rho-factor helicase activity depends on magnesium concentration and energy coupling to NTP hydrolysis. J Biol Chem *265*, 5440-5447.

Brennan, R.G., and Link, T.M. (2007). Hfq structure, function and ligand binding. Curr Opin Microbiol *10*, 125-133.

Briani, F., Ghisotti, D., and Deho, G. (2000). Antisense RNA-dependent transcription termination sites that modulate lysogenic development of satellite phage P4. Mol Microbiol *36*, 1124-1134.

Browning, D.F., and Busby, S.J. (2016). Local and global regulation of transcription initiation in bacteria. Nature reviews Microbiology *14*, 638-650.

Bubunenko, M., Baker, T., and Court, D.L. (2007). Essentiality of ribosomal and transcription antitermination proteins analyzed by systematic gene replacement in Escherichia coli. J Bacteriol *189*, 2844-2853.

Burgess, R.R., and Travers, A.A. (1970). Escherichia coli RNA polymerase: purification, subunit structure, and factor requirements. Federation proceedings *29*, 1164-1169.

Burmann, B.M., Knauer, S.H., Sevostyanova, A., Schweimer, K., Mooney, R.A., Landick, R., Artsimovitch, I., and Rosch, P. (2012). An alpha helix to beta barrel domain switch transforms the transcription factor RfaH into a translation factor. Cell *150*, 291-303.

Burmann, B.M., Schweimer, K., Luo, X., Wahl, M.C., Stitt, B.L., Gottesman, M.E., and Rosch, P. (2010). A NusE:NusG complex links transcription and translation. Science *328*, 501-504.

Burns, C.M., and Richardson, J.P. (1995). NusG is required to overcome a kinetic limitation to Rho function at an intragenic terminator. Proc Natl Acad Sci U S A *92*, 4738-4742.

Burova, E., Hung, S.C., Sagitov, V., Stitt, B.L., and Gottesman, M.E. (1995). Escherichia coli NusG protein stimulates transcription elongation rates in vivo and in vitro. J Bacteriol *177*, 1388-1392.

Campagne, S., Marsh, M.E., Capitani, G., Vorholt, J.A., and Allain, F.H. (2014). Structural basis for -10 promoter element melting by environmentally induced sigma factors. Nat Struct Mol Biol *21*, 269-276.

Campbell, E.A., Muzzin, O., Chlenov, M., Sun, J.L., Olson, C.A., Weinman, O., Trester-Zedlitz, M.L., and Darst, S.A. (2002). Structure of the bacterial RNA polymerase promoter specificity sigma subunit. Mol Cell *9*, 527-539.

Canals, A., Uson, I., and Coll, M. (2010). The structure of RNA-free Rho termination factor indicates a dynamic mechanism of transcript capture. J Mol Biol *400*, 16-23.

Cardinale, C.J., Washburn, R.S., Tadigotla, V.R., Brown, L.M., Gottesman, M.E., and Nudler, E. (2008a). Termination factor Rho and its cofactors NusA and NusG silence foreign DNA in E. coli. Science *320*, 935-938.

Cardinale, C.J., Washburn, R.S., Tadigotla, V.R., Brown, L.M., Gottesman, M.E., and Nudler, E. (2008b). Termination Factor Rho and Its Cofactors NusA and NusG Silence Foreign DNA in E. coli. Science *320*, 935-938.

Carlomagno, M.S., and Nappo, A. (2003). NusA modulates intragenic termination by different pathways. Gene *308*, 115-128.

Chakraborty, A., Wang, D., Ebright, Y.W., Korlann, Y., Kortkhonjia, E., Kim, T., Chowdhury, S., Wigneshweraraj, S., Irschik, H., Jansen, R., *et al.* (2012). Opening and closing of the bacterial RNA polymerase clamp. Science *337*, 591-595.

Chao, Y., Papenfort, K., Reinhardt, R., Sharma, C.M., and Vogel, J. (2012). An atlas of Hfq-bound transcripts reveals 3' UTRs as a genomic reservoir of regulatory small RNAs. EMBO J *31*, 4005-4019.

Chauvier, A., Picard-Jean, F., Berger-Dancause, J.C., Bastet, L., Naghdi, M.R., Dube, A., Turcotte, P., Perreault, J., and Lafontaine, D.A. (2017). Transcriptional pausing at the translation start site operates as a critical checkpoint for riboswitch regulation. Nature communications *8*, 13892.

Chen, B., Zhang, T., Bond, T., and Gan, Y. (2015). Development of quantitative structure activity relationship (QSAR) model for disinfection byproduct (DBP) research: A review of methods and resources. Journal of hazardous materials *299*, 260-279.

Chen, C.Y., and Richardson, J.P. (1987). Sequence elements essential for rho-dependent transcription termination at lambda tR1. J Biol Chem *262*, 11292-11299.

Chinnappan, R., Dube, A., Lemay, J.F., and Lafontaine, D.A. (2013). Fluorescence monitoring of riboswitch transcription regulation using a dual molecular beacon assay. Nucleic Acids Res *41*, e106.

Ciampi, M.S. (2006). Rho-dependent terminators and transcription termination. Microbiology *152*, 2515-2528.

Ciampi, M.S., Alifano, P., Nappo, A.G., Bruni, C.B., and Carlomagno, M.S. (1989). Features of the rho-dependent transcription termination polar element within the hisG cistron of Salmonella typhimurium. J Bacteriol *171*, 4472-4478.

Ciampi, M.S., Schmid, M.B., and Roth, J.R. (1982). Transposon Tn10 provides a promoter for transcription of adjacent sequences. Proc Natl Acad Sci U S A *79*, 5016-5020.

Colonna, B., and Hofnung, M. (1981). rho Mutations restore lamB expression in E. coli K12 strains with an inactive malB region. Molecular & general genetics : MGG *184*, 479-483.

Condon, C., French, S., Squires, C., and Squires, C.L. (1993). Depletion of functional ribosomal RNA operons in Escherichia coli causes increased expression of the remaining intact copies. EMBO J *12*, 4305-4315.

Cook, H., and Ussery, D.W. (2013). Sigma factors in a thousand E. coli genomes. Environmental microbiology *15*, 3121-3129.

Cook, K.B., Hughes, T.R., and Morris, Q.D. (2015). High-throughput characterization of protein-RNA interactions. Briefings in functional genomics *14*, 74-89.

Corbino, K.A., Barrick, J.E., Lim, J., Welz, R., Tucker, B.J., Puskarz, I., Mandal, M., Rudnick, N.D., and Breaker, R.R. (2005). Evidence for a second class of S-adenosylmethionine riboswitches and other regulatory RNA motifs in alpha-proteobacteria. Genome Biol *6*, R70.

Czyz, A., Mooney, R.A., Iaconi, A., and Landick, R. (2014). Mycobacterial RNA polymerase requires a U-tract at intrinsic terminators and is aided by NusG at suboptimal terminators. mBio *5*, e00931.

D'Heygere, F., Rabhi, M., and Boudvillain, M. (2013). Phyletic distribution and conservation of the bacterial transcription termination factor Rho. Microbiology *159*, 1423-1436.

D'Heygere, F., Schwartz, A., Coste, F., Castaing, B., and Boudvillain, M. (2015). Monitoring RNA Unwinding by the Transcription Termination Factor Rho from Mycobacterium tuberculosis. Methods Mol Biol *1259*, 293-311.

Dann, C.E., 3rd, Wakeman, C.A., Sieling, C.L., Baker, S.C., Irnov, I., and Winkler, W.C. (2007). Structure and mechanism of a metal-sensing regulatory RNA. Cell *130*, 878-892.

Dar, D., and Sorek, R. (2018). High-resolution RNA 3'-ends mapping of bacterial Rho-dependent transcripts. Nucleic Acids Res.

Darst, S.A., Feklistov, A., and Gross, C.A. (2014). Promoter melting by an alternative sigma, one base at a time. Nat Struct Mol Biol *21*, 350-351.

Davis, M.C., Kesthely, C.A., Franklin, E.A., and MacLellan, S.R. (2017). The essential activities of the bacterial sigma factor. Canadian journal of microbiology *63*, 89-99.

De Crombrugghe, B., Adhya, S., Gottesman, M., and Pastan, I. (1973). Effect of Rho on transcription of bacterial operons. Nature: New biology *241*, 260-264.

de Hoon, M.J., Makita, Y., Nakai, K., and Miyano, S. (2005). Prediction of Transcriptional Terminators in Bacillus subtilis and Related Species. PLoS computational biology *1*, e25.

deHaseth, P.L., Zupancic, M.L., and Record, M.T., Jr. (1998). RNA polymerase-promoter interactions: the comings and goings of RNA polymerase. J Bacteriol *180*, 3019-3025.

Dutta, D., Shatalin, K., Epshtein, V., Gottesman, M.E., and Nudler, E. (2011). Linking RNA polymerase backtracking to genome instability in E. coli. Cell *146*, 533-543.

Edwards, A.N., Patterson-Fortin, L.M., Vakulskas, C.A., Mercante, J.W., Potrykus, K., Vinella, D., Camacho, M.I., Fields, J.A., Thompson, S.A., Georgellis, D., *et al.* (2011). Circuitry linking the Csr and stringent response global regulatory systems. Mol Microbiol *80*, 1561-1580.

Enemark, E.J., and Joshua-Tor, L. (2006). Mechanism of DNA translocation in a replicative hexameric helicase. Nature *442*, 270-275.

Epshtein, V., Dutta, D., Wade, J., and Nudler, E. (2010). An allosteric mechanism of Rho-dependent transcription termination. Nature *463*, 245-249.

Epshtein, V., Mironov, A.S., and Nudler, E. (2003). The riboswitch-mediated control of sulfur metabolism in bacteria. Proc Natl Acad Sci U S A *100*, 5052-5056.

Erie, D.A., Hajiseyedjavadi, O., Young, M.C., and von Hippel, P.H. (1993). Multiple RNA polymerase conformations and GreA: control of the fidelity of transcription. Science *262*, 867-873.

Esyunina, D., Agapov, A., and Kulbachinskiy, A. (2016). Regulation of transcriptional pausing through the secondary channel of RNA polymerase. Proc Natl Acad Sci U S A *113*, 8699-8704.

Faus, I., and Richardson, J.P. (1990). Structural and functional properties of the segments of lambda cro mRNA that interact with transcription termination factor Rho. J Mol Biol *212*, 53-66.

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

Forti, F., Polo, S., Lane, K.B., Six, E.W., Sironi, G., Deho, G., and Ghisotti, D. (1999). Translation of two nested genes in bacteriophage P4 controls immunity-specific transcription termination. J Bacteriol *181*, 5225-5233.

Freeman, W.H. (2012). Molecular Biology: Principles and Practic, Fifth Edition edn.

Friedman, D.I., and Baron, L.S. (1974). Genetic characterization of a bacterial locus involved in the activity of the N function of phage lambda. Virology *58*, 141-148.

Friedman, D.I., and Court, D.L. (1995). Transcription antitermination: the lambda paradigm updated. Mol Microbiol *18*, 191-200.

Gaal, T., Ross, W., Estrem, S.T., Nguyen, L.H., Burgess, R.R., and Gourse, R.L. (2001). Promoter recognition and discrimination by EsigmaS RNA polymerase. Mol Microbiol *42*, 939-954.

Gall, A.R., Datsenko, K.A., Figueroa-Bossi, N., Bossi, L., Masuda, I., Hou, Y.M., and Csonka, L.N. (2016). Mg2+ regulates transcription of mgtA in Salmonella Typhimurium via translation of proline codons during synthesis of the MgtL peptide. Proc Natl Acad Sci U S A *113*, 15096-15101.

Gall, A.R., Hegarty, A.E., Datsenko, K.A., Westerman, R.P., SanMiguel, P., and Csonka, L.N. (2018). High-level, constitutive expression of the mgtC gene confers increased thermotolerance on Salmonella enterica serovar Typhimurium. Mol Microbiol.

Geiselmann, J., Yager, T., Gill, S., Calmettes, P., and von Hippel, P. (1992). Physical properties of the Escherichia coli transcription termination factor rho. 1. Association states and geometry of the rho hexamer. Biochemistry *31*, 111-121.

Giuliodori, A.M., Di Pietro, F., Marzi, S., Masquida, B., Wagner, R., Romby, P., Gualerzi, C.O., and Pon, C.L. (2010). The cspA mRNA is a thermosensor that modulates translation of the cold-shock protein CspA. Mol Cell *37*, 21-33.

Gocheva, V., Le Gall, A., Boudvillain, M., Margeat, E., and Nollmann, M. (2015). Direct observation of the translocation mechanism of transcription termination factor Rho. Nucleic Acids Res *43*, 2367-2377.

Gogol, E.P., Seifried, S.E., and von Hippel, P.H. (1991). Structure and assembly of the Escherichia coli transcription termination factor rho and its interaction with RNA. I. Cryoelectron microscopic studies. J Mol Biol *221*, 1127-1138.

Goldenberg, D., Azar, I., and Oppenheim, A.B. (1996). Differential mRNA stability of the cspA gene in the cold-shock response of Escherichia coli. Mol Microbiol *19*, 241-248.

Goldstein, J., Pollitt, N.S., and Inouye, M. (1990). Major cold shock protein of Escherichia coli. Proc Natl Acad Sci U S A *87*, 283-287.

Gong, F., and Yanofsky, C. (2003). Rho's role in transcription attenuation in the tna operon of E. coli. Methods Enzymol *371*, 383-391.

Gottesman, S., McCullen, C.A., Guillier, M., Vanderpool, C.K., Majdalani, N., Benhammou, J., Thompson, K.M., FitzGerald, P.C., Sowa, N.A., and FitzGerald, D.J. (2006). Small RNA regulators and the bacterial response to stress. Cold Spring Harb Symp Quant Biol *71*, 1-11.

Govantes, F., Molina-Lopez, J.A., and Santero, E. (1996). Mechanism of coordinated synthesis of the antagonistic regulatory proteins NifL and NifA of Klebsiella pneumoniae. J Bacteriol *178*, 6817-6823.

Graves, E.T., Duboc, C., Fan, J., Stransky, F., Leroux-Coyau, M., and Strick, T.R. (2015). A dynamic DNA-repair complex observed by correlative single-molecule nanomanipulation and fluorescence. Nat Struct Mol Biol *22*, 452-457.

Greive, S.J., and von Hippel, P.H. (2005). Thinking quantitatively about transcriptional regulation. Nat Rev Mol Cell Biol *6*, 221-232.

Grylak-Mielnicka, A., Bidnenko, V., Bardowski, J., and Bidnenko, E. (2016). Transcription termination factor Rho: a hub linking diverse physiological processes in bacteria. Microbiology *162*, 433-447.

Guerin, M., Robichon, N., Geiselmann, J., and Rahmouni, A.R. (1998). A simple polypyrimidine repeat acts as an artificial Rho-dependent terminator in vivo and in vitro. Nucleic Acids Res *26*, 4895-4900.

Guijarro, J.A., Cascales, D., Garcia-Torrico, A.I., Garcia-Dominguez, M., and Mendez, J. (2015). Temperature-dependent expression of virulence genes in fish-pathogenic bacteria. Frontiers in microbiology *6*, 700.

Guo, X., Myasnikov, A.G., Chen, J., Crucifix, C., Papai, G., Takacs, M., Schultz, P., and Weixlbaumer, A. (2018). Structural Basis for NusA Stabilized Transcriptional Pausing. Mol Cell *69*, 816-827 e814.

Gusarov, I., and Nudler, E. (2001). Control of intrinsic transcription termination by N and NusA: the basic mechanisms. Cell *107*, 437-449.

Gutierrez-Preciado, A., Henkin, T.M., Grundy, F.J., Yanofsky, C., and Merino, E. (2009). Biochemical features and functional implications of the RNA-based T-box regulatory mechanism. Microbiology and molecular biology reviews : MMBR *73*, 36-61.

Gutierrez, P., Kozlov, G., Gabrielli, L., Elias, D., Osborne, M.J., Gallouzi, I.E., and Gehring, K. (2007). Solution structure of YaeO, a Rho-specific inhibitor of transcription termination. J Biol Chem *282*, 23348-23353.

Ha, K.S., Toulokhonov, I., Vassylyev, D.G., and Landick, R. (2010). The NusA N-terminal domain is necessary and sufficient for enhancement of transcriptional pausing via interaction with the RNA exit channel of RNA polymerase. J Mol Biol *401*, 708-725.

Hammann, C., and Westhof, E. (2007). Searching genomes for ribozymes and riboswitches. Genome Biol *8*, 210.

Harinarayanan, R., and Gowrishankar, J. (2003). Host factor titration by chromosomal R-loops as a mechanism for runaway plasmid replication in transcription termination-defective mutants of Escherichia coli. J Mol Biol *332*, 31-46.

Herbert, K.M., Zhou, J., Mooney, R.A., Porta, A.L., Landick, R., and Block, S.M. (2010). E. coli NusG inhibits backtracking and accelerates pause-free transcription by promoting forward translocation of RNA polymerase. J Mol Biol *399*, 17-30.

Hitchens, T.K., Zhan, Y., Richardson, L.V., Richardson, J.P., and Rule, G.S. (2006). Sequence-specific Interactions in the RNA-binding Domain of Escherichia coli Transcription Termination Factor Rho. J Biol Chem *281*, 33697-33703.

Hockensmith, J.W., Kubasek, W.L., Vorachek, W.R., and von Hippel, P.H. (1986). Laser cross-linking of nucleic acids to proteins. Methodology and first applications to the phage T4 DNA replication system. J Biol Chem *261*, 3512-3518.

Hollands, K., Proshkin, S., Sklyarova, S., Epshtein, V., Mironov, A., Nudler, E., and Groisman, E.A. (2012). Riboswitch control of Rho-dependent transcription termination. Proc Natl Acad Sci U S A *109*, 5376-5381.

Holmqvist, E., Li, L., Bischler, T., Barquist, L., and Vogel, J. (2018). Global Maps of ProQ Binding In Vivo Reveal Target Recognition via RNA Structure and Stability Control at mRNA 3' Ends. Mol Cell *70*, 971-982 e976.

Holmqvist, E., and Vogel, J. (2018). RNA-binding proteins in bacteria. Nature reviews Microbiology.

Holmqvist, E., Wright, P.R., Li, L., Bischler, T., Barquist, L., Reinhardt, R., Backofen, R., and Vogel, J. (2016). Global RNA recognition patterns of post-transcriptional regulators Hfq and CsrA revealed by UV crosslinking in vivo. EMBO J *35*, 991-1011.

Hsu, L.M., Vo, N.V., and Chamberlin, M.J. (1995). Escherichia coli transcript cleavage factors GreA and GreB stimulate promoter escape and gene expression in vivo and in vitro. Proc Natl Acad Sci U S A *92*, 11588-11592.

Hu, K., and Artsimovitch, I. (2017). A Screen for rfaH Suppressors Reveals a Key Role for a Connector Region of Termination Factor Rho. mBio *8*.

Huppertz, I., Attig, J., D'Ambrogio, A., Easton, L.E., Sibley, C.R., Sugimoto, Y., Tajnik, M., Konig, J., and Ule, J. (2014). iCLIP: protein-RNA interactions at nucleotide resolution. Methods *65*, 274-287.

Huttenhofer, A., and Noller, H.F. (1994). Footprinting mRNA-ribosome complexes with chemical probes. EMBO J *13*, 3892-3901.

Hwang, H., and Myong, S. (2014). Protein induced fluorescence enhancement (PIFE) for probing protein-nucleic acid interactions. Chemical Society reviews *43*, 1221-1229.

Ingham, C.J., Dennis, J., and Furneaux, P.A. (1999). Autogenous regulation of transcription termination factor Rho and the requirement for Nus factors in Bacillus subtilis. Mol Microbiol *31*, 651-663.

Jager, S., Fuhrmann, O., Heck, C., Hebermehl, M., Schiltz, E., Rauhut, R., and Klug, G. (2001). An mRNA degrading complex in Rhodobacter capsulatus. Nucleic Acids Res *29*, 4581-4588.

Jager, S., Hebermehl, M., Schiltz, E., and Klug, G. (2004). Composition and activity of the Rhodobacter capsulatus degradosome vary under different oxygen concentrations. Journal of molecular microbiology and biotechnology *7*, 148-154.

Jin, D.J., Burgess, R.R., Richardson, J.P., and Gross, C.A. (1992). Termination efficiency at rho-dependent terminators depends on kinetic coupling between RNA polymerase and rho. Proc Natl Acad Sci U S A *89*, 1453-1457.

Kalarickal, N.C., Ranjan, A., Kalyani, B.S., Wal, M., and Sen, R. (2010). A bacterial transcription terminator with inefficient molecular motor action but with a robust transcription termination function. J Mol Biol *395*, 966-982.

Kalyani, B.S., Muteeb, G., Qayyum, M.Z., and Sen, R. (2011). Interaction with the nascent RNA is a prerequisite for the recruitment of Rho to the transcription elongation complex in vitro. J Mol Biol *413*, 548-560.

Kang, J.Y., Mooney, R.A., Nedialkov, Y., Saba, J., Mishanina, T.V., Artsimovitch, I., Landick, R., and Darst, S.A. (2018). Structural Basis for Transcript Elongation Control by NusG Family Universal Regulators. Cell.

Kapanidis, A.N., Margeat, E., Ho, S.O., Kortkhonjia, E., Weiss, S., and Ebright, R.H. (2006). Initial transcription by RNA polymerase proceeds through a DNA-scrunching mechanism. Science *314*, 1144-1147.

Kapanidis, A.N., Margeat, E., Laurence, T.A., Doose, S., Ho, S.O., Mukhopadhyay, J., Kortkhonjia, E., Mekler, V., Ebright, R.H., and Weiss, S. (2005). Retention of Transcription Initiation Factor sigma(70) in Transcription Elongation: Single-Molecule Analysis. Mol Cell *20*, 347-356.

Kavita, K., de Mets, F., and Gottesman, S. (2018). New aspects of RNA-based regulation by Hfq and its partner sRNAs. Curr Opin Microbiol *42*, 53-61.

Kim, J.N., and Breaker, R.R. (2008). Purine sensing by riboswitches. Biology of the cell *100*, 1-11.

Kohler, R., Mooney, R.A., Mills, D.J., Landick, R., and Cramer, P. (2017). Architecture of a transcribing-translating expressome. Science *356*, 194-197.

Komissarova, N., Becker, J., Solter, S., Kireeva, M., and Kashlev, M. (2002). Shortening of RNA:DNA hybrid in the elongation complex of RNA polymerase is a prerequisite for transcription termination. Mol Cell *10*, 1151-1162.

Konig, J., Zarnack, K., Luscombe, N.M., and Ule, J. (2012). Protein-RNA interactions: new genomic technologies and perspectives. Nature reviews Genetics *13*, 77-83.

Korn, L.J., and Yanofsky, C. (1976). Polarity suppressors increase expression of the wild-type tryptophan operon of Escherichia coli. J Mol Biol *103*, 395-409.

Koslover, D.J., Fazal, F.M., Mooney, R.A., Landick, R., and Block, S.M. (2012). Binding and translocation of termination factor rho studied at the single-molecule level. J Mol Biol *423*, 664-676.

Kriner, M.A., and Groisman, E.A. (2015). The bacterial transcription termination factor Rho coordinates Mg homeostasis with translational signals. J Mol Biol.

Kriner, M.A., and Groisman, E.A. (2017). RNA secondary structures regulate three steps of Rho-dependent transcription termination within a bacterial mRNA leader. Nucleic Acids Res *45*, 631-642.

Kriner, M.A., Sevostyanova, A., and Groisman, E.A. (2016). Learning from the Leaders: Gene Regulation by the Transcription Termination Factor Rho. Trends Biochem Sci *41*, 690-699.

Krummel, B., and Chamberlin, M.J. (1992). Structural analysis of ternary complexes of Escherichia coli RNA polymerase. Deoxyribonuclease I footprinting of defined complexes. J Mol Biol *225*, 239-250.

Kupper, H., Sekiya, T., Rosenberg, M., Egan, J., and Landy, A. (1978). A rho-dependent termination site in the gene coding for tyrosine tRNA su3 of Escherichia coli. Nature *272*, 423-428.

La Farina, M., Izzo, V., Costa, M.A., Barbier, R., Duro, G., Vitale, M., and Mutolo, V. (1990). Readthrough transcription occurs at the rho dependent signal F1 TIV in suppressor cells. Nucleic Acids Res *18*, 865-870.

Larson, M.H., Greenleaf, W.J., Landick, R., and Block, S.M. (2008). Applied force reveals mechanistic and energetic details of transcription termination. Cell *132*, 971-982.

Larson, M.H., Mooney, R.A., Peters, J.M., Windgassen, T., Nayak, D., Gross, C.A., Block, S.M., Greenleaf, W.J., Landick, R., and Weissman, J.S. (2014). A pause sequence enriched at translation start sites drives transcription dynamics in vivo. Science *344*, 1042-1047.

Lau, L.F., Roberts, J.W., and Wu, R. (1982). Transcription terminates at lambda tR1 in three clusters. Proc Natl Acad Sci U S A *79*, 6171-6175.

Lau, L.F., Roberts, J.W., and Wu, R. (1983). RNA polymerase pausing and transcript release at the lambda tR1 terminator in vitro. J Biol Chem *258*, 9391-9397.

Lawson, M.R., Ma, W., Bellecourt, M.J., Artsimovitch, I., Martin, A., Landick, R., Schulten, K., and Berger, J.M. (2018). Mechanism for the Regulated Control of Bacterial Transcription Termination by a Universal Adaptor Protein. Mol Cell.

Le, T.T., Yang, Y., Tan, C., Suhanovsky, M.M., Fulbright, R.M., Jr., Inman, J.T., Li, M., Lee, J., Perelman, S., Roberts, J.W., *et al.* (2018). Mfd Dynamically Regulates Transcription via a Release and Catch-Up Mechanism. Cell *172*, 344-357 e315.

Leela, J.K., Syeda, A.H., Anupama, K., and Gowrishankar, J. (2013). Rho-dependent transcription termination is essential to prevent excessive genome-wide R-loops in Escherichia coli. Proc Natl Acad Sci U S A *110*, 258-263.

Lesnik, E.A., Sampath, R., Levene, H.B., Henderson, T.J., McNeil, J.A., and Ecker, D.J. (2001). Prediction of rho-independent transcriptional terminators in Escherichia coli. Nucleic Acids Res *29*, 3583-3594.

Liberman, J.A., Suddala, K.C., Aytenfisu, A., Chan, D., Belashov, I.A., Salim, M., Mathews, D.H., Spitale, R.C., Walter, N.G., and Wedekind, J.E. (2015). Structural analysis of a class III preQ1 riboswitch reveals an aptamer distant from a ribosome-binding site regulated by fast dynamics. Proc Natl Acad Sci U S A *112*, E3485-3494.

Lo, Y.C., Rensi, S.E., Torng, W., and Altman, R.B. (2018). Machine learning in chemoinformatics and drug discovery. Drug discovery today *23*, 1538-1546.

Lowery-Goldhammer, C., and Richardson, J.P. (1974). An RNA-dependent nucleoside triphosphate phosphohydrolase (ATPase) associated with rho termination factor. Proc Natl Acad Sci U S A *71*, 2003-2007.

Ma, C., Yang, X., and Lewis, P.J. (2016). Bacterial Transcription as a Target for Antibacterial Drug Development. Microbiology and molecular biology reviews : MMBR *80*, 139-160.

Mader, U., Nicolas, P., Depke, M., Pane-Farre, J., Debarbouille, M., van der Kooi-Pol, M.M., Guerin, C., Derozier, S., Hiron, A., Jarmer, H., *et al.* (2016). Staphylococcus aureus Transcriptome Architecture: From Laboratory to Infection-Mimicking Conditions. PLoS Genet *12*, e1005962.

Masse, E., Escorcia, F.E., and Gottesman, S. (2003). Coupled degradation of a small regulatory RNA and its mRNA targets in Escherichia coli. Genes Dev *17*, 2374-2383.

Matsumoto, Y., Shigesada, K., Hirano, M., and Imai, M. (1986). Autogenous regulation of the gene for transcription termination factor rho in Escherichia coli: localization and function of its attenuators. J Bacteriol *166*, 945-958.

McSwiggen, J.A., Bear, D.G., and von Hippel, P.H. (1988). Interactions of Escherichia coli transcription termination factor rho with RNA. I. Binding stoichiometries and free energies. J Mol Biol *199*, 609-622.

Mellin, J.R., Koutero, M., Dar, D., Nahori, M.A., Sorek, R., and Cossart, P. (2014). Riboswitches. Sequestration of a two-component response regulator by a riboswitch-regulated noncoding RNA. Science *345*, 940-943.

Menouni, R., Champ, S., Espinosa, L., Boudvillain, M., and Ansaldi, M. (2013). Transcription termination controls prophage maintenance in Escherichia coli genomes. Proc Natl Acad Sci U S A *110*, 14414-14419.

Miller, O.L., Jr., Hamkalo, B.A., and Thomas, C.A., Jr. (1970). Visualization of bacterial genes in action. Science *169*, 392-395.

Mitra, P., Ghosh, G., Hafeezunnisa, M., and Sen, R. (2017). Rho Protein: Roles and Mechanisms. Annu Rev Microbiol *71*, 687-709.

Mogridge, J., Mah, T.F., and Greenblatt, J. (1998). Involvement of boxA nucleotides in the formation of a stable ribonucleoprotein complex containing the bacteriophage lambda N protein. J Biol Chem *273*, 4143-4148.

Mondal, S., Yakhnin, A.V., Sebastian, A., Albert, I., and Babitzke, P. (2016). NusA-dependent transcription termination prevents misregulation of global gene expression. Nature Microbiology *1*, 15007.

Mooney, R.A., Darst, S.A., and Landick, R. (2005). Sigma and RNA polymerase: an on-again, off-again relationship? Mol Cell *20*, 335-345.

Mooney, R.A., Davis, S.E., Peters, J.M., Rowland, J.L., Ansari, A.Z., and Landick, R. (2009a). Regulator trafficking on bacterial transcription units in vivo. Mol Cell *33*, 97-108.

Mooney, R.A., Schweimer, K., Rosch, P., Gottesman, M., and Landick, R. (2009b). Two structurally independent domains of E. coli NusG create regulatory plasticity via distinct interactions with RNA polymerase and regulators. J Mol Biol *391*, 341-358.

Morgan, W.D., Bear, D.G., and von Hippel, P.H. (1983). Rho-dependent termination of transcription. I. Identification and characterization of termination sites for transcription from the bacteriophage lambda PR promoter. J Biol Chem *258*, 9553-9564.

Morita, T., Nishino, R., and Aiba, H. (2017). Role of the terminator hairpin in the biogenesis of functional Hfq-binding sRNAs. RNA *23*, 1419-1431.

Moses, P.B., and Model, P. (1984). A rho-dependent transcription termination signal in bacteriophage f1. J Mol Biol *172*, 1-22.

Murakami, K.S. (2015). Structural biology of bacterial RNA polymerase. Biomolecules *5*, 848-864.

Murakami, K.S., Masuda, S., and Darst, S.A. (2002). Structural basis of transcription initiation: RNA polymerase holoenzyme at 4 A resolution. Science *296*, 1280-1284.

Murigneux, V., Sauliere, J., Roest Crollius, H., and Le Hir, H. (2013). Transcriptome-wide identification of RNA binding sites by CLIP-seq. Methods *63*, 32-40.

Nadiras, C., Eveno, E., Schwartz, A., Figueroa-Bossi, N., and Boudvillain, M. (2018a). A multivariate prediction model for Rho-dependent termination of transcription. Nucleic Acids Res.

Nadiras, C., Schwartz, A., Delaleau, M., and Boudvillain, M. (2018b). Evaluating the Effect of Small RNAs and Associated Chaperones on Rho-Dependent Termination of Transcription In Vitro. Methods Mol Biol *1737*, 99-118.

NandyMazumdar, M., and Artsimovitch, I. (2015). Ubiquitous transcription factors display structural plasticity and diverse functions: NusG proteins - Shifting shapes and paradigms. BioEssays : news and reviews in molecular, cellular and developmental biology *37*, 324-334.

Nehrke, K.W., Zalatan, F., and Platt, T. (1993). NusG alters rho-dependent termination of transcription in vitro independent of kinetic coupling. Gene Expr *3*, 119-133.

Nicolas, P., Mader, U., Dervyn, E., Rochat, T., Leduc, A., Pigeonneau, N., Bidnenko, E., Marchadier, E., Hoebeke, M., Aymerich, S., *et al.* (2012). Condition-dependent transcriptome reveals high-level regulatory architecture in Bacillus subtilis. Science *335*, 1103-1106.

Nowatzke, W.L., Keller, E., Koch, G., and Richardson, J.P. (1997). Transcription termination factor Rho is essential for Micrococcus luteus. J Bacteriol *179*, 5238-5240.

Nudler, E., Mustaev, A., Lukhtanov, E., and Goldfarb, A. (1997). The RNA-DNA hybrid maintains the register of transcription by preventing backtracking of RNA polymerase. Cell *89*, 33-41.

Opalka, N., Brown, J., Lane, W.J., Twist, K.A., Landick, R., Asturias, F.J., and Darst, S.A. (2010). Complete structural model of Escherichia coli RNA polymerase from a hybrid approach. PLoS Biol *8*.

Opalka, N., Chlenov, M., Chacon, P., Rice, W.J., Wriggers, W., and Darst, S.A. (2003). Structure and function of the transcription elongation factor GreB bound to bacterial RNA polymerase. Cell *114*, 335-345.

Pan, T., and Sosnick, T. (2006). RNA folding during transcription. Annual review of biophysics and biomolecular structure *35*, 161-175.

Pani, B., Banerjee, S., Chalissery, J., Abishek, M., Loganathan, R.M., Suganthan, R.B., and Sen, R. (2006a). Mechanism of inhibition of Rho-dependent transcription termination by bacteriophage P4 protein Psu. J Biol Chem *281*, 26491-26500.

Pani, B., Banerjee, S., Chalissery, J., Muralimohan, A., Loganathan, R.M., Suganthan, R.B., and Sen, R. (2006b). Mechanism of inhibition of Rho-dependent transcription termination by bacteriophage P4 protein Psu. J Biol Chem *281*, 26491-26500.

Pani, B., and Nudler, E. (2017). Mechanistic insights into transcription coupled DNA repair. DNA repair *56*, 42-50.

Pani, B., Ranjan, A., and Sen, R. (2009). Interaction surface of bacteriophage P4 protein Psu required for complex formation with the transcription terminator Rho. J Mol Biol *389*, 647-660.

Park, J.S., Marr, M.T., and Roberts, J.W. (2002). E. coli Transcription repair coupling factor (Mfd protein) rescues arrested complexes by promoting forward translocation. Cell *109*, 757-767.

Park, J.S., and Roberts, J.W. (2006). Role of DNA bubble rewinding in enzymatic transcription termination. Proc Natl Acad Sci U S A *103*, 4870-4875.

Patteson, J.B., Cai, W., Johnson, R.A., Santa Maria, K.C., and Li, B. (2018). Identification of the Biosynthetic Pathway for the Antibiotic Bicyclomycin. Biochemistry *57*, 61-65.

Penno, C., Sharma, V., Coakley, A., O'Connell Motherway, M., van Sinderen, D., Lubkowska, L., Kireeva, M.L., Kashlev, M., Baranov, P.V., and Atkins, J.F. (2015). Productive mRNA stem loop-mediated transcriptional slippage: Crucial features in common with intrinsic terminators. Proc Natl Acad Sci U S A *112*, E1984-1993.

Peters, J.M., Mooney, R.A., Grass, J.A., Jessen, E.D., Tran, F., and Landick, R. (2012). Rho and NusG suppress pervasive antisense transcription in Escherichia coli. Genes Dev *26*, 2621-2633.

Peters, J.M., Mooney, R.A., Kuan, P.F., Rowland, J.L., Keles, S., and Landick, R. (2009). Rho directs widespread termination of intragenic and stable RNA transcription. Proc Natl Acad Sci U S A *106*, 15406-15411.

Peters, J.M., Vangeloff, A.D., and Landick, R. (2011). Bacterial transcription terminators: the RNA 3'-end chronicles. J Mol Biol *412*, 793-813.

Petushkov, I., Esyunina, D., Mekler, V., Severinov, K., Pupov, D., and Kulbachinskiy, A. (2017). Interplay between sigma region 3.2 and secondary channel factors during promoter escape by bacterial RNA polymerase. The Biochemical journal *474*, 4053-4064.

Pichoff, S., Alibaud, L., Guedant, A., Castanie, M.P., and Bouche, J.P. (1998). An Escherichia coli gene (yaeO) suppresses temperature-sensitive mutations in essential genes by modulating Rho-dependent transcription termination. Mol Microbiol *29*, 859-869.

Platt, T. (1981). Termination of transcription and its regulation in the tryptophan operon of E. coli. Cell *24*, 10-23.

Portman, J.R., and Strick, T.R. (2018). Transcription-Coupled Repair and Complex Biology. J Mol Biol.

Potter, K.D., Merlino, N.M., Jacobs, T., and Gollnick, P. (2011). TRAP binding to the Bacillus subtilis trp leader region RNA causes efficient transcription termination at a weak intrinsic terminator. Nucleic Acids Res *39*, 2092-2102.

Potts, A.H., Vakulskas, C.A., Pannuri, A., Yakhnin, H., Babitzke, P., and Romeo, T. (2017). Global role of the bacterial post-transcriptional regulator CsrA revealed by integrated transcriptomics. Nature communications *8*, 1596.

Prasch, S., Jurk, M., Washburn, R.S., Gottesman, M.E., Wohrl, B.M., and Rosch, P. (2009). RNA-binding specificity of E. coli NusA. Nucleic Acids Res *37*, 4736-4742.

Proshkin, S., Mironov, A., and Nudler, E. (2014). Riboswitches in regulation of Rho-dependent transcription termination. Biochim Biophys Acta *1839*, 974-977.

Proshkin, S., Rahmouni, A.R., Mironov, A., and Nudler, E. (2010). Cooperation between translating ribosomes and RNA polymerase in transcription elongation. Science *328*, 504-508.

Rabhi, M., Espeli, O., Schwartz, A., Cayrol, B., Rahmouni, A.R., Arluison, V., and Boudvillain, M. (2011a). The Sm-like RNA chaperone Hfq mediates transcription antitermination at Rho-dependent terminators. Embo J *30*, 2805-2816.

Rabhi, M., Gocheva, V., Jacquinot, F., Lee, A., Margeat, E., and Boudvillain, M. (2011b). Mutagenesis-based evidence for an asymmetric configuration of the ring-shaped transcription termination factor Rho. J Mol Biol *405*, 497-518.

Rabhi, M., Rahmouni, A.R., and Boudvillain, M. (2010a). Transcription termination factor Rho: a ring-shaped RNA helicase from bacteria. In RNA helicases, E. Jankowsky, ed. (Cambridge (UK): RSC Publishing), pp. 243-271.

Rabhi, M., Tuma, R., and Boudvillain, M. (2010b). RNA remodeling by hexameric RNA helicases. RNA biology *7*, 655-666.

Raghavan, R., Groisman, E.A., and Ochman, H. (2011). Genome-wide detection of novel regulatory RNAs in E. coli. Genome Res *21*, 1487-1497.

Raghunathan, N., Kapshikar, R.M., Leela, J.K., Mallikarjun, J., Bouloc, P., and Gowrishankar, J. (2018). Genome-wide relationship between R-loop formation and antisense transcription in Escherichia coli. Nucleic Acids Res *46*, 3400-3411.

Ranjan, A., Sharma, S., Banerjee, R., Sen, U., and Sen, R. (2013). Structural and mechanistic basis of anti-termination of Rho-dependent transcription termination by bacteriophage P4 capsid protein Psu. Nucleic Acids Res *41*, 6839-6856.

Ray-Soni, A., Bellecourt, M.J., and Landick, R. (2016). Mechanisms of Bacterial Transcription Termination: All Good Things Must End. Annual review of biochemistry *85*, 319-347.

Richardson, J.P. (1982). Activation of rho protein ATPase requires simultaneous interaction at two kinds of nucleic acid-binding sites. J Biol Chem *257*, 5760-5766.

Richardson, J.P. (1991). Preventing the synthesis of unused transcripts by Rho factor. Cell *64*, 1047-1049.

Richardson, J.P. (2002). Rho-dependent termination and ATPases in transcript termination. Biochim Biophys Acta *1577*, 251-260.

Richardson, J.P., and Greenblatt, J.F. (1996). Control of RNA chain elongation and termination. In Escherischia coli and Salmonella: Cellular and Molecular Biology, F. Neidhardt, R.I. Curtiss, J. Ingraham, E. Lin, K. Low, B. Magasanik, W. Reznikov, M. Riley, M. Schaechter, and H. Umbarger, eds. (Washington D.C.: ASM Press), pp. 822-848.

Richardson, L.V., and Richardson, J.P. (1996). Rho-dependent termination of transcription is governed primarily by the upstream Rho utilization (rut) sequences of a terminator. J Biol Chem *271*, 21597-21603.

Roberts, J., and Park, J.S. (2004). Mfd, the bacterial transcription repair coupling factor: translocation, repair and termination. Curr Opin Microbiol *7*, 120-125.

Roberts, J.W. (1969). Termination factor for RNA synthesis. Nature *224*, 1168-1174.

Roberts, J.W., Shankar, S., and Filter, J.J. (2008). RNA polymerase elongation factors. Annu Rev Microbiol *62*, 211-233.

Rossi, J., Egan, J., Hudson, L., and Landy, A. (1981). The tyrT locus: termination and processing of a complex transcript. Cell *26*, 305-314.

Ruff, E.F., Record, M.T., Jr., and Artsimovitch, I. (2015). Initial events in bacterial transcription initiation. Biomolecules *5*, 1035-1062.

Ruteshouser, E.C., and Richardson, J.P. (1989). Identification and characterization of transcription termination sites in the Escherichia coli lacZ gene. J Mol Biol *208*, 23-43.

Rutherford, S.T., Lemke, J.J., Vrentas, C.E., Gaal, T., Ross, W., and Gourse, R.L. (2007). Effects of DksA, GreA, and GreB on transcription initiation: insights into the mechanisms of factors that bind in the secondary channel of RNA polymerase. J Mol Biol *366*, 1243-1257.

Saecker, R.M., Record, M.T., Jr., and Dehaseth, P.L. (2011). Mechanism of bacterial transcription initiation: RNA polymerase - promoter binding, isomerization to initiation-competent open complexes, and initiation of RNA synthesis. J Mol Biol *412*, 754-771.

Santangelo, T.J., and Artsimovitch, I. (2011). Termination and antitermination: RNA polymerase runs a stop sign. Nature reviews Microbiology *9*, 319-329.

Santangelo, T.J., and Roberts, J.W. (2004). Forward translocation is the natural pathway of RNA release at an intrinsic terminator. Mol Cell *14*, 117-126.

Santiago-Frangos, A., Kavita, K., Schu, D.J., Gottesman, S., and Woodson, S.A. (2016). C-terminal domain of the RNA chaperone Hfq drives sRNA competition and release of target RNA. Proc Natl Acad Sci U S A *113*, E6089-E6096.

Saxena, S., Myka, K.K., Washburn, R., Costantino, N., Court, D.L., and Gottesman, M.E. (2018). Escherichia coli transcription factor NusG binds to 70S ribosomes. Mol Microbiol.

Schwartz, A., Rabhi, M., Jacquinot, F., Margeat, E., Rahmouni, A.R., and Boudvillain, M. (2009). A stepwise 2'-hydroxyl activation mechanism for the bacterial transcription termination factor Rho helicase. Nat Struct Mol Biol *16*, 1309-1316.

Schwartz, A., Walmacq, C., Rahmouni, A.R., and Boudvillain, M. (2007). Noncanonical interactions in the management of RNA structural blocks by the transcription termination rho helicase. Biochemistry *46*, 9366-9379.

Sedlyarova, N., Rescheneder, P., Magan, A., Popitsch, N., Rziha, N., Bilusic, I., Epshtein, V., Zimmermann, B., Lybecker, M., Sedlyarov, V., *et al.* (2017). Natural RNA Polymerase Aptamers Regulate Transcription in E. coli. Mol Cell *67*, 30-43 e36.

Sedlyarova, N., Shamovsky, I., Bharati, B.K., Epshtein, V., Chen, J., Gottesman, S., Schroeder, R., and Nudler, E. (2016). sRNA-Mediated Control of Transcription Termination in E. coli. Cell *167*, 111-121 e113.

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

Sevostyanova, A., and Groisman, E.A. (2015). An RNA motif advances transcription by preventing Rho-dependent termination. Proc Natl Acad Sci U S A.

Shankar, S., Hatoum, A., and Roberts, J.W. (2007). A transcription antiterminator constructs a NusA-dependent shield to the emerging transcript. Mol Cell *27*, 914-927.

Shashni, R., Qayyum, M.Z., Vishalini, V., Dey, D., and Sen, R. (2014). Redundancy of primary RNA-binding functions of the bacterial transcription terminator Rho. Nucleic Acids Res *42*, 9677-9690.

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

Silverman, M.S. (1974). Immunologic response and periodontal disease. The Journal of the North Carolina Dental Society *57*, 21-24.

Singleton, M.R., Dillingham, M.S., and Wigley, D.B. (2007). Structure and mechanism of helicases and nucleic acid translocases. Annual review of biochemistry *76*, 23-50.

Skordalakes, E., and Berger, J.M. (2003). Structure of the Rho transcription terminator: mechanism of mRNA recognition and helicase loading. Cell *114*, 135-146.

Skordalakes, E., and Berger, J.M. (2006). Structural Insights into RNA-Dependent Ring Closure and ATPase Activation by the Rho Termination Factor. Cell *127*, 553-564.

Smirnov, A., Forstner, K.U., Holmqvist, E., Otto, A., Gunster, R., Becher, D., Reinhardt, R., and Vogel, J. (2016). Grad-seq guides the discovery of ProQ as a major small RNA-binding protein. Proc Natl Acad Sci U S A *113*, 11591-11596.

Smith, M.N., Crane, R.A., Keates, R.A., and Wood, J.M. (2004). Overexpression, purification, and characterization of ProQ, a posttranslational regulator for osmoregulatory transporter ProP of Escherichia coli. Biochemistry *43*, 12979-12989.

Soares, E., Schwartz, A., Nollmann, M., Margeat, E., and Boudvillain, M. (2014). The RNA-mediated, asymmetric ring regulatory mechanism of the transcription termination Rho helicase decrypted by time-resolved Nucleotide Analog Interference Probing (trNAIP). Nucleic Acids Res *42*, 9270-9284.

Stanssens, P., Remaut, E., and Fiers, W. (1986). Inefficient translation initiation causes premature transcription termination in the lacZ gene. Cell *44*, 711-718.

Stebbins, C.E., Borukhov, S., Orlova, M., Polyakov, A., Goldfarb, A., and Darst, S.A. (1995). Crystal structure of the GreA transcript cleavage factor from Escherichia coli. Nature *373*, 636-640.

Steinmetz, E.J., and Platt, T. (1994). Evidence supporting a tethered tracking model for helicase activity of Escherichia coli Rho factor. Proc Natl Acad Sci U S A *91*, 1401-1405.

Steitz, J.A., and Jakes, K. (1975). How ribosomes select initiator regions in mRNA: base pair formation between the 3' terminus of 16S rRNA and the mRNA during initiation of protein synthesis in Escherichia coli. Proc Natl Acad Sci U S A *72*, 4734-4738.

Stenz, L., Francois, P., Whiteson, K., Wolz, C., Linder, P., and Schrenzel, J. (2011). The CodY pleiotropic repressor controls virulence in gram-positive pathogens. FEMS immunology and medical microbiology *62*, 123-139.

Stewart, V., Landick, R., and Yanofsky, C. (1986). Rho-dependent transcription termination in the tryptophanase operon leader region of Escherichia coli K-12. J Bacteriol *166*, 217-223.

Stracy, M., and Kapanidis, A.N. (2017). Single-molecule and super-resolution imaging of transcription in living bacteria. Methods *120*, 103-114.

Strauss, M., Vitiello, C., Schweimer, K., Gottesman, M., Rosch, P., and Knauer, S.H. (2016). Transcription is regulated by NusA:NusG interaction. Nucleic Acids Res *44*, 5971-5982.

Sudarsan, N., Hammond, M.C., Block, K.F., Welz, R., Barrick, J.E., Roth, A., and Breaker, R.R. (2006). Tandem riboswitch architectures exhibit complex gene control functions. Science *314*, 300-304.

Sullivan, S.L., and Gottesman, M.E. (1992). Requirement for E. coli NusG protein in factor-dependent transcription termination. Cell *68*, 989-994.

Sundararaman, B., Zhan, L., Blue, S.M., Stanton, R., Elkins, K., Olson, S., Wei, X., Van Nostrand, E.L., Pratt, G.A., Huelga, S.C., *et al.* (2016). Resources for the Comprehensive Discovery of Functional RNA Elements. Mol Cell *61*, 903-913.

Svetlov, V., Belogurov, G.A., Shabrova, E., Vassylyev, D.G., and Artsimovitch, I. (2007). Allosteric control of the RNA polymerase by the elongation factor RfaH. Nucleic Acids Res *35*, 5694-5705.

Svetlov, V., and Nudler, E. (2011). Clamping the clamp of RNA polymerase. EMBO J *30*, 1190-1191.

Takemoto, N., Tanaka, Y., and Inui, M. (2015). Rho and RNase play a central role in FMN riboswitch regulation in Corynebacterium glutamicum. Nucleic Acids Res *43*, 520-529.

Takemoto, N., Tanaka, Y., Inui, M., and Yukawa, H. (2014). The physiological role of riboflavin transporter and involvement of FMN-riboswitch in its gene expression in Corynebacterium glutamicum. Applied microbiology and biotechnology *98*, 4159-4168.

Thomsen, N.D., and Berger, J.M. (2009). Running in reverse: the structural basis for translocation polarity in hexameric helicases. Cell *139*, 523-534.

Thomsen, N.D., Lawson, M.R., Witkowsky, L.B., Qu, S., and Berger, J.M. (2016). Molecular mechanisms of substrate-controlled ring dynamics and substepping in a nucleic acid-dependent hexameric motor. Proc Natl Acad Sci U S A *113*, E7691-E7700.

Toulokhonov, I., Artsimovitch, I., and Landick, R. (2001). Allosteric control of RNA polymerase by a site that contacts nascent RNA hairpins. Science *292*, 730-733.

Toulokhonov, I., and Landick, R. (2003). The flap domain is required for pause RNA hairpin inhibition of catalysis by RNA polymerase and can modulate intrinsic termination. Mol Cell *12*, 1125-1136.

Toulokhonov, I., Zhang, J., Palangat, M., and Landick, R. (2007). A central role of the RNA polymerase trigger loop in active-site rearrangement during transcriptional pausing. Mol Cell *27*, 406-419.

Unniraman, S., Prakash, R., and Nagaraja, V. (2002). Conserved economics of transcription termination in eubacteria. Nucleic Acids Res *30*, 675-684.

Vakulskas, C.A., Potts, A.H., Babitzke, P., Ahmer, B.M., and Romeo, T. (2015). Regulation of bacterial virulence by Csr (Rsm) systems. Microbiology and molecular biology reviews : MMBR *79*, 193-224.

Valabhoju, V., Agrawal, S., and Sen, R. (2016). Molecular Basis of NusG-mediated Regulation of Rho-dependent Transcription Termination in Bacteria. J Biol Chem *291*, 22386-22403.

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

Vassylyeva, M.N., Svetlov, V., Dearborn, A.D., Klyuyev, S., Artsimovitch, I., and Vassylyev, D.G. (2007). The carboxy-terminal coiled-coil of the RNA polymerase beta'-subunit is the main binding site for Gre factors. EMBO reports *8*, 1038-1043.

Vieu, E., and Rahmouni, A.R. (2004a). Dual role of boxB RNA motif in the mechanisms of termination/antitermination at the lambda tR1 terminator revealed in vivo. J Mol Biol *339*, 1077-1087.

Vieu, E., and Rahmouni, A.R. (2004b). Dual role of boxB RNA motif in the mechanisms of termination/antitermination at the lambda tR1 terminator revealed in vivo. J Mol Biol *339*, 1077-1087.

Vvedenskaya, I.O., Vahedian-Movahed, H., Bird, J.G., Knoblauch, J.G., Goldman, S.R., Zhang, Y., Ebright, R.H., and Nickels, B.E. (2014). Interactions between RNA polymerase and the "core recognition element" counteract pausing. Science *344*, 1285-1289.

Wagner, E.G., and Romby, P. (2015). Small RNAs in bacteria and archaea: who they are, what they do, and how they do it. Advances in genetics *90*, 133-208.

Walmacq, C., Rahmouni, A.R., and Boudvillain, M. (2004). Influence of substrate composition on the helicase activity of transcription termination factor Rho: reduced processivity of Rho hexamers during unwinding of RNA-DNA hybrid regions. J Mol Biol *342*, 403-420.

Walmacq, C., Rahmouni, A.R., and Boudvillain, M. (2006). Testing the steric exclusion model for hexameric helicases: substrate features that alter RNA-DNA unwinding by the transcription termination factor Rho. Biochemistry *45*, 5885-5895.

Wang, X., Ji, S.C., Jeon, H.J., Lee, Y., and Lim, H.M. (2015). Two-level inhibition of galK expression by Spot 42: Degradation of mRNA mK2 and enhanced transcription termination before the galK gene. Proc Natl Acad Sci U S A *112*, 7581-7586.

Wang, Y., and von Hippel, P.H. (1993). Escherichia coli transcription termination factor rho. I. ATPase activation by oligonucleotide cofactors. J Biol Chem *268*, 13940-13946.

Washburn, R.S., and Gottesman, M.E. (2015). Regulation of transcription elongation and termination. Biomolecules *5*, 1063-1078.

Waters, L.S., and Storz, G. (2009). Regulatory RNAs in bacteria. Cell *136*, 615-628.

Watson, J., Baker, T., Bell, S., Gann, A., LeVine, M.J., and Losick, R. (2009). Biologie moléculaire du gène, 6 edn.

Wei, B., Shin, S., LaPorte, D., Wolfe, A.J., and Romeo, T. (2000). Global regulatory mutations in csrA and rpoS cause severe central carbon stress in Escherichia coli in the presence of acetate. J Bacteriol *182*, 1632-1640.

Wei, R.R., and Richardson, J.P. (2001). Identification of an RNA-binding Site in the ATP binding domain of Escherichia coli Rho by H2O2/Fe-EDTA cleavage protection studies. J Biol Chem *276*, 28380-28387.

Weixlbaumer, A., Leon, K., Landick, R., and Darst, S.A. (2013). Structural basis of transcriptional pausing in bacteria. Cell *152*, 431-441.

Wek, R.C., Sameshima, J.H., and Hatfield, G.W. (1987). Rho-dependent transcriptional polarity in the ilvGMEDA operon of wild-type Escherichia coli K12. J Biol Chem *262*, 15256-15261.

Wheeler, E.C., Van Nostrand, E.L., and Yeo, G.W. (2018). Advances and challenges in the detection of transcriptome-wide protein-RNA interactions. Wiley interdisciplinary reviews RNA *9*.

Winkelman, J.T., and Gourse, R.L. (2017). Open complex DNA scrunching: A key to transcription start site selection and promoter escape. BioEssays : news and reviews in molecular, cellular and developmental biology *39*.

Wu, A.M., Christie, G.E., and Platt, T. (1981). Tandem termination sites in the tryptophan operon of Escherichia coli. Proc Natl Acad Sci U S A *78*, 2913-2917.

Yakhnin, A.V., and Babitzke, P. (2002). NusA-stimulated RNA polymerase pausing and termination participates in the Bacillus subtilis trp operon attenuation mechanism invitro. Proc Natl Acad Sci U S A *99*, 11067-11072.

Yakhnin, A.V., and Babitzke, P. (2010). Mechanism of NusG-stimulated pausing, hairpin-dependent pause site selection and intrinsic termination at overlapping pause and termination sites in the Bacillus subtilis trp leader. Mol Microbiol *76*, 690-705.

Yakhnin, H., Babiarz, J.E., Yakhnin, A.V., and Babitzke, P. (2001). Expression of the Bacillus subtilis trpEDCFBA operon is influenced by translational coupling and Rho termination factor. J Bacteriol *183*, 5918-5926.

Yanofsky, C. (2007). RNA-based regulation of genes of tryptophan synthesis and degradation, in bacteria. RNA *13*, 1141-1154.

Yarnell, W.S., and Roberts, J.W. (1999). Mechanism of intrinsic transcription termination and antitermination. Science *284*, 611-615.

Yu, X., Horiguchi, T., Shigesada, K., and Egelman, E.H. (2000). Three-dimensional reconstruction of transcription termination factor rho: orientation of the N-terminal domain and visualization of an RNA-binding site. J Mol Biol *299*, 1279-1287.

Zalatan, F., Galloway-Salvo, J., and Platt, T. (1993). Deletion analysis of the Escherichia coli rho-dependent transcription terminator trp t'. J Biol Chem *268*, 17051-17056.

Zhang, A., Wassarman, K.M., Rosenow, C., Tjaden, B.C., Storz, G., and Gottesman, S. (2003). Global analysis of small RNA and mRNA targets of Hfq. Mol Microbiol *50*, 1111-1124.

Zhang, G., Campbell, E.A., Minakhin, L., Richter, C., Severinov, K., and Darst, S.A. (1999). Crystal structure of Thermus aquaticus core RNA polymerase at 3.3 A resolution. Cell *98*, 811-824.

Zhang, J., and Landick, R. (2016). A Two-Way Street: Regulatory Interplay between RNA Polymerase and Nascent RNA Structure. Trends Biochem Sci *41*, 293-310.

Zhu, A.Q., and von Hippel, P.H. (1998). Rho-dependent termination within the trp t' terminator. I. Effects of rho loading and template sequence. Biochemistry *37*, 11202-11214.

Zuber, P.K., Artsimovitch, I., NandyMazumdar, M., Liu, Z., Nedialkov, Y., Schweimer, K., Rosch, P., and Knauer, S.H. (2018). The universally-conserved transcription factor RfaH is recruited to a hairpin structure of the non-template DNA strand. eLife *7*.

Zuo, Y., and Steitz, T.A. (2015). Crystal structures of the E. coli transcription initiation complexes with a complete bubble. Mol Cell *58*, 534-540.

Zwiefka, A., Kohn, H., and Widger, W.R. (1993). Transcription termination factor rho: the site of bicyclomycin inhibition in Escherichia coli. Biochemistry *32*, 3564-3570.

Cédric Nadiras

Etude des mécanismes de reconnaissance du transcrit dans la terminaison de la transcription Rho-dépendante

Rho est un facteur protéique bactérien organisé en anneau homo-hexamèrique qui induit la terminaison de la transcription. Rho se fixe aux transcrits naissants au niveau d'un site *Rut* (*Rho-utilization*) libre à partir duquel il transloque le long de l'ARN (5'→3') de façon ATP-dépendante pour rattraper le complexe d'élongation de la transcription et induire la dissociation de celui-ci. Il est généralement admis que les sites de fixation de Rho présentent une richesse en Cytosines et une pauvreté en Guanines, ainsi qu'une relative pauvreté en structures secondaires. Les études génomiques ou transcriptomiques n'ont pas dégagé d'éléments consensus ou de règles permettant de prédire les sites de terminaison Rho-dépendants. En combinant approches biochimiques et bioinformatiques, j'ai tenté de comprendre les mécanismes par lesquels Rho reconnait les transcrits. J'ai identifié un ensemble de déterminants de séquence qui, pris ensemble, possèdent un bon pouvoir prédictif et que j'ai utilisé pour construire le premier modèle computationnel capable de prédire la terminaison Rho-dépendante à l'échelle des génomes d'*E. coli* et *Salmonella*. J'ai caractérisé *in vitro* certains de ces terminateurs, en particulier dans les régions 5'UTR, avec l'espoir qu'ils soient impliqués dans des mécanismes de régulation conditionnelle. J'ai identifié des candidats dont l'activité pourrait être sous le contrôle de facteurs comme des petits ARN non codants (sRNA) ou la température. J'ai également développé une méthode fluorogénique pour détecter facilement la terminaison Rho-dépendante *in vitro* et ai commencé à adapter l'approche CLIP-seq à l'étude du transcriptome Rho-dépendant chez *Salmonella*. Collectivement, mes travaux offrent de nouveaux outils d'analyse et de prédiction de la terminaison Rho-dépendante, une meilleure cartographie des sites d'action de Rho chez *E. coli* et *Salmonella*, ainsi que de nouvelles pistes d'étude du rôle de Rho dans l'expression conditionnelle du génome.

Mots clés : Rho, transcription, terminaison, *Rut*, prédiction, Clip-seq, *Molecular beacons*.

Study of transcript recognition mechanisms in Rho-dependent termination of transcription

Rho is a ring-shaped bacterial factor that induces termination of transcription. Rho binds to nascent transcripts at a free *Rut* (*Rho-utilization*) site from which Rho moves along the RNA in an ATP-dependent fashion to catch up with and dissociate the transcription elongation complex. It is generally believed that the *Rut* sites are, respectively, rich and poor in Cytosines and Guanines as well as relatively poor in secondary structures. Studies at the genomic or transcriptomic scale have not revealed any stronger consensus features or rules for predicting potential Rho-dependent termination sites. By combining biochemical and bioinformatics approaches, I have explored the mechanisms by which Rho recognizes transcripts to induce transcription termination. I have identified a complex set of sequence determinants which, taken together, have good predictive power and which I used to build the first computational model able to predict Rho-dependent termination at the scale of *Escherichia coli* and *Salmonella* genomes. I have characterized *in vitro* some of these terminators, particularly in 5'UTRs, with the hope that they will be involved in conditional regulatory mechanisms. I have identified several candidates whose activity may be under the control of factors such as small non-coding RNAs (sRNA) or temperature. I have also developed a fluorogenic method to easily detect Rho-dependent termination *in vitro* and have begun to adapt the CLIP-seq approach to the study of the Rho-dependent transcriptome in *Salmonella*. Collectively, my work offers new tools for the analysis and prediction of Rho-dependent termination, a better mapping of the sites of probable Rho action in *E. coli* and *Salmonella*, as well as several lines of investigation of the role of Rho in the conditional expression of bacterial genomes.

Keywords: Rho, transcription, termination, *Rut*, predict, Clip-seq, *Molecular beacons*.

Centre de Biophysique Moléculaire
CNRS UPR 4301
Avenue de la recherche Scientifique