



#### En vue de l'obtention du

## **DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE**

**Délivré par** l'Université Toulouse III - Paul Sabatier **Discipline ou spécialité :** Microbiologie

## Présentée et soutenue par Kounthéa PHOK

**Le** 18 mars 2011

**Titre :** Identification et caractérisation de nouveaux ARN non codant chez l'archée hyperthermophile *Pyrococcus abyssi* 

#### **JURY**

Claude Gutierrez Pascale Romby Bruno Charpentier Marie-Claude Serre Hélène Touzet Béatrice Clouet d'Orval Christine Gaspin Professeur, Université Paul Sabatier, Toulouse Directrice de recherche, Strasbourg Professeur, Université Henri Poincaré, Nancy Chargée de Recherche, Orsay Directrice de recherche, Lille Chargée de recherche, Toulouse Directrice de recherche, Castanet-Tolosan Président
Examinateur
Rapporteur
Rapporteur
Rapporteur
Co-directrice de thèse
Co-directrice de thèse

**Ecole doctorale :** Ecole doctorale BSB Biologie-Santé-Biotechnologies

#### Unités de recherche :

Laboratoire de Microbiologie et de Génétique Moléculaires UMR5100 CNRS-UPS Laboratoire de Biométrie et Intelligence Artificielle UR0878 INRA

Co-directrices de thèse : Béatrice Clouet d'Orval et Christine Gaspin

## Remerciements

Je tiens à remercier Béatrice Clouet d'Orval et Christine Gaspin pour m'avoir permis de travailler sur ce projet de recherche. Leur exigence et leur rigueur scientifique ont été très importants pour la qualité de ce travail. Bien que ne les ayant pas sollicitées aussi souvent qu'elles ne l'auraient voulu (à leur grand désarroi), elles ont toujours su me stimuler lorsque je rencontrais des difficultés. Elles ont été présentes et m'ont toujours soutenue dans tous les moments difficiles liés ou pas à cette thèse.

Je remercie Marie-Claude Serre, Hélène Touzet et Bruno Charpentier d'avoir accepté d'être rapporteurs de ce manuscrit. Leurs commentaires pertinents ont joué un rôle très important pour l'élaboration de ce mémoire et je leur en suis absolument reconnaissante. Je remercie aussi très sincèrement Pascale Romby d'avoir accepté d'être examinateur de cette thèse.

Merci à Claude Gutierrez d'avoir bien voulu présider ce jury de thèse, pour son soutien et ses encouragements depuis mon M2R jusqu'à la finalisation de ma thèse.

Je tiens à remercier A.J. Carpousis pour m'avoir accueillie dans son équipe afin de réaliser cette thèse dans des conditions idéales pour tout doctorant. Son niveau scientifique, ses remarques toujours pertinentes et ses encouragements ont été des sources de motivation pour mener à bien ce travail. Culinairement parlant, son « One bowl chocolate cake » et ses brochettes de canard aux fruits confits sont un délice !

J'ai absolument apprécié d'avoir travaillé avec Annick Moisan. Elle a été une très bonne formatrice malgré mon niveau plus que novice en bio-informatique. Sa curiosité a stimulé ma propre curiosité afin de faire « parler » les ARN. Merci aussi de m'avoir fait une petite place dans son bureau lors de mes déplacements occasionnels à l'INRA. Merci aussi aux membres du laboratoire BIA avec qui j'ai passé de très bons moments pendant ma thèse.

Bien sur, je ne peux pas oublier tous les membres et anciens membres de l'équipe Carpousis (que des filles ! sauf A.J.) : Cathy (pour sa présence et son soutien permanents depuis le M2R !), Leo (qui veille toujours au bon fonctionnement de l'équipe), Marie (pour sa bonheur humeur quotidienne et pour les nombreuses discussions scientifiques et extra-scientifiques), Vanessa (toutes les discussions qui m'ont souvent permis de relativiser), les différents stagiaires passés (Inès, Aurélie, Isabelle, Aurélie, Félix, Rémi, Endrick, Camille, Amélie et Benjamin) et bien sur, les Carpousettes : Dana (les « mains magiques ») et Soraya (la « grande sœur »).

Dana et Soraya, je vous avais promis plusieurs pages de remerciements pour tout ce que vous avez pu faire pour moi mais il ne serait pas convenable que ces remerciements soient aussi longs que mon manuscrit, j'espère que ces quelques lignes témoigneront de l'importance que vous avez pu m'apporter pendant toute cette thèse que ça soit pour les manipulations, la rédaction de la thèse et tout ce qui se passait en parallèle.

Je suis particulièrement reconnaissante envers Patricia Bordes. Tu as cru en moi lors de mon stage de M2R, tu m'as soutenue jusqu'au bout de cette thèse dans les hauts comme dans les bas. Beaucoup d'autres personnes m'ont aussi soutenue et je leur remercie de tout mon cœur : Philou (le papa poule des M2R), Carine (et sa bonne humeur), Marie-Pierre (toujours à me redonner le sourire), JC (le grand frère et le grand sage), Jérôme R. (mon voisin Eaunois !), Violette (et tous les conseils que tu as pu me donner pour les manips), Nathalie (mon binôme préféré pour les concerts), Chantal (doyenne des ITA, respect), Anne (pour le papier bulles à éclater), Jérôme C. (l'autre voisin Eaunois), Stéphane L., Hervé S. et tous les autres.

Bien sur, mon binôme de toujours : la petite et pénible Audrey. On est devenues inséparables depuis le M2R et notre amitié continue à grandir. On a partagé de très bons moments comme les mauvais mais toujours pour en sortir la tête haute !!! Merci d'avoir été à mes côtés et j'espère que je vais pouvoir continuer à t'apporter tout le soutien qu'il te faudra pour que tu puisses finir ta thèse en bonne et due forme !

Je n'oublie pas les autres étudiants, post-docs et assimilés avec qui j'ai pu partager des rires, des discussions et autres : mes chouchous Jipé et Yoyo, Cécile, Romain, Clément, les deux Mathieu, Laure, Elsa, Sophie, Delphine, Virginie, Cheikh, Fanny, la grande Audrey, Coralie, Cathy, Marie, Emilien et tous les autres.

Un clin d'œil aux collègues du badminton, très bonne initiative de la part de Guillaume, ça m'a bien aidée à surmonter tout le stress de la fin de thèse.

Merci à tous mes professeurs et encadrants de stage qui m'ont encouragée à continuer dans cette voie.

Un autre clin d'œil à l'association Alpha-T, j'ai été très honorée d'avoir pu faire vivre cette association de doctorants en ayant été secrétaire puis présidente. Ça a été une excellente expérience d'organiser le colloque inter-doctorants !!!

Je n'oublie pas tous mes amis de toujours : Stéphanie, François, Guy, Céline, les « greluches » : Dorothée, Claire, Carole, Audrey, Laetitia ; les « quiches » : Lucie, Emilie, Julie, Laure, Anaïs, Céline ; Guillaume, Fred, Lionel, Cédric .... Merci pour tout !!!

Trinity, Clairass, Cawole (et son Flowent), Audrey, Lulu, Milou, Jiji, Lolotte, Nanou et la quiche Céline: vous avez été présents dans tous les bons moments comme dans les mauvais, on a vécu un bon nombre d'aventures ensemble et j'espère que ça continuera jusqu'à nos vieux jours!!! Merci d'avoir été là pour moi, de me rappeler qu'il n'y a pas que la thèse dans la vie et que les amis sont indispensables pour tenir le coup!

Un grand merci à ma famille : mes parents pour leur soutien et leur confiance dans cette folle aventure qu'est la thèse, ma petite sœur Theary et mon petit frère Thea-Rong, qui m'ont très souvent rappelée à mon rôle de grande sœur malgré le peu de temps que je leur accordais au profit de ma thèse... Aux familles de mes oncles et tantes en France et au Cambodge qui n'ont jamais compris dans quel domaine j'étudiais mais qui me manifestaient leur soutien malgré tout ! Je remercie également les amis du restaurant : Camille et Antoinette, Jean-Paul et Chantal, et tous les autres... Et je suis immensément reconnaissante à la famille de mon ami : Marie et sa fille, Françoise, Jean-Paul et Frédéric et sa famille, pour leur soutien pendant l'absence de JB, leur présence et leurs encouragements.

Et pour terminer cette liste, je ne peux oublier Jean-Baptiste, pour sa patience, son soutien, ses encouragements et son amour qui ont su me redonner confiance dans les moments les plus difficiles et ce, malgré la distance entre Kaboul et Toulouse qui nous sépare depuis le 5 octobre 2010... Le temps défile et le retour en Avril approche à grands pas ! La patience est une qualité indispensable pour faire face au monde militaire !

#### Résumé

Les ARN non codant (ARNnc) possèdent des fonctions vitales dans les trois domaines du vivant (eucaryotes, bactéries et archées). Les plus connus sont les acteurs de la machinerie traductionnelle (ARNr et ARNt) mais de nouveaux ARNnc régulateurs de l'expression génique ont été identifiés et caractérisés à ce jour. Des génomes d'organismes modèles eucaryotes et bactériens ont été criblés systématiquement pour la recherche d'ARNnc aussi bien par des approches biologiques que bioinformatiques. Cependant, peu d'études du même type sont parues chez les génomes d'archées. L'objectif de cette thèse est d'identifier de nouveaux ARNnc dans le génome d'une archée hyperthermophile de l'ordre des Thermococcales, Pyrococcus abyssi. Une approche bioinformatique combinant deux méthodes complémentaires a été utilisée pour cribler les régions intergéniques de la séquence génomique de cette archée. Cette recherche a abouti à la prédiction de nombreuses régions candidates dont 82 ont été retenues pour une validation expérimentale par Northern Blot. Des transcrits ont été détectés pour 24 prédictions dont 16 sont de nouveaux ARN qui n'ont jamais été décrits. Pour chaque nouvel ARN, les extrémités 5' et 3' ont été identifiées par des analyses d'extension d'amorce et en Circular RACE. Des motifs promoteurs ont été proposés pour neuf d'entre eux. Une analyse comparative des séquences étendue à l'ensemble des Thermococcales, intégrant l'étude de la conservation de séquences et la réalisation d'alignements des séquences orthologues, a amené à proposer des structures secondaires conservées pour huit ARN. L'ensemble des données acquises a permis d'apporter des arguments conduisant à émettre des hypothèses quant à la fonction de ces ARN dans la cellule. Deux d'entre eux sont capables de former des complexes ribonucléoprotéiques (RNP) avec L7Ae, protéine ribosomale composante de RNP déjà caractérisés chez les archées, suggérant que ces ARN feraient partie de complexes RNP fonctionnels.

Ce travail de thèse a donc conduit à la détection de nouveaux ARNnc qui apparaissent être de bons candidats pour réguler l'expression génique chez l'archée *P. abyssi* et de manière plus générale chez les Thermococcales.

#### **Abstract**

Non coding RNAs (ncRNAs) play vital roles in all kingdoms of life (Eukarya, Bacteria and Archaea). In addition to the key factors of translation apparatus, rRNAs and tRNAs, a large variety of gene expression regulatory ncRNAs have been identified and charaterized. Lots of computational and biological screens have been used for eukaryal and bacterial genomes to search new ncRNAs but fewer studies have been published for archaeal ones. The aim of this project is to identify new ncRNAs in a hyperthermophile archaeon belonging to the order of Thermococcales, Pyrococcus abyssi. A computational screen combining two complementary methods was used to predict ncRNAs genes into intergenic regions of P. abyssi's genome. 82 predicted regions were selected and tested by Northern blotting analysis. 24 of them are transcribed into RNAs and 16 appeared to be new ncRNAs that have never been described before this work. For each ncRNA, 5' and 3' ends were identified by primer extension and Circular RACE analysis. Promoters have been predicted for nine of them. A comparative analysis extend to the Thermococcales genomes, comprising similar sequences searches and multiple alignments with ortholog sequences, conducted to proposed conserved secondary structures for eight of them. All these data enabled us to propose putative functions for these RNAs in the cell. Moreover, two of them are able to form RNP (ribonucleoproteins) complexes with the multifunctional ribosomal protein L7Ae, a compount of known archaeal RNP complexes, suggesting that these RNAs may play roles into functionnal RNP complexes.

This thesis project allowed detecting new ncRNAs that might play regulatory functions for genomic expression in the archaeon *P. abyssi* and in a broad manner in the Thermococcales.

#### **Abréviations**

A: adénine

ADN: acide désoxyribonucléique

ADNc : acide désoxyribonucléique complémentaire

ARN : acide ribonucléique

ARNm: acide ribonucléique messager

ARNnc : ARN non codant ARN pol : ARN polymérase

ARNr : acide ribonucléique ribosomique ARNt : acide ribonucléique de transfert

ATP: adénosine 5' triphosphate

bp/kb/Mb : paire de base/kilobase/Mégabase

C : cytosine °C : degré Celsius Cpm : coups par minute

ddNTP: didésoxyribonucléotide triphosphate

DNase: désoxyribonucléase

dNTP: désoxyribonucléotide triphosphate

G: guanine

LUCA: Last Universal Common Ancestor ou dernier ancêtre commun universel

LSU : large subunit min : minute

NTP: ribonucléotide triphosphate

nt : nucléotide

ORF: phase ouverte de lecture (Open Reading Frame)

PCR : amplification en chaine par une polymérase (polymerase chain reaction)

pré-ARNm : précurseur de l'ARN messager pré-ARNr : précurseur de l'ARN ribosomique pré-ARNt : précurseur de l'ARN de transfert

R: purine

RNase : ribonucléase RT : reverse transcriptase

s : secondeS : Svedbergsb : simple brin

snRNA: petit ARN nucléaire (Small nuclear RNA)

snRNP: particule ribonucléoprotéique contenant un ou des petits ARN nucléaires

(Small nuclear Ribonucleoprotein Particule)

snoRNA: petit ARN nucléolaire (Small nucleolar RNA)

snoRNP: particule ribonucléoprotéique contenant un ou des petits ARN nucléolaires

(Small nucleolar ribonucleoprotein particule)

spliceosome : complexe nucléaire ribonucléoprotéique dans lequel ont lieu les réactions

d'épissage

sRNA: petit ARN (Small RNA)

sRNP : particule ribonucléoprotéique contenant un ou des petits ARN (Small

ribonucleoprotein particule)

 $SSU: small\ subunit$ 

T: thymine

U (UE) : unité enzymatique

U: uracile

UTR: untranslated region

Ψ : pseudouridine Y : pyrimidine

# Sommaire

Remerciements	1
Résumé	3
Abstract	
Abréviations	
SommaireAvant-propos	
Introduction	
I. Le domaine des archées	
A. Découverte	12
B. Diversité et classification	13
Des phénotypes et des écosystèmes variés	13
2. Les principaux phyla et leur classification	15
3. Des outils biotechnologiques et des modèles d'études	16
4. Des outils génétiques limités	17
C. Caractéristiques générales	19
1. Les caractéristiques cellulaires	19
a) Le métabolisme	19
b) Parois et membranes cellulaires	20
c) Les éléments génétiques	21
(1) Les éléments transposables	21
(2) Les plasmides	22
(3) Les virus	23
d) L'immunité cellulaire : les CRISPR	24
2. Les systèmes cellulaires principaux	24
a) La réplication, la réparation et la recombinaison	24
b) La transcription	26
(1) La machinerie de transcription	26
(2) Les facteurs de transcription	27
(3) Organisation des promoteurs	29
c) La traduction	30
(1) Les ARNm et l'initiation de la traduction	31
(2) Maturation et modification des ARNr et les ARNt	

	d)	Dégradation des ARNm	40
	e)	Ciblage des protéines à la membrane : l'ARN SRP	42
D.	Les	Thermococcales et Pyrococcus abyssi	43
	a)	Caractéristiques générales	43
	b)	Les génomes séquencés	43
	c)	Les éléments génétiques	44
	d)	Pyrococcus abyssi, notre modèle d'étude	45
II. L A.		RN régulateursARN régulateurs chez les bactéries	
	a)	Les ARN régulateurs modulant l'activité ou la quantité de protéines	48
	b)	Les éléments cis-régulateurs : les riboswitch	49
	c)	Les ARN régulateurs antisens	51
	(1	) Les ARN régulateurs cis-antisens	51
	(2	) Les ARN régulateurs <i>trans</i> -antisens	53
	d)	Les CRISPR	54
2	. Le	es ARN non codant régulateurs chez les archées	56
	a)	Recherche d'ARN non codant chez les archées	57
	(1	) Les approches bioinformatiques	57
	(2	) Les approches dites « biologiques »	61
	b)	Fonctions des nouveaux ARN non codant chez les archées	63
	(1	) ARNnc en cis	63
	(2	) Des ARNnc <i>cis</i> -antisens	64
	(3	) Des ARNnc issus de régions intergéniques	65
	(4	) Les protéines Sm et les ARN	65
	(5	) L'immunité cellulaire : fonction des CRISPR	66
		tifs du projet de thèse	
Résul			
I. P A.		etions bio-informatiques	
1.		incipe	
2		ésultats	
В.		alyse comparative de génomes : l'outil RNAsim	
1		incipe	
2		pplication au génome de <i>P. abyssi</i> et résultats	

C. Critères de sélection et combinaison des deux approches	74
II. Détection et caractérisation de nouveaux ARN chez <i>P. abyssi</i>	
B. Les transcrits supérieurs à 500 nt	76
C. Les transcrits inférieurs à 500 nt	77
Les ARN guides de modification à boites H/ACA	77
2. Les ARN issus des loci CRISPR	79
3. Les ARN issus de loci uniques	82
4. Les ARN issus de loci répétés	84
Discussion  A. Les limites des méthodes bioinformatiques et expérimentales	
B. Les ARN guides de modification à boîtes H/ACA et leur maturation	94
C. Le système CRISPR chez P. abyssi	94
D. Des ARNnc chez P. abyssi	96
1. Des ARNnc cis-régulateurs	96
2. Relation ARN régulateurs et transposases	98
E. Conclusion	99
Références bibliographiques	101
Annexes	

# **Avant-propos**

A la fin des années 50, Francis Crick a émis l'hypothèse du dogme de la biologie moléculaire représenté par le modèle schématique de la conservation et de l'utilisation de l'information génétique. Sur ce modèle, l'ARN est considéré comme molécule instable qui permet de transmettre l'information génétique de l'ADN à la machinerie de traduction ou qui est une ossature inactive (représentée par les ARNr et les ARNt) des machineries de traduction (Crick, 1970). Il est maintenant reconnu que l'ARN est fonctionnel, lui conférant le terme d'ARN non codant (ARNnc). Ce concept d'ARNnc date de la découverte initiale des ARN ribosomiques (ARNr). En effet, l'ARNr a longtemps été considéré comme ayant un simple rôle structural mais par la détermination de la structure cristallisée de la grande sousunité ribosomique, il a été montré que l'ARNr 23S (28S chez les eucaryotes) est fonctionnel dans le centre peptidyl-transférase du ribosome (Ban et al., 2000; Nissen et al., 2000). La phylogénie comparative des ARNr a permis de différencier le monde des archées, ou Archaea, de celui des bactéries (Woese and Fox, 1977). Les différentes études ont mis en évidence des similarités avec les domaines eucaryote et bactérien menant à de nombreux scénarios évolutifs entre les trois domaines du vivant (pour revue (Cavicchioli, 2010)). La découverte du domaine des archées a mis à l'ordre du jour le concept de LUCA (the last universal common ancestor) qui est considéré comme le dernier ancêtre commun à partir duquel les trois grands domaines du vivant reconnus actuellement se seraient différenciés : Archaea (archées), Bacteria (bactéries) et Eucarya (eucaryotes) (Forterre et al., 2005).

L'importance des ARNnc s'est révélée dans les trois domaines du vivant aussi bien chez les bactéries, les eucaryotes que chez les archées. Les ARNnc assurent diverses fonctions qui sont déterminées par leur structure tridimensionnelle. Ils sont retrouvés associés avec des protéines formant des complexes ribonucléoprotéiques (RNP) où l'ARN peut porter l'activité catalytique ou servir de « guide » pour les réactions telles que la modification de nucléotides ou le clivage d'une cible à un site spécifique

Chez les bactéries, les ARN non codant font généralement de 50 nt à 500 nt. *Escherichia coli* coderait 80 à 100 ARNnc (Gottesman and Storz, 2010). La fonction biologique est connue pour une vingtaine d'entre eux (pour revue, (Waters and Storz, 2009)). Chez les eucaryotes, les snoARN (small nucleolar) à boîtes C/D et H/ACA, les microARN,

les siARN (small interfering) et plus récemment, les piARN (piwiARN) ont été caractérisés (Aravin et al., 2007; Bushati and Cohen, 2007; Hannon et al., 2006). Les ARNnc bactériens ou eucaryotes sont retrouvés impliqués dans des processus de régulation transcriptionnelle ou post-transcriptionnelle de l'expression génique (pour revue (Costa, 2007; Gottesman and Storz, 2010)).

Chez les archées, deux familles majeures d'ARNnc ont été identifiés : les ARN guides de modification à boîtes C/D et H/ACA qui font partie de systèmes cellulaires de modification des ARNr et des ARNt (Bachellerie et al., 2002; Dennis et al., 2001; Gaspin et al., 2000; Muller et al., 2008; Omer et al., 2000) et les ARN issus de loci CRISPR qui sont des composants de systèmes d'interférence dirigés contre des acides nucléiques étrangers, les systèmes CRISPR/Cas et CRISPR/Cmr impliqués dans l'immunité cellulaire, qui ont aussi été caractérisés chez les bactéries (pour revue (Deveau et al., 2010)). De plus, d'autres ARN ont été identifiés sans être caractérisés et leur(s) fonction(s) cellulaire reste encore inconnue (Jager et al., 2009; Klein et al., 2002; Schattner, 2002; Soppa et al., 2009; Straub et al., 2009; Tang et al., 2002a; Tang et al., 2005; Wurtzel et al., 2010; Zago et al., 2005). Les travaux de recherche effectués au cours de cette thèse ont pour objectifs d'identifier de nouvelles familles d'ARNnc chez l'archée hyperthermophile *Pyrococcus abyssi* (*P. abyssi*) et de les caractériser par des critères de synthénie et de conservation en séquence.

Ce manuscrit est organisé en trois principales parties. La première partie introductive apporte les principales caractéristiques des domaines des archées et des ARNnc régulateurs. Une deuxième partie concerne les résultats obtenus. Enfin une troisième partie permet de discuter des résultats obtenus et d'annoncer les perspectives de ce travail.

# Introduction

# I. Le domaine des archées

#### A. Découverte

Durant plusieurs décennies, le monde du vivant était divisé en deux grands règnes : les eucaryotes et les bactéries. Il a fallu attendre le développement des études de phylogénie moléculaire pour remettre en question cette dichotomie. En 1977, sur la base de la comparaison de séquences des ARNr de la petite sous-unité du ribosome (16S et 18S) de plusieurs organismes, C. Woese et G. Fox ont proposé une nouvelle taxonomie du monde du vivant (Woese and Fox, 1977). Les ARNr abondants et essentiels pour toutes cellules, sont conservés au cours de l'évolution. Ils sont composés de domaines conservés et variables qui permettent de calculer une distance évolutive entre les organismes vivants. Ces analyses de phylogénie moléculaire ont permis de mettre en évidence qu'un groupe de microorganismes capables de produire du méthane constitueraient un domaine du vivant à part, bien qu'ils aient été longtemps confondus avec des bactéries. Ainsi, il a été établi un arbre universel du vivant divisé en trois groupes taxonomiques distincts : les « archaebacteria », les « eubacteria » et les « eukaryota » (Woese and Fox, 1977). Le domaine « archaebacteria » fut renommé une vingtaine d'années plus tard, Archaea (Woese et al., 1990). Trois domaines du vivant sont maintenant reconnus : Archaea (archées), Bacteria (bactéries) et Eucarya (eucaryotes).

La première archée de type halophile fut découverte en 1880 et des méthanogènes furent isolées dans les années 30, révélant ainsi que le domaine des archées fait l'objet d'études depuis plus de 130 années (pour revue, (Cavicchioli, 2010)). Aujourd'hui, plus de 300 souches ont été identifiées et plus d'une centaine de génomes sont séquencés, assemblés, annotés et disponibles publiquement dans la banque de données Genbank hébergée sur le site du NCBI. La taille des génomes est variable, pour le plus petit à 490 kb pour environ 590 ORF prédites (*N. equitans*, (Huber et al., 2003)) à un des plus grands à 5,75 Mb pour environ 4700 ORF (*M. acetivorans*, (Galagan et al., 2002)). Les systèmes moléculaires des archées et des eucaryotes partagent diverses caractéristiques ce qui permet d'élucider la certains mécanismes moléculaires eucaryotes. En effet les systèmes d'archées sont souvent utilisés et étudiées comme modèles moléculaires simplifiés.



Figure 1: Exemples de niches écologiques d'archées. (A) Etangs salés de couleur rose de la Baie de San Francisco (source : http://commons.wikimedia.org/wiki/File:Salt\_ponds\_SF\_Bay\_(dro!d). jpg ). (B) Fumeur noir du site Ashadze (source : Ifremer, actuenvironnement.com). (C) Fumeur blanc du site « Champagne », volcan Eifuku sous-marin (source : http://oceanexplorer.noaa.gov/explorations/04fire/logs/hirez/chimneys\_hirez.jpg).

#### B. Diversité et classification

#### 1. Des phénotypes et des écosystèmes variés

Le domaine des archées est constitué de microorganismes extrêmement variés phénotypiquement et souvent associés à des « microbes de l'extrême ». Un certain nombre d'entre elles ont été découvertes dans des milieux extrêmes allant des sources chaudes terrestres jusqu'aux lacs hypersalés. Elles peuvent constituer jusqu'à 20% de la biomasse terrestre totale (DeLong and Pace, 2001). Sur la base des caractéristiques connues sur les archées cultivable en laboratoire et de leurs habitats d'origine, les archées présentent différents phénotypes tels que les halophiles, les thermophiles et hyperthermophiles, les acidophiles, les alcalophiles, les archées réductrices d'ammonium et d'azote ou encore les méthanogènes.

Les archées halophiles ont la capacité de croitre en présence de fortes concentrations en sels pouvant aller jusqu'à 5,2 moles/L de NaCl ( 300 g/L) (Baliga et al., 2004). Elles ont été isolées à partir de milieux hypersalés tels que le Grand lac salé (Utah), le lac Owens (Californie), le lac Rose (Sénégal) ou encore de la Mer Morte où a été identifiée *Haloarcula marismortui* (Baliga et al., 2004). Certaines de ces archées sont capables de synthétiser des pigments protéiques colorés qui donnent une couleur à l'eau, pour exemple, *Halobacterium salinarum* produit l'halorhodopsine qui donne une couleur rose caractéristique aux étangs salés observés dans la baie de San Fransisco (**Figure 1**) (Oesterhelt and Stoeckenius, 1971). Par ailleurs, d'autres archées halophiles ont aussi un caractère alcalophile qui leur permettent de croitre dans des milieux basiques (pH supérieur à 8) tels *Natronomonas pharaonis*, isolée dans le lac Gabara en Egypte (Gonzalez et al., 2010).

Les archées acidophiles croissent à faible pH, jusqu'à pH=0 pour *Ferroplasma acidarmanus* (Macalady et al., 2004). Les Thermoplasmales et les Sulfolobales ont été isolées dans des environnements à hautes températures extrêmement acides telles que des sources chaudes à production de sulfures de Iron Mountain en Californie (voir références citées dans (Macalady et al., 2004)).

Les archées thermophiles et hyperthermophiles sont cultivables respectivement à de hautes températures, de 60 à 80°C pour les thermophiles et au-delà de 80°C pour les hyperthermophiles (Prieur et al., 2004). La température de croissance maximum (113°C) est observée pour *Pyrolobus fumarii* (Blochl et al., 1997). Ces archées appartiennent aux phyla

des Euryarchaeotes telles que les Thermococcales (décrites dans la section « les Thermococcales et Pyrococcus abyssi) et des Crénarchaeotes telles que les Sulfolobales (voir la section « Les principaux phyla » ci-après). Certaines Thermococcales sont aussi qualifiées de piézophiles (ou barophiles) car elles ont été isolées dans des grands fonds océaniques dont l'environnement est soumis à de hautes pressions. C'est notamment le cas de Pyrococcus CH1, isolée sur des cheminées hydrothermales appelées « fumeurs noirs » sur le site Ashadze à 4080 m de profondeur de 41 MPa de pression hydrostatique (Figure 1) (Zeng et al., 2009). Il existe aussi des archées qualifiées de psychrophiles isolées dans des lacs de la Région Arctique ou encore dans des sédiments océaniques, qui sont capables de réduire des molécules contenant de l'azote telles que les méthanogènes (pour revue, (Margesin and Miteva, 2010)). Celles-ci sont retrouvées dans des habitats anoxiques et ont une respiration anaérobie stricte avec la capacité de réduire le dioxyde de carbone pour former du méthane (Deppenmeier, 2002). La méthanogénèse est la capacité à produire du méthane à partir de nombreux substrats tels que de simples molécules carbonées (dioxyde de carbone -CO<sub>2</sub>-, monoxyde de carbone -CO-), le méthanol, les méthylamines ou encore l'acétate, en condition d'anaérobiose. Ce processus biologique est d'une importance majeure dans la production naturelle de méthane. En effet, il est estimé que le méthane biologique représenterait ~85% de la totalité du méthane disponible (Cavicchioli, 2010). Methanocaldococcus jannashii est une des méthanogènes largement caractérisées. Son génome figure parmi les premiers génomes d'archées séquencés (Bult et al., 1996). Celle-ci a été isolée à proximité d'un « fumeur blanc » à 2600 m de profondeur dans la zone est de l'Océan Pacifique (Figure 1) (Bult et al., 1996).

Ces différentes données disponibles dans la littérature illustrent la diversité des archées. Bien que les archées soient souvent qualifiées d'extrêmophiles, certaines d'entre elles ont été retrouvées dans des niches écologiques moins « exotiques » tels que les eaux usées ou encore le tractus intestinal animal et humain (Dridi et al., 2009). Les archées sont donc ubiquitaires (Chaban et al., 2006). Par ailleurs, il n'existe pas encore d'archée décrite possédant un caractère pathogène. Malgré leur présence dans le système gastro-intestinal humain, les tissus vaginaux ou la cavité buccale, il n'est pas démontré que les archées soient pathogènes (voir références citées dans (Vianna et al., 2006)). Cependant, certaines travaux concernant l'étude de la population microbienne dans les tissus affectées par des parodonties et des endodonties (inflammation des gencives) évoquent une corrélation de ces pathologies avec la présence de *Methanobrevibacter oralis* dans des échantillons infectés (Vianna et al., 2006).

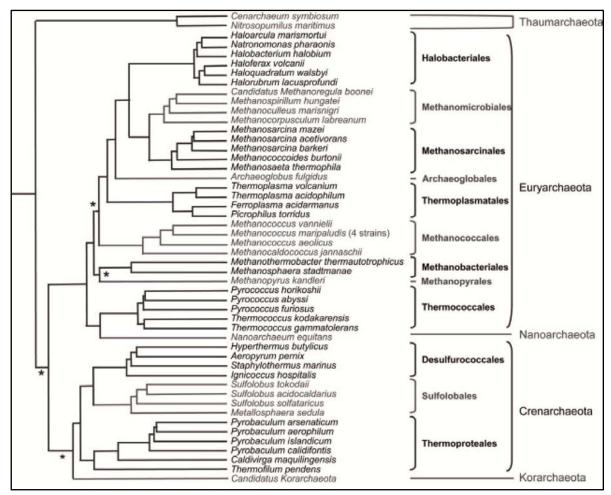


Figure 2 : Phylogénie des archées (tiré de (Gribaldo and Brochier, 2009)). Arbre consensus phylogénétique des archées basé sur des analyses phylogénomiques récentes. La taille des branches ne représente pas les distances évolutives. Les trois phyla connus (Crenarchaeota, Euryarchaeota et Thaumarchaeota) et les deux phyla additionnels proposés sont représentés (Korarchaeota et Nanoarchaeota). Les astérisques indiquent les racines mal résolues.

#### 2. Les principaux phyla et leur classification

Différentes classifications du domaine des archées ont été proposées en fonction de leurs caractères phénotypiques principaux ou suivant leur phylogénie. Six groupes phénotypiques sont proposés (voir ci-avant) (Ulrih et al., 2009). La classification phylogénétique apparait plus aisée à mettre en œuvre, notamment par le nombre grandissant de génomes séquencés. De nombreuses classifications basées sur la comparaison de gènes (gène codant l'ARN 16S pour exemple) permettent de distinguer deux principaux phyla : les Euryarchaeota et les Crenarchaeota (Pace, 2009; Woese and Fox, 1977; Woese et al., 1990). Cependant, la comparaison d'un seul gène n'est pas un signal phylogénétique assez fort pour obtenir un arbre avec des racines clairement résolues (Brochier-Armanet et al., 2008a; Robertson et al., 2005). La classification la plus récente a été proposée par Gribaldo et Brochier en 2009 suite à la compilation de nombreux arbres phylogénétiques basés sur la comparaison des systèmes macromoléculaires conservés chez les archées (les appareils de transcription, de traduction, de dégradation des ARN -exosome-, le système SRP -impliqué dans la signalisation des protéines pour leur sécrétion-) (Gribaldo and Brochier, 2009). Cette compilation propose les deux phyla majeurs avec l'émergence d'un troisième phylum, les Thaumarchaeota.

Les *Thaumarchaeota* (ou Thaumarchaeotes) apparaissent comme la lignée la plus ancienne (**Figure 2**). Deux ordres sont retrouvés : les Nitrosopumilales et les Cenarchaeales (Brochier-Armanet et al., 2008a). Ce phylum pourrait apporter des informations importantes sur un possible ancêtre commun entre les archées et les eucaryotes. En effet, ils apparaissent comme la seule lignée d'archées possédant une ADN topoisomérase de classe IB (TopoIB) qui apparait très proche de celle des eucaryotes et n'avait pas encore été identifiée chez les archées (Brochier-Armanet et al., 2008b). Cette caractéristique commune *Thaumarchaeota*/eucaryotes évoque une apparition plus ancienne de ce phylum par rapport aux autres phyla (Brochier-Armanet et al., 2008a).

Les *Crenarchaeota* (ou Crenarchaeotes) comprennent trois ordres : les Desulfurococcales, les Sulfolobales et les Thermoproteales (**Figure 2**). On y regroupe des espèces thermophiles et hyperthermophiles, autotrophes et hétérotrophes, anaérobies strictes à microaérophile. Parmi les genres hyperthermophiles, on retrouve entre autre les *Sulfolobus*, *Desulfurococcus*, les *Thermofilum*, les *Pyrobaculum* et les *Pyrolobus*. Les Sulfolobales sont les membres les plus étudiés de ce phylum. Ce sont de rares hyperthermophiles capables de croitre en aérobiose et dont le métabolisme est chimio-organotrophe (pouvant utiliser des

sucres comme source énergétique) facilitant ainsi leur culture en laboratoire. Les Crenarchaeotes possèdent des caractéristiques qui leurs sont propres : diverses analyses comparatives de génomes montrent que les Crénarchaeotes sont dépourvues d'histones et qu'elles possèdent une organisation de l'opéron des ADNr qui leur est propre (voir section traduction) (Gao and Gupta, 2007; She et al., 2001).

Les Euryarchaeota (ou Euryarchaeotes) ont une répartition écologique plus large que les Crenarchaeota. Elles regroupent des méthanogènes, des halophiles extrêmes et des thermophiles extrêmes, réductrices de sulfure ou encore réductrices de dioxyde de carbone. Elles sont classées dans les ordres des Halobacteriales, des Méthanomicrobiales, des Méthanosarcinales, les Archaeoglobales, les Thermoplasmales, les Methanococcales, les Methanobacteriales, les Methanopyrales et les Thermococcales (Figure 2). De part la disponibilité d'un grand nombre de génomes séquencés et annotés, ce phylum est le mieux caractérisé. En particulier, les Halobacteriales comprennent des archées halophiles qui sont aisément cultivables en laboratoire telle que Haloferax volcanii, et pour lesquelles il existe des outils génétiques pour inactiver ou déléter des gènes (pour revue (Allers and Mevarech, 2005)). Les Methanosarcinales comptent parmi elles Methanosarcina mazeï dont le transcriptome a été obtenu par séquençage à haut débit (Jager et al., 2009) et Methanocaldococcus jannaschii qui fut la première archée dont le génome a été entièrement séquencé et annoté (Bult et al., 1996). L'ordre des Thermococcales, avec des temps de génération courts et des résistances à divers stress, font d'elles d'excellents modèles de laboratoire pour l'étude physiologique et enzymatique des archées hyperthermophiles.

Il est difficile de classer *Nanoarchaeum equitans* qui est l'unique représentant du phylum des *Nanoarchaeota* (ou Nanoarchaeotes) (**Figure 2**). Cette archée serait la plus petite cellule vivante connue à ce jour et possède le plus petit génome décrit (490 kb et environ 400 gènes) (Huber et al., 2003). C'est un symbiote/parasite obligatoire du Crenarchaeote hyperthermophile *Ignicoccus hospitalis*. Les *Nanoarchaeota* apparaissent comme une lignée ayant évoluée plus rapidement que les autres, ceci étant une caractéristique typique des organismes parasitaires et symbiotiques (Brochier et al., 2005). De même, les *Korarchaeota*, avec des données génomiques encore peu nombreuses, serait une lignée anciennement divergente des *Crenarchaeota* (**Figure 2**).

#### 3. Des outils biotechnologiques et des modèles d'études

De part leurs phénotypes variés, les archées présentent un fort intérêt biotechnologique. En effet, les protéines d'archées sont dotées de propriétés non conventionnelles de stabilité et d'activité dans des conditions extrêmes (van den Burg, 2003). Les dérivés de ces enzymes sont aussi bien utilisés dans les industries agro-alimentaires et biologiques que dans les laboratoires de recherche (pour revue, (de Miguel Bouzas et al., 2006)). En effet, le développement de la réaction de polymérisation en chaine de l'ADN a été une des avancées les plus importantes au cours de ces 25 dernières années. Plusieurs dérivés d'ADN polymérases d'archées sont maintenant régulièrement utilisés pour leur thermostabilité ou encore pour leur fidélité et leurs activités additionnelles telle que l'activité exonucléase (pour revue (van den Burg, 2003)). D'autres enzymes hyperthermophiles trouvent une application en génie génétique, comme la ligature de fragments d'ADN par les ADN ligases, mais aussi dans les techniques impliquant la coupure spécifique ou non de protéines, comme le séquençage peptidique ou le clivage de certains acides aminés.

Les propriétés de stabilité des protéines sont aussi exploitées pour l'étude des systèmes moléculaires tels que la réplication, la transcription et la traduction. En effet, les archées représentent de bons modèles d'étude notamment pour les systèmes de réplication de l'ADN car elles possèdent une version simplifiée et homologue de la machinerie cœur de réplication eucaryote (Duggin and Bell, 2006). Par ailleurs, la résolution de la structure cristallisée de l'ARN polymérase de *S. solfataricus* a révélé que celle-ci représentait également une structure simplifiée de l'ARN polymérase II eucaryote, permettant ainsi d'utiliser le modèle d'archée afin de comprendre les bases mécanistiques de la transcription chez les eucaryotes (Hirata et al., 2008). C'est aussi par la même approche que la résolution de la structure de la grande sous-unité du ribosome de *H. marismotui* a permis une meilleure compréhension des bases moléculaires du mécanisme de traduction (Ban et al., 2000).

#### 4. Des outils génétiques limités

Les premières archées découvertes étaient retrouvées dans des environnements présentant des conditions de vie extrêmes telles que de fortes salinités, l'anaérobiose, des températures élevées, une forte pression hydrostatique rendant leur mise en culture en laboratoire difficile. Dans sa dernière revue, R. Cavicchioli cite pour exemple l'archée halophile *Haloquadratum walsbyi* qui fut découverte en 1980 mais les conditions d'isolement et de culture ne furent maîtrisées qu'en 2004 (voir références citées dans (Cavicchioli, 2010)).

De part ces difficultés de mise en culture en laboratoire, les outils génétiques tels que des systèmes de transfert de gènes n'ont été développés que très récemment et ne sont applicables que pour certaines archées. Par ailleurs, très peu de marqueurs de sélection sont connus rendant le développement d'outils génétiques difficile. En effet, beaucoup d'antibiotiques utilisés comme marqueurs de sélection chez les bactéries sont inefficaces chez les archées. Pour exemple, les antibiotiques ciblant les membranes bactériennes tels que l'ampicilline n'ont aucun effet sur les membranes d'archées qui sont dépourvues de peptidoglycane (Allers and Mevarech, 2005). Toutefois, des techniques de transformation ont été développées chez les archées. De nombreux éléments génétiques chez les archées du genre Sulfolobus (plasmides et virus) sont maintenant connus et l'observation de transfert de gènes via un mécanisme de conjugaison ou encore par infection virale ont permis la maitrise d'un certain nombre de techniques de transfert de gènes (voir références citées dans (Sato et al., 2003)). Les méthodes de transformation sont diverses : l'électroporation apparait très efficace pour les Sulfolobus mais inefficace pour M. acetivorans, P. abyssi et les archées halophiles qui ne peuvent supporter d'être mises en présence d'une concentration en NaCl inférieure à 1 M (pour revue (Allers and Mevarech, 2005)); la transformation au chlorure de calcium a été développée chez M. voltae et adaptée avec succès chez T. kodakaraensis (Sato et al., 2003); la transfection d'ADN est également possible chez H. volcanii via l'utilisation de PolyEthylène Glycol (PEG).

Les archées halophiles sont capables de croitre en aérobiose sur des milieux solides, elles représentent alors le groupe phénotypique d'archées le plus facilement manipulable (pour revue (Soppa, 2006)). Par ailleurs, la mise en évidence d'un système de transfert de gène naturel chez ces archées a permis le développement d'outils génétiques tels que des vecteurs, des systèmes de sélection et des gènes rapporteurs (Allers and Mevarech, 2005; Cavicchioli, 2010; Soppa, 2006). Ainsi, il est possible de réaliser des délétions de gènes dans le chromosome de *H. volcanii* par la technique de « Pop-In/Pop-Out » dont le principe repose sur un système de recombinaison homologue (Allers and Mevarech, 2005; Hammelmann and Soppa, 2008). La première caractérisation phénotypique suite à la délétion d'ARNnc a été réalisée chez cette archée apportant ainsi les premiers indices sur les rôles biologiques des ARNnc chez les archées (Hammelmann and Soppa, 2008; Soppa et al., 2009; Straub et al., 2009).

### C. Caractéristiques générales

#### 1. Les caractéristiques cellulaires

#### a) Le métabolisme

Il est souvent décrit que les systèmes opérationnels des archées sont similaires à ceux décrits chez les bactéries (voir références citées dans (Allers and Mevarech, 2005)). Le nombre grandissant de génomes séquencés et annotés et l'obtention de protéomes de génomes d'archées apportent une quantité d'informations qui est notamment exploitée pour la construction des différentes voies métaboliques.

En fonction de leur respiration (aérobie ou anaérobie), les différents groupes d'archées possèdent des gènes qui leur sont propres. Pour exemple, les gènes de biosynthèse de l'hème (composé retrouvé dans de nombreuses protéines oxydo-reductrices) ne sont pas retrouvés chez les archées du genre *Pyrococcus* (Cohen et al., 2003) mais retrouvés dans les génomes de Sulfolobales et de méthanogènes (Brochier-Armanet et al., 2009).

Certains systèmes de translocation membranaire sont associés à la production d'un gradient d'ions qui permet de fournir de l'énergie aux processus métaboliques. Comme les bactéries, les archées possèdent des systèmes de transport de type ABC transporteurs (ATP binding cassette) impliqués dans l'import de sucres et polymères de sucres, de peptides ou encore d'acides aminés ((Noll et al., 2008) et pour revue (Ellen et al., 2010)). On retrouve aussi des systèmes de transport « secondaires » comprenant les systèmes membranaires de type symport et antiport qui sont importants pour les archées extrêmophiles (acidophiles et basophiles) afin de maintenir le pH cytoplasmique adéquat et créant une force protomotrice nécessaire à la production d'énergie (pour revue (Slonczewski et al., 2009)).

Le catabolisme de composés carbohydrates (sucres) et de peptides produit de l'énergie sous forme d'ATP. Les enzymes participant aux différentes voies de glucogénèse sont différentes en fonction des sous-groupes d'archées ainsi qu'en fonction de leur capacité à dégrader les variétés de sucres. Ainsi, les Thermococcales (notamment *T. kodakaraensis*), utiliseraient des enzymes impliquées dans la voie Embden-Meyerhof modifiée alors que chez le crénarchaeote *S. solfataricus*, on retrouverait plutôt des enzymes impliquées dans la voie d'Entner-Doudoroff (pour revue (Berg et al., 2010)).

Le catabolisme des peptides/protéines serait orchestré par de nombreuses peptidases/protéinases putatives qui sont généralement associées à des transporteurs de type ABC. On retrouve aussi le complexe de dégradation ubiquitaire, le protéasome (complexe multicatalytique de protéases, 20S chez *H. volcanii* (Humbard et al., 2010b), ou 26S chez *P. abyssi* (Cohen et al., 2003)). Bien que peu conservées chez les archées, des protéines de type ubiquitine-like (impliquées dans le ciblage de protéines au protéasome) seraient aussi retrouvées chez *H. volcanii* (Humbard et al., 2010a). Les archées sont aussi dotées de lipases, tel est le cas de *T. sibiricus* où des lipases putatives sont retrouvées permettant d'hydrolyser des substrats tryacyglycérides (donnant des molécules de glycérol et des chaines d'acides gras, le glycérol serait utilisé comme substrat énergétique) (Mardanov et al., 2009).

L'analyse comparative de génomes réalisée par (Peregrin-Alvarez et al., 2009) a montré que les protéines impliquées dans les voies de biosynthèse des acides nucléiques (purines et pyrimidines) sont conservées sous différentes isoformes chez les archées (pour revue, (Makarova et al., 1999)). Les voies de biosynthèse d'acides aminés sont en fonction de l'auxotrophie des différentes archées. Ces voies de biosynthèse impliquent différents ensembles de protéines selon les cellules (pour revue (Makarova et al., 1999)). Les méthanogènes font partie des rares organismes capables de synthétiser les acides aminés 21 et 22, la sélénocystéine et la pyrrolysine respectivement ((Rother and Krzycki, 2010) et voir références cités dans (Cavicchioli, 2010)). Ces acides aminés ont été caractérisés pour leur implication dans la suppression de codon stop mais leur synthèse semblerait aussi être étroitement liée au métabolisme de conversion énergétique uniquement retrouvé chez les archées méthanogènes : la méthanogénèse (Rother and Krzycki, 2010). Chez les archées méthanogènes, la quasi-totalité des protéines impliquées dans ce processus contiennent les acides aminés pyrrolysine et sélénocystéine (pour revue (Rother and Krzycki, 2010)).

#### b) Parois et membranes cellulaires

Les archées sont des microorganismes qui ont été longtemps confondus avec les bactéries de part leur petite taille, l'absence d'organites et de membrane nucléaire. La composition de leur paroi et de leur membrane cellulaire a permis de les distinguer des bactéries.

Contrairement aux bactéries, les parois d'archées sont dépourvues de peptidoglycane justifiant la résistance à l'attaque du lysozyme. Certaines archées de type Gram-positif

possèdent une paroi composée d'une variété de polymères complexes tels que la pseudomuréine retrouvée chez les méthanogènes. D'autres archées de type Gram-négatif présentent une couche protéique ou glycoprotéique externe à leur membrane cytoplasmique (pour revue et voir référence citées dans (Forterre et al., 2002)). Elles présentent par ailleurs une grande variété de morphologie suggérant l'existence de mécanismes de morphogénèse encore inconnus (Forterre et al., 2002). Comme les bactéries, certaines d'entre elles sont de forme ovoïde ou encore en bâtonnet dont le diamètre ou la longueur varie entre 0.6 μm et 8 μm. D'autres possèdent des morphologies atypiques en forme de disques ou encore de polygones tel que *Haloquadratum walsbyi* une archée halophile qui se présente sous la forme de carrés (Bolhuis et al., 2004).

Au niveau de leur membrane plasmique, la caractéristique remarquable est due à la nature de leurs lipides : les phospholipides bactériens/eucaryotes sont composés à la fois de molécules de glycérol liées par des liaisons esters à des acides gras linéaires de 16 à 18 atomes de carbone (partie hydrophobique) alors que chez les archées, les parties hydrophobes sont composées de chaines de molécules d'isoprène liées aux molécules de glycérol par des liaisons éthers (pour revue (Matsumi et al., 2010)). Parmi ces molécules lipidiques membranaires, on retrouve entre autres le diphytanylglycérol diéther (ou l'archaéol) et le diphytanyldiglycérol tétraéther (ou la caldarchaéol). Le caldarchaéol possède une structure typique uniquement retrouvée chez les archées : celui-ci est composé de deux groupements hydrophiles reliés par de longues molécules d'alcool isopréniques (jusqu'à 40 atomes de carbone) permettant la structuration de la membrane en une seule couche (voir références citées dans (Matsumi et al., 2010)). Par ailleurs, la composition lipidique de la membrane est variable d'une archée à une autre mais varie aussi en fonction des conditions environnementales (pour revue (Matsumi et al., 2010)). En effet, Matsumi et ses collaborateurs citent notamment pour exemple les hyperthermophiles dont les ratios lipidiques changent en fonction de la température de croissance.

#### c) Les éléments génétiques

#### (1) Les éléments transposables

Les archées possèdent un grand nombre et une grande variété d'éléments transposables : des séquences d'insertions (IS pour <u>i</u>nsertion <u>s</u>equence) et des éléments

transposables miniatures répétés et inversés (MITE pour <u>m</u>iniature <u>i</u>nverted-repeat <u>t</u>ransposable <u>e</u>lements) (Filee et al., 2007). Chez les archées, les IS sont généralement de courts segments d'ADN spécifiques (jusqu'à ~2 kb) contenant une (IS partiels) à deux ORF (qui codent la transposase responsable de la mobilité de l'IS) parfois flanquées par de courtes séquences inversées répétées (IR pour inverted repeat). Les MITE sont des IS non autonomes car ils sont délétés d'une portion ou de toutes les ORF codant la transposase (voir référence citée dans (Filee et al., 2007)).

L'étude réalisée par Filée et ses collaborateurs montrent la diversité des IS/MITE dans les 28 génomes séquencés disponibles au moment de l'étude (23 Euryarchaeotes et 5 Crenarchaeotes). Ce travail a mis en évidence que la distribution des IS/MITE était aléatoire avec une grande variation du nombre d'IS d'une espèce à l'autre (jusqu'à plusieurs centaines par génome). Par ailleurs, la majeure partie des IS identifiées serait apparentée aux familles d'IS bactériens suggérant un possible transfert latéral de gènes entre les archées et les bactéries.

Les premiers événements de transposition ont été mis en évidence chez des halophiles extrêmes (*H. salinarum* et *H. volcanii*) possédant un grand nombre d'IS suggérant que les éléments transposables joueraient un rôle primordial dans le réarrangement de gènes dans les génomes d'halophiles. Le modèle le plus intriguant correspond au génome de *S. solfataricus* pour lequel 443 IS sont identifiées (dont 297 sont partiels, source : IS Finder <a href="http://www-is.biotoul.fr/">http://www-is.biotoul.fr/</a>). Ces IS seraient impliquées notamment dans des évènements de délétion, des réarrangements chromosomiques et des inactivations de gènes par insertion (voir références citées dans (Filee et al., 2007)).

La régulation de la transposition chez les archées est peu connue. Il existe néanmoins des études suggérant que des mécanismes de régulation seraient orchestrés par des ARN non codant retrouvés chez les archées *S. solfataricus*, *P. furiosus* ou encore *M. mazeï* (Klein et al., 2002; Soppa et al., 2009; Straub et al., 2009; Tang et al., 2002a; Tang et al., 2005; Wurtzel et al., 2010; Zago et al., 2005), comme ceux caractérisés pour les éléments transposables bactériens IS *10* et IS *30* (aussi annotés Tn *10* et Tn *30*) (Arini et al., 1997; Case et al., 1989).

#### (2) Les plasmides

Les archées possèdent des plasmides sous forme « cryptique » (portant un gène unique) ou sous forme de méga-plasmides (réplicons secondaires en plus du chromosome)

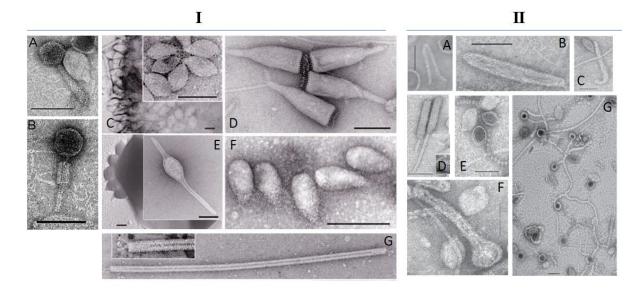


Figure 3 : Diversité de morphotypes de virus d'archées (images de microscopie électronique à transmission tirées de (Prangishvili et al., 2006; Sime-Ngando et al., 2010)).

- (I) Morphotypes de virus identifiés et caractérisés : A-B. Morphologie semblable à des bactériophages. C. Virus fusiforme SSV1 (Sulfolobus spindle-shaped virus1) et sa sortie de la cellule hôte. B. Virus ABV en forme de bouteille (Acidianus bottle-shaped virus). C. Virus ATV en forme bi-caudée (Acidianus two tailed virus) et sa sortie de la cellule hôte. D. Virus SNVD Forme de goutte d'eau (Sulfolobus neozealandicus droplet-shaped virus). E. Virus SIRV1 en forme de bâtonnet (Sulfolobus islandicus rod-shaped virus 1).
- (II) Morphotypes exceptionnels identifiés à partir d'échantillons du lac Retba (Sénégal) comprenant des virus sous formes : A. d'« épingle à cheveux », B. de bacilles, C. de crochets, D. de corps cylindriques rattachés à une tige, E. de chaînettes de globules, F. de « têtard », G. de complexes de filaments terminés par des corps sphériques.

(voir références citées dans (Soler et al., 2010)) Deux stratégies de réplication de plasmides sont connues chez les archées : la réplication en « cercle roulant » ou « rolling-circle » et la réplication en mode « theta » (désignant les réplications uni ou bidirectionnelles). La réplication en cercle roulant a notamment été décrite pour l'unique plasmide pGT5 de la Thermococcale *Pyrococcus abyssi* (Benbouzid-Rollet et al., 1997; Charbonnier et al., 1992; Erauso et al., 1996). Les plasmides d'archées les plus caractérisés sont retrouvés chez les *Sulfolobus* et ont servi au développement d'outils génétiques (voir références citées dans (Soler et al., 2007; Soler et al., 2010)).

#### (3) Les virus

Des virus ne présentant aucune similarité avec les virus des deux autres domaines sont capables d'infecter les cellules d'archées (Prangishvili et al., 2006b). Jusqu'en 2009, il était établi que leur génome était à ADN double brin (Prangishvili et al., 2006a). Une étude plus récente montre que le virus HRPV-1 de *Halorubrum*, isolé dans un marais salant en Italie, possède un génome à ADN simple brin (Pietila et al., 2009). Jusqu'à présent, environ une cinquantaine virus de Crénarchaeotes et d'Euryarchaeotes ont été identifiés et caractérisés (Pietila et al., 2009). Une différence majeure entre les virus de ces deux phyla est notable : seuls quelques virus de Crénarchaeotes sont capables d'induire un cycle lytique chez leurs hôtes alors que la majorité de ceux des Euryarchaeotes sont lytiques (virus appelés virions ou virus particule-like, VLP) (voir références citées dans (Pietila et al., 2009) et pour revue (Prangishvili et al., 2006a)).

Une autre distinction propre aux virus d'archées est leur diversité morphologique. Ils présentent des morphotypes singuliers: fusiformes (Fuselloviridae), bi-caudés (Bicaudaviridae), en forme de goutte ou de citron (Guttaviridae), de bouteille ou d'ampoule (Ampullaviridae), de bâtonnet ou de bacilles (Rudiviridae et Claraviridae), filamenteux (Lipothrixviridae) ou sphériques (Globuloviridae) (Figure 3) (pour revue (Prangishvili et al., 2006a)). Très récemment, d'autres nouveaux morphotypes ont été isolés à partir d'échantillons prélevés du lac salé Retba (Péninsule du Cap Vert au Sénégal) montrant la grande diversité morphologiques des virus d'archées : en forme d'« épingle à cheveux », de bacilles, de crochets, de corps cylindriques rattachés à une tige, de chaînettes de globules, de « têtard » ou encore de complexes de filaments terminés par des corps sphériques (Figure 3) (Sime-Ngando et al., 2010).

Protéine ou Fonction	Bactéries	Archées		Eucaryotes
ADN polymérase principales	Famille C	Familles B et D (Euryarchaeaotes)	Familles B (Crenarchaeotes)	Famille B
Protéines se liant à l'origine de réplication	DnaA	ORC/Cdc6		ORC + Cdc6
Hélicase réplicative	DnaB	MCM		MCM
Protéine SSB	SSB	RPA		RPA
Primase	DnaG	PriSL		Polα/primases Pri
Clamp	β-clamp	PCNA		PCNA

Tableau 1 : Composition des réplisomes dans les trois domaines du vivant. Les principales protéines décrites dans le texte sont indiquées avec leur fonction respective.

#### d) L'immunité cellulaire : les CRISPR

Une caractéristique partagée avec les bactéries est la présence de loci composés d'éléments répétés dans la quasi-totalité des génomes d'archées nommés CRISPR pour « Clustered Regularly Interspaced Palindromic Repeats » (Mojica et al., 2005). Ce sont des séquences d'ADN composées de séquences répétées directes en tandem (séquences DR) de 21 à 48 bp séparées par des séquences non répétées de 26 à 72 nt appelés « espaceurs » (pour revue (Deveau et al., 2010)). Ces séquences espaceurs présenteraient des similarités avec des éléments extra chromosomiques tels que les génomes de virus (des virus d'archées : les rudivirus et les fusellovirus) et des plasmides (Bolotin et al., 2005; Mojica et al., 2005; Pourcel et al., 2005). Un lien entre les séquences espaceurs et des séquences virales suggère que les cassettes CRISPR seraient impliqués dans un système de défense contre l'invasion d'acides nucléiques étrangers via un mécanisme semblable à l'interférence à ARN connue chez les eucaryotes (Bolotin et al., 2005; Lillestol et al., 2006). Cette hypothèse a été renforcée suite au séquençage complet du génome du virus ATV : neuf régions de son génome sont similaires à des séquences espaceurs de CRISPR de S. solfataricus P2 (Shah et al., 2009). Par ailleurs, des études de recherche d'ARNnc réalisées chez A. fulgidus et S. solfataricus ont mis en évidence que ces loci CRISPR étaient transcrits en ARN (Tang et al., 2002a; Tang et al., 2005). Une étude chez P. furiosus a montré que les ARN transcrits à partir des loci CRISPR (ARN Cr) seraient directement impliqué dans un mécanisme d'interférence à ARN (détaillé dans le chapitre ARN régulateurs) (Hale et al., 2008).

### 2. Les systèmes cellulaires principaux

#### a) La réplication, la réparation et la recombinaison

La réplication de l'ADN est constituée de trois étapes principales : l'initiation qui se fait à une origine de réplication, l'élongation faisant intervenir des ADN polymérases et la terminaison (pour revue, (Grabowski and Kelman, 2003)).

Chez certaines archées, la structure du chromosome et le mode de réplication sont similaires aux bactéries. Elles possèdent un à plusieurs chromosomes circulaires (*Haloarcula marismortui* en possède deux (Baliga et al., 2004)). Ces chromosomes circulaires peuvent

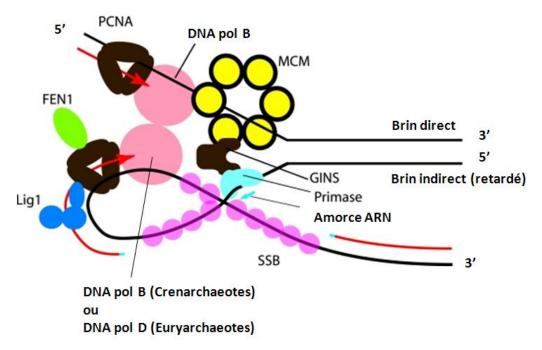


Figure 4: Modèle de l'architecture du réplisome chez les archées lors du processus d'élongation (Tiré et adapté de (Barry and Bell, 2006)). Les brins d'ADN matrices sont représentés en noir (Brins direct 5'-3' et indirect 3'-5') et les brins d'ADN nouvellement synthétisés sont représentés en rouge. Les ADN polymérases (de la famille B ou D) sont représentées par des pastilles roses, l'hélicase réplicative MCM est représentée par un anneau de pastilles jaunes, les protéines SSB de liaison à l'ADN simple brin sont représentées par des pastilles mauves, le complexe de clamp PCNA est représenté par des anneaux de pastilles ovales noires, la primase est indiquée en bleu clair, les protéines accessoires FEN1 et Lig1 sont respectivement indiquées en vert et en bleu foncé. Les auteurs de ce schéma proposent que le complexe GINS servirait de « pont moléculaire » entre la primase et l'hélicase MCM et les deux ADN polymérases seraient couplées.

contenir une à trois origines de réplication (ORB pour « Origin Recognition Boxes »)(pour revue (Barry and Bell, 2006)). Les premières études du processus de réplication chez les archées ont été réalisées dans les génomes du genre *Pyrococcus* (Myllykallio et al., 2000) et continuent à attiser un grand intérêt, notamment chez les Sulfolobales qui contiennent plusieurs ORB ((Duggin et al., 2008) et pour revue (Duggin and Bell, 2006)). Ces études ont montré une réplication bidirectionnelle avec des vitesses similaires à celles observées chez les bactéries (Myllykallio et al., 2000).

La machinerie de réplication (réplisome) de l'ADN d'archées est souvent décrite comme une machinerie eucaryote simplifiée (Tableau 1 et Figure 4) (Duggin and Bell, 2006). L'initiation de la réplication aux sites ORB se fait grâce à des facteurs d'initiation protéiques similaires aux facteurs eucaryotes: la protéine ORC/Cdc6 (origin recognition complex) et l'hélicase réplicative MCM (Mini Chromosome maintenance) dont les gènes sont retrouvés flanquant à ORB ((Myllykallio and Forterre, 2000) et pour revue (Barry and Bell, 2006)). De plus, le réplisome comprend les protéines ubiquitaires SSB (Single-stranded binding proteins) protégeant l'ADN simple brin, une primase homologue aux primases eucaryotes PriS et PriL qui synthétise des amorces ARN nécessaires à la réplication du brin retardé de l'ADN, un complexe homologue au complexe GINS eucaryote qui permettrait d'établir une interaction entre l'hélicase MCM et la primase, des ADN polymérases (décrites ci-après), et des protéines accessoires intervenant dans la processivité de la réplication (homologues au PCNA eucaryote, à l'endonucléase FEN1 et à l'ADN ligase I) (Figure 4) (pour revues (Barry and Bell, 2006; Duggin and Bell, 2006)).

Les ADN polymérases d'archées sont ubiquitaires et classées en six familles (A, B, C, D, X et Y) (Ohmori et al., 2001). Les Euryarchaeotes possèdent des ADN polymérases des familles B et D qui seraient respectivement impliquées dans la réplication des brins direct et indirect (ou retardé) (**Figure 4**) (Henneke et al., 2005). Ces deux familles d'ADN polymérases seraient aussi retrouvées dans les phyla des Korarchaeotes, Nanoarchaeotes et Thaumarchaeotes (Tahirov et al., 2009). Chez les Crenarchaeotes, seules des ADN polymérases de la famille B sont retrouvées en plusieurs copies suggérant que ces différentes ADN polymérases seraient spécifiques à la réplication de l'un ou l'autre des deux brins (Duggin and Bell, 2006). Par ailleurs, les ADN polymérases d'archées possèdent une activité 3' → 5' exonucléase qui joue un rôle dans les mécanismes de relecture ou « proofreading » (voir références citées dans (de Koning et al., 2010)). De plus, des mécanismes permettent de corriger les erreurs survenant durant et/ou après le processus de réplication qui sont les systèmes de « Mismatch repair » (chez quelques rares Euryarchaeotes), d'« excision repair »

(comprenant le « Base excision repair » et le « Nucleotide excision repair ») (pour revue (de Koning et al., 2010)) et le mécanisme ubiquitaire de recombinaison homologue impliquant RadA, l'homologue des protéines RecA bactérienne et Rad51 eucaryote (Haldenby et al., 2009).

La terminaison de la réplication chez P. abyssi semblerait s'effectuer dans la région opposée à l'origine de réplication bien qu'un site terminus de réplication ne soit pas encore identifié (Myllykallio et al., 2000). Très récemment, il a été mis en évidence que les génomes d'archées possèderaient un système de résolution de dimères de chromosomes comprennent des orthologues des recombinases XerC/D bactériennes (Cortez et al., 2010; Duggin et al., 2010; Sherratt, 2003). Chez P. abyssi, la protéine XerA a été identifiée comme homologue aux recombinases bactériennes XerC/D (Cortez et al., 2010). Celle-ci est capable de réaliser une recombinaison spécifique de site sur un plasmide portant la séquence dif-like prédite chez P. abyssi. Comme chez E. coli, ce site dif-like serait retrouvé dans la région terminus de réplication. En revanche, chez S. solfataricus où trois origines de réplication sont identifiées sur le chromosome, il a été montré que la terminaison de la réplication et la résolution des dimères étaient deux événements distincts (Duggin et al., 2010). La terminaison de la réplication se ferait plutôt sur un mode eucaryote, en d'autres termes, dans des zones aléatoires de collision des fourches de réplication. Cependant, un homologue de la protéine XerD a été aussi identifié. Des analyses expérimentales ont permis de mettre en évidence l'existence d'un site dif-like distant des zones de collision de fourches de réplication (Duggin et al., 2010).

#### b) La transcription

#### (1) La machinerie de transcription

La machinerie de transcription chez les archées peut être comparée à une version simplifiée de celle de l'ARN polymérase II des eucaryotes comprenant une ARN polymérase II-like, les facteurs de transcription principaux TBP (TATA-Binding protein), TFB (transcription factor B) et les facteurs accessoires TFE, et TFS (respectivement orthologues aux facteurs eucaryotes TBP, TFIIB, TFIIEα et TFIIS) (**Tableau 2**) (Grohmann et al., 2009a).

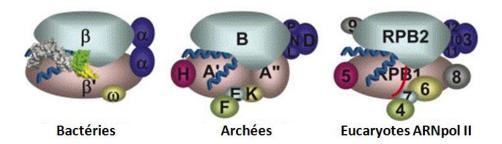


Figure 5 : Représentation schématique des ARN polymérases des trois domaines (Tiré de (Shibata et al., 2007)). Les sous-unités et les facteurs auxiliaires listés dans le Tableau 2 sont représentés.

		Bactéries	Archées	Eucaryotes (ARN polII)
Sous-unités de l'ARN polymérase		β'	A'/A''	RPB1
	Cœur catalytique	β	B'/B''	RPB2
	J 1			
		α	D	RPB3
			N	RPB10
		α	L	RPB11
	Plateforme		P	RPB12
$\frac{N}{N}$	d'assemblage			
ľĄĮ				
de				
ités		ω	K	RPB6
nn			Н	RPB5
-sn			F	RPB4
Sc	Facteurs auxiliaires		E	RPB7
			TFS*	RPB9
			$\mathbf{G}^{\circ}$	RPB8
			TBP	TBP
_	Initiation de la transcription		TFB	TFIIB
Facteurs de transcription			TFEα	TFIIΕα
		2222	TFS*	TFIIS
		σ 2.3-3.2	TFEβ/Rpc34	TFIIEβ
		σ 1.2-2.3		TFIIEα TFIIEβ
		0 1.2-2.3	Spt4	Spt4
	Elementina	NusG	Spt5	Spt-5
cte		NusA	NusA	Spec
Fa	Elongation			

Tableau 2 : Composition des principaux composants de la machinerie de transcription dans les trois domaines du vivant (adapté de (Grohmann and Werner, 2010)). Les sous-unités de l'ARN polymérase sont indiquées suivant leur groupe fonctionnel. TFS\*: le facteur TFS est homologue à la sous-unité RPB9 de l'ARN pol II ainsi qu'au facteur d'élongation TFIIS. G°: la sous-unité G est uniquement retrouvée dans les génomes de Crénarchaeotes hyperthermophiles.

L'ARN polymérase d'archées (aRNAP) comporte 10 à 14 sous-unités ((Grohmann et al., 2009a, b) et pour revues (Grohmann and Werner, 2010a; Werner, 2007)). Celles-ci peuvent être décomposées en trois groupes fonctionnels : (i) le cœur catalytique comprenant les sous-unités A'/A'' et B'/B'' (70% de la masse totale du complexe); (ii) la plateforme d'assemblage (sous-unités D, N, L et P) et (iii) des sous-unités auxiliaires (K, H, F, E, TFS et G) qui permettent de stabiliser le complexe enzyme-ADN-ARN en présence de facteurs de transcription auxiliaires (**Figure 5 et Tableau 2**) (Grohmann and Werner, 2010a).

Le cœur catalytique des ARN polymérases des trois domaines est conservé (**Tableau**) (Grohmann and Werner, 2010b). Cependant, les ARN polymérases d'archées et eucaryotes possèdent des caractéristiques qui leurs sont propres et qui les distinguent du domaine bactérien avec de un nombre de sous-unités auxiliaires plus important (Figure 5 et Tableau 2). Parmi eux, les deux sous-unités F/E apparaissent comme les éléments structuraux retrouvés dans les ARN pol d'archées et d'eucaryotes (Grohmann and Werner, 2010b). Ces deux sous-unités seraient impliquées dans les trois étapes fondamentales de la transcription comprenant l'initiation, l'élongation et la terminaison (pour revues (Grohmann and Werner, 2010a, b)). Lors de l'initiation de la transcription, les sous-unités F/E permettraient de faciliter la séparation des deux brins d'ADN et ainsi de moduler le chargement de l'ARN pol au niveau des promoteurs. De plus, leur interaction avec la matrice ARN néosynthétisée permettrait de stabiliser le complexe d'élongation et ainsi d'augmenter la processivité de l'enzyme (Hirtreiter et al., 2010b). Cette interaction serait aussi impliquée dans le phénomène de « pause » lors de la terminaison de la transcription. Les sous-unités F/E seraient capables de reconnaitre la transcription de stretchs de T (le signal polyU), signalant ainsi la terminaison et déstabilisant le complexe d'élongation (pour revue (Grohmann and Werner, 2010a)). En effet une étude récente réalisée chez Thermococcus kodakaraensis a mis en évidence des terminateurs de transcription caractérisés par des stretchs de 6 à 8 nucléotides T (Santangelo et al., 2009).

#### (2) Les facteurs de transcription

L'initiation de la transcription fait intervenir deux facteurs de transcription principaux : les protéines TBP (TATA box binding protein) et TFB (transcription factor B), respectivement homologues aux protéines eucaryotes TBP et TFIIB. Un troisième facteur non essentiel, TFE (homologue à TFIIEα eucaryote), stimulerait modérément la transcription à

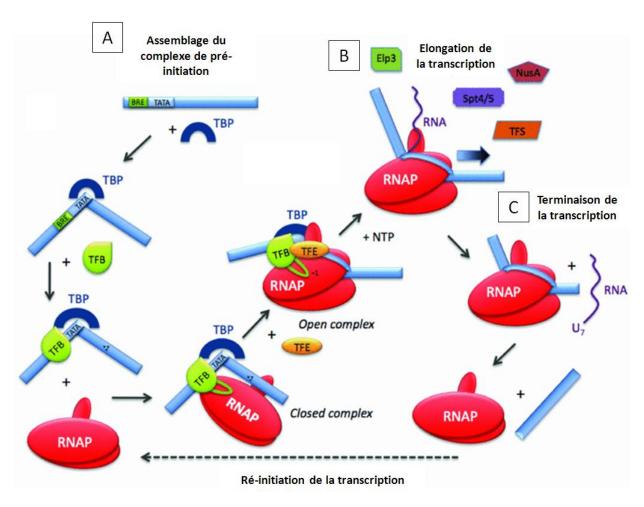


Figure 6 : Schématisation du processus de transcription chez les archées (tiré et adapté de (Grohmann et al., 2009)). Les trois étapes principales ((A) Initiation, (B) Elongation et (C) Terminaison) sont représentées en fonction des données de caractérisation disponibles. Les facteurs de régulation de la transcription connus sont représentés en fonction de leur implication dans les différentes étapes du processus.

partir de promoteurs faibles (voir références citées dans la revue (Grohmann et al., 2009a)). La protéine TBP se lie à la boîte TATA engendrant une courbure du promoteur et la formation du complexe d'initiation (**Figure 6**). Le facteur TFB est composé de trois domaines : un domaine C-terminal, qui se fixe sur le motif BRE du promoteur (B recognition element) permettant son interaction avec le complexe TBP-TATA et stabilisant le complexe de pré-initiation. Son domaine N-terminal recrute et interagit avec l'aRNAP et son domaine B-finger joue un rôle dans la stimulation de la réaction de la transcription (**Figure 6**). Après formation du complexe de pré-initiation, les sous-unités F/E de l'aRNAP stimule la séparation des deux brins d'ADN (**Figure 6**) (Grohmann and Werner, 2010a, b).

L'élongation et la terminaison de la transcription sont encore peu caractérisées chez les archées. Les génomes d'archées codent pour des homologues des facteurs d'élongation eucaryotes et bactériens TFS/TFIIS, NusG/Spt5 et NusA (Lange and Hausner, 2004). La protéine TFS (homologue de TFIIS eucaryote) caractérisée chez M. thermolithotrophicus serait impliquée dans les mécanismes de fidélité et de relecture du processus de transcription (Lange and Hausner, 2004). Très récemment, la structure cristallisée du complexe Spt4/5 de M. jannashii a été résolue (Hirtreiter et al., 2010a). Ces facteurs sont homologues aux Spt4/5 eucaryotes et Spt5 présente des homologies au niveau de deux domaines du facteur NusG bactérien (Hirtreiter et al., 2010a). Les auteurs ont suggéré que Spt4/5 chez les archées réguleraient l'élongation de la transcription par modulation des interactions aRNAP-ARN et aRNAP-ADN probablement en facilitant la distinction entre l'ADN et l'ARN par l'aRNAP (Hirtreiter et al., 2010a). La protéine NusA est retrouvée dans les génomes bactériens et d'archées (Ha et al., 2010). Chez les bactéries NusA est impliquée dans la régulation de l'élongation et de la terminaison de la transcription (Ha et al., 2010). Chez les archées, la caractérisation de NusA basée sur l'obtention de la structure cristallisée de NusA de A. pernix a montré sa capacité à se lier à des ARN synthétiques (Shibata et al., 2007; Tanaka et al., 2007).

Très récemment, une étude bio-informatique a révélé que des archées des phyla des Crenarchaeotes et des Thaumarchaeotes possèderaient un homologue de la sous-unité Rpc34 de l'ARN pol III eucaryote (responsable de la transcription des ARN 5S, ARNt et quelques ARN non codant) (Blombach et al., 2009). Le contexte génomique de ce gène suggère par ailleurs qu'il serait probablement co-transcrit avec des gènes codant pour des protéines impliquées dans la maturation et la modification d'ARNr et d'ARNt (Blombach et al., 2009). Il est à noter que bien que l'aRNP soit plus proche de l'ARN pol II eucaryotes, les génomes d'archées possèdent aussi des régulateurs de transcription proches de ceux décrits chez les

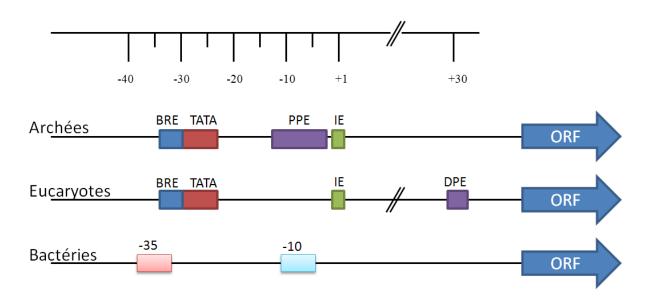


Figure 7 : Organisation des promoteurs dans les trois domaines du vivant (adapté de (Soppa, 1999)). Les éléments sont alignés par rapport à l'échelle donnée. Le site +1 représente le point +1 de l'initiation de la transcription. BRE : TFB Recognition Element, TATA : boite TATA, PPE : Proximal Promoter Element, IE : Initiator Element.

Groupes	Motif BRE	Boite TATA	Références
Halophiles	ND	TTTWWW (-29/-24)	(Soppa, 1999)
Methanogènes	ND	YTTATATA (-30-23)	(Soppa, 1999)
Sulfolobales	RNWAAW (-36/-31)	YTTTTAAA (-30-23)	(Bell et al., 1999;
			Soppa, 1999)
Pyrococcus furiosus	VRAAA (-36/-32)	TTWWWAW (-30-23)	(van de Werken et
			al., 2006)
Pyrococcus abyssi	RAAANN (-36/-32)	TTWWWWA (-30/-23)	(Cohen et al.,
(prédiction)			2003)
Methanorsarcina	NRAAA (-37/-33)	TTTATATA	(Jager et al.,
mazeï		Ou TTTWWATA (-30/-23)	2009)

**Tableau 3 : Séquences consensus des motifs promoteurs chez les archées.** N = A/C/G/T, W = A/T, Y = C/T, R = G/A, V = C/G/A. Leurs positions sont indiquées par rapport au site de départ de transcription +1.

bactéries (pour revue (Bell and Jackson, 2001)). Ces régulateurs transcriptionnels seraient impliqués dans le contrôle de l'expression génique de systèmes métaboliques spécifiques ou lorsque les organismes sont confrontés à des stress environnementaux (Jager et al., 2009; van de Werken et al., 2006).

### (3) Organisation des promoteurs

Les promoteurs chez les archées comprennent généralement deux éléments de séquences majeurs, la boîte TATA (aussi appelée boîte A) et l'élément de reconnaissance B (BRE pour B recognition element) et deux éléments moins bien caractérisés qui sont l'élément initiateur (IE pour Initiator Element) et l'élément promoteur proximal (PPE pour proximal promoter element) (Figure 7) (Soppa, 1999b). La boîte TATA est une séquence riche en A/T généralement centrée à la position -25 par rapport au site +1 de transcription (Reiter et al., 1990). L'élément BRE est une séquence de 6 à 8 nt riche en purine située juste en amont de la boîte TATA où se fixe la protéine TFB (Figure 7). Cet élément augmente le recrutement de la TBP au niveau du promoteur (Qureshi and Jackson, 1998) et définit l'orientation de la machinerie transcriptionnelle (Bell et al., 1999). Les séquences additionnelles IE et PPE seraient impliquées dans la régulation de l'initiation la transcription au promoteur. La séquence IE est souvent retrouvée à la position du départ de transcription ou à proximité, elle est généralement composée de dinucléotides pyrimidine-purine. L'élément PPE chevauchant est généralement riche en A/T situé entre les positions -12 et -1 par rapport au site de départ de la transcription (Soppa, 1999b). Les éléments IE et PPE ne seraient conservés que dans un sous-groupe de promoteurs et ils apparaissent moins conservés que les boîtes TATA et BRE (Kessler et al., 2004) suggérant que le nombre de facteurs nécessaires à l'initiation de la transcription sont différents en fonction des sous-groupes d'archées (Soppa, 1999a).

Différentes études statistiques de fréquence de nucléotide ont révélé que les séquences consensus et les positions des motifs promoteurs BRE/TATA sont différents en fonction du sous groupe d'archées (**Tableau 3**) (Jager et al., 2009; Soppa, 1999a; Torarinsson et al., 2005; van de Werken et al., 2006). Ces faibles conservations en séquences et en positions ont été corrélées aux génomes contenant des transcrits dits « leaderless », transcrits dont les séquences 5'UTR (untranslated regions) sont plus courtes (voir section synthèse protéique/traduction) (Torarinsson et al., 2005). Par ailleurs certains génomes d'archées

codent pour plusieurs paralogues de facteurs de transcription tel que *T. kodakaraensis* où l'on retrouve deux protéines TFB (notées TFB1 et TFB2) (Santangelo et al., 2007). Leur activité est redondante et l'une ou l'autre des deux protéines est suffisante pour la transcription *in vitro*. Bien qu'il n'y ait aucune différence pour la reconnaissance des promoteurs (d'ARNr, d'ARNt ou encore d'ARNm) par l'un ou l'autre des deux TFB, il semblerait que leur efficacité de reconnaissance soit dépendante des conditions de concentrations en sel utilisées dans le système de transcription *in vitro* (Santangelo et al., 2007). Les auteurs de ce travail ont également suggéré que les archées possédant plusieurs paralogues de protéines régulatrices (notamment les halophiles) pourraient utiliser différentes combinaisons du complexe d'initiation en fonction de la variabilité des motifs promoteurs (Bell and Jackson, 2001; Santangelo et al., 2007).

Comme les bactéries, les archées possèdent aussi des motifs promoteurs auxiliaires qui sont généralement associés à des régulateurs transcriptionnels répondant à des conditions de croissance spécifiques. Pour exemple, l'analyse du régulon glycolytique de *P. furiosus* a révélé la présence d'un motif de liaison répété et inversé en aval des motifs BRE/TATA (TATCAC-N<sub>5</sub>-GTGATA) (van de Werken et al., 2006) et l'analyse du transcriptome obtenu par séquençage à haut débit de *M. mazeï* a révélé que les gènes impliqués dans la réponse au nitrogène sont composés d'un motif consensus situé de part et d'autre des motifs de transcription basale (AGGAGGCA-N<sub>27</sub>-AAAGCTA) (Jager et al., 2009).

### c) La traduction

La compréhension des mécanismes traductionnels dans les trois domaines du vivant a été facilitée notamment par l'obtention et l'analyse de structures cristallisées de sous-unités des ribosomes dont la première fut celle de la grande sous-unité ribosomique de l'archée *Haloarcula marismortui* (Ban et al., 2000; Nissen et al., 2000). Ce n'est que très récemment qu'un couplage entre la transcription et la traduction comme chez les bactéries a été révélé chez les archées par l'observation de polysomes extraits de *T. kodakaraensis* (French et al., 2007).

La traduction est décomposée en deux processus essentiels : le couplage des acides aminés à leurs ARNt respectifs et le décodage de l'ARNm orchestré par les ribosomes (une petite sous-unité (30S pour les bactéries et les archées, 40S pour les eucaryotes) et une grande sous-unité (50S pour les bactéries et les archées et 60S pour les eucaryotes)). La traduction

Type d'ARNm	Acteurs nécessaires à l'initiation de la traduction			
SD AUG ARNm —	SSU 30S	Facteurs d'initiation IF  ARNt <sup>Met</sup>		
AUG ARNm —	SSU 30S	Facteurs d'initiation IF ARNt <sup>Met</sup>		
AUG —ARNm ——	LSU 50S SSU 30S	ARNt <sup>Met</sup> chargé dans la particule 70S		

Tableau 4 : Caractéristiques principales des trois mécanismes d'initiation de la traduction connus chez les archées en fonction des ARNm. Trois types d'ARN sont identifiés chez les archées : des ARNm possédant un 5'UTR portant un motif SD, des ARNm avec un 5'UTR dépourvu du motif SD et des ARNm leaderless (Colonne Type d'ARNm, de haut en bas). Les mécanismes sont encore mal connus bien que des acteurs soient déjà identifiés. Les sous-unités du ribosome sont notées SSU 30S et LSU 50S.

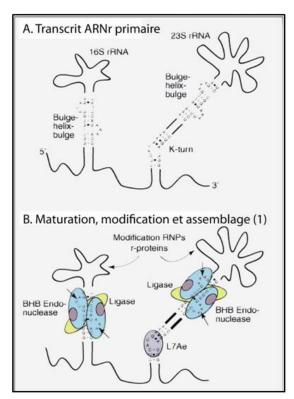
requiert de plus des facteurs protéiques auxiliaires. Des facteurs d'initiation sont identifiés chez les archées (aIF1, aIF1A, aIF2, aIF5B et aIF5B) sans que leurs rôles respectifs ne soient encore bien définis (Benelli and Londei, 2009). Pour l'élongation, les archées possèdent les facteurs d'élongation aEF1α et aEF2, homologues des protéines bactériennes EF-Tu et EF-G respectivement (eEF1A et eEF2 chez les eucaryotes) (voir référence citées dans (de Koning et al., 2010)). Très récemment, deux structures cristallisées obtenues chez *Aeropyrum pernix* ont révélé que des protéines homologues à Pelota et eEF1α eucaryotes seraient impliquées dans le contrôle de l'élongation lors de pause du ribosome (Kobayashi et al., 2010).

### (1) Les ARNm et l'initiation de la traduction

Chez les archées, les ARNm partagent des similarités avec les ARNm bactériens. Ils sont souvent polycistroniques : 50 à 70% des gènes d'un génome donné sont retrouvés organisés sous forme d'opérons polycistroniques contenant 2 à 3 gènes (Torarinsson et al., 2005). Contrairement aux eucaryotes, les ARNm d'archées sont généralement dépourvus d'introns. Leur extrémité 5' ne porte pas de coiffe méthylguanosine et leur extrémité 3' ne comprend pas de queue poly(A) de type eucaryote. Cependant, certains ARNm possèderaient de courtes queues hétéropolymériques riches en nucléotide A (voir références citées dans la revue (Evguenieva-Hackenberg and Klug, 2009)).

Deux études réalisées chez les deux crénarchaeotes, *S. solfataricus* et *P. aerophilum*, ont mis en évidence qu'il existait deux types d'ARNm, avec ou sans région 5'UTR (Slupska et al., 2001; Tolstrup et al., 2000). L'observation de transcrits « leaderless » a été étendue à d'autres génomes : le crenarchaeote *A. pernix* et les euryarchaeotes *Halobacterium* sp, *M. kandleri*, *M. thermoautotrophicum*, *P. horikoshii* et *T. volcanium* chez lesquels ces transcrits représenteraient de 31 à 74 % des transcrits totaux par génome (Brenneis et al., 2007; Fuglsang, 2004). Cependant, l'analyse globale du transcriptome par séquençage à haut débit de *M. mazeï* a montré que la majorité des ARNm séquencés possèderait de longs 5'UTR (Jager et al., 2009). Ces différentes données suggèrent la présence de différents mécanismes d'initiation au sein des archées.

Parmi les ARNm ayant une région 5'UTR, certains possèdent un motif Shine-Dalgarno (SD) ou Ribosome Binding Site (RBS) riche en résidus purines et centré en moyenne à 9 nt en amont du codon d'initiation (Torarinsson et al., 2005). Un motif SD consensus d'archées est difficilement définissable car ce motif est très variable d'un génome à



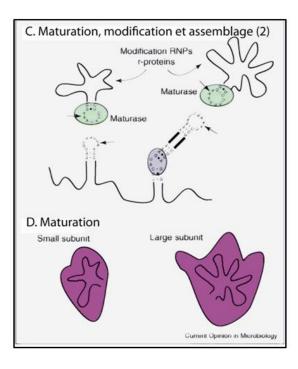


Figure 8 : Schématisation de la maturation et modification concomitantes des ARNr au cours de l'assemblage des ribosomes chez S. solfataricus. (Tiré de (Dennis and Omer, 2005)). Ce processus est composé de quatre étapes. (A) Le transcrit primaire ARNr est composé de deux hélices contenant les motifs BHB servant à la maturation des séquences des ARNr pré-16S et pré 23S. Il est à noter qu'un motif k-turn se forme au niveau de l'hélice de maturation de l'ARN 23S qui permet à la protéine ribosomale L7Ae de se lier (en violet). Celle-ci serait impliquée dans le recrutement des complexes de maturation/assemblage des particules ribosomiques. (B) Le clivage initial du motif BHB est réalisé par les endonucléases d'épissage spécifiques (en bleu). (C) La réaction de clivage est suivie de la ligature (par une ligase encore non caractérisée représentée en jaune) des introns (espaceurs internes et externes de l'ARN primaire) et des exons (pré-ARNr). Les processus de maturation des pré-ARNr ne sont pas encore très bien caractérisés et impliqueraient des endonucléases (maturase) telles que aDim2p (D) et des exonucléases.

un autre (Torarinsson et al., 2005). Il est généralement complémentaire aux 7 derniers nucléotides de l'extrémité 3' de l'ARNr 16S d'un génome donné (Torarinsson et al., 2005). L'initiation de la traduction des ces transcrits semblerait suivre un modèle bactérien : (i) la reconnaissance du codon d'initiation par la SSU 30S implique l'appariement entre le motif SD et l'extrémité 3' de l'ARNr 16S, (ii) les facteurs d'initiation auxiliaires aIF1A, aIF1 et aIF2 (aIF pour archaeal Initiation Factor) et l'ARNt<sup>Met</sup> initiateur viennent former le complexe d'initiation avec la SSU et (iii) la LSU 50S se rajoute au complexe afin de former la particule ribosomique 70S et de poursuivre la traduction par l'élongation (**Figure 8**) (pour revue (Benelli and Londei, 2009)). Chez les archées halophiles, certains ARNm possédant une région 5'UTR ne contiendraient pas de motif SD ou un autre motif conservé, suggérant un nouveau mécanisme d'initiation de la traduction (**Tableau 4**) (Hering et al., 2009).

Le mécanisme d'initiation d'ARNm leaderless chez les archées est peu caractérisé. Il semblerait aussi être retrouvé chez les eucaryotes et les bactéries (voir références citées dans (Brenneis et al., 2007; Hering et al., 2009; Malys and McCarthy, 2010)). Contrairement au mécanisme décrit ci-dessus, l'initiation de transcrits leaderless impliquerait le chargement de la particule ribosomique 70S non dissociée (**Tableau 4**). Ce chargement serait dépendant de la présence de l'ARNt<sup>Met</sup> initiateur dans la particule 70S afin de former le complexe d'initiation sur l'ARNm (voir références citées dans (Nakagawa et al., 2010)).

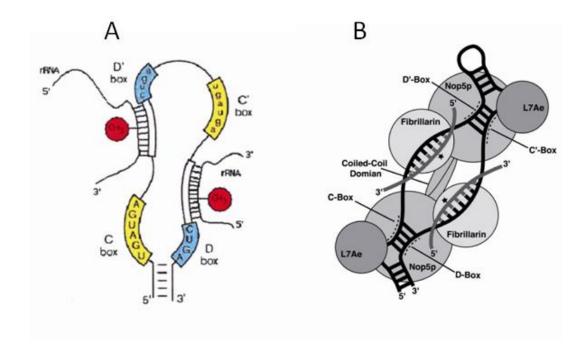
### (2) Maturation et modification des ARNr et les ARNt

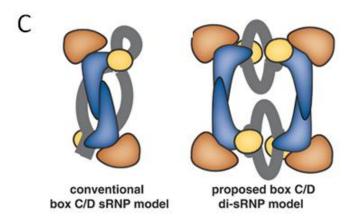
### (a) Les ARNr

Les premières études chez les archées ont mis en évidence que les gènes d'ARNr 16S et 23S étaient organisés sous forme d'opéron contenant des séquences absentes des ARNr matures, notamment chez les Crenarchaeotes *S. acidocaldarius*, *D. mobilis* et *A. pernix* ((Nomura et al., 1998) et pour revue (Garrett et al., 1991)). Certains opérons d'euryarchaeotes ribosomiques contiennent des gènes d'ARNt (Gao and Gupta, 2007; She et al., 2001). Ces ARNr sont transcrits sous forme d'un ARN primaire de taille moyenne de 5000 nt (**Figure 8**) (Dennis and Omer, 2005). Peu de données sont disponibles afin d'élucider les mécanismes de maturation des ARNr d'archées contrairement aux eucaryotes et aux bactéries. Deux structures secondaires sous forme de longues hélices encadrent les ARNr 16S et 23S d'archées formant un motif en « bulge-helix-bulge » (BHB) (**Figure 8**). Celui-ci est composé

d'une hélice de 4 nt séparée par deux protubérances de 3 nt (parfois 4nt) situées sur les brins opposés (Kjems and Garrett, 1988; Thompson and Daniels, 1990). Une endonucléase d'épissage identifiée chez Desulfurococcus mobilis reconnaît et coupe ce motif aux positions symétriques dans les deux protubérances générant des extrémités 2'-3'cyclic phosphate et 5'hydroxyl (Kjems and Garrett, 1988). De plus, l'hélice formée à la base du pré-ARNr 23S de Sulfolobus solfataricus contient un motif ARN structural en «Kink-turn» qui permet la liaison de la protéine ribosomique multifonctionnelle L7Ae (voir ci-après) (Tang et al., 2002b). La liaison de cette protéine serait importante pour l'organisation du complexe de maturation et permettrait de recruter les différents facteurs de maturation (Figure 8). Après action de l'endonucléase d'épissage, les produits de coupure sont ligaturés au niveau de leurs extrémités 2'-3'cyclic phosphate et 5'-hydroxyl par une ligase encore non caractérisée qui génère des intermédiaires circulaires, pré-16S et pré-23S (Figure 8) (pour revue (Lykke-Andersen et al., 1997)). Les réactions de maturation générant les formes matures à partir des formes circulaires restent à ce jour, non identifiées. Cependant, très récemment, l'homologue aDim2p de la protéine Dim2 de la levure Saccharomyces cerevisiae a été caractérisé chez P. horikoshii. Dim2 est connue pour son activité endonucléolytique sur les ARNr précurseurs (voir références citées dans (Jia et al., 2010)). La résolution de la structure du complexe de aDim2p de P. horikoshii avec l'ARNr 16S suggère que celle-ci serait impliquée dans la reconnaissance et la maturation de l'extrémité 3' de l'ARNr 16S chez P. horikoshii (Jia et al., 2010).

Les ARN ribosomiques font l'objet de modifications de nucléotides comprenant des réactions de méthylation (de bases ou de riboses), de déamination, d'isomérisation, de thiolation et de réduction de base (pour revue (Limbach et al., 1994)). Les nucléotides modifiés sont majoritairement localisés dans les régions fonctionnelles des ARNr leur suggérant un rôle pour la fidélité et l'efficacité de la traduction bien que la fonction spécifique de chaque nucléotide modifié soit encore mal connue (Decatur and Fournier, 2002). Chez les eucaryotes, les deux modifications les plus fréquentes retrouvées dans les ARNr sont des 2'-O-ribose méthylations et des pseudourydilations qui impliquent des complexes ribonucléoprotéiques contenant des ARNnc appelés ARN guides de modification à boîtes C/D et H/ACA respectivement (pour revues, (Dennis and Omer, 2005; Dennis et al., 2001; Omer et al., 2000; Omer et al., 2003)). De plus ces ARN guides pourraient jouer le rôle de chaperon pour un repliement fonctionnel de l'ARNr par le biais d'appariements des séquence guides avec l'ARNr précurseur naissant pendant le processus d'assemblage du ribosome (Dennis et al., 2001).





**Figure 9 : Caractéristiques des ARN guides de modification à boîtes C/D.** (A) Schéma simplifié d'un ARN guide C/D, les boites C et C' sont indiquées en jaune, les boites D et D' en bleu. Le 5ème nucléotide 2'-O-méthylé est indiqué en amont des boites D et D' (tiré de (Omer et al., 2000)). (B) Complexe RNP C/D. Chaque paire C/D et C'/D' forment un motif k-turn sur lequel se lie la protéine L7Ae. Les autres composants du complexe RNP sont indiqués. (C) Modèles des complexes RNP conventionnel et di-sRNP (tiré de (Bleichert et al., 2009)), les fribrillarines sont représentées en bleu, L7Ae en jaune, Nop5 en orange et l'ARN guide en gris.

### 2'-O-ribose méthylation par les complexes RNP à ARN guides à boîtes C/D

Le nombre élevé de sites de méthylation (67 au total) de l'ARNr de *S. solfataricus* (Noon et al., 1998) et l'identification de protéines homologues aux protéines composant les complexes RNP eucaryotes (Amiri, 1994; Lafontaine et al., 1998; Lafontaine and Tollervey, 1999) ont suggéré la présence d'une machinerie moléculaire de modification des ARNr similaire à celles des eucaryotes (Gaspin et al., 2000; Omer et al., 2000). En effet, chez les eucaryotes, il existe 50 à 100 sites de méthylation du ribose tandis que les bactéries n'en possèdent qu'un nombre restreint (4 sites chez *E. coli*) (voir références citées dans (Noon et al., 1998)).

Les ARN guides de modification à boîtes C/D d'archées sont des ARN courts d'environ 50 à 60 nt. Ils sont caractérisés par les boîtes à motifs conservés C (UGAUGA) et D (CUGA) qui sont localisés près des extrémités 5' et 3' de la molécule ainsi que par les motifs homologues C' et D' localisés dans la partie médiane de la molécule (Figure 9A) (Bachellerie et al., 2002; Cavaille and Bachellerie, 1998; Kiss, 2001). Les régions guides sont localisées en 5' des boîtes D et D' et la méthylation est dirigée sur le cinquième nucléotide de la séquence cible appariée en amont des boîtes D' ou D (loi du N+5) (Figure 9A) (Cavaille et al., 1996; Kiss-Laszlo et al., 1996; Nicoloso et al., 1996). Ces ARN font partie de complexes RNP comprenant les protéines homologues aux protéines des RNP C/D eucaryotes : L7Ae (Snu13 chez la levure et 15.5kD chez l'humain) (Watkins et al., 2000), Nop5 (ou aNop56/58 car similaire à la fois aux deux protéines eucaryotes Nop56 et Nop58) et la méthyltransférase fibrillarine (aussi appelée aFib et aNop1 car similaire à NOP1) (Figure 9B) (Reichow et al., 2007). Dans une réaction dynamique, ces complexes utilisent la complémentarité en séquence d'environ 10 nt entre les régions guides portées par l'ARN C/D et les régions cibles dans les ARNr et les ARNt (Clouet d'Orval et al., 2001; Dennis et al., 2001; Omer et al., 2000; Renalier et al., 2005). Le complexe catalyse ainsi l'insertion d'un groupement méthyl par la fibrillarine à la position 2'-O du ribose d'un nucléotide spécifique de la région cible (Figure 9A).

Contrairement à leurs équivalents eucaryotes, les complexes archées RNP guides C/D ont pu être reconstitués et sont actifs *in vitro* (Bortolin et al., 2003; Omer et al., 2002). Pour une 2'-O-ribose méthylation efficace, les appariements des boîtes C/D et C'/D' sont nécessaires pour l'assemblage symétrique de deux complexes contenant les trois protéines cœur L7Ae, Nop5 et aFib en complexe RNP (**Figure 9C**). Ces appariements C/D et C'/D' forment les motifs Kink turn (décrit ci-après) capable de fixer la protéine L7Ae. Des études

biochimiques (Rashid et al., 2003; Tran et al., 2003) et la résolution de la structure tridimensionnelle d'un complexe RNP guide C/D chez *M. jannaschii* montrent que le complexe RNP est composé d'un ARN guide C/D et deux copies de chacune des trois protéines cœur (**Figure 9C**). La microscopie électronique et la reconstruction *ab initio* grâce à l'analyse en « single particule » ont permis d'affiner la structure tridimensionnelle du complexe RNP guide C/D. Ainsi, il a pu être déterminé que non pas deux mais quatre séries de chacune des protéines L7Ae, Nop5 et aFib composent un complexe RNP guide C/D suggérant fortement l'implication de deux ARN guides C/D (**Figure 9C**). Ainsi, les complexes RNP à guides C/D seraient organisés en complexes di-sRNP dont l'activité enzymatique serait encore plus efficace (**Figure 9C**) (Bleichert et al., 2009). Plus récemment, une structure tridimensionnelle en présence d'un substrat ARN et du cofacteur S-Adénosylméthionine a été résolue. Par cette résolution, un modèle de catalyse croisée a été suggéré : la fibrillarine d'un sRNP modifierait le substrat apparié au sRNP opposé (Xue et al., 2010).

### Pseudouridylation par les complexes RNP à ARN guides à boîtes H/ACA

La pseudouridine (ψ), isomère de l'uridine, est un nucléotide modifié retrouvé dans la les ARNt et les ARNr. Chez les bactéries, quatre familles de pseudouridine synthases sont responsables de l'isomérisation site-spécifique de l'uridine en pseudouridine (Hamma and Ferre-D'Amare, 2006). Chez les eucaryotes et les archées, il existe une classe de pseudouridine synthases qui dépendent d'un ARNnc, formant un complexe RNP. Ce complexe RNP comprend un ARN guide à boîtes H/ACA, portant les séquences guides complémentaires à la cible ARN, et quatre protéines chez les archées : Gar1, aNop10, L7Ae (NH2P chez les eucaryotes) et aCbf5 (NAF57 chez les mammifères et la dyskerine chez l'humain), responsable de la catalyse (Lafontaine et al., 1998; Zebarjadian et al., 1999).

Généralement les ARN guides à boîtes H/ACA chez les archées sont composés d'une ou plusieurs tiges-boucle séparées par une région simple brin. Par analogie aux ARN guides à boîtes H/ACA eucaryotes, les ARN guides H/ACA d'archées sont dépourvus de la boîte H (ANANNA; N = G, A, T ou C). La boîte ACA (parfois AUA) est retrouvée dans une queue terminale, exactement à trois nucléotides de l'extrémité 3' de l'ARN (**Figure 10**) (Dennis and Omer, 2005; Kiss et al., 2010; Muller et al., 2008). Les tiges-boucles contiennent des boucles internes qui correspondent aux poches de pseudouridylation. L'ARN guide H/ACA spécifie l'uridine à modifier en ciblant l'ARNr en s'appariant de façon complémentaire sur la séquence bipartite de 10 à 15 nt située dans la poche de pseudouridylation. L'uridine non appariée est

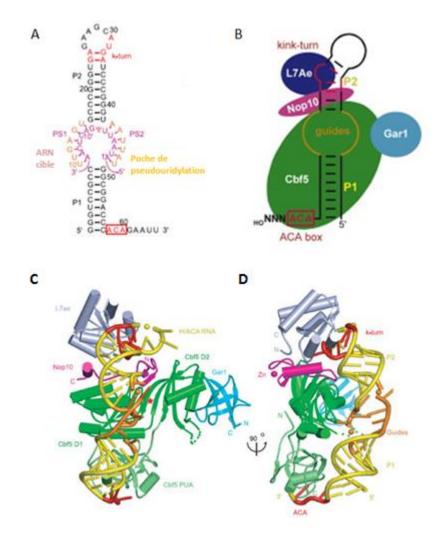
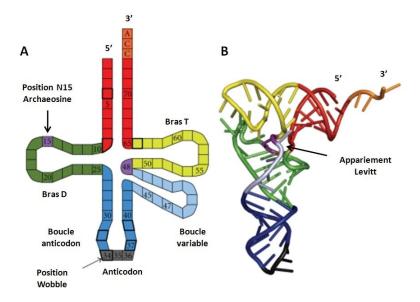


Figure 10: Caractéristiques des ARN guides de modification à boîtes H/ACA chez les archées. (A) Schéma simplifié d'un ARN H/ACA. La boîte ACA est indiquée, l'ARNr cible est représenté en violet dans les poches de pseudouridylation en orange où la réaction s'effectue. (Tiré de (Li and Ye, 2006)). (B) Complexe RNP H/ACA constitué d'un ARN en noir dont les poches de pseudourydilation sont en orange et des quatre protéines: L7Ae en bleu foncé, Nop10 en rose, Cbf5 en vert et Gar1 en bleu clair (Tiré de (Kiss et al., 2010)). (C) et (D) Caractéristiques structurales des RNP guides H/ACA (Tiré de (Li and Ye, 2006)), les domains catalytique et PUA de Cbf5 sont représentés en vert foncé et vert clair respectivement, Nop10 est en rose, la L7Ae est en gris, Gar1 en bleu, l'ARN guide H/ACA est représenté en jaune avec le motif K-turn et la boite ACA en rouge et les séquences guides sont en orange.

positionnée au centre de la poche sert de substrat à la pseudouridine synthase, aCbf5 (Baker et al., 2005). La distance entre l'uridine ciblée et les boîtes H ou ACA de l'ARN guide est d'environ 14 nt. De la même manière que les ARN guides à boîtes C/D, les ARN guides à boîtes H/ACA sont constitués d'un ou plusieurs motifs Kink-turn liant la protéine L7Ae (**Figure 10A**).

Chez les archées, les premiers ARN guides de modification à boîtes H/ACA ont été détectés chez P. furiosus, A. fulgidus, S. solfataricus (Klein et al., 2002; Tang et al., 2002a; Tang et al., 2005). La résolution de la structure des RNP guides H/ACA a été facilitée par la reconstitution de ces complexes actifs in vitro comprenant un ARN H/ACA d'archées contenant une simple structure tige boucle-boucle interne-tige boucle transcrite in vitro avec les protéines purifiées recombinantes Cbf5, L7Ae, Nop10 et Gar1 (Figure 10B) (Baker et al., 2005; Charpentier et al., 2005). Ces expériences ont permis de mettre en évidence que les protéines Cbf5 et L7Ae sont capables de se lier directement à l'ARN guide H/ACA. La protéine L7Ae se lie au motif Kink-turn formé dans la partie apicale de la tige boucle indépendamment des autres protéines. La protéine Cbf5 interagit directement avec les ARN guides via la boîte ACA et requiert les protéines Gar1 et Nop10 (Figure 10C et D). Ces différentes interactions ont été confirmées par la structure cristalline du complexe ARN guide-protéines (Duan et al., 2009; Li and Ye, 2006; Liang et al., 2009). L'ARN guide H/ACA forme une tige proximale (P1) suivi de la boîte ACA, de la poche de pseudouridylation et d'une tige distale (P2) contenant le motif Kink-turn. La boîte ACA et la tige P1 interagissent avec le domaine de pseudouridine synthase et archaeosine transglycosylase (PUA) de Cbf5, tandis que la tige P2 se retrouve à la surface composite formée par Nop10, L7Ae et le sous-domaine D1 de Cbf5 (Li and Ye, 2006; Liang et al., 2007). L'interaction entre le complexe Cbf5-Nop10-L7Ae et l'ARN guide positionne la poche de pseudouridylation près du site actif de Cbf5. Cependant, les nucléotides formant la boucle de pseudouridylation ont peu d'interaction avec les protéines du complexes contrairement aux tiges basale et apicale (Figure 10C et D) (Duan et al., 2009; Li and Ye, 2006; Liang et al., 2009). La structure cristalline de la particule RNP H/ACA en présence de substrat ARN lié au site actif montre que le substrat est recruté par un appariement séquence spécifique avec l'ARN guide et des contacts protéiques esssentiels. La liaison du substrat entraîne une réorganisation spatiale de la poche de pseudouridilation et de la tige basale de l'ARN guide ainsi qu'un mouvement du domain PUA (Duan et al., 2009).



**Figure 11 : Représentation des structures** (A) secondaire en trèfle et (B) tertiaire en forme de « L » d'un ARNt (ARNt de levure) (adapté de (de Koning et al., 2010)). Les différents bras sont représentés : en rouge, le bras accepteur contenant l'extrémité 3' CCA en orange où l'acide aminé est chargé, en bleu, le bras anticodon contenant la boucle anticodon en gris impliquée dans le décodage des codons de l'ARNm, en vert, jaune et bleu clair, les bras D, T et la boucle variable respectivement. La position Wobble N34 ainsi que les positions N15 et N48 de l'appariement Levitt sont indiquées.

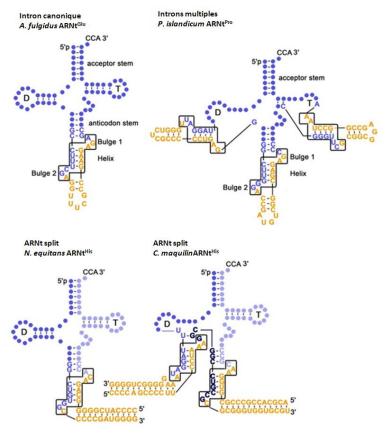


Figure 12 : Diversité des ARNt maturés par l'endonucléase d'épissage chez les archées (Tiré de (Heinemann et al., 2010)). L'endonucléase reconnait et clive le motif BHB encadré. Les séquences représentées en bleu correspondent aux futurs ARNt matures et en jaune sont représentées les séquences qui sont éliminées lors de la maturation.

#### (b) Les ARNt

#### Les introns

Chez les archées les transcrits primaires d'ARNt (pré-ARNt) font l'objet de nombreux processus de maturation pour former des ARNt matures et fonctionnels (**Figure 11**). Ils comprennent des étapes de maturation des extrémités 5' et 3', l'épissage d'intron(s), la modification de résidus nucléotidiques et l'addition de -CCA- à leur extrémité 3' pour la plupart des ARNt (pour revue, (Heinemann et al., 2010)).

Environ 15% des pré-ARNt d'archées contiennent des introns (Sugahara et al., 2007; Tocchini-Valentini et al., 2009). Ces introns sont de taille variable comprise entre 16 et 44 nt et peuvent être retrouvés à diverses positions dans les pré-ARNt. Dans certains cas, on peut retrouver jusqu'à trois introns au sein d'un même précurseur (Figure 12) (Sugahara et al., 2007). Les jonctions intron-exon forment un motif structural en « BHB » qui est le domaine de reconnaissance spécifique de l'endonucléase d'épissage spécifique identique à celui observé dans les introns d'ARNr (voir ci-avant) ((Fabbri et al., 1998) et pour revue (Heinemann et al., 2010)). De manière surprenante, des ARNt précurseurs fragmentés ont été découverts chez N. equitans non détectable par des algorithmes classiques de recherche de gènes d'ARNt. En plus des 34 gènes d'ARNt classiques sans intron et quatre gènes d'ARNt contenant des introns retrouvés dans le génome petit et compact de N. equitans, six gènes fragmentés ou « split tRNA gènes » ont été découverts (Randau et al., 2005a; Randau et al., 2005b) (Figure 12). Chaque «demi-» gène possède son propre promoteur et n'est pas nécessairement situé à proximité de sa moitié complémentaire, parfois même sur des brins différents. La moitié 3' d'un ARNt est précédée d'une séquence leader complémentaire à la séquence située en aval de sa paire 5' (Figure 12). Ces séquences inversées complémentaires forment alors une hélice de 12 à 14 nt formant une jonction de type motif « BHB » relâché (Randau et al., 2005b). Dans tous les cas de figures, la ligase responsable de la ligature des deux jonctions exon-intron n'est pas identifiée à ce jour.

### La RNase P

La maturation de l'extrémité 3' du pré-ARNt d'archées est réalisée par la ribonucléase RNase Z et la maturation de l'extrémité 5' par la ribonucléase P (RNase P), retrouvée dans les trois domaines du vivant. Le complexe RNase P est un RNP qui possède une activité essentielle et Mg2+-dépendante (Altman et al., 1993; Frank and Pace, 1998). Bien que l'activité catalytique soit portée par l'ARN de la RNase P (« RPR » pour « RNase P RNA »),

la qualifiant de « ribozyme », les cofacteurs protéiques de la RNase P (RPP pour RNase P proteins) jouent aussi des rôles primordiaux. La RNase P d'archées est constituée d'un RPR et d'au moins quatre RPP homologues aux RPP eucaryotes (Fukuhara et al., 2006). Chez les archées, les premières caractérisations de la RNase P ont été faites chez *H. volcanii* et *S. acidocaldarius* (Darr et al., 1990; LaGrandeur et al., 1994; Lawrence et al., 1987). Généralement, les RPR des archées ne sont pas actives *in vitro* en l'absence de leurs protéines associées. Bien que de nombreux travaux de résolution de structure de chacun des partenaires protéiques de *P. horikoshii* sont disponibles (pour revue, (Walker and Engelke, 2006)), aucune structure cristalline d'une particule RNase P complète n'est connue. Cependant, les interactions entre facteurs protéiques ont été mise en évidence chez *P. horikoshii* par la technique de double-hybride (Kifusa et al., 2005). Récemment, un travail a caractérisé la RNase P native chez l'archée *Methanococcus maripaludis*, constituée d'une RPR et des quatre RPP dont la protéine ribosomique L7Ae. L'ajout de L7Ae dans la RNase P reconstituée *in vitro* accélère la réaction de clivage des pré-ARNt (Cho et al., 2010).

Une fois de plus, l'archée *Nanoarchaeum equitans* s'est révélée particulière du fait de l'absence de gène codant la RNase P (Randau et al., 2005a). Cependant, cette absence semble compensée par la présence de promoteurs dits « stricts », révélés par analyse génomique, permettant d'initier la transcription des ARNt en leur extrémité 5' mature (Randau et al., 2008).

### Modifications de nucléotides dans les ARNt

Dans les trois domaines du vivant, les ARNt contiennent jusqu'à un quart de leurs résidus modifiés (voir références citées dans (Vinayak and Pathak, 2010)). Plus de 120 modifications d'ARN (comprenant celles concernant les ARNt et les ARNr) sont répertoriées dans la base de données MODOMICS (Czerwoniec et al., 2009). Ces modifications permettent à la molécule d'ARN d'adopter une conformation canonique en L et modulent les interactions avec les acteurs de la traduction tels que les aminoacyl-tRNA synthétases et différents facteurs protéiques (voir références citées dans (Grosjean et al., 2008)). Ces modifications et leurs enzymes respectifs ne seront pas détaillés ici, seules quelques particularités spécifiques liées aux archées seront développées.

Les deux guanosines hypermodifiés ubiquitaires des ARNt sont la queuosine (Q) (position wobble N34 de l'anticodon chez les eucaryotes et les bactéries) et l'archaeosine (G<sup>+</sup>) (position N15 faisant partie de l'appariement Levitt entre la boucle D et la boucle variable chez les archées) (**Figure 11**) (pour revues, (de Bernis et al., 2003; de Koning et al., 2010;

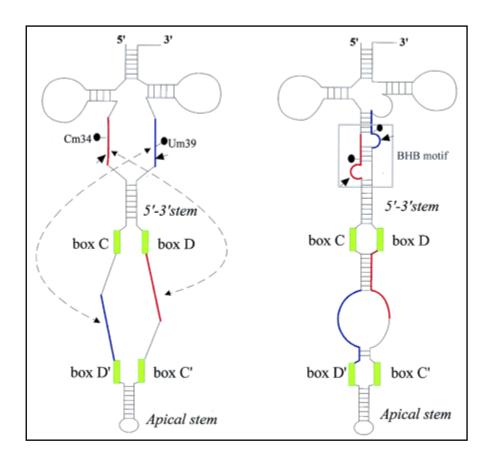


Figure 13 : Caractéristique de l'ARNt<sup>Trp</sup> chez les archées (Tiré de (Clouet d'Orval et al., 2001). La première modification par la première 2'-O-méthylation effectuée à la position C34 ou U39 engendre la structuration de l'ARN afin de former le motif BHB permettant le clivage par l'endonucléase d'épissage. Les boîtes C/D et C'/D'sont représentées en vert, Les flèches courtes indiquent les sites de clivage de l'endoribonucléase et les longues flèches en pointillés indiquent les séquences complémentaires de l'ARN guide C/D aux séquences cibles de l'ARNt (représentées en rouge et en bleu).

Vinayak and Pathak, 2010)). Les interactions entre les bras D et T permettraient la stabilisation de la structure tertiaire en forme de « L » des ARNt, notamment dans les organismes thermophiles afin de prévenir la dénaturation thermique (Phillips et al., 2010). La queuosine et l'archaeosine sont des dérivés du même précurseur, le GTP (voir références citées dans (Phillips et al., 2008)). Chez les bactéries et les eucaryotes, la conversion du GTP en Q nécessite plusieurs réactions chimiques complexes qui font intervenir des enzymes spécifiques dont une ARNt-guanine transglycosylase (TGT) ou encore les protéines QueCDEF caractérisés chez *B. subtilis* (voir références citées dans (Roth et al., 2007)). Les réactions enzymatiques permettant de convertir le GTP en G<sup>+</sup> chez les archées ne sont pas encore toutes caractérisées. Seul un isoforme arcTGT de la TGT est aussi retrouvé (Phillips et al., 2010).

Contrairement aux ARN guides C/D eucaryotes, les ARN guides C/D d'archées peuvent cibler des méthylations à des positions spécifiques des ARNt (Clouet-d'Orval et al., 2005; Clouet d'Orval et al., 2001; Dennis et al., 2001; Grosjean et al., 2008; Omer et al., 2000; Omer et al., 2003). Par prédiction bio-informatique, un ARN guide C/D donné pourrait cibler plusieurs ARNt différents (Dennis et al., 2001). Une particularité est à relever dans le cas du pré-ARNt Trp. Son propre intron code pour l'ARN guide à boîtes C/D responsable de la 2'O-méthylation de ses positions C34 et U39 suggérant un système de modification en *cis* (**Figure 13**) (Clouet-d'Orval et al., 2005; Singh et al., 2008). Egalement, il est suggéré que la position \(\psi 35\) de certains ARNt Tyr soit modifiée par un ARN guide à boîtes H/ACA (Muller et al., 2008).

### (c) Le motif Kink-turn et la protéine L7Ae

Les différents ARN stables essentiels pour une traduction fidèle et efficace, tels que les ARNr, ARNt, ARN de la RNase P, et ARN guides C/D et H/ACA, possèdent des structures tridimensionnelles définies par des motifs ARN particuliers permettant un empilement ordonné et assemblé d'appariements non-Watson-Crick (Leontis et al., 2006; Moore, 1999). Ils confèrent aux ARN une structure tertiaire qui serait déterminante pour leur fonction (liaison à des métaux ou des protéines). De nombreux motifs ARN sont répertoriés à ce jour comprenant entre autres : le motif Kink-turn (K-turn), le C-motif, le motif sarcine-ricine ou encore le motif E-loop (pour revue (Leontis and Westhof, 2003)).

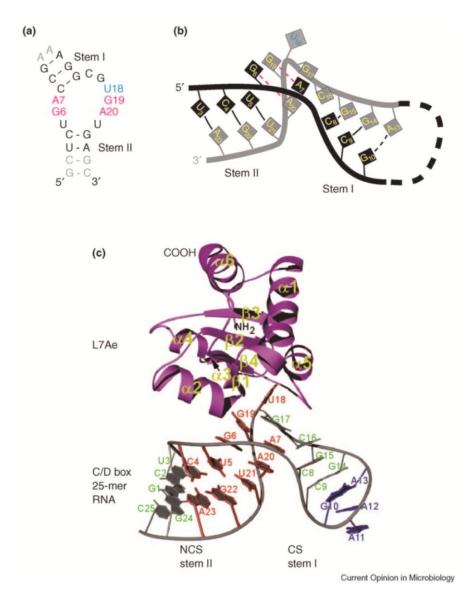


Figure 14: Caractéristiques du motif K-turn (tiré de (Dennis and Omer, 2005)). (a) Structure secondaire d'un motif k-turn d'un oligonucléotide de 25 nt contenant les boîtes caractéristiques d'un ARN guide C/D (C = CUGA, D = GUGAUGA). Cet oligonucléotide a été utilisé dans les co-cristaux avec la protéine L7Ae (Moore et al., 2004). (b) Représentation schématique du motif K-turn dans la structure cristallisée. Les nucléotides G6/A20 et A7/G19 forment les appariements en ciseaux caractéristiques du K-turn engendrant un « coude » dans l'ARN où l'on observe le nucléotide protubérant U18. (c) Structure cristallisée de la protéine L7Ae représentée en violet de A. fulgidus liée au motif K-turn (Moore et al., 2004).

Le motif K-turn et son dérivé Kink-loop (K-loop) sont retrouvés dans de nombreuses architectures structurales d'ARN dans les trois domaines du vivant. Ils sont caractérisés comme site de reconnaissance à des protéines et ils participent à des interactions intramoléculaires multiples ((Klein et al., 2001; Lescoute et al., 2005; Nolivos et al., 2005; Vidovic et al., 2000) et pour revue, (Tiedge, 2006)).

Le K-turn est un motif structural tridimensionnel caractérisé par une boucle asymétrique interne (dans les K-turn) ou apicale (dans les K-loop) encadrée par une hélice canonique (avec au minimum trois appariements de type Watson Crick) et une hélice non canonique composée d'au moins trois appariements dont deux interactions AG caractéristiques dites en « ciseau ». Ce dernier agencement de nucléotides engendre une courbure de l'ARN en « coude » (kink) avec en protubérance les nucléotides de la boucle (Lescoute et al., 2005). Le motif K-turn existe dans un équilibre dynamique sous la forme d'une structure ouverte, typiquement une simple hélice contenant un bulge, ou une conformation hautement ordonnée en coude dans laquelle l'ARN se courbe formant un angle compris entre 45° et 63° (**Figure 14**) (Moore et al., 2004). Ce motif structural est le site de liaison de la protéine ribosomique multifonctionnelle L7Ae, composante des RNP guides C/D et H/ACA mentionnés précédemment, chez les archées. Cette protéine semble aussi se lier à d'autres ARN cellulaires de *S. solfataricus* donnant une fonction importante au motif ARN en K-turn (Dennis and Omer, 2005; Zago et al., 2005).

### d) Dégradation des ARNm

Les mécanismes de maturation des ARN dits stables et de dégradation des ARN messagers sont des étapes fondamentales de la régulation de l'expression des gènes. Ces mécanismes sont encore mal connus chez les archées (pour revue (Evguenieva-Hackenberg and Klug, 2009)).

Un exosome de type eucaryote est retrouvé chez les archées à l'exception des archées du genre *Methanococcus* et les halophiles telles que *Halobacterium sp.* ((Koonin et al., 2001; Ng et al., 2010) et voir références citées dans (Evguenieva-Hackenberg and Klug, 2009)). Chez les eucaryotes, l'exosome est un complexe multi-protéique composé de protéines impliquées dans la maturation d'ARNnc (ARNr, ARNt, ARN nucléaires et nucléolaires) et la dégradation dans le sens 3'→ 5' d'ARNm (voir références citées dans (Evguenieva-Hackenberg and Klug, 2009; Ng et al., 2010)). Le complexe cœur de l'exosome d'archées est

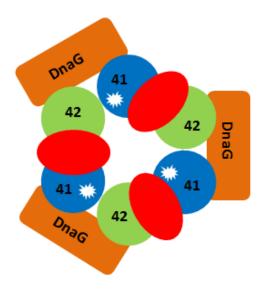
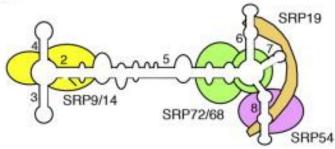


Figure 15 : Vue générale schématique de la composition de l'exosome d'archées. L'anneau hexamérique composé par une alternance de protéines Rrp41 et Rrp42 est représenté en vert et bleu. Les trois protéines additionnelles Csl4 et/ou Rrp4 sont représentées par des pastilles ovales rouges. Les protéines DnaG interagissant avec l'exosome sont représentées en orange. Les sites catalytiques portés par les protéines Rrp41 sont représentés en blanc.

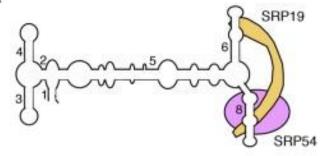
composé de trois exemplaires de chacune des protéines Rrp41 et Rrp42 qui s'alternent afin de s'organiser en anneau hexamérique (Figure 15) (voir référence citée dans (Roppelt et al., 2010b)). Une coiffe hétéro-trimérique, composée de trois protéines Rrp4 et/ou Csl4, est associée au dessus de l'anneau (Figure 15) (voir référence citée dans (Roppelt et al., 2010b)). Cet arrangement spatial est similaire à celui de l'exosome eucaryote et de la polynucléotide phosphorylase bactérienne (PNPase) (pour revue (Evguenieva-Hackenberg and Klug, 2009)). La caractérisation fonctionnelle in vitro de l'exosome de S. solfataricus a montré son implication dans la dégradation d'ARNm dans le sens 3'→ 5' (Portnoy et al., 2005). Rrp41 et Rrp42 contiennent un domaine RNase PH-like suggérant leur implication dans la dégradation d'ARN simple brin dans le sens 3'→ 5' (Koonin et al., 2001). En effet, celles-ci sont suffisantes pour l'activité de dégradation et les sites catalytiques sont exclusivement situés sur les sous-unités Rrp41 ((Lorentzen et al., 2005) et pour revue (Evguenieva-Hackenberg and Klug, 2009)). Par ailleurs, il a été montré chez S. solfataricus que l'anneau héxamérique Rrp41-Rrp42 est doté d'une activité d'addition de nucléotides di-phosphates permettant d'ajouter une queue hétéropolymérique riche en nucléotides A (Portnoy et al., 2005; Portnoy and Schuster, 2006) aux ARN simples brins à dégrader suggérant que l'exosome d'archées aurait un mécanisme de dégradation similaire à celui porté par la PNPase bactérienne (Portnoy et al., 2005). Les protéines Rrp4 et Csl4 possèdent des domaines de liaison à l'ARN: les domaines S1 et KH ou S1 et Ribbon-zinc respectivement (Koonin et al., 2001). L'activité de dégradation est modulée en présence des protéines Rrp4 et/ou Csl4 (Walter et al., 2006). L'exosome d'archées possède trois sous-unités additionnelles de protéines annotées DnaG (Figure 15) (Roppelt et al., 2010a). Récemment, les co-sédimentations de DnaG et de Rrp41 dans les extraits membranaires de cellules de S. solfataricus suggèreraient que l'exosome serait localisé à la membrane (Roppelt et al., 2010a). Cette hypothèse est d'autant plus appuyée par des expériences d'immuno-marquage de ces mêmes protéines à la périphérie cellulaire permettant ainsi de supposer que la dégradation des ARNm se ferait à la membrane cellulaire. Ces observations peuvent être mises en parallèle avec la présence de la Ribonucléase E du dégradosome de la bactérie E. coli à la membrane ((Khemici et al., 2008)).

Les archées du genre *Methanococcus* (*M. maripaludis* et *M. jannashii*) et les halophiles ne possèdent pas de gènes codant pour des protéines d'un complexe exosome-like. Les halophiles seraient cependant pourvues de la RNase R (voir références citées dans (de Koning et al., 2010)). De plus, il semblerait que la polyadénylation ne serait observée que chez des archées possédant un exosome (Portnoy et al., 2005; Portnoy and Schuster, 2006). La RNase R apparait comme l'exoribonucléase majeure et essentielle à la viabilité cellulaire

# Eukarya



### Archaea



Bacteria (E. coli)

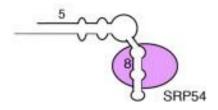


Figure 16 : Schématisation des particules SRP dans les trois domaines du vivant. (tiré de (Calo and Eichler, 2010)). L'ARN 7S est représenté en noir, les protéines SRP sont représentées par des formes colorées : SRP54 en violet, SRP19 en marron, SRP9/14 en jaune et SRP72/68 en vert.

de l'halophile *H. volcanii*. Cependant, celle-ci n'est pas capable de dégrader les substrats ARN contenant des structures secondaires (Portnoy and Schuster, 2006).

Très récemment, des RNases possédant une activité exoribonucléase impliquées dans les dégradations d'ARN dans le sens  $5' \rightarrow 3'$  ont été caractérisées chez *P. abyssi*, *T. kodakaraensis* et *S. solfataricus*. Celles-ci font toutes parties de la famille des metallo-β-lactamases de type β-CASP (Clouet-d'Orval et al., 2010; Hasenohrl et al., 2011).

### e) Ciblage des protéines à la membrane : l'ARN SRP

La particule de reconnaissance du signal (SRP pour Signal Recognition Particle) est un complexe RNP qui reconnait le peptide signal des chaines peptidiques naissantes sortant des ribosomes afin de les cibler aux systèmes de translocation aux membranes. Dans les systèmes eucaryotes où ce système a été intensivement étudié, la SRP (aussi appelée MRP chez les eucaryotes) retarde la traduction de la protéine jusqu'à ce que le ribosome lié à la SRP soit associé au récepteur de la SRP (SR pour SRP receptor) situé dans la membrane (Koch et al., 2003). La protéine est transloquée de la membrane de manière co-traductionnelle pour finir dans un autre compartiment cellulaire ou dans l'espace inter-membranaire (Shan and Walter, 2005). Chez les eucaryotes, les protéines sont ciblées à la membrane du réticulum endoplasmique. Chez les archées, la SRP cible les protéines à la membrane cytoplasmique (Zwieb and Eichler, 2002). Chez les bactéries, il existe de nombreux systèmes de ciblage aux membranes mais la SRP est le principal système utilisé pour l'incorporation de protéines dans la membrane interne (voir références citées dans (Calo and Eichler, 2010)).

La SRP d'archées est plus ressemblante à son homologue eucaryote (**Figure 16**). Elle est composée d'une molécule d'ARN 7S dont la structure secondaire est similaire à celle des eucaryotes (Kaine, 1990). Deux des six protéines composant la SRP eucaryote, SRP19 et SRP54, font également partie de la SRP d'archées. Certains aspects de la SRP d'archées sont propres à leur domaine d'après les caractéristiques spécifiques observées lors de la reconstitution *in vitro* de cette particule ((Bhuiyan et al., 2000; Diener and Wilson, 2000; Tozik et al., 2002) et pour revue (Calo and Eichler, 2010)). Les complexes SRP du Crenarchaeote *Acidianus ambivalens* et des Euryarchaeotes *A. fulgidus*, *P. furiosus* et *H. volcanii* ont été reconstitués *in vitro* (Bhuiyan et al., 2000; Diener and Wilson, 2000; Moll et al., 1999; Tozik et al., 2002). L'étude de Bhuiyan et ses collaborateurs a montré que l'ARN SRP de *M. jannaschii* était capable de former un complexe RNP recombinant avec la SRP54

Organisme	P. abyssi GE5	P. furiosus DSM3638	P. horikoshii OT3	T. gammatolerans EJ3	T. kodakaraensis KOD1	T. onnurineus NA1	T. sibiricus MM 739
Taille du génome (bp)	1765118 (C) 3444 (P)	1908256	1738505	2045438	2088737	1847607	1845800
Nombre de CDS	1896	2125	1955	2157	2306	1976	2035
% de régions codantes	93.1% (C)	93.8%	95.0%	94.0%	93.2%	91.7%	90.0%
Contenu en GC%	44.71% (C) 43.44% (P)	40.77%	41.88%	53.56%	52.0%	51.72%	40.20%
Nombre d'ARNt	46	46	46	46	46	46	46
Nombre	2 x 5S	2 x 5S	2 x 5S	2 x 5S	2 x 5S	2 x 5S	2 x 5S
d'ARNr	16S	16S	16S	16S	16S	16S	16S
u ARNI	23S	23S	23S	23S	23S	23S	23S
ARN 7S (SRP)	1	1	1	1	1	1	1
Eléments IS	7 partiels	40 dont 10 partiels	3 partiels	0	8 dont 7 partiels	ND	7 dont 5 partiels
Virus intégrés	0	0	2	2	4	ND	1
Éléments CRISPR	4	7	6	3	3	5	1
T. opt de croissance (°C)	96	100	98	85-90	95	80	64-84
Habitat	Fonds marins, cheminées hydrothermales	Fonds marins volcaniques	Fonds marins, cheminées hydrothermales	Fonds marins, cheminées hydrothermales	Solfatares	Fonds marins, sources chaudes	Réserves d'hydrocar bures sibériques
Localisation	Océan Pacifique	Italie	Océan Pacifique (Japon)	Mexique	Japon	PACMANUS	Due-i-
géographique	géographique (Iles Fidji)						Russie
Références	(Cohen et al., 2003; Erauso et al., 1996)	(Robb et al., 2001)	(Kawarabayasi et al., 1998)	(Zivanovic et al., 2009)	(Fukui et al., 2005)	(Lee et al., 2008)	(Mardanov et al., 2009)

**Tableau 5 : Caractéristiques générales des Thermococcales dont le génome est séquencé et annoté.** (C) et (P) désignent respectivement le chromosome et le plasmide. CDS : Coding Sequence. IS : Insertion sequence. CRISPR : Clustered Regularly Interspaced Palindromic Repeats.

humaine indépendamment de la présence de SRP19 (Bhuiyan et al., 2000), suggérant ainsi la conservation structurale et fonctionnelle du complexe SRP entre archées et eucaryotes.

### D. Les Thermococcales et *Pyrococcus abyssi*

### a) Caractéristiques générales

Les Thermococcales constituent un ordre des Euryarchaeotes, phylum majeur des archées (Figure 2). 320 espèces (dont 17 ne sont pas encore classées) sont réparties dans trois genres : les Pyrococcus (68 espèces), les Thermococcus (232 espèces) et les Paleococcus (3 espèces) (source : Taxonomy browser, NBCI). Les archées de cet ordre ont été isolées à partir d'écosystèmes sous-marins à proximité de cheminées hydrothermales, de solfatares (régions volcaniques où l'on retrouve des dépôts de soufre), mais aussi dans des réservoirs d'hydrocarbures (**Tableau 5**). Les Thermococcales sont qualifiées d'hyperthermophiles lorsque leur température de croissance est supérieure à 80°C. Elles sont capables de croître dans des milieux en anaérobiose avec un métabolisme organo-hétérotrophe (capacité à utiliser des composés organiques complexes tels que des mixtures d'acides aminés, des maltodextrines) couplé à l'utilisation de soufre élémentaire (S<sup>0</sup>) comme accepteur d'électrons (voir références citées dans (Mardanov et al., 2009; Zivanovic et al., 2009)). En plus de leur caractère hyperthermophile, certaines Thermococcales se distinguent par d'autres phénotypes extrêmes : P. abyssi, Pyrococcus CH1 et T. barophilus sont qualifiées de barophiles par leur capacité à croître sous de hautes pressions hydrostatiques (Cohen et al., 2003; Zeng et al., 2009). T. gammatolerans EJ3 apparait être l'organisme le plus résistant à des irradiations allant jusqu'à 30 kilos Gray (sachant qu'une dose de 10 Gray est suffisante pour tuer un homme) (Zivanovic et al., 2009).

### b) Les génomes séquencés

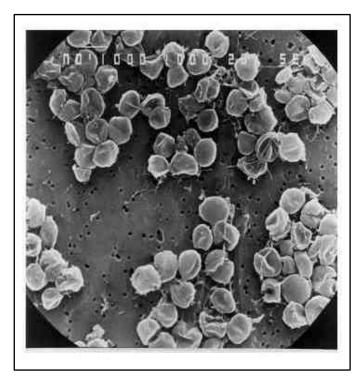
On compte sept génomes de Thermococcales entièrement séquencés, annotés et disponibles publiquement comprenant trois *Pyrococcus*: *P. abyssi* GE5, *P. furiosus* DSM3638, *P. horikoshii* OT3, et quatre *Thermococcus*: *Thermococcus kodakaraensis* KOD1, *T. onnurineus* NA1, *T. gammatolerans* EJ3 et *T. sibiricus* MM 739 (**Tableau 5** et références correspondantes). Tous possèdent un chromosome circulaire mais seule *P. abyssi* présente un

réplicon additionnel au chromosome : le plasmide pGT5 (Erauso et al., 1996). La taille des chromosomes varie de 1,77 à 2,09 Mb contenant de 90 à 95% séquences codantes (CDS) (**Tableau 5**). Le contenu en GC est supérieur à 50% pour les *Thermococcus* à l'exception de *T. sibiricus* pour lequel le pourcentage en GC est plus proche de celui des *Pyrococcus* (environ 42%). Chacun des sept génomes possède une copie de l'opéron ribosomique 16S-23S, deux gènes d'ARNr 5S séparés et 46 gènes d'ARNt.

Les deux analyses comparatives de génomes les plus récentes montrent que les Thermococcales partagent environ 1100 gènes conservés suggérant que ce cœur de gènes serait présent dans le génome de l'ancêtre commun aux *Pyrococcus* et aux *Thermococcus* (Mardanov et al., 2009; Zivanovic et al., 2009). Ces gènes coderaient en majeure partie des protéines hypothétiques conservées mais aussi des protéines impliquées dans le métabolisme des acides aminés, le métabolisme énergétique, le métabolisme des acides nucléiques ou encore dans les systèmes de transports membranaires (Zivanovic et al., 2009). Par ailleurs, chaque genre possède aussi leur propre ensemble de gènes communs qui illustrent la séparation évolutive des deux genres de Thermococcales les plus étudiés (Gunbin et al., 2009; Mardanov et al., 2009; Zivanovic et al., 2009).

### c) Les éléments génétiques

Les éléments génétiques ont été prédits dans six des sept génomes séquencés (**Tableau 5**). L'analyse de chacun des génomes met en évidence la présence de séquences d'insertion (IS) dans tous les génomes de Thermococcales (de trois à 40 IS) à l'exception de *T. gammatolerans* qui n'en possède aucun (Zivanovic et al., 2009). L'absence d'IS dans un de ces génomes reflèteraient que ces éléments mobiles ne jouent pas de rôle dans l'histoire évolutive des génomes de Thermococcales contrairement à ce qu'il a été suggéré pour les halophiles (Filee et al., 2007; Zivanovic et al., 2009). Il est à noter que la famille IS200/605 est représentée dans les cinq autres génomes (source : IS Finder <a href="http://www-is.biotoul.fr/">http://www-is.biotoul.fr/</a>) suggérant qu'il existerait une origine phylogénétique commune (Mahillon and Chandler, 1998). Une partie des IS annotées sont de type « partielles » correspondant à des éléments orfB (dépourvus de la copie de orfA qui sont les deux gènes d'une transposase) (**Tableau 5**). Aucune donnée n'indique si ces éléments isolés agissent (i) dans des phénomènes de transposition ou (ii) par activation en trans par une autre transposase de la même famille (Filee et al., 2007).



**Figure 17 : Cellules de** *Pyrococcus abyssi*. Photographie en microscopie électronique. Tiré du site web de l'Ifremer

P. horikoshii, T. gammatolerans, T. kodakaraensis et T. sibiricus possèdent des génomes de virus intégrés dans leurs chromosome (**Tableau 5**). Leurs intégrations sitespécifiques au niveau de gènes d'ARNt se feraient sur un mécanisme médié par une intégrase comparable aux éléments génétiques chez les Sulfolobus (Fukui et al., 2005; Mardanov et al., 2009; Zivanovic et al., 2009). Cette hypothèse est notamment renforcée par la prédiction du gène d'une intégrase chez T. sibiricus qui serait homologue à l'intégrase du virus SSV1 de Sulfolobus shibatae (Mardanov et al., 2009).

Une à sept copies d'éléments CRISPR sont retrouvées dans les génomes des Thermococcales séquencés (**Tableau 5**). La transcription de ces loci (ou cassettes) CRISPR a été vérifiée expérimentalement chez *P. furiosus* et *P. abyssi* suggérant qu'un mécanisme de défense à ARN serait actif chez ces deux Thermococcales ((Hale et al., 2008) et ce travail de thèse). Comme évoqué précédemment, les CRISPR sont impliqués dans des mécanismes défensifs afin de protéger les cellules contre l'invasion d'acides nucléiques d'origine étrangère tels les génomes de phages ou des plasmides (Bolotin et al., 2005; Mojica et al., 2005; Pourcel et al., 2005). A ce jour, il n'est connu qu'un seul virus de Thermococcales : le virus PAV1 de *P. abyssi* GE23 dont le génome a été séquencé très récemment (Geslin et al., 2007).

### d) Pyrococcus abyssi, notre modèle d'étude

Pyrococcus abyssi a été collectée au sein d'une source hydrothermale à 2000 m de profondeur au Nord du bassin Fidjien dans l'océan Pacifique. Les cellules de *P. abyssi* sont des coques sphériques de 0,8 à 2 μm de diamètre (**Figure 17**). Cette archée est un microorganisme anaérobie strict, chimio-hétérotrophe et capable de réduire le soufre. Elle est qualifiée d'hyperthermophile avec un optimum de croissance à 96°C (Prieur et al., 1995). En condition de pression atmosphérique, elle est capable de croitre entre 67 et 102°C avec un temps de génération de 33 min à 96°C. À pression in situ de 20MPa, la température de croissance peut-être augmentée jusqu'à 3°C de plus. Son chromosome est répliqué environ en 45min (Myllykallio et al., 2000). Le génome de la souche GE5 est entièrement séquencé et annoté ((Cohen et al., 2003) ; GenBank AL096836). Il est composé d'un chromosome circulaire d'environ 1,77 Mb et d'un plasmide de 3,5 kb (Charbonnier et al., 1992). L'annotation du génome a mis en évidence 1764 cadres ouverts de lecture dont environ 50% n'ont pas encore de fonction identifiée (Cohen et al., 2003).

P. abyssi s'est révélée comme un modèle d'étude pour l'identification de l'origine de réplication (Myllykallio and Forterre, 2000) et pour la résolution des dimères de chromosomes (Cortez et al., 2010). Cette archée possède le plasmide pGT5 qui a servi comme modèle pour l'étude de la topologie de l'ADN ainsi que pour la réplication en cercle roulant (Charbonnier et al., 1992; Erauso et al., 1996). De plus, les protéines de P. abyssi servent souvent de modèle pour définir des structures cristallines de protéines comme cela a été le cas pour la structure de l'exosome en présence de substrat ARN (Navarro et al., 2008). Le génome de P. abyssi a aussi été exploité pour la recherche d'ARN guides de modification à boîtes C/D par notre équipe et l'équipe d'Ariana Omer (Gaspin et al., 2000; Omer et al., 2000) et d'ARN guides à boîtes H/ACA par l'équipe de C. Branlant (Muller et al., 2008). Le pourcentage en G+C de son génome est de 44,7%, le qualifiant ainsi de riche en base A et T. Ce biais de composition en nucléotide a été exploité pour la recherche d'ARNnc dans le génome de P. furiosus (Klein et al., 2002).

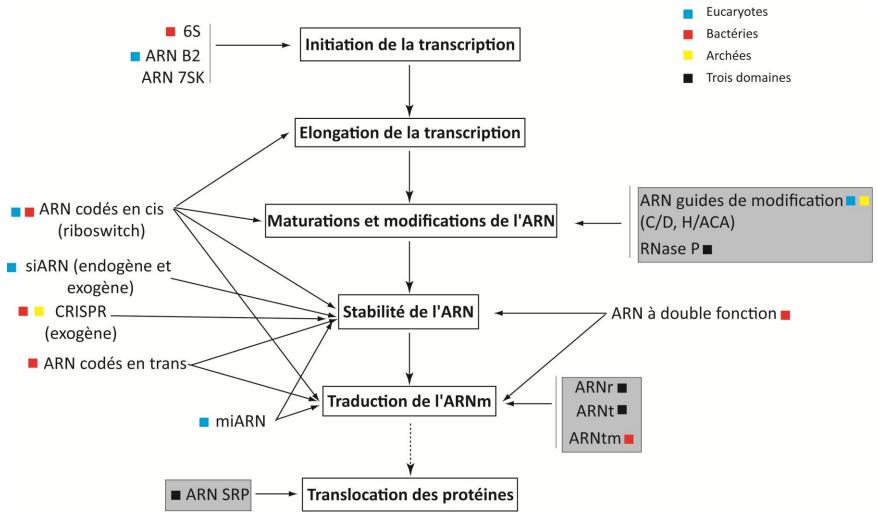


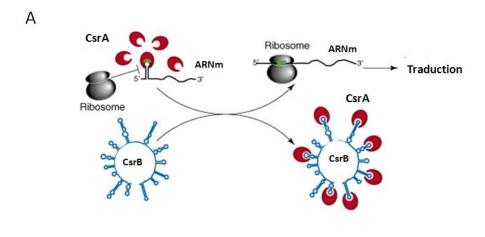
Figure 18 : Variété des ARNnc dans l'expression génique (adapté de (Storz et al., 2005)). Les principales étapes de l'expression génique sont représentées. Les ARNnc participant à ces processus sont indiqués. Les ARNnc grisés ont été présentés dans les chapitres précédents à l'exception de l'ARNtm des bactéries. La distribution de ces ARN est schématisée par des carrés colorés annotés sur la figure.

## II. Les ARN régulateurs

Le contrôle de l'expression génique se fait à différents niveaux depuis la transcription des gènes en ARNm jusqu'à la traduction de ces ARNm en protéines. Il fait intervenir différents types de molécules telles que des protéines et des ARN régulateurs. Le nombre d'ARN régulateurs ne cesse d'augmenter et apparait de plus en plus élevé (pour revues, (Costa, 2007; Waters and Storz, 2009)). Ils sont impliqués dans tous les processus de l'expression génique comprenant la modulation de la transcription, de la traduction, de la stabilité des ARNm ou encore de l'extinction de l'expression de gènes (**Figure 18**). Ces diverses fonctions sont associées à différents mécanismes comme les changements de conformation spatiale de l'ARN, la liaison à des protéines, l'appariement à d'autres ARN ou des interactions avec l'ADN.

A ce jour, plus de 1446 familles d'ARN, répertoriés dans la base de données Rfam (Version 10.0, janvier 2010). Elles sont définies en fonction de leur classe fonctionnelle (les ARNnc, les éléments structurés *cis*-régulateurs et les ARN « self-splicing » capables de catalyser des évènements d'auto-épissage) et de leur conservation en structure secondaire (Gardner et al., 2009). Cependant, avec le nombre grandissant d'ARNnc mis en évidence, une classification universelle apparait difficile. En effet, il est possible de classer les ARNnc en fonction de leur taille (les plus petits ARNnc font une vingtaine de nucléotides et les plus longs peuvent atteindre plusieurs milliers de nt tels que les « large RNA » eucaryotes) (Costa, 2007), de leur mode d'action (modulation de l'activité ou de la quantité de protéines, par appariement à leur cible…) ou encore de leur implication dans un processus biologique (pathogénicité, réponse à des stress ou l'exemple des CRISPR dans l'immunité cellulaire des bactéries et des archées) (Gottesman and Storz, 2010).

Chez les eucaryotes, les ARN non codant sont très diversifiés. On retrouve : les petits ARN nucléaires (snRNA pour small nuclear RNA) impliqués dans les phénomènes d'épissage, les petits ARN nucléolaires (snoRNA pour small nucleolar RNA) participant aux modifications de nucléotides d'autres ARN, les gRNA (guide RNA) intervenant dans les processus d'« editing », des ARN impliqués dans l'inactivation du chromosome X (par exemple, *Xist* et *Tsix*) ou encore des ARN associés aux loci génomiques parentaux et les plus documentés sont représentés par la famille des petits ARNnc constitués par les miARN (micro RNA), les siARN (small interfering RNA) et les piARN (Piwi-interacting RNA) (voir



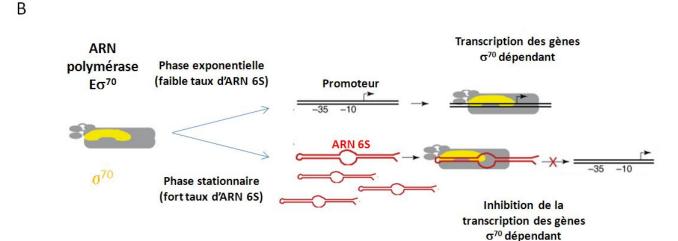


Figure 19 : Deux exemples d'ARNnc modulant l'activité protéique chez les bactéries. (A) Action de l'ARN CsrB sur la protéine CsrA. La protéine CsrA (en rouge) inhibe la traduction de certains gènes via sa fixation à la partie 5'UTR contenant le motif GGA (représenté en vert sur la tige-boucle). En présence d'un fort taux de l'ARN CsrB (en bleu), celui-ci séquestre CsrA via ses 22 sites de fixation GGA, permettant d'antagoniser l'activité de répression de la traduction de CsrA (adapté de (Babitzke and Romeo, 2007)). (B) Mécanisme de régulation porté par l'ARN 6S. En phase exponentielle de croissance, l'ARN polymérase  $E\sigma^{70}$  (ARN polymérase en gris et  $\sigma^{70}$  en jaune) transcrit les gènes contenant les promoteurs  $\sigma^{70}$ -dépendant. Lors de l'entrée en phase stationnaire, l'ARN 6S (en rouge) s'accumule dans la cellule et permet ainsi de titrer l'ARN pol  $\sigma^{70}$  réduisant la transcription des gènes non indispensables à la cellule dans cette phase de croissant (adapté de (Wassarman, 2007)).

références citées dans et pour revues (Choudhuri, 2010; Zhou et al., 2010)). Par choix, cette section a pour objectif de donner une vue globale sur la fonction régulatrice que peuvent porter les ARNnc chez les bactéries et les archées par des exemples précis et documentés.

### A. Les ARN régulateurs chez les bactéries

Les ARNnc bactériens ou « sRNA » (small RNA) sont de taille variable allant de 50 à 500 nt. Leur nombre varie d'un génome bactérien à un autre allant d'une centaine pour E. coli à plusieurs centaines pour d'autres (pour revue récente, (Gottesman and Storz, 2010)). Bien que leur nombre ne cesse de croître, la fonction de seulement quelques dizaines d'entre eux est caractérisée à ce jour. On peut distinguer les ARN régulateurs liant directement des protéines (afin de moduler leur activité et/ou leur quantité dans la cellule), les ARN régulateurs antisens qui s'apparient à leur cible ARN, les riborégulateurs (plus connus sous le terme de « riboswitch ») et les ARN codés par les loci CRISPR. De plus, certains ARNnc présenteraient une double fonction en codant aussi pour des peptides (pour revues (Brantl, 2007; Gottesman, 2005; Gottesman and Storz, 2010; Romby and Charpentier, 2010)). Cette classe d'ARN régulateurs remet en question l'appellation d'ARN non codant. Par exemple, l'ARNIII de la bactérie Staphylococcus aureus régule l'expression de facteurs de virulence et de transcription en s'appariant à leur ARNm (Boisset et al., 2007). Il code également pour une exotoxine de 26 acides aminés, la δ-hémolysine (pour revue, (Romby and Charpentier, 2010)). Il en est de même pour l'ARN SgrS d'E. coli qui inhibe la traduction de l'ARNm ptsG codant pour un transporteur de sucre-phosphate. SgrS code aussi pour une protéine SgrT de 43 acides aminés, connue pour renforcer la régulation exercée par l'ARN SgrS en régulant négativement la consommation de glucose via une inhibition directe ou indirecte de la protéine PtsG (Wadler and Vanderpool, 2007).

# a) Les ARN régulateurs modulant l'activité ou la quantité de protéines

Des ARNnc régulateurs ont été caractérisés pour leur capacité à se lier à des protéines afin de contrecarrer leurs activités. Chez les bactéries, les ARNnc CsrB et CsrC et l'ARN 6S possèdent cette caractéristique (**Figure 19**).

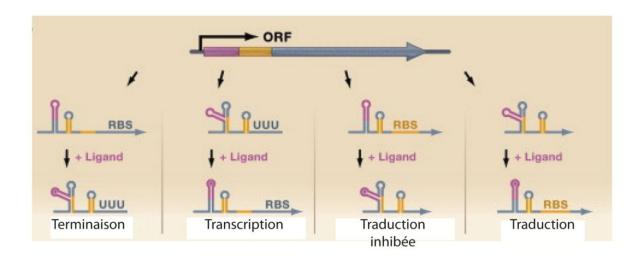


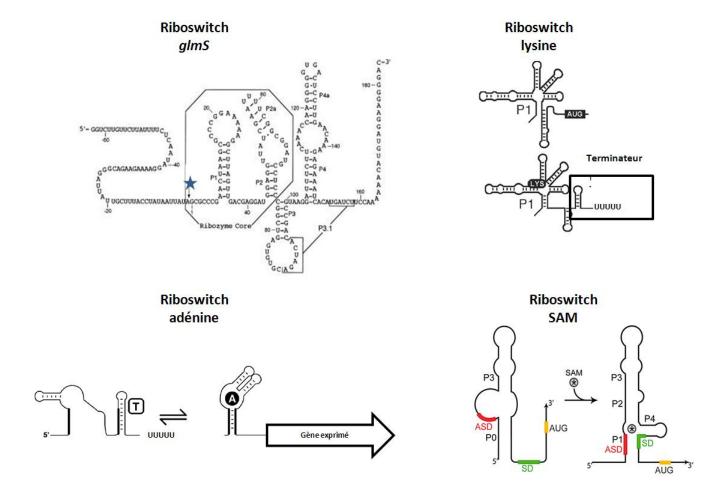
Figure 20 : Caractéristiques des riboswitch chez les bactéries (Tiré de (Waters and Storz, 2009)). Les riboswitch sont composé de deux domaines : un domaine aptamère (ou senseur) en rose et une plateforme d'expression en jaune. Les régulations transcriptionnelles et traductionnelles peuvent être inhibées ou activées. Terminaison : la fixation du ligand induit la formation d'un terminateur de transcription. Transcription : la fixation du ligand déstabilise la structure du terminateur. Inhibition de la traduction : la fixation du ligand le site de fixation des ribosomes (RBS) n'est plus accessible. Traduction : la fixation du ligand libère le RBS et permet la traduction.

La protéine CsrA est une protéine de liaison aux ARN. Elle participe à la régulation de l'utilisation de sources carbonées et de la motilité bactérienne en entrée en phase stationnaire (pour revue, (Babitzke and Romeo, 2007)). Les dimères de CsrA se lient aux motifs GGA situés dans les parties 5'UTR des ARNm cibles et affectent ainsi la stabilité et/ou la traduction des ARNm. Les deux ARNnc CsrB et CsrC contiennent chacun 22 et 13 motifs GGA respectivement. Ils agissent alors en séquestrant la protéine CsrA et l'empêchant de se lier aux 5'UTR des ARNm. L'expression de ces ARNnc s'effectue sous le contrôle d'un système à deux composants dépendant du milieu de croissance et sous le contrôle de la protéine CsrD, capable de recruter la RNase E pour dégrader les ARNm. Des homologues de ces ARNnc et de CsrA ont été également caractérisés chez d'autres bactéries notamment les *Salmonella*, les *Pseudomonas* et les *Vibrio* (pour revue, (Lapouge et al., 2008)).

L'ARN 6S d'*E. coli* a été l'un des premiers ARNnc détecté et séquencé (**Figure 19**). De part sa structure primaire et secondaire, il est capable de mimer un promoteur afin de lier et de séquestrer l'ARN polymérase bactérienne contenant le facteur de transcription  $\sigma^{70}$ . L'ARN 6S s'accumule dans les cellules en phase stationnaire de croissance et s'associe à l'ARN polymérase  $E\sigma^{70}$ . Il ne peut s'associer à  $E\sigma^{S}$  qui est la forme holoenzyme de l'ARN polymérase de la phase stationnaire. L'interaction entre l'ARN 6S et  $E\sigma^{70}$  inhibe ainsi la transcription des gènes dont les promoteurs sont  $\sigma^{70}$ -dépendants et par conséquent augmente l'expression des gènes  $\sigma^{S}$ -dépendant (pour revue, (Wassarman and Storz, 2000)). De plus, l'ARN 6S sert aussi de matrice pour la transcription d'un petit ARN de 14-20 nt noté pARN dont le rôle n'est pas encore connu (voir références citées dans (Waters and Storz, 2009)).

## b) Les éléments cis-régulateurs : les riboswitch

Les molécules d'ARN ont la capacité d'adopter des structures secondaires variées qui peuvent efficacement réguler leur propre expression génique au niveau transcriptionnel ou traductionnel. La modulation de leur structure se fait suite à la liaison à des molécules *trans*-régulatrices (métabolites, ARNt non chargés, protéines) ou encore lors de changements environnementaux (température ou ions divalents) (pour revues récentes (Blouin et al., 2009; Smith et al., 2010; Wachter, 2010)). Ces éléments ARN sont codés en *cis* dans les régions 5'UTR d'ARNm et sont connus sous le terme de « riboswitch », riborégulateurs ou encore ribocommutateurs.



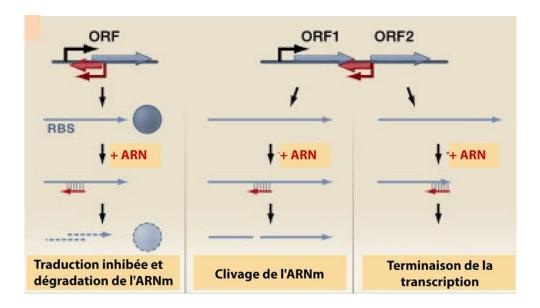
**Figure 21 : Quatre exemples de riboswitch caractérisés chez les bactéries.** Le riboswitch *glmS* permet la maturation de l'ARNm glmS par auto-clivage, le site de clivage localisé en amont du nucléotide +1 est indiqué par une étoile (adapté de (Winkler et al., 2004)). Le riboswitch lysine contrôle l'atténuation de la transcription : suite à la liaison de la lysine au domaine aptamère, un terminateur de transcription se structure, inhibant ainsi la transcription du gène (tiré de (Blouin et al., 2010)). Le riboswitch adénine contrôle positivement l'expression du gène en aval : un terminateur de transcription est présent en 5' de l'ORF, suite à la liaison de l'adénine, le terminateur se défait et la transcription du gène est activée (tiré de (Blouin et al., 2009)). Le riboswitch SAM (S-Adénosylméthionine) contrôle la traduction de l'ARNm : en absence du SAM, la séquence complémentaire au Shine-Dalgarno (notée ASD en rouge pour anti-Shine-Dalgarno) est retrouvée dans une région non structurée, en présence de SAM, cette région ASD forme une hélice avec le site Shine Dalgarno (en vert), rendant le site SD inaccessible aux ribosomes (tiré de (Smith et al., 2010)).

Les riboswitch sont composés de deux domaines : une région aptamère et une plateforme d'expression (**Figure 20**). La région aptamère (ou domaine senseur) est capable de reconnaitre un ligand ou un signal environnemental. La plateforme d'expression peut adopter des structures secondaires qui altèrent les processus de la transcription dans le cas d'atténuation transcriptionnelle ou de la traduction dans le cas d'atténuation traductionnelle (**Figure 20**) (pour revues récentes (Blouin et al., 2009; Smith et al., 2010; Wachter, 2010)).

Différents types de riboswitch sont connus : (i) les riboswitch senseurs de changements environnementaux (les ARN thermosenseurs), (ii) les riboswitch senseurs de métabolites (par exemple, la thiamine pyrophosphate (TPP), la flavine mononucléotide (FMN), l'adénosyl-cobalamine, la S-adénosyl méthionine (SAM) des nucléobases (comme l'adénine ou la guanine), les acides aminés (comme la glycine ou la lysine/L-box), les dérivés de sucres (glucosamine-6-phosphate) ou encore les ions divalents (Mg<sup>2+</sup>)), (iii) les riboswitch régulés par les ARNt (les ARN T-box) et (iv) les riboswitch senseurs de protéines (pour revue (Blouin et al., 2009)) (**Figure 21**).

Les terminateurs Rho-indépendant sont des éléments ARN composés d'une tigeboucle riche en GC suivi d'un stretch de 6 à 9 résidus uridine. Ils trouvés dans les régions 3' terminales de gènes ou d'opérons mais également au niveau des plateformes d'expression des riboswitchs dans le cas d'atténuateur transcriptionnel (**Figure 20**). Dans le cas d'atténuateur traductionnel, l'initiation de la traduction est contrôlée par des changements de la structure secondaire. Les séquences Shine-Dalgarno (ou RBS pour Ribosome Binding Site) et le codon d'initiation AUG peuvent être séquestrés dans des structures d'ARN alternatives. Ce type de contrôle est aussi associé au phénomène de « pause » permettant d'attendre que l'ARNm soit de pleine taille avant le démarrage de la traduction (**Figure 20**). Les riboswitch des bactéries à Gram positif sont plutôt de type transcriptionnel et ceux des bactéries à Gram négatif de type traductionnel (pour revue (Romby and Charpentier, 2010)). Cette différence fonctionnelle peut être expliquée par la prévalence des ARNm polycistroniques trouvés dans les deux types bactériens. En effet, les bactéries à Gram positif possèdent des clusters de gènes plus larges que les Gram négatif, clusters qui gagnent à être régulés par une atténuation de transcription (Romby and Charpentier, 2010).

En plus des mécanismes de transcription et de traduction, les riboswitch peuvent induire directement la dégradation d'ARNm naissant. La séquence leader du gène *glmS* peut adopter une structure alternative de type ribozyme après liaison au cofacteur glucosamine-6-phosphate induisant son auto-coupure. Les fragments générés sont ensuite dirigés vers les voies de dégradation classique des ARN (**Figure 21**) (Collins et al., 2007).



**Figure 22 : Modes d'action des ARNnc** *cis*-antisens chez les bactéries (Tiré de (Waters and Storz, 2009)). Les ARN régulateurs (en rouge) sont codés en *cis*-antisens à des ORF ou des opérons (en bleu). L'appariement parfait peut induire une traduction de l'ARNm cible ou son clivage ou la terminaison de la transcription.

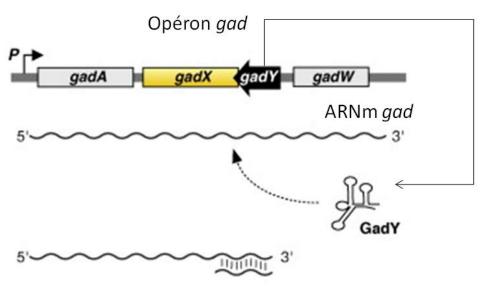
Le riboswitch de type thermosenseur localisé dans le 5'UTR du transcrit du facteur de virulence *prfA* chez la bactérie *Listeria monocytogenes* est particulièrement intéressant. En dessous de 30°C, la région 5'UTR adopte une structure secondaire de façon à rendre inaccessible le site Shine-Dalgarno, empêchant ainsi sa traduction. A une température plus élevée (37°C), cette structure secondaire en « épingle à cheveux » est partiellement déstabilisée permettant ainsi le chargement du ribosome et l'initiation de la traduction. Cette structure peut aussi être modulée par un appariement en phase exponentielle de croissance, avec l'ARN régulateur SreA de 229 nt. Ce régulateur étant identifié, d'autre part, pour être un riboswitch liant la <u>S-a</u>denosyl<u>m</u>ethionine (SAM) contrôlant l'expression des gènes situés en aval (Imo2419-2418-2417) (Henkin, 2008; Roth and Breaker, 2009). Cette étude menée sur SreA a mis en évidence la double fonctionnalité d'un riboswitch de type SAM agissant en *cis* sur l'expression de gènes áloignés (Loh et al., 2009).

#### c) Les ARN régulateurs antisens

Le mécanisme de régulation de l'expression des gènes par des ARN régulateur s'effectue par appariement avec leur cible ARN. Deux types d'ARN régulateurs se distinguent : (i) les ARN *cis*-antisens codés par la même région que leur cible sur le brin opposé et présentant des complémentarités parfaites avec leur cible et (ii) les ARN *trans*-antisens codés dans une région génomique différente de leur cible présentant une complémentarité imparfaite avec un ou plusieurs ARNm cibles (pour revues récentes (Gottesman and Storz, 2010; Waters and Storz, 2009)).

#### (1) Les ARN régulateurs *cis*-antisens

Les ARN cis-antisens sont des ARN codés en cis sur le brin d'ADN opposé à leur cible partageant ainsi des régions étendues de complémentarité parfaite (**Figure 22**) (Brantl, 2007). Chez les bactéries, les exemples d'ARN cis-antisens les plus étudiés sont ceux retrouvés sur les plasmides ou les éléments génétiques mobiles. La majorité des ARNnc cis-antisens codés par les bactériophages, les plasmides et les transposons ont pour fonction de maintenir le nombre de copies suffisant de l'élément génétique mobile (EGM) (pour revues (Brantl, 2007; Wagner et al., 2002)). Cette maintenance est réalisée grâce à différents

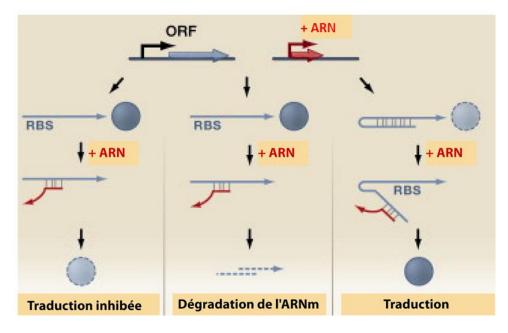


GadY s'apparie en 3' permettant la stabilisation des transcrits *gadX* 

**Figure 23 : Stabilisation de l'ARNm** *gadX* par l'ARNnc GadY (tiré de (Frohlich and Vogel, **2009**)). Le gène de GadY (en noir) est codé en antisens au gène *gadX* (en jaune) présentant ainsi une complémentarité de plus de 30 nt à l'ARNm et permettant ainsi de stabiliser l'ARNm.

mécanismes tels que l'inhibition de la traduction de transposase (ARN-OUT de Tn10). De plus, des ARNnc cis-antisens sont connus pour leur rôle d'antitoxine (pour revue (Fozo et al., 2008a)). Ils agissent par appariement à des ARNm codant pour des protéines qui sont toxiques pour les cellules. Un des systèmes toxine-antitoxine les plus caractérisés est celui porté par les gènes hok-sok du plasmide R1. Le gène hok, pour « host killing », code pour une petite protéine qui cause des dommages membranaires amenant à la mort cellulaire. Dans les cellules contenant ce plasmide, la traduction de l'ARNm hok est réprimée par l'ARN Sok, pour « suppressor of killer ». Lorsque le plasmide est perdu, la stabilité différentielle entre les ARN Sok et *hok* engendre une diminution rapide des quantités d'ARNnc Sok permettant ainsi l'expression de la protéine Hok et la mort cellulaire (pour revue (Storz et al., 2005)). Certains ARNnc antitoxines codés sur le chromosome partagent des homologies aux ARNnc des systèmes toxine-antitoxine portés par les plasmides tels que le système Hok-Sok (Fozo et al., 2008a). D'autres encore sont situés dans des régions acquises grâce à des EGM (Waters and Storz, 2009). Ces différentes observations suggèrent que les ARNnc antitoxines et leurs gènes d'ARNm toxines seraient acquis par transfert horizontal (pour revue (Fozo et al., 2008b)). Les ARN cis-antisens codés par le chromosome sont aussi caractérisés pour leur rôle dans l'induction de la dégradation et/ou la répression de la traduction d'ARNm codant pour des protéines qui sont toxiques à de fortes concentrations. Par exemple, chez E. coli, les gènes de deux ARNnc OhsC (oppresion of hydrophobic ORF by sRNA) et IstR (inhibitor of SOSinduced toxicity by RNA) sont adjacents à des opérons codant pour des protéines potentiellement toxiques. Bien que ces ARNnc ne soient pas de véritables ARN cis-antisens, ils contiennent tous les deux des régions parfaitement complémentaires (19 à 23 nt) aux ARNm issus des opérons ysdAB et ohsC respectivement (pour revues (Fozo et al., 2008a; Fozo et al., 2008b)).

Un autre groupe d'ARN *cis*-antisens est connu pour moduler l'expression de gènes d'un opéron. Certains de ces ARN sont codés dans les régions complémentaires aux régions inter-ORF d'opérons. Pour exemple, l'ARNnc antisens GadY d'*E. coli* qui est capable de s'apparier à l'ARNm *gadXW* engendre le clivage du duplex entre *gadX* et *gadW*. GadY peut aussi stabiliser la quantité de *gadX*: les gènes *gadY* et *gadX* se chevauchent en 3', ceci permet à l'ARNnc GadY de s'apparier sur 30 nt de la région 3'UTR de l'ARNm *gadX* lui conférant ainsi une stabilité (**Figure 23**) (Opdyke et al., 2004; Storz et al., 2004; Tramonti et al., 2008).



**Figure 24 : Caractéristiques des ARNnc** *trans*-antisens (Tiré de (Waters and Storz, 2009)). L'ARN régulateur (en rouge) est codé dans un locus génomique différent de sa cible (en bleu). Il présente des appariements imparfaits à sa cible ARNm. Trois régulations sont possibles : une inhibition de la traduction par appariement au site RBS, un ciblage de l'ARN cible à sa dégradation ou la libération de sites RBS permettant ainsi sa traduction.

# (2) Les ARN régulateurs trans-antisens

Les ARN *trans*-antisens sont fonctionnellement comparables aux microARN eucaryotes (miARN) (Aiba, 2007; Gottesman, 2005). La majorité des ARNnc antisens codés en *trans* sont exprimés sous des conditions de croissance spécifiques notamment lors de conditions de stress oxydatifs, de stress membranaires ou encore lors de variations de quantités de glucose dans le milieu (pour revues, (Gorke and Stulke, 2008; Gottesman, 2005; Gottesman and Storz, 2010; Papenfort and Vogel, 2009)). Il n'y a aucune corrélation entre les localisations génomiques des ARNnc codés en *trans* et leur cible. Un ARNnc donné peut avoir plusieurs cibles ARNm (Gottesman, 2005). Les ARN *trans*-antisens partagent une complémentarité limitée et imparfaite à leurs cibles ARNm (environ 10 à 25 nt) (**Figure 24**). Un « cœur » d'appariement minimal est critique pour le mécanisme de régulation de l'expression de protéines *via* l'inhibition de la traduction et/ou la dégradation de l'ARNm.

Les premiers ARNnc de ce type agissent en se liant à la partie 5'-UTR des ARNm cibles en empêchant l'accessibilité au RBS (**Figure 24**). D'autres sont connus pour inhiber la traduction via un appariement bien en amont de l'AUG des gènes réprimés. Pour citer quelques exemples caractérisés, les ARN MicA et RybB sont capables de réprimer l'expression des gènes codant pour des protéines de la membrane externe bactérienne (pour revues, (Guillier et al., 2006; Papenfort and Vogel, 2009)) : l'ARN OxyS régule négativement l'expression du facteur de transcription codé par le gène *rpoS* (pour revue, (Majdalani et al., 2005)) ou encore l'ARN RyhB dont l'expression augmente en condition de carence en fer, est capable de réprimer l'expression de gènes codant pour des protéines contenant du fer (pour revues, (Masse et al., 2007; Papenfort and Vogel, 2009)).

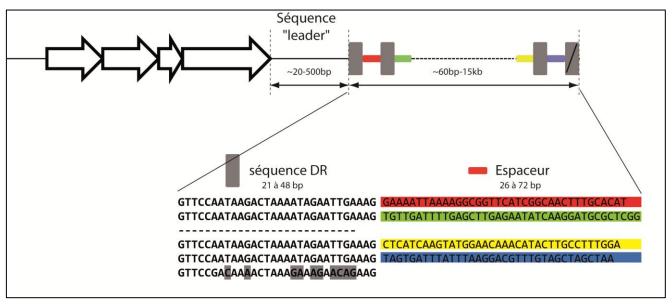
Pour la majeure partie des interactions ARNnc-ARNm caractérisées, l'inhibition de la liaison des ribosomes permet de réduire la quantité de protéines alors que la dégradation du duplex ARNnc-ARNm permettrait d'augmenter la répression engendrée et ainsi de rendre la réaction irréversible (Morita et al., 2008). Chez les entérobactéries, après appariement de l'ARNnc à sa cible, le duplex ARNnc-ARNm serait amené au complexe de dégradation des ARN, le dégradosome, contenant l'endonucléase RNase E (Vogel, 2009). La protéine Hfq dite chaperonne est souvent nécessaire aux régulations médiées par les ARNnc codés en *trans*, probablement pour faciliter l'interaction ARN-ARN (pour revue (Aiba, 2007)). Hfq est une protéine homologue aux protéines Sm et Sm-like qui sont impliquées dans l'épissage et dans la dégradation des ARN chez les eucaryotes. Hfq est active sous forme d'un anneau hexamérique. Ses fonctions ne sont pas encore totalement élucidées bien que plusieurs

spéculations ont été proposées. Elle favoriserait le remodelage des interactions ARN-ARN en défaisant les structures secondaires inhibitrices. Elle servirait de plateforme de « rencontre » entre l'ARNnc et sa cible en augmentant la concentration locale des ARN. Cependant, certains ARNnc sont capables de former des duplexes *in vitro* avec leurs cibles en l'absence d'Hfq (Maki et al., 2008) bien qu'*in vivo* ces ARNnc soient incapables de réguler leurs cibles ARNm dans des souches mutantes pour *hfq*. Tous les ARNnc caractérisés de la classe des ARN régulateurs en *trans* co-immunoprécipitent avec Hfq permettant ainsi d'identifier de nouveaux ARNnc chez *E. coli* et *Salmonella enterica serovar Typhimurium* (Sittka et al., 2008; Zhang et al., 2003). En facilitant les appariements, Hfq jouerait un rôle dans la régulation des ARNnc en modulant leur quantité dans la cellule (pour revue, (Aiba, 2007)). En effet, la majorité des ARNnc chez *E. coli* apparaissent moins stable en l'absence d'Hfq suggérant que celle-ci protègerait les ARNnc de la dégradation. Bien que tous les ARNnc codés en *trans* caractérisés chez *E.* coli aient recours à Hfq pour la régulation de leurs cibles, il existe néanmoins des ARNnc de ce type qui ne nécessitent pas de protéine chaperon (Lee and Groisman, 2010).

#### d) Les CRISPR

Chez les organismes eucaryotes, les petits ARN de la famille des miARN et siARN sont connus pour jouer le rôle de guide dans le mécanisme d'interférence à ARN en ciblant des ARN endogènes ou exogènes. Depuis peu, un mécanisme d'interférence à ARN a été proposé chez les bactéries et les archées. Ce système impliquerait les loci CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats). D'après les bases de données, ces loci seraient retrouvés dans 40 % des génomes de bactéries séquencés et dans la presque totalité des génomes d'archées séquencés. Depuis leur découverte en 1987, différents noms leur ont été attribués : SRSR (Short Regularly Spaced Repeat), DR (Direct Repeat), DVR (Direct Variant Repeats), TREP (Tandem Repeat), LCTR (Long Clusters of Tandem Repeats) et SPIDR (Spacers Interspaced Direct Repeats) (pour revue (Deveau et al., 2010)).

Les loci CRISPR sont des régions d'ADN contenant une région « leader » de taille variable (de 20 à 534 bp) suivie de séquences répétées directes en tandem et palindromiques (séquences DR) de 21 à 48 bp qui sont séparées par des séquences non répétées de 26 à 72 nt appelées séquences « espaceurs ». La séquence leader est riche en bases A et T (Horvath et



**Figure 25 : Organisation d'un locus CRISPR (adapté de (Deveau et al., 2010)).** Les loci CRISPR sont caractérisés par des séquences répétées directes palindromiques (en gris) qui sont régulièrement séparées par des séquences espaceurs (en couleur) de séquence unique. Ces séquences sont précédées d'une séquence dite « leader » qui serait déterminante pour la transcription d'ARN Cr et de l'acquisition *de novo* de séquences espaceurs. On retrouve fréquemment des opérons de gènes *cas* représentés par des flèches.

al., 2008; Jansen et al., 2002a) et jouerait le rôle de séquence promotrice (**Figure 25**) (Hale et al., 2008; Jansen et al., 2002a; Lillestol et al., 2006; Tang et al., 2002a; Tang et al., 2005).

Dans un locus (ou cassette) donné, la taille des répétitions est conservée, à l'exception des dernières répétitions qui sont souvent dégénérées (Horvath et al., 2008; Jansen et al., 2002a; Jansen et al., 2002b). Le nombre de répétitions et de séquences espaceurs dans un locus CRISPR donné est variable d'une souche à l'autre d'une même espèce. On peut dénombrer de deux à 375 répétitions pour un seul locus (**Figure 25**) (pour revue (Deveau et al., 2010)). Plusieurs cassettes CRISPR peuvent être présentes dans un même génome. (Gottesman and Storz, 2010; Grissa et al., 2007a; Horvath et al., 2009).

Depuis leur découverte en 1987, plusieurs fonctions leur ont été attribuées de la partition des réplicons bactériens aux processus de réparation de l'ADN et/ou à la stabilité du génome ((Makarova et al., 2002; Mojica et al., 1995) et pour revue, (Sorek et al., 2008)). Ce n'est qu'en 2005 qu'il a été montré que les séquences des espaceurs présentaient des similarités avec des éléments extra-chromosomiques tels que les phages et les plasmides (Bolotin et al., 2005; Mojica et al., 2005; Pourcel et al., 2005). Un lien entre les séquences espaceurs et les phages fut proposé suggérant que les cassettes CRISPR seraient impliqués dans un système de défense contre l'invasion d'acides nucléiques étrangers via un mécanisme semblable à l'interférence à ARN connue chez les eucaryotes (Bolotin et al., 2005; Marraffini and Sontheimer, 2008). L'isolement de bactéries présentant une résistance à l'infection de bactériophages et la caractérisation de leurs loci CRISPR ont permis de mettre en évidence que ces bactéries étaient capables d'intégrer de nouvelles séquences espaceurs dans les cassettes CRISPR (Barrandon et al., 2008; Deveau et al., 2008). Ces séquences espaceurs présentent une identité parfaite à la séquence génomique du phage d'origine et seraient intégrées dans la région 5' du locus CRISPR entre la séquence leader et la première séquence DR. Par ailleurs, ce mécanisme de défense serait aussi dirigé contre les plasmides conjugatifs chez des souches de Staphylocoques (Marraffini and Sontheimer, 2008).

Fréquemment, on retrouve à proximité des loci CRISPR des gènes associés aux CRISPR (gènes *cas* pour CRISPR-associated) (**Figure 25**) (Jansen et al., 2002a). Ces gènes sont généralement retrouvés sous forme d'opéron. Le nombre de gènes *cas* peut varier de 4 à plus de 20 (Barrangou et al., 2007; Haft et al., 2005; Sorek et al., 2008). Une étude bioinformatique de ces gènes provenant de 200 génomes séquencés a permis de les classifier en 45 familles (Haft et al., 2005). Six gènes *cas* « cœur » (*cas1* à *cas6*) sont généralement retrouvés dans la majorité des organismes analysés avec de nombreuses exceptions et le gène *cas1* semble être présent dans la totalité des génomes pourvus de CRISPR (Haft et al., 2005).

Dans chaque organisme, cet ensemble de gènes « cœur » est souvent accompagné par d'autres gènes additionnels dont les *cse*, *csy*, *csn*, *csd*, *cmr* pour n'en citer que quelques uns (Haft et al., 2005). Ces familles de gènes comprennent 2 à 6 gènes co-ségrégent et leur nom proviennent le plus souvent de l'organisme dans lequel ils ont été identifiés (pour exemple, les *cse* pour *cas* <u>s</u>ubtype *Escherichia coli* genes) (pour revue, (Deveau et al., 2010)).

Un mécanisme d'acquisition de nouveaux espaceurs appelé mécanisme d'adaptation et un mécanisme de défense appelé mécanisme d'interférence commencent à émerger à partir de données actuelles (Deveau et al., 2010). Le mécanisme d'adaptation n'est pas encore très bien caractérisé notamment lors des étapes de sélection et d'intégration de nouvelles séquences espaceurs dans les cassettes CRISPR (Deveau et al., 2008; Makarova et al., 2006; Tyson and Banfield, 2008; Vestergaard et al., 2008). Les proto-espaceurs (séquences sélectionnées) seraient constitués d'un motif PAM (protospacer adjacent motifs) qui serait important dans le mécanisme d'adaptation. Par ailleurs, ce motif serait impliqué dans la détermination de l'orientation de l'espaceur à intégrer. L'acquisition de nouveaux espaceurs est observée en 5' des cassettes CRISPR directement en amont de la séquence leader (Barrangou et al., 2007; Deveau et al., 2008; Pourcel et al., 2005) suggérant que cette séquence leader jouerait potentiellement un rôle important dans l'acquisition de nouveaux espaceurs (Mojica et al., 2009). Quant au mécanisme d'interférence orchestré par le système CRISPR/cas, il serait similaire au système d'interférence aux siARN des eucaryotes (Makarova et al., 2006; Sorek et al., 2008). Le système CRISPR/cas caractérisé chez la bactérie Staphylococcus epidermidis serait impliqué dans la prise en charge de l'ADN exogène (plasmide) (Marraffini and Sontheimer, 2008).

## 2. Les ARN non codant régulateurs chez les archées

En comparaison aux deux autres domaines du vivant, la description d'ARNnc régulateurs chez les archées est très peu documentée. En effet, à l'exception des ARNr, des ARNt et des ARN guides de modification à boîtes C/D et H/ACA, d'autres ARNnc ont été mis en évidence mais peu de données sur leur fonction et mode d'action sont disponibles (Jager et al., 2009; Klein et al., 2002; Schattner, 2002; Soppa et al., 2009; Straub et al., 2009; Tang et al., 2002a; Tang et al., 2005; Wurtzel et al., 2010; Zago et al., 2005).

Cette section est divisée en deux parties : une première partie fera un état de l'art sur les méthodes bioinformatiques et les approches biologiques qui ont été utilisées pour

rechercher des ARNnc dans les génomes d'archées et la deuxième partie aura pour objectif de décrire les différents ARNnc identifiés tout en apportant les hypothèses de fonction données dans la littérature.

#### a) Recherche d'ARN non codant chez les archées

Les premiers travaux de recherche d'ARNnc chez les archées ont émergé après une étude montrant l'existence de 67 sites de méthylation dans les ARNr du crenarchaeote Sulfolobus solfataricus. Ainsi par des méthodes de spectrométrie de masse et de chromatographie, la présence d'ARN guides de modification a été suggérée (Noon et al., 1998). Des approches bioinformatiques ont permis d'identifier des homologues des ARNnc guides de modifications eucaryotes dans les génomes d'archées disponibles (Gaspin et al., 2000; Grosjean et al., 2008; Klein et al., 2002; Muller et al., 2008; Omer et al., 2000; Schattner, 2002). Une autre approche dite « biologique » a permis d'élargir cette recherche en construisant des banques d'ADNc par shotgun cloning ou RNomics (Tang et al., 2002a; Tang et al., 2005), approche utilisée avec succès notamment pour l'identification des ARN guides à boîtes C/D et H/ACA chez la souris (Huttenhofer et al., 2001). De plus, une étude a permis l'identification de nouveaux ARNnc chez S. solfataricus par construction d'une banque d'ADNc à partir d'ARNnc qui sont co-précipités avec la protéine L7Ae (Zago et al., 2005). Plus récemment, avec le développement des approches globales par séquençage à haut débit, de nouveaux ARN ont également été identifiés chez Methanosarcina mazeï et Sulfolobus solfataricus (Jager et al., 2009; Wurtzel et al., 2010).

#### (1) Les approches bioinformatiques

Les méthodes bio-informatiques sont de deux types : (i) les approches *ab initio* qui comprennent l'analyse du biais de composition en nucléotides, la minimisation de l'énergie libre de structure en ARN et la recherche de famille d'ARNnc connues et (ii) les approches par analyse comparative de génomes.

#### (a) L'approche ab initio

L'approche ab initio (du latin « au départ ») permet de rechercher des gènes candidats dans un génome. La méthode par le biais de composition en nucléotide s'appuie sur l'analyse des caractéristiques intrinsèques de la séquence génomique afin de prédire des candidats ARNnc propres au génome analysé. L'analyse de 764 organismes procaryotes pour lesquels les molécules d'ARN stables structurées (ARNr et ARNt) présentent un contenu intramoléculaire en nucléotides G et C plus élevé que le reste du génome (Galtier and Lobry, 1997) a servi de base pour prédire de nouveaux ARNnc par l'exploitation du biais de composition en nucléotides chez M. jannashii et P. furiosus (Klein et al., 2002; Schattner, 2002). Klein et ses collaborateurs ont modélisé le biais de composition par un modèle de Markov caché à deux états : un état « ARN » et un état « AUTRE ». En d'autres termes, les probabilités émises pour l'état « AUTRE » correspondent à un faible contenu en GC dans tout le génome alors que les probabilités émises pour l'état « ARN » sont définies par un contenu élevé en GC. Dans un premier temps, ils ont montré qu'il n'y a pas de corrélation entre la température optimale de croissance et le contenu en nucléotides G et C de 52 génomes procaryotes disponibles en 2001. Cependant, lorsque l'on regarde la différence du contenu en GC des ARNt par rapport à celui du génome en fonction de la température optimale de croissance, cette différence est élevée non seulement pour les génomes d'hyperthermophiles annotés mais aussi pour quelques organismes mésophiles suggérant que le crible par le biais de composition en nucléotides serait aussi applicable à ces organismes. Par l'exploitation de ce biais, 43 régions différentes ont été proposées chez M. jannashii comprenant les 37 ARNt connus, les ARNr, l'ARN de la RNase P, l'ARN 7S SRP et 9 régions correspondant à des candidats pouvant coder pour de nouveaux ARNnc (Schattner, 2002). Chez P. furiosus, l'analyse par le biais de composition en nucléotides n'a permis de détecter que seulement 67 % des ARNt connus, ce qui a motivé les auteurs à développer une approche complémentaire par analyse comparative (Klein et al., 2002).

L'analyse du biais de composition en nucléotides est parfois associée à la recherche de signatures d'ARNnc connus tels que des signaux de transcription (promoteurs et terminateurs) ou de structure propre aux familles d'ARN connues. Sur la base des alignements des ARN d'une même famille, il existe des logiciels qui permettent de créer des profils covariants qui servent ensuite à effectuer des recherches automatisées dans un génome. Ce type d'approche a permis de concevoir le programme dédié à la recherche systématique d'ARNt: tRNAScan-SE, programme qui combine trois méthodes de recherche d'ARNt. Lors de son développement, son taux de réussite représentait 99 à 100% de gènes d'ARNt détectés dans une séquence d'ADN avec moins d'un « faux-positif » par 15 Gb (Lowe and Eddy, 1997).

Cette approche a également permis de rechercher les ARN guides à boîtes C/D dans les génomes eucaryotes puis d'archées (Lowe and Eddy, 1997; Omer et al., 2000). L'alignement des séquences de 17 ARN de cette famille a défini d'une part, une matrice donnant la fréquence de chaque nucléotide à chaque position dans les boîtes caractéristiques de ces ARN (les boîtes C, D', C' et D), et d'autre part, un score de distribution des tailles des séquences ciblant les ARNr (8 à 11nt), proposant alors un modèle qui a permis de cribler les génomes d'archées disponibles en 2002 dont ceux des trois Pyrococcales *P. abyssi*, *P. furiosus* et *P. horikoshii* (Omer et al., 2000). La prédiction d'ARN sur la base de critères connus a aussi été appliquée au génome de l'archée halophile *Haloferax volcanii* (Grosjean et al., 2008). Les auteurs ont utilisés les logiciels Patscan et Darn! <a href="http://carlit.toulouse.inra.fr/Darn/">http://carlit.toulouse.inra.fr/Darn/</a>. Ces logiciels sont des outils de recherche automatisés qui permettent de rechercher des motifs dans une séquence génomique à l'aide d'un descripteur qui modélise la famille d'ARN recherché, ainsi les ARN guides à boîtes C/D et H/ACA ont été prédits chez *H. volcanii*.

Parmi les outils utilisés pour rechercher des loci CRISPR, on peut notamment citer Tandem Repeat Finder (Benson, 1999) et Location Uniform poly-Nucleotide Areas (LUNA) (Lillestol et al., 2006; Peng et al., 2003). L'outil Patscan a aussi servi pour identifier des CRISPR contenant au minimum trois à quatre répétitions (Godde and Bickerton, 2006; Jansen et al., 2002b). Le logiciel PYGRAM combiné aux données générées par Patscan a permis entre autre de visualiser toutes les séquences répétées sur un génome. Plus récemment, deux logiciels dédiés à la recherche de CRISPR ont été proposés : CRT et PILER-CR (http://www.room220.com/crt; (Edgar, 2007)). Tous deux présentent des capacités de calcul rapide et semblent donner des résultats corrects par rapport aux définitions des CRISPR (répétitions et espaceurs, voir section du manuscrit correspondante). En 2007, le site web CRISPRFinder développé à l'IGM à Orsay a été mis en ligne publiquement (http://crispr.upsud.fr). Il est spécialisé dans l'identification automatique de loci CRISPR. Il présente l'avantage de détecter les loci CRISPR courts, de définir les extrémités exactes, d'extraire les séquences des espaceurs, d'interroger les banques de données Genbank pour l'étude des espaceurs et de posséder leur propre base de données répertoriant les loci CRISPR détectés chez les bactéries et les archées. Plus récemment, un nouvel outil nommé CRISPI a été proposé pour identifier les systèmes CRISPR/Cas (http://crispi.genouest.org/). Cette base de données intègre les outils PYGRAM, CGView et WebLogo et aussi le service Blast comme dans CRISPRFinder (Grissa et al., 2007b).

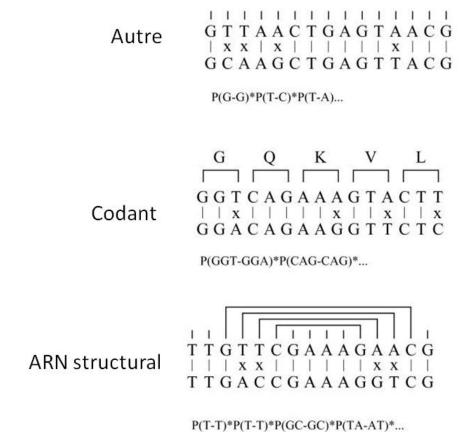


Figure 26 : Schématisation des trois états définis par QRNA (Tiré de (Rivas and Eddy, 2001)). Trois alignements deux à deux peuvent être classés suivant l'indépendance des bases (autre), les séquences codantes (codant) et les ARN structurés (ARN structural).

# (b) L'analyse comparative de génomes

L'analyse comparative des génomes est devenue un outil de choix pour la recherche d'ARNnc par l'augmentation du nombre de génomes séquencés. Le principe de cette approche repose sur l'hypothèse d'une conservation en séquence et en structure d'ARN homologues. Différents outils pour ce type de recherche *in silico* ont été développés à ce jour pour la recherche dans les génomes eucaryotes, bactériens et d'archées.

L'outil QRNA a été utilisé avec succès sur les génomes de *E. coli* et de *P. furiosus* (Klein et al., 2002; Rivas and Eddy, 2001). Sur la base de l'alignement de deux séquences, il consiste à l'analyse de profil de substitution de nucléotides conservés et à classer les séquences correspondent en (i) région codante, (ii) région contenant un ARN structural ou (iii) région autre (**Figure 26**). En effet, une région codante aura un profil de substitution synonyme (un triplet différent correspond à un même acide aminé), tandis qu'un ARN structural conservé aura tendance à avoir un profil de changements compensatoires qui resteraient compatibles à la structure du même ARN. Par ce crible, 73 régions candidates ont été proposées chez *P. furiosus* en comparaison aux génomes des autres Pyrococcales : *P. abyssi* et *P. horikoshii*. Parmi ces régions, 45 des 46 ARNt ont été prédits, ainsi que tous les ARNr et l'ARN 7S. Des filtres manuels ont été appliqués afin d'éliminer toute prédiction recouvrant partiellement ou entièrement des ORF annotées et en ne sélectionnant que les prédictions de séquence supérieure à 50 nt. Au final, 17 candidats ARN ont été proposés (Klein et al., 2002).

Afin de rechercher des ARN guides de modification à boîtes H/ACA, Muller et ses collaborateurs ont utilisé une approche par analyse comparative de génomes combinée à une recherche de motifs caractéristiques qui a permis d'identifier cinq candidats putatifs conservés dans les génomes des trois *Pyrococcus* séquencés (Muller et al., 2008). Une première étape avait consisté à extraire toutes les régions intergéniques des trois génomes. Puis une analyse statistique a permis d'isoler les régions intergéniques conservées sur la base de l'observation d'une conservation en séquence d'au moins 18 nt ou de 21 nt en autorisant la présence d'une seule substitution dans la séquence. Puis la recherche de motifs a été effectuée grâce à l'outil RNAMOT qui permet de rechercher la présence des motifs structuraux caractéristiques des ARN guides à boîtes H/ACA : deux hélices séparées par une boucle interne (modélisant la poche de pseudourydilation) (Gautheret and Lambert, 2001).

Les nombreuses approches bioinformatiques développées se sont le plus souvent limitées à la recherche d'ARNnc dans les régions intergéniques. Cependant, certaines régions

annotées ORF (Open Reading Frame) pourraient potentiellement correspondre à des gènes d'ARNnc. De plus, les gènes d'ARNnc antisens ou chevauchant partiellement des ORF ne peuvent pas être détectés par ce type d'approche.

# (2) Les approches dites « biologiques »

Chez les bactéries, les premières détections d'ARNnc ont été faites sur de simples cribles génétiques puis, la technologie des puces à ADN s'est étendue à la recherche d'ARNnc par la construction de puces recouvrant les régions intergéniques. Chez les organismes eucaryotes, la recherche d'ARNnc a explosé avec le développement des approches par la construction et le clonage de librairies d'ADNc codant pour des ARNnc préalablement purifiés sur gel en fonction de leur taille. Des méthodes par co-précipitation avec des protéines ont aussi été élaborées. Enfin, les approches les plus récentes utilisées avec succès dans des organismes des trois domaines du vivant sont celles qui utilisent la technologie de séquençage à haut débit (pour revues (Altuvia, 2007; Sharma and Vogel, 2009; Vogel and Sharma, 2005)). Chez les archées, quelques travaux par des approches biologiques ont été publiés. Ces approches biologiques par « shotgun-cloning » (reverse transcription suivi de clonage et séquençage) ou par séquençage direct (RNA-seq). Les ARN analysés proviennent (i) d'une co-immunopurification avec des protéines, ou (ii) d'une sélection en fonction de leur taille ou (iii) correspondent aux ARN totaux cellulaires.

Omer et ses collaborateurs ont utilisé une approche biologique par coimmunoprécipitation des ARN avec les protéines cœur de la RNP guide à boites C/D : la
fibrillarine et NOP56/58. L'isolement des ARN des ces complexes et leur amplification par
RT-PCR puis de leur clonage et leur séquençage a permis d'identifier 104 ARNnc. Parmi
ceux-ci, 18 d'entre eux possèdent les caractéristiques des ARN guides C/D et les autres clones
représentant des fragments des ARNr 16S, 23S et 5S (Omer et al., 2000). De même, Zago et
ses collaborateurs ont co-purifié tous les ARN associés avec la protéine L7Ae chez *S.*solfataricus. Les 128 clones ainsi générés correspondent à 45 séquences uniques. Ainsi, 32
ARN autres que les ARNr, ARNt et ARN 7S (SRP) furent identifiés. Sur la base de leur
séquence, de leur motifs structuraux et de leur localisation génomique, ces ARN ont été
classés en 4 groupes : (i) les ARN de séquences et/ou de motifs structuraux connus, (ii) les
ARN sens, (iii) les ARN retrouvés dans les régions intergéniques (IGR) et (iv) les ARN
antisens (Zago et al., 2005).

Le fractionnement des ARN par leur taille a permis de mettre en évidence de nouveaux ARN dont les miARN et des siARN chez les eucaryotes. Typiquement, les protocoles incluent un fractionnement des ARN totaux sur gel d'acrylamide sélectionnant les ARN selon leur taille. Les ARN sont ensuite extraits puis utilisés comme matrice pour construire des librairies d'ADNc qui sont ensuite séquencés. Trois études chez les archées ont utilisé cette approche pour l'identification de nouveaux ARNnc. Cette méthode a été appliquée aux génomes de A. fulgidus (Tang et al., 2002a), S. solfataricus (Tang et al., 2005) et H. volcanii (Straub et al., 2009). Les deux travaux de Tang et collaborateurs se sont focalisés sur des ARN de taille comprises entre 50 et 500 nt alors que Straub et collaborateurs ont purifié les ARN de H. volcanii de taille comprise entre 130 et 460 nt. Ces trois études ont ensuite utilisé la même stratégie pour la reverse-transcription des ARN isolés sur gel et se sont servies d'un crible par hybridation de colonies sur filtres pour éliminer les ARNnc déjà connus (ARNr, ARNt, ARN 7S). Après analyse des différents clones obtenus (élimination des redondances et des ARN retrouvés dans des régions annotées comme ORF), 182 ARNnc ont été identifiés dans les 3 études confondues dont 86 pour A. fulgidus, 57 pour S. solfataricus et 39 chez H. volcanii. L'expression de la plupart de ces ARN a été validée par une analyse en Northern Blot. Ces trois études ont permis de mettre en évidence des transcrits issus de cassettes CRISPR ainsi que des ARNnc portant des motifs structuraux connus (ARN guides de modification à boîtes C/D et H/ACA) et d'autres ARNnc (retrouvés dans des IGR, antisens à des ORF ou encore chevauchant des ORF). Cette approche ne permet d'identifier qu'une fraction des ARNnc de la cellule selon la taille des ARN choisie et elle est limitée par le choix des conditions physiologiques de croissance des cellules et exclut un certain nombre d'ARN structurés peu amplifiés lors des étapes de reverse-transcription.

L'approche par séquençage à haut débit a récemment émergé comme méthode de choix pour l'analyse globale des transcriptomes et donc pour l'identification d'ARNnc. Contrairement à l'approche par RNomics, le séquençage à haut débit nécessite moins de temps et de mise en œuvre. Elle n'utilise pas la technologie de séquençage de type Sanger mais d'autres technologies telles que le pyroséquençage, la terminaison cyclique réversible ou encore le séquençage par ligature. Cette approche est connue sous différentes appellations : deep sequencing, high troughput sequencing ou encore par abus de langage, RNA-sequencing (RNA-seq). Quelque soit la technologie utilisée, des millions d'ADNc peuvent être analysés à partir du moment où tous les ARN sont reverse-transcrits en ADNc et amplifiés via des oligonucléotides adaptateurs (pour revue, (Metzker, 2010)). L'utilisation de cette approche nécessite un traitement bioinformatique des données. Par la technologie de pyroséquençage au

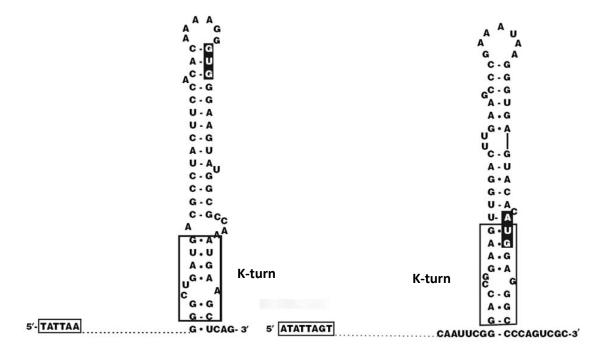


Figure 27 : Deux exemples d'ARN cis-régulateurs putatifs co-précipités avec la protéine L7Ae chez S. solfataricus (tiré de (Zago et al., 2005)). Ces deux ARNnc sont transcrits à partir de loci chevauchant les extrémités 5' d'ORF annotées. Les codons d'initiation de la traduction sont surlignés en noir, les motifs promoteurs putatifs sont encadrés à l'extrémité 5'. Les motifs k-turn retrouvés dans ces ARN sont encadrés sur la structure secondaire proposée.

454 (Roche FLX sequencer), Jäger et ses collaborateurs ont présenté une cartographie génomique globale pour les sites de départ de transcription et pour les gènes d'ARNnc chez M. mazeï cultivée dans différentes conditions (Jager et al., 2009). Un maximum de 94 % des ADNc obtenus sont des séquences de lecture de plus de 18 bp. Une moyenne de 30% des ADNc obtenus sont cartographiés dans les IGR. Un total de 248 candidats pour des ARNnc a été proposé. 57 des 77 candidats sélectionnés aléatoirement ont été vérifiés expérimentalement par Northern Blot. Un seul ARN guide de modification à boîtes C/D a été identifié par cette approche, ceci a été justifié par soit une limite de détection, soit un biais de préparation de la librairie d'ADNc. La seconde technologie utilisée est la terminaison cyclique réversible (Illumina/Solexa's GA<sub>II</sub>). Wurtzel et ses collaborateurs ont appliqué cette méthode de séquençage sur des cellules de S. solfataricus P2 cultivées dans différentes conditions de croissance. Par cette approche, ils ont présenté une cartographie globale du génome ainsi qu'une cartographie de tous les départs de transcription. Ils ont aussi identifié 310 candidats ARNnc comprenant la totalité ARNnc préalablement identifiés par les deux travaux de 2005 : les ARNnc associés aux transposons, les ARNnc issus de loci CRISPR et les ARN guides de modification à boîtes C/D (Tang et al., 2005; Zago et al., 2005) ainsi que l'ARN du ribozyme RNase P qui n'avait pas été annoté dans le génome de S. solfataricus. 60% des ARNnc détectés par cette méthode correspondent à des ARNnc cis-antisens (Wurtzel et al., 2010).

#### b) Fonctions des nouveaux ARN non codant chez les archées

Différentes classes d'ARNnc ont été déterminées en fonction de leur localisation génomique : (i) chevauchant, ou (ii) antisens à des ORF annotés, (iii) dans des régions intergéniques ou (iv) dans des loci CRISPR. De plus, les travaux réalisés par séquençage à haut débit ont mis en évidence des transcrits dont les gènes contiendraient aussi des ORF qui ont été annotées pour coder de petites protéines. Ces observations ont conduits à l'hypothèse que ces ARN pourraient être bifonctionnels (Jager et al., 2009; Wurtzel et al., 2010).

#### (1) ARNnc en cis

Un grand nombre et une grande variété de riboswitch ont été identifiés chez les bactéries ainsi que chez quelques eucaryotes. Plusieurs travaux bioinformatiques évoquent la conservation dans les trois domaines du vivant du riboswitch *thi*-box dont le ligand est la



Figure 28: Deux exemples d'ARN cis-antisens à des ORF de transposases détectés chez S. solfataricus (Tiré de (Zago et al., 2005)). Les séquences de deux ARN sR126 et sR127 sont représentés en complémentaire aux deux ORF de transposases ISC 1439 et ISC 1359. Les codons d'initiation des ARNm des transposases ISC1439 et ISC1359 sont surlignés en noir.

thiamine pyrophosphate TPP (Barrick and Breaker, 2007). Des travaux chez des archées méthanogènes de l'ordre des Methanosarcinales évoquent des régulations de l'expression de gènes par des séquences de type « leader » retrouvées dans les 5'UTR de gènes qui sont impliqués dans des processus de biogénèse du méthane (Anderson et al., 2009; Bose et al., 2009). Chez H. volcanii, la présence d'une structure secondaire dans la région 5'UTR d'un ARNm serait impliquée dans la régulation de la traduction du ARNm suggérant que cette région 5'UTR aurait une fonction de riborégulateur (Brenneis and Soppa, 2009). Très récemment, l'analyse globale du transcriptome de l'archée Methanosarcina mazeï par séquençage à haut débit a mis en évidence que des ARNm étaient pourvus d'une région 5'-UTR (pouvant atteindre 500 nt) suggérant que des riboswitchs existeraient chez les Methanosarcinales (Jager et al., 2009). Des ARNnc chez A. fulgidus et S. solfataricus sont codés par des regions chevauchantes la partie 5' d'ORF annotées pour des transposases hypothétiques, une hydrogène lyase (Tang et al., 2002a; Zago et al., 2005). De plus, certains d'entre eux contiennent un motif K-turn et forment un complexe RNP en présence de la protéine L7Ae (Figure 27) (Zago et al., 2005). Cette observation peut être mise en parallèle à deux études caractérisant deux types de riboswitch chez la bactérie B. subtilis pour lesquels un motif K-turn était observé. Bien que B. subtilis ne possède pas la protéine L7Ae, ce motif permet la liaison in vitro de la protéine L7Ae sur le riboswitch, suggérant que le motif K-turn permettrait la liaison à d'autres protéines qui serait importante dans la stabilisation du complexe ligand-aptamère (Blouin and Lafontaine, 2007; Heppell and Lafontaine, 2008).

#### (2) Des ARNnc cis-antisens

De nombreux ARN cis-antisens ont été identifiés aussi bien chez P. furiosus (Klein et al., 2002), S. solfataricus (Tang et al., 2005; Wurtzel et al., 2010; Zago et al., 2005), H. volcanii (Straub et al., 2009) et M. mazeï (Jager et al., 2009). Ces ARN cis-antisens sont aussi bien antisens aux parties 5' et 3' des ORF annotées ou encore antisens dans les parties centrales des ORF. La plupart de ces ARN sont antisens à des ORF annotées comme protéines hypothétiques. Chez S. solfataricus, une vingtaine d'ARNnc identifiés sont cis-antisens à des gènes de transposases suggérant qu'un mécanisme similaire à celui de Tn10 chez E. coli (Case et al., 1989) existerait aussi chez les archées et permettrait ainsi de contrôler les événements de transposition (Figure 28). Par ailleurs, il a été observé chez M. mazeï que l'expression de ces ARNnc cis-antisens à des transposases serait différentielle en fonction des

conditions de croissance limitantes ou non en nitrogène (Jager et al., 2009). Bien que des fonctions associées à ces ARNnc soient proposées, il n'existe pas encore d'étude de caractérisation fonctionnelle de ces ARN.

# (3) Des ARNnc issus de régions intergéniques

Des ARNnc issus de régions intergéniques ont été identifiés chez *A. fulgidus* (Tang et al., 2002a), *S. solfataricus* (Zago et al., 2005) et chez *H. volcanii* (Straub et al., 2009). Pour quelques uns d'entre eux, des cibles potentielles ont été proposées, notamment des ARNm de transposases ou encore, des ARNm codant des protéines putatives impliquées dans le métabolisme du phosphate inorganique et du sulfure. Chez *H. volcanii*, les premières caractérisations phénotypiques de deux souches, chacune délétée pour deux ARNnc révèlent des variations en quantités de certaines protéines suggérant que des mécanismes de régulation seraient altérés par l'absence de ces ARNnc dans des conditions de culture à haute température et à faible concentration de sels et des analyses globales de leurs protéomes (Straub et al., 2009).

#### (4) Les protéines Sm et les ARN

Les protéines Sm et Sm-like (Lsm) sont retrouvées dans les trois domaines du vivant et sont connues pour être impliquées dans le métabolisme des ARN. La protéine liant les ARNnc bactériens, Hfq, ne partage pas de similarités avec les protéines Lsm eucaryotes au niveau primaire mais possède des similarités au niveau de sa structure tridimensionnelle. Le domaine eucaryote présente une plus grande diversité en protéines de la famille Sm/Lsm. On peut dénombrer jusqu'à 18 protéines Sm et Lsm différentes de fonctions caractérisées dans l'épissage des ARNm, la maturation des histones, la maintenance des télomères ou encore dans la dégradation des ARNm (Thore et al., 2003). Des protéines Lsm sont aussi retrouvées chez les archées de l'ordre des Archaeoglobales, chez les Pyrococcales et des archées halophiles suggérant l'existence d'un ancêtre commun avec les eucaryotes (Salgado-Garrido et al., 1999). Les génomes d'archées peuvent coder une à trois protéines Lsm (Lsm1 à Lsm3) à l'exception de *M. jannashii* qui code pour une protéine Hfq-like (Khusial et al., 2005). Les études de ces protéines chez les archées apportent généralement des informations sur leurs caractéristiques structurales et indiquent souvent que ces protéines sont capables de se lier aux

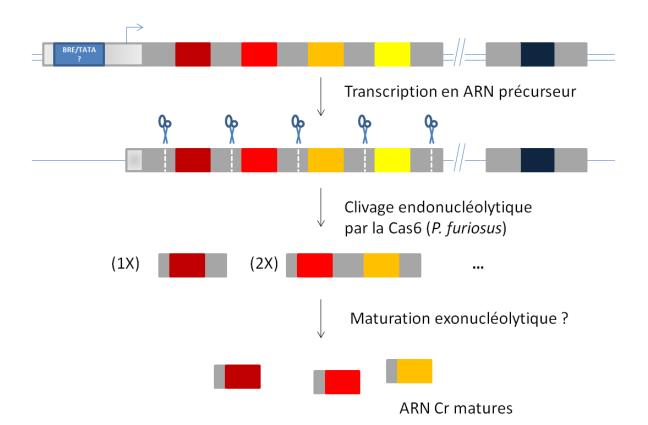


Figure 29 : Modèle de maturation des ARN Cr mature (adapté de (Carte et al., 2008)). Le locus CRISPR est transcrit à partir de la séquence leader représentée en gris clair. Les séquences espaceurs correspondent aux rectangles colorés, les séquences DR sont en gris foncés. L'ARN précurseur transcrit subit des clivages endonucléolytiques représentés par des ciseaux au et des pointillés blancs situés dans les séquences DR. Différents intermédiaires de maturations font l'objet de maturations endo et exonucléolytiques par des protéines non caractérisées. Les ARN Cr matures sont composés d'une portion de séquence DR et de la séquence espaceur associée.

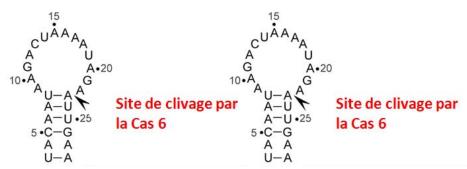


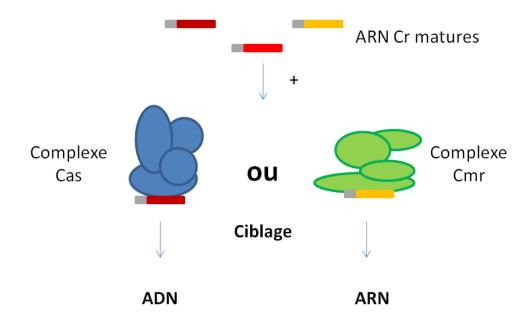
Figure 30 : Représentation schématique de la structure secondaire potentielle formée par une séquence DR palindromique chez *P. furiosus*. Le site de clivage par l'endonucléase Cas6 est indiqué par une pointe de flèche.

ARN (voir références citées dans (Fischer et al., 2010)). Très récemment, une étude sur la caractérisation fonctionnelle *in vivo* et *in vitro* de la protéine Lsm chez *H. volcanii* montre que cette protéine forme des homoheptamères et est capable de lier des ARN (ARN synthétiques et ARNt) (Fischer et al., 2010). Des expériences de co-immunoprécipation ont identifiées 33 partenaires protéiques (protéines ribosomiques, facteurs d'élongation, ARNt synthétases, chaperons et ribonucléases et 20 ARN différents. La liaison entre Lsm et ces ARN suggère que cette protéine d'archées aurait une fonction similaire à la protéine Hfq cependant, aucune cible aux ARN détectés n'a été trouvée par cette approche (Fischer et al., 2010).

#### (5) L'immunité cellulaire : fonction des CRISPR

La présence de clusters de séquences répétées en tandem a été observée pour la première fois chez les archées dans les génomes des halophiles *H. volcanii* et *H. mediterranei* (Mojica et al., 1995). L'observation de clusters similaires a été par la suite rapportée suite au séquençage du génome de *M. jannashii* (Bult et al., 1996). Le nombre grandissant de génomes séquencés a permis de rechercher plus globalement la représentation de ces loci dans les génomes bactériens et d'archées révélant ainsi que ces loci CRISPR sont présents dans près de 88% des génomes d'archées et 39% des génomes bactériens disponibles au moment des analyses (Godde and Bickerton, 2006; Kunin et al., 2007).

La première observation de transcription de cassettes CRISPR a été faite chez l'archée A. fulgidus lors de la recherche systématique d'ARNnc par RNomics (Tang et al., 2002a). Des analyses par détection de transcrits en Northern Blot ont révélé un profil atypique où plusieurs transcrits correspondent à un multiple de la taille de l'espèce la plus courte. Ces analyses suggèrent alors que ces régions sont transcrites sous forme d'un long précurseur ARN subissant des étapes de maturation (**Figure 29**). Des résultats similaires ont été obtenus chez d'autres archées : S. solfataricus et S. acidocaldaricus (Lillestol et al., 2009; Tang et al., 2005). Seules les études réalisées chez S. acidocaldaricus suggèrent que des transcrits antisens aux loci CRISPR sont aussi exprimés (Lillestol et al., 2009). Des analyses de détection par Northern Blot révèlent un profil d'expression particulier où la forme ARN Cr mature est détectée mais aussi deux autres espèces intermédiaires 2X et 1X, l'une correspondrait à deux séquences espaceurs et l'autre à une seule séquence espaceur encadrée



**Figure 31 : Schématisation des mécanismes d'action portés par les ARNCr.** En fonction de la nature de l'acide nucléique ciblé, les ARN Cr seront retrouvés en complexes avec des protéines de type Cas représentées en bleu (contre de l'ADN) ou Cmr représentées en vert (contre de l'ARN). Les abréviations indiquées sont Cas pour CRISPR associated et Cmr pour Cas module RAMP.

par des portions de séquences DR (Hale et al., 2008). L'ARN Cr mature final correspondrait à un espaceur et une portion de séquence DR en 5' (**Figure 29**).

Chez les *Sulfolobus*, des motifs promoteurs ont été prédits dans la séquence leader de trois cassettes CRISPR. Ces motifs promoteurs seraient constitués des boîtes BRE et TATA (définies précédemment) localisées à une distance de 25 nt en amont du site +1 de transcription. Ce site +1 serait localisé entre 17 et 32 nt en amont de la première répétition des cassettes CRISPR (Lillestol et al., 2009).

Comme cité précédemment, un ensemble cœur de 6 gènes *cas* est généralement retrouvé dans les génomes contenant des cassettes CRISPR (Haft et al., 2005; Makarova et al., 2006). Seules quelques protéines codées par les gènes *cas* ont été caractérisées. Parmi les protéines Cas cœur, la Cas1 de *S. solfataricus* P2 est capable de se lier à de l'ADN et de l'ARN et seule la fonction de la Cas6 a été démontrée chez *P. furiosus*. Cas6 est l'endonucléase participant à la maturation de l'ARN Cr précurseur en ARN Cr : elle est capable de se lier et de couper la jonction hélice-boucle des structures en tige-boucle formées par les séquences DR palindromiques et de produire des ARN Cr pré-matures (2X et 1X cités précédemment) de 38-45 nt et de 43-46 nt en fonction de la taille de la séquence espaceur produite (**Figure 30**) (Carte et al., 2008). Une seconde et dernière étape de maturation serait de nature exonucléolytique afin de produire l'ARN Cr mature constitué de la séquence espaceur et d'une portion de 8 nt de la séquence DR en 5' (Carte et al., 2008; Hale et al., 2008).

A ce jour, deux fonctions biologiques sont attribuées aux systèmes CRISPR chez les archées : un système d'interférence à ARN CRISPR/Cmr chez *P. furiosus* (Hale et al., 2009) et un système d'interférence à ADN CRISPR/Cas chez les archées du phylum des Crenarchaeotes (**Figure 31**) (Gudbergsdottir et al., 2011; Shah et al., 2009). L'étude réalisée chez *P. furiosus* a montré que l'existence d'un système CRISPR constitué d'ARN Cr matures associés à une sous-famille de protéines Cas codées par un ensemble de six gènes *cmr* (pour Cas module RAMP et RAMP pour repeat-associated mysterious proteins). Ce complexe RNP serait capable de cibler des ARN synthétiques. Similairement aux ARN à interférence eucaryotes, l'ARN Cr servirait de guide de clivage : quelque soit la taille de l'ARN Cr mature, le clivage est toujours observé à une distance de 14 nt en amont de l'extrémité 3' de l'ARN Cr (Hale et al., 2008). En parallèle du système à interférence à l'ARN, le système d'interférence à ADN a été suggéré chez les archées acidothermophiles du phylum des Crenarchaeotes (Shah and Garrett, 2011; Shah et al., 2009). Les séquences espaceurs provenant des loci CRISPR de 12 souches ont été comparées aux séquences d'ADN

génomique de virus et de plasmides de Crenarchaeotes disponibles (cinq familles de virus, une famille de plasmide conjugatif et une famille de plasmide cryptique) (Shah et al., 2009). Près de 30% des séquences espaceurs correspondent à des régions d'ADN génomiques des éléments génétiques comparés. De plus, l'analyse de la distribution des séquences espaceurs révèlent que les deux brins d'ADN génomique étranger peuvent être ciblés suggérant que les futures séquences espaceurs intégrées seraient sélectionnées de façon aléatoire et ne proviendraient pas de la transcription inverse de transcrits viraux ou plasmidiques. En effet, certaines séquences espaceurs correspondent à des régions intergéniques des génomes viraux et plasmidiques suggérant que le mécanisme d'interférence ne cible pas des transcrits ARN mais directement des séquences d'ADN étranger (Shah et al., 2009). La coexistence des deux systèmes à interférence à ARN et à ADN a été observée chez les Sulfolobus (Gudbergsdottir et al., 2011; Shah and Garrett, 2011). Par ailleurs, l'étude effectuée par l'équipe de R. A. Garrett a renforcé l'hypothèse de l'existence de changements dynamiques de la séquence des loci CRISPR. En effet, lors de l'introduction de vecteurs portant les gènes correspondant à des séquences espaceurs (d'origine virale ou plasmidique), des évènements de délétion de ces séquences espaceurs sont observés (Gudbergsdottir et al., 2011).

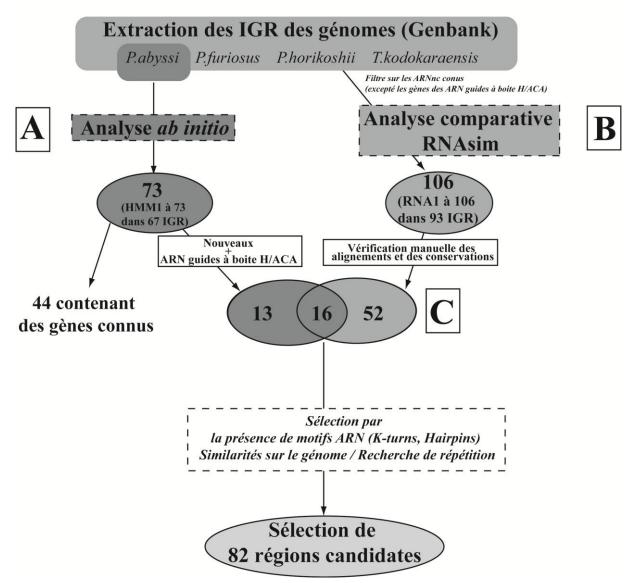
# III. Objectifs du projet de thèse

L'annotation du génome de *P. abyssi* et les différents travaux de recherche de familles d'ARNnc ont permis de prédire et/ou valider l'existence de gènes d'ARNnc comprenant les gènes des ARNr (deux ARNr 5S, ARNr 16S, ARNr 23S), le gène de l'ARN de la SRP, le gène de l'ARN de la RNase P, 46 gènes d'ARNt, 59 gènes d'ARN guides à boîtes C/D (Gaspin et al., 2000; Omer et al., 2000), 7 gènes d'ARN guides à boîtes H/ACA (Muller et al., 2008) et quatre cassettes CRISPR (Portillo and Gonzalez, 2009). Par ailleurs, les travaux de recherche d'ARNnc réalisés chez *P. furiosus* ont proposé que les génomes de Thermococcales contiendraient d'autres ARNnc (Klein et al., 2002). Ce travail de thèse a consisté à rechercher de nouveaux ARNnc dans le génome de *P. abyssi* par des outils bio-informatiques développés dans notre équipe qui utilisent le caractère AT riche du génome et l'analyse comparative entre trois autres génomes de Thermococcales (*P. furiosus*, *P. horikoshii* et *T. kodakaraensis*). L'expression en ARN de certaines des regions prédites a été validée expérimentalement par différentes techniques de biologie moléculaire. Sur la base de prédictions de modèles de structure secondaire et de l'étude des contextes génomiques, des fonctions biologiques ont pu être été suggérées pour les nouveaux ARNnc de *P. abyssi* identifiés.

# Résultats

Ce projet est né d'une collaboration étroite entre deux équipes utilisant des approches complémentaires pour appréhender le monde des ARN chez les archées dont l'une développe des méthodes et des outils bio-informatiques et l'autre utilise des approches biochimiques et de biologie moléculaire. Cette collaboration avait déjà permis d'identifier et de caractériser des ARN guides de modification à boite C/D dans le génome annoté et séquencé de *P. abyssi* (Clouet-d'Orval et al., 2005). Cette thèse a pour objectif d'identifier de nouveaux ARNnc chez l'archée hyperthermophile *P. abyssi* par une approche bio-informatique et de valider leur existence par une étude expérimentale.

Les résultats sont présentés en deux parties distinctes. La première correspond à la description des analyses bioinformatique, aux résultats obtenus et à la procédure de sélection des candidats. La deuxième décrit la validation expérimentale et la caractérisation de nouveaux ARNnc chez *P. abyssi*. La majeure partie de cette section « Travaux de recherche » fera prochainement l'objet d'une publication dont l'état actuel est en cours d'écriture.



**Figure 1: Approche bioinformatique par l'utilisation de deux méthodes de prédictions.** (A) Résultats obtenus par l'analyse *ab initio.* (B) Résultats obtenus par l'analyse comparative de génomes RNAsim. (C) Comparaison et sélection des résultats.

# I. Prédictions bio-informatiques

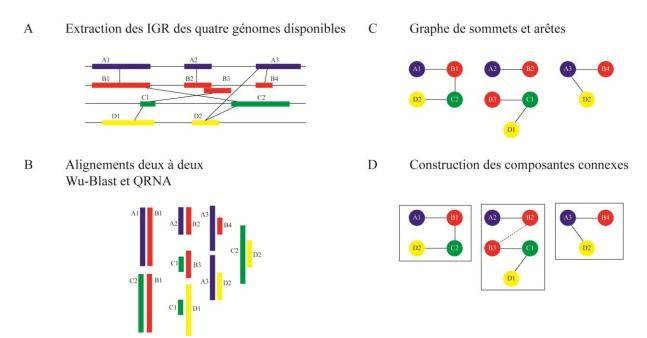
La recherche de nouveaux ARN non codant réalisée dans ce travail s'appuie sur une approche bioinformatique qui combine l'utilisation de deux méthodes de prédiction : l'analyse *ab initio* et l'analyse comparative de génome (**Figure 1**). Nous avons restreint notre analyse aux régions intergéniques (IGR) c'est-à-dire aux régions d'ADN double brin du génome ne comportant pas de phase ouverte de lecture ou ORF annotée.

Cette section « prédictions bio-informatiques » est divisée en trois parties : la première décrit la méthode *ab initio* utilisée, la deuxième concerne la description de l'approche comparative et enfin la troisième partie détaille les critères qui nous ont conduit à sélectionner les candidats pour la validation expérimentale.

# A. Méthode ab initio

## 1. Principe

Les méthodes *ab initio* permettent d'explorer les propriétés intrinsèques du génome étudié afin de prédire des candidats ARNnc spécifiques au génome analysé. Dans le cadre de cette étude, nous avons choisi d'exploiter le biais de composition en nucléotides existant entre les régions codantes et les ARNnc connus du génome. En effet, dans le génome de *P. abyssi*, comme dans la plupart des génomes d'archées AT-riches, une classe d'ARN (ARNt, ARNr...) montre un biais de composition significatif. La segmentation du génome en deux classes de régions a été réalisée à l'aide d'un modèle de Markov caché à deux états (HMM pour Hidden Markov Model). L'HMM utilisé est basé sur le même principe que celui utilisé chez *P. furiosus* (Klein et al., 2002). Deux états sont définis et représentent respectivement un état ARN (forte composition en GC) et un état AUTRE (faible composition en GC). Les probabilités d'émission pour l'état ARN ont été paramétrées sur le contenu en GC des ARNt et des ARNr et celles émises pour l'état AUTRE, sur le contenu en GC de la totalité du génome. Cependant, l'HMM appliqué dans notre étude diffère de celui utilisé pour *P. furiosus* (Klein et al., 2002) par les valeurs des probabilités de transition entre états, considérant un plus grand nombre attendu de candidats.



**Figure 2 : Principe de l'outil RNAsim.** (A) Les régions intergéniques sont extraites et alignées deux à deux par Wu-Blast. (B) QRNA sélectionne les régions de structures conservées. (C) Conversion des alignements en graphes. (D) Création de composantes connexes et fusion des sommets voisins sur une même séquence génomique (B2 et B3).

#### 2. Résultats

La segmentation réalisée dans le génome de *P. abyssi* a permis de proposer 73 régions de contenu en GC plus élevé que le reste du génome dans 67 IGR, annotées de HMM1 à HMM73 (**Figure 1A**). 51 des 73 régions prédites correspondent à des gènes d'ARNnc déjà annotés dans Genbank (NCBI). Elles comprennent les quatre gènes d'ARNr (5Sa, 5Sb, 16S et 23S), les 46 gènes d'ARNt, le gène de l'ARN composant de la RNase P et l'ARN SRP. Les 22 régions restantes ont été analysées à l'aide d'autres critères qui seront détaillés ci-après.

# B. L'analyse comparative de génomes : l'outil RNAsim

#### 1. Principe

L'autre approche que nous avons appliquée est basée sur l'utilisation de l'outil QRNA développé par (Rivas and Eddy, 2001). Cependant, QRNA possède un inconvénient majeur qui est l'utilisation limitée à deux génomes. En effet, plus le nombre de génomes comparé est élevé, plus forte sera la prédiction des candidats pour des ARNnc de séquences et structures conservées. L'outil RNAsim développé au sein de l'équipe a déjà permis de réaliser avec succès une analyse comparative sur plus de deux génomes chez la bactérie *Staphylococcus aureus* (Geissmann et al., 2009). RNAsim s'appuie sur une représentation par un graphe pour étendre la comparaison à plusieurs génomes et sur QRNA pour sélectionner les alignements mettant en évidence des régions potentiellement structurées.

RNAsim est constituée d'une première étape qui est l'extraction de toutes les IGR des génomes grâce aux coordonnées génomiques fournies par l'annotation de la base de données GenBank (Benson et al., 2009; Benson et al., 2008) (**Figure 2A**). Un filtre est utilisé permettant d'éliminer toutes les séquences annotées comme ARNnc connus et répertoriés dans Genbank. Ces IGR font l'objet d'une comparaison itérative par WU-Blast (Lopez et al., 2003) donnant des alignements deux à deux qui sont ensuite filtrés par QRNA (**Figures 2B**). Un graphe représentant l'ensemble des alignements QRNA est alors construit, où chaque sommet représente la séquence d'un alignement et chaque arête dessinée entre deux sommets signifie qu'un alignement existe entre deux séquences. Ce graphe est ensuite modifié afin de construire un graphe plus simple dans lequel des sommets ont été fusionnés lorsqu'il existe un chevauchement complet ou partiel entre les régions représentées par deux sommets différents



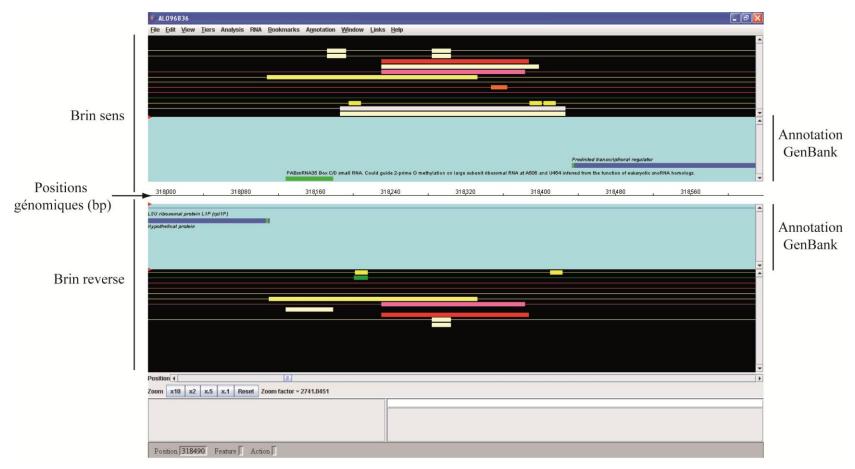
**Figure 3 : Sortie d'un résultat RNAsim.** (A) Tableau d'une composante connexe indiquant les organismes retenus, le dut et la fin de la région prédite (begin et end), la longueur des séquences, leur pourcentage en G et C (G+C%), les liens pour les fichiers Fasta et GFF. D'autres liens sont accessibles : le tableau des synthénies (Fichier des gènes voisins), l'alignement multiple proposé par MultAlin et la structure CARNAC. (B) et (C) Alignements multiples respectivement sélectionné (B) ou non (C) pour la validation expérimentale.

ou bien lorsque les deux séquences d'un alignement sont exactement consécutives (**Figure 2C**). Chaque composante connexe (ensemble de sommets tel que de chaque sommet on puisse aller à tout autre en parcourant les arêtes) de ce graphe représente alors un ensemble de séquences présentant entre elles des similarités en séquence et des structures consensus potentielles (**Figure 2D**).

Les résultats RNAsim se présentent sous la forme d'un tableau de composantes connexes, triées en fonction de la taille croissante des composantes. Les régions candidates de chaque composante connexe sont présentées avec les informations suivantes (**Figure 3A**): leurs coordonnées génomiques, leur longueur, leur contenu en GC (%), les séquences FASTA et GFF (General-Finding Format). Pour chaque composante connexe, RNAsim propose également un alignement multiple des éléments à l'aide de MultAlin qui permet de vérifier la similarité de séquences entre les régions conservées (Corpet, 1988) et un tri manuel permet de sélectionner les alignements multiples significatifs (**Figure 3B & 3C**).

## 2. Application au génome de P. abyssi et résultats

RNAsim peut donc s'utiliser dans le cadre de l'analyse comparative de plus de deux génomes. Lorsque nous avons débuté notre étude, trois autres génomes de Thermococcales étaient disponibles publiquement : les génomes de *P. furiosus* DMS3638, de *P. horikoshii* OT3 et de *T. kodakaraensis* KOD1. Les régions intergéniques de ces quatre génomes ont été extraites et les loci identifiés comme ARNnc connus ont été filtrés (ARNr, ARNt et ARN guides de modification à boîtes C/D). Il est à noter que lorsque nous avons débuté l'étude, les ARN guides à boîtes H/ACA n'étaient pas encore enregistrés dans GenBank. Les comparaisons par WU-Blast et QRNA ont été réalisées avec les paramètres par défaut. Les résultats sont disponibles sur la page web RNAsim disponible via le lien suivant : <a href="http://snp.toulouse.inra.fr/~gaspin/cgibin/RNAsim/">http://snp.toulouse.inra.fr/~gaspin/cgibin/RNAsim/</a> AC THERMOCOCCALES.html. Au total, 579 séquences ont été produites pour les quatre génomes comparés avec un total de 106 séquences pour *P. abyssi* nommées de RNA1 à RNA106 (Figure 1B). Pour les 106 prédictions proposées par l'analyse comparative, une sélection manuelle a été effectuée sur la base des alignements multiples donnés par MultAlin pour chaque composante connexe, conduisant à retenir 65 candidats.



**Figure 4 : ApolloRNA, environnement d'annotation de génomes.** Zoom dans une région génomique où l'on retrouve le gène de l'ARN guide de modification à boîtes H/ACA Pab35 : sont indiqués sur fond bleu, l'annotation du génome tirée de GenBank, sur fond noir, les données produites dans cette étude, le candidat HMM (en rose), le candidat RNAsim (en rouge), l'annotation de (KLEIN *et al.* 2002) (en blanc), <u>la prédiction d'</u>un motif K-turn (en orange), <u>la prédiction lesde</u> motifs promoteurs consensus (en vert) ou partiel de type TATA (en en jaune).

Candidat	Candidat	Annotation GenBank	Critères			
HMM	RNAsim					
Gènes codant pour les ARN guides de modification à boîtes H/ACA						
HMM7	RNA12	Pab19*	ΔG			
HMM11	RNA16	Pab21*/SnoR9°	$\Delta G$			
HMM12	RNA18	Pab160*/Pf9°	-			
HMM15	RNA19	Pab35*/HgcF°	KT, ΔG			
HMM16	RNA22	Pab40*/HgcG°	KT, ΔG			
HMM48	RNA50	Pab91*	KT, ΔG			
HMM58	RNA60	Pab105*/HgcE°	ΔG			
Loci CRISPR						
HMM4	RNA7	CDICDD 1	-			
-	RNA8	CRISPR1	-			
-	RNA25	CRISPR2	-			
HMM72	RNA74	CRISPR4	-			
-	RNA75	CKISFK4	-			
Candidats retrouvés par les deux approches						
HMM6	RNA11	-	-			
HMM23	RNA28	-	-			
HMM31	RNA34	-	-			
HMM34	RNA38	-	-			
HMM40	RNA45	-	-			
Autres prédictions retrouvés						
-	RNA4	PfQ13°	ΔG			
HMM29	RNA33	SscA/PfQ11°	ΔG			
HMM33	RNA37	Pf5°	ΔG, TB			
-	RNA56	PfQ6°	KT			
HMM61	-	Pf2°	_			
HMM45	-	PfQ12°	-			
-	RNA70	PfQ2°	_			

**Tableau 1 : Comparaison des prédictions obtenues par les deux méthodes.** Les résultats congruents sont regroupés. Pour plus de précisions, se référer au texte. Les critères indiqués sont KT (motif k-turn), TB (Tige-boucle) et  $\Delta G$  (énergie libre). Les gènes d'ARNnc annotés par \* ont été étudiés dans (MULLER *et al.* 2008) et par ° dans (KLEIN *et al.* 2002).

## C. Critères de sélection et combinaison des deux approches

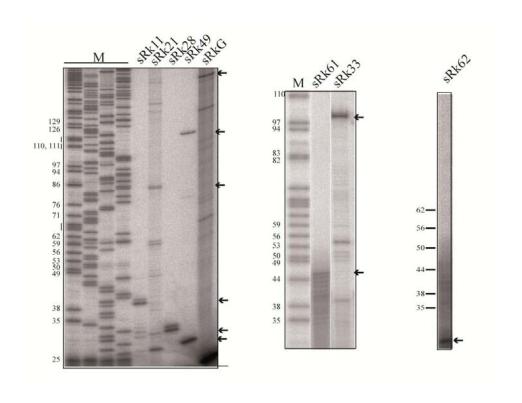
En complément des prédictions obtenues par les deux méthodes utilisées, nous avons recherché des critères caractérisant les ARNnc tels que les motifs K-turn/K-loop, la structuration possible de régions en tige-boucles et le critère thermodynamique. Ce dernier a été utilisé pour explorer la capacité du génome à former des structures stables en utilisant une fenêtre glissante le long du génome. Par ailleurs, nous avons recherché la présence de motifs promoteurs putatifs dans le génome de *P. abyssi*. Les séquences utilisées sont de type BRE/TATA défini par (Cohen et al., 2003) à l'aide de 19 promoteurs cartographiés chez *P. furiosus* (RAAANNTTWWWWA où R = G ou A, N = A, C, G ou T et W = A ou T). Nous avons aussi recherché des régions partielles composées uniquement de la boite TATA (TTWWWWAA). Tous ces résultats ont ensuite été visualisés à l'aide du logiciel ApolloRNA (**Figure 4**) qui est un environnement d'annotation permettant d'avoir une vision intégrée de l'ensemble des données produites pour sélectionner les candidats (**Figure 1C**) (Lewis et al., 2002).

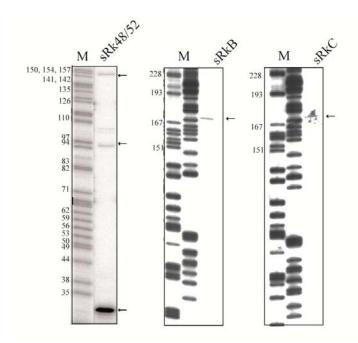
Par ailleurs, les 22 régions HMM ont été comparées aux 65 candidats RNAsim afin d'éliminer une possible redondance des prédictions retrouvées par les deux méthodes. La congruence des résultats est donnée dans le **Tableau 1**. Les sept ARN guides de modification à boîtes H/ACA identifiés par (Muller et al., 2008) ont été retrouvés par les deux méthodes. La prédiction de ces ARN valide l'approche bio-informatique utilisée. Il est à noter que seuls cinq de ces ARN avaient été prédits par (Klein et al., 2002). Ceci peut s'expliquer par la différence des probabilités de transition. Tous ces gènes présentent une énergie libre remarquable excepté pour Pab160 et des motifs de type Kink-turn sont prédits pour trois d'entre eux.

La comparaison des prédictions des deux méthodes montrent que deux candidats HMM4 et HMM72 et cinq candidats RNA7, 8, 25, 74 et 75 correspondent à trois loci CRISPR qui n'étaient pas encore annotés au début de notre étude. De plus, cinq autres régions prédites sont trouvées pas les deux approches. Nous avons aussi comparé nos prédictions à celles proposées chez *P. furiosus* mettant en évidence que six de nos prédictions seraient similaires à celles retrouvées par (Klein et al., 2002). Enfin, des candidats présentent des similarités en séquence avec d'autres régions retrouvées sur le génome de *P. abyssi*. Ces régions similaires que nous avons qualifiées de « répétées » ont aussi été ajoutées à nos candidats ARN. Les différents critères utilisés nous ont conduit à proposer 82 régions ayant le potentiel de générer des ARNnc (**Figure 1C**).

Nom	Prédiction	Prédiction Fin	Taille (nt)					Conservation			
	Début		Е	ES	S	5'-end	Promoteur	Gènes adjacents	Brin	chez les Thermococcales	Prédiction
					Lo	ci de trans	crits supérie	urs à 500 nt			
sRk 2	16318	16461	> 700	> 700	> 700	N. D.	-	PAB2341/PAB0014	$\leftarrow\leftarrow\rightarrow$	Pho, Pfu, Tko	R
sRk 21	355801	355883	> 700	> 700	> 700	355831	-	PAB2107/PAB2106	<b>←</b> ←←	Pho, Pfu, Tga	R
sRk 54	1198159	1198249	> 500	> 500	> 500	N. D.	-	PAB1560/PAB1559	←→←	Tous Excepté <i>Tko</i>	R
sRk G	1093067	1093124	310/ 420	420	420	1093281	-	PAB1612/PAB0757	$\leftarrow\leftarrow\rightarrow$	Tga, Ton	Н
sRk62	1397186	1397290	> 500	> 500	> 500	1397237	-	PAB0959/PAB0960	$\rightarrow \rightarrow \rightarrow$	Tous Excepté <i>Tga</i>	R
					Lo	oci de trans	scrits inférieu	ırs à 500 nt			
					(1)	ARN guid	es de modificat	ion à boîtes H/ACA			
Pab19	204858	205038	130/ 190	190	130/ 190	N. D.	Cons (79 nt)	PAB0133/PABr02	$\rightarrow \rightarrow \rightarrow$	Tous	H, R
Pab21	230516	230634	60/ 118	60/ 118	60/ 118	N. D.	TATA	PABsnRNA19/PABsnRN A10	$\leftarrow \leftarrow \rightarrow$	Pho, Pfu, Tko	H, R
Pab35	318187	318426	160	160	160	N. D.	TATA	PABsnRNA35/PAB0211	$\rightarrow \rightarrow \rightarrow$	Tous	H, R
Pab40	382342	382606	200	200	200	N. D.	TATA	PAB0262/PAB0263	$\rightarrow \rightarrow \rightarrow$	Tous	H, R
Pab91	1074502	1074589	60	60	60	N. D.	Cons	PAB1625/PAB1624	$\leftarrow\leftarrow\leftarrow$	Pho, Pfu, Tko	H, R
Pab105	1335653	1335801	160	160	160	N. D.	Cons	PAB0917/PAB2426	$\longleftrightarrow \leftarrow$	Pho, Pfu	H, R
Pab160	289023	289153	55/ 65	55/ 65	55/ 65	N. D.	-	PAB0191/PAB0191.1n	$\rightarrow \longleftarrow \rightarrow$	Pho, Pfu	H, R
			ı	ı		Г	(2) Loci CRIS	SPR		<del></del>	ı
Cr 1-1 CRISPR1	149430	147916	60	60	60	149408	Cons (53nt)	PAB0095/PABt02	<b>→</b> ←←	Tous (répétitions)	H, R
Cr 4-4 CRISPR4	1760062	760062 1761854		60	60	1760285	- Cons (53nt)	PAB1170/PABt46	$\longleftrightarrow \to$		H, R
Cr 4-12 CRISPR4				60	60	1760832					H, R
(3) Loci unique											
sRk 11	197248	197534	50	50	50	197484	TATA	PAB2227/PAB2402	$\leftarrow\leftarrow\leftarrow$	Tous	H, R
sRk 28	527697	527833	70/ 100	70/ 100	-	527798	-	PAB1992/PAB1991	<b>←</b> ←←	Tous excepté <i>Tsi</i>	H, R
sRk 33	636804	636904	100	100	100	636804	Cons	PAB1916/PAB0465	$\longleftrightarrow \to$	Tous	H, R
sRk 61	1348633	1348700	60/ 95	60/ 95	60/ 95	1348615	-	PAB1455/PAB0921	$\longleftrightarrow \to$	Pho, Tsi	R
			1	T	1	T	(4) Loci répé	ités		T	1
sRk B	809887	810256	220/ 350	220/ 290/ 350	-	810048	Cons	PAB1794/PAB0571	$\longleftrightarrow \to$	-	Н
sRk C	1612986	1613347	145/ 220	145/ 220	145	1613195	Cons	PAB1080.4n/PAB1080.5n	<b>←</b> ←←	-	Н
sRk48	985849	986002	190	190/ 130	190	985805	Cons	PAB0686/PAB0686.1n	$\rightarrow \rightarrow \rightarrow$	Pho, Pfu, Tko, Tsi	H, R
sRk 49	1067535	1067886	150/ 220	150/ 220	-	1067923	TATA (60nt)	PAB0740/PAB0741	$\rightarrow \longleftarrow \rightarrow$	Pho	R
sRk 52	1104008	1104286	190	190/ 130	190	1104056	Cons	PABt30/PAB0766	$\longleftrightarrow \to$	Pho, Pfu, Tko, Tsi	H, R

**Tableau 2 : Caractéristiques des loci exprimant les ARNnc détectés par Northern Blot.** Les candidats sont classés en fonction de leur taille (supérieure '>' ou inférieure '<' à 500 nt). Quatre sous-groupes sont définis pour les ARN inférieurs à 500nt : les ARN guides de modification à boîtes H/ACA, les loci CRISPR, les loci uniques et les loci répétés. Pour chaque locus étudié, les positions indiquées sont celles prédites par les outils bio-informatiques, les tailles correspondent aux résultats obtenus en Northern Blot, l'extrémité 5' a été déterminée par des analyses par des extensions d'amorces et/ou par Circular-RACE, les motifs promoteurs sont indiqués lorsqu'ils ont été prédits entre 15 et 30 nt en amont de l'extrémité 5' (ou à défaut, la distance est indiquée), les gènes adjacents aux loci sont donnés ainsi que leur orientation, la conservation en séquence chez les autres Thermococcales est indiquée avec les annotations suivantes : *P. horikoshii (Pho)*, *P. furiosus (Pfu)*, *T. kodakaraensis (Tko)*, *T. gammatolerans (Tga)*, *T. sibiricus (Tsi)*, les outils de prédictions sont représentées par H pour HMM (approche *ab initio*) et R pour RNAsim (analyse comparative).





**Figure 5 : Extension d'amorce des transcrits détectés.** Analyse des produits d'extension d'amorce obtenus pour les ARN étudiés. Les réactions ont été réalisées sur les préparations d'ARN totaux extraits de cellules cultivées en entrée en phase stationnaire, puis migrées sur un gel 6% PAGE dénaturant. Les oligonucléotides utilisés sont indiqués dans l'annexe 2 et les positions des arrêts de rétro-transcription obtenus sont indiquées dans le Tableau 2.

## II. Détection et caractérisation de nouveaux ARN chez P. abyssi

# A. Expression de 24 régions prédites par l'approche bioinformatique

Les résultats bio-informatiques obtenus nous ont amené à sélectionner 82 loci candidats. Afin de valider expérimentalement leur existence, nous avons tout d'abord choisi d'analyser la transcription de chaque locus par une analyse en Northern Blot.

Des préparations d'ARN totaux extraits de cellules cultivées en phase exponentielle (E), en entrée en phase stationnaire (ES) et en phase stationnaire (S) ont été utilisées. Des oligonucléotides complémentaires aux brins direct et indirect ont été définis selon la séquence de chaque région candidate. Leur taille est comprise entre 20 et 25 nucléotides pour obtenir une température de fusion (Tm ou température de « melting ») définie entre 50 et 65 °C afin de pouvoir réaliser toutes les hybridations à une température fixée à 50°C. La séquence de chaque oligonucléotide a fait l'objet d'un crible sur le génome de *P. abyssi* à l'aide du logiciel Vector NTI® (version7.0, Invitrogen) afin de vérifier qu'elle soit bien spécifique à la région ciblée. Ainsi nous avons pu déterminer des oligonucléotides spécifiques à chacune des séquences. Certaines régions candidates ont des régions similaires répétées sur le génome. Dans ces cas précis, au moins deux couples d'oligonucléotides spécifiques (complémentaires aux deux brins) ont été déterminés afin de valider le résultat observé en Northern Blot. Les oligonucléotides dessinés pour cette étude sont répertoriés dans le tableau de l'annexe 2. Au total, nous avons effectué plus de 220 analyses par Northern Blot.

Vingt-quatre loci se sont avérés être transcrits en ARN stables, ces résultats sont regroupés dans le **Tableau 2**. Pour chaque candidat transcrit, les positions indiquées correspondent aux régions proposées par les prédictions bio-informatiques. Les tailles données sont celles des ARN observés sur les membranes de Northern Blot par rapport à la taille du marqueur de taille. Pour chaque candidat, des extensions d'amorce ont été réalisées avec les oligonucléotides ayant permis de détecter des signaux de transcription en Northern Blot. Les résultats de ces extensions d'amorces sont indiqués dans les figures correspondant à chaque candidat analysé ci-après. Pour chaque région étudiée, les séquences promotrices proposées proviennent de la recherche *in silico* de motifs entiers (Cons : BRE/TATA) ou partiels (TATA) localisés en amont des séquences transcrites. Elles sont indiquées lorsque

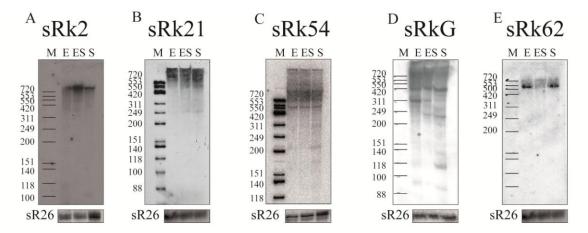


Figure 6 : Northern Blot des transcrits de taille supérieure à 500 nt : (A) sRk2, (B) sRk21, (C) sRk54, (D) sRkG, (E) sRk62. L'analyse a été réalisée avec les ARN totaux de *P. abyssi* extraits de cellules en phase exponentielle (E), en entrée en phase stationnaire (ES) et en phase stationnaire de croissance (S). Le marqueur de taille (M) (ΦX174 digéré par Hinf1) est radiomarqué en 5' au γ-ATP [32]P. Pour chaque membrane, une hybridation témoin a été effectuée à l'aide d'une sonde dirigée contre l'ARN guide de modification à boite C/D sR26

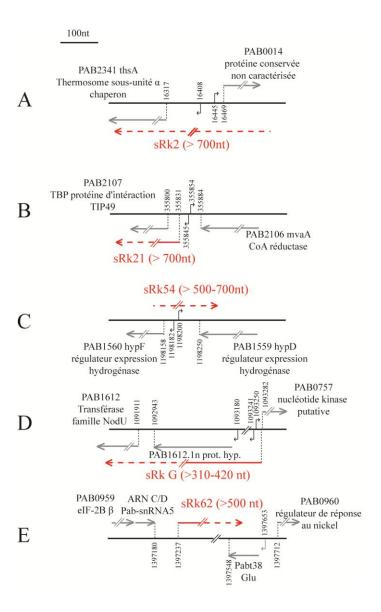


Figure 7: Localisation génomique des loci transcrits en ARN de taille supérieure à 500 nt. (A) sRk2, (B) sRk21, (C) sRk54, (D) sRkG, (E) sRk62. Le locus transcrit est représenté par une flèche rouge en pointillés. Les ORF annotées dans GenBank sont indiquées et représentées par des flèches grises. Les éléments promoteurs prédits sont identifiables par un crochet fléché de couleur noire pour les motifs consensus BRE/TATA et grise pour les motifs partiels TATA. La représentation respecte l'échelle du génome.

leur localisation est comprise entre 15 et 30 bp en amont de l'extrémité 5' déterminée (sauf pour certains cas où la distance est indiquée).

Sur la base de la taille des ARNnc bactériens qui sont compris de 40 à 500 nt (Huttenhofer et al., 2002), nous avons définis deux groupes en fonction de la taille des transcrits observés sur les membranes de Northern Blot (**Tableau 2**). Le premier groupe de taille « supérieure à 500 nt » est composé de cinq loci. Le second groupe « inférieure à 500 nt » est décomposé en quatre classes : (i) les ARN guides de modification à boite H/ACA, (ii) les transcrits provenant de loci CRISPR, (iii) les transcrits provenant de régions « uniques » et non annotées et (iv) les transcrits provenant de régions répétées sur le génome (**Tableau 2**).

La caractérisation des loci exprimés, présentée avec l'étude de leur expression, a permis d'identifier les gènes adjacents ainsi que la synthénie observée entre les six autres génomes de Thermococcales séquencées : *P. horikoshii*, *P. furiosus*, *T. kodakaraensis*, *T. sibiricus*, *T. gammatolerans* et *T. onnurineus*. La conservation des séquences et l'analyse des covariations ont permis de proposer un modèle de structure secondaire pour certains ARNnc candidats. L'analyse des contextes génomiques et les modèles de structure secondaire nous ont fourni quelques éléments permettant de proposer une fonction pour les ARN détectés par analogie avec les types de régulation exercés par des ARNnc chez les bactéries qui sont décrites dans la littérature.

# B. Les transcrits supérieurs à 500 nt

Nous retrouvons dans ce groupe les loci sRk2, sRk21, sRk54 et sRkG et sRk62. Ils ont été prédits par l'outil RNAsim à l'exception de sRkG qui a été proposé par la méthode *ab initio* (**Tableau 2**).

sRk2 et sRk21 sont des transcrits de taille supérieure à 700 nt. Leur expression est observée dans les préparations d'ARN totaux extraits de cellules dans les trois phases de croissance (**Figures 6A et 6B**). L'annotation du génome montre que leur localisation génomique respective chevauche les ORF PAB2341 et PAB2107 (**Figures 7A et 7B**). Cette observation suggère que les transcrits observés pourraient correspondre aux ARNm de ces ORF dont les tailles minimales prédites sont de 1653 nt et 1326 nt respectivement.

sRk54 est un transcrit de taille supérieure à 500 nt (**Figure 6C**). Son locus est en antisens d'un opéron annoté constitué d'ORF codant des régulateurs d'expression

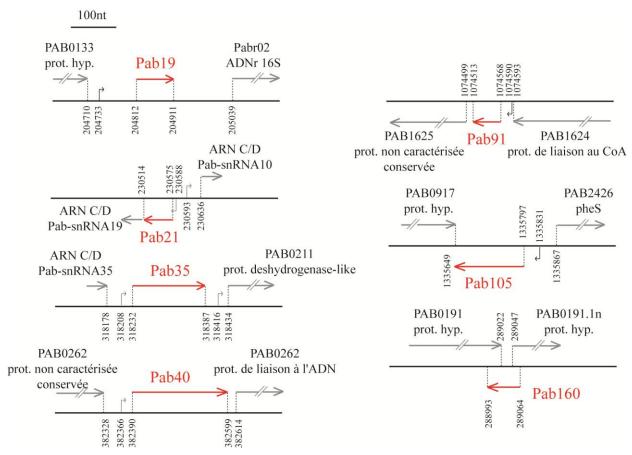


Figure 8 : Localisations génomiques des gènes des sept ARN guides de modification à boite H/ACA. Les annotations sont décrites dans la Figure 6.

	BRE TATA
Cons	RAAANNTTWWWWA
Pab21	TTCCCTTCTTCACCGATAACGTTATAACCCATTTTTAAATTTAATGTGAT
Pab35	CATAGCTCATTTGGTAATGATTTTTAAATCCTTAACCGTAAATATGAATG
Pab40	TGCTTCTTCCATAGTCACTA <mark>TTATAAA</mark> ACTTCGGGTCTTATTCTCATATG <u>'</u>
Pab91	TAGGCTCGAAAGGGAAAGAGTTTAAAATCATAATCTCAAGGTAATAATG
Pab105	GGATGTCAACTTTAAAAATCCTTTTGGGTAAGATTTAATTGTCCGACCTG
Pab160	TTTTATTGATGGTATAAAAG <mark>TTTTTAG</mark> CCTCGAAGTAAATAAGATATATG

Figure 9 : Alignement des séquences de 50 nt situées en amont des extrémités 5' de chaque ARN guide de modification à boite H/ACA. Alignement avec le motif consensus BRE en noir et TATA en rouge où R = G ou A, N = A, C, G ou T et W = A ou T.

d'hydrogénases (**Figure 7C**). Nous ne disposons d'aucune donnée supplémentaire pour proposer une hypothèse de fonction de ce transcrit. Cependant, il est possible que nous ayons détecté un transcrit d'une nouvelle ORF non annotée.

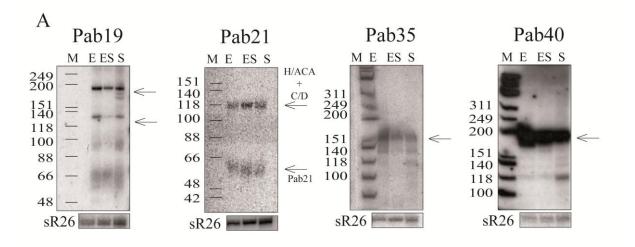
Pour sRkG, plusieurs transcrits de taille variable sont observés sur Northern Blot avec une abondance différente en fonction de la phase de croissance : un transcrit de 310 nt est abondant uniquement en phase exponentielle et un autre de 420 nt est présent dans les trois phases de croissance (**Figure 6D**). L'annotation du génome a été corrigée récemment et indique la présence de l'ORF PAB1612.1n dont la taille serait de 204 bp (**Figure 7D**). Il est alors possible que nous ayons détecté un ARNm putatif issu de cette ORF de taille supérieure à celle qui a été ré-annotée.

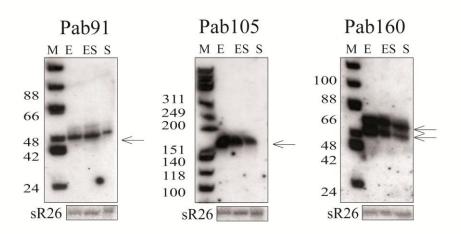
sRk62 est transcrit en un ARN d'environ 500 nt (**Figure 6E**). Ce locus est situé dans une IGR de 584 bp où sont retrouvés les gènes codant pour un ARNt et d'un ARN guide de modification à boîtes C/D (**Figure 7E**). Une extension d'amorce a été réalisée pour ce locus permettant d'identifier le nucléotide +1 de ce transcrit qui n'inclut pas les séquences des ARN annotés en amont (**Figure 5**). Aucun élément promoteur n'est identifiable dans les séquences situées en amont. Nous avons effectué une recherche de similarité de séquence via l'outil Blast en ne se limitant qu'à la séquence prédite de ce locus (**Annexe 3**) et de façon surprenante, cette recherche a abouti à l'identification de 3 autres répétitions chez *P. abyssi*, 4 chez *T. kodakaraensis*, 15 chez *T. sibiricus*, 2 chez *P. furiosus* et 2 chez *T. onnurineus*. Par ailleurs, les contextes génomiques de ces répétitions semblent souvent associer ce locus répété avec une ORF codant une transposase de la famille IS605. Cette association a déjà été notée chez *S. solfataricus* mais aucune caractérisation fonctionnelle n'a été faite (Wurtzel et al., 2010; Zago et al., 2005).

## C. Les transcrits inférieurs à 500 nt

## 1. Les ARN guides de modification à boites H/ACA

Sept régions retrouvées par nos deux outils bio-informatiques correspondent aux gènes codant pour les ARN guides de modification à boîtes H/ACA. Ces ARN ont été décrits dans (Muller et al., 2008). La recherche de motifs que nous avons réalisée a permis de proposer des séquences promotrices putatives situées à une distance de 13 à 34 nt en amont de l'extrémité





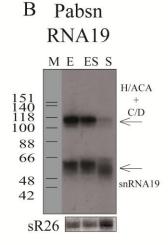


Figure 10 : Northern Blot (A) des transcrits des gènes des sept ARN guides de modification à boite H/ACA et (B) de l'ARN guide de modification à boite C/D adjacent à Pab21. Les annotations sont décrites dans la Figure 6.

5' annotée pour cinq de ces ARN (**Figure 9**) à l'exception de Pab19 dont le motif consensus est prédit à 79 nt en amont de l'extrémité 5' et de Pab160. Aucun promoteur putatif n'est proposé pour Pab160 (**Figure 9**). Nous avons alors aligné les séquences de 50 nt situées en amont de l'extrémité 5' définie de ces ARN par (Muller et al., 2008) et nous proposons la présence d'un motif promoteur partiel TATA dégénéré (TTWWWWA) qui serait retrouvé à 23 nt en amont du +1 de Pab160 (**Figure 10**).

La détection de ces ARN par Northern Blot est illustrée dans la Figure 8A. Pour quatre d'entre eux, une seule espèce sans intermédiaire est détectée alors que deux espèces différentes sont présentes pour Pab19, Pab21 et Pab160. Ces analyses en Northern Blot sont comparables aux résultats de (Muller et al., 2008) excepté pour Pab19 pour lequel la validation par Northern Blot n'était pas publiée. Pour celui-ci, nous observons deux transcrits d'environ 130 nt et 180 nt (Figure 8A). Le plus grand semble être plus abondant. Ce résultat suggère que Pab19 pourrait résulter de la maturation d'un ARN précurseur afin de donner la forme H/ACA mature. La recherche de motifs promoteurs putatifs a permis d'identifier une séquence de type consensus BRE/TATA (Figure 9). Cependant, ce promoteur putatif s'avère être assez éloigné (- 79 nt) de l'extrémité 5' de l'ARN proposée par (Muller et al., 2008). Par ailleurs, le locus de Pab19 est situé en amont du gène de l'ARNr 16S pour lequel aucun motif promoteur n'est prédit. Cette configuration semblerait suggérer qu'un même et unique précurseur ARN incluant l'ARN H/ACA et pré-ARN 16S serait transcrit à partir du promoteur prédit et que des événements de maturation successifs généreraient l'ARN H/ACA et l'ARNr pré-16S. Afin de vérifier cette hypothèse, il serait nécessaire de réaliser une extension d'amorce du pré-ARN 16S.

Deux transcrits d'environ 60 nt et 118 nt sont observés pour Pab21 (**Figure 8A**). Ce résultat est identique à celui obtenu par (Muller et al., 2008). Le contexte génomique indique que Pab21 est adjacent au locus codant un ARN guide de modification à boîtes C/D, PabsnRNA19 (**Figure 9**). Ces deux ARN seraient co-transcrits en un ARN précurseur de 118 nt qui subirait des processus de maturation afin de donner les deux ARN matures. Ici, l'espèce à 60 nt correspondrait à Pab21 (**Figure 8A**). Nous avons effectué une autre analyse par Northern Blot à l'aide d'un oligonucléotide complémentaire à l'ARN PabsnRNA19 (**Figure 8B**). Deux espèces sont également observées, l'une de 118 nt, taille similaire au résultat précédent et l'autre d'environ 55 nt qui correspondrait à l'ARN guide à boite C/D. Il est à noter qu'en fonction de l'oligonucléotide utilisé (complémentaire à l'ARN H/ACA ou au C/D), l'expression du transcrit précurseur montre un signal d'une intensité différente en

```
CRISPR1 149430-147916 brin(-)
GTTCCAATAAGACTAAAATAGAATTGAAAG TTATATCTTCTTCGGGGAAGGAAAAGCCCATGTATCG
GTTCCAATAAGACTAAAATAGAATTGAAAG AAAGAGTAGCTTGAGAGTTCTGCTTCCACTAGCAAAA
GTTCCAATAAGACTAAAATAGAATTGAAAG TTGAGAATAGTTGAAAAGCTCCACACGGAGCTTGTTG
GTTCCAATAAGACTAAAATAGAATTGAAAG CCTCTTTTTAGCGTTAATGCTGCAAGGATTGCTGCAA
GTTCCAATAAGACTAAAATAGAATTGAAAG TCTGCAGGGAGCTTAAGAACAAACGCCGATGAAGTGT
GTTCCAATAAGACTAAAATAGAATTGAAAG GAAGACCACACTGGAGTAGGTCTCTTGAAGAGTTCT
GTTCCAATAAGACTAAAATAGAATTGAAAG AGGGCTTGGAGCTGAGAGCCACAAGCCCTCAACCCTCTTTCAA
GTTCCAATAAGACTAAAATAGAATTGAAAG TAAACCGTGCTCTCAATACGCATCCCCCCTAAAAAA
GTTCCAATAAGACTAAAATTGAAAG AATGATGCTGTGTGCTTCGAGCTTAACGTGGAACGTC
GTTCCAATAAGACTAAAATAGAATTGAAAG GATGAGGAAGGCCGGGGTCGAGCAGTGGGCAATAAAC
GTTCCAATAAGACTAAAATAGAATTGAAAG TATGACGATGCGGTCTGGGTATATCTCGAGGACTACC
GTTCCAATAAGACTAAAATAGAATTGAAAG TAATCCTCTATTTCTCTTATGCTCATTCTTTCAAGTT
GTTCCAATAAGACTAAAATAGAATTGAAAG GGGGCTCCGCCGGGCTACCCCAGCCGCCCGGGCCCGT
GTTCCAATAAGACTAAAATAGAATTGAAAG GCAACTTATCTCTTATTTGAGCCGGATCGGTAACCGG
GTTCCAATAAGACTAAAATAGAATTGAAAG CGGCCCTGAGTCGAGGGATTTTTCGCTTTTCTAACCC
GTTCCAATAAGACTAAAATAGAATTGAAAG CTGCATGCTCGTATAGAACATGGAGAATAAACTTAGG
GTTCCAATAAGACTAAAATAGAATTGAAAG GAGTTTAAGCCTCATTGGAAGAAACCGATCAGGAAAA
GTTCCAATAAGACTAAAATAGAATTGAAAG CGGTTCTGAAGCCATAGACGGCTATTCCGGCCCTACTGGA
GT-CCAATAAGACTAAAATAGAATTGAAAG AGTGTTATAGTGTAGTTTGGGTATGTGCTTCCGAAT
GTTCCAATAAGACTAAAAATAGAATTGAAAG GCAAAAGCGCGAGGTTCTGAACACTGTTCTTCAGAGC
GTTCCAATAAGACTAGAATAGAATTGAAAG CGAGGACGAATGAGCTCGCAACGGCGAGTGCACCAACCG
CRISPR2 482830-482332 brin (-)
GTTTCCGTAGAACTTAGTAGTGGGAAAG TCGATTGAACTGATTGCGTCAATGAACCCATGCACTTTCTCCTG
GTTTCCGTAGAACCTAGTGTGGAAAG ACAGATACTAACGGAATTACAAGAGCAGCAGGGTGGTTT
GTTTCCGTAGAACTTAGTAGTGGGAAAG TGTGCCCTGTGATTTTAGCTATGAGTTGGGGAGAAACTCCT
GTTTCCGTAGAACTTAGTAGTGTGGAAAG GGGTTCTTCCTGGGGTCAATCCTGTCTAACGTGTAGGCCCTGA
GTTTCCGTAGAACTTAGTAGTGTGGAAAG ACCTCCTCTCTCACCCACTCAACCCA
GTTTCCGTAGAACT AATAGTGTGGAAAG CCGGCGGCGTGGTAATTCAGAGCGACATATTCGTAGT
GTTTCCGTAGAACT AATAGTAT GATTG CTTGATGTTAGGTTCGGAAGGAAACCAGCGAAATCGATGAGGAAG
CRIPR3 1090355-1090058 brin (-)
GTTCCAATAAGACTAAAATAGAATTGAAAG GCGTTTATTCGCGGGGCGGATCCGAGGAAAGTGGAGCTGAGGAGAGTGTTTA/
                              /CTAGGAGTTGCTGGTGGTGATAATGGAGGAGTTAAGGTTTACAGCCTCTCAACCTG
GTTCCAATAAGACTAAAATAGAATTGAAAC AAGCCAAAGAAAAGTTCAATCTGTAGGAAATGCGCGT
GTTCCAATAAGACTAAAATAGAATTGAAA G TAACAAACTAATTTGTAATAAAAAAAGAAGGCATCAAATT
CRISP4 1760062 1761985 brin (+)
GTTCCAATAAGACTAAAATAGAATTGAAAG TCTATCTCGGAAAGCAGTAATCTTGCTAGAGTAGTCT
GTTCCAATAAGACTAAAATAGAATTGAAAG CTAACACAATTGTCTGAAAGTTACTGTAAATTCCACG
GTTCCAATAAGACTAAAATAGAATTGAAAG GCCGAAAGGCCGGCGCTGAGGATCTGTTTCATCGGTT
GTTCCAATAAGACTAAAATAGAATTGAAAG CGGGAGTGTTGAGACGGTGAGGATATCACGGGTGGACACTGGGAAT
GTTCCAATAAGACTAAAATAGAATTGAAAG GGAACCCCTAGGGGGTCTAAGGCGCCGGGGCCAACGC
GTTCCAATAAGACTAAAATAGAATTGAAAG CCCGTCTGATCTACGTAGAATCTGCGTGTTGCCCCTGT
GTTCCAATAAGACTAAAATAGAATTGAAAG CGTTGGGCTTCACCTCAACCGGCTTCACGTCTGTTGGA
GTTCCAATAAGACTAAAATAGAATTGAAAG ACATTTTTTTGCAAAAAATTTTTCAAAAAAATCGGCCTAA
GTTCCAATAAGACTAAAATAGAATTGAAAG TGAGCCTTCATCAACTTCATCAACTTCAAAT
GTTCCAATAAGACTAAAATAGAATTGAAAG TAATTGTTACATGGTGATGGCGTCTTGGCGAAATGGA
GTTCCAATAAGACTAAAATAGAATTGAAAG CCACTTAAATTCCGCCGAAGCTATTTCGAATATCTT
GTTCCAATAAGACTAAAATAGAATTGAAAG GAAATCATTGAGGGACTCGTCGGCCCAGAGGTCGGAA
GTTCCAATAAGACTAAAATAGAATTGAAAG TGGCGAGCTTTCCCCTGCCGGCCCTCGCTATGAAGAG
GTTCCAATAAGACTAAAATAGAATTGAAAG TCAACTTACGCGCAATACCATCAAGAACACTATTCTCC
GTTCCAATAAGACTAAAATAGAATTGAAAG CTATACCGAGTTCATATAAGAGGTTCACTAAGAAAGCAATAA
GTTCCAATAAGACTAAAATAGAATTGAAAG TTGCTCCCACATTGAGGAAGTTAGTTGGGAGTGCGA
GTTCCAATAAGACTAAAATAGAATTGAAAG TCGATCCGCCCGGGCTGGTATATCTTCGCCGACTTGA
GTTCCAATAAGACTAAAATAGAATTGAAAG CCGCAGTAGCCAGCTCTTGTCCGCCGTTGCCCCCTAT
GTTCCAATAAGACTAAAATAGAATTGAAAG GCAGTTCCAATTGCGAGTAAAGGCCCTCCGTTTCTTC
GTTCCAATAAGACTAAAATAGAATTGAAAG TTTTCTGATATAAACGATGTAGTGTTCGCTTAGGCTT
GTTCCAATAAGACTAAAATAGAATTGAAAG ACTGGCTCGATCGAAGGCTTAATCGTTGCAGATGCAT
GTTCCAATAAGACTAAAATAGAATTGAAAG CCTCATTCCAGTCCGTTCCCGAGAGGTTCGTCAGGGGCAG
GTTCCAATAAGACTAAAATAGAATTGAAAG GAGTTCCCGCTGACCCTCCTCGCCTGCACCAGTAACGT
GTTCCAATAAGACTAAAATGAAATGAAAG GAAAATTAAAAGGCGGTTCATCGGCAACTTTGCACAT
GTTCCAATAAGACTAAAATAGAATTGAAAG TGTTGATTTTGAGCTTGAGAATATCAAGGATGCGCTCGG
GTTCCAATAAGACTAAAATAGAATTGAAAG CTCATCAAGTATGGAACAAACATACTTGCCTTTGGA
GTTCCAATAAGACTAAAATAGAATTGAAAA TAGTGATTTATTTAAGGACGTTTGTAGCTAA
GTTCCGACAAAACTAAAGAAAGAACAGAAG TAGCGGGCCCGGCGGGATTTGAACCCGCGACCTTCG
                             GGCCACGGGCCCACAAATTCTTTAAAAATACTGCTGGTAATAAAGCTTGCG
```

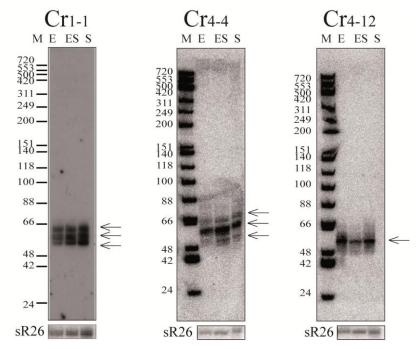
Figure 11 : Alignement des séquences DR de chaque cassette CRISPR de *P. abyssi*. Pour chaque cassette, les positions sont indiquées ainsi que le brin sur lequel elles sont retrouvées. Les séquences DR sont en gras et les séquences espaceurs en police normale. Les séquences DR dégénérées sont en bleu et les nucléotides non conservés sont surlignés en gris. La séquence en orange correspond aux nucléotides en aval de chaque cassette.

fonction de la phase de croissance (**Figure 8A et 8B**). Ces résultats montreraient que ces deux ARN impliqués dans les modifications des ARNr seraient co-transcrits sous la forme d'un précurseur bicistronique, voire co-régulés. Un résultat similaire avait aussi été observé pour les orthologues de ces ARN chez *P. furiosus* : un même oligonucléotide permettait de détecter la transcription des deux ARN homologues à ceux de *P. abyssi* (Klein et al., 2002).

## 2. Les ARN issus des loci CRISPR

Parmi les quatre loci annotés dans Genbank, trois (CRIPSR 1, 2 et 4) sont détectés par nos outils de prédiction (**Tableau 1**). La détection de ces loci par l'outil RNAsim proviendrait de la conservation des séquences palindromiques répétées en tandem présentes dans les cassettes CRISPR (Godde and Bickerton, 2006; Jansen et al., 2002a; Jansen et al., 2002b).

L'annotation des cassettes CRISPR chez P. abyssi proposée par (Portillo and Gonzalez, 2009) diffère de celle donnée dans GenBank. Nous proposons alors une autre annotation basée sur une caractéristique répertoriée des cassettes CRISPR : les dernières séquences DR en 3' d'une cassette CRISPR sont généralement dégénérées (Horvath et al., 2008; Jansen et al., 2002a; Jansen et al., 2002b). Nous avons réalisé un alignement de toutes les séquences DR pour chaque cassette CRISPR (Figure 11). Les CRISPR1 et 4 sont respectivement composés de 23 et 29 séquences DR de motif commun de 30 nt (GTTCCAATAAGACTAAAATAGAATTGAAAG). Quatre séquences DR dégénérées sont observées pour le CRISPR1 et trois pour le CRISPR4. Le CRISPR3 n'est constitué que de quatre séquences DR de motif similaire à celles des CRISPR1 et 4 dont deux sont dégénérées. Le CRISPR2 est composé de huit séquences DR d'un motif différent des autres cassettes de 29 nt (GTTTCCGTAGAACTTAGTAGTGTGGAAAG) dont deux sont dégénérées. Par l'identification de ces séquences dégénérées, nous proposons que les cassettes CRISPR1, 2 et 3 seraient situées sur le brin antisens du génome et la cassette CRISPR4 serait sur le brin sens. Des analyses complémentaires en Northern Blot ont été effectuées pour les cassettes 1 et 4, à l'aide d'oligonucléotides dirigés contre les espaceurs situés entre les séquences DR dégénérées. Aucun transcrit n'est détecté comme c'est le cas chez P. furiosus pour lequel aucun signal de transcription n'avait été observé à partir des séquences espaceurs situées dans les parties 3' des cassettes CRISPR (Hale et al., 2008). Ce résultat conforte l'orientation de ces cassettes CRISPR1 et 4 que nous proposons.



**Figure 12 : Northern Blot des transcrits CRISPR.** Résultats obtenus à l'aide d'oligonucléotides complémentaires aux séquences espaceurs 1 du CRISPR1 (Cr1-1), 4 et 12 du CRISPR4 (Cr4-4 et CR4-12 respectivement). Les annotations sont décrites dans la Figure 6.

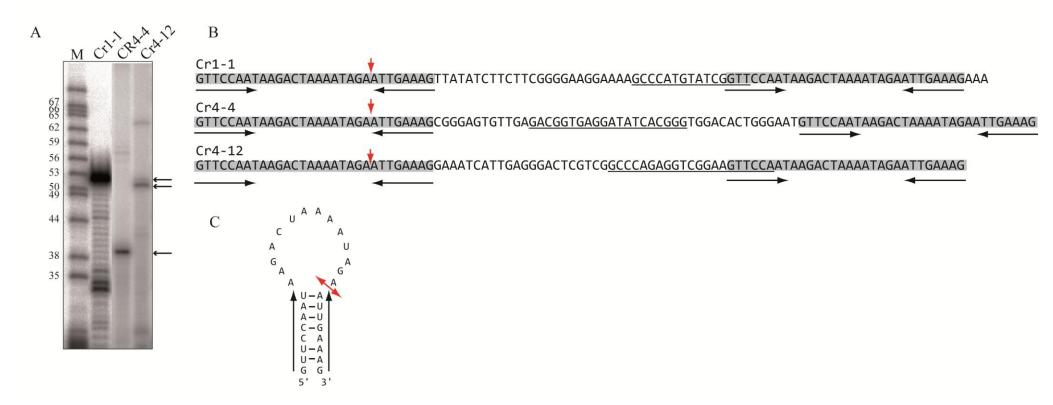


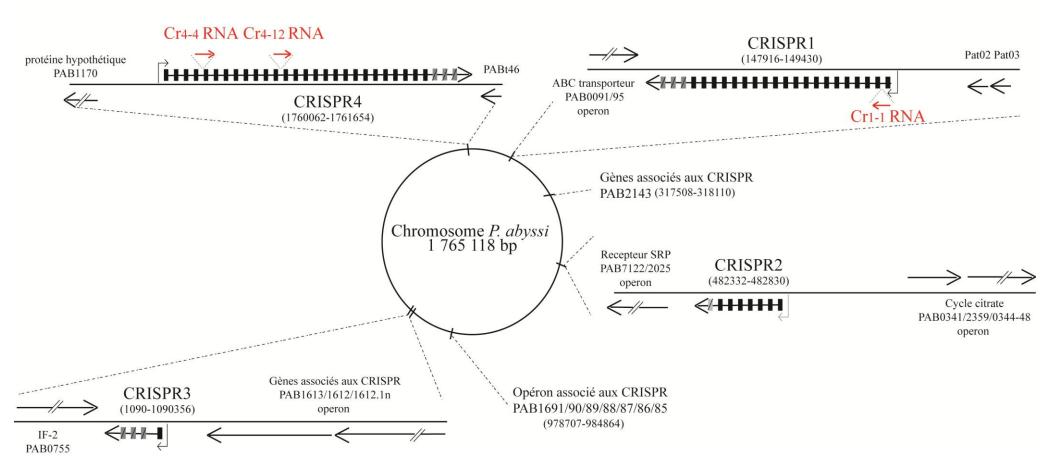
Figure 13: Extensions d'amorce des transcrits CRISPR. (A) Analyse sur gel dénaturant 6%PAGE, les arrêts observés sont indiqués par des flèches. (B) Cartographie des extrémités 5' observées, les oligonucléotides utilisés sont soulignés et l'extrémité 5' est indiquée par une flèche rouge. Les séquences DR sont surlignées en gris et les séquences palindromiques sont indiquées par des flèches inversées. (C) Modèle de structure secondaire proposée. Les séquences palindromiques sont indiquées par des flèches noires et le site de clivage par l'endoribonucléase Cas6 chez *P. furiosus* est indiqué en rouge (CARTE *et al.* 2008).

La détection de transcrits provenant de ces loci par des analyses en Northern Blot faites à l'aide de sondes dirigées contre les séquences espaceurs situées à proximité de la séquence leader ont permis l'observation de transcrits correspondant aux espaceurs 1 du CRISPR 1 noté Cr1-1 et aux espaceurs 4 et 12 du CRISPR 4 notés Cr4-4 et Cr4-12 (Figure 12). Aucun signal de transcription n'est détecté pour l'espaceur 2 du CRISPR2. Le résultat de Northern Blot obtenu pour Cr1-1 et Cr4-4 montre la détection de plusieurs intermédiaires dont la taille varie entre 50 et 66 nt suggérant que ces ARN seraient issus de processus de maturation tel qu'il a été proposé chez P. furiosus par (Hale et al., 2008) et chez les Sulfolobales par (Lillestol et al., 2009). Ces transcrits d'environ 60 nt ont la taille attendue pour des transcrits issus de cassettes CRISPR (Hale et al., 2008; Tang et al., 2002a; Tang et al., 2005). Ils apparaissent stables quelque soit la phase de croissance de P. abyssi (Figure 12). Par ailleurs, le génome de *P. abyssi* possède une autre cassette CRISPR annotée CRISPR 3. Nous avons aussi analysé l'expression de ce locus par Northern Blot à l'aide d'un oligonucléotide dirigé contre son espaceur 2. Ce locus ne présente aucun signe visible de transcription. En conclusion, parmi les quatre loci CRISPR annotés chez P. abyssi, nous avons montré l'expression en ARN Cr de trois séquences espaceurs provenant des CRISPR 1 et 4.

Nous avons déterminé l'extrémité 5' de chacun des transcrits Cr par des expériences d'extension d'amorce (Figure 13). Ces extrémités correspondent au nucléotide adjacent à la séquence palindromique de la séquence DR (Figures 13A et 13B) qui a été identifiée comme site de clivage de l'endoribonucléase Cas6 chez P. furiosus caractérisée récemment (Figure 13C) (Carte et al., 2008). Un seul arrêt de transcription inverse est obtenu pour Cr1-1 et Cr4-4 alors que des signaux multiples sont observés par Northern Blot (Figure 12). Cette hétérogénéité de taille pourrait provenir d'événements de maturation partiels ou incomplets à leur extrémité 3'. Dans nos expériences, nous n'avons pas observé de transcription provenant des brins antisens contrairement à ce qui a été observé chez S. acidocaldaricus (Lillestol et al., 2009). Ces résultats montrent bien que les cassettes CRISPR 1 et 4 produisent des ARN Cr d'environ 60 nt. Cependant, aucun précurseur n'est observé par analyse en Northern Blot comme c'est le cas chez A. fulgidus entre autres (Tang et al., 2002a). Afin de vérifier la présence de précurseurs d'ARN Cr des CRISPR 1 et 4, nous avons effectué des extensions d'amorce avec des oligonucléotides de 20 nt complémentaires aux dix premiers nucléotides de la première séquence DR et aux dix nucléotides situés en amont de celle-ci (Figure 14B). Un arrêt de transcription inverse est observé à 35 nt en amont de la première DR (Figure



Figure 14: Caractéristiques des cassettes CRISPR chez les Thermococcales. (A) Extension d'amorce réalisée sur les ARN Cr précurseurs (B) Cartographie des d'extrémités 5' des ARN précurseurs (indiquées par une flèche et nucléotide +1 indiqué en rouge). (C) Alignement des séquences situées en amont de la première séquence DR étendue à aux cassettes CRISPR de *P. abyssi* (PabCRISPR1, 2, 3 et 4), *P. furiosus* (Pf.1 à 8), *P. horikoshii* (Ph.1 à 4) et *T. kodakaraensis* (Tk.1 à 3). Les séquences DR sont surlignées en gris, les motifs consensus BRE/TATA et le motif (AAAAA) conservé sont indiqués. Les séquences AT riches de PabCRISPR2 sont indiquées en rouge.



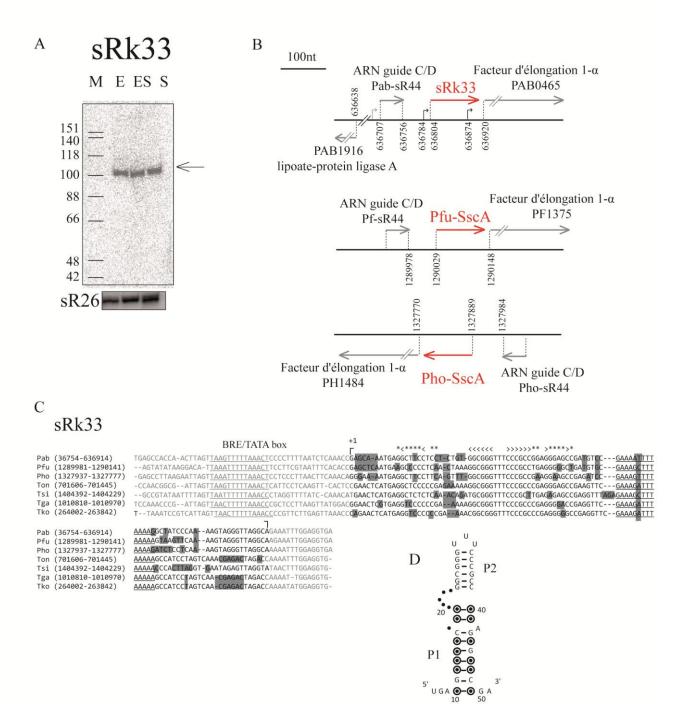
**Figure 15 : Contexte génomique des loci CRISPR et des gènes associés.** Sont indiqués sur le chromosome circulaire de *P. abyssi* : les orientations des cassettes CRISPR et les annotations des opérons et des gènes *cas* et *cmr* associés au système CRISPR. Les séquences DR sont illustrées par des boites colorées en noir et les séquences DR dégénérées sont en gris foncé. Les transcrits Cr que nous avons détectés sont indiqués en rouge pour les loci correspondants (Cr1-1, Cr4-4 et Cr4-12). La représentation est respectée et proportionnelle à l'échelle du génome.

**14A**). Il semblerait alors qu'un transcrit précurseur subirait un ou plusieurs processus de maturation pour générer des ARN Cr comme cela est suggéré chez *P. furiosus* (Carte et al., 2008).

L'alignement des séquences situées en amont de la première séquence DR de chaque cassette permet de mettre en évidence un motif (AAAAA) excepté pour le CRISPR3. Lorsque l'on étend cet alignement à 18 autres loci CRISPR de Thermococcales (Figure 14), on observe que ce motif est conservé. Celui-ci est retrouvé dans la séquence des ARN Cr précurseurs et il pourrait correspondre à un site de liaison à une protéine du système CRISPR et/ou à un site de maturation. Par la recherche de motifs promoteurs putatifs, nous proposons un motif de consensus BRE/TATA situé 10 nt en amont de l'extrémité 5' des précurseurs Cr déterminés (Figures 14A et 14B). Cet alignement étendu à d'autres cassettes CRISPR montre qu'un motif promoteur est observé pour la totalité des séquences alignées excepté pour le locus CRISPR2 (Figure 14C). Bien qu'aucun motif promoteur ne soit prédit pour celui-ci, trois motifs riches en nucléotides A et T sont observés.

Chez P. abyssi, le CRISPR 3 diffère des autres loci en trois points : (i) l'absence du motif (AAAAA) (**Figure 14C**), (ii) la taille de la cassette plutôt courte en comparaison aux trois autres et (iii) aucun signal de transcription n'est détecté. Ces observations suggèrent que le CRISPR3 pourrait être le reliquat d'une ancienne cassette CRISPR et peut être mis en parallèle aux changements évolutifs rapides et dynamiques des CRISPR due aux expositions à des parasites tels que les phages (Barrangou et al., 2007; Deveau et al., 2010).

Contrairement à la majorité des cassettes CRISPR bactériens ou d'archées (Jansen et al., 2002a), on ne retrouve aucun gène *cas* à proximité des loci chez *P. abyssi* à l'exception de l'ORF PAB1613 annotée comme homologue de l'endonucléase Cas6 situé en amont du CRISPR 3 (**Figure 15**). Cependant, l'annotation du génome révèle l'existence d'un opéron qui semblerait coder pour des protéines du système cas chez *P. abyssi* (**Figure 15**) bien qu'il soit complètement éloigné des localisations génomiques des loci CRISPR. L'expression de transcrits Cr suggère qu'il existerait un système dynamique de défense orchestrée par les CRISPR 1 et 4 quelque soit la phase de croissance de la cellule. De plus, la recherche d'homologie des espaceurs à l'aide de l'outil Blast n'a pas aboutit à la détection de séquence d'éléments génétiques connus tel que le virus PAV1, unique virus de *P. abyssi* séquencé à ce jour.



**Figure 16 : Caractéristiques du locus sRk33.** (A) Northern Blot. Les annotations sont décrites dans la Figure 6. (B) Contexte génomique chez *P. abyssi* (sRk33), *P. furiosus* (Pfu-SscA) et *P. horikoshii* (Pho-SscA). Les annotations sont décrites dans la Figure 7. (C) Alignement des séquences orthologues à sRk33 retrouvées dans les génomes des sept Thermococcales. La séquence de sRk33 conservée est représentée en noir comprenant les nucléotides variant surlignés en gris, les nucléotides adjacents au locus sont en caractères gris. Le motif promoteur BRE/TATA proposé est souligné et il est indiqué sur la séquence, les nucléotides formant les structures hélicales dans la structure secondaire proposée sont indiqués par '<' ou '>' et les changements compensatoires sont notés par '\*'. (D) Structure secondaire proposée sur la base de l'alignement de séquences. Les deux structures hélicales sont annotées P1 et P2. Les nucléotides variants dans l'alignement sont représentés par un point et les changements compensatoires sont représentés par un point dans un cercle.

## 3. Les ARN issus de loci uniques

Quatre loci ne présentant aucune autre séquence similaire dans le génome de *P. abyssi* expriment des transcrits dont deux d'entre eux montrent une expression différentielle en fonction de la phase de croissance (sRk11 et sRk28). Par ailleurs, nous avons identifié des éléments promoteurs putatifs dans deux des quatre loci par rapport à l'extrémité 5' estimée par extension d'amorce effectuée sur les préparations d'ARN totaux (**Tableau 2 et Figure 5**).

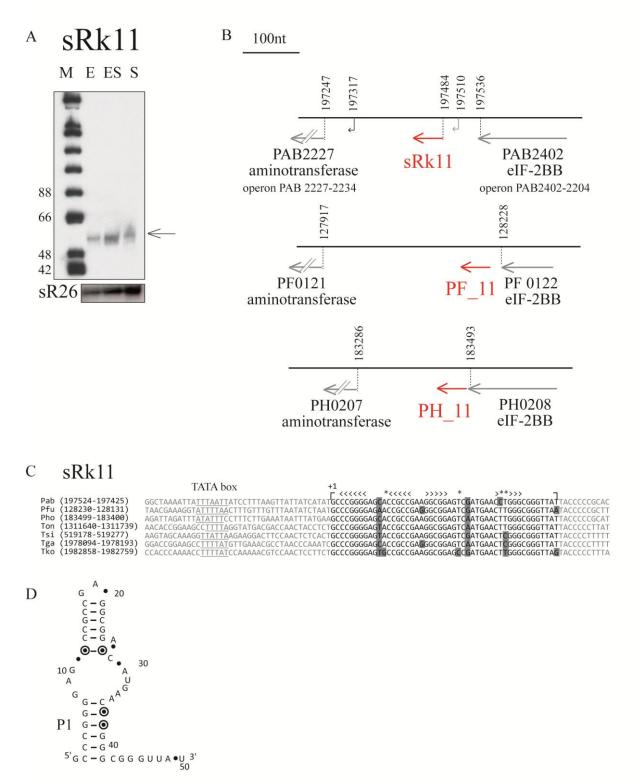
#### sRk33

sRk33 est transcrit en un ARN d'environ 100 nt exprimé de manière stable dans les trois phases de croissance (**Figure 16A**). Son extrémité 5' a été identifiée par extension d'amorce (**Figure 5**). Un motif promoteur de consensus BRE/TATA est proposé 19 nt en amont du nucléotide +1 (**Figures 5 et 16B**). L'orthologue de cet ARN a été précédemment identifié et validé chez *P. furiosus* (Klein et al., 2002). Dans la base de données Rfam (version 10.1, 2010 Sanger institute), il est annoté SscA pour « secondary structure conserved ». sRk33/SscA semblerait être uniquement retrouvé dans les génomes de Thermococcales où il est toujours codé en aval d'un gène codant pour un ARN guide à boîtes C/D et en amont d'un opéron de deux ORF annotées respectivement pour le facteur d'élongation de la traduction EF1α et la protéine ribosomale S10P (rps10p), tous deux jouant un rôle dans la liaison de l'ARNt au ribosome (**Figure 16B**).

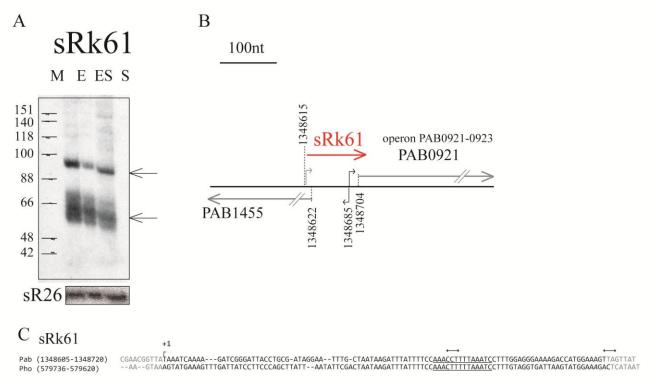
La structure secondaire en 5' proposée se composerait de deux hélices fortement conservées séparées par une boucle asymétrique. Sur la base de l'alignement des séquences avec les orthologues retrouvés chez les Thermococcales, nous avons pu identifier 13 changements compensatoires qui renforcent le modèle proposé (**Figures 16C et 16D**). La séquence primaire ne permet pas de proposer une structure secondaire de la région 3' par la présence de stretchs de nucléotides A et U (**Figure 16C**). Nous pouvons observer que cette partie 3' non structurée contient un motif promoteur putatif de consensus BRE/TATA qui pourrait correspondre au motif promoteur du gène PAB0465.

## sRk11

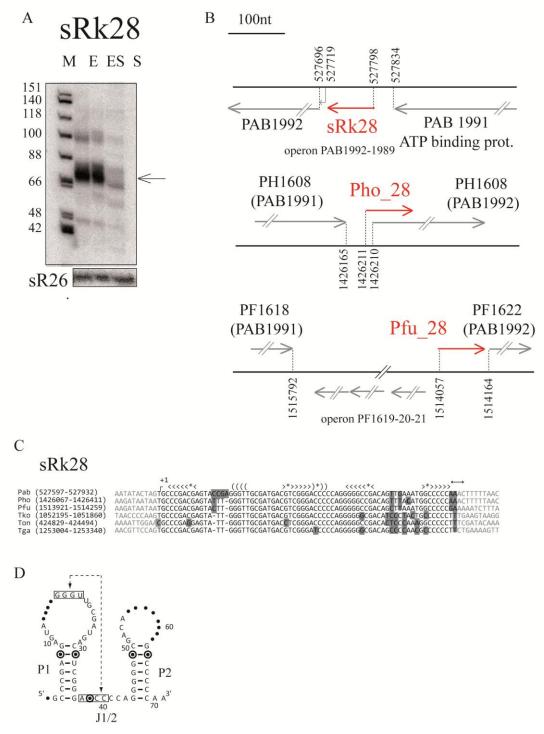
sRk11 est transcrit à partir du brin inverse d'une région intergénique de 289 bp. L'analyse par Northern Blot a permis la détection d'un transcrit stable de 50 nt qui semble plus abondant dans la préparation d'ARN totaux de cellules en entrée en phase stationnaire



**Figure 17 : Caractéristiques du locus sRk11.** (A) Northern Blot. (B) Contexte génomique chez *P. abyssi* (sRk11), *P. furiosus* (PF\_11) et *P. horikoshii* (PH\_11). (C) Alignement des séquences orthologues retrouvées chez les six autres Thermococcales. (D) Structure secondaire proposée sur la base de l'alignement de séquences. Les différentes annotations sont décrites dans la figure 16.



**Figure 18 : Caractéristiques du locus sRk61.** (A) Northern Blot. (B) Contexte génomique chez *P. abyssi* (C) Alignement avec la séquence orthologue retrouvée chez P. *horikoshii*. Les différentes annotations sont décrites dans la figure 16.



**Figure 19 : Caractéristiques du locus sRk28.** (A) Northern Blot. (B) Contexte génomique chez *P. abyssi* (sRk28), *P. furiosus* (Pfu\_28) et *P. horikoshii* (Pho\_11). (C) Alignement des séquences orthologues retrouvées chez cinq autres Thermococcales. (D) Structure secondaire proposée sur la base de l'alignement de séquences. Les différentes annotations sont décrites dans la figure 16.

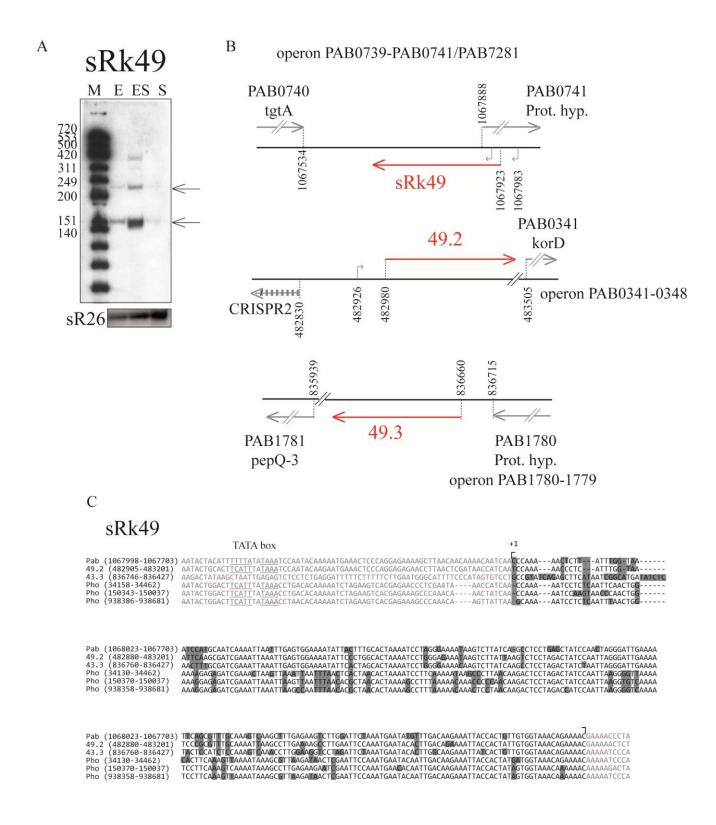
(**Figure 17A**). Il est à noter que son locus génomique, composé d'ORF annotées pour une aminotransférase et un facteur de traduction, semble être conservé chez *P. horikoshii* et *P. furiosus* (**Figure 17B**). La détermination de son extrémité 5' a été réalisée et une séquence correspondant au motif promoteur partiel TATA est proposée à 24 nt en amont du nucléotide +1 (**Figure 5 et Figure 17C**). Des orthologues de cet ARN ont pu être identifiés par une recherche Blast dans tous les génomes de Thermococcales à l'exception de *T. gammatolerans* (**Figure 17C**). La structure secondaire proposée est composée d'une boucle asymétrique séparant deux hélices fortement conservées contenant une mutation compensatoire (**Figure 17D**).

#### sRk61

L'analyse de l'expression de sRk61 par Northern Blot à l'aide d'un seul oligonucléotide permet de distinguer deux transcrits distincts d'environ 60 et 90 nt (**Figure 18A**). Cette observation suggère que ces ARN feraient l'objet de processus de maturation ou seraient sujet à une dégradation. Ce locus est situé dans une région intergénique en amont d'un opéron prédit annoté pour coder des protéines non caractérisées conservées chez les archées (**Figure 18B**). La recherche d'éléments promoteurs a permis de proposer un motif TATA qui semble cependant trop proche de l'extrémité 5' déterminée par extension d'amorce. Un autre motif TATA est retrouvé dans le locus de sRk61 et pourrait correspondre à l'élément promoteur de l'ORF PAB0921. De par sa séquence riche en nucléotides A et T et de l'unique conservation de ce locus chez *P. furiosus* qui ne permet pas de faire une analyse de covariation précise (**Figure 18C**), il nous a été difficile de proposer une structure secondaire stable pour les transcrits de ce locus.

## sRk28

sRk28 est transcrit en un ARN d'environ 70 nt très abondant en phase exponentielle et en début de phase stationnaire et quasi-absent en phase stationnaire (**Figure 19A**). Une espèce d'environ 100 nt de faible intensité est aussi visible suggérant la présence d'un transcrit précurseur (**Figure 19A**). Le locus génomique de sRk28 est retrouvé dans une IGR située entre deux gènes d'un même opéron (**Figure 19B**). Une seule extrémité 5' a été identifiée par extension d'amorce suggérant que les deux transcrits détectés par Northern Blot seraient différents en leurs extrémités 3' (**Figure 5**). La recherche d'un promoteur putatif en amont du nucléotide +1 n'a pas permis de mettre en évidence un motif semblable au consensus BRE/TATA ou motif TATA partiel. Cette absence de motif promoteur peut signifier que



**Figure 20 : Caractéristiques du locus sRk49.** (A) Northern Blot. (B) Contexte génomique des trois répétitions retrouvées chez *P. abyssi.* (C) Alignement des séquences orthologues retrouvées *P. abyssi.* (Pab, 49.2 et 49.3) et *P. horikoshii.* (Pho). Les différentes annotations sont décrites dans la figure 16.

sRk28 serait inclus dans l'opéron polycistronique ou alors que l'expression de ce locus serait contrôlée par des éléments promoteurs nouveaux et inconnus.

L'alignement des séquences orthologues retrouvées chez six Thermococcales dont *P. abyssi* (**Figure 19C**) permet de proposer une structure secondaire conservée composée de deux tige-boucles P1 et P2 (**Figure 19D**) où l'on trouve cinq mutations compensatoires. On peut aussi noter la présence de nucléotides pouvant former un pseudo-nœud : nous proposons que la séquence conservée GGGU de la tige P1 pourrait s'apparier au motif AYCC (Y = C ou U) retrouvé dans la séquence linéaire reliant les deux modules P1 et P2 (**Figure 18D**).

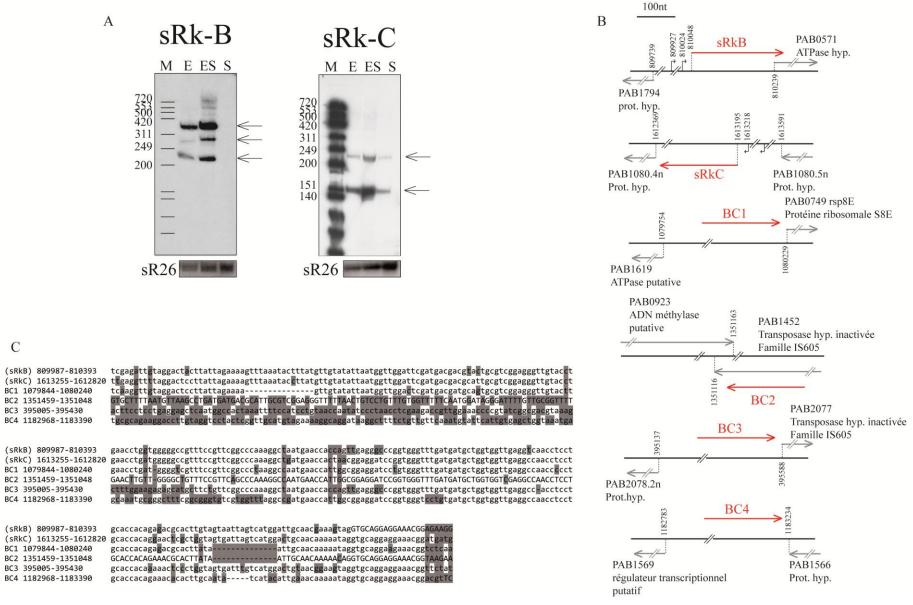
Lorsque le locus génomique de sRk28 et de ses gènes adjacents est comparé aux régions synthéniques chez les autres Thermococcales, il apparait que chez *P. furiosus*, cet opéron putatif serait interrompu par une insertion de trois ORF sur le brin inverse et que la séquence orthologue de sRk28 serait chevauchante à la partie 5' de l'ORF orthologue de PAB1992 (PH1608) (**Figure 19B**). Une recherche de similarité de séquence par Blastx suggère que PAB1992 possèderait un domaine QueC qui est retrouvé dans la protéine QueC de bactéries du genre *Bacillus*, connue pour son rôle dans la conversion du GTP en queuosine, nucléotide modifié retrouvé dans les ARNt des bactéries. De plus, l'expression de QueC semblerait être régulée par des éléments ARN codés en *cis* (Roth et al., 2007).

## 4. Les ARN issus de loci répétés

Nous avons classé dans ce groupe les transcrits des loci sRk49, sRkB/sRkC, sRk48/sRk52, candidats pour lesquels ont été repérés des séquences similaires dans le génome soit parmi les autres candidats prédits ou soit par recherche avec Blast contre le génome de *P. abyssi*. Ils présentent tous une expression plus abondante dans les cellules en entrée en phase stationnaire.

### sRk49

L'outil RNAsim a proposé une région candidate que nous avons nommée sRk49. La recherche de similarités par Blast a permis de détecter deux autres loci de séquences similaires (**Figures 20A et 20B**). Trois loci existent donc pour cette région, que nous avons respectivement appelés 49.1 (correspondant à la prédiction), 49.2 et 49.3 (**Figure 20B**). L'alignement des trois séquences a permis d'identifier une région variable où on observe une insertion de 6 nt pour 49.3 (**Figure 20C**). Nous avons alors dessiné un oligonucléotide spécifique à 49.3 et complémentaire à cette région pour effectuer une analyse en Northern



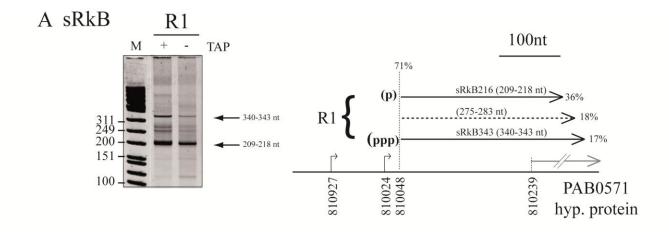
**Figure 21 : Caractéristiques des loci sRkB et sRkC.** (A) Northern Blot. (B) Contexte génomique des six répétitions retrouvées chez *P. abyssi.* (C) Alignement des séquences répétées. Les différentes annotations sont décrites dans la figure 16.

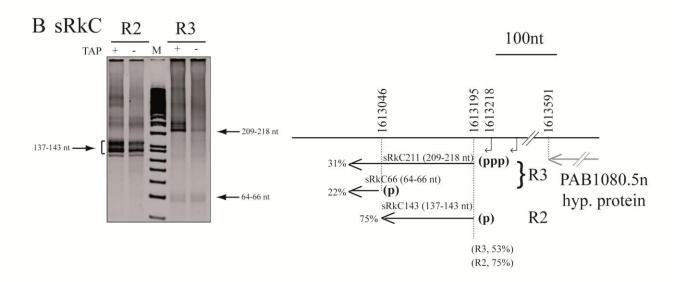
Blot. Aucun signal de transcription n'a été détecté suggérant que ce transcrit ne serait pas exprimé. Par ailleurs, les loci 49.1 et 49.2 ne présentent pas une variabilité en séquence importante pour dessiner un oligonucléotide spécifique pour chacun d'entre. L'analyse par Northern Blot a été effectuée à l'aide d'un oligonucléotide qui ne permet donc pas de discerner si de l'un et/ou l'autre des deux loci sont exprimés. Deux transcrits d'environ 150 nt et 220 nt sont détectés et semblent plus abondants en entrée en phase stationnaire (**Figure 20A**). L'extension d'amorce réalisée révèle deux extrémités 5' (**Tableau 1**). Ceci suppose que les ARN observés en Northern Blot pourraient résulter de processus de maturation. Le séquençage par C-RACE (voir ci-après) des transcrits détectés est en cours d'analyse et devrait permettre de distinguer l'expression des deux loci. Nous avons recherché des similarités de séquences dans les génomes des autres Thermococcales. Nous avons pu mettre en évidence que trois répétitions similaires étaient retrouvées dans le génome de *P. horikoshii* (**Figure 20C**). L'analyse des contextes génomiques chez *P. horikoshii* révèle une configuration semblable au locus 49.2 où les loci seraient à proximité de cassettes CRISPR (environ à 150 nt en amont).

## sRkB et sRkC

Les loci sRkB et sRkC sont très similaires au niveau de leurs séquences nucléotidiques. Ils sont transcrits en ARN stables de taille variable allant de 145 à 344 nt, exprimés en phase exponentielle et en entrée en phase stationnaire de croissance et totalement absents en phase stationnaire (**Figure 21A**). La recherche de similarités par l'outil Blast révèle que *P. abyssi* possèderait quatre autres loci similaires à sRkB et sRkC (**Figures 21B et 21C**). Les loci BC1, BC3 et BC4 sont retrouvés dans des régions intergéniques contrairement à BC2 qui chevauche la partie 3' d'une ORF annotée pour transposase de la famille IS605 (**Figure 21B**). Aucun signal de transcription des autres loci n'est détecté par Northern blot. L'observation de ces différentes espèces de transcrits issus des loci *sRkB* et *sRkC* suggère la présence d'un grand précurseur ARN qui ferait l'objet de processus de maturation ou de dégradation.

sRkB est situé dans sur le brin sens d'une IGR de 500 bp dont les ORF adjacentes de part et d'autre sont PAB1794 et PAB0571. La première ORF coderait pour une protéine hypothétique conservée chez les archées et la seconde, coderait pour une ATPase hypothétique possédant un domaine conservé de fonction inconnue (**Figure 21B**). sRkC est quant à lui localisé sur le brin inverse d'une IGR étendue sur 1223 bp flanquée par deux ORF







**Figure 22 : Analyse en C-RACE.** (A) Caractéristiques schématiques des transcrits sRkB (A) et sRkC (B). (C) Cartographie des transcrits observés pour les deux loci. La nature des extrémités est indiquée par (p) pour 5' monophosphate et (ppp) pour 5' triphosphate. Le % des clones séquencés de l'ARN correspondant sont indiqués. Les différentes annotations sont décrites dans la figure 16.

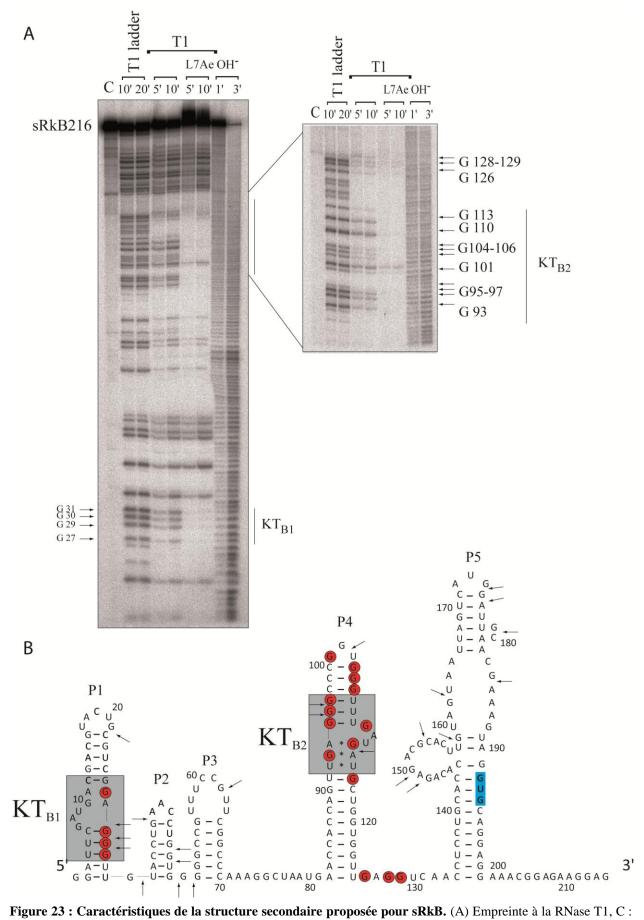
PAB1080.4n et PAB1080.5n situées toutes deux sur le même brin que sRkC et codant pour des protéines hypothétiques (**Figure 21B**).

Afin de mieux caractériser les transcrits issus des loci sRkB et sRkC, (Figure 21A), nous avons défini les extrémités 5' et 3' des différents transcrits en les séquençant par la technique d'amplification rapide d'extrémités d'ADNc à partir de matrice ARN circulaires (C-RACE pour Circular-Rapid Amplification of cDNA ends) (Annexe 5). Cette technique permet également de distinguer les transcrits primaires dont l'extrémité est de nature 5'triphosphate des transcrits matures (5'-monophosphate) grâce à un traitement ou non des ARN par l'enzyme Tobacco acid pyrophosphatase (TAP). Cette enzyme possède la propriété de convertir l'extrémité 5' triphosphate d'un ARN primaire en 5'-monophosphate (Bensing et al., 1996). Cette conversion permet ensuite de ligaturer l'extrémité 5' monophosphate à l'extrémité 3'-OH grâce à l'activité de la T4 RNA ligase, produisant des ARN circularisés. Deux pools d'ARN totaux circularisés traités (+TAP) ou non (-TAP) servent alors de matrices pour une amplification par RT-PCR. Les ADNc obtenus à partir des deux pools d'ARN + et -TAP sont migrés sur gel en parallèle et comparés. Ils sont ensuite clonés puis séquencés. Lorsqu'un même produit d'amplification est détecté à partir des C-RACE des deux pools d'ARN (+ et -TAP) cela signifie que les ARN qui ont servi de matrices ont une extrémité 5'monophosphate, contrairement aux ARN précurseurs 5'-triphosphates pour lesquels des ADNc ne seront amplifiés qu'à partir du pool traité à l'enzyme TAP.

Les résultats obtenus pour chacun des loci sRkB et sRkC sont illustrés dans la Figure 22. Un couple d'oligonucléotides (R1) strictement spécifique a été utilisé pour amplifier les transcrits du locus B (Figure 22A). 71 % des clones séquencés ont leur extrémité 5' cartographiée à un nucléotide unique à la position 810048 du génome, identique à l'extrémité 5' identifiée par extension d'amorce (Figures 22A, 22C). 17% des clones séquencés ont une taille comprise entre 340 et 343 nt. Ces clones sont amplifiés à partir d'ARN circulaires préalablement traités par la TAP indiquant que les ARN servant de matrice ont leur extrémité 5' de nature triphosphate, les ARN sont notés sRkB343 (Figure 22A). 36% des clones séquencés ont une taille définie entre 209 et 218 nt (la majeure partie des clones séquencés sont de 216 nt). Ces clones sont amplifiés à partir des deux préparations d'ARN traités ou non par la TAP suggérant que l'extrémité 5' des ARN qui ont servi de matrice est 5'monophosphate, ces ARN sont notés sRkB216. Par la nature de leur extrémité 5', il est possible que sRkB343 soit un transcrit précurseur qui subirait des processus de maturation (tel que la conversion du 5' triphosphate en 5' monophosphate) ou de dégradation pour donner sRk216. Parmi les clones séquencés, nous avons retrouvé des séquences de tailles comprises

entre 275 et 283 nt dont les extrémités 3' sont variables. Celles-ci pourraient correspondre à des intermédiaires de maturation ou à des produits de dégradation. De plus, bien que l'extrémité 5' soit parfaitement identique pour sRkB343 et sRkB216, leurs extrémités 3' sont différentes et chevauchent toutes les deux l'ORF PAB0571 sur une longueur de 25 à 150 nt (Figures 22A et 22C). Nous n'avons pas pu prédire la présence d'un motif promoteur pour PAB0571. Nous avons alors réalisé une extension d'amorce à l'aide d'un oligonucléotide complémentaire à l'ARNm de cette ORF. Plusieurs signaux sont déterminés dont le plus grand correspond à l'extrémité 5' de nos ARN (Figure en annexe 4). Bien que nous n'ayons pas de données supplémentaires, ce résultat pourrait suggérer que PAB0571 serait co-transcrit avec les ARN sRk343 et sRk216.

Pour le locus sRkC (Figure 22B), une première analyse par C-RACE notée R3 a permis de séquencer deux classes de clones majoritaires. La première correspond à 53% des clones séquencés dont leur extrémité 5' est cartographiée à la position 1613195 du génome. Cette extrémité est identique à celle obtenue par l'analyse en extension d'amorce (Figures 5, 22B et 22C). Ces clones d'une taille comprise entre 211 et 218 nt (211 nt étant majoritaire) ne sont amplifiés qu'à partir de la préparation d'ARN traitée par la TAP suggérant que les ARN sRkC211 qui ont servi de matrice seraient 5' triphosphate. La deuxième classe comprend 22 % des clones séquencés. Leur extrémité 5' est cartographiée à la position 1613046 du génome. Ces clones ont une taille comprise entre 64 et 66 nt (Figures 22B et 22C). L'amplification de ces clones a été obtenue à partir des deux préparations d'ARN totaux traités et non traités par la TAP suggérant que la matrice ARN sRkC66 serait 5' monophosphate. Le produit d'amplification est faiblement visible sur le gel d'analyse de RT-PCR (Figure 22B) et aucun oligonucléotide n'a pu permettre la détection de l'expression de sRkC66 par Northern Blot (Figure 21A). Ces observations suggèrent que sRkC66 ne serait présent qu'en faible quantité dans les préparations d'ARN totaux. Sa détection par C-RACE pourrait être expliquée par sa taille : plus une matrice est courte et plus sont amplification sera rapide. L'analyse de l'expression du locus sRkC par Northern Blot avait mis en évidence la présence d'un autre transcrit d'environ 145 nt (Figure 21A) mais la C-RACE R3 n'a pas conduit à l'amplification de ce transcrit. Nous avons choisi un autre couple d'oligonucléotides afin de réaliser la C-RACE R2 qui a permis de séquencer des clones pouvant correspondre à ce transcrit de 145 nt. Un total de 75% de ces clones séquencés sont de taille comprise entre 137 et 143 nt et leur extrémité 5' monophosphate est à la même position que celle de sRkC211. De façon surprenante, son extrémité 3' est située à environ 4 nt en amont de l'extrémité 5' déterminée pour sRkC66. L'analyse des résultats de C-RACE R2 et R3 pour le



**Figure 23 : Caractéristiques de la structure secondaire proposée pour sRkB.** (A) Empreinte à la RNase T1, C : contrôle ARN, T1 ladder : digestion totale de sRkB216 par la RNase T1 en 10' et 20'. T1 : digestion partielle de sRkB216 par la RNase T1 en absence ou en présence de la protéine L7Ae à 400 nM. OH : hydrolyse alcaline. Les résidus G sont indiqués sur le gel séquence. (B) Structure secondaire proposée pour sRkB216. Les flèches indiquent les sites de digestion de la RNase T1, les résidus G protégés sont indiqués dans un cercle rouge, les deux motifs kturn proposés sont indiqués dans un cadre gris. Les nucléotides surlignés en bleu correspondent au codon d'initiation annoté pour le transcrit de l'ORF PAB0571.

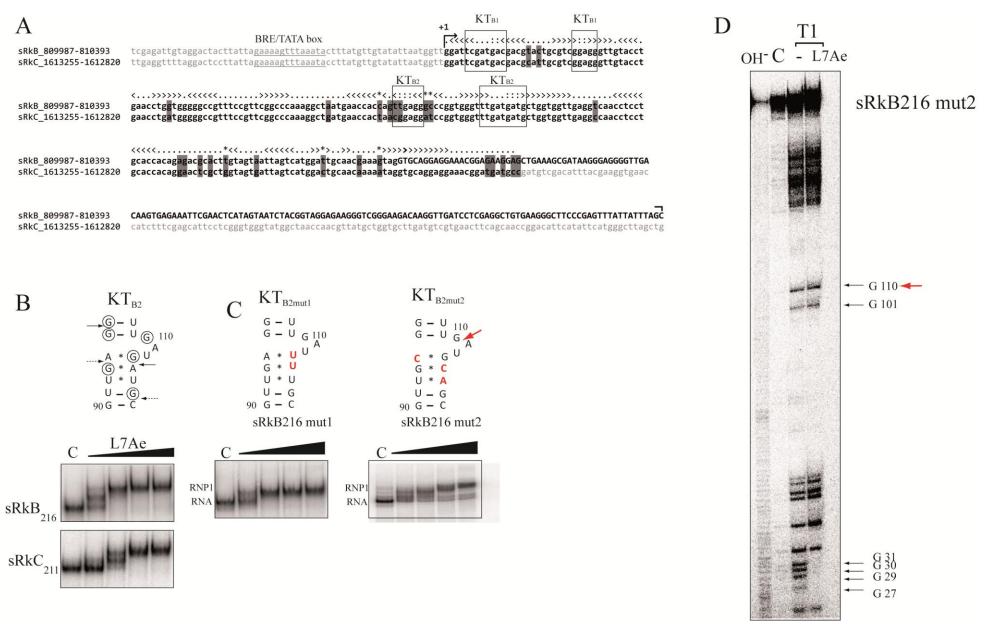


Figure 24 : Caractérisation des motifs K-turn. (A) Alignement des loci sRkB et sRkC. Les K-turn proposés sont indiqués. (B) et (C) Expériences de retard sur gel. Les ARN sRkB<sub>216</sub>, sRkC<sub>211</sub> et les ARN sRk<sub>B216</sub> mutés dans le motif KTB2 sont radiomarqués en 5' au γ-ATP [32]P puis sont mis en présence de concentrations croissantes de la protéine L7Ae recombinante : 0 (C), 50, 100, 200 et 400 nM. Les nucléotides mutés sont représentés en caractères rouges. (C) Empreinte à la RNase T1 du mutant sRkB216 KTB2mut2.

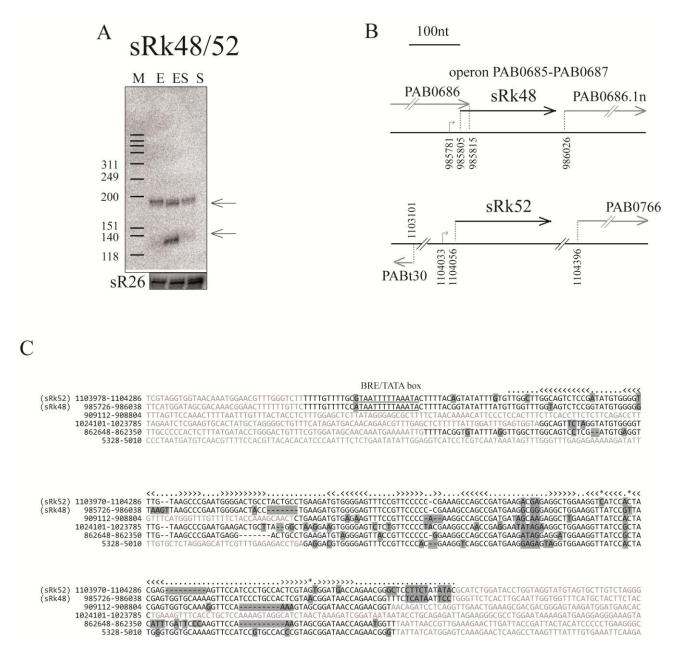
locus sRkC suggère que celui-ci serait transcrit en un ARN précurseur sRkC211 qui pourrait faire l'objet de processus de maturation afin de donner les ARN sRkC143 et sRkC66.

Des analyses par extension d'amorce ont été effectuées et confirment les extrémités 5' des transcrits sRkB et sRkC définies par C-RACE. De plus, nous pouvons proposer un motif promoteur consensus BRE/TATA pour les deux loci situé à 24 nt en amont du nucléotide +1 déterminé (**Figure 22C**).

Sur la base de l'alignement des séquences des transcrits de sRkB et sRkC, un modèle de structure secondaire est proposé pour sRkB216 (**Figure 23**). Cette structure est constituée de cinq tige-boucles notées de P1 à P5 et nous proposons la présence de deux motifs k-turn KTB1 et KTB2 localisés dans P1 et P4 respectivement. Par ailleurs, la recherche de critères lors de la prédiction bio-informatique avait prédit la présence d'un motif K-turn et la présence d'une structure en tige-boucle. L'alignement des séquences des loci sRkB et sRkC montre que les motifs K-turn retrouvés dans sRkB seraient aussi conservés dans sRkC (**Figure 24A**). Nous avons alors testé la capacité des deux ARN à former des complexes RNP *in vitro* avec la protéine L7Ae se liant spécifiquement aux motifs K-turn. Des expériences de retard sur gel ont été réalisées avec des transcrits *in vitro* de sRkB216 radiomarqués et des concentrations croissantes en protéine L7Ae de *P. abyssi* recombinante (**Figure 24B**). Nous pouvons observer que deux types de complexes RNP (RNP1 et RNP2) sont formés en présence de 50 mM et 100 mM de L7Ae pour sRkB216 et sRkC211 respectivement. Au-delà de ces concentrations, seul le complexe RNP2 est visible.

Nous avons réalisé une empreinte de sRkB216 à la RNase T1 en présence ou en absence de la protéine L7Ae (Figure 23). Nous observons 18 positions fortement protégées de la digestion par la RNase T1 en présence de 400 nM de L7Ae : les nucléotides G27/29/30 correspondant au premier motif KTB1 proposé les résidus protégés et G93/95/96/97/110/113/116 seraient impliqués dans le motif KTB2. Il est à noter que les G126/128/129 sont aussi protégés de la digestion par la RNase T1 mais le modèle de structure secondaire que nous proposons ne permet pas d'expliquer la protection de ces résidus par la L7Ae mais l'implication de ces résidus dans une structure tertiaire pourrait justifier leur protection.

Nous avons poursuivi la caractérisation du motif K-turn KTB2 de sRkB216. Deux ARN sRkB216 mutants sont produits *in vitro* et portent chacun des mutations différentes au niveau des nucléotides impliqués dans le motif : KTB2mut1 (GA-AG en GU-AU) et KTB2mut2 (GA-AG en GC-CG) (**Figure 24C**). Nous avons regardé leur capacité à se lier à la



**Figure 25 : Caractéristiques des loci sRk48 et sRk52.** (A) Northern Blot. (B) Contexte génomique. (C) Alignement des séquences répétées retrouvées dans le génome de *P. abyssi*. Les différentes annotations sont décrites dans la figure 16.

protéine L7Ae. Les deux mutants sont toujours capables de former un complexe *in vitro* avec L7Ae mais en comparaison à l'ARN sRkB216, seul le RNP1 est visible. Nous avons réalisé une empreinte à la RNase T1 de sRkB216-KTB2mut2 et nous observons que le résidu G110 est digéré par la RNase T1 (**Figure 24D**). Ces résultats montrent que les mutations apportées dans le motif KTB2 réduisent l'efficacité de fixation de L7Ae sur l'ARN et confirment que sRkB216 possèderait un second motif k-turn. Le motif KTB1que nous proposons est celui modélisé dans la tige-boucle P1. Des analyses complémentaires sont actuellement en cours afin de caractériser ce second motif.

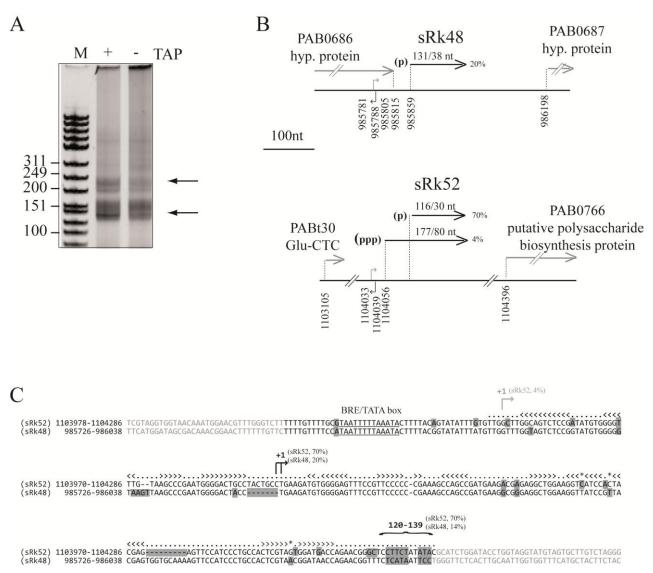
Nous avons proposé une structure secondaire pour sRkB216 qui inclut les 25 premiers nucléotides de l'ORF PAB0571. Le codon start annoté GUG semblerait être impliqué dans les nucléotides appariés dans la structure de la tige-boucle P5. Cette configuration évoque un mécanisme de régulation post-transcriptionnelle en *cis* de l'ARNm de PAB0571.

#### sRk48/52

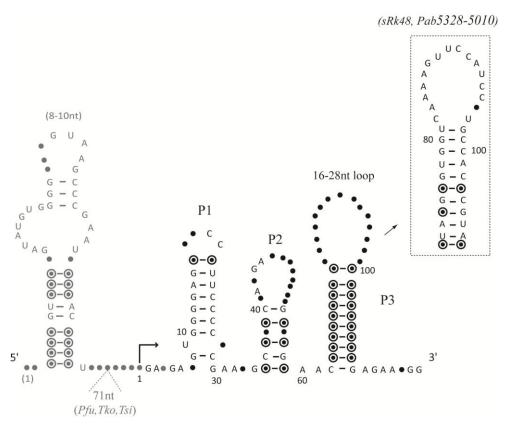
La classe d'ARN issus de loci répétés comprend deux candidats détectés par l'analyse RNAsim. Ces deux candidats font partie d'un groupe de six séquences répétées retrouvées sur le génome de *P. abyssi* (**Figure 25B**). Par la même démarche que pour sRk49 et sRkB/C, nous avons défini des oligonucléotides spécifiques à chacun des loci permettant de ne détecter des transcrits que pour les loci sRk48 et sRk52. De façon surprenante, le profil de détection observé en Northern Blot est semblable pour les deux loci (Figure 25A), suggérant qu'une cross-hybridation pourrait expliquer ces profils similaires. L'un et/ou l'autre des deux loci semblent être transcrits en ARN stables de 130 et 190 nt (**Figure 25A**). Les deux types d'ARN sont exprimés de façon différentielle en fonction de la phase de croissance : le transcrit à 190 nt est plus abondant en phase exponentielle et en entrée en phase stationnaire qu'en phase stationnaire et le transcrit de 130 nt ne semble être exprimé que dans les cellules en entrée en phase stationnaire. Ces transcrits peuvent résulter soit de processus de maturation à partir d'un ARN précurseur, soit de processus de dégradation.

sRk48 est un locus situé dans une région intergénique entre deux ORF d'un opéron de 4 gènes annotés comme protéines hypothétiques (**Figure 25B**). sRk52 est quant à lui retrouvé dans une IGR, flanqué d'un gène d'un ARNt en 5' et d'une ORF en 3' annotée pour coder une protéine hypothétique impliquée dans la biosynthèse de polysaccharide.

Un motif consensus BRE/TATA est prédit à 23 nt en amont de l'extrémité 5'. Afin de mieux définir les extrémités et la nature des transcrits issus de ces deux loci, nous avons réalisé une analyse en C-RACE en utilisant un couple d'oligonucléotides permettant



**Figure 26 : Analyse en C-RACE.** (A) et (B) Caractéristiques schématiques des transcrits sRk48 et sRk52. (C) Cartographie des transcrits observés pour les deux loci. Les différentes annotations sont décrites dans la figure 16.



**Figure 27 : Structure secondaire proposée pour sRk48 et sRk52** sur la base de l'alignement des séquences conservées chez les Thermococcales (Figure en Annexe). L'extension de la séquence 5' observée pour sRk52 (190 nt) est représentée en gris. L'insertion de 9 nt observée chez sRk48 permet d'étendre la tige P3 par d'autres appariements. Les différentes annotations sont décrites dans la figure 16.



Figure 28 : Alignement des séquences similaires à sRk48 et sRk52 retrouvées chez les autres Thermococcales. Les séquences retrouvées dans des régions intergéniques sont représentées en caractères majuscules, les séquences retrouvées dans les ORF annotées, en minuscules, les ORF et les loci génomiques sont indiqués entre parenthèses. Les insertions d'environ 70 nt observées chez *P. furiosus*, *T. kodakaraensis* et *T. sibiricus* sont indiquées entre les caractères '/\'.

d'amplifier simultanément les transcrits issus des deux loci. Les résultats de C-RACE sont schématisés dans la **Figure 26**. De façon surprenante, 70% des clones séquencés correspondent au transcrit sRk52 d'une taille d'environ 130 nt alors que seulement 14% correspondent au transcrit de taille similaire du locus sRk48, tous deux ayant une extrémité 5' monophosphate. Contrairement à ce qui a été observé en Northern Blot, seulement 4% des clones séquencés ont une taille de 180 nt en moyenne dont l'extrémité 5' est triphosphate. Le transcrit de 190 nt détecté pour sRk48 en Northern Blot semble alors résulter d'une cross-hybridation. En effet, lorsque l'on aligne les séquences de deux loci, très peu de nucléotides divergent au niveau de leur séquence primaire excepté deux insertions de 9 nt et de 7 nt observées pour sRk48 et sRK52 respectivement (**Figure 26C**).

Nous avons regardé si des séquences similaires étaient retrouvées dans les génomes des 6 autres Thermococcales séquencées et la recherche par l'outil Blast a mis en évidence l'existence de ces séquences dans les génomes de *P. horikoshii*, *P. furiosus*, *T. kodakaraensis* et *T. sibiricus* (**Figure 27**). Sur la base de l'alignement obtenu, nous avons prédit une structure secondaire pour cette famille d'ARN (**Figure 28**). 45 changements compensatoires sont observés et permettent de renforcer la structure secondaire proposée. Des variabilités telles que des insertions sont indiquées, une première de 71 nt observée chez *P. furiosus*, *T. kodakaraensis* et *T. sibiricus*, une seconde observée au niveau de la tige-boucle P3 permettant de rallonger la longueur de la tige tout en conservant le nombre de nucléotide formant la boucle dans le domaine P3. Les nombreux changements compensatoires observés et la conservation phylogénétique pourraient être corrélés avec une conservation de ces ARN dans la cellule. Par ailleurs, comme pour sRkB et sRkC, les séquences similaires à sRk48 et sRk52 retrouvées chez les autres Thermococcales semblent être retrouvées dans des ORF annotées pour des transposases putatives (**Figure 27**).

# **Discussion**

L'objectif de ce travail de recherche était d'identifier de nouveaux ARNnc dans les régions intergéniques du génome de l'archée hyperthermophile P. abyssi. Pour cela, nous avons utilisé une approche bio-informatique combinant deux méthodes : une analyse ab initio qui exploite le biais de composition en nucléotides du génome et une analyse comparative des régions intergéniques des quatre génomes de Thermococcales disponibles au début de ce travail. Cette approche bio-informatique nous a permis de définir 82 régions candidates pouvant potentiellement coder des ARNnc. Nous avons détecté un signal de transcription pour 24 d'entre elles. Nous avons classé ces 24 régions en fonction de la taille de leur transcrits supérieurs ou inférieurs à 500 nt. Cinq régions sont transcrites en ARN de taille supérieure à 500nt et 19 en ARN de taille comprise entre 50 et 300 nt. Nous avons divisés ces dernières régions en sous-classes en fonction de leur localisation génomique : (i) sept regions correspondent aux gènes des ARN guides de modification à boîtes H/ACA, (ii) trois sont des loci du système CRISPR/Cas, (iii) quatre sont situés à proximité ou chevauchent en 5' une ORF (sRk28, 33, 61 et B) et (iv) quatre sont localisés dans des régions intergéniques (sRk11, sRkC, sRk49.1/49.2 et sRk48/52). A l'exception de sRkB et sRkC, des séquences similaires sont retrouvées dans au moins un autre génome de Thermococcales.

# A. Les limites des méthodes bioinformatiques et expérimentales

La rapidité de mise en œuvre de la méthode de prédiction bioinformatique a été un atout majeur de cette étude. Comme toute méthode d'analyse, ces prédictions sont biaisées par la limite des outils utilisés. En effet, l'analyse *ab initio* se base sur les différentes caractéristiques d'ARNnc connus (contenu en GC, motifs promoteurs, motifs structuraux caractéristiques...) et l'analyse comparative de génomes est gérée par l'outil QRNA (Rivas and Eddy, 2001). Ces deux méthodes d'analyse sont conçues pour la recherche d'ARNnc structurés et conservés. Par exemple les ARN guides de modification à boîte C/D, peu structurés, n'ont pas été détectés. Outre les limites propres aux outils utilisés, nous avons effectué un tri manuel pour s'affranchir des alignements douteux des régions de chaque composante connexe implémentant un paramètre arbitraire à notre méthode. Finalement, la base des recherches *in silico* se repose sur l'annotation des génomes proposée sur NCBI qui peut être erronée avec certains régions intergéniques codantes et inversement. De plus, notre recherche n'est certainement pas exhaustive puisqu'il existe chez les bactéries des ARN

régulateurs qui peuvent aussi coder des petits peptides tels que l'ARNIII de *S. aureus* (Romby and Charpentier, 2010). La recherche et l'identification de ce type d'ARN régulateurs apparaissent complexes. Enfin, notre stratégie ne permet pas de prédire des candidats ARNnc *cis*-antisens à des ORF annotées puisque nous nous sommes limités aux régions intergéniques.

Parmi les 82 prédictions proposées par notre approche bio-informatique, nous avons détectés des transcrits provenant de 24 régions. L'absence de détection de signal de transcription pour les 58 régions restantes ne signifie pas que ces loci ne sont pas transcrits. Le choix des sondes oligonucléotidiques d'hybridation et les différentes conditions de culture des cellules dont sont extraits les ARN sont à l'origine de divers biais. Tout d'abord le choix de sondes spécifiques utilisées en Northern blot est une étape critique. Nous avons choisi des sondes de 20 à 25 nt spécifiques aux régions ciblées à la température d'hybridation que nous avons fixée à 50°C afin de maintenir des conditions de stringence classiquement utilisées pour la détection de transcrits cellulaires. De plus, les sondes utilisées ne recouvrent pas la totalité de la prédiction établie pour chaque région candidate. Il est donc probable que certaines régions transcrites ne soient pas détectées dû au choix des sondes. Les préparations d'ARN totaux utilisées proviennent de cellules cultivées en condition de laboratoire (milieu VSM, incubation en anaérobiose à 95°C, annexe 5 pXVIII) et prélevées dans trois phases de croissance : en phase exponentielle, en entrée en phase stationnaire et en phase stationnaire de croissance. Il est possible que l'expression de certains ARN soit dépendante de condition de culture particulière, de stress par exemple. Dans notre étude, nous n'avons pas abordé cette question au niveau expérimentale contrairement à l'étude effectuée chez l'archée, M. mazeï, où 99 ARNnc sont identifiés dans des conditions de croissance limitées en azote et seulement 36 ARNnc à des concentrations élevées en azote (Jager et al., 2009). De plus, la détection sur membrane par hybridation avec une sonde radiomarquée est limitée par la quantité d'ARN présent sur la membrane ; en dessous d'une certaine concentration, il n'y aura pas de signal visible. En effet chez H. volcanii, l'ARNnc sRNA<sub>63</sub> codé sur un de ses plasmides avait été révélé par RNomics alors que la vérification de son expression par Northern Blot n'a pas permis de détecter de signal de sa transcription suggérant que cet ARN serait exprimé en quantité inférieure au seuil de détection par Northern Blot (Straub et al., 2009). En conclusion, notre étude est non exhaustive et a permis de détecter 24 régions exprimées soit environ pour 30% de nos prédictions. Pour les 70% restantes, nos techniques d'analyse ne permettent pas de détecter d'expression. Pour une analyse plus large, une étude du transcriptome à grande échelle semble l'outil de choix pour avoir une vue d'ensemble des ARNnc présents chez *P. abyssi*.

# B. Les ARN guides de modification à boîtes H/ACA et leur maturation

Premièrement, la détection des sept ARN guides de modification à boîtes H/ACA répertoriés par (Muller et al., 2008) valide nos méthodes de prédiction et certifie nos techniques expérimentales de validation, en particulier la qualité de nos préparations d'ARN totaux de P. abyssi. De plus, plusieurs transcrits de tailles différentes ont été détectés pour Pab19, Pab21 et Pab160 suggérant qu'ils feraient l'objet de processus de maturation. Cependant, peu de données sont disponibles sur ces processus pour expliquer nos observations. Les contextes génomiques de Pab19 et Pab21 indiquent que leurs gènes sont respectivement adjacents à l'opéron codant les ARNr et à un gène d'un ARN guide à boîtes C/D. Nos analyses par Northern Blot suggèrent que Pab21 serait co-transcrit avec l'ARN guide à boîtes C/D, Pab-snRNA19. La co-transcription des homologues de ces ARN avait déjà été suggérée chez P. furiosus (Klein et al., 2002). Nos résultats évoquent la présence d'un précurseur commun. La définition précise des extrémités 5' et 3' des ARN matures permettrait de définir leurs sites de maturation. Il serait également intéressant de regarder si Pab19 serait co-transcrit avec les ARNr adjacents. De nouveaux éléments sur les mécanismes de maturation pour les ARN guides à boîtes H/ACA restent à élucider telle que l'identification de sites de maturation et les ribonucléases impliquées.

## C. Le système CRISPR chez *P. abyssi*

L'outil RNAsim a proposé des candidats ARN retrouvés dans trois des quatre loci CRISPR. Nous nous sommes donc engagés dans l'analyse détaillée des cassettes CRISPR de *P. abyssi* et nous avons constaté une ambigüité entre les annotations proposées par (Portillo and Gonzalez, 2009) et par GenBank. Ceci nous a conduit à analyser précisément les séquences de chaque cassette et de constater que des séquences répétées dégénérées n'étaient pas toujours placées en 3' de l'annotation proposée. En effet il est noté dans la littérature que des séquences directes répétées en 3' des cassettes CRISPR divergent des séquences directes

en amont par un ou plusieurs nucléotides (Horvath et al., 2008; Jansen et al., 2002a; Jansen et al., 2002b). Sur cette base, nous proposons que le locus CRISPR1 soit codé sur le brin sens et que les trois loci CRISPR2, 3 et 4 le soient sur le brin inverse du génome.

Nous avons montré que seules les cassettes CRISPR 1 et 4 sont transcrites en ARN de taille variable de 50 à 60 nt. D'après nos résultats d'extension d'amorce, leur extrémité 5' semble bien définie suggérant que leur variabilité de taille serait due à leur extrémité 3' hétérogène comme observé par ailleurs (Brouns et al., 2008; Carte et al., 2008). De plus, nous avons détecté de longs transcrits en faible quantité (observables en extension d'amorce seulement) qui semblent être des précurseurs des ARNCr observés comme il avait été montré chez P. furiosus et les Sulfolobales (Carte et al., 2008; Lillestol et al., 2009). Ceci corrobore avec la nature des extrémités 5' cartographiées des ARNCr qui correspondent exactement au site de clivage de l'endonucléase Cas6 caractérisée chez P. furiosus (Carte et al., 2008). En effet l'homologue de l'endonucléase Cas6 (PAB1613) est présent dans le génome de P. abyssi et est situé à proximité de la cassette CRISPR3 qui par ailleurs semble silencieuse. Il est à noter que PAB1613 est retrouvé dans un court opéron composé de deux ORF. Son équivalent chez les autres Thermococcales est retrouvé dans des opérons de gènes cas beaucoup plus large constitués de 4 à 19 ORF. Cependant, aucune autre donnée ne permet d'émettre d'hypothèse sur la fonction et l'originalité des gènes cas chez P. abyssi et de proposer un modèle de maturation complet des ARN Cr chez P. abyssi.

Aucun signal de transcription provenant des CRISPR2 et 3 n'a été observé suggérant qu'ils ne sont pas ou peu exprimés dans les conditions expérimentales testées. Il est possible que ces loci soient des reliquats d'anciennes cassettes CRISPR de part leur nombre réduit de répétitions directes et la dégénérescence des séquences de celles-ci. Cette observation peut être mise en parallèle aux changements dynamiques et évolutifs des CRISPR décrits par ailleurs. En effet, une étude récente chez la bactérie *Enterococcus feacalis* a montré que des souches peuvent émerger avec la perte de loci CRISPR (Palmer and Gilmore, 2010). De plus, cette hypothèse est renforcée par une étude très récente suggérant un phénomène « d'auto-immunité » : la présence d'une séquence génomique similaire à la séquence d'un espaceur provoquerait un phénomène de délétion de la cassette CRISPR correspondante (Dyall-Smith, 2011).

Chez les Thermococcales, nous avons analysé l'alignement des séquences « leader » situées en amont des cassettes CRISPR et mis en évidence la présence d'un motif riche en A conservé. Ce motif pourrait être impliqué dans le mécanisme d'acquisition de nouveaux proto-espaceurs en servant de déterminant pour leur insertion. En effet, chez les bactéries S.

thermophilus et Yersinia pestis, cette acquisition semblerait se faire préférentiellement en 5' des cassettes CRISPR, juste en aval des séquences dites « leader » (Deveau et al., 2008; Pourcel et al., 2005).

Les ARN de type CRISPR interviennent dans des mécanismes encore mal caractérisés de défense des cellules contre les phages et les plasmides. A ce jour, 3% des séquences espaceurs de loci CRISPR bactériens et d'archées répertoriés présentent des séquences similaires à des séquences de génomes de phages ou de plasmides séquencés (Marraffini and Sontheimer, 2008). Par une analyse BLASTn, nous avons pu constater que deux séquences espaceurs (séquences 7 et 19) du CRISPR1 de *P. abyssi* présentent de courtes régions de similarité (14 et 17 nt respectivement) avec des ORF annotées dans le génome du virus PAV1. L'analyse de l'expression de ces séquences espaceurs lors d'une infection par le virus PAV1 pourraient apporter des évidences de fonction immunitaire du système CRISPR chez *P. abyssi*.

## D. Des ARNnc chez *P. abyssi*

Il est clairement montré que des ARNnc sont importants pour la croissance dans certaines conditions. Chez *H. volcanii*, la délétion de gènes de deux ARNnc a des effets sur la croissance des souches mutantes selon la concentration en sel. La comparaison de leurs profils protéomiques par rapport à celui de la souche sauvage a mis en évidence un certain nombre de protéines dont l'expression était influencée par l'absence de ces ARNnc. Chez *H. volcanii*, ces ARNnc possèderaient donc des fonctions régulatrices en condition de stress salin (Straub et al., 2009). Dans le cadre de notre étude, nous ne disposions pas d'outil génétique permettant d'établir la fonction des nouveaux ARNnc identifiés dans ce travail. Nous nous sommes donc contentés d'effleurer leur fonction en analysant leur contexte génomique, leur structure secondaire putative et leur conservation dans la phylogénie des Thermococcales.

## 1. Des ARNnc cis-régulateurs

Parmi les loci exprimés, certains d'entre eux, sRk61, sRk28 et sRkB, sont localisés à proximité ou chevauchent une phase ouverte de lecture (ORF) qui pourrait donc correspondre à des régions 5'UTR d'ARNm. Cette configuration laisse penser à celle de certains ARN *cis*-régulateurs (Jager et al., 2009). Des analyses globales du transcriptome chez l'archée *M*.

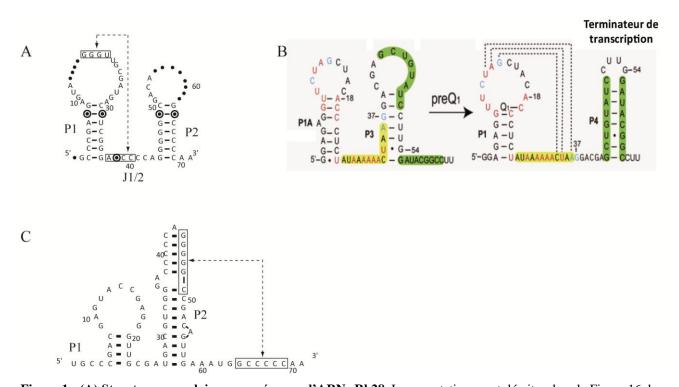


Figure 1 : (A) Structure secondaire proposée pour l'ARN sRk28. Les annotations sont décrites dans la Figure 16 du chapitre Résultats. Un pseudo nœud est proposé, celui-ci est représenté par les nucléotides encadrés. (B) Modèle de changement de structure secondaire du riboswitch PreQ1 contrôlant l'expression de l'opéron queCDEF chez la bactérie B. subtilis (tiré de (Kang et al., 2009)). Les nucléotides surlignés en jaune et vert sont impliqués dans le réarrangement de la structure secondaire en présence ou en absence du ligand preQ1. (C) Structure secondaire alternative proposée pour l'ARN sRk28. Dans ce modèle, les nucléotides encadrés participeraient à la formation d'un pseudo nœud.

mazeï confirment la présence de région 5'UTR dans certains ARNm jusqu'à 500 nt qui peuvent être le lieu de régulations post-transcriptionnelles (Jager et al., 2009). Il n'est donc pas exclu que des ARN de type riboswitch soient stables dans la cellule. Par ailleurs, la recherche d'ARN réalisée chez la bactérie *S. aureus* par l'outil RNAsim avait permis d'identifier et de confirmer l'existence de riboswitch annotés et conservés chez les bactéries de la même famille (Geissmann et al., 2009).

Nous avons observé que l'expression du loci sRk28 est sensible à la phase de croissance. Ce locus se situe en amont de PAB1992 appartenant à un opéron composé de quatre ORF (PAB1989/1990/1991/1992). PAB1992 est annotée comme codant une protéine hypothétique appartenant à la famille des COG0037 (pour Cluster of Ortholog Groups) qui regroupe des ATPases de la super famille PP loop importante dans le contrôle du cycle cellulaire et entre autre, l'ATPase QueC de B. subtilis dont l'expression est régulée de manière transcriptionnelle par le riboswitch preQ1 (Roth et al., 2007). Le gène queC fait partie de l'opéron que CDEF impliqué dans la biosynthèse de la guanine hypermodifiée, la queuosine. Un des intermédiaires de la voie de biosynthèse de ce nucléotide est le nucléotide guanine modifié preQ1 qui est le ligand du riborégulateur qui se fixe sur l'aptamère. Après une recherche de similarité via l'outil Blastx, PAB1992 serait apparenté à l'ORF codant la QueC de différents Bacilles. Chez les archées, la guanosine modifié, preQ1, est un intermédiaire de la voie de biosynthèse de l'archaeosine, guanosine hypermodifiée présent en position 15 de tous les ARNt d'archées. Il est concevable que la transcription en ARNm de PAB1992 soit régulée de la même manière par un riborégulateur. Ainsi, nous proposons une structure secondaire pour sRk28 constituée de deux tiges boucles avec un pseudo-nœud (Figure 1A) semblable à celle du riboswitch preQ1 bactérien (Figure 1B) (Gregson et al., 1993; Kang et al., 2009; Phillips et al., 2008). De plus, il est possible de dessiner une structure alternative pour sRk28 (Figure 1C). L'ensemble de ces données permet d'envisager un rôle de type riboswitch pour sRk28. On pourrait à présent définir si sRk28 est capable de se lier à une guanosine modifié de type preQ1 et de regarder les variations de structure secondaire de sRk28 en présence et absence de ligand.

D'après nos résultats, sRkB est un ARNnc recouvrant en 5' l'ORF PAB0571 adjacente. Pour essayer de comprendre ce chevauchement, nous avons identifié l'extrémité 5' de l'ARNm de PAB0571 et montré que les mêmes signaux de transcription seraient utilisés par sRkB et l'ARNm de PAB0571. De plus, la structure secondaire pour la forme longue de l'ARN sRkB englobe les 25 premiers nucléotides de l'ARNm PAB0571 incluant trois nucléotides d'un codon d'initiation potentiel (Annexe 4). PAB0571 est annotée comme une

protéine hypothétique possédant un domaine ATPase, conservé chez les bactéries et les archées, de la superfamille des ATPases AAA+ possédant un répertoire varié de fonctions cellulaires telles que les processus de dégradation des protéines et la réparation de l'ADN (Snider and Houry, 2008). Le fait que sRkB possède au moins deux motifs K-turn dans sa structure secondaire indique que cet ARN est potentiellement un ARN fonctionnel. En effet, le motif K-turn est retrouvé dans de nombreux ARN fonctionnels bien connus tels que les ARNr, les ARN guides de modification à boîtes C/D et H/ACA et plus récemment l'ARN de la RNase P (Cho et al., 2010). Au sein de ces différents ARN, le motif K-turn est un motif structural en coude souvent stabilisé mais pas toujours par la fixation d'une protéine, par exemple L7Ae chez les archées (voir Introduction, pages 37-38). Il est maintenant établi que ce motif est aussi retrouvé dans des éléments riboswitch de type L-box et de type SAM retrouvés chez B. subtilis (Blouin and Lafontaine, 2007; Heppell and Lafontaine, 2008). D'autre part, chez S. solfataricus, la protéine L7Ae co-immunoprécipite avec quatre ARNnc qui seraient issus de la transcription de loci chevauchant les parties 5' d'ORF dont certaines sont annotées comme transposases. Des structures secondaires avaient été proposées pour trois d'entre eux qui incluent des motifs K-turn comme attendu et des structures en tigeboucles masquant les codons d'initiation des protéines adjacentes (Zago et al., 2005). Ces différents exemples suggèrent que les transcrits de type riboswitch peuvent être retrouvés dans les ARN cellulaires sous forme stable et pourraient être des produits de maturation de transcrit précurseur. Il est montré que la ribonucléase RNase Y de B. subtilis est capable d'induire une coupure au niveau du riboswitch yitJ de type transcriptionnel (Shahbabian et al., 2009). De plus, la détection de transcrits stables correspondant à des ribocommutateurs est observée chez E. coli et S. aureus (Beaume et al., 2010; Kawano et al., 2005) et il est montré que chez L. monocytogenes, le riboswitch SreA de type SAM est aussi fonctionnel en trans en s'appariant avec la région 5'-UTR de l'ARNm prfA pour réguler négativement l'expression de la protéine PfrA (Loh et al., 2009).

## 2. Relation ARN régulateurs et transposases

Dans notre étude, nous constatons une relation entre la localisation des loci exprimant des ARNnc (sRk62, sRkB, sRkC, SRk48 et sRk52) et des gènes annotés codant des transposases. Cette observation avait déjà été faite chez d'autres archées et bactéries. Il est établi que chez les bactéries, les éléments transposables Tn10 et Tn30 ont leur mobilité

contrôlé par des ARNnc régulateurs de type *cis*-antisens par complémentarité à la région 5'UTR (masquant les signaux de traduction) ou codante respectivement pour Tn10 et Tn30 (Ma and Simons, 1990) (Arini et al., 1997; Case et al., 1989). De plus, la protéine Hfq, protéine liant de nombreux ARN régulateurs bactériens, semble jouer un rôle dans la régulation de la transposition de Tn10 (Ross et al., 2010). De la même manière, l'identification de 10 ARNnc antisens aux régions codantes annotées comme transposases chez *S. solfataricus* suggère leur importance dans la régulation de la transposition chez les archées (Tang et al., 2002a; Tang et al., 2005). De manière surprenante, tous les nouveaux ARNnc identifiés chez *P. abyssi* ayant un lien avec des ORF de transposase correspondent à la séquence sens des ORF et non pas antisens comme les exemples décrits ci-dessus. Certains transcrits de ce type avaient aussi été identifiés chez *S. solfataricus* par immunoprécipitation avec la protéine L7Ae (Zago et al., 2005). Il semblerait donc que nos observations ne soient pas fortuites, cependant en l'état actuel de nos connaissances, il est difficile de proposer un rôle de ces ARNnc dans la régulation de la transposition.

## E. Conclusion

L'identification des cibles et des partenaires protéiques des ARNnc est essentielle à la compréhension de leur fonction biologique. Le défi actuel est d'identifier les cibles des ARN régulateurs. La recherche de cibles chez les bactéries repose sur la caractérisation des mécanismes de régulation déjà connus permettant de faire des analogies de fonctions. Cependant, cette recherche reste encore difficile à mettre en œuvre : certaines interactions ARN régulateur-ARN cible présentent une région de complémentarité courte (environ 10 nt) rendant l'identification de cible difficile. De plus, certains ARN régulateurs ont plusieurs cibles (Bouvier et al., 2008). Chez les archées, à l'exception des ARN guides de modification qui ont été largement caractérisés, encore aucun mécanisme de régulation de l'expression des gènes par des ARNnc n'est connu. Il est important de noter que des protéines retrouvées associées à des ARNnc tel que Hfq faisant partie de la famille des protéines Sm ou Sm-like sont présentes dans les trois domaines du vivant. Une étude récente a montré que la protéine Sm-like chez H. volcanii pouvait se lier à des ARN (Fisher et al., 2010). L'archée, P. abyssi, possède également une protéine de type Sm-like dont la structure cristalline a été caractérisée en complexe avec un ARN synthétique (Thore et al., 2003). L'implication de cette protéine dans le monde des ARNnc chez P. abyssi pourrait être établi par des expériences de coimmunoprécipitation et permettrait peut-être d'apporter des informations sur les transcrits ciblés par les ARNnc identifiés. Enfin la présence d'ARNnc régulateurs chez les archées hyperthermophiles engendrent un certain nombre de question en particulier celle de la stratégie adoptée par ces micro-organismes pour stabiliser les interactions ARN-ARN à l'origine de la régulation par des ARNnc à haute température. La question de la stabilité a déjà été abordée pour les ARNr des organismes hyperthermophiles. Il a été proposé une corrélation entre le nombre de modifications par des ARN guide de modification et la température de croissance : plus la température est élevée et plus le nombre de modification serait grand (Dennis et al., 2001). Les méthylations du ribose des ARN permettraient de stabiliser les structures d'ARN ou les interactions avec les protéines ribosomiques à haute température. Cette stratégie pourrait aussi favoriser des interactions intermoléculaires. En l'état actuel de nos connaissances et avec le peu d'outils génétiques disponibles, il est encore extrêmement difficile d'établir une fonction des ARNnc stables identifiés dans cette étude.

# Références bibliographiques

- Aiba, H., 2007, Mechanism of RNA silencing by Hfq-binding small RNAs. Curr Opin Microbiol 10, 134-139.
- Allers, T., Mevarech, M., 2005, Archaeal genetics the third way. Nat Rev Genet 6, 58-73.
- Altman, S., Wesolowski, D., Puranam, R.S., 1993, Nucleotide sequences of the RNA subunit of RNase P from several mammals. Genomics 18, 418-422.
- Altuvia, S., 2007, Identification of bacterial small non-coding RNAs: experimental approaches. Curr Opin Microbiol 10, 257-261.
- Amiri, K.A., 1994, Fibrillarin-like proteins occur in the domain Archaea. J Bacteriol 176, 2124-2127.
- Anderson, K.L., Apolinario, E.E., MacAuley, S.R., Sowers, K.R., 2009, A 5' leader sequence regulates expression of methanosarcinal CO dehydrogenase/acetyl coenzyme A synthase. J Bacteriol 191, 7123-7128.
- Aravin, A.A., Hannon, G.J., Brennecke, J., 2007, The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. Science 318, 761-764.
- Arini, A., Keller, M.P., Arber, W., 1997, An antisense RNA in IS30 regulates the translational expression of the transposase. Biol Chem 378, 1421-1431.
- Babitzke, P., Romeo, T., 2007, CsrB sRNA family: sequestration of RNA-binding regulatory proteins. Curr Opin Microbiol 10, 156-163.
- Bachellerie, J.P., Cavaille, J., Huttenhofer, A., 2002, The expanding snoRNA world. Biochimie 84, 775-790.
- Baker, D.L., Youssef, O.A., Chastkofsky, M.I., Dy, D.A., Terns, R.M., Terns, M.P., 2005, RNA-guided RNA modification: functional organization of the archaeal H/ACA RNP. Genes Dev 19, 1238-1248.
- Baliga, N.S., Bonneau, R., Facciotti, M.T., Pan, M., Glusman, G., Deutsch, E.W., Shannon,
  P., Chiu, Y., Weng, R.S., Gan, R.R., Hung, P., Date, S.V., Marcotte, E., Hood, L., Ng,
  W.V., 2004, Genome sequence of Haloarcula marismortui: a halophilic archaeon from
  the Dead Sea. Genome Res 14, 2221-2234.
- Ban, N., Nissen, P., Hansen, J., Moore, P.B., Steitz, T.A., 2000, The complete atomic structure of the large ribosomal subunit at 2.4 A resolution. Science 289, 905-920.
- Barrandon, C., Spiluttini, B., Bensaude, O., 2008, Non-coding RNAs regulating the transcriptional machinery. Biol Cell 100, 83-95.
- Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D.A., Horvath, P., 2007, CRISPR provides acquired resistance against viruses in prokaryotes. Science 315, 1709-1712.
- Barrick, J.E., Breaker, R.R., 2007, The distributions, mechanisms, and structures of metabolite-binding riboswitches. Genome Biol 8, R239.
- Barry, E.R., Bell, S.D., 2006, DNA replication in the archaea. Microbiol Mol Biol Rev 70, 876-887.
- Beaume, M., Hernandez, D., Farinelli, L., Deluen, C., Linder, P., Gaspin, C., Romby, P., Schrenzel, J., Francois, P., 2010, Cartography of methicillin-resistant S. aureus transcripts: detection, orientation and temporal expression during growth phase and stress conditions. PLoS One 5, e10725.
- Bell, S.D., Jackson, S.P., 2001, Mechanism and regulation of transcription in archaea. Curr Opin Microbiol 4, 208-213.
- Bell, S.D., Kosa, P.L., Sigler, P.B., Jackson, S.P., 1999, Orientation of the transcription preinitiation complex in archaea. Proc Natl Acad Sci U S A 96, 13662-13667.
- Benbouzid-Rollet, N., Lopez-Garcia, P., Watrin, L., Erauso, G., Prieur, D., Forterre, P., 1997, Isolation of new plasmids from hyperthermophilic Archaea of the order Thermococcales. Res Microbiol 148, 767-775.

- Benelli, D., Londei, P., 2009, Begin at the beginning: evolution of translational initiation. Res Microbiol 160, 493-501.
- Bensing, B.A., Meyer, B.J., Dunny, G.M., 1996, Sensitive detection of bacterial transcription initiation sites and differentiation from RNA processing sites in the pheromone-induced plasmid transfer system of Enterococcus faecalis. Proc Natl Acad Sci U S A 93, 7794-7799.
- Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Sayers, E.W., 2009, GenBank. Nucleic Acids Res 37, D26-31.
- Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Wheeler, D.L., 2008, GenBank. Nucleic Acids Res 36, D25-30.
- Benson, G., 1999, Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 27, 573-580.
- Berg, I.A., Kockelkorn, D., Ramos-Vera, W.H., Say, R.F., Zarzycki, J., Hugler, M., Alber, B.E., Fuchs, G., 2010, Autotrophic carbon fixation in archaea. Nat Rev Microbiol 8, 447-460.
- Bhuiyan, S.H., Gowda, K., Hotokezaka, H., Zwieb, C., 2000, Assembly of archaeal signal recognition particle from recombinant components. Nucleic Acids Res 28, 1365-1373.
- Bleichert, F., Gagnon, K.T., Brown, B.A., 2nd, Maxwell, E.S., Leschziner, A.E., Unger, V.M., Baserga, S.J., 2009, A dimeric structure for archaeal box C/D small ribonucleoproteins. Science 325, 1384-1387.
- Blochl, E., Rachel, R., Burggraf, S., Hafenbradl, D., Jannasch, H.W., Stetter, K.O., 1997, Pyrolobus fumarii, gen. and sp. nov., represents a novel group of archaea, extending the upper temperature limit for life to 113 degrees C. Extremophiles 1, 14-21.
- Blombach, F., Makarova, K.S., Marrero, J., Siebers, B., Koonin, E.V., van der Oost, J., 2009, Identification of an ortholog of the eukaryotic RNA polymerase III subunit RPC34 in Crenarchaeota and Thaumarchaeota suggests specialization of RNA polymerases for coding and non-coding RNAs in Archaea. Biol Direct 4, 39.
- Blouin, S., Lafontaine, D.A., 2007, A loop loop interaction and a K-turn motif located in the lysine aptamer domain are important for the riboswitch gene regulation control. RNA 13, 1256-1267.
- Blouin, S., Mulhbacher, J., Penedo, J.C., Lafontaine, D.A., 2009, Riboswitches: ancient and promising genetic regulators. Chembiochem 10, 400-416.
- Boisset, S., Geissmann, T., Huntzinger, E., Fechter, P., Bendridi, N., Possedko, M., Chevalier, C., Helfer, A.C., Benito, Y., Jacquier, A., Gaspin, C., Vandenesch, F., Romby, P., 2007, Staphylococcus aureus RNAIII coordinately represses the synthesis of virulence factors and the transcription regulator Rot by an antisense mechanism. Genes Dev 21, 1353-1366.
- Bolhuis, H., Poele, E.M., Rodriguez-Valera, F., 2004, Isolation and cultivation of Walsby's square archaeon. Environ Microbiol 6, 1287-1291.
- Bolotin, A., Quinquis, B., Sorokin, A., Ehrlich, S.D., 2005, Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. Microbiology 151, 2551-2561.
- Bortolin, M.L., Bachellerie, J.P., Clouet-d'Orval, B., 2003, In vitro RNP assembly and methylation guide activity of an unusual box C/D RNA, cis-acting archaeal pretRNA(Trp). Nucleic Acids Res 31, 6524-6535.
- Bose, A., Kulkarni, G., Metcalf, W.W., 2009, Regulation of putative methyl-sulphide methyltransferases in Methanosarcina acetivorans C2A. Mol Microbiol 74, 227-238.
- Bouvier, M., Sharma, C.M., Mika, F., Nierhaus, K.H., Vogel, J., 2008, Small RNA binding to 5' mRNA coding region inhibits translational initiation. Mol Cell 32, 827-837.

- Brantl, S., 2007, Regulatory mechanisms employed by cis-encoded antisense RNAs. Curr Opin Microbiol 10, 102-109.
- Brenneis, M., Hering, O., Lange, C., Soppa, J., 2007, Experimental characterization of Cisacting elements important for translation and transcription in halophilic archaea. PLoS Genet 3, e229.
- Brenneis, M., Soppa, J., 2009, Regulation of translation in haloarchaea: 5'- and 3'-UTRs are essential and have to functionally interact in vivo. PLoS One 4, e4484.
- Brochier-Armanet, C., Boussau, B., Gribaldo, S., Forterre, P., 2008a, Mesophilic Crenarchaeota: proposal for a third archaeal phylum, the Thaumarchaeota. Nat Rev Microbiol 6, 245-252.
- Brochier-Armanet, C., Gribaldo, S., Forterre, P., 2008b, A DNA topoisomerase IB in Thaumarchaeota testifies for the presence of this enzyme in the last common ancestor of Archaea and Eucarya. Biol Direct 3, 54.
- Brochier-Armanet, C., Talla, E., Gribaldo, S., 2009, The multiple evolutionary histories of dioxygen reductases: Implications for the origin and evolution of aerobic respiration. Mol Biol Evol 26, 285-297.
- Brochier, C., Gribaldo, S., Zivanovic, Y., Confalonieri, F., Forterre, P., 2005, Nanoarchaea: representatives of a novel archaeal phylum or a fast-evolving euryarchaeal lineage related to Thermococcales? Genome Biol 6, R42.
- Brouns, S.J., Jore, M.M., Lundgren, M., Westra, E.R., Slijkhuis, R.J., Snijders, A.P., Dickman, M.J., Makarova, K.S., Koonin, E.V., van der Oost, J., 2008, Small CRISPR RNAs guide antiviral defense in prokaryotes. Science 321, 960-964.
- Bult, C.J., White, O., Olsen, G.J., Zhou, L., Fleischmann, R.D., Sutton, G.G., Blake, J.A., FitzGerald, L.M., Clayton, R.A., Gocayne, J.D., Kerlavage, A.R., Dougherty, B.A., Tomb, J.F., Adams, M.D., Reich, C.I., Overbeek, R., Kirkness, E.F., Weinstock, K.G., Merrick, J.M., Glodek, A., Scott, J.L., Geoghagen, N.S., Venter, J.C., 1996, Complete genome sequence of the methanogenic archaeon, Methanococcus jannaschii. Science 273, 1058-1073.
- Bushati, N., Cohen, S.M., 2007, microRNA functions. Annu Rev Cell Dev Biol 23, 175-205. Calo, D., Eichler, J., 2010, Crossing the membrane in Archaea, the third domain of life. Biochim Biophys Acta.
- Carte, J., Wang, R., Li, H., Terns, R.M., Terns, M.P., 2008, Cas6 is an endoribonuclease that generates guide RNAs for invader defense in prokaryotes. Genes Dev 22, 3489-3496.
- Case, C.C., Roels, S.M., Jensen, P.D., Lee, J., Kleckner, N., Simons, R.W., 1989, The unusual stability of the IS10 anti-sense RNA is critical for its function and is determined by the structure of its stem-domain. EMBO J 8, 4297-4305.
- Cavaille, J., Bachellerie, J.P., 1998, SnoRNA-guided ribose methylation of rRNA: structural features of the guide RNA duplex influencing the extent of the reaction. Nucleic Acids Res 26, 1576-1587.
- Cavaille, J., Nicoloso, M., Bachellerie, J.P., 1996, Targeted ribose methylation of RNA in vivo directed by tailored antisense RNA guides. Nature 383, 732-735.
- Cavicchioli, R., 2010, Archaea timeline of the third domain. Nat Rev Microbiol.
- Chaban, B., Ng, S.Y., Jarrell, K.F., 2006, Archaeal habitats--from the extreme to the ordinary. Can J Microbiol 52, 73-116.
- Charbonnier, F., Erauso, G., Barbeyron, T., Prieur, D., Forterre, P., 1992, Evidence that a plasmid from a hyperthermophilic archaebacterium is relaxed at physiological temperatures. J Bacteriol 174, 6103-6108.
- Charpentier, B., Muller, S., Branlant, C., 2005, Reconstitution of archaeal H/ACA small ribonucleoprotein complexes active in pseudouridylation. Nucleic Acids Res 33, 3133-3144.

- Cho, I.M., Lai, L.B., Susanti, D., Mukhopadhyay, B., Gopalan, V., 2010, Ribosomal protein L7Ae is a subunit of archaeal RNase P. Proc Natl Acad Sci U S A 107, 14573-14578.
- Choudhuri, S., 2010, Small noncoding RNAs: biogenesis, function, and emerging significance in toxicology. J Biochem Mol Toxicol 24, 195-216.
- Clouet-d'Orval, B., Gaspin, C., Mougin, A., 2005, Two different mechanisms for tRNA ribose methylation in Archaea: a short survey. Biochimie 87, 889-895.
- Clouet-d'Orval, B., Rinaldi, D., Quentin, Y., Carpousis, A.J., 2010, Euryarchaeal beta-CASP proteins with homology to bacterial RNase J Have 5'- to 3'-exoribonuclease activity. J Biol Chem 285, 17574-17583.
- Clouet d'Orval, B., Bortolin, M.L., Gaspin, C., Bachellerie, J.P., 2001, Box C/D RNA guides for the ribose methylation of archaeal tRNAs. The tRNATrp intron guides the formation of two ribose-methylated nucleosides in the mature tRNATrp. Nucleic Acids Res 29, 4518-4529.
- Cohen, G.N., Barbe, V., Flament, D., Galperin, M., Heilig, R., Lecompte, O., Poch, O., Prieur, D., Querellou, J., Ripp, R., Thierry, J.C., Van der Oost, J., Weissenbach, J., Zivanovic, Y., Forterre, P., 2003, An integrated analysis of the genome of the hyperthermophilic archaeon Pyrococcus abyssi. Mol Microbiol 47, 1495-1512.
- Collins, J.A., Irnov, I., Baker, S., Winkler, W.C., 2007, Mechanism of mRNA destabilization by the glmS ribozyme. Genes Dev 21, 3356-3368.
- Corpet, F., 1988, Multiple sequence alignment with hierarchical clustering. Nucleic Acids Res 16, 10881-10890.
- Cortez, D., Quevillon-Cheruel, S., Gribaldo, S., Desnoues, N., Sezonov, G., Forterre, P., Serre, M.C., 2010, Evidence for a Xer/dif system for chromosome resolution in archaea. PLoS Genet 6, e1001166.
- Costa, F.F., 2007, Non-coding RNAs: lost in translation? Gene 386, 1-10.
- Crick, F., 1970, Central dogma of molecular biology. Nature 227, 561-563.
- Czerwoniec, A., Dunin-Horkawicz, S., Purta, E., Kaminska, K.H., Kasprzak, J.M., Bujnicki, J.M., Grosjean, H., Rother, K., 2009, MODOMICS: a database of RNA modification pathways. 2008 update. Nucleic Acids Res 37, D118-121.
- Darr, S.C., Pace, B., Pace, N.R., 1990, Characterization of ribonuclease P from the archaebacterium Sulfolobus solfataricus. J Biol Chem 265, 12927-12932.
- de Bernis, L., Sherratt, D.R., AbouZahr, C., Van Lerberghe, W., 2003, Skilled attendants for pregnancy, childbirth and postnatal care. Br Med Bull 67, 39-57.
- de Koning, B., Blombach, F., Brouns, S.J., van der Oost, J., 2010, Fidelity in archaeal information processing. Archaea 2010.
- de Miguel Bouzas, T., Barros-Velazquez, J., Villa, T.G., 2006, Industrial applications of hyperthermophilic enzymes: a review. Protein Pept Lett 13, 645-651.
- Decatur, W.A., Fournier, M.J., 2002, rRNA modifications and ribosome function. Trends Biochem Sci 27, 344-351.
- DeLong, E.F., Pace, N.R., 2001, Environmental diversity of bacteria and archaea. Syst Biol 50, 470-478.
- Dennis, P.P., Omer, A., 2005, Small non-coding RNAs in Archaea. Curr Opin Microbiol 8, 685-694.
- Dennis, P.P., Omer, A., Lowe, T., 2001, A guided tour: small RNA function in Archaea. Mol Microbiol 40, 509-519.
- Deppenmeier, U., 2002, The unique biochemistry of methanogenesis. Prog Nucleic Acid Res Mol Biol 71, 223-283.
- Deveau, H., Barrangou, R., Garneau, J.E., Labonte, J., Fremaux, C., Boyaval, P., Romero, D.A., Horvath, P., Moineau, S., 2008, Phage response to CRISPR-encoded resistance in Streptococcus thermophilus. J Bacteriol 190, 1390-1400.

- Deveau, H., Garneau, J.E., Moineau, S., 2010, CRISPR/Cas system and its role in phage-bacteria interactions. Annu Rev Microbiol 64, 475-493.
- Diener, J.L., Wilson, C., 2000, Role of SRP19 in assembly of the Archaeoglobus fulgidus signal recognition particle. Biochemistry 39, 12862-12874.
- Dridi, B., Henry, M., El Khechine, A., Raoult, D., Drancourt, M., 2009, High prevalence of Methanobrevibacter smithii and Methanosphaera stadtmanae detected in the human gut using an improved DNA detection protocol. PLoS One 4, e7063.
- Duan, J., Li, L., Lu, J., Wang, W., Ye, K., 2009, Structural mechanism of substrate RNA recruitment in H/ACA RNA-guided pseudouridine synthase. Mol Cell 34, 427-439.
- Duggin, I.G., Bell, S.D., 2006, The chromosome replication machinery of the archaeon Sulfolobus solfataricus. J Biol Chem 281, 15029-15032.
- Duggin, I.G., Dubarry, N., Bell, S.D., 2010, Replication termination and chromosome dimer resolution in the archaeon Sulfolobus solfataricus. EMBO J.
- Duggin, I.G., McCallum, S.A., Bell, S.D., 2008, Chromosome replication dynamics in the archaeon Sulfolobus acidocaldarius. Proc Natl Acad Sci U S A 105, 16737-16742.
- Dyall-Smith, M., 2011, Dangerous weapons: a cautionary tale of CRISPR defence. Mol Microbiol 79, 3-6.
- Edgar, R.C., 2007, PILER-CR: fast and accurate identification of CRISPR repeats. BMC Bioinformatics 8, 18.
- Ellen, A.F., Zolghadr, B., Driessen, A.M., Albers, S.V., 2010, Shaping the archaeal cell envelope. Archaea 2010, 608243.
- Erauso, G., Marsin, S., Benbouzid-Rollet, N., Baucher, M.F., Barbeyron, T., Zivanovic, Y., Prieur, D., Forterre, P., 1996, Sequence of plasmid pGT5 from the archaeon Pyrococcus abyssi: evidence for rolling-circle replication in a hyperthermophile. J Bacteriol 178, 3232-3237.
- Evguenieva-Hackenberg, E., Klug, G., 2009, RNA degradation in Archaea and Gramnegative bacteria different from Escherichia coli. Prog Mol Biol Transl Sci 85, 275-317.
- Fabbri, S., Fruscoloni, P., Bufardeci, E., Di Nicola Negri, E., Baldi, M.I., Attardi, D.G., Mattoccia, E., Tocchini-Valentini, G.P., 1998, Conservation of substrate recognition mechanisms by tRNA splicing endonucleases. Science 280, 284-286.
- Filee, J., Siguier, P., Chandler, M., 2007, Insertion sequence diversity in archaea. Microbiol Mol Biol Rev 71, 121-157.
- Fischer, S., Benz, J., Spath, B., Maier, L.K., Straub, J., Granzow, M., Raabe, M., Urlaub, H., Hoffmann, J., Brutschy, B., Allers, T., Soppa, J., Marchfelder, A., 2010, The archaeal Lsm protein binds to small RNAs. J Biol Chem 285, 34429-34438.
- Forterre, P., Brochier, C., Philippe, H., 2002, Evolution of the Archaea. Theor Popul Biol 61, 409-422.
- Forterre, P., Gribaldo, S., Brochier, C., 2005, [Luca: the last universal common ancestor]. Med Sci (Paris) 21, 860-865.
- Fozo, E.M., Hemm, M.R., Storz, G., 2008a, Small toxic proteins and the antisense RNAs that repress them. Microbiol Mol Biol Rev 72, 579-589, Table of Contents.
- Fozo, E.M., Kawano, M., Fontaine, F., Kaya, Y., Mendieta, K.S., Jones, K.L., Ocampo, A., Rudd, K.E., Storz, G., 2008b, Repression of small toxic protein synthesis by the Sib and OhsC small RNAs. Mol Microbiol 70, 1076-1093.
- Frank, D.N., Pace, N.R., 1998, Ribonuclease P: unity and diversity in a tRNA processing ribozyme. Annu Rev Biochem 67, 153-180.
- French, S.L., Santangelo, T.J., Beyer, A.L., Reeve, J.N., 2007, Transcription and translation are coupled in Archaea. Mol Biol Evol 24, 893-895.

- Fuglsang, A., 2004, Compositional nonrandomness upstream of start codons in archaebacteria. Gene 332, 89-95.
- Fukuhara, H., Kifusa, M., Watanabe, M., Terada, A., Honda, T., Numata, T., Kakuta, Y., Kimura, M., 2006, A fifth protein subunit Ph1496p elevates the optimum temperature for the ribonuclease P activity from Pyrococcus horikoshii OT3. Biochem Biophys Res Commun 343, 956-964.
- Fukui, T., Atomi, H., Kanai, T., Matsumi, R., Fujiwara, S., Imanaka, T., 2005, Complete genome sequence of the hyperthermophilic archaeon Thermococcus kodakaraensis KOD1 and comparison with Pyrococcus genomes. Genome Res 15, 352-363.
- Galagan, J.E., Nusbaum, C., Roy, A., Endrizzi, M.G., Macdonald, P., FitzHugh, W., Calvo, S., Engels, R., Smirnov, S., Atnoor, D., Brown, A., Allen, N., Naylor, J., Stange-Thomann, N., DeArellano, K., Johnson, R., Linton, L., McEwan, P., McKernan, K., Talamas, J., Tirrell, A., Ye, W., Zimmer, A., Barber, R.D., Cann, I., Graham, D.E., Grahame, D.A., Guss, A.M., Hedderich, R., Ingram-Smith, C., Kuettner, H.C., Krzycki, J.A., Leigh, J.A., Li, W., Liu, J., Mukhopadhyay, B., Reeve, J.N., Smith, K., Springer, T.A., Umayam, L.A., White, O., White, R.H., Conway de Macario, E., Ferry, J.G., Jarrell, K.F., Jing, H., Macario, A.J., Paulsen, I., Pritchett, M., Sowers, K.R., Swanson, R.V., Zinder, S.H., Lander, E., Metcalf, W.W., Birren, B., 2002, The genome of M. acetivorans reveals extensive metabolic and physiological diversity. Genome Res 12, 532-542.
- Galtier, N., Lobry, J.R., 1997, Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes. J Mol Evol 44, 632-636.
- Gao, B., Gupta, R.S., 2007, Phylogenomic analysis of proteins that are distinctive of Archaea and its main subgroups and the origin of methanogenesis. BMC Genomics 8, 86.
- Gardner, P.P., Daub, J., Tate, J.G., Nawrocki, E.P., Kolbe, D.L., Lindgreen, S., Wilkinson, A.C., Finn, R.D., Griffiths-Jones, S., Eddy, S.R., Bateman, A., 2009, Rfam: updates to the RNA families database. Nucleic Acids Res 37, D136-140.
- Garrett, R.A., Dalgaard, J., Larsen, N., Kjems, J., Mankin, A.S., 1991, Archaeal rRNA operons. Trends Biochem Sci 16, 22-26.
- Gaspin, C., Cavaille, J., Erauso, G., Bachellerie, J.P., 2000, Archaeal homologs of eukaryotic methylation guide small nucleolar RNAs: lessons from the Pyrococcus genomes. J Mol Biol 297, 895-906.
- Gautheret, D., Lambert, A., 2001, Direct RNA motif definition and identification from multiple sequence alignments using secondary structure profiles. J Mol Biol 313, 1003-1011.
- Geissmann, T., Chevalier, C., Cros, M.J., Boisset, S., Fechter, P., Noirot, C., Schrenzel, J., Francois, P., Vandenesch, F., Gaspin, C., Romby, P., 2009, A search for small noncoding RNAs in Staphylococcus aureus reveals a conserved sequence motif for regulation. Nucleic Acids Res 37, 7239-7257.
- Geslin, C., Gaillard, M., Flament, D., Rouault, K., Le Romancer, M., Prieur, D., Erauso, G., 2007, Analysis of the first genome of a hyperthermophilic marine virus-like particle, PAV1, isolated from Pyrococcus abyssi. J Bacteriol 189, 4510-4519.
- Godde, J.S., Bickerton, A., 2006, The repetitive DNA elements called CRISPRs and their associated genes: evidence of horizontal transfer among prokaryotes. J Mol Evol 62, 718-729.
- Gonzalez, O., Oberwinkler, T., Mansueto, L., Pfeiffer, F., Mendoza, E., Zimmer, R., Oesterhelt, D., 2010, Characterization of growth and metabolism of the haloalkaliphile Natronomonas pharaonis. PLoS Comput Biol 6, e1000799.
- Gorke, B., Stulke, J., 2008, Carbon catabolite repression in bacteria: many ways to make the most out of nutrients. Nat Rev Microbiol 6, 613-624.

- Gottesman, S., 2005, Micros for microbes: non-coding regulatory RNAs in bacteria. Trends Genet 21, 399-404.
- Gottesman, S., Storz, G., 2010, Bacterial Small RNA Regulators: Versatile Roles and Rapidly Evolving Variations. Cold Spring Harb Perspect Biol.
- Grabowski, B., Kelman, Z., 2003, Archeal DNA replication: eukaryal proteins in a bacterial context. Annu Rev Microbiol 57, 487-516.
- Gregson, J.M., Crain, P.F., Edmonds, C.G., Gupta, R., Hashizume, T., Phillipson, D.W., McCloskey, J.A., 1993, Structure of the archaeal transfer RNA nucleoside G\*-15 (2-amino-4,7-dihydro- 4-oxo-7-beta-D-ribofuranosyl-1H-pyrrolo[2,3-d]pyrimidine-5-carboximi dam ide (archaeosine)). J Biol Chem 268, 10076-10086.
- Gribaldo, S., Brochier, C., 2009, Phylogeny of prokaryotes: does it exist and why should we care? Res Microbiol 160, 513-521.
- Grissa, I., Vergnaud, G., Pourcel, C., 2007a, The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. BMC Bioinformatics 8, 172.
- Grissa, I., Vergnaud, G., Pourcel, C., 2007b, CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. Nucleic Acids Res 35, W52-57.
- Grohmann, D., Hirtreiter, A., Werner, F., 2009a, Molecular mechanisms of archaeal RNA polymerase. Biochem Soc Trans 37, 12-17.
- Grohmann, D., Hirtreiter, A., Werner, F., 2009b, RNAP subunits F/E (RPB4/7) are stably associated with archaeal RNA polymerase: using fluorescence anisotropy to monitor RNAP assembly in vitro. Biochem J 421, 339-343.
- Grohmann, D., Werner, F., 2010a, Cycling through transcription with the RNA polymerase F/E (RPB4/7) complex: structure, function and evolution of archaeal RNA polymerase. Res Microbiol.
- Grohmann, D., Werner, F., 2010b, Hold on!: RNA polymerase interactions with the nascent RNA modulate transcription elongation and termination. RNA Biol 7, 310-315.
- Grosjean, H., Gaspin, C., Marck, C., Decatur, W.A., de Crecy-Lagard, V., 2008, RNomics and Modomics in the halophilic archaea Haloferax volcanii: identification of RNA modification genes. BMC Genomics 9, 470.
- Gudbergsdottir, S., Deng, L., Chen, Z., Jensen, J.V., Jensen, L.R., She, Q., Garrett, R.A., 2011, Dynamic properties of the Sulfolobus CRISPR/Cas and CRISPR/Cmr systems when challenged with vector-borne viral and plasmid genes and protospacers. Mol Microbiol 79, 35-49.
- Guillier, M., Gottesman, S., Storz, G., 2006, Modulating the outer membrane with small RNAs. Genes Dev 20, 2338-2348.
- Gunbin, K.V., Afonnikov, D.A., Kolchanov, N.A., 2009, Molecular evolution of the hyperthermophilic archaea of the Pyrococcus genus: analysis of adaptation to different environmental conditions. BMC Genomics 10, 639.
- Ha, K.S., Toulokhonov, I., Vassylyev, D.G., 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.
- Haft, D.H., Selengut, J., Mongodin, E.F., Nelson, K.E., 2005, A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. PLoS Comput Biol 1, e60.
- Haldenby, S., White, M.F., Allers, T., 2009, RecA family proteins in archaea: RadA and its cousins. Biochem Soc Trans 37, 102-107.
- Hale, C., Kleppe, K., Terns, R.M., Terns, M.P., 2008, Prokaryotic silencing (psi)RNAs in Pyrococcus furiosus. RNA 14, 2572-2579.

- Hale, C.R., Zhao, P., Olson, S., Duff, M.O., Graveley, B.R., Wells, L., Terns, R.M., Terns, M.P., 2009, RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex. Cell 139, 945-956.
- Hamma, T., Ferre-D'Amare, A.R., 2006, Pseudouridine synthases. Chem Biol 13, 1125-1135.
- Hammelmann, M., Soppa, J., 2008, Optimized generation of vectors for the construction of Haloferax volcanii deletion mutants. J Microbiol Methods 75, 201-204.
- Hannon, G.J., Rivas, F.V., Murchison, E.P., Steitz, J.A., 2006, The expanding universe of noncoding RNAs. Cold Spring Harb Symp Quant Biol 71, 551-564.
- Hasenohrl, D., Konrat, R., Blasi, U., 2011, Identification of an RNase J ortholog in Sulfolobus solfataricus: implications for 5'-to-3' directional decay and 5'-end protection of mRNA in Crenarchaeota. RNA 17, 99-107.
- Heinemann, I.U., Soll, D., Randau, L., 2010, Transfer RNA processing in archaea: unusual pathways and enzymes. FEBS Lett 584, 303-309.
- Henkin, T.M., 2008, Riboswitch RNAs: using RNA to sense cellular metabolism. Genes Dev 22, 3383-3390.
- Henneke, G., Flament, D., Hubscher, U., Querellou, J., Raffin, J.P., 2005, The hyperthermophilic euryarchaeota Pyrococcus abyssi likely requires the two DNA polymerases D and B for DNA replication. J Mol Biol 350, 53-64.
- Heppell, B., Lafontaine, D.A., 2008, Folding of the SAM aptamer is determined by the formation of a K-turn-dependent pseudoknot. Biochemistry 47, 1490-1499.
- Hering, O., Brenneis, M., Beer, J., Suess, B., Soppa, J., 2009, A novel mechanism for translation initiation operates in haloarchaea. Mol Microbiol 71, 1451-1463.
- Hirata, A., Klein, B.J., Murakami, K.S., 2008, The X-ray crystal structure of RNA polymerase from Archaea. Nature 451, 851-854.
- Hirtreiter, A., Damsma, G.E., Cheung, A.C., Klose, D., Grohmann, D., Vojnic, E., Martin, A.C., Cramer, P., Werner, F., 2010a, Spt4/5 stimulates transcription elongation through the RNA polymerase clamp coiled-coil motif. Nucleic Acids Res 38, 4040-4051.
- Hirtreiter, A., Grohmann, D., Werner, F., 2010b, Molecular mechanisms of RNA polymerase-the F/E (RPB4/7) complex is required for high processivity in vitro. Nucleic Acids Res 38, 585-596.
- Horvath, P., Coute-Monvoisin, A.C., Romero, D.A., Boyaval, P., Fremaux, C., Barrangou, R., 2009, Comparative analysis of CRISPR loci in lactic acid bacteria genomes. Int J Food Microbiol 131, 62-70.
- Horvath, P., Romero, D.A., Coute-Monvoisin, A.C., Richards, M., Deveau, H., Moineau, S., Boyaval, P., Fremaux, C., Barrangou, R., 2008, Diversity, activity, and evolution of CRISPR loci in Streptococcus thermophilus. J Bacteriol 190, 1401-1412.
- Huber, H., Hohn, M.J., Stetter, K.O., Rachel, R., 2003, The phylum Nanoarchaeota: present knowledge and future perspectives of a unique form of life. Res Microbiol 154, 165-171.
- Humbard, M.A., Miranda, H.V., Lim, J.M., Krause, D.J., Pritz, J.R., Zhou, G., Chen, S., Wells, L., Maupin-Furlow, J.A., 2010a, Ubiquitin-like small archaeal modifier proteins (SAMPs) in Haloferax volcanii. Nature 463, 54-60.
- Humbard, M.A., Reuter, C.J., Zuobi-Hasona, K., Zhou, G., Maupin-Furlow, J.A., 2010b, Phosphorylation and methylation of proteasomal proteins of the haloarcheon Haloferax volcanii. Archaea 2010, 481725.
- Huttenhofer, A., Brosius, J., Bachellerie, J.P., 2002, RNomics: identification and function of small, non-messenger RNAs. Curr Opin Chem Biol 6, 835-843.

- Huttenhofer, A., Kiefmann, M., Meier-Ewert, S., O'Brien, J., Lehrach, H., Bachellerie, J.P., Brosius, J., 2001, RNomics: an experimental approach that identifies 201 candidates for novel, small, non-messenger RNAs in mouse. EMBO J 20, 2943-2953.
- Jager, D., Sharma, C.M., Thomsen, J., Ehlers, C., Vogel, J., Schmitz, R.A., 2009, Deep sequencing analysis of the Methanosarcina mazei Go1 transcriptome in response to nitrogen availability. Proc Natl Acad Sci U S A 106, 21878-21882.
- Jansen, R., Embden, J.D., Gaastra, W., Schouls, L.M., 2002a, Identification of genes that are associated with DNA repeats in prokaryotes. Mol Microbiol 43, 1565-1575.
- Jansen, R., van Embden, J.D., Gaastra, W., Schouls, L.M., 2002b, Identification of a novel family of sequence repeats among prokaryotes. OMICS 6, 23-33.
- Jia, M.Z., Horita, S., Nagata, K., Tanokura, M., 2010, An archaeal Dim2-like protein, aDim2p, forms a ternary complex with a/eIF2 alpha and the 3' end fragment of 16S rRNA. J Mol Biol 398, 774-785.
- Kaine, B.P., 1990, Structure of the archaebacterial 7S RNA molecule. Mol Gen Genet 221, 315-321.
- Kang, M., Peterson, R., Feigon, J., 2009, Structural Insights into riboswitch control of the biosynthesis of queuosine, a modified nucleotide found in the anticodon of tRNA. Mol Cell 33, 784-790.
- Kawano, M., Reynolds, A.A., Miranda-Rios, J., Storz, G., 2005, Detection of 5'- and 3'-UTR-derived small RNAs and cis-encoded antisense RNAs in Escherichia coli. Nucleic Acids Res 33, 1040-1050.
- Kessler, A., Brinkman, A.B., van der Oost, J., Prangishvili, D., 2004, Transcription of the rodshaped viruses SIRV1 and SIRV2 of the hyperthermophilic archaeon sulfolobus. J Bacteriol 186, 7745-7753.
- Khemici, V., Poljak, L., Luisi, B.F., Carpousis, A.J., 2008, The RNase E of Escherichia coli is a membrane-binding protein. Mol Microbiol 70, 799-813.
- Khusial, P., Plaag, R., Zieve, G.W., 2005, LSm proteins form heptameric rings that bind to RNA via repeating motifs. Trends Biochem Sci 30, 522-528.
- Kifusa, M., Fukuhara, H., Hayashi, T., Kimura, M., 2005, Protein-protein interactions in the subunits of ribonuclease P in the hyperthermophilic archaeon Pyrococcus horikoshii OT3. Biosci Biotechnol Biochem 69, 1209-1212.
- Kiss-Laszlo, Z., Henry, Y., Bachellerie, J.P., Caizergues-Ferrer, M., Kiss, T., 1996, Site-specific ribose methylation of preribosomal RNA: a novel function for small nucleolar RNAs. Cell 85, 1077-1088.
- Kiss, T., 2001, Small nucleolar RNA-guided post-transcriptional modification of cellular RNAs. EMBO J 20, 3617-3622.
- Kiss, T., Fayet-Lebaron, E., Jady, B.E., 2010, Box H/ACA small ribonucleoproteins. Mol Cell 37, 597-606.
- Kjems, J., Garrett, R.A., 1988, Novel splicing mechanism for the ribosomal RNA intron in the archaebacterium Desulfurococcus mobilis. Cell 54, 693-703.
- Klein, D.J., Schmeing, T.M., Moore, P.B., Steitz, T.A., 2001, The kink-turn: a new RNA secondary structure motif. EMBO J 20, 4214-4221.
- Klein, R.J., Misulovin, Z., Eddy, S.R., 2002, Noncoding RNA genes identified in AT-rich hyperthermophiles. Proc Natl Acad Sci U S A 99, 7542-7547.
- Kobayashi, K., Kikuno, I., Kuroha, K., Saito, K., Ito, K., Ishitani, R., Inada, T., Nureki, O., 2010, Structural basis for mRNA surveillance by archaeal Pelota and GTP-bound EF1alpha complex. Proc Natl Acad Sci U S A 107, 17575-17579.
- Koch, H.G., Moser, M., Muller, M., 2003, Signal recognition particle-dependent protein targeting, universal to all kingdoms of life. Rev Physiol Biochem Pharmacol 146, 55-94.

- Koonin, E.V., Wolf, Y.I., Aravind, L., 2001, Prediction of the archaeal exosome and its connections with the proteasome and the translation and transcription machineries by a comparative-genomic approach. Genome Res 11, 240-252.
- Kunin, V., Sorek, R., Hugenholtz, P., 2007, Evolutionary conservation of sequence and secondary structures in CRISPR repeats. Genome Biol 8, R61.
- Lafontaine, D.L., Bousquet-Antonelli, C., Henry, Y., Caizergues-Ferrer, M., Tollervey, D., 1998, The box H + ACA snoRNAs carry Cbf5p, the putative rRNA pseudouridine synthase. Genes Dev 12, 527-537.
- Lafontaine, D.L., Tollervey, D., 1999, Nop58p is a common component of the box C+D snoRNPs that is required for snoRNA stability. RNA 5, 455-467.
- LaGrandeur, T.E., Huttenhofer, A., Noller, H.F., Pace, N.R., 1994, Phylogenetic comparative chemical footprint analysis of the interaction between ribonuclease P RNA and tRNA. EMBO J 13, 3945-3952.
- Lange, U., Hausner, W., 2004, Transcriptional fidelity and proofreading in Archaea and implications for the mechanism of TFS-induced RNA cleavage. Mol Microbiol 52, 1133-1143.
- Lapouge, K., Schubert, M., Allain, F.H., Haas, D., 2008, Gac/Rsm signal transduction pathway of gamma-proteobacteria: from RNA recognition to regulation of social behaviour. Mol Microbiol 67, 241-253.
- Lawrence, N.P., Richman, A., Amini, R., Altman, S., 1987, Heterologous enzyme function in Escherichia coli and the selection of genes encoding the catalytic RNA subunit of RNase P. Proc Natl Acad Sci U S A 84, 6825-6829.
- Lee, E.J., Groisman, E.A., 2010, An antisense RNA that governs the expression kinetics of a multifunctional virulence gene. Mol Microbiol 76, 1020-1033.
- Leontis, N.B., Lescoute, A., Westhof, E., 2006, The building blocks and motifs of RNA architecture. Curr Opin Struct Biol 16, 279-287.
- Leontis, N.B., Westhof, E., 2003, Analysis of RNA motifs. Curr Opin Struct Biol 13, 300-308.
- Lescoute, A., Leontis, N.B., Massire, C., Westhof, E., 2005, Recurrent structural RNA motifs, Isostericity Matrices and sequence alignments. Nucleic Acids Res 33, 2395-2409.
- Lewis, S.E., Searle, S.M., Harris, N., Gibson, M., Lyer, V., Richter, J., Wiel, C., Bayraktaroglir, L., Birney, E., Crosby, M.A., Kaminker, J.S., Matthews, B.B., Prochnik, S.E., Smithy, C.D., Tupy, J.L., Rubin, G.M., Misra, S., Mungall, C.J., Clamp, M.E., 2002, Apollo: a sequence annotation editor. Genome Biol 3, RESEARCH0082.
- Li, L., Ye, K., 2006, Crystal structure of an H/ACA box ribonucleoprotein particle. Nature 443, 302-307.
- Liang, B., Zhou, J., Kahen, E., Terns, R.M., Terns, M.P., Li, H., 2009, Structure of a functional ribonucleoprotein pseudouridine synthase bound to a substrate RNA. Nat Struct Mol Biol 16, 740-746.
- Liang, X.H., Liu, Q., Fournier, M.J., 2007, rRNA modifications in an intersubunit bridge of the ribosome strongly affect both ribosome biogenesis and activity. Mol Cell 28, 965-977.
- Lillestol, R.K., Redder, P., Garrett, R.A., Brugger, K., 2006, A putative viral defence mechanism in archaeal cells. Archaea 2, 59-72.
- Lillestol, R.K., Shah, S.A., Brugger, K., Redder, P., Phan, H., Christiansen, J., Garrett, R.A., 2009, CRISPR families of the crenarchaeal genus Sulfolobus: bidirectional transcription and dynamic properties. Mol Microbiol 72, 259-272.
- Limbach, P.A., Crain, P.F., McCloskey, J.A., 1994, Summary: the modified nucleosides of RNA. Nucleic Acids Res 22, 2183-2196.

- Loh, E., Dussurget, O., Gripenland, J., Vaitkevicius, K., Tiensuu, T., Mandin, P., Repoila, F., Buchrieser, C., Cossart, P., Johansson, J., 2009, A trans-acting riboswitch controls expression of the virulence regulator PrfA in Listeria monocytogenes. Cell 139, 770-779.
- Lopez, R., Silventoinen, V., Robinson, S., Kibria, A., Gish, W., 2003, WU-Blast2 server at the European Bioinformatics Institute. Nucleic Acids Res 31, 3795-3798.
- Lorentzen, E., Walter, P., Fribourg, S., Evguenieva-Hackenberg, E., Klug, G., Conti, E., 2005, The archaeal exosome core is a hexameric ring structure with three catalytic subunits. Nat Struct Mol Biol 12, 575-581.
- Lowe, T.M., Eddy, S.R., 1997, tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 25, 955-964.
- Lykke-Andersen, J., Aagaard, C., Semionenkov, M., Garrett, R.A., 1997, Archaeal introns: splicing, intercellular mobility and evolution. Trends Biochem Sci 22, 326-331.
- Ma, C., Simons, R.W., 1990, The IS10 antisense RNA blocks ribosome binding at the transposase translation initiation site. EMBO J 9, 1267-1274.
- Macalady, J.L., Vestling, M.M., Baumler, D., Boekelheide, N., Kaspar, C.W., Banfield, J.F., 2004, Tetraether-linked membrane monolayers in Ferroplasma spp: a key to survival in acid. Extremophiles 8, 411-419.
- Mahillon, J., Chandler, M., 1998, Insertion sequences. Microbiol Mol Biol Rev 62, 725-774. Majdalani, N., Vanderpool, C.K., Gottesman, S., 2005, Bacterial small RNA regulators. Crit Rev Biochem Mol Biol 40, 93-113.
- Makarova, K.S., Aravind, L., Galperin, M.Y., Grishin, N.V., Tatusov, R.L., Wolf, Y.I., Koonin, E.V., 1999, Comparative genomics of the Archaea (Euryarchaeota): evolution of conserved protein families, the stable core, and the variable shell. Genome Res 9, 608-628.
- Makarova, K.S., Aravind, L., Grishin, N.V., Rogozin, I.B., Koonin, E.V., 2002, A DNA repair system specific for thermophilic Archaea and bacteria predicted by genomic context analysis. Nucleic Acids Res 30, 482-496.
- Makarova, K.S., Grishin, N.V., Shabalina, S.A., Wolf, Y.I., Koonin, E.V., 2006, A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. Biol Direct 1, 7.
- Maki, K., Uno, K., Morita, T., Aiba, H., 2008, RNA, but not protein partners, is directly responsible for translational silencing by a bacterial Hfq-binding small RNA. Proc Natl Acad Sci U S A 105, 10332-10337.
- Malys, N., McCarthy, J.E., 2010, Translation initiation: variations in the mechanism can be anticipated. Cell Mol Life Sci.
- Mardanov, A.V., Ravin, N.V., Svetlitchnyi, V.A., Beletsky, A.V., Miroshnichenko, M.L., Bonch-Osmolovskaya, E.A., Skryabin, K.G., 2009, Metabolic versatility and indigenous origin of the archaeon Thermococcus sibiricus, isolated from a siberian oil reservoir, as revealed by genome analysis. Appl Environ Microbiol 75, 4580-4588.
- Margesin, R., Miteva, V., 2010, Diversity and ecology of psychrophilic microorganisms. Res Microbiol.
- Marraffini, L.A., Sontheimer, E.J., 2008, CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. Science 322, 1843-1845.
- Masse, E., Salvail, H., Desnoyers, G., Arguin, M., 2007, Small RNAs controlling iron metabolism. Curr Opin Microbiol 10, 140-145.
- Matsumi, R., Atomi, H., Driessen, A.J., van der Oost, J., 2010, Isoprenoid biosynthesis in Archaea Biochemical and evolutionary implications. Res Microbiol.

- Metzker, M.L., 2010, Sequencing technologies the next generation. Nat Rev Genet 11, 31-46.
- Mojica, F.J., Diez-Villasenor, C., Garcia-Martinez, J., Almendros, C., 2009, Short motif sequences determine the targets of the prokaryotic CRISPR defence system. Microbiology 155, 733-740.
- Mojica, F.J., Diez-Villasenor, C., Garcia-Martinez, J., Soria, E., 2005, Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. J Mol Evol 60, 174-182.
- Mojica, F.J., Ferrer, C., Juez, G., Rodriguez-Valera, F., 1995, Long stretches of short tandem repeats are present in the largest replicons of the Archaea Haloferax mediterranei and Haloferax volcanii and could be involved in replicon partitioning. Mol Microbiol 17, 85-93.
- Moll, R., Schmidtke, S., Schafer, G., 1999, Domain structure, GTP-hydrolyzing activity and 7S RNA binding of Acidianus ambivalens ffh-homologous protein suggest an SRP-like complex in archaea. Eur J Biochem 259, 441-448.
- Moore, P.B., 1999, Structural motifs in RNA. Annu Rev Biochem 68, 287-300.
- Moore, T., Zhang, Y., Fenley, M.O., Li, H., 2004, Molecular basis of box C/D RNA-protein interactions; cocrystal structure of archaeal L7Ae and a box C/D RNA. Structure 12, 807-818.
- Morita, T., Maki, K., Yagi, M., Aiba, H., 2008, Analyses of mRNA destabilization and translational inhibition mediated by Hfq-binding small RNAs. Methods Enzymol 447, 359-378.
- Muller, S., Leclerc, F., Behm-Ansmant, I., Fourmann, J.B., Charpentier, B., Branlant, C., 2008, Combined in silico and experimental identification of the Pyrococcus abyssi H/ACA sRNAs and their target sites in ribosomal RNAs. Nucleic Acids Res 36, 2459-2475.
- Myllykallio, H., Forterre, P., 2000, Mapping of a chromosome replication origin in an archaeon: response. Trends Microbiol 8, 537-539.
- Myllykallio, H., Lopez, P., Lopez-Garcia, P., Heilig, R., Saurin, W., Zivanovic, Y., Philippe, H., Forterre, P., 2000, Bacterial mode of replication with eukaryotic-like machinery in a hyperthermophilic archaeon. Science 288, 2212-2215.
- Nakagawa, S., Niimura, Y., Miura, K., Gojobori, T., 2010, Dynamic evolution of translation initiation mechanisms in prokaryotes. Proc Natl Acad Sci U S A 107, 6382-6387.
- Navarro, M.V., Oliveira, C.C., Zanchin, N.I., Guimaraes, B.G., 2008, Insights into the mechanism of progressive RNA degradation by the archaeal exosome. J Biol Chem 283, 14120-14131.
- Ng, C.L., Waterman, D.G., Antson, A.A., Ortiz-Lombardia, M., 2010, Structure of the Methanothermobacter thermautotrophicus exosome RNase PH ring. Acta Crystallogr D Biol Crystallogr 66, 522-528.
- Nicoloso, M., Qu, L.H., Michot, B., Bachellerie, J.P., 1996, Intron-encoded, antisense small nucleolar RNAs: the characterization of nine novel species points to their direct role as guides for the 2'-O-ribose methylation of rRNAs. J Mol Biol 260, 178-195.
- Nissen, P., Hansen, J., Ban, N., Moore, P.B., Steitz, T.A., 2000, The structural basis of ribosome activity in peptide bond synthesis. Science 289, 920-930.
- Nolivos, S., Carpousis, A.J., Clouet-d'Orval, B., 2005, The K-loop, a general feature of the Pyrococcus C/D guide RNAs, is an RNA structural motif related to the K-turn. Nucleic Acids Res 33, 6507-6514.
- Noll, K.M., Lapierre, P., Gogarten, J.P., Nanavati, D.M., 2008, Evolution of mal ABC transporter operons in the Thermococcales and Thermotogales. BMC Evol Biol 8, 7.

- Nomura, N., Sako, Y., Uchida, A., 1998, Molecular characterization and postsplicing fate of three introns within the single rRNA operon of the hyperthermophilic archaeon Aeropyrum pernix K1. J Bacteriol 180, 3635-3643.
- Noon, K.R., Bruenger, E., McCloskey, J.A., 1998, Posttranscriptional modifications in 16S and 23S rRNAs of the archaeal hyperthermophile Sulfolobus solfataricus. J Bacteriol 180, 2883-2888.
- Oesterhelt, D., Stoeckenius, W., 1971, Rhodopsin-like protein from the purple membrane of Halobacterium halobium. Nat New Biol 233, 149-152.
- Ohmori, H., Friedberg, E.C., Fuchs, R.P., Goodman, M.F., Hanaoka, F., Hinkle, D., Kunkel, T.A., Lawrence, C.W., Livneh, Z., Nohmi, T., Prakash, L., Prakash, S., Todo, T., Walker, G.C., Wang, Z., Woodgate, R., 2001, The Y-family of DNA polymerases. Mol Cell 8, 7-8.
- Omer, A.D., Lowe, T.M., Russell, A.G., Ebhardt, H., Eddy, S.R., Dennis, P.P., 2000, Homologs of small nucleolar RNAs in Archaea. Science 288, 517-522.
- Omer, A.D., Ziesche, S., Decatur, W.A., Fournier, M.J., Dennis, P.P., 2003, RNA-modifying machines in archaea. Mol Microbiol 48, 617-629.
- Omer, A.D., Ziesche, S., Ebhardt, H., Dennis, P.P., 2002, In vitro reconstitution and activity of a C/D box methylation guide ribonucleoprotein complex. Proc Natl Acad Sci U S A 99, 5289-5294.
- Opdyke, J.A., Kang, J.G., Storz, G., 2004, GadY, a small-RNA regulator of acid response genes in Escherichia coli. J Bacteriol 186, 6698-6705.
- Pace, N.R., 2009, Mapping the tree of life: progress and prospects. Microbiol Mol Biol Rev 73, 565-576.
- Palmer, K.L., Gilmore, M.S., 2010, Multidrug-Resistant Enterococci Lack CRISPR-cas. MBio 1.
- Papenfort, K., Vogel, J., 2009, Multiple target regulation by small noncoding RNAs rewires gene expression at the post-transcriptional level. Res Microbiol 160, 278-287.
- Peng, X., Brugger, K., Shen, B., Chen, L., She, Q., Garrett, R.A., 2003, Genus-specific protein binding to the large clusters of DNA repeats (short regularly spaced repeats) present in Sulfolobus genomes. J Bacteriol 185, 2410-2417.
- Peregrin-Alvarez, J.M., Sanford, C., Parkinson, J., 2009, The conservation and evolutionary modularity of metabolism. Genome Biol 10, R63.
- Phillips, G., Chikwana, V.M., Maxwell, A., El-Yacoubi, B., Swairjo, M.A., Iwata-Reuyl, D., de Crecy-Lagard, V., 2010, Discovery and characterization of an amidinotransferase involved in the modification of archaeal tRNA. J Biol Chem 285, 12706-12713.
- Phillips, G., El Yacoubi, B., Lyons, B., Alvarez, S., Iwata-Reuyl, D., de Crecy-Lagard, V., 2008, Biosynthesis of 7-deazaguanosine-modified tRNA nucleosides: a new role for GTP cyclohydrolase I. J Bacteriol 190, 7876-7884.
- Pietila, M.K., Roine, E., Paulin, L., Kalkkinen, N., Bamford, D.H., 2009, An ssDNA virus infecting archaea: a new lineage of viruses with a membrane envelope. Mol Microbiol 72, 307-319.
- Portillo, M.C., Gonzalez, J.M., 2009, CRISPR elements in the Thermococcales: evidence for associated horizontal gene transfer in Pyrococcus furiosus. J Appl Genet 50, 421-430.
- Portnoy, V., Evguenieva-Hackenberg, E., Klein, F., Walter, P., Lorentzen, E., Klug, G., Schuster, G., 2005, RNA polyadenylation in Archaea: not observed in Haloferax while the exosome polynucleotidylates RNA in Sulfolobus. EMBO Rep 6, 1188-1193.
- Portnoy, V., Schuster, G., 2006, RNA polyadenylation and degradation in different Archaea; roles of the exosome and RNase R. Nucleic Acids Res 34, 5923-5931.

- Pourcel, C., Salvignol, G., Vergnaud, G., 2005, CRISPR elements in Yersinia pestis acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. Microbiology 151, 653-663.
- Prangishvili, D., Forterre, P., Garrett, R.A., 2006a, Viruses of the Archaea: a unifying view. Nat Rev Microbiol 4, 837-848.
- Prangishvili, D., Garrett, R.A., Koonin, E.V., 2006b, Evolutionary genomics of archaeal viruses: unique viral genomes in the third domain of life. Virus Res 117, 52-67.
- Prieur, D., Erauso, G., Geslin, C., Lucas, S., Gaillard, M., Bidault, A., Mattenet, A.C., Rouault, K., Flament, D., Forterre, P., Le Romancer, M., 2004, Genetic elements of Thermococcales. Biochem Soc Trans 32, 184-187.
- Prieur, D., Erauso, G., Jeanthon, C., 1995, Hyperthermophilic life at deep-sea hydrothermal vents. Planet Space Sci 43, 115-122.
- Qureshi, S.A., Jackson, S.P., 1998, Sequence-specific DNA binding by the S. shibatae TFIIB homolog, TFB, and its effect on promoter strength. Mol Cell 1, 389-400.
- Randau, L., Munch, R., Hohn, M.J., Jahn, D., Soll, D., 2005a, Nanoarchaeum equitans creates functional tRNAs from separate genes for their 5'- and 3'-halves. Nature 433, 537-541.
- Randau, L., Pearson, M., Soll, D., 2005b, The complete set of tRNA species in Nanoarchaeum equitans. FEBS Lett 579, 2945-2947.
- Randau, L., Schroder, I., Soll, D., 2008, Life without RNase P. Nature 453, 120-123.
- Rashid, R., Aittaleb, M., Chen, Q., Spiegel, K., Demeler, B., Li, H., 2003, Functional requirement for symmetric assembly of archaeal box C/D small ribonucleoprotein particles. J Mol Biol 333, 295-306.
- Reichow, S.L., Hamma, T., Ferre-D'Amare, A.R., Varani, G., 2007, The structure and function of small nucleolar ribonucleoproteins. Nucleic Acids Res 35, 1452-1464.
- Reiter, W.D., Hudepohl, U., Zillig, W., 1990, Mutational analysis of an archaebacterial promoter: essential role of a TATA box for transcription efficiency and start-site selection in vitro. Proc Natl Acad Sci U S A 87, 9509-9513.
- Renalier, M.H., Joseph, N., Gaspin, C., Thebault, P., Mougin, A., 2005, The Cm56 tRNA modification in archaea is catalyzed either by a specific 2'-O-methylase, or a C/D sRNP. RNA 11, 1051-1063.
- Rivas, E., Eddy, S.R., 2001, Noncoding RNA gene detection using comparative sequence analysis. BMC Bioinformatics 2, 8.
- Robertson, C.E., Harris, J.K., Spear, J.R., Pace, N.R., 2005, Phylogenetic diversity and ecology of environmental Archaea. Curr Opin Microbiol 8, 638-642.
- Romby, P., Charpentier, E., 2010, An overview of RNAs with regulatory functions in grampositive bacteria. Cell Mol Life Sci 67, 217-237.
- Roppelt, V., Hobel, C.F., Albers, S.V., Lassek, C., Schwarz, H., Klug, G., Evguenieva-Hackenberg, E., 2010a, The archaeal exosome localizes to the membrane. FEBS Lett 584, 2791-2795.
- Roppelt, V., Klug, G., Evguenieva-Hackenberg, E., 2010b, The evolutionarily conserved subunits Rrp4 and Csl4 confer different substrate specificities to the archaeal exosome. FEBS Lett 584, 2931-2936.
- Ross, J.A., Wardle, S.J., Haniford, D.B., 2010, Tn10/IS10 transposition is downregulated at the level of transposase expression by the RNA-binding protein Hfq. Mol Microbiol 78, 607-621.
- Roth, A., Breaker, R.R., 2009, The structural and functional diversity of metabolite-binding riboswitches. Annu Rev Biochem 78, 305-334.
- Roth, A., Winkler, W.C., Regulski, E.E., Lee, B.W., Lim, J., Jona, I., Barrick, J.E., Ritwik, A., Kim, J.N., Welz, R., Iwata-Reuyl, D., Breaker, R.R., 2007, A riboswitch selective

- for the queuosine precursor preQ1 contains an unusually small aptamer domain. Nat Struct Mol Biol 14, 308-317.
- Rother, M., Krzycki, J.A., 2010, Selenocysteine, pyrrolysine, and the unique energy metabolism of methanogenic archaea. Archaea 2010.
- Salgado-Garrido, J., Bragado-Nilsson, E., Kandels-Lewis, S., Seraphin, B., 1999, Sm and Sm-like proteins assemble in two related complexes of deep evolutionary origin. EMBO J 18, 3451-3462.
- Santangelo, T.J., Cubonova, L., James, C.L., Reeve, J.N., 2007, TFB1 or TFB2 is sufficient for Thermococcus kodakaraensis viability and for basal transcription in vitro. J Mol Biol 367, 344-357.
- Santangelo, T.J., Cubonova, L., Skinner, K.M., Reeve, J.N., 2009, Archaeal intrinsic transcription termination in vivo. J Bacteriol 191, 7102-7108.
- Sato, T., Fukui, T., Atomi, H., Imanaka, T., 2003, Targeted gene disruption by homologous recombination in the hyperthermophilic archaeon Thermococcus kodakaraensis KOD1. J Bacteriol 185, 210-220.
- Schattner, P., 2002, Searching for RNA genes using base-composition statistics. Nucleic Acids Res 30, 2076-2082.
- Shah, S.A., Garrett, R.A., 2011, CRISPR/Cas and Cmr modules, mobility and evolution of adaptive immune systems. Res Microbiol 162, 27-38.
- Shah, S.A., Hansen, N.R., Garrett, R.A., 2009, Distribution of CRISPR spacer matches in viruses and plasmids of crenarchaeal acidothermophiles and implications for their inhibitory mechanism. Biochem Soc Trans 37, 23-28.
- Shahbabian, K., Jamalli, A., Zig, L., Putzer, H., 2009, RNase Y, a novel endoribonuclease, initiates riboswitch turnover in Bacillus subtilis. EMBO J 28, 3523-3533.
- Shan, S.O., Walter, P., 2005, Molecular crosstalk between the nucleotide specificity determinant of the SRP GTPase and the SRP receptor. Biochemistry 44, 6214-6222.
- Sharma, C.M., Vogel, J., 2009, Experimental approaches for the discovery and characterization of regulatory small RNA. Curr Opin Microbiol 12, 536-546.
- She, Q., Singh, R.K., Confalonieri, F., Zivanovic, Y., Allard, G., Awayez, M.J., Chan-Weiher, C.C., Clausen, I.G., Curtis, B.A., De Moors, A., Erauso, G., Fletcher, C., Gordon, P.M., Heikamp-de Jong, I., Jeffries, A.C., Kozera, C.J., Medina, N., Peng, X., Thi-Ngoc, H.P., Redder, P., Schenk, M.E., Theriault, C., Tolstrup, N., Charlebois, R.L., Doolittle, W.F., Duguet, M., Gaasterland, T., Garrett, R.A., Ragan, M.A., Sensen, C.W., Van der Oost, J., 2001, The complete genome of the crenarchaeon Sulfolobus solfataricus P2. Proc Natl Acad Sci U S A 98, 7835-7840.
- Sherratt, D.J., 2003, Bacterial chromosome dynamics. Science 301, 780-785.
- Shibata, R., Bessho, Y., Shinkai, A., Nishimoto, M., Fusatomi, E., Terada, T., Shirouzu, M., Yokoyama, S., 2007, Crystal structure and RNA-binding analysis of the archaeal transcription factor NusA. Biochem Biophys Res Commun 355, 122-128.
- Sime-Ngando, T., Lucas, S., Robin, A., Tucker, K.P., Colombet, J., Bettarel, Y., Desmond, E., Gribaldo, S., Forterre, P., Breitbart, M., Prangishvili, D., 2010, Diversity of virushost systems in hypersaline Lake Retba, Senegal. Environ Microbiol.
- Singh, S.K., Gurha, P., Gupta, R., 2008, Dynamic guide-target interactions contribute to sequential 2'-O-methylation by a unique archaeal dual guide box C/D sRNP. RNA 14, 1411-1423.
- Sittka, A., Lucchini, S., Papenfort, K., Sharma, C.M., Rolle, K., Binnewies, T.T., Hinton, J.C., Vogel, J., 2008, Deep sequencing analysis of small noncoding RNA and mRNA targets of the global post-transcriptional regulator, Hfq. PLoS Genet 4, e1000163.

- Slonczewski, J.L., Fujisawa, M., Dopson, M., Krulwich, T.A., 2009, Cytoplasmic pH measurement and homeostasis in bacteria and archaea. Adv Microb Physiol 55, 1-79, 317.
- Slupska, M.M., King, A.G., Fitz-Gibbon, S., Besemer, J., Borodovsky, M., Miller, J.H., 2001, Leaderless transcripts of the crenarchaeal hyperthermophile Pyrobaculum aerophilum. J Mol Biol 309, 347-360.
- Smith, A.M., Fuchs, R.T., Grundy, F.J., Henkin, T.M., 2010, Riboswitch RNAs: regulation of gene expression by direct monitoring of a physiological signal. RNA Biol 7, 104-110.
- Snider, J., Houry, W.A., 2008, AAA+ proteins: diversity in function, similarity in structure. Biochem Soc Trans 36, 72-77.
- Soler, N., Justome, A., Quevillon-Cheruel, S., Lorieux, F., Le Cam, E., Marguet, E., Forterre, P., 2007, The rolling-circle plasmid pTN1 from the hyperthermophilic archaeon Thermococcus nautilus. Mol Microbiol 66, 357-370.
- Soler, N., Marguet, E., Cortez, D., Desnoues, N., Keller, J., van Tilbeurgh, H., Sezonov, G., Forterre, P., 2010, Two novel families of plasmids from hyperthermophilic archaea encoding new families of replication proteins. Nucleic Acids Res 38, 5088-5104.
- Soppa, J., 1999a, Normalized nucleotide frequencies allow the definition of archaeal promoter elements for different archaeal groups and reveal base-specific TFB contacts upstream of the TATA box. Mol Microbiol 31, 1589-1592.
- Soppa, J., 1999b, Transcription initiation in Archaea: facts, factors and future aspects. Mol Microbiol 31, 1295-1305.
- Soppa, J., 2006, From genomes to function: haloarchaea as model organisms. Microbiology 152, 585-590.
- Soppa, J., Straub, J., Brenneis, M., Jellen-Ritter, A., Heyer, R., Fischer, S., Granzow, M., Voss, B., Hess, W.R., Tjaden, B., Marchfelder, A., 2009, Small RNAs of the halophilic archaeon Haloferax volcanii. Biochem Soc Trans 37, 133-136.
- Sorek, R., Kunin, V., Hugenholtz, P., 2008, CRISPR--a widespread system that provides acquired resistance against phages in bacteria and archaea. Nat Rev Microbiol 6, 181-186.
- Storz, G., Altuvia, S., Wassarman, K.M., 2005, An abundance of RNA regulators. Annu Rev Biochem 74, 199-217.
- Storz, G., Opdyke, J.A., Zhang, A., 2004, Controlling mRNA stability and translation with small, noncoding RNAs. Curr Opin Microbiol 7, 140-144.
- Straub, J., Brenneis, M., Jellen-Ritter, A., Heyer, R., Soppa, J., Marchfelder, A., 2009, Small RNAs in haloarchaea: identification, differential expression and biological function. RNA Biol 6, 281-292.
- Sugahara, J., Yachie, N., Arakawa, K., Tomita, M., 2007, In silico screening of archaeal tRNA-encoding genes having multiple introns with bulge-helix-bulge splicing motifs. RNA 13, 671-681.
- Tahirov, T.H., Makarova, K.S., Rogozin, I.B., Pavlov, Y.I., Koonin, E.V., 2009, Evolution of DNA polymerases: an inactivated polymerase-exonuclease module in Pol epsilon and a chimeric origin of eukaryotic polymerases from two classes of archaeal ancestors. Biol Direct 4, 11.
- Tanaka, H., Umehara, T., Inaka, K., Takahashi, S., Shibata, R., Bessho, Y., Sato, M., Sugiyama, S., Fusatomi, E., Terada, T., Shirouzu, M., Sano, S., Motohara, M., Kobayashi, T., Tanaka, T., Tanaka, A., Yokoyama, S., 2007, Crystallization of the archaeal transcription termination factor NusA: a significant decrease in twinning under microgravity conditions. Acta Crystallogr Sect F Struct Biol Cryst Commun 63, 69-73.

- Tang, T.H., Bachellerie, J.P., Rozhdestvensky, T., Bortolin, M.L., Huber, H., Drungowski, M., Elge, T., Brosius, J., Huttenhofer, A., 2002a, Identification of 86 candidates for small non-messenger RNAs from the archaeon Archaeoglobus fulgidus. Proc Natl Acad Sci U S A 99, 7536-7541.
- Tang, T.H., Polacek, N., Zywicki, M., Huber, H., Brugger, K., Garrett, R., Bachellerie, J.P., Huttenhofer, A., 2005, Identification of novel non-coding RNAs as potential antisense regulators in the archaeon Sulfolobus solfataricus. Mol Microbiol 55, 469-481.
- Tang, T.H., Rozhdestvensky, T.S., d'Orval, B.C., Bortolin, M.L., Huber, H., Charpentier, B., Branlant, C., Bachellerie, J.P., Brosius, J., Huttenhofer, A., 2002b, RNomics in Archaea reveals a further link between splicing of archaeal introns and rRNA processing. Nucleic Acids Res 30, 921-930.
- Thompson, L.D., Daniels, C.J., 1990, Recognition of exon-intron boundaries by the Halobacterium volcanii tRNA intron endonuclease. J Biol Chem 265, 18104-18111.
- Thore, S., Mayer, C., Sauter, C., Weeks, S., Suck, D., 2003, Crystal structures of the Pyrococcus abyssi Sm core and its complex with RNA. Common features of RNA binding in archaea and eukarya. J Biol Chem 278, 1239-1247.
- Tiedge, H., 2006, K-turn motifs in spatial RNA coding. RNA Biol 3, 133-139.
- Tocchini-Valentini, G.D., Fruscoloni, P., Tocchini-Valentini, G.P., 2009, Processing of multiple-intron-containing pretRNA. Proc Natl Acad Sci U S A 106, 20246-20251.
- Tolstrup, N., Sensen, C.W., Garrett, R.A., Clausen, I.G., 2000, Two different and highly organized mechanisms of translation initiation in the archaeon Sulfolobus solfataricus. Extremophiles 4, 175-179.
- Torarinsson, E., Klenk, H.P., Garrett, R.A., 2005, Divergent transcriptional and translational signals in Archaea. Environ Microbiol 7, 47-54.
- Tozik, I., Huang, Q., Zwieb, C., Eichler, J., 2002, Reconstitution of the signal recognition particle of the halophilic archaeon Haloferax volcanii. Nucleic Acids Res 30, 4166-4175.
- Tramonti, A., De Canio, M., De Biase, D., 2008, GadX/GadW-dependent regulation of the Escherichia coli acid fitness island: transcriptional control at the gadY-gadW divergent promoters and identification of four novel 42 bp GadX/GadW-specific binding sites. Mol Microbiol 70, 965-982.
- Tran, E.J., Zhang, X., Maxwell, E.S., 2003, Efficient RNA 2'-O-methylation requires juxtaposed and symmetrically assembled archaeal box C/D and C'/D' RNPs. EMBO J 22, 3930-3940.
- Tyson, G.W., Banfield, J.F., 2008, Rapidly evolving CRISPRs implicated in acquired resistance of microorganisms to viruses. Environ Microbiol 10, 200-207.
- Ulrih, N.P., Gmajner, D., Raspor, P., 2009, Structural and physicochemical properties of polar lipids from thermophilic archaea. Appl Microbiol Biotechnol 84, 249-260.
- van de Werken, H.J., Verhees, C.H., Akerboom, J., de Vos, W.M., van der Oost, J., 2006, Identification of a glycolytic regulon in the archaea Pyrococcus and Thermococcus. FEMS Microbiol Lett 260, 69-76.
- van den Burg, B., 2003, Extremophiles as a source for novel enzymes. Curr Opin Microbiol 6, 213-218.
- Vestergaard, G., Shah, S.A., Bize, A., Reitberger, W., Reuter, M., Phan, H., Briegel, A., Rachel, R., Garrett, R.A., Prangishvili, D., 2008, Stygiolobus rod-shaped virus and the interplay of crenarchaeal rudiviruses with the CRISPR antiviral system. J Bacteriol 190, 6837-6845.
- Vianna, M.E., Conrads, G., Gomes, B.P., Horz, H.P., 2006, Identification and quantification of archaea involved in primary endodontic infections. J Clin Microbiol 44, 1274-1282.

- Vidovic, I., Nottrott, S., Hartmuth, K., Luhrmann, R., Ficner, R., 2000, Crystal structure of the spliceosomal 15.5kD protein bound to a U4 snRNA fragment. Mol Cell 6, 1331-1342.
- Vinayak, M., Pathak, C., 2010, Queuosine modification of tRNA: its divergent role in cellular machinery. Biosci Rep 30, 135-148.
- Vogel, J., 2009, A rough guide to the non-coding RNA world of Salmonella. Mol Microbiol 71, 1-11.
- Vogel, J., Sharma, C.M., 2005, How to find small non-coding RNAs in bacteria. Biol Chem 386, 1219-1238.
- Wachter, A., 2010, Riboswitch-mediated control of gene expression in eukaryotes. RNA Biol 7, 67-76.
- Wadler, C.S., Vanderpool, C.K., 2007, A dual function for a bacterial small RNA: SgrS performs base pairing-dependent regulation and encodes a functional polypeptide. Proc Natl Acad Sci U S A 104, 20454-20459.
- Wagner, E.G., Altuvia, S., Romby, P., 2002, Antisense RNAs in bacteria and their genetic elements. Adv Genet 46, 361-398.
- Walker, S.C., Engelke, D.R., 2006, Ribonuclease P: the evolution of an ancient RNA enzyme. Crit Rev Biochem Mol Biol 41, 77-102.
- Walter, P., Klein, F., Lorentzen, E., Ilchmann, A., Klug, G., Evguenieva-Hackenberg, E., 2006, Characterization of native and reconstituted exosome complexes from the hyperthermophilic archaeon Sulfolobus solfataricus. Mol Microbiol 62, 1076-1089.
- Wassarman, K.M., Storz, G., 2000, 6S RNA regulates E. coli RNA polymerase activity. Cell 101, 613-623.
- Waters, L.S., Storz, G., 2009, Regulatory RNAs in bacteria. Cell 136, 615-628.
- Watkins, N.J., Segault, V., Charpentier, B., Nottrott, S., Fabrizio, P., Bachi, A., Wilm, M., Rosbash, M., Branlant, C., Luhrmann, R., 2000, A common core RNP structure shared between the small nucleoar box C/D RNPs and the spliceosomal U4 snRNP. Cell 103, 457-466.
- Werner, F., 2007, Structure and function of archaeal RNA polymerases. Mol Microbiol 65, 1395-1404.
- Woese, C.R., Fox, G.E., 1977, Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci U S A 74, 5088-5090.
- Woese, C.R., Kandler, O., Wheelis, M.L., 1990, Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc Natl Acad Sci U S A 87, 4576-4579.
- Wurtzel, O., Sapra, R., Chen, F., Zhu, Y., Simmons, B.A., Sorek, R., 2010, A single-base resolution map of an archaeal transcriptome. Genome Res 20, 133-141.
- Xue, S., Wang, R., Yang, F., Terns, R.M., Terns, M.P., Zhang, X., Maxwell, E.S., Li, H., 2010, Structural basis for substrate placement by an archaeal box C/D ribonucleoprotein particle. Mol Cell 39, 939-949.
- Zago, M.A., Dennis, P.P., Omer, A.D., 2005, The expanding world of small RNAs in the hyperthermophilic archaeon Sulfolobus solfataricus. Mol Microbiol 55, 1812-1828.
- Zebarjadian, Y., King, T., Fournier, M.J., Clarke, L., Carbon, J., 1999, Point mutations in yeast CBF5 can abolish in vivo pseudouridylation of rRNA. Mol Cell Biol 19, 7461-7472.
- Zeng, X., Birrien, J.L., Fouquet, Y., Cherkashov, G., Jebbar, M., Querellou, J., Oger, P., Cambon-Bonavita, M.A., Xiao, X., Prieur, D., 2009, Pyrococcus CH1, an obligate piezophilic hyperthermophile: extending the upper pressure-temperature limits for life. ISME J 3, 873-876.

- Zhang, A., Wassarman, K.M., Rosenow, C., Tjaden, B.C., Storz, G., Gottesman, S., 2003, Global analysis of small RNA and mRNA targets of Hfq. Mol Microbiol 50, 1111-1124.
- Zhou, H., Hu, H., Lai, M., 2010, Non-coding RNAs and their epigenetic regulatory mechanisms. Biol Cell 102, 645-655.
- Zivanovic, Y., Armengaud, J., Lagorce, A., Leplat, C., Guerin, P., Dutertre, M., Anthouard, V., Forterre, P., Wincker, P., Confalonieri, F., 2009, Genome analysis and genomewide proteomics of Thermococcus gammatolerans, the most radioresistant organism known amongst the Archaea. Genome Biol 10, R70.
- Zwieb, C., Eichler, J., 2002, Getting on target: the archaeal signal recognition particle. Archaea 1, 27-34.

# **Annexes**

## Annexe 1 : Tableau des prédictions bioinformatiques en fonction des régions intergéniques (IG).

Les résulats ab initio sont notées de HMM1 à HMM73, les résultats RNAsim de RNA1 à RNA75 (suite au tri manuel). Les ARNnc annotés dans GenBank sont répertoriés dans les IGR correspondantes, les critères annotés sont win (énergie libre), KT (k-turn) et hair (tige-boucle).

	IGR			prédicti	on		gène	s ARN annot	és			m	otif	
début	fin	taille	approche	début	fin	taille	nom	début	fin	taille	type	début	fin	taille
62	124	124	RNA1	53	124	72								
			HMM1	4923	5006	84	ARNt_Pro_CGG	4930	5007	78	win	4891	4996	106
4891	5794	904					RNApfQ12y	5061	5187	127	win	5022	5085	64
4691	3794	904									KT	5185	5201	17
											KT	5304	5320	17
16 318	16 468	151	RNA2	16 318	16 461	144								
49 486	49 576	91	RNA3	49 486	49 576	91								
<i>EC</i> 001	57.420	5.40	HMM2	56896	57228	333	ARNr_7S	56930	57319	390	win	56891	57158	268
56891	57438	548					sR46	57376	57438	63	KT	57343	57363	21
			HMM3	58858	58987	130	RNAabPab1	58858	58987	130	win	58762	58938	177
58667	59425	759	RNA4	59344	59422	79	PfQ13	59345	59432	88	KT	58836	58852	17
											win	59256	59400	145
122 701	123 499	799	RNA5	122 730	123 156	427	PfQ1	123 034	123 174	141				
134 579	134 829	251	RNA.1.78	134 622	134 821	200								
140 571	140 776	206	RNA.1.78	140 614	140 752	139					KT	140 755	140 771	17
			HMM4	148488	148524	37					win	149818	150029	212
			RNA6	148044	148188	145	CRISPR1							
			RNA7	148457	148597	141								
147876	150137	2262	RNA8	149316	149538	223								
			RNA.1.78	149631	149772	142								
			HMM5	149920	150087	168	ARNt_Gly_TCC	149919	149996	78				
							ARNt_Phe_GAA	150004	150080	77				
160 557	161 262	706	RNA9	160 602	160 663	62					KT	160 831	160 847	17
100 337	101 202	700									KT	161 024	161 040	17
164 353	164 473	121	RNA10	164 355	164 473	119								
197248	197535	288	HMM6	197440	197484	45								

			RNA11	197408	197477	70								
			HMM7	204785	206984	2200	ARNr_16S00	205039	206541	1503	win	204711	209872	5162
			RNA12	204858	205038	181	ARNt_Ala_TGC	206606	206683	78				
							Pab19	204812	204911					
204711	209954	5244	RNA13	206684	206808	125								
			HMM8	207025	208718	1694	ARNr_23S	206831	209848	3018	KT	207555	207575	21
			HMM9	208738	208841	104								
			HMM10	208903	209908	1006								
			RNA16	230516	230634	119	SnoR9	230449	230575	127	win	230420	230568	149
230328	230652	325					sR9_(sR19)	230464	230514	51				
230320	230032	323					sR2_(sR10)	230636	230687	52				
			HMM11	230521	230575	55	Pab21	230518	230575					
253 502	253 644	143	RNA.1.78	253 505	253 644	140								
259 798	259 917	120	RNA17	259 798	259 911	114					win	259 880	259 956	77
289023	289477	455	HMM12	288995	289065	71	Pf9	288990	289067	78				
209023	209411	433	RNA18	289023	289153	131	Pab160	288993	289064					
			HMM13	296662	296752	91	ARNt_His_GTG	296677	296753	77	win	296604	296699	96
296604	297202	599									KT	296948	296962	15
230004	291202	399									win	297128	297202	75
											win	297128	297284	157
298690	299029	340	HMM14	298926	298960	35					KT	298911	298925	15
			RNA19	318187	318426	240	sR29_(sR35)	318129	318178	50	win	318111	318332	222
318111	318433	323	HMM15	318231	318383	153	HgcF	318231	318398	168	KT	318348	318364	17
							Pab35	318 232	318 387					
337 697	337 835	139	RNA20	337 698	337 835	138								
355 801	355 883	83	RNA21	355 801	355 883	83								
202220	202622	204	RNA22	382342	382606	265	HgcG	382332	382604	273	win	382329	382564	236
382329	382622	294	HMM16	382389	382595	207	Pab40	382 390	382 599		KT	382407	382427	21
383 522	383 608	87	RNA23	383 538	383 602	65								
409 915	410 361	447	RNA.1.78	410 231	410 345	115					KT	410 190	410 205	16
410 788	411 080	293	RNA24	410 819	410 963	145					KT	411 061	411 076	16
414887	415416	530	HMM17	414997	415079	83	sR42_(sR30)	414888	414945	58	win	414886	415056	171

							ARNt_Val_TAC	415004	415080	77	win	414892	415061	170
415015	41.602.4	220	HMM18	415995	416207	213	ARNr_5Sa	415992	416113	122	win	415915	416190	276
415915	416234	320					ARNt_Asp_GTC	416131	416208	78				
419 399	419 655	257	RNA.1.78	419 497	419 648	152					hair	419 555	419 607	53
			RNA25	482 452	482 600	149	CRISPR 3				KT	482 324	482 339	16
482 140	483 504	1365	RNA.15.3	482 623	482 883	261								
			RNA26	483 043	483 192	150								
463403	464159	757	HMM19	463510	463591	82	ARNt_Val_CAC	463515	463592	78	win	463430	463539	110
403403	404139	131					sR28_(sR11)	463661	463715	55	hair	464077	464113	37
488 912	488 990	79	RNA27	488 912	488 989	78								
503955	504054	100	HMM20	503923	504025	103	ARNt_Ser_TGA	503940	504023	84	win	503955	504018	64
506286	506451	166	HMM21	506286	506421	136	ARNt_Leu_TAG	506290	506377	88	win	506286	506344	59
507916	508035	120	HMM22	507936	508014	79	ARNt_Pro_GGG	507939	508015	77	win	507916	508033	118
527607	507022	137	HMM23	527748	527797	50					win	527635	527732	98
527697	527833	137	RNA28	527697	527833	137								
541 779	542 064	286	RNA.67.1	541 982	542 057	76	sR58_(sR38)	541 774	541 823	50				
543 328	543 434	107	RNA29	543 360	543 434	75					KT	543 427	543 448	22
559 480	559 599	120	RNA30	559 492	559 599	108					hair	559 534	559 585	52
574 035	574 192	158	RNA31	574 038	574 187	150					win	574 074	574 169	96
374 033	3/4 192	130									win	574 074	574 183	110
578049	578400	352	HMM24	578135	578295	161	ARNt_Thr_TGT	578139	578215	77	win	578049	578225	177
378049	378400	332					ARNt_Pro_TGG	578219	578296	78				
579031	579271	241	HMM25	579081	579156	76	ARNt_Thr_GGT	579081	579157	77	win	579031	579108	78
600 291	600 397	107	RNA32	600 291	600 397	107					KT	600 389	600 409	21
601836	601973	138	HMM26	601870	601955	86	ARNt_Leu_TAA	601870	601956	87	win	601836	601905	70
001030	001973	130									KT	601973	601989	17
622465	622517	53	HMM27	622465	622491	27					win	628935	629216	282
630572	631056	485	HMM28	630584	630750	167	ARNt_Met_CAT_3	630583	630660	78	win	630572	630688	117
030372	031030	403					ARNt_Asn_GTT	630669	630744	76				
636639	636919	281	RNA33	636764	636919	156	SscA	636803	636898	96	win	636694	636835	142
030037	030717	201	HMM29	636814	636861	48	PfQ11b	636848	636922	75				

							sR44_(sR13)	636707	636756	50				
638552	638732	181	HMM30	638585	638674	90	ARNt_Ser_GGA	638589	638675	87	win	638552	638621	70
639636	639775	140	HMM31	639643	639668	26								
			RNA34	639639	639725	87								
644589	644826	238	HMM32	644646	644762	117	ARNt_Met_CAT_2	644645	644753	109	win	644589	644717	129
044369	044620	236	RNA35	644754	644820	67								
683 491	683 670	180	RNA36	683 491	683 640	150								
690023	690234	212	RNA37	690024	690232	209	Pf5	690083	690229	147	win	690023	690172	150
090023	090234	212	HMM33	690100	690205	106					hair	690084	690137	54
746 889	747 282	394	RNA.32.1	746 891	746 983	93								
757 510	757 822	313	RNA.1.78	757 764	757 822	59								
762 342	762 883	542	RNA.1.13	762 374	762 450	77	sR15_(sR14)	762 547	762 597	51	KT	762 710	762 731	22
784602	784852	251	HMM34	784567	784643	77								
784002	764632	231	RNA38	784607	784688	82								
790 959	791 042	84	RNA39	790 959	791 021	63								
794964	795122	159	HMM35	795016	795100	85	ARNt_Lys_CTT	795024	795101	78	win	794964	795093	130
798 283	798 652	370	RNA40	798 388	798 529	142								
800525	800819	295	HMM36	800617	800741	125	ARNr_5Sb	800607	800728	122	win	800525	800677	153
809740	010220	499	HMM37	810068	810153	86	RNAabB	809887	810256	370	win	809988	810144	157
809740	810238	499									KT	810045	810062	18
823158	823285	128	HMM38	823192	823283	92	ARNt_Leu_CAG	823197	823284	88	win	823158	823250	93
833 304	834 095	792	RNA41	834 000	834 132	133					KT	834 019	834 035	17
850 682	851 000	319	RNA42	850 691	850 837	147								
852 219	852 533	315	RNA.1.13	852 257	852 323	67								
			RNA43	862 333	862 529	197					win	862 332	862 463	132
862 245	862 713	469	PfQ12e	862 405	862 521	117					hair	862 412	862 458	47
			RNA44	862 631	862 707	77					hair	862 520	862 563	44
877339	877788	450	HMM39	877493	877666	174	RNAabPab4	877493	877666	174	win	877422	877614	193
883980	884229	250	HMM40	884020	884046	27								
			RNA45	884120	884228	109								
889 831	889 947	117	RNA.23.1	889 835	889 943	109					KT	889 818	889 834	17

			HMM41	894664	894747	84	ARNt_Thr_CGT	894672	894748	77	win	894628	894909	282
894628	895181	554	HMM42	894802	894982	181	ARNt_Ser_GCT	894809	894895	87				
							ARNt_Cys_GCA	894909	894983	75				
			RNA46	908 833	908 975	143								
908 761	909 187	427	PfQ12a	908 859	908 976	118					hair	909 096	909 149	54
											KT	909 172	909 187	16
913 306	913 453	148	RNA47	913 309	913 447	139								
0.41702	0.41002	201	HMM43	941727	941810	84	ARNt_Tyr_GTA	941729	941806	78	win	941702	941789	88
941702	941902	201									hair	941820	941886	67
055110	055270	270	HMM44	955113	955211	99	ARNt_Leu_GAG	955112	955199	88	win	954975	955140	166
955110	955379	270					sR54_(sR26)	955254	955307	54	KT	955300	955316	17
967 001	967 080	80	RNA.3.3	967 001	967 079	79								
005.016	006 107	202	RNA48	985 849	986 002	154					win	985 816	985 900	85
985 816	986 197	382									hair	985 920	985 976	57
999 719	999 939	221	RNA.24.2	999 844	999 913	70								
1 000 591	1 000 764	174	RNA.92.1	1 000 591	1 000 660	70								
			HMM45	1023449	1023530	82	ARNt_Ala_CGC	1023448	1023524	77	win	1023371	1023485	115
1023371	1024152	782					PfQ12b	1023862	1023958	97				
							RNAabPfQ12b	1023891	1024016	126				
1029893	1030018	126	HMM46	1029938	1030019	82	ARNt_Arg_CCG	1029944	1030018	75	win	1029893	1029943	51
			PfQ12A											
1041958	1042356	399	HMM47	1042253	1042325	73	sR53	1042253	1042312	60	win	1042236	1042310	75
1041936	1042330	377									KT	1042276	1042291	16
1 067 535	1 067 887	353	RNA49	1 067 710	1 067 857	148								
1 071 146	1 071 283	138	RNA.1.78	1 071 151	1 071 272	122								
1074500	1074592	93	HMM48	1074510	1074569	60	Pab91	1074509	1074576	68	win	1074413	1074540	128
1074300	1074392	93	RNA50	1074502	1074589	88					KT	1074531	1074551	21
1090026	1090461	436	HMM49	1090093	1090121	29								
1092912	1002201	370	HMM50	1093067	1093124	58	RNAabi2	1093067	1093124	58	KT	1093029	1093047	19
1092912	1093281	3/0									win	1093206	1093274	69
1 101 373	1 101 419	47	RNA51	1 101 361	1 101 414	54					win	1 101 391	1 101 742	352
1101046	1101123	78	HMM51	1101069	1101100	32					win	1100950	1101080	131

			HMM52	1103025	1103188	164	ARNt_Ile_GAT	1103024	1103101	78	win	1102992	1103116	125
440000	110100		RNA.1.13	1103205	1103345	141	ARNt_Glu_CTC	1103105	1103182	78	win	1103991	1104143	153
1102992	1104395	1404	RNA52	1104008	1104286	279					hair	1104177	1104224	48
			PfQ12c	1104114	1104179	66								
1180479	1180867	389	HMM53	1180516	1180598	83	ARNt_Gly_CCC	1180522	1180599	78	win	1180416	1180518	103
1102704	1102225	452	HMM54	1183074	1183173	100	RNAabD	1183102	1183233	132	win	1183003	1183157	155
1182784	1183235	452									hair	1183163	1183225	63
1 119 620	1 119 903	284	RNA53	1 119 660	1 119 799	140								
1 125 308	1 125 815	508	RNA.12.1	1 125 316	1 125 400	85								
1 198 159	1 198 249	91	RNA54	1 198 159	1 198 249	91								
1 213 403	1 213 657	255	RNA.1.78	1 213 403	1 213 647	245								
1261094	1261533	440	HMM55	1261324	1261407	84	ARNt_Arg_CCT	1261323	1261400	78	win	1261225	1261352	128
1 218 159	1 218 279	121	RNA55	1 218 183	1 218 247	65								
1 219 998	1 220 557	560	RNA56	1 220 201	1 220 350	150	PfQ6	1 220 213	1 220 293	81	KT	1 220 210	1 220 224	15
1 226 956	1 227 146	191	RNA.1.78	1 226 963	1 227 104	142								
1 235 360	1 235 891	532	RNA.17.1	1 235 373	1 235 514	142								
1 254 311	1 254 409	99	RNA57	1 254 313	1 254 409	97								
1 271 361	1 271 473	113	RNA58	1 271 361	1 271 472	112								
1 283 174	1 283 263	90	RNA.20.1	1 283 175	1 283 263	89								
1 302 401	1 302 721	321	RNA.13.2	1 302 411	1 302 557	147								
1215200	1216100	802	HMM56	1315530	1315699	170	ARNt_Ala_GGC	1315533	1315610	78	win	1315455	1315631	177
1315398	1316199	802					ARNt_Val_GAC	1315623	1315700	78				
							PfQ6							
1330474	1330731	258	HMM57	1330477	1330629	153	ARNt_Trp_CCA	1330476	1330624	149	win	1330474	1330538	65
1330474	1550751	236					sR40_(sR50)	1330539	1330583	45				
1 333 783	1 333 859	77	RNA59	1 333 783	1 333 858	76								
1335653	1335866	214	HMM58	1335652	1335805	154	HgcE	1335630	1335760	131	win	1335653	1335758	106
1555055	1333000	214	RNA60	1335653	1335801	149	Pab105	1 335 649	1 335 797					
1376060	1376189	130	HMM59	1376125	1376221	97	ARNt_Ser_CGA	1376125	1376211	87	win	1376060	1376170	111
1385911	1386591	681	HMM60	1386027	1386107	81	ARNt_Lys_TTT	1386026	1386103	78	win	1385925	1386071	147
1388645	1388741	97	HMM61	1388655	1388706	52	Pf2	1388644	1388707	64				

1207127	1207711	585	HMM62	1397549	1397629	81	ARNt_Glu_TTC	1397548	1397625	78	win	1397431	1397579	149
1397127	1397711	383	RNA62	1397186	1397290	105	sR11_(sR05)	1397130	1397180	51				
1412532	1413388	857	HMM63	1413210	1413299	90	ARNt_Leu_CAA	1413209	1413296	88	KT	1412731	1412747	17
1 458 076	1 458 190	115	RNA.51.1	1 458 077	1 458 190	114					hair	1 458 141	1 458 164	24
1 459 491	1 459 631	141	RNA63	1 459 510	1 459 626	117					win	1 459 608	1 459 748	141
1428186	1428363	178	HMM64	1428209	1428285	77	ARNt_Gln_CTG	1428204	1428279	76	win	1428082	1428217	136
1466417	1466780	364	HMM65	1466353	1466476	124	ARNt_Arg_GCG	1466395	1466472	78	win	1466231	1466418	188
1467507	1468869	1262	HMM66	1468719	1468803	85	ARNt_Gly_GCC	1468718	1468795	78	KT	1468625	1468641	17
1467507	1408809	1363					sR48_(sR41)	1468657	1468706	50	win	1468584	1468715	132
1 477 222	1 477 300	79	RNA.6.2	1 477 227	1 477 293	67					hair	1 477 235	1 477 289	55
1 478 375	1 478 525	151	RNA64	1 478 385	1 478 529	145					win	1 478 484	1 478 542	59
1 484 006	1 484 083	78	RNA.4.4	1 484 025	1 484 083	59								
1 499 979	1 500 163	185	RNA65	1 499 982	1 500 129	148								
1 516 033	1 516 335	303	RNA.37.1	1 516 034	1 516 180	147					hair	1 516 301	1 516 325	25
			HMM67	1522139	1522215	77	ARNt_Gln_TTG	1522145	1522220	76	win	1522080	1522152	73
1522022	1523046	1025									KT	1522085	1522105	21
1322022	1323040	1023									KT	1522516	1522530	15
											win	1522989	1523086	98
1 578 481	1 578 913	433	RNA66	1 578 824	1 578 902	79								
1 592 700	1 592 815	116	RNA.1.12	1 592 729	1 592 807	79					hair	1 592 720	1 592 748	29
1 606 729	1 606 870	142	RNA.1.78	1 606 730	1 606 823	94								
1 610 195	1 610 397	203	RNA67	1 610 197	1 610 273	77								
			RNA68	1 613 567	1 613 716	150					hair	1 612 190	1 612 233	44
							RNAabC	1 612 986	1 613 347	362	win	1 612 694	1 612 775	82
1 611 281	1 614 553	3273					PfQ3				win	1 612 911	1 613 053	143
1 011 201	1 014 333	3213									hair	1 612 994	1 613 061	68
											KT	1 613 028	1 613 048	21
											KT	1 613 180	1 613 197	18
1 629 176	1 629 552	377	RNA.1.78	1 629 285	1 629 441	157								
1640351	1640506	156	HMM68	1640395	1640476	82	ARNt_Met_CAT_1	1640404	1640481	78	win	1640319	1640416	98
10.0001	1010500										win	1640353	1640418	66

1670871	1671597	727	HMM69	1670937	1670994	58					KT	1671027	1671048	22
1685598	1685785	188	HMM70	1685684	1685766	83	ARNt_Arg_TCT	1685683	1685759	77	win	1685609	1685689	81
1691488	1691610	123	HMM71	1691496	1691542	47								
1 715 267	1 715 350	84	RNA69	1 715 287	1 715 350	64								
			RNA70	1 720 605	1 720 715	111	PfQ2	1 720 559	1 720 621	63				
1 720 546	1 720 988	443	RNA71	1 720 769	1 720 821	53								
			RNA.1.12	1 720 913	1 720 988	76								
1 730 534	1 730 819	286	RNA72	1 730 555	1 730 636	82								
1 754 418	1 754 772	355	RNA73	1 754 421	1 754 563	143								
			HMM72	1760368	1760406	39								
			RNA74	1760267	1760411	145	CRISPR4							
1759565	1762044	2480	RNA75	1760842	1761020	179								
			RNA.1.5	1761296	1761355	60								
			HMM73	1761922	1761997	76	ARNt_Arg_TCG	1761923	1761997	75	win	1761848	1761934	87

## Annexe 2 : Tableaux des oligonucléotides dans cette étude. Les séquences indiquées sont orientées du 5' au 3'.

Oligonucléotides utilisés pour les analyses en Northern Blot.

région cible	complémentaire au brin sens	complémentaire au brin reverse
RNA1	ATCCCTTTTTACTCCACCTTGAA	TTCAAGGTGGAGTAAAAAGGGAT
RNA2	ACAACAGGTAACTAAAACTTAG	CTAAGTTTTAGTTACCTGTTGT
rna2R2	CACCCCACCTCCCTTATTTT	
RNA3	TTTCTATTTATTTAGTTTTGAA	TTCAAAACTAAATAAATAGAAA
RNA4	GTCCCTTCGCCCAAACATAA	TTATGTTTGGGCGAAGGGAC
rna4D2	ATGGTGACGGCCAGCTTAA	
RNA5.1	TGTCACATTATGTCATGCAC	GTGCATGACATAATGTGACA
RNA5.2	GCTCTTATCTGAATGAACATATA	TATATGTTCATTCAGATAAGAGC
RNA10	CCAGATGTGTGGCGCAAAGC	GCTTTGCGCCACACATCTGG
RNA11	GGCGGAGTCGATGAACCTGG	CCAGGTTCATCGACTCCGCC
RNA12	CCGACGCCGCCGTCTTGACC	GGTCAAGACGGCGGCGTCGG
RNA16	TGATGTGGTGATGAGCACGC	GCGTGCTCATCACCACATCA
RNA16R2		TCCCTTCTTCACCGATAACG
pab21minus	AATCGTGAACCGGGGGTT	AACCCCGGTTCACGATT
RNA17	CACTAGAGTATCAATTTAATG	CATTAAATTGATACTCTAGTG
RNA18	TAAGCCGAGCTATGACTCCCC	GGGGAGTCATAGCTCGGCTTA
RNA19	GCTCATCGAACCCGAACACC	GGTGTTCGGGTTCGATGAGC
RNA19R2		GTCATCCGCCATAGCTCATT
pab35minus	GATTGTACCCTCGCTGTCG	CGACAGCGAGGGTACAATC
RNA20	TCCAAGACCAAGAAAATCCC	GGGATTTTCTTGGTCTTGGA
RNA21	GGAAGAAAGTCGGGGGTGA	TCACCCCGACTTTTCTTCC
RNA22	GCACCTGGTCTCACTAATCG	CGATTAGTGAGACCAGGTGC
RNA23	TAACTTGACCTAAGCATTCA	TGAATGCTTAGGTCAAGTTA
RNA24	AGACATAATTTGAATTGTCGA	TCGACAATTCAAATTATGTCT
RNA27	TCGTCCAGTTAGTTAGAATA	TATTCTAACTAACTGGACGA
RNA28	CTAGTGCCCGACGAGTACCG	CGGTACTCGTCGGGCACTAG
RNA28A	CGAGGGTTGCGATGACGT	ACGTCATCGCAACCCTCG
RNA28B	GAAATATACTAGTGCCCGAC	GTCGGGCACTAGTATATTTC
RNA33	GAAACCCGCCACAGAGGAGG	CCTCCTCTGTGGCGGGTTTC
SscA	GAGGAAGCCTCATTTGCTCG	
PfQ11b	TTTCTGCCTAACCCTACTT	
RNA34	CCCGCCGAAGTAACACAAAA	TTTTGTGTTACTTCGGCGGG
RNA35	TCCGTGAATGAGTAATTCCG	CGGAATTACTCATTCACGGA
RNA36	TCTTTTAAGATTTCTACGCTATC	GATAGCGTAGAAATCTTAAAAGA
RNA37	CGTGGTCTAACCTACTCCCG	CGGGAGTAGGTTAGACCACG
RNA38	GTAATTAACCCGCAAACCTC	GAGGTTTGCGGGTTAATTAC
RNA39	TAATAACTAAAGGTGAGTAA	TTACTCACCTTTAGTTATTA

suite		
région cible	complémentaire au brin sens	complémentaire au brin reverse
RNA41	TATTACCTCAATGATCTCACCC	GGGTGAGATCATTGAGGTAATA
RNA42	GGTACTACGAGGCCTATGGA	TCCATAGGCCTCGTAGTACC
RNA44	AAGTTCTCAAGATTGCCCCC	GGGGCAATCTTGAGAACTT
RNA45	CCCTCCAGACGTTTATAGAAG	CTTCTATAAACGTCTGGAGGG
RNA47	AGGGTTTTGTAGTAGAATGTC	GACATTCTACTACAAAACCCT
RNA49	CCTGAGCTATCCAACTAGGG	CCCTAGTTGGATAGCTCAGG
RNA50	TCTCACACCTCCGGGATACG	CGTATCCCGGAGGTGTGAGA
RNA51	AAAAATTGGGGTTAGAGGGA	TCCCTCTAACCCCAATTTTT
RNA53	AAAGCTTCCTTATGAACGGG	CCCGTTCATAAGGAAGCTTT
RNA54	CTCTCCTCTCGTTCCCAACT	AGTTGGGAACGAGAGAGAG
RNA55	TTAGATGTAAATCGAAGCGG	CCGCTTCGATTTACATCTAA
RNA56	TTAATGCACATATTTGTGCGTT	AACGCACAAATATGTGCATTAA
RNA57	CACCAGTAAATGCTTCCTA	TAGGAAGCATTTACTGGTG
RNA58	AACTCACCGGAGAATCTTCCCAT	ATGGGAAGATTCTCCGGTGAGTT
RNA59	ACCTTGACCTTGGCTAATCGGG	CCCGATTAGCCAAGGTCAAGGT
RNA60	CGGGCCTATACCACTTCAAT	ATTGAAGTGGTATAGGCCCG
RNA61	CTTATTAGCAAATTCCTATCGC	GCGATAGGAATTTGCTAATAAG
RNA62	ATGGCCGTTTCAGAATTACTGC	GCAGTAATTCTGAAACGGCCAT
RNA63	AATTCAACTACGCCCATGTAT	ATACATGGGCGTAGTTGAATT
RNA64	CGCATTCCGATTCTCGCTTATC	GATAAGCGAGAATCGGAATGCG
RNA65	GTCCAATTAGAATCTCGACTT	AAGTCGAGATTCTAATTGGAC
RNA66	CAGTTCCACACATAATGCTC	GAGCATTATGTGTGGAACTG
RNA67	ACGTCAAAACTTGAGGTGAT	ATCACCTCAAGTTTTGACGT
RNA68	GAGCCAGACACAAAGACGATAG	CTATCGTCTTTGTGTCTGGCTC
RNA69	TTCATACTTAAGAAAATTACG	CGTAATTTCTTAAGTATGAA
RNA70	GAGAGGAGCTCTCAAGGATAGAA	TTCTATCCTTGAGAGCTCCTCTC
RNA71	ACCCCTAGTTCATTTCTCC	GGAGAAATGAACTAGGGGT
RNA72	GGGATGTCTTTGTTCTGACGTAACC	GGTTACGTCAGAACAAAGACATCCC
RNA73	ATATGCTCATTAGCTTCGTTACT	AGTAACGAAGCTAATGAGCATAT
abi1D	GGTGAGGGTGAGGCTATCGGC	GGTGAGGGTGAGGCTATCGGC
abi2D	TTGTGAGGAAGAGGGAAGC	GCTTCCCTCCTCTCACAA
Pf2D	CAAATTATTATTGCGCGGCT	AGCCGCGCAATAATAATTTG
PfQ3D	AGGAACATCGTTACAACGGC	GCCGTTGTAACGATGTTCCT
Pab0571ol1	ACTCTCTGCGGTCTCCCT	
Pab0571o12	TTTCCCAATCCTCCCTTACG	
pab0921	GGTAGTAGCGAGGCAAATGC	

Oligonucléotides utilisés pour les analyses en Northern Blot des loci répétés

locus	nom oligo	complémentaire au brin sens	complémentaire au brin reverse
	49.1	CCTGAGCTATCCAACTAGGG	CCCTAGTTGGATAGCTCAGG
sRk49	49.2	CCIGAGCIAICCAACIAGGG	CCCIAGIIGGAIAGCICAGG
	49.3	ATCGGCATGATATCTCAACTTT	
sRkB	В1	CAGCTCCTTCTCCGTTTCCT	AGGAAACGGAGAAGGAGCTG
sRkC	D6C2	AGCATCATCAAACCCACCGG	GGTGGTTGAGGCCAACCTCC
BC1	I1	AGAATACATAACAAAAGTATT	GCAATTATAAGTGCGTTTCTGT
BC2	II1	GAAAAATATTATCCTCGCATGG	CCATGCGAGGATAATATTTTC
вс3	A1	AACCGTTTCCTCCTGCACCT	AGGTGCAGGAGGAAACGGTT
BC4	D6	CAGGACCCACCGGATCCTC	GAGGATCCGGTGGGTCCTG
RNA48	PfQ12D/RNA48D	GGATAACCTTCCAGCCTCCC	
	PfQ12D2	TTATCCGTTACGAGTGGCAG	
	PfQ12D+9	GGGATGGAACTTTTGCACCA	
RNA52	PfQ12C/RNA52D	GATGACCTTCCAGCCTCTCGTC	
	PfQ12C2	GAGTGGCAGGGATGGAACTC	
48/52.1	PfQ12A		GGATAATCTTCAAGCCTCTTGCT
48/52.2	PfQ12B		GGATAACCTTCCAGCCTTCTATA
48/52.3	PfQ12E/RNA43R		GATAACCTTCCATCCTCCTATCT
48/52.4	PfQ12Y		GGATAACCTTCCACCTACTCTCC

Oligonucléotides utilisés pour les analyses en C-RACE des loci répétés

	tides diffises pour les diffuilyses en	
nom	RT/PCR1	PCR2
sRk49	CCTGAGCTATCCAACTAGGG	ATTGAAAATTCAGCGTTTGC
sRkB (R1)	ATGACTAATTACTACAAGTGCGTCT	AAAGCGATAAGGGAGGGGTT
sRkC (R3)	GAACTCGCTGGTAGTGATTAGT	GGACTGCAACAAAAATAGGTGCAGG
sRkC (R2)	CAGGACCCACCGGATCCTC	TGATGCTGGTGGTTGAGGC
sRk48/52	GAACGGAAACTCCCCACATC	CCCCGAAAGCCAGCCGATG

Oligonucléotides utilisés pour les analyses en Northern Blot des cassettes CRISPR.

Casette	locus	complémentaire au brin sens	complémentaire au brin revers	
	Cr1-1	GCCCATGTATCGGTTCCAA	TTGGAACCGATACATGGGC	
CRISPR1	Cr1-22		TGCACTCGCCGTTGCGAGCT	
	Cr1-L		GTCTTATTGGAACAGGGGCC	
CRISPR2	RNA25	AATCCTGTCTAACGTGTAGG	CCTACACGTTAGACAGGATT	
	CR2SP2		CCTCCTCTCACCCACTCA	
CRISPR3	CR3SP1		GCATTTCCTACAGATTGAAC	
	Cr4-12	CCCGTGATATCCTCACCGTC	GACGGTGAGGATATCACGGG	
CRISPR4	Cr4-4	TGGAACTTCCGACCTCTGGGC	GCCCAGAGGTCGGAAGTTCCA	
	Cr4-28	GAAGGTCGCGGGTTCAAATC		
	Cr4-L	GTCTTATTGGAACAGGGCGT		

# Oligonucléotides utilisés pour l'amplification des matrices PCR pour les analyses en retard sur gel.

	oligont 5'	oligont3'
sRkB216	5' CCGGGAATTCTAATACGACTCACTATA -GGATTCGATGACGACGTACTG 3'	CAGGAGGAAACGGAGAAGGAGC
sRkB216- KTmut1	5' CCGGGAATTCTAATACGACTCACTATA -GGATTCGATGACGACGTACTG 3'	ATCAAACCCACCGGGCCCTCAACTGGTG
	GGGCCCGGTGGGTTTGATTTTGCTGGTG	CAGGAGGAACGGAGAAGGAGC
sRkB216- KTmut2	5' CCGGGAATTCTAATACGACTCACTATA -GGATTCGATGACGACGTACTG 3'	ATCAAACCCACCGGGCCCGCAACTGGTG
	GGGCCCGGTGGGTTTGATGCAGCTGGTG	CAGGAGGAACGGAGAAGGAGC
sRkC211	5'CCGGGAATTCTAATACGACTCACTATA -GGTTGGATTCGATGACGACG 3'	TCATCCGTTTCCTCCTGCACC

## Annexe 3 : <u>Recherche de séquences similaires à sRk62 dans les génomes des</u> Thermococcales étudiées

#### Pyrococcus abyssi complete genome

```
Features flanking this part of subject sequence:
  60 bp at 5' side: Translation initiation factor eIF-2, beta subunit, putative
  422 bp at 3' side: nikR nickel responsive regulator
            TGCTTTTCTCATTTTTACTTTTTCTTGAAAGTCTCGCCCTTTAGGGTGGGGATGCAGT
            Sbjct 202736
           TGCTTTTCTCATTTTTACTTTTTCTTCTGAAAGTCTCGCCCTTTAGGGTGGGGATGCAGT
            AATTCTGAAACGGCCATGCTCTTTAAGTTTTTGCAACTTAGTAA 105
Ouerv 61
            Sbjct 202796 AATTCTGAAACGGCCATGCTCTCTTAAGTTTTTGCAACTTAGTAA
Features flanking this part of subject sequence:
  44 bp at 5' side: Hypothetical protein
  186 bp at 3' side: mmdB methylmalonyl-coA decarboxylase, beta chain
Query 18
            CTTTTCTTCTGAAAGTCTCGCCCTTTAGGGTGGGGATGCAGTAATTCTGAAACGCCCAT
            Sbjct 251962
           CTTTTTCTTTTGAAAG-CTCGTCCTTTAGGGTGGGGATGCGGTAATTCTGAAACGGCCAT 252020
Ouerv 78
            GCTCTCTTAAGTTTTTGCAACTTAGTAA 105
            Sbjct 252021 GCCCTTTTAAGTTTTTGCAACTTAGTAA 252048
Features flanking this part of subject sequence:
  189 bp at 5' side: Heat shock protein/ Zn-dependent protease with chaperone ...
  2 bp at 3' side: aat alanine aminotransferase
Query 15
            TTAC-TTTTTCTTGAAAGTCTCGCCCTTTAGGGTGGGGATGCAGTAATTCTGAAACGG 73
            Sbjct 184551
           TTACATTTTCTTTTGAGAACCTCGCCCTTTAGGGTTGAGATGCAGTAATTCTGAAACAG 184492
            CC 75
Ouerv 74
Sbjct 184491 CC 184490
Thermococcus kodakarensis KOD1
Features in this part of subject sequence:
  predicted nucleic acid-binding protein, containing PIN do...
  hypothetical protein
           TTTTACTTTT-TCTTCTGAAAGTCTCGCCCTTTAGGGTGGGGATGCAGTAAT
Ouerv 13
            Sbjct 741802
           TTTTACTTTTTCTTCTGAAAGCCTCGCCCTTTAGGGCGGGGATGCAGTAAT 741853
Features in this part of subject sequence:
  predicted nucleic acid-binding protein, containing PIN do...
  hypothetical membrane protein
Query 15
            TTACTTTTT-CTTCTGAAAGTCTCGCCCTTTAGGGTGGGGATGCAGTAAT
            Sbjct 426380 TTACTTTTTCTTTTGAAAGCCTCGCCCTTCAGGGCGGGGATGCAGTAAT
                                                    426331
Features in this part of subject sequence:
  predicted nucleic acid-binding protein, containing PIN do...
            TTTACTTTTT-CTTCTGAAAGTCTCGCCCTTTAGGGTGGGGATGCAGTAAT
            Sbict 249688
Features in this part of subject sequence:
predicted nucleic acid-binding protein, containing PIN do...
            TTTTTACTTTTCTGAAAGTCTCGCCCTTTAGGGTGGGGA 54
Query 12
```

#### Thermococcus sibiricus MM 739

```
Features flanking this part of subject sequence:
  74 bp at 5' side: Archaeal ATPase
70 bp at 3' side: IS element ISTsi3 orfA-like, probably transposase
Query 1
             TGCTTTTCTCATT----TTTACTTTTTCTTGAAAGTCTCGCCCTTTAGGGTGGGGATG 56
                         TGCATTTCTCATCCTTCTTTACTTTTTCTTTTGAATGCCTTGTCCTTTAGGGCGGGGATG 255152
Sbjct 255093
Query 57
             CAGTAATTCTGAAAC 71
             Sbjct 255153 CAGTAATTCCAAAAC 255167
Features flanking this part of subject sequence:
  82 bp at 5' side: hypothetical protein
  140 bp at 3' side: hypothetical protein
              TTACTTTTT-CTTCTGAAAGTCTCGCCCTTTAGGGTGGGGATGCAGTAAT 63
Sbjct 1736015 TTACTTTTTTGAAAGCCTCGCCCTTTAAGGTGGGGATGCAGTAAT 1735966
Features in this part of subject sequence:
  hypothetical protein
             TTTTACTTTTCTTCTGAAAGTCTCGCCCTTTAGGGTGGGGA 54
Query 13
             Sbjct 465815 TTTTACTTTTTCTTTTGAAAGCCTCGCCCTTCAGGGCGGGGA 465856
Features flanking this part of subject sequence:
  135 bp at 5' side: Uracil-DNA glycosylase
29 bp at 3' side: 2-Phosphoglycerate mutase (PGM)
Query 14
              TTTACTTTT-CTTCTGAAAGTCTCGCCCTTTAGGGTGGGGA 54
              Sbjct 1420248 TTTACTTTTTTTTTTGAAAACCTCGCCCTTTAGGGTGGGGA 1420289
Features in this part of subject sequence:
  hypothetical protein
             TTACTTTTT-CTTCTGAAAGTCTCGCCCTTTAGGGTGGGGA 54
Ouerv 15
Features flanking this part of subject sequence:
  183 bp at 5' side: Inosine-5'-monophosphate dehydrogenase related protein III 21 bp at 3' side: Adenylate kinase
             TTACTTTTCTTCT-GAAAGTCTCGCCCTTTAGGGTGGGGA 54
Query 15
             Sbjct 761512 TTACTTTTCTTTTGAAAGCCTCGCCCTTTAGGGCGGGGA 761472
Features flanking this part of subject sequence:
  170 bp at 5' side: Pyrroline-5-carboxylate reductase
  4 bp at 3' side: Serine hydroxymethyltransferase
             CATTT--TTACTTTTTCTTCTGAAAGTCTCGCCCTTTAGGGTGGGGA 54
             Sbjct 581132 CATTTCCTTACTTTTCTTTTGAAAGCCTCGCCGTTTACGATGGGGA 581086
Features in this part of subject sequence:
  hypothetical protein
             TTACTTTTTCTTCTGAAAGTCTCGCCCTTTAGGGTGGG 52
Query 15
             Sbjct 806043 TTACTTTTTTTTTTGAAAGCCTCGCCCTTAAGGGCGGG 806080
Features in this part of subject sequence:
  hypothetical protein
             TTACTTTTCTTCTGAAAGTCTCGCCCTTTAGGGTGGGGATG 56
Ouerv 15
Sbjct 734343 TTGCTTTTTTTGAAAACCTTGCCCTTTAGGGCGGGGATG 734302
```

```
Features flanking this part of subject sequence:
  151 bp at 5' side: ParB-like nuclease, putative
  127 bp at 3' side: Proteasome subunit beta precursor
             TGCTT-TTCTCA-TTTTTACTTTTTCTTCTGAAAGTCTCGCCCTTTAGGGTGGGG 53
             Sbjct 810539 TGCTTGTTCTTAGTTTTTACCACCCCTTCTGAAA-TCTCGCCCTTTAGGGTGGGG 810592
 Features flanking this part of subject sequence:
  138 bp at 5' side: hypothetical protein
  138 bp at 3' side: hypothetical protein
              TTAC-TTTTCTTCTGAAAGTCTCGCCCTTTAGGGTGG 51
Query 15
              TTACATTTTTCTTTTGAAAGCCTCGCCCTTTAGGATGG 1525281
Sbjct 1525244
Features flanking this part of subject sequence:
  65 bp at 5' side: Molybdopterin biosynthesis protein moeb
  291 bp at 3' side: hypothetical protein
              ATTTTTACTTTTTCTTGAAAGTCTCGCCCTTTAGGGTGGGGATGCA 58
Ouerv 11
              Sbjct 1550331 ATTGTGACTTTTTCTTTTTCAAGCCTTGCCCTTTAGGGTAGGGAGGCA 1550284
Features flanking this part of subject sequence:
  1251 bp at 5' side: hypothetical protein
  56 bp at 3' side: Predicted transcription regulator, YpuH-like family
             TTTTCTTCTGAAAGTCTCGCCCTTTAGGGTGGGG 53
Ouerv 20
             Sbjct 652173 TTTTCTTCCGAAAACCTCACCCTTTAGGGTGGGG 652140
Features flanking this part of subject sequence:
  66 bp at 5' side: NDP-sugar dehydrogenase
91 bp at 3' side: hypothetical protein
Ouerv 28
              TGAAAGTCTCGCCCTTTAGGGTGGGGATGCAGTAA 62
              Sbjct 1818773 TGAAAGTCTTGTCCTTTAGGGTGGAGATGTAATAA 1818807
Features flanking this part of subject sequence:
  1 bp at 5' side: Indolepyruvate: ferredoxin oxidoreductase (IOR), subunit ...
69 bp at 3' side: Manganese-dependent transcription regulator
             CATTTTTACTTTTCTTCTGAAAGTCTCGCCCTTTAGGGTGGGG 53
Ouerv 10
Sbjct 122716 CATTTTTATCTTCCTTCCAAAAATCTCGCCCTTTAGGATCGG 122759
Features flanking this part of subject sequence:
  84 bp at 5' side: hypothetical protein
3 bp at 3' side: Proton/glutamate symporter, SDF family
Ouerv 6
             TTCTCATTTTTACTTTTT-CTTCTGAAAGTCTCGCCCTTTAGGGT 49
             Sbjct 222610 TTCTC-TTCTTACTTTTTCCTTTGAATGTTCCACCCTTTAGGGT 222567
Features in this part of subject sequence:
  hypothetical protein
              ACTTTT-TCTTCTGAAAGTCTCGCCCTTTAGGGTGGGG 53
Ouerv 17
              Sbjct 1693914 ACTTTTCTCTTTTTCAAGCATCGTCCTTTAGGGTGGGG 1693877
Pyrococcus horikoshii OT3
Features in this part of subject sequence:
  219aa long hypothetical 50S ribosomal protein L1
  147aa long hypothetical protein
             TTACTTTTCTTCT-GAAAGTCTCGCCCTTTAGGGTGGGGATGCAGTAAT 63
Query 15
             Sbjct 522973 TTACTTTTTTTTTTTAAAGCCTCGCCCTTTAGGGCGGGGATGCAGTAAT 523022
Features in this part of subject sequence:
```

219aa long hypothetical 50S ribosomal protein L1

Query	15	TTACTTTTT-CTTCTGAAAGTCTCGCCCTTTAGGGTGGGGATGCAGTAAT	63
Sbjct	567087	TTACTTTTTCTTTCGAAAGCCTCGCCCTTTAGGGCGGGGATGCAGTAAT	567038

### Pyrococcus furiosus DSM 3638

```
Features flanking this part of subject sequence:

157 bp at 5' side: hypothetical protein

35 bp at 3' side: hypothetical protein

Query 18 CTTTTCTCTGAAAGTCTCGCCCTTTAGGGTGGGGATGCAGTAATTCTGAA 69
```

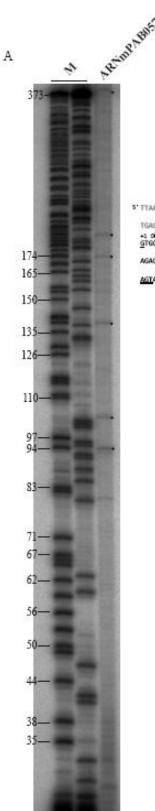
Features flanking this part of subject sequence:  $\frac{188 \text{ bp at 5' side: putative ABC transporter (ATP-binding protein)}}{60 \text{ bp at 3' side: hypothetical protein}}$ 

Sbjct 753547 CTTTTTCTTTTGAGAGCTTCGCCTTTTAGGATGAGGATGGTGTAATTCTGAA 753598

### Thermococcus onnurineus NA1

```
Features flanking this part of subject sequence:
8 bp at 5' side: hypothetical protein
6 bp at 3' side: phosphohydrolase
```

Query	15	TTACTTTTCTTCTGAAAGTCTCGCCCTTTAGGGTGGGGA			
Sbict	392720	TTACTTTTCTCTTGAGAGCCTCGCCCTTTAGGGCGGGGA	392759		





Les arrêts de rétro-transcription (indiqué par une flèche) obtenus ont été reportés sur la séquence de PAB0571 (en noir) précédée de la région intergénique (en gris). L'oligonucléotide utilisé est complémentaire à la séquence soulignée. L'arrêt de RT dont la taille est la plus longue correspondrait à l'extrémité 5' des ARN issus du locus sRkB.

## Cultures de P. abyssi

(fournies par Arnaud Hecker, IGM Orsay)

L'archaea hyperthermophile *Pyrococcus abyssi* est cultivée en condition anaérobie dans du milieu VSM (*Vent Sulfothermophiles Medium*) complémenté par du soufre (environ 2%, m/v) avant utilisation. Ce milieu est constitué d'une solution SM apportant les sels majeurs et d'une solution nutritive contenant de la résazurine (voir composition ci-dessous). La résazurine est un indicateur coloré d'oxydo-réduction qui permet de s'assurer d'une bonne anaérobiose en devenant incolore.

En général, le volume de culture utilisé est de 50 mL en fioles pénicilline stériles pour de petites cultures, ou de 250 mL en bouteilles Wheaton pour de plus gros volumes. Les fioles ou bouteilles sont fermées hermétiquement et l'anaérobiose obtenue par réalisation de plusieurs cycles de dégazage à l'azote (N<sub>2</sub>). A ce stade, les dernières traces d'oxygène présentes dans le milieu sont éliminées par l'ajout d'une solution de Na<sub>2</sub>S à 5% (environ 1/500ème, v/v). La résazurine initialement rose tout au long du processus devient incolore. L'ensemencement des milieux est réalisé à l'aide d'une souche pure de *P. abyssi* GE5 souche Orsay (dilution au 1/25°) et l'incubation réalisée sous agitation (180 rpm) dans un incubateur Infors® Haute Température à 95°C pendant 18 heures. Les cellules en phase stationnaire (environ 1.10<sup>8</sup> cellules/mL) sont alors dénombrées sous microscope photonique à l'aide d'une cellule de Thomas. Pour l'obtention de cellules en phase exponentielle, la culture est réalisée dans les mêmes conditions à partir d'une culture fraiche en phase stationnaire (dilution au 1/25°). La croissance est régulièrement suivie par comptage puis la culture stoppée lorsque cette dernière atteint une densité d'environ 4.10<sup>7</sup> cellules/mL de milieu.

Les cellules sont alors centrifugées 10 min. à 5000 g puis lavées plusieurs fois dans de l'eau de mer artificielle (ASW 1X, voir ci-dessous) afin d'éliminer le soufre qui a tendance à sédimenter et les dernières traces de milieu. Les culots cellulaires obtenus sont alors conservés à -20°C jusqu'à utilisation.

## Milieux de culture et solutions

## Milieu SM 4X

NaCl 80 g KCl 1 g NaBr 0,2 g H<sub>3</sub>BO<sub>3</sub> 0,08 g SrCl<sub>2</sub> 0,04 g C<sub>6</sub>H<sub>5</sub>Na<sub>3</sub>O<sub>7</sub>, 2 H<sub>2</sub>0 2 g H<sub>2</sub>O qsp 1 litre

Autoclaver

## Milieu VSM (Vent Sulfothermophiles Medium)

750 mL
9 g
3 mL
3 g
12 g
15 mL
3 mL
1 mL
qsp 3 litres

Ajuster le pH à 6,8 avec HCl Autoclaver

## Agent réducteur

 $\begin{array}{ccc} \text{Na}_2\text{S}, 9 \text{ H2O} & 5\text{g} \\ \text{H}_2\text{O} & 100 \text{ mL} \end{array}$ 

Adjuster le pH à 7,5 avec H<sub>2</sub>SO<sub>4</sub>

Filtrer sur membrane 0,22 μm, dégazer et conserver sous N<sub>2</sub>

## ASW 4X (Artificial See Water)

NaCl	88 g
KCl	1 g
$KH_2PO_4$	0,8 g
$H_3BO_3$	0,08 g
KI	0,08 g
SrCl <sub>2</sub> , 6 H <sub>2</sub> 0	0,04 g
$H_2O$	qsp 1 litre

Autoclaver

## EXTRACTION D'ARN TOTAL AU TRIZOL

## 1. Homogénéisation

• Homogénéiser 0,5 g de cellules (conservées à -80°C) avec 7mL de Trizol

## 2. Extraction

- Incuber les échantillons homogénéisés 10 min à 65°C en vortexant régulièrement = mélange rosé
- Faire des aliquots de 1 ml en tubes Eppendorf
- Ajouter du chloroforme saturé en eau (1/3 H2O + 2/3 CHCl3)
   200 μL CHCl3 pour 1 ml de Trizol.
   Incuber 2 à 3 minutes à RT tout en vortexant régulièrement = mélange rose laiteux
- Centrifuger 15 min à 14000 rpm à 4°C

Après centrifugation, le mélange est séparé en une phase inférieure rose, la phase de phénolchloroforme, une interphase, et une phase aqueuse supérieure sans couleur. L'ARN reste exclusivement dans la phase aqueuse. Le volume de la phase aqueuse est environ 60% du volume de Trizol utilisé pour l'homogénéisation.

- Transférer la phase aqueuse dans un nouveau tube
- Précipiter l'ARN (phase aqueuse) avec de l'isopropanol.
   500 μL d'isopropanol pour 1 ml de Trizol.
   Incuber à 10 minutes à RT
- Centrifuger 15 minutes à 14000 rpm à 4°C

Le précipité d'ARN, souvent invisible avant centrifugation, forme un culot colloïdal (texture proche de la gélatine) du côté et au fond du tube.

- Jeter le surnageant
- Lavez le culot d'ARN une fois avec de l'éthanol à 75%, 500μL d'éthanol à 75% pour 1 ml de Trizol
- Centrifuger à 7 500 g maximum 5 minutes à 4°C
- Séchage des culots **5 min** au SpeedVac (position medium, 3°étage)

## 3. Digestion de l'ADN et des protéines

• Reprendre chaque culot dans de l'eau mQ stérile170µL eau par culot

- Pour chaque tube : ajouter
  - o 20 μL **RQ1 DNase Buffer**
  - o 10 μL **RQ1 DNase**

Incuber 30 min à 37°C

- Pour chaque tube : ajouter
  - 20 μL de SDS 10%
  - 5 μL de protéinase K 20 μg/μL

Incuber 20 min à 37°C

## 4. Précipitation de l'ARN

- Pour chaque tube : ajouter
  - o 225 µL Phenol Chloroforme Isoamylalcool

Mélanger par retournement

Centrifuger 5 min à 14000 rpm à 4°C

Pooler les surnageants 2 à 2

- Pour chaque tube : ajouter
  - 20 μL NaOAcetate 7,5 M (pour 3M final)
  - o 600 µL éthanol absolu froid (2vol d'éthanol pour 1vol de prep)

Précipiter 20 min mini à -80°C

Centrifuger 20 min à 14000 rpm à 4°C

- Laver le culot avec 800 μL éthanol 70% froid Centrifuger 20 min à 14000 rpm à 4°C (10 min mini)
- Sécher les culots au SpeedVac **2 min** position medium (ou 10min à RT) Conserver les culots à **-80**°C

## 5. Dissolution de l'ARN

 Reprendre un culot dans 40 μL d'eau mQ stérile Mesurer au Nanodrop Ajuster la concentration à 10 μg/μL Aliquoter par 5 μL et conserver à -80°C

## 6. Présence/pureté des ARN

- Préparer : Gel d'agarose 1% + TBE 0,5X
- Diluer l'ARN au 1/10
  - $\circ$  1 µL ARN 1/10 + 6 µL **FUD**
  - $\circ$  4 µL ARN 1/10 + 4 µL **FUD**

Dénaturer 2 min à 90°C

- Déposer :
  - o 5 μL du Marqueur de taille ADN Smart Ladder
  - o Totalité des échantillons
  - o Migration 1h à 50 V
  - o Révélation 15 min en SybrSafe

## NORTHERN BLOT

## 1. Migration des ARN

Préparer un gel séquence 16x24cm (lxH) 6%PAGE dénaturant.

Echantillons:

Ladder  $1\mu L \sim 30000 cpm + 4\mu L FUD$ ARN (E, ES et S)  $1\mu L a 10\mu g/\mu L + 4\mu L FUD$ Dénaturer 1 min à 96°C puis refroidir dans la glace

Préchauffer le gel pendant 15min à 900V

Migration à 800V pendant 40 min

## 2. Transfert sur membrane Hybond N+ (GE Healthcare)

#### Matériel:

4 feuilles de papier Wathman ® 15x20cm (lxH) 1 membrane HybondN+ 15x20cm (lxH) TBE 0,5X (environ 6-7L) Cuve de transfert en liquide

Préparer le sandwich :

2 éponges, 2 papiers W, gel, membrane, 2 papiers W, 2 éponges

Transfert 2h à 1A dans du TBE 0,5X

A la fin du transfert, sécher la membrane puis cross-linker les ARN et le marqueur dans le four cross-linker (3<sup>ème</sup> étage) 4 min à 1200J/cm².

Exposer face à un écran PhosphorImager 1h puis révéler => le marqueur radiomarqué sert de témoin de transfert.

## 3. <u>Hybridation</u>

#### Matériel:

Membranes d'ARN

Tubes à hybridation

Four à hybridation (50°C, 4trs/min)

Tampon d'hybridation : pour 200 mL de tampon

- 50 mL Ssc 20X
- 4 mL Denhart's solution 50X
- 20 mL SDS 10%
- 1 mL d'ADN de sperme de harreng à 10mg/mL (Sigma)
- H2O Ultrafiltrée qsp 200mL

Tampon de lavage: 0,1 X Ssc/0.1%SDS

Pour 2 L: 10 mL Ssc 20X 20 mL SDS 10%

- 1) Préchauffer le four à hybridation à 50°C pendant le marquage des sondes
- 2) Mettre la membrane à pré-hybrider dans 10 ml de tampon d'hybridation 30' (minimum) à 50°C (4trs/min)
- 3) Rajouter la sonde radiomarquée à 10<sup>6</sup> cpm final
- 4) Mettre à hybrider O/N à 50°C (16h minimum)
- 5) Récolter le tampon d'hybridation dans un tube Falcon pour une utilisation ultérieure si nécessaire. (pendant ce temps, faire refroidir le four jusqu'à température ambiante RT)
- 6) Rincer la membrane dans le tube à l'aide de ~50 mL de tampon de lavage.
- 7) Mettre ~10-20 mL de tampon dans le tube contenant la membrane et laver 10 min à RT en remettant sur la roue du four à hybridation. Jeter le tampon
- 8) Renouveler une 2<sup>ème</sup> fois le lavage.
- 9) Mettre à sécher sur benchkotte pendant minimun 15 min sous lampe de paillasse.
- 10) Mettre dans une pochette transparente et exposer O/N face à un écran PhosphoRimager.

## **EXTENSION D'AMORCE**

#### Matériel:

ARN P.ab 10 μg/μL
Oligonucléotide radiomarqué dilué à 50 000 cpm/μL
AMV retro transcriptase 10U/μL Promega
Tampon AMV 5X
dNTP 10 mM
H<sub>2</sub>O UF
Acétate d'ammonium 7,5 M
Glycogène 10γ
Ethanol 100% froid

## Préparation des mix :

Ethanol 70% froid

## Mix1 pour chaque réaction (10 μL):

• 1  $\mu$ L ARN Pab 10 $\mu$ g/ $\mu$ L  $\rightarrow$  10  $\mu$ g final • 2  $\mu$ L Tampon AMV 5X  $\rightarrow$  1X final

• 1  $\mu L$  Oligo\* 50 000 cpm/ $\mu L$   $\rightarrow$  50 000 cpm final

• 6 μL H<sub>2</sub>O UF

## Mix2 pour chaque réaction (10μL):

•  $2 \mu L$  Tampon AMV 5X  $\rightarrow 1X$  final

• 2  $\mu$ L dNTP 10 mM  $\rightarrow$  1mM final (dans 20  $\mu$ L de réaction)

• 5 μL H2O UF

• 1  $\mu$ L AMV RTase 10U/ $\mu$ L  $\rightarrow$  10 U

- 1) Dénaturation du mix1 1' à 95°C puis refroidir dans la glace
- 2) Ajouter 10 µL de mix2 et incuber 1h30 à 42°C
- 3) Centrifuger brièvement
- 4) Précipiter en ajoutant par tube :
  - 1 μL de glycogène 10γ
  - 15  $\mu$ L Acétate d'ammonium 7,5 M  $\rightarrow$  3 M final
  - 50  $\mu$ L d'Ethanol 100% froid  $\rightarrow$  2,5 vol

Précipiter 20' minimum à -20°C (possibilité de le laisser O/N)

- 5) Centrifuger 20' à 14 000 rpm à 4°C, ôter le surnageant
- 6) Laver le culot avec 100 µL Ethanol 70% froid
- 7) Centrifuger 10' à 14 000 rpm à 4°C, ôter le surnageant
- 8) Répéter les étapes 6 et 7
- 9) Mettre le culot à sécher au Speedvac 10' minimum
- 10) Reprendre le culot dans 8 µL de FUD et compter la radioactivité

Migration sur gel séquence 6% acrylamide 19 :1 ; 8M Urée ; 1X TBE (16 x 22 cm) MarkerΦX174/Hinf\* 1μL + 9μL H<sub>2</sub>O UF + 10 μL FUD Dénaturer 1' à 95°C

Déposer par puits : 5 μL de MarkerΦX174/Hinf\*

4 µL de chaque échantillon

Migration en TBE 1X 800V (ou moins) pendant 30 min (BPB au ¾ du gel).

## Transcription froide in vitro

## Pour un tube réactionnel:

Buffer TRB 10 X	1X	10 μL
DTT 100 mM	10 mM	10 μL
H2O UF		qsq 100 μL
Mix rNTPs AUCG 10 mM	2 mM	20 μL
RNasine (Promega) 40 U/µL	0,8 U	2 μL
Matrice PCR	1 μg	x μL
T7 RNA pol 80 U/μL	4 U	5 μL
		100 μL

Incubation O/N à 37°C

### Traitement à la DNase

Ajouter 2 μL de DNase RQ1 (1 U/μL soit 0,02 U final) Incubation 30 min à 37°C

## Purification sur colonne P30 RNase free

#### Extraction Phénol/Chloroforme

Ajouter du phénol/chloroforme ultrapure Invitrogen vol/vol soit 100  $\mu$ L Vortex Centrifuger 5 min à 14000 rpm à 4°C Prélever le surnageant (~100 $\mu$ L)

## Précipitation Acétate d'ammonium

Ajouter 1  $\mu$ L glycogène 20 mg/mL 67  $\mu$ L d'NH4 acétate à 7,5 M pH 8 (3 M final) 420  $\mu$ L d'éthanol froid absolu (2,5 vol)

Précipiter au minimum 20 min à -20°C

Centrifuger 20 min 14000 rpm 4°C, ôter le surnageant Laver le culot avec 500  $\mu L$  d'éthanol froid 70 % Centrifuger 20 min 14000 rpm 4°C, ôter le surnageant Sécher le culot d'ARN au Speedvac (~20min) Reprendre le culot dans 20  $\mu L$  H2O UF Doser au Nanodrop

## Purification de la protéine L7Ae recombinante en condition native pour 150 ml de culture

## Préparation des billes de Co:

- Prélever 500 µl de billes Co dans un Falcon
- Centrifuger 2' à 700g, enlever le surnageant
- Ajouter 5 ml de WB
- Centrifuger 5' à 700g, enlever le surnageant
- Ajouter 5 ml de WB
- Centrifuger 5' à 700 g, enlever le surnageant. OK pour 250 μl de billes (150 ml de culture)

### **Purification:**

- Préparer 5 ml de WI<sub>1</sub> + 3.75 mg de lysozyme + ½ pastille inhibiteur de protéases
- Ajouter 2 ml au culot, homogénéiser
- Mettre dans un tube Falcon 20' à 4°C
- 3 × congélation/décongélation successives à l'azote liquide
- Sonication 5×30 sec
- Répartir dans des tubes de 1.5 ml (1 ml par tube)
- Centrifuger 20 ' à 15000g à 4°C
- Récupérer les surnageants
- Chauffer 20' à 70°C
- Centrifuger 10' à 15000g à 4°C
- Récupérer les surnageants
- Chauffer 5' à 70°C
- Centrifuger 10' à 15000g à 4°C. Garder 10 μl S
- Pooler les surnageants
- Ajouter les surnageants aux 250 µl de billes
- Sur la roue 20' à T°C ambiante
- Centrifuger 5' à 700g. Garder FT
- Ajouter 5 ml de WB, 10' sur la roue
- Centrifuger 5' à 700g. Garder L<sub>1</sub>
- Ajouter 5 ml de WB
- Centrifuger 5' à 700g. Garder L<sub>2</sub>
- Reprendre les billes avec 1 ml de WB sur la colonne
- Ajouter 5 ml de WI<sub>1</sub>. Garder L<sub>3</sub>
- Ajouter 5 ml de WI<sub>2</sub>. Garder L<sub>4</sub>
- Elutions : 2×100 μl imidazole 200 mM 2×100 μl imidazole 300 mM

5×100 µl imidazole 500 mM

## Circular-RACE

## Adapté du protocole élaboré par K. Gonzales, C. Lorenz, C. Condon, R., Shroeder et E. Charpentier

## Conversion des ARN tri-phosphate en mono-phosphate

ARN +TAP : 20 µg d'ARN totaux sont traités par 25 U de Tobacco acid Pyrophosphatase (TAP)

(Tebu-Bio), en présence de 1X de tampon TAP.

ARN –TAP : en absence de TAP Incubation de 45 min à 37°C

Extration Phénol-Chloroforme et précipitation en présence d'acétate de sodium à 3M final.

## Circularisation des ARN

1000 ng d'ARN + ou − TAP sont utilisés pour une ligature à la T4 RNA ligase selon les recommandations du fournisseur.

Incubation de 30 min à 37°C

Extration Phénol-Chloroforme et précipitation en présence d'acétate de sodium à 3M final.

#### Rétro-transcription

Voir protocole extension d'amorce avec un primer R1 spécifique.

#### **PCR**

Amplification avec la Crimson Taq à l'aide du couple de primer R1 (identique à la RT) et R2 (flanquant au primer R1)

Purification des produits de PCR sur colonne Qiaquick (Qiagen)

#### Clonage

A partir du produit de PCR purifié et à l'aide du kit TOPO TA-cloning suivant les recommandations du fournisseur (Invitrogen)

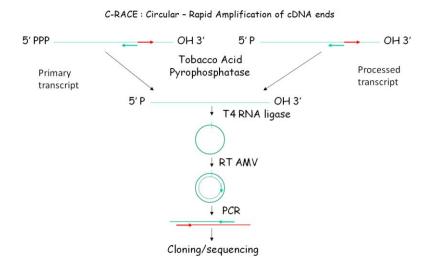
#### Séquençage

Les colonies sélectionnées sont envoyées par plaques de 96 puits.

La société MWG produit, purifie et séquence les plasmides.

### Analyse des résultats obtenus

Grâce à l'outil Blast contre le génome de P. abyssi GE5



Schématisation des différentes étapes de la C-RACE. Le Primer R1 est représenté en vert et le primer R2 en rouge.

## Résumé

Les ARN non codant (ARNnc) possèdent des fonctions vitales dans les trois domaines du vivant (eucaryotes, bactéries et archées). Les plus connus sont les acteurs de la machinerie traductionnelle (ARNr et ARNt) mais de nouveaux ARNnc régulateurs de l'expression génique ont été identifiés et caractérisés à ce jour. Des génomes d'organismes modèles eucaryotes et bactériens ont été criblés systématiquement pour la recherche d'ARNnc aussi bien par des approches biologiques que bioinformatiques. Cependant, peu d'études du même type sont parues chez les génomes d'archées. L'objectif de cette thèse est d'identifier de nouveaux ARNnc dans le génome d'une archée hyperthermophile de l'ordre des Thermococcales, Pyrococcus abyssi. Une approche bioinformatique combinant deux méthodes complémentaires a été utilisée pour cribler les régions intergéniques de la séquence génomique de cette archée. Cette recherche a abouti à la prédiction de nombreuses régions candidates dont 82 ont été retenues pour une validation expérimentale par Northern Blot. Des transcrits ont été détectés pour 24 prédictions dont 16 sont de nouveaux ARN qui n'ont jamais été décrits. Pour chaque nouvel ARN, les extrémités 5' et 3' ont été identifiées par des analyses d'extension d'amorce et en Circular RACE. Des motifs promoteurs ont été proposés pour neuf d'entre eux. Une analyse comparative des séquences étendue à l'ensemble des Thermococcales, intégrant l'étude de la conservation de séquences et la réalisation d'alignements des séquences orthologues, a amené à proposer des structures secondaires conservées pour huit ARN. L'ensemble des données acquises a permis d'apporter des arguments conduisant à émettre des hypothèses quant à la fonction de ces ARN dans la cellule. Deux d'entre eux sont capables de former des complexes ribonucléoprotéiques (RNP) avec L7Ae, protéine ribosomale composante de RNP déjà caractérisés chez les archées, suggérant que ces ARN feraient partie de complexes RNP fonctionnels.

Ce travail de thèse a donc conduit à la détection de nouveaux ARNnc qui apparaissent être de bons candidats pour réguler l'expression génique chez l'archée *P. abyssi* et de manière plus générale chez les Thermococcales.