

Université Pierre & Marie Curie - Paris 6
Ecole Doctorale Complexité du Vivant
ED 515
7 Quai Saint-Bernard, case 32
75252 PARIS cedex 05

Unité de Biologie Moléculaire du Gène chez
les Extrémophiles
Institut Pasteur, Bâtiment Fernbach
25 rue du Dr. Roux
75724 PARIS cedex 15

THESE DE DOCTORAT DE L'UNIVERSITE PIERRE ET MARIE CURIE

Spécialité
Biologie moléculaire

Présentée par
Mlle. Chloë DANIOUX

Pour obtenir le grade de
DOCTEUR de L'UNIVERSITE PIERRE ET MARIE CURIE

Régulateurs transcriptionnels chez les archées hyperthermophiles et leurs virus : analyse moléculaire, fonctionnelle et génétique

Soutenue le 17 janvier 2014 devant le jury composé de :

Pr. Vincent MARECHAL
Pr. Daniel CHARLIER
Pr. Marc NADAL
Dr. Didier FLAMENT
Dr. Evelyne RICHEL
Pr. Guennadi SEZONOV

Président du jury
Rapporteur
Rapporteur
Examinateur
Examinatrice
Directeur de thèse

Université Pierre & Marie Curie - Paris 6
Ecole Doctorale Complexité du Vivant
ED 515
7 Quai Saint-Bernard, case 32
75252 PARIS cedex 05

Unité de Biologie Moléculaire du Gène chez
les Extrémophiles
Institut Pasteur, Bâtiment Fernbach
25 rue du Dr. Roux
75724 PARIS cedex 15

THESE DE DOCTORAT DE L'UNIVERSITE PIERRE ET MARIE CURIE

Spécialité
Biologie moléculaire

Présentée par
Mlle. Chloë DANIOUX

Pour obtenir le grade de
DOCTEUR de L'UNIVERSITE PIERRE ET MARIE CURIE

Régulateurs transcriptionnels chez les archées hyperthermophiles et leurs virus : analyse moléculaire, fonctionnelle et génétique

Soutenue le 17 janvier 2014 devant le jury composé de :

Pr. Vincent MARECHAL
Pr. Daniel CHARLIER
Pr. Marc NADAL
Dr. Didier FLAMENT
Dr. Evelyne RICHEL
Pr. Guennadi SEZONOV

Président du jury
Rapporteur
Rapporteur
Examineur
Examinatrice
Directeur de thèse

Remerciements

Je tiens d'abord à remercier tous les membres de mon jury. Merci à Daniel Charlier et Marc Nadal, d'avoir accepté de relire et corriger ce manuscrit, ainsi qu'à Didier Flament et Evelyne Richet d'avoir examiné ce travail et enfin Vincent Maréchal d'avoir accepté d'être le président de mon jury.

J'aimerais remercier Patrick Forterre, chef de l'Unité de Biologie Moléculaire du Gène chez les extrémophiles et David Prangishvili, chef de laboratoire, pour leur accueil chaleureux et m'avoir permis de réaliser mon travail de thèse au sein de l'unité. Merci aussi pour votre passion communicative pour les Archées et leurs Virus!

J'aimerais tout particulièrement exprimer ma sincère gratitude et ma reconnaissance à mon directeur de thèse Guennadi Sezonov. Merci de m'avoir guidé et conseillé et merci pour tout ce que tu m'as appris pendant ces trois ans de travail.

J'aimerais également remercier chaleureusement tous les membres, passés et présents, de mon laboratoire. Merci d'avoir rendu mon quotidien si agréable tout au long de ma thèse grâce à votre gentillesse et votre bonne humeur.

Nicole, un énorme merci pour ta gentillesse et ta précieuse expertise. Te cotoyer chaque jour était un vrai plaisir. Nuno, obrigada pour ton aide et ton amitié. Qu'est-ce qu'il était bon ce hot dog à Tinglev! Carole, merci beaucoup pour ta sympathie et ton énergie ! Merci aussi à Elena et Sukhi pour avoir été des collègues si sympas. Merci à tous pour votre aide et vos conseils ainsi que tous les moments qu'on a passé ensemble.

Matthieu, mon binôme, il faudrait que je réussisse à faire chanter le virage entier pour te remercier...Ces trois ans n'auraient vraiment pas été les mêmes sans toi. Merci du fond du cœur pour tout ce que l'on a partagé ensemble.

Merci beaucoup Ana pour ton aide, ton soutien et tes conseils. Tu es la bonne fée de ce laboratoire!

J'aimerais remercier mes collaborateurs sans qui ce travail n'aurait pas vu le jour. Merci à Iñaki Guijarro pour les discussions enrichissantes que nous avons eu. Merci à Eric Le Cam, Alexandra Joubert et Olivier Piétrement pour votre travail et vos discussions intéressantes. Thank you to Qunxin She and its team for their great collaboration and for teaching me a lot about genetics.

J'aimerais remercier mes proches, famille et amis, pour votre soutien et votre présence tout au long de ma thèse. Je peux compter sur des personnes précieuses et je vous en suis éternellement reconnaissante. Parce que la vie à côté de la thèse est aussi importante et que grâce à vous, elle était réussie !

Merci énormément à vous Papa et Maman. Mes retours en Bretagne auprès de vous étaient de véritables bouffées d'air ! Vous rendre fiers de moi est la meilleure façon que j'ai trouvée pour vous témoigner tout l'amour que je vous porte. Merci Morgan d'avoir été un petit frère qui a toujours su m'encourager et me motiver. Merci à tous les trois d'avoir toujours cru en moi et d'avoir toujours été là pour moi.

Merci à tous mes amis, les Bretons, les Parisiens, les (faux) Grenoblois et les autres! Merci pour tout!

Gaëtan, je ne sais pas quels mots utiliser. Tu es un véritable ange et aussi le seul géotechnicien capable de citer Sulfolobus et la TATA-box! Merci pour ta présence cela m'a porté et motivé, que ça soit depuis un autre continent ou simplement depuis la pièce d'à côté...

Table des matières

REMERCIEMENTS	5
TABLE DES MATIERES	7
RESUME	10
ABSTRACT	11
LISTE DES ABRÉVIATIONS	12
LISTE DES FIGURES	14
LISTE DES TABLEAUX	17
PARTIE I – INTRODUCTION	19
1. ARCHEES	21
1. <i>Découverte des archées et émergence du domaine Archaea</i>	21
2. <i>Organisation générale de la cellule chez les archées</i>	22
3. <i>Diversité des archées</i>	24
a. D'extrêmophiles spécialisées à mésophiles ubiquitaires	24
b. Taxonomie des archées	26
4. <i>Les Sulfolobales, des crénarchées modèles</i>	28
2. VIRUS D'ARCHEES	31
1. <i>Famille Rudiviridae</i>	34
a. Structure des virions	34
b. Organisation du génome	35
c. Réplication de l'ADN	37
d. Transcription de l'ADN	39
e. Cycle de l'infection	44
2. <i>Famille Lipothrixviridae</i>	46
a. Structure du virion et protéines	46
b. Organisation du génome	47
3. LA TRANSCRIPTION CHEZ LES ARCHEES	49
1. <i>ARNs polymérase ADN-dépendantes dans les trois domaines du vivant</i>	49
2. <i>ARN polymérase d'archées</i>	52
a. Le noyau minimum de l'ARN polymérase archée	52
b. Les sous-unités auxiliaires	54
3. <i>L'analyse fonctionnelle de l'ARN polymérase d'archées</i>	55
a. Le complexe de pré-initiation chez les archées	55
i. Les promoteurs chez les archées	55
ii. La TATA-box binding protéine – TBP	56
iii. Le facteur de transcription B – TFB	57
b. L'initiation de la transcription	58
c. L'élongation de la transcription	59
d. La terminaison de la transcription	60
4. LA REGULATION DE LA TRANSCRIPTION CHEZ LES ARCHEES	63
1. <i>Caractérisation générale des mécanismes de régulation de l'expression génique chez les archées</i>	63
a. Régulation par des facteurs généraux de la transcription	63
b. Régulation par des facteurs liant des séquences ADN spécifiques	64
i. Les répresseurs transcriptionnels chez les archées	64
ii. Les activateurs transcriptionnels chez les archées	64
c. Régulation post-transcriptionnelle et dégradation des ARNm	65

d.	La régulation par les histones chez les archées	65
2.	<i>Deux groupes majoritaires de régulateurs transcriptionnels chez les archées – les familles HTH et RHH</i> 66	
a.	Les domaines de liaison HTH et wHTH	66
b.	Les régulateurs HTH de la famille Lrp chez les archées	67
i.	Mécanismes de contrôle et exemples de régulateurs Lrp archées.....	67
1.	Les répresseurs Lrp chez les archées	67
2.	Les activateurs Lrp chez les archées.....	69
ii.	Analyse structurale des protéines Lrp archées	73
iii.	Interaction de Lrp avec les séquences cibles	73
c.	Les régulateurs HTH de la famille MarR.....	74
d.	Les autres régulateurs HTH chez les archées.....	76
e.	La famille <i>Ribbon-Helix-Helix</i> des régulateurs transcriptionnels.....	78
i.	Présentation générale des protéines de la famille RHH	80
1.	Structure du domaine RHH	80
2.	Spécificités de liaison à l'ADN des protéines RHH et de leurs séquences cibles	81
ii.	Les régulateurs RHH chez les archées.....	83
3.	<i>Les protéines aux doigts de zinc, des régulateurs transcriptionnels putatifs chez les archées</i>	84
5.	GENETIQUE DES <i>SULFOLOBUS</i>	89
1.	<i>Marqueurs de sélection</i>	90
a.	Sélection uracile	90
b.	Sélection lactose.....	92
c.	Sélection avec l'antibiotique simvastatine	93
2.	<i>Modèles génétiques de Sulfolobus - souches réceptrices</i>	94
a.	<i>Sulfolobus solfataricus</i>	94
b.	<i>Sulfolobus acidocaldarius</i>	95
c.	<i>Sulfolobus islandicus</i>	97
3.	<i>Vecteurs de clonage de Sulfolobus</i>	98
a.	Virus SSV1	98
b.	Vecteurs plasmidiques.....	103
i.	Vecteurs basés sur le plasmide conjugatif pNOB8.....	103
ii.	Plasmides non conjugatifs	104
4.	<i>etConstruction de mutants KO chez les Sulfolobus</i>	107
a.	Génétique de <i>Sulfolobus solfataricus</i> et <i>Sulfolobus acidocaldarius</i> : construction de mutants KO	107
b.	Génétique de <i>Sulfolobus islandicus</i> : construction de mutants KO	112
i.	Remplacement Allélique (AR)	112
ii.	Remplacement du marqueur et recircularisation (MRL) et intégration du plasmide et ségrégation (PIS) 114	
iii.	Insertion de marqueur et délétion du gène cible (MID)	118
5.	<i>Système du gène rapporteur, surexpression et expression contrôlée</i>	120
6.	<i>Découvertes et progrès récents</i>	121
	PARTIE II – RESULTATS	123
1.	OBJECTIFS DE LA THESE	125
2.	DEVELOPPEMENT D'UN SYSTEME GENETIQUE POUR LA SOUCHE MODELE <i>S. ISLANDICUS LAL14/1</i>	127
1.	<i>Contexte de l'étude</i>	127
2.	<i>Etablissement de l'approche génétique pour la souche S. islandicus LAL14/1</i>	127
a.	Isolation de mutants <i>pyrEF</i> -.....	127
b.	Construction d'un mutant <i>S. islandicus LAL14/1 ΔpyrEF</i> par l'approche de remplacement allélique (AR) ...	128
3.	<i>Complémentation du phénotype Ura- par le plasmide pHZ2 chez S. islandicus LAL14/1-CD</i>	129
4.	<i>Construction d'un mutant ΔCRISPR_1</i>	129
5.	<i>Construction d'un mutant ΔtopR2</i>	133
3.	STA1, UN REGULATEUR TRANSCRIPTIONNEL CODE PAR <i>SULFOLOBUS ISLANDICUS</i>	155

1.	<i>Contexte de l'étude</i>	155
2.	<i>Le régulon hypothétique de Sta1</i>	156
3.	<i>Raisonnement génétique</i>	158
4.	<i>Construction du plasmide pSTA1-123 pour réaliser le KO du gène sta1</i>	158
4.	NOUVELLES CIBLES ET MECANISME INEDIT DE REGULATION TRANSCRIPTIONNELLE PAR UNE PROTEINE DE LA FAMILLE RHH	
	CODEE PAR LE VIRUS D'ARCHEES HYPERTHERMOPHILES SIRV1	163
1.	<i>Résumé</i>	163
2.	<i>Présentation de l'étude - introduction</i>	163
3.	<i>Matériels et Méthodes</i>	166
a.	Clonage, expression et purification de la protéine	166
b.	Oligonucléotides.....	166
c.	Expériences d'EMSA (<i>Electrophoretic gel mobility shift assays</i>)	166
d.	Agarose Gel EMSA (AG-EMSA).....	166
e.	Expériences de transcription <i>in vitro</i> (IVT).....	167
f.	<i>Footprinting</i> à la DNase I	170
g.	Méthodes bio-informatiques – Recherche <i>in silico</i> des cibles SvtR	170
h.	Microscopie	170
i.	Complexes ADN-protéines.....	170
ii.	Microscopie électronique à transmission (MET)	171
iii.	Microscopie à force atomique (AFM)	171
4.	<i>Résultats</i>	171
a.	Identification d'un site de liaison consensus de SvtR	171
b.	Identification <i>in silico</i> de tous les gènes candidats susceptibles d'être régulés par SvtR	174
c.	Les expériences EMSA révèlent une reconnaissance spécifique de certaines des nouvelles cibles prédites par la protéine SvtR	175
d.	Analyse fonctionnelle des nouvelles cibles SvtR par l'approche de transcription <i>in vitro</i>	177
e.	Le mécanisme d'action de SvtR inclut l'oligomérisation sur l'ADN.....	181
i.	Positions des sites de liaison de la protéine SvtR comparé aux régions promotrices de ses gènes cibles.....	181
ii.	La polymérisation de SvtR sur l'ADN cible lui permet de couvrir la région entre son site de fixation jusqu'à la séquence promotrice du gène <i>gp30</i>	182
iii.	L'approche de microscopie confirme l'hypothèse d'oligomérisation de SvtR	186
f.	Conclusions.....	187
5.	PO6, PREMIER REGULATEUR TRANSCRIPTIONNEL PUTATIF AU DOIGT DE ZINC CHEZ LES ARCHEES	189
1.	<i>Contexte de l'étude</i>	189
2.	<i>La protéine AFV1p06 est capable de lier l'ADN</i>	190
3.	<i>La protéine AFV1p06 lie l'ADN de façon spécifique</i>	191
4.	<i>Rôle biologique de la protéine AFV1p06</i>	193
	PARTIE III – DISCUSSION ET PERSPECTIVES.....	205
1.	ÉTABLISSEMENT DU NOUVEAU MODELE CELLULAIRE CHEZ LES <i>SULFOLOBALES</i> – <i>SULFOLOBUS ISLANDICUS</i> LAL14/1.....	208
1.	<i>Le génome de S. islandicus LAL14/1</i>	208
2.	<i>Etablissement du modèle de génétique et les premières retombées</i>	209
2.	ÉTABLISSEMENT DU MODELE VIRAL INFECTANT LES <i>SULFOLOBALES</i> – VIRUS SIRV1/2.....	213
1.	<i>Vision globale de l'activité transcriptionnelle chez SIRV1/2 à l'aide de transcriptome</i>	214
2.	<i>L'originalité du mécanisme de régulation par SvtR</i>	217
	PARTIE IV – CONCLUSIONS	223
	REFERENCES.....	227
	ANNEXES	265

Résumé

Chez les *Archaea*, tous les processus informationnels, transcription incluse, sont effectués par des protéines proches de celles des *Eukarya*. Alors que la machinerie transcriptionnelle des archées a été bien caractérisée structurellement et fonctionnellement, très peu d'informations sont disponibles sur la régulation de son activité. En travaillant à la fois avec des modèles cellulaires (crénarchée hyperthermophile *Sulfolobus islandicus*) et viraux, nous avons pu réaliser une étude approfondie de trois régulateurs transcriptionnels et mieux comprendre les mécanismes de régulation transcriptionnelle chez les archées. Au cours de cette thèse, deux régulateurs viraux, SvtR et AFV1p06, et un régulateur cellulaire, Sta1, ont été étudiés. Concernant SvtR, codé par le virus SIRV1 qui infecte *S. islandicus*, nous avons poursuivi la recherche précédente, qui avait permis de déterminer sa structure et sa fonction, en nous focalisant sur la caractérisation de l'ensemble de ses cibles dans le génome viral et sur l'étude de son mécanisme d'action. Pour cela, la séquence du site consensus reconnu par SvtR a été établie à l'aide de la mutagenèse systématique d'un de ses sites déjà caractérisés. Ce site est présent dans les promoteurs de dix gènes de SIRV1 montrant que SvtR pourrait réguler l'activité de plus de 20% des gènes viraux. Ses cibles incluent tous les gènes codant pour les protéines de la capsid virale. L'analyse fonctionnelle réalisée sur une partie des sites de liaison de SvtR a permis de démontrer qu'il s'agit d'un régulateur polyvalent agissant, selon la cible, en tant que activateur ou répresseur transcriptionnel. En prenant comme modèle le promoteur du gène *gp30*, nous avons pu démontrer par plusieurs approches que la régulation de ce promoteur inclut la polymérisation de la protéine depuis son site de liaison principal jusqu'à la *TATA-box* du promoteur. Il s'agit d'un mécanisme de régulation de transcription à distance original et inédit chez les archées. La structure de l'autre régulateur viral étudié, AFV1p06, codé par le virus AFV1 qui infecte *Acidianus hospitalis*, révèle la présence au sein de cette protéine d'un domaine en doigt de zinc C2H2, considéré jusqu'à présent comme spécifique des eucaryotes. Nous avons démontré la capacité d'AFV1p06 à se lier à l'ADN avec une préférence pour les régions riches en GC. AFV1p06 est la première *DNA binding* protéine d'archées de ce type caractérisée *in vitro*. Le troisième régulateur transcriptionnel, Sta1, est codé par le génome des *Sulfolobales*. Il est capable d'activer la transcription de gènes viraux, ainsi que du gène chromosomique *radA* en réponse à un dommage à l'ADN. Pour comprendre son rôle dans la cellule, nous avons tenté de réaliser, sans succès, un mutant *knock-out* du gène *sta1* de *S. islandicus* RYE15A, ce qui indique que le gène *sta1* serait un gène essentiel. L'étude de l'interaction hôte-virus sur le modèle *S. islandicus* LAL14/1 est un des sujets principaux de notre laboratoire. Le séquençage du génome de cette souche a ouvert la voie pour établir un système génétique. Plusieurs mutants KO de LAL14/1 (*pyrEF-*; Δ CRISPR1) ont été construits. L'impossibilité d'inactiver un autre gène candidat, *topR2*, codant pour une reverse gyrase, indique qu'il s'agit d'un gène essentiel. La construction du mutant Δ CRISPR1 est la première étape pour obtenir un dérivé de LAL14/1 dépourvu de système CRISPR, un mutant très utile pour mieux comprendre l'implication des CRISPRs dans le phénotype de résistance de LA14/1 au virus SIRV1 et leur rôle chez les archées en général. L'ensemble des résultats de cette thèse contribue à la meilleure compréhension du fonctionnement moléculaire chez les archées et leurs virus.

Abstract

In *Archaea* cells all information processes, including transcription, are performed by the *Eukarya*-like proteins. While the transcriptional machinery of archaea has been well characterized structurally and functionally, very few information concerning the regulation of its activity is available. By working with both cell (crenarchaeota *Sulfolobus islandicus*) and viral models, we have performed an in-depth study of three transcriptional regulators: two viral regulators, SvtR and AFV1p06, and a cell regulator Sta1. The obtained results allow to better understand the mechanisms of transcriptional regulation in archaea. Concerning the protein SvtR encoded by the virus SIRV1 that infects *S. islandicus*, we continued the research project that had identified its structure and function. We were focused on identification and characterization of all of SvtR targets in the viral genome and on the study of the mechanisms of regulation. For this purpose, we established the sequence of consensus site recognized by SvtR using systematic mutagenesis of one of its previously characterized binding sites. This site is present in the promoters of 10 genes meaning that SvtR may regulate the activity of more than 20% of SIRV1 genes. Its targets include all known genes encoding proteins of the viral capsid. Functional analysis of SvtR has demonstrated that, according to the target, this protein is a versatile regulator acting as transcriptional activator or repressor. Taking as a model the *gp30* gene promoter, we demonstrated by several approaches that regulation of this promoter includes the polymerization of the protein from its primary binding site towards the TATA-box. Such a mechanism of transcriptional regulation is new in archaea. Second, we performed a structural analysis of the protein AFV1p06 encoded by the virus AFV1 which infects *Acidianus hospitalis*. The structural analysis of AFV1p06 revealed the presence of a C2H2 zinc finger domain regarded hitherto as specific to eukaryotes. We demonstrated that AFV1p06 has ability to bind specifically to DNA sequences rich in GC. AFV1p06 is the first archaeal DNA binding protein with zinc finger domain characterized *in vitro*. The third transcriptional regulator, Sta1 is encoded by the genome of *Sulfolobales*. The protein RadA is able to activate the transcription of viral as well as chromosomal genes in response to DNA damage. To understand its role in the cell, we attempted, without success, to knockout the *sta1* gene in *S. islandicus* RYE15A. This result indicates that the *sta1* gene is probably essential. The strain *S. islandicus* LAL14 /1 is a model strain to study host-virus interaction in archaea. The sequencing of the genome of this strain opened the way to establish a genetic system for this model and allowed us to construct knockout mutants for several LAL14/1 genes (*pyrEF*; Δ *CRISPR1*). Our unsuccessful attempts to inactivate *topR2*, another candidate gene encoding reverse gyrase indicate that *topR2* function could be essential. The construction of the Δ *CRISPR1* mutant opens the way to obtain a derivative of LAL14/1 entirely lacking the CRISPR system. Such a mutant will be very useful for the future studies of function and role of CRISPRs in archaea in general but also will allow to verify the hypothesis of involvement of CRISPRs in the phenotype of resistance of LA14/1 to SIRV1. All the results of this thesis contribute to an improved understanding of molecular mechanisms in archaeal cells and their viruses.

Liste des abréviations

A : adénine

Alba : acetylation lowers binding affinity

ADN : acide désoxyribonucléique

AR : allelic replacement/remplacement allélique

ARN : acide ribonucléique

ARN : acide ribonucléique messenger

ARNP: ARN polymérase ADN-dépendante

ARNsn: petit acide ribonucléique nucléaire

ARNt: acide ribonucléique de transfert

ARNr : acide ribonucléique ribosomique

AvtR : acidianus viral transcriptional regulator

AG-EMSA: agarose gel electrophoretic mobility shift assay

ATP: adenosine-5'-triphosphate

BRE: transcription factor B responsive element/élément de réponse au facteur de transcription B

C : cytosine

CO₂: dioxyde de carbone

Da : daltons

db : double brin

dUTPase: dUTP pyrophosphatase

EMSA: electrophoretic mobility shift assay

G : guanine

HTH: helix-turn-helix

INR: transcription initiator

ITR: inverted terminal repeats/repetitions terminales inversées

IVT : *in vitro* transcription

KO : knock-out

Lrp : leucine-responsive regulatory protein

MALDI-TOF: matrix-assisted laser desorption ionization time-of-flight mass spectrometer

MCP : major capsid protein/protéine majeure de capside

MRL : marker replacement and looping out/remplacement du marqueur et recircularisation

MID : marker insertion and unmarked target gene deletion/insertion de marqueur et délétion du gène cible

nm : nanomètre

nt : nucléotide

NTP: nucléoside triphosphate

ORF: open reading frame/cadre de lecture ouverte

PAGE – EMSA : polyacrylamide gel electrophoresis - electrophoretic mobility shift assay

PCNA: proliferating cell nuclear antigen

p.i.: post infection

PIC: pre-initiation complex/complexe de pré-initiation de la transcription

PIS : plasmid integration and segregation/recombinaison *pop-in/pop-out*

pb : paire de bases

PCR : polymerase chain reaction

RHH: ribbon-helix-helix

RMN : résonance magnétique nucléaire

sb : simple brin

SELEX: systematic evolution of ligands by exponential enrichment

SIRV1/2 : sulfolobus islandicus rod-shaped virus 1 et 2

Sta1 : sulfolobus transcriptional activator 1

SvtR : sulfolobus virus transcription regulator

T : thymine

TBP: TATA-box binding protein/protéine de liaison à la *TATA-box*

TEC: transcription elongation complex/complexe d'élongation de la transcription

TFB: transcription factor B/facteur de transcription B

TFE: transcription factor E/facteur de transcription E

U: uridine

UAS: upstream activation sequence/séquence d'activation en amont

wt : wild type

X-gal : 5-bromo-4-chloroindol-3-yl β -D-galactopyranoside

ZNF : zinc finger

5-FOA : 5' Fluoro-orotic acid

Liste des figures

FIGURE 1 : L'ARBRE PHYLOGENETIQUE UNIVERSEL, MONTRANT LES TROIS DOMAINES DU VIVANT	21
FIGURE 2 : COMPARAISON DE LA STRUCTURE CHIMIQUE DES MEMBRANES DES <i>ARCHAEA</i> ET DES <i>BACTERIA</i>	24
FIGURE 3 : PREDOMINANCE DES ARCHEES DANS LES MILIEUX EXTREMOPHILES PAR RAPPORT AUX BACTERIES....	25
FIGURE 4 : COMPARAISON DES PROPRIETES MOLECULAIRES ENTRE LES PHYLAs D'ARCHEES	26
FIGURE 5 : UNE CELLULE <i>S. ISLANDICUS</i> TYPIQUE POSSEDE UNE FORME IRREGULIERE EN LOBES.	29
FIGURE 6 : VIRUS DES HYPERTHERMOPHILES CRENARCHAEOTA ET EURYARCHAEOTA, REPRESENTANT LES DIFFERENTES FAMILLES, SOUS MICROSCOPIE ELECTRONIQUE A TRANSMISSION	33
FIGURE 7 : RELATIONS GENOMIQUES ENTRE LES MEMBRES DE LA FAMILLE DES <i>RUDIVIRIDAE</i>	35
FIGURE 8 : LA CARTE DE TRANSCRIPTION DES VIRUS SIRV1 ET SIRV2.....	40
FIGURE 9 : EXEMPLES DE SEQUENCES NUCLEOTIDIQUES DES PROMOTEURS SIRV1 ET SIRV2.	41
FIGURE 10 : CYCLE DE L'INFECTION DU <i>SULFOLOBUS ISLANDICUS</i> ROD-SHAPED VIRUS 2 (SIRV2).	45
FIGURE 11 : LES « LIGAMENVIRALES ».....	46
FIGURE 12 : ALIGNEMENT DES GENOMES DES <i>BETALIPOTHRIXVIRIDAE</i>	47
FIGURE 13 : STRUCTURES DES ARN POLYMERASES-ADN DEPENDANTES A SOUS-UNITES MULTIPLES DE (A) <i>SACHAROMYCES CEREVISIAE</i> , (B) <i>SULFOLOBUS SHIBATAE</i> ET (C) <i>THERMUS AQUATICUS</i>	50
FIGURE 14 : ORGANISATION ET ARCHITECTURE DE L'ARN POLYMERASE.....	53
FIGURE 15 : COMPARAISON ENTRE L'ORGANISATION DES PROMOTEURS ARCHEE, EUCARYOTE ET BACTERIEN ...	56
FIGURE 16 : LE CYCLE DE LA TRANSCRIPTION CHEZ LES <i>ARCHAEA</i>	57
FIGURE 17 : L'INITIATION DE LA TRANSCRIPTION CHEZ LES <i>ARCHAEA</i>	59
FIGURE 18 : VUE D'ENSEMBLE DES MECANISMES MOLECULAIRES DE LA REGULATION DE LA TRANSCRIPTION ET DES LOCALISATIONS DE LIAISON OBSERVEES	65
FIGURE 19 : STRUCTURES CARACTERISTIQUES DES PROTEINES DE TYPE LRP CHEZ LES <i>ARCHAEA</i>	68
FIGURE 20 : MECANISME DE REGULATION TRANSCRIPTIONNEL PAR PTR2 DE <i>METHANOCOCCUS JANNASCHII</i> ...	70
FIGURE 21 : SEQUENCE DE LA REGION PROMOTRICE DE GENE <i>SS-LRPB</i> AVEC LES TROIS SITES DE RECONNAISSANCE ET LEUR DISTANCE RELATIVE DU SITE D'INITIATION DE LA TRANSCRIPTION	71
FIGURE 22 : REPRESENTATION DES DIFFERENTS COMPLEXES FORMES APRES LA LIAISON DE <i>Ss-LRPB</i> A SON PROPRE PROMOTEUR ET LES COURBURES INDUITES ASSOCIEES	72
FIGURE 23 : REPRESENTATION DES STRUCTURES CONSERVEES ET LIAISON A L'ADN DES PROTEINES RHH	81
FIGURE 24 : MECANISME DE LIAISON A L'ADN DE METJ DE LA FAMILLE RHH	82
FIGURE 25 : ALIGNEMENT MULTIPLE DE PUTATIFS REGULATEURS TRANSCRIPTIONNELS AVEC DES DOMAINES DE LIAISON A L'ADN EN DOIGT DE ZINC (C2H2) PARTAGES PAR DES VIRUS CRENARCHEES.....	84
FIGURE 26 : STRUCTURES DE PROTEINES PORTANT UN DOMAINE EN DOIGT DE ZINC (C2H2)	85
FIGURE 27 : VOIE DE LA BIOSYNTHESE DES PYRIMIDINES ET DE LA CONVERSION DU 5-FOA EN 5-FU	91

FIGURE 28 : TRANSFORMATION DU DOUBLE MUTANT <i>S. ISLANDICUS</i> E233S Δ PYREF Δ LAC _S PAR LES PLASMIDES PHZ2 ET PHZ2LAC _S	93
FIGURE 29 : CIBLE DE LA SIMVASTATINE DANS LA VOIE DE BIOSYNTHESE DE LA MEMBRANE DES ARCHEES.	94
FIGURE 30 : CARTE GENETIQUE DU VIRUS SSV1	99
FIGURE 31 : CARTE GENETIQUE DU PLASMIDE PMJ03.	102
FIGURE 32 : CARTE GENETIQUE DU PLASMIDE CONJUGATIF PNOB8	104
FIGURE 33 : CARTE GENETIQUE DU PLASMIDE PRN1	105
FIGURE 34 : CARTES GENOMIQUES DE PLASMIDES DERIVES DU PLASMIDE PHZ1 DE <i>SULFOLOBUS</i>	106
FIGURE 35 : SCHEMA DE LA STRATEGIE DE DELETION POUR LE GENE <i>AMYA</i> CHEZ <i>S. SOLFATARICUS</i>	108
FIGURE 36 : RECOMBINAISON DU PLASMIDE ADN VIA UN UNIQUE EVENEMENT DE <i>CROSSING-OVER</i> AVEC LE CHROMOSOME DE <i>S. SOLFATARICUS</i> PBL2025	109
FIGURE 37 : SCHEMA DE CONSTRUCTION D'UN MUTANT DE DELETION CHEZ <i>S. ACIDOCALDARIUS</i> UTILISANT LE PLASMIDE Δ 2PYREF	110
FIGURE 38 : DESCRIPTION DE TROIS DIFFERENTES METHODES POUR OBTENIR DES MUTANTS DE DELETIONS CHEZ <i>S. ACIDOCALDARIUS</i>	111
FIGURE 39 : MECANISME DE RECOMBINAISON MIS EN ŒUVRE PAR L'APPROCHE DE REMPLACEMENT ALLELIQUE	113
FIGURE 40 : MODULES DE RECOMBINAISON DE CHAQUE PLASMIDE <i>KNOCK-OUT</i>	113
FIGURE 41 : CONSTRUCTION DE MUTANTS DE DELETION <i>LAC_S</i> PAR L'APPROCHE MRL (A) ET PIS (B)	117
FIGURE 42 : SCHEMA DE CONSTRUCTION DE MUTANT VIA LA METHODE MID	119
FIGURE 43 : EFFICACITE DE TRANSFORMATION DE LA SOUCHE <i>S. ISLANDICUS</i> LAL14/1-CD PAR LE PLASMIDE D'EXPRESSION PHZ2	129
FIGURE 44 : CARTE GENETIQUE DE LA REGION CRISPR_1 DELETEDE PAR L'APPROCHE <i>POP-IN/POP-OUT</i> . (A) CARTE GENETIQUE DE LA REGION CRISPR_1	131
FIGURE 45 : ANALYSE PAR PCR DES MUTANTS OBTENUS	132
FIGURE 46 : REGION DU LOCUS DU GENE <i>TOPR2</i> CHEZ <i>SULFOLOBUS ISLANDICUS</i>	133
FIGURE 47 : LES PRINCIPALES ETAPES DU <i>KNOCK-OUT</i> DU GENE <i>STA1</i>	157
FIGURE 48 : CONSTRUCTION DU PLASMIDE PSTA1-123	158
FIGURE 50 : VERIFICATION PAR PCR DES COLONIES INITIALES DE TRANSFORMANTS	160
FIGURE 51 : COLORATION AU X-GAL DES COLONIES OBTENUES A PARTIR DES CULTURES DE TRANSFORMANTS	161
FIGURE 52 : STRUCTURES DES DIMERES DES PROTEINES SVTR (A) ET COPG (B) DE LA FAMILLE DES PROTEINES RHH	165
FIGURE 53 : LE REGULON CONTROLE PAR LA PROTEINE SVTR DANS LE GENOME DE SIRV1 CONNU AU DEPART DE CETTE ETUDE	172
FIGURE 54 : COMPARAISON DE SEQUENCES ENTRE LES SITES RECONNUS PAR SVTR DES PROMOTEURS DES GENES <i>GP30</i> ET <i>GP08</i>	172

FIGURE 55 : ANALYSE MUTATIONNELLE DES NUCLEOTIDES DU SITE DE LIAISON DE SVTR PAR TRANSITIONS	173
FIGURE 56 : DEUX GROUPES DE CIBLES DE SVTR DANS LE GENOME DE SIRV1 REVELES PAR L'APPROCHE EMSA	176
FIGURE 57 : SCHEMA D'EXPLICATION DE L'APPROCHE DE TRANSCRIPTION <i>IN VITRO</i>	177
FIGURE 58 : TRANSCRIPTION <i>IN VITRO</i> DE L'ACTIVITE DE SVTR SUR LES GENES CIBLES IDENTIFIES	178
FIGURE 59 : EXPRESSION DES GENES <i>GP08</i> ET <i>GP30</i> AU COURS DE L'INFECTION DE <i>S. ISLANDICUS</i> LAL14/1 PAR LE VIRUS SIRV2.....	179
FIGURE 60 : TRANSCRIPTION <i>IN VITRO</i> DE L'ACTIVITE DE SVTR SUR LE PROMOTEUR DU GENE <i>GP30</i>	180
FIGURE 61 : POLYMERISATION DE LA PROTEINE SVTR SUR L'ADN, POUR LE GENE <i>GP30</i> DE SIRV1	183
FIGURE 62 : SITES DE LIAISON SECONDAIRES IMPLIQUES DANS LA POLYMERISATION DE LA PROTEINE SVTR DANS LE CAS DE LA CIBLE <i>GP30</i>	185
FIGURE 63 : COMPLEXES SVTR-ADN SOUS MICROSCOPE ELECTRONIQUE (ABCDE) ET A FORCE ATOMIQUE (FGH).....	187
FIGURE 64 : REGULON DE SVTR DANS LE GENOME DE SIRV1	188
FIGURE 65 : VIRUS AFV1 DE LA FAMILLE DES <i>LIPOTHRIXVIRIDAE</i> , SOUS MICROSCOPIE ELECTRONIQUE A TRANSMISSION	189
FIGURE 66 : STRUCTURE DE LA PROTEINE AFV1P06	190
FIGURE 67 : LA PROTEINE AFV1P06 PEUT SE LIER A L'ADN DOUBLE BRIN ET NON SIMPLE BRIN.....	191
FIGURE 68 : COMPARAISON DU BINDING A L'ADN NE CONTENANT AUCUN NUCLEOTIDE G (POLYA) (A) ET LES CONTENANT EN EXCES (POLYG) (B) DE LA PROTEINE AFV1P06 PAR PAGE-EMSA	191
FIGURE 69 : LES OLIGONUCLEOTIDES « DSATCOMB » ET « DSCGCOMB »	192
FIGURE 71 : RECHERCHE DE CIBLES POTENTIELLES DE LA PROTEINE AFV1P06 DANS LE GENOME DU VIRUS AFV1	194
FIGURE 72 : ORGANISATION DES SEQUENCES CRISPRS CHEZ <i>S. ISLANDICUS</i> LAL14/1	211
FIGURE 73 : CARTE DU GENOME DE SIRV2 MONTRANT L'EXPRESSION TEMPORELLE DES GENES VIRAUX	214
FIGURE 74 : PROFILS D'EXPRESSION DU GENE <i>GP08</i> CODANT POUR LE REGULATEUR SVTR ET DE SES GENES CIBLES	216
FIGURE 75 : MODELES DE REGULATION DE SVTR POUR SON PROPRE GENE (AGIT COMME UN REPRESSEUR) ET UN GENE CODANT UNE PROTEINE DE CAPSIDE (AGIT COMME UN ACTIVATEUR).....	219
FIGURE 76 : MODELE DE COMPACTION DE L'ADN PAR LA PROTEINE SVTR	221

Liste des tableaux

TABLEAU 1 : COMPARAISON DES PRINCIPALES CARACTERISTIQUES DES <i>ARCHAEA</i> , <i>BACTERIA</i> ET <i>EUKARYA</i>	22
TABLEAU 2 : TAXONOMIE ET PROPRIETES DES VIRUS INFECTANT LES <i>CRENARCHAEOTA</i>	32
TABLEAU 3 : CORRESPONDANCE DES ANNOTATIONS DES GENES ENTRE LES VIRUS SIRV1 ET SIRV2	36
TABLEAU 4 : FONCTIONS CODEES PAR LE GENOME SIRV2	38
TABLEAU 5 : EXPRESSION GENIQUE DE SIRV2 PENDANT L'INFECTION DE <i>S. ISLANDICUS</i> LAL14/1.	43
TABLEAU 6 : LES SOUS-UNITES CONSERVEES DES ARN POLYMERASES ET DES FACTEURS GENERAUX DE TRANSCRIPTION DANS LES TROIS DOMAINES DE VIE	51
TABLEAU 7 : REGULATEURS TRANSCRIPTIONNELS DECRITS EXPERIMENTALEMENT CHEZ LES <i>SULFOLOBUS</i>	75
TABLEAU 8 : REGULATEURS TRANSCRIPTIONNELS <i>RIBBON-HELIX-HELIX</i> CARACTERISES ET LEURS FONCTIONS	79
TABLEAU 9 : LES OUTILS GENETIQUES DISPONIBLES CHEZ LES <i>SULFOLOBALES</i>	92
TABLEAU 10 : PERTINENCE DES DIFFERENTES SOUCHES ET ESPECES <i>SULFOLOBUS</i> COMME DESTINATAIRES POUR DES SYSTEMES GENETIQUES.....	96
TABLEAU 11 : LISTE CHRONOLOGIQUE DE TOUTES LES CONSTRUCTIONS BASIQUES DE VECTEURS NAVETTES <i>SULFOLOBUS-E. COLI</i>	101
TABLEAU 12 : APPROCHES <i>KNOCK-OUT</i> POUR L'ETUDE DES GENES DE <i>S. ISLANDICUS</i> REY15A IMPLIQUES DANS LA REPARATION ET LA REPLICATION DE L'ADN	115
TABLEAU 13 : OLIGONUCLEOTIDES UTILISES POUR LES EXPERIENCES.....	167
TABLEAU 14 : LA LISTE DE L'ENSEMBLE DES SITES DE LIAISON DE SVTR DANS LE GENOME DE SIRV1 (PREDITS OU CONFIRMES).	175
TABLEAU 15 : CIBLES RECONNUES PAR LA PROTEINE SVTR DANS LE GENOME DE SIRV1	181

Partie I – Introduction

1. Archées

1. Découverte des archées et émergence du domaine *Archaea*

Avec les *Bacteria* et les *Eukarya*, les *Archaea* forment un des trois domaines du vivant (Figure 1). La découverte de ce groupe d'organismes a été retardée de plusieurs décennies car les cellules d'archées ont l'apparence typique des cellules bactériennes et peuvent facilement être confondues avec ces dernières. Ce sont les travaux de G. Fox et C. Woese réalisés à la fin des années 70 (Woese et Fox, 1977) qui ont révélé l'existence insoupçonnée des archées. Les conclusions qui découlent de ces travaux ont marqué un tournant dans notre vision de l'évolution et de la phylogénie des organismes cellulaires. Ces chercheurs ont introduit des critères objectifs de classification du monde du vivant, fondés sur des approches moléculaires impliquant l'étude des séquences d'ADN ou d'ARN. Cela a bouleversé la classification dichotomique (eucaryote/procaryote) du vivant largement admise à cette époque. S'inspirant des idées de L. Pauling et d'E. Zuckerkandl, les pères de la phylogénie moléculaire (Zuckerkandl et Pauling, 1965), C. Woese a été le premier à développer des approches moléculaires fiables, applicables à tous les organismes cellulaires et permettant de ré-analyser sur cette base, l'arbre du vivant. Il a comparé les séquences des ARN ribosomiaux (16S chez les procaryotes et 18S chez les eucaryotes), molécule informationnelle présente chez tous les organismes cellulaires (Fox *et al.* 1977) pour établir le nouvel arbre universel du vivant. Son analyse a confirmé que les eucaryotes et les bactéries formaient deux domaines distincts et a mis en évidence un troisième groupe incluant des organismes procaryotiques précédemment classés d'après l'organisation morphologique de leurs cellules parmi les bactéries. Il s'agit de méthanogènes, et certains représentants des halophiles et des hyperthermophiles. Le métabolisme particulier des organismes de ce nouveau groupe, rappelant par certains paramètres les conditions de la Terre primitive, ont conduit au départ (et à tort) à la perception des archées comme des organismes anciens et à les nommer *Archaeobacteria*, du grec « *arkhaia* » pour ancien (Woese *et al.* 1977).

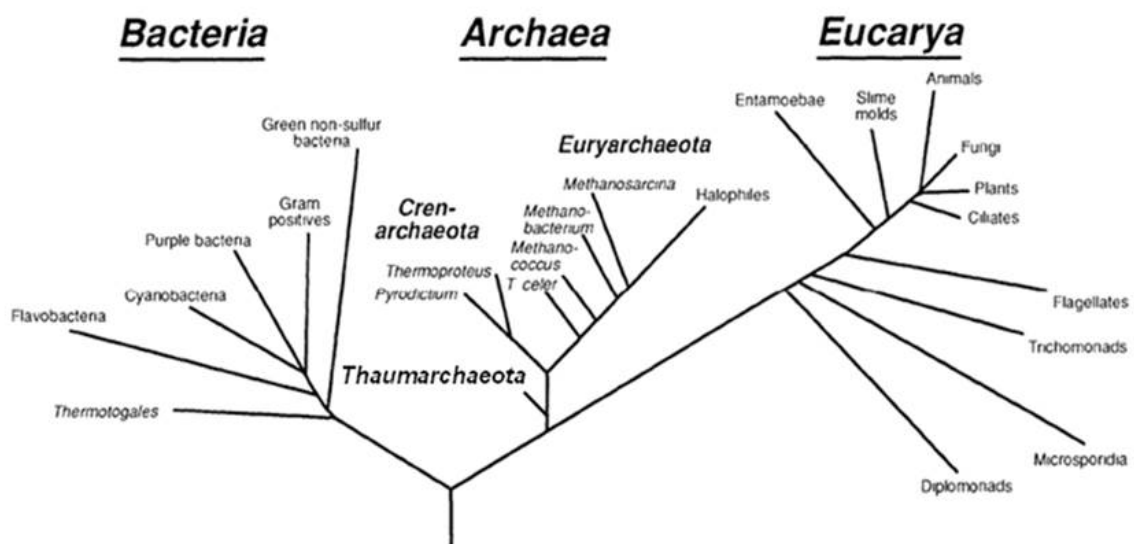


Figure 1 : L'arbre phylogénétique universel, montrant les trois domaines du vivant (modifié d'après Woese *et al.*, 1990 et Gribaldo and Brochier-Armanet., 2006).

La découverte ultérieure d'autres groupes d'archées comme les archées halophiles (Magrum, Luehrsen *et al.* 1978) ou les archées hyperthermophiles (Stetter 1996) a permis de s'apercevoir que les archées ne constituent pas un groupe plus ancien comparé aux eucaryotes ou aux bactéries. Le domaine *Archaeobacteria* fut renommé plus tard, *Archaea* (Woese *et al.* 1990). Désormais, trois domaines du vivant sont reconnus : *Archaea* (archées), *Bacteria* (bactéries) et *Eukarya* (eucaryotes) (Figure 1). L'histoire fascinante de la découverte des archées a été décrite en détails dans plusieurs ouvrages, notamment récemment dans « Microbes de l'enfer » de Patrick Forterre et l'influence des travaux de C. Woese et W. Zillig sur la science et la biologie ont été discutés dans un papier récent (Albers *et al.*, 2013).

2. Organisation générale de la cellule chez les archées

D'un point de vue morphologique, les cellules d'archées sont des cellules procaryotes typiques : elles ne possèdent ni noyau ni organites. En étant un groupe monophylétique, toutes les archées partagent des traits moléculaires principaux communs présentés dans le Tableau 1. L'analyse de ces données indique clairement le caractère mosaïque de l'organisation moléculaire des cellules d'archées qui « empruntent » certaines fonctions (et les gènes correspondants) au domaine des bactéries et d'autres à celui des eucaryotes.

Tableau 1 : Comparaison des principales caractéristiques des *Archaea*, *Bacteria* et *Eukarya*. Les traits caractéristiques des archées, soit uniques, soit partagés avec les eucaryotes ou avec les bactéries, sont résumés. Comme les bactéries, les cellules des archées sont de petite taille et ne possèdent ni de noyau ni d'organites, il est généralement impossible de les différencier des cellules bactériennes sans faire au préalable une analyse moléculaire ou chimique (modifié d'après Gribaldo and Brochier-Armanet, 2006; Prangishvili *et al.*, 2006).

	ARCHAEA	BACTERIA	EUKARYA
Taille	1-4µm	1-4µm	> 5µm
Matériel génétique	Chromosome circulaire	Chromosome circulaire	Plusieurs chromosomes linéaires
Organelles	Non	Non	Oui
Noyau	Non	Non	Oui
Système CRISPR	Oui	Oui	Non
Liaison des lipides	Ether	Ester	Ester
Chaînes carbonées	Isoprènes	Acides gras	Acides gras
Tête phosphate	L-glycérol-1-phosphate	D-glycérol-3-phosphate	D-glycérol-3-phosphate
Réplication/transcription/traduction	Type eucaryote	Bactérien	Eucaryote
ARN polymérase	1 type, 13 sous-unités	1 type, 4 sous-unités	5 types, 12 sous-unités
Régulation de la transcription	Type bactérien	Bactérien	Eucaryote
Histones	Oui	Non	Oui
ARNr	5S, 16S, 23S	5S, 16S, 23S	5S, 8S, 18S, 28S
Synthèse des protéines	Ribosome 70S	Ribosome 70S	Ribosome 70S et 80S
Méthanogénèse	Oui	Non	Non
Pathogène pour l'homme	Non	Oui	Oui

Il est admis que, globalement, les protéines impliquées dans les principaux processus informationnels chez les archées, comme la réplication (Barry and Bell, 2006; de Koning *et al.* 2010), la transcription (Werner and Grohmann, 2011), la traduction (Londei, 2005; Steitz, 2008) et la réparation de l'ADN (Kelman and White, 2005; Fujikane *et al.*, 2010) sont codées par des gènes orthologues proches de ceux des eucaryotes.

En outre, les archées compactent leur génome en utilisant des protéines assez similaires aux histones des eucaryotes (Reeve *et al.*, 2004; Wilkinson *et al.*, 2010) ou d'autres protéines impliquées dans la structuration des chromosomes (Sandman and Reeve, 2005).

D'autre part, les protéines opérationnelles des archées, comme par exemple celles impliquées dans le métabolisme cellulaire de base, sont davantage liées, d'un point de vue évolutif, aux protéines bactériennes. Aussi, l'organisation générale de la cellule archée rappelle fortement celle de la cellule bactérienne. La taille de la cellule, l'absence de noyau, les chromosomes circulaires de taille d'environ 2-3 Mb, l'organisation des gènes en opérons se rapportent clairement aux bactéries. D'autres détails du fonctionnement cellulaire chez les archées rappellent ceux décrits chez les bactéries. Par exemple, comme chez les bactéries, il y a le couplage entre la transcription et la traduction chez les archées (French *et al.* 2007). Egalement en amont du premier codon, une séquence est détectée chez les archées ayant le même rôle que la séquence Shine-Dalgarno chez les bactéries.

Comme il sera discuté plus loin (Chapitre I.4), la régulation de la transcription chez les archées est réalisée par des protéines régulatrices très proches de celles connues chez les bactéries.

Pour terminer, les archées et les bactéries partagent un système d'immunité acquise assurée par les séquences CRISPRs et les gènes *cas*. Les CRISPRs inactivent les virus et d'autres éléments extra chromosomiques procurant aux archées et aux bactéries une immunité contre l'infection par ces ADNs mobiles et infectieux.

Les archées possèdent également des propriétés uniques qu'elles ne partagent avec aucun autre domaine (Tableau 1). Ainsi, la composition chimique de leur membrane cellulaire est unique. Les phospholipides constituant la membrane des archées sont composés de longues chaînes d'alcool isopréniques, liées au glycérol-3-phosphate par des liaisons éther tandis que chez les bactéries et les eucaryotes, il s'agit d'acides gras liées au glycérol-1-phosphate par une liaison ester (Figure 2).

De plus, chez certaines archées, la membrane lipidique bicouche est remplacée partiellement par une membrane monocouche formée de tétra-éthers atypiques présentant deux têtes hydrophiles reliées par leurs chaînes alkyles.

La composition chimique particulière et la structure en monocouche de la membrane des archées lui donnent une très forte résistance aux facteurs de stress, surtout à haute température (Valentine 2007).

Enfin, le peptidoglycane présent dans la majorité des parois cellulaires bactériennes est absent chez les archées, à l'exception de certaines espèces méthanogènes porteuses du pseudo peptidoglycane, une molécule structurellement proche du peptidoglycane (Kandler *et al.* 1978). Chez la majorité des archées, une structure spécifique appelée S-layer remplace le peptidoglycane de l'enveloppe cellulaire. Le S-layer n'est pas une caractéristique exclusive aux archées, puisque certaines bactéries

et cyanobactéries en sont aussi pourvues. Composé de protéines spécifiques formant une structure rigide semi-cristalline, le S-layer procure aux cellules archées leur forme et leur intégrité dans des milieux extrêmes (Kandler *et al.* 1998; Ellen *et al.* 2010; Albers and Meyer. 2011).

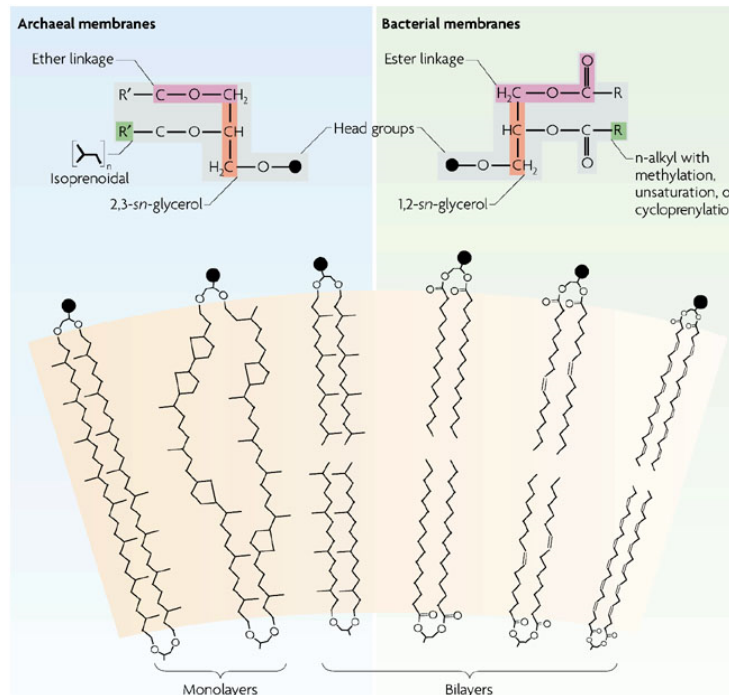


Figure 2 : Comparaison de la structure chimique des membranes des *Archaea* et des *Bacteria*. Les phospholipides des membranes des archées sont composés de longues chaînes d'alcool isopréniques, liées au glycérol-3-phosphate par des liaisons éther tandis que chez les bactéries, il s'agit d'acides gras liées au glycérol-1-phosphate par une liaison ester. La membrane lipidique bicouche est remplacée partiellement par une membrane monocouche formée des tétra-éthers atypiques présentant deux têtes hydrophiles reliées par leurs chaînes alkyles (d'après Valentine, 2007).

3. Diversité des archées

a. D'extrêmophiles spécialisées à mésophiles ubiquitaires

Les archées présentent une grande diversité, en matière d'habitats, de morphologies et de voies métaboliques codées par leurs génomes. Initialement, à l'exception peut-être de certaines méthanogènes, la plupart des espèces d'archées cultivables ont été isolées dans des habitats extrêmes, d'où l'association courante des archées aux organismes extrêmophiles.

Par exemple, la première archée jamais décrite a été isolée en 1878 par W.G. Farlow dans un environnement hautement salé, présumé stérile, (Farlow, 1878). Depuis, des archées halophiles ont été trouvées dans la mer Morte et dans des lacs et sols hypersalins (Lawson *et al.*, 1996). Les archées dites hyperthermophiles sont capables de vivre dans des conditions de pH et de température extrêmes, comme le présente la Figure 3. Cette résistance s'explique le plus probablement par la composition unique et la structure inhabituelle de leurs membranes (voir la section 1.1.2). Actuellement, le record de thermophilie, est établi par la souche 116 de *Methanopyrus kandleri*, d'une source hydrothermale profonde qui est capable de croître à 122°C, sous une forte pression

atmosphérique (Takai *et al.*, 2008). De plus, les archées codent pour des voies métaboliques uniques leur permettant d'utiliser des sources d'énergies inorganiques et de produire des métabolites inhabituels. Ainsi les archées méthanogènes sont les seuls organismes sur Terre capable de transformer le CO₂ en méthane (méthanogénèse) (Liu *et al.*, 2008). Les écosystèmes abritant des méthanogènes incluent des biotopes très divers allant des sources chaudes géothermales, sédiments marins et mangroves aux systèmes digestifs de la plupart des animaux

La surreprésentation des archées dans des endroits extrêmes a contribué à donner à ces organismes une image quelque peu exotique, même si leurs habitats sont loin d'être réduits aux abysses marins et volcans en éruption. Il ne faut pas oublier que des bactéries extrémophiles existent également (Antranikian *et al.*, 2005) même si elles n'égalent jamais les archées par leur résistance aux conditions extrêmes (Figure 3).

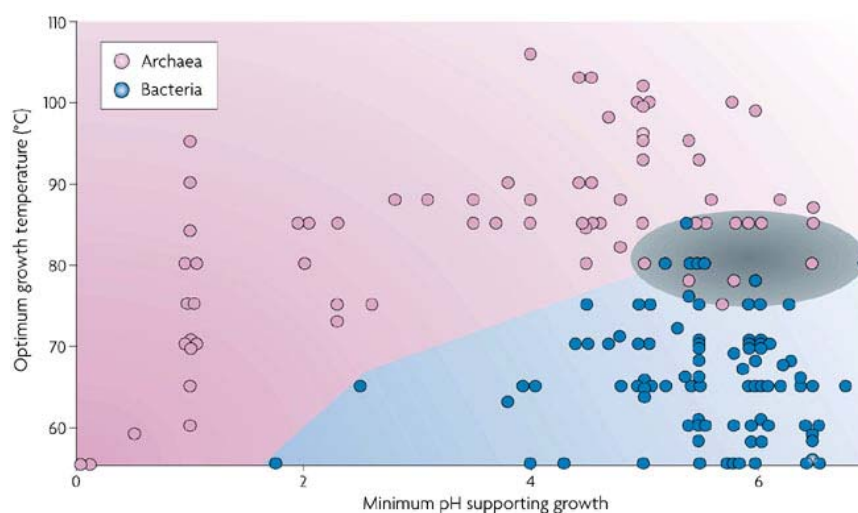


Figure 3 : Prédominance des archées dans les milieux extrémophiles par rapport aux bactéries. Sur l'axe des X : valeurs de pH de 0 à 7; sur l'axe Y : température de croissance optimale de 65°C à 110°C. Les domaines et plages de température/pH sont indiqués en rose (*Archaea*) et en bleu (*Bacteria*). La région grise indique la gamme de conditions convenant à la fois aux archées et aux bactéries. La prédominance des archées dans les environnements très chauds et très acides apparaît clairement (Valentine 2007).

C'est essentiellement le développement des approches méta-génomiques (DeLong, 2005) qui ont permis au début du XXI^{ème} siècle de mettre en évidence la large présence d'archées dans des écosystèmes très divers et non extrêmes (Mardanov and Ravin, 2012), comme les océans (DeLong, 2005), les sédiments océaniques (Knittel *et al.*, 2005), les sols (Leininger *et al.*, 2006), les eaux usées (Tabatabaei *et al.*, 2010), le tractus digestif humain (Gill *et al.*, 2006) ou encore la peau humaine (Probst *et al.*, 2013)

Depuis lors, la prédominance des archées mésophiles dans les différents habitats et le rôle qu'elles jouent dans les cycles importants des différents habitats a été réévaluée. La présence des archées dans tous les principaux écosystèmes de notre planète pointe clairement qu'elles jouent un rôle écologique majeur sur Terre. Les archées sont impliquées dans plusieurs étapes du cycle de l'azote, par la dénitrification ou la nitrification des substrats (Prosser *et al.*, 2012; Liu *et al.*, 2010; Schleper *et al.*, 2010)

L'implication des archées dans l'oxydation de l'ammonium a été découverte récemment (Jia *et al.*, 2010). Il a été pensé pendant longtemps que seul un groupe particulier de bactéries était capable de réaliser cette transformation chimique. Les nouvelles données montrent que cette réaction peut être également effectuée par les archées regroupées dans le phylum *Thaumarchaeota* (Brochier-Armanet *et al.*, 2008) et ceci grâce à la présence des gènes spécifiques *amoA*, *B* et *C* (Schleper, 2012). Dans certains écosystèmes, le rôle des archées dans le cycle de l'azote semble très important. Ainsi, dans certains sols, les gènes *amoA* d'archées peuvent être jusqu'à 3000 fois plus abondants que le même gène d'origine bactérienne (Leininger *et al.*, 2006).

b. Taxonomie des archées

Depuis leur reconnaissance en tant que troisième domaine du vivant, la classification des archées a été basée principalement sur l'analyse des séquences d'ARNr 16S. Le domaine *Archaea* est divisé en trois principaux phyla: *Euryarchaeota*, *Crenarchaeota* et *Thaumarchaeota* découvertes récemment (Figure 4).

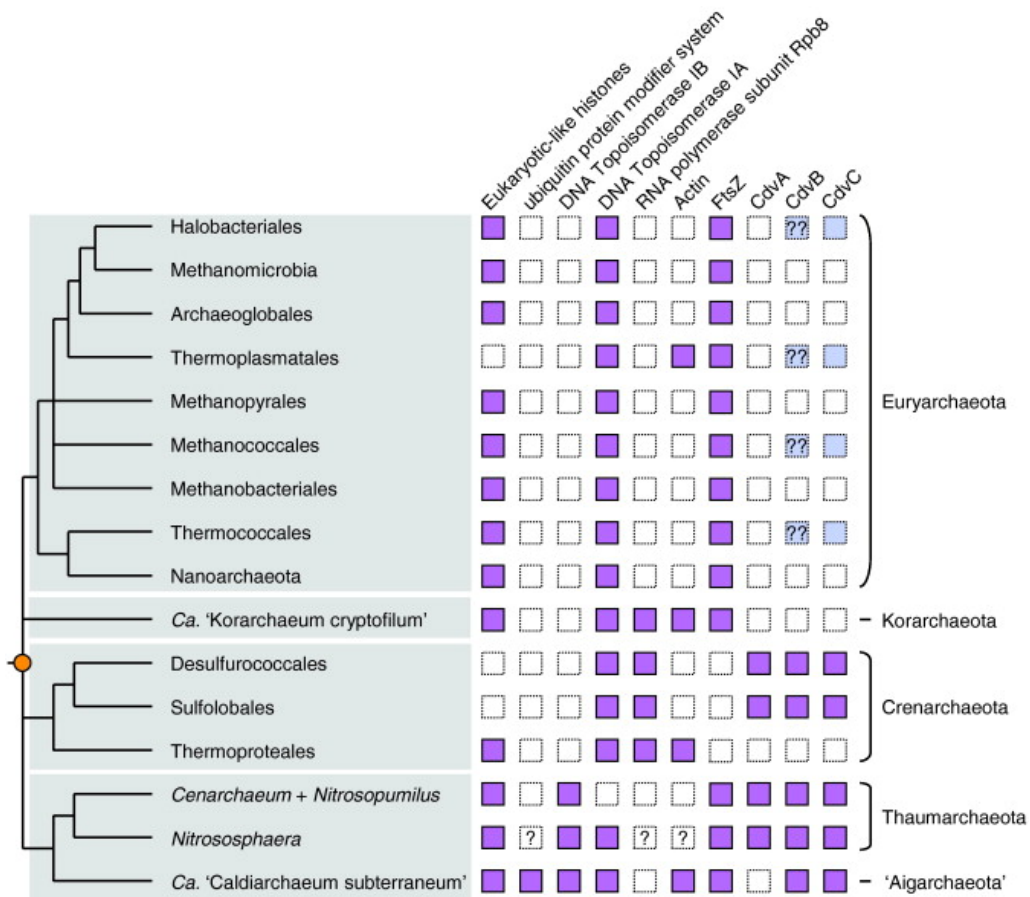


Figure 4 : Comparaison des propriétés moléculaires entre les phyla d'archées. Les composants des systèmes impliqués dans la division cellulaire chez les archées (FtsZ, CdvBC, et actine), la maintenance de l'ADN (topoisomérases IA et IB, histones), l'ubiquitination des protéines et la transcription (Rpb8) systèmes sont répartis inégalement entre les trois phyla principaux (*Crenarchaeota*, *Euryarchaeota* et *Thaumarchaeota*) et les trois phyla proposés (*Korarchaeota*, *Aigarchaeota* et *Nanoarchaeota* – inclut dans les *Euryarchaeota*) (Pester *et al.* 2011).

Les *Euryarchaeota* (du grec « euryos » pour « diverses ») sont phénotypiquement et métaboliquement hétérogènes, dont la majeure partie sont des méthanogènes hyperthermophiles et des halophiles. Ce phylum est divisé en huit ordres : *Archaeoglobi*, *Halobacteria*, *Methanobacteria*, *Methanococci*, *Methanomicrobia*, *Methanopyri*, *Thermococci* et *Thermoplasmata*.

Les *Crenarchaeota* (du grec « crenos » pour « origine ») sont représentées uniquement par des espèces hyperthermophiles. Actuellement, il existe quatre ordres : *Caldisphaerales*, *Desulfurococcales*, *Sulfolobales* and *Thermoproteales*.

La découverte de l'ubiquité de « crénarchées » mésophiles, l'analyse de la présence de gènes codant pour les ARNr 16S spécifiques dans les séquences métagénomiques et le séquençage du premier génome d'une « crénarchée mésophile », *Cenarchaeum symbiosum* ont mis en évidence l'existence dans l'arbre phylogénétique d'une branche supplémentaire aux deux grands groupes d'archées déjà établis (*Eury-* et *Crenarchaeota*). Ainsi, en 2008, l'existence du troisième phylum des *Archaea* est postulée: les *Thaumarchaeota* (du grec « thauma », pour « merveilleux ») (Brochier-Armanet *et al.*, 2008)

La classification des archées reste toujours un sujet de controverses (Gribaldo *et al.* 2012). L'existence de plusieurs autres phyla chez les archées a été suggérée dans certaines publications. La liste des « candidats » inclut les phyla *Korarchaeota* (Barns *et al.* 1996), *Nanoarchaeota* (Huber *et al.* 2002), et plus récemment *Aigarchaeota* (Nunoura *et al.* 2011).

Le phylum des *Korarchaeota* a été proposé sur la base de la phylogénie de l'ARNr 16S car, à l'époque, aucune espèce cultivable de ce groupe n'était disponible (Elkins *et al.* 2008). La vaste répartition géographique de ce groupe a été par la suite montrée par les analyses PCR menées dans plusieurs environnements marins et terrestres géographiquement isolés, comme le Parc du Yellowstone (Barns *et al.* 1994), des sources chaudes d'Islande (Hjorleifsdottir *et al.* 2001) ou le bassin du Guayamas dans le golfe de Californie (Teske *et al.* 2002). Le séquençage du premier génome de *Korarchaeota*, *Korarchaeum cryptophilum*, a été fait à partir d'une culture enrichie (Elkins *et al.* 2008). Son analyse pointe la proximité entre les *Korarchaeota* et les *Crenarchaeota* qui, très probablement, ne formeraient qu'un seul phylum. Des analyses supplémentaires seront nécessaires pour établir la taxonomie définitive de ce groupe au sein des *Archaea*.

L'existence du phylum *Nanoarchaeota* (du grec « nano » pour « petit »), suggérée en 2002 (Hohn *et al.* 2002), n'a pas été acceptée unanimement par la communauté scientifique. Ce phylum n'est composé que d'un seul organisme, *Nanoarchaeota equitans*, parasite obligatoire de la crénarchée *Ignicoccus hospitalis* (Huber *et al.* 2002). Des traces d'ARNr 16S proche de celui de *Nanoarchaeota equitans* ont été détectées dans des biotopes hyperthermophiles du Parc du Yellowstone (USA), de la caldera d'Uzon (Russie) et dans les sources chaudes sous-marines profondes de la dorsale Pacifique Est (Hohn *et al.* 2002). L'analyse phylogénétique plus récente, portant sur l'ensemble des cadres ouverts de lecture (ORF) du génome de *Nanoarchaeota equitans*, contredit l'observation faite à partir de l'analyse des séquences 16S et définit ce phylum comme une possible lignée présentant une évolution accélérée des *Euryarchaeota* (Brochier *et al.* 2005).

Le phylum des *Aigarchaeota* (du grec « agia » pour « aube ») est le dernier phylum en date proposé (Nunoura *et al.* 2011). Il se compose actuellement des séquences des crénarchées thermophiles du Groupe I, dont le génome composite d'une des espèces, *Caldiarchaeum subterraneum* est

actuellement disponible (Nunoura *et al.* 2011). Ce génome possède certaines caractéristiques uniques, comme la présence de gènes homologues aux gènes impliqués dans l'ubiquitination des protéines chez les eucaryotes (Figure 4). Ce phylum pourrait former un groupe-sœur du phylum *Thaumarchaeota* (Nunoura *et al.* 2011).

L'identification des principaux taxons des *Archaea* est essentielle pour parfaire notre connaissance de leur diversité et de leur évolution. Quinze ans après le séquençage du premier génome d'archée, le palier symbolique des cent génomes d'archées séquencés est maintenant franchi (Brochier-Armanet *et al.* 2011) mais notre connaissance de ce domaine et de sa composition taxonomique reste incomplète.

4. Les *Sulfolobales*, des crénarchées modèles

La compréhension de l'organisation et du fonctionnement de la cellule des archées, leur rôle écologique, leur origine et leur évolution sont indispensables pour décrire et caractériser toutes les formes de vie sur Terre. La résolution de la question de l'origine de la vie sur la planète est également impensable sans inclure les archées dans les modèles biologiques à étudier. La forte différence de fonctionnement entre les archées appartenant aux différents phyla ne permet pas de proposer un modèle unique, une sorte de « *E. coli* » des archées, pour tenter à répondre aux questions posées. Chez les *Euryarchaeota*, plusieurs organismes comme *Methanococcus maripaludis* (Haydock *et al.* 2004; Hendrickson *et al.*, 2008), *Methanosarcina* (Welander and Metcalf, 2008), *Halobacterium salinarum* (Ng *et al.*, 2008; Peck *et al.*, 2000) et *Haloferax volcanii* (Robinson *et al.*, 2005) sont établis comme modèles pour les méthanogènes et les halophiles (Leigh *et al.*, 2011). Concernant les *Thaumarchaeota*, le progrès dans l'étude de ce phylum souffre pour le moment de l'absence d'un modèle cultivable établi, malgré une rapide avancée dans l'exploration de cette branche très importante des archées. Enfin, parmi les *Crenarchaeota*, les *Sulfolobales* sont probablement les mieux caractérisées et communément utilisées comme un modèle représentant des *Crenarchaeota* hyperthermophiles.

Les *Sulfolobales* regroupent des archées hyperthermophiles, acidophiles et aérobies obligatoires, ayant une croissance optimale autour de 80°C et pH 3. Ces organismes sont présents dans des sources chaudes acides et soufrées, à travers la planète. Le premier membre des *Sulfolobales*, *S. acidocaldarius*, a été décrit par T. Brock en 1972 (Brock *et al.* 1972) et isolé à partir d'une source chaude du parc naturel du Yellowstone (Etats-Unis). Plus tard, différents membres ont été caractérisés à partir de sources chaudes acides dans le monde entier, comme *S. solfataricus* (Naples, Italie; Zillig *et al.*, 1980), *S. islandicus* (Islande; Zillig *et al.*, 1994) ou *S. tokodaii* (Japon; Suzuki *et al.*, 2002).

Les *Sulfolobus* sont des organismes chimiolithotrophes, capables d'utiliser des composés soufrés, comme source d'énergie et de fixer le carbone à partir du CO₂ gazeux, en respiration aérobie. (Figure 5). Cependant, ils peuvent également croître comme hétéroorganotrophes, en utilisant des composés organiques, comme source de carbone et d'énergie, ce qui facilite leur maintien et leur croissance en laboratoire. En effet, le temps de doublement est compris entre quatre et six heures et permet une croissance rapide en 48-72 heures. Des milieux solides sont disponibles, grâce à l'utilisation d'un polymère spécifique, un hétéropolysaccharide (Gellan Gum, nommé commercialement Phytigel® ou Gelrite®) supportant, contrairement à l'agar-agar, l'exposition aux fortes températures.

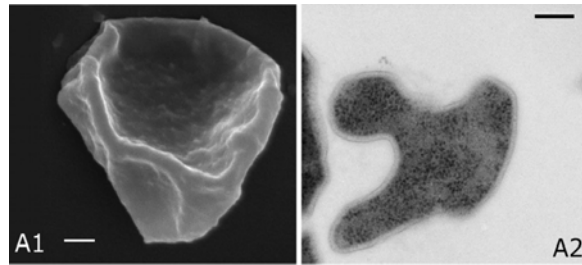


Figure 5 : Une cellule *S. islandicus* typique possède une forme irrégulière en lobes. Vues prises sous microscopie électronique à balayage (A1) et à transmission (A2). La barre d'échelle représente 200nm (Bize *et al.*, 2009).

Grâce à une croissance rapide dans les conditions de laboratoire, les *Sulfolobus* apparaissent comme des modèles très répandus pour l'étude des processus biologiques vitaux (réplication (Zhang *et al.*, 2010); transcription (Peeters *et al.*, 2009); traduction (La Teana *et al.*, 2013)) et du métabolisme des crénarchées (Garrett *et al.*, 2007). Les génomes des souches *Sulfolobus* communément utilisées ont été séquencés et sont disponibles : *S. acidocaldarius* (Chen *et al.*, 2005), *S. solfataricus* P2 (She *et al.*, 2001), *S. tokodaii* (Kawarabayasi *et al.*, 2001), *Metallosphaera sedula* (Auernik *et al.*, 2008) et les souches *S. islandicus* REY15A et HVE10/4 (Guo *et al.*, 2011). La souche modèle *S. islandicus* LAL14/1, utilisée dans notre laboratoire pour l'étude des interactions hôte-virus a été récemment séquencée (Jaubert *et al.*, 2013) et l'établissement de l'approche génétique pour cette souche fait partie de mon travail doctoral dont les résultats seront présentés dans le chapitre II.2.

Les espèces *Sulfolobus* ont été la source d'isolation et de caractérisation d'un grand nombre d'éléments génétiques, comme les virus et les plasmides (Zillig *et al.*, 1996; Prangishvili *et al.*, 1998, 2006; Peng *et al.*, 2000; Greve *et al.*, 2004). Ces éléments génétiques sont intensément étudiés et certains ont été utilisés pour le développement d'outils génétiques (Berkner and Lipps, 2008) qui seront discutés plus loin dans ce manuscrit.

2. Virus d'archées

Comme pour les membres des autres domaines de vie, les cellules d'archées peuvent être infectées par les virus; par ailleurs, ces virus sont spécifiques à chaque domaine. A l'exception du phylum des *Thaumarchaeota*, les différentes espèces de virus infectant les *Euryarchaeota* et les *Crenarchaeata* sont décrites en abondance depuis la découverte du premier virus infectant les archées halophiles (Torsvik and Dundas, 1974). Dans les dix dernières années, un effort important a été entrepris pour étudier les virus d'archées et actuellement, quarante-deux espèces virales ont été décrites (Pina *et al.*, 2011). Dans les sections suivantes, je me focaliserai exclusivement sur les virus des crénarchées, puis, pour illustrer les détails, plus particulièrement sur deux familles de virus dont deux membres, SIRV1/2 et AFV1 ont été des modèles d'études dans mon travail doctoral.

De manière générale, il est considéré que les virus sont apparus tôt dans l'évolution de la vie et qu'ils ont joué un rôle primordial dans les transitions évolutives majeures (Forterre 2006). Aussi, l'étude des virus peut apporter des informations importantes permettant de mieux comprendre les mécanismes principaux de l'évolution en générale et celle du domaine *Archaea* en particulier (Forterre *et al.*, 2013).

Les virus d'archées ont été isolés à partir de deux types d'habitats : les environnements géothermaux extrêmes, dont les températures dépassent les 80°C et les eaux hypersalines (Rachel *et al.*, 2002; Prangishvili *et al.*, 2006; Prangishvili 2013).

Bien que le nombre de virus d'archées connus et étudiés soit limité, leur diversité morphologique est exceptionnelle et inclut des morphotypes uniques. Face à cette diversité, ces virus ont été répartis en huit familles bien que certains restent non classés (Tableau 2). Les morphotypes des virions sont montrés dans la Figure 6. Le morphotype tête-queue observé chez les virus infectant les *Euryarchaeota* est absent des virus des *Crenarchaeota* (Prangishvili *et al.* 2005; Comeau *et al.* 2008; Forterre *et al.* 2009; Pina *et al.* 2011). Un grand nombre des virus infectant les crénarchées présentent un morphotype en bâtonnet (*Rudiviridae*, *Lipothrixviridae*) ou fusiforme (*Fuselloviridae*). Quelques morphotypes inédits dans le monde vivant peuvent être observés, comme la forme de bouteille du virus ABV (Figure 6A) (Haring *et al.* 2005) ou de citron du virus ATV (Figure 6B) (Haring *et al.* 2005). Ce dernier virus présente la particularité, unique dans le monde vivant, de pouvoir développer ses deux extrémités hors des cellules hôtes (Haring *et al.* 2005).

Les virus infectant les crénarchées semblent être plutôt du type lysogène; très peu de ces virus provoquent la mort massive des cellules infectées (Zillig *et al.* 1996; Bize *et al.* 2009).

Parmi les virus d'archées, seulement deux espèces possèdent un génome ADN simple brin (Pietilä *et al.*, 2009; Mochizuki *et al.*, 2012) alors que les autres sont des virus à ADN double brin (Tableau 2). De la même façon que leurs morphotypes sont uniques, les virus de crénarchées sont exceptionnels également d'un point de vue génomique : plus de 90% de leurs gènes ne montrent aucune homologie avec les séquences disponibles dans les banques de données publiques (Prangishvili *et al.*, 2006).

Tableau 2 : Taxonomie et propriétés des virus infectant les *Crenarchaeota* (d'après Pina *et al.*, 2011; Peng *et al.*, 2012 et Prangishvili and Garrett, 2005)

Morphologie	Organisme	Hôte	Taille (en kb)	Référence
Bouteille	Famille Ampullaviridae ABV	<i>Acidianus</i>	Linéaire 23.8kb	(Haring, Rachel <i>et al.</i> , 2005)
	Famille Bicaudaviridae ATV	<i>Acidianus</i>	Circulaire 62.7kb	(Haring, Vestergaard <i>et al.</i> , 2005)
Fusiforme	Famille Fuselloviridae SSV1	<i>Sulfolobus</i>	Circulaire 15.5kb	(Schleper, Kubo <i>et al.</i> , 1992)
	SSV2	<i>Sulfolobus</i>	Circulaire 14.8kb	(Stedman, She <i>et al.</i> , 2003)
	SS-K1	<i>Sulfolobus</i>	Circulaire 17.4kb	(Wiedenheft, Stedman <i>et al.</i> , 2004)
	SSVRH	<i>Sulfolobus</i>	Circulaire 16.5kb	(Wiedenheft, Stedman <i>et al.</i> , 2004)
	SSV4	<i>Sulfolobus</i>	Circulaire 15.1kb	(Peng 2008)
	SSV5	<i>Sulfolobus</i>	Circulaire 15.3kb	(Redder, Peng <i>et al.</i> , 2009)
	SSV6	<i>Sulfolobus</i>	Circulaire 15.7kb	(Redder, Peng <i>et al.</i> , 2009)
	SSV7	<i>Sulfolobus</i>	Circulaire 17.6kb	(Redder, Peng <i>et al.</i> , 2009)
	ASV1	<i>Acidianus</i>	Circulaire 24.8kb	(Redder, Peng <i>et al.</i> , 2009)
Linéaire	Famille Lipothrixviridae Genre <i>Alphalipothrixvirus</i> TTV1	<i>Thermoproteus</i>	Linéaire 13.7kb	(Janekevic, Wunderl <i>et al.</i> , 1983)
	Genre <i>Betalipothrixvirus</i> SIFV	<i>Sulfolobus</i>	Linéaire 40.9kb	(Arnold, Zilig <i>et al.</i> , 2000)
	AFV3	<i>Acidianus</i>	Linéaire 40.4kb	(Vestergaard, Shah <i>et al.</i> , 2008)
	AFV6	<i>Acidianus</i>	Linéaire 39.6kb	(Vestergaard, Shah <i>et al.</i> , 2008)
	AFV7	<i>Acidianus</i>	Linéaire 36.9kb	(Vestergaard, Shah <i>et al.</i> , 2008)
	AFV8	<i>Acidianus</i>	Linéaire 38.2kb	(Vestergaard, Shah <i>et al.</i> , 2008)
	AFV9	<i>Acidianus</i>	Linéaire 41.2kb	(Bize, Peng <i>et al.</i> , 2008)
	Genre <i>Gammalipothrixvirus</i> AFV1	<i>Acidianus</i>	Linéaire 21.1kb	(Bettstetter, Peng <i>et al.</i> , 2003)
	Genre <i>Deltalipothrixvirus</i> AFV2	<i>Acidianus</i>	Linéaire 31.2kb	(Haring, Vestergaard <i>et al.</i> , 2005)
	Famille Rudiviridae SIRV1	<i>Sulfolobus</i>	Linéaire 32.3kb	(Prangishvili, Arnold <i>et al.</i> , 1999)
	SIRV2	<i>Sulfolobus</i>	Linéaire 35.5kb	(Prangishvili, Arnold <i>et al.</i> , 1999)
	ARV1	<i>Acidianus</i>	Linéaire 24.7kb	(Haring, Vestergaard <i>et al.</i> , 2005)
	SRV	<i>Stygiolobus</i>	Linéaire 17.6kb	(Vestergaard, Shah <i>et al.</i> , 2008)
	Bacille	Famille Clavaviridae APBV	<i>Aeropyrum</i>	Circulaire 5.3kb
Sphérique	Famille Globuloviridae PSV	<i>Pyrobaculum</i>	Linéaire 28.3kb	(Haring, Peng <i>et al.</i> , 2004)
	TTSV1	<i>Thermoproteus</i>	Linéaire 20.9kb	(Ahn, Kim <i>et al.</i> , 2006)
Gouttelette	Famille Guttaviridae SNDV	<i>Sulfolobus</i>	Circulaire 20kb	(Arnold, Ziese <i>et al.</i> , 2000)
Non classés	Isomériques STIV	<i>Sulfolobus</i>	Circulaire 17.7kb	(Rice, Tang <i>et al.</i> , 2004)
	STIV2	<i>Sulfolobus</i>	Circulaire 17.6kb	(Happonen, Redder <i>et al.</i> , 2010)
	Inconnu HAV2	Non déterminé	N.d 17.7kb	(Garett, Prangishvili <i>et al.</i> , 2010)
	Ovoïde APOV1	<i>Aeropyrum</i>	Circulaire	(Mochizuki, Sako <i>et al.</i> , 2011)
	Spindle-shaped APSV1	<i>Aeropyrum</i>	Circulaire 38kd	(Mochizuki, Sako <i>et al.</i> , 2011)
	STSV1	<i>Sulfolobus</i>	Circulaire 75.3kb	(Xiang, Chen <i>et al.</i> , 2005)

Les virions avec une forme linéaire sont vraisemblablement prédominants dans beaucoup d'écosystèmes extrêmes (Snyder *et al.*, 2007; Held and Whitaker, 2009). Les virus linéaires isolés et caractérisés qui infectent les crénarchées hyperthermophiles du genre *Sulfolobus*, *Acidianus*, *Stygiolobus* et *Thermoproteus* ont été classés en deux familles selon la structure de leur virion et la composition de leurs génomes : les *Rudiviridae* et les *Lipothrixviridae* (Figure 6C et 6D). De récentes analyses génomiques et structurales ont démontré un lien évolutif fort entre les membres de ces deux familles. Les protéines majeures de capsides des virions forment des hélices parallèles et ces structures sont très similaires chez les deux groupes, malgré une faible conservation de séquences. Récemment, ces résultats ont conduit à la proposition de réunir les *Rudiviridae* et les *Lipothrixviridae* sous un seul ordre, les *Ligamenvirales* (Prangishvili and Krupovic, 2012).

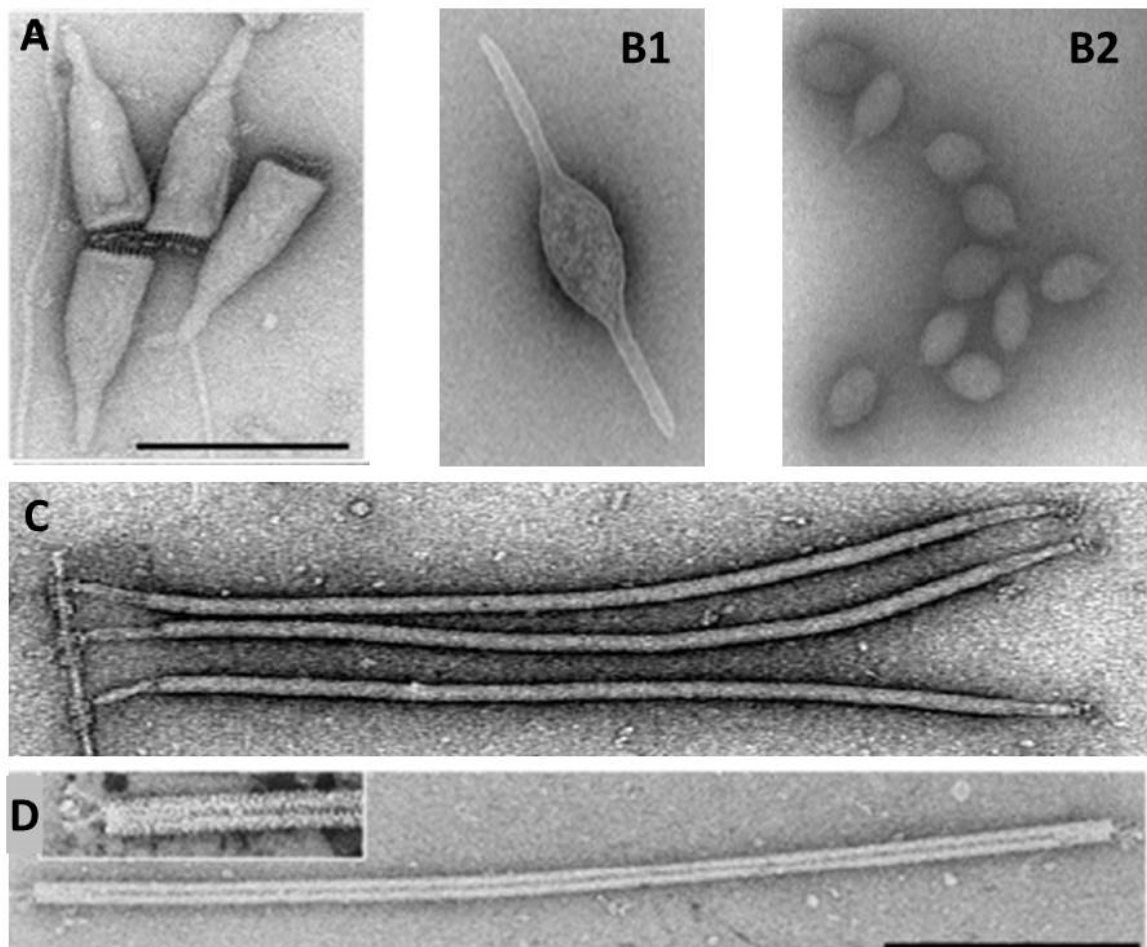


Figure 6 : Virus des hyperthermophiles Crenarchaeota et Euryarchaeota, représentant les différentes familles, sous microscopie électronique à transmission. (A) Ampullaviridae : ABV; (B1 et B2) : Bicaudaviridae : ATV, virions et forme mature du virion; (C) Lipothrixviridae : AFV1; (D) Rudiviridae : SIRV2. Les barres représentent 100nm ((F, H, J) et 200nm (A, E) (modifié depuis Prangishvili *et al.*, 2006; Peng *et al.*, 2012; Pina *et al.*, 2011).

1. Famille *Rudiviridae*

Les membres de la famille des *Rudiviridae* sont parmi les mieux étudiés des virus des crénarchées. En effet, leurs cellules hôtes sont facilement cultivables dans les conditions de laboratoire et l'efficacité de l'infection virale est suffisamment bonne pour pouvoir maintenir la production virale dans un milieu synthétique. Ils sont bien caractérisés d'un point de vue structural (Prangishvili *et al.* 1999), moléculaire (Birkenbihl *et al.* 2001; Kessler *et al.* 2004; Guilliere *et al.* 2009) et fonctionnel (Bize *et al.* 2009; Quax *et al.* 2010; Quax *et al.* 2011) et sont de plus en plus utilisés comme modèles pour décrire et comprendre les interactions hôte-virus (Peng *et al.* 2004; Bize *et al.* 2009; Quax *et al.* 2011) ou la régulation de la transcription (Guillière *et al.*, 2009).

La famille *Rudiviridae* inclut à présent quatre membres, notamment les virus **SIRV1** et **2** (*Sulfolobus islandicus rod-shaped virus 1* et **2**) (Prangishvili *et al.*, 1999), **ARV1** (*Acidianus rod-shaped virus 1*) (Haring *et al.*, 2005) et **SRV** (*Stygiolobus rod-shaped virus*) (Vestergaard *et al.*, 2008). Tous proviennent de sources d'eau chaude en Europe (Islande, Italie et Portugal). Récemment, selon GenBank®, un génome supplémentaire de rudivirus a été séquencé. Ce virus **SMRV1** (*Sulfolobales Mexican rudivirus 1*) a été retrouvé à partir d'une source chaude au Mexique (Servin-Garcidueñas *et al.*, 2013)

Sur les cinq isolats de rudivirus, SIRV1 et SIRV2 sont les plus proches, avec 73% d'identité de séquence de leur génome complet (Figure 7). De plus, SIRV1 et SIRV2 infectent des souches *S. islandicus* proches. La plupart des connaissances actuelles sur la famille *Rudiviridae* proviennent des études réalisées avec SIRV1 et SIRV2. Etant donné la similitude entre ces deux virus, les résultats obtenus avec un virus peuvent être directement transférable à un autre.

Les virus SIRV sont les plus étudiés dans la famille des *Rudiviridae* et peuvent être considérés comme les virus « types » de cette famille. En outre, ils m'ont servi de modèles au cours de mon travail de thèse. Aussi, pour la suite, je présenterai surtout les propriétés de ces deux représentants de la famille des *Rudiviridae*.

a. Structure des virions

La composition des virions et l'organisation structurale des rudivirus ont été étudiées par des approches biochimiques *in vitro* et par observation sous microscope électronique. Les virions de SIRV1 et SIRV2, membres types de la famille, sont des particules non enveloppées ayant la forme de bâtonnet d'une longueur variant de 610 à 910nm et d'une largeur de 23nm (Prangishvili *et al.*, 1999; Vestergaard *et al.*, 2005). Le corps du virion est formé d'une super-hélice composée d'un ADN db génomique recouvert d'une seule protéine de capsid de 20kDa (protéine P134). La structure cristalline de la protéine majeure de capsid, P134 (MCP pour « *major capsid protein* ») révèle une topologie en faisceau à quatre hélices (Figure 7) (Szymczyma *et al.*, 2009). Sa structure est presque identique à la protéine majeure de capsid du virus AFV1 des *Lipothrixviridae*, malgré l'absence de similarité de séquences entre les gènes correspondants de ces deux virus (Figure 11b). Les extrémités des virions portent trois courtes fibres, probablement impliquées dans l'adsorption des virions à la surface des cellules de l'hôte (Steinmetz *et al.*, 2008). La plus grande protéine des virus SIRV1-2, P1070 est un composant de ces fibres. Outre P134 et P1070, deux autres protéines codées par les ORFs 488 et 564 ont été trouvées présentes, bien qu'en très petite quantité dans le virion de SIRV2 (Vestergaard *et al.*, 2008) (Tableau 4). Les quatre protéines de capsides sont codées par tous les

membres de la famille *Rudiviridae* et le degré de conservation de séquence est très élevé, surtout pour la protéine majeure de capside (83-95% d'identité). La protéine P1070 est codée par le gène *gp30* chez le virus SIRV1 et fait l'objet d'étude de mon travail de thèse.

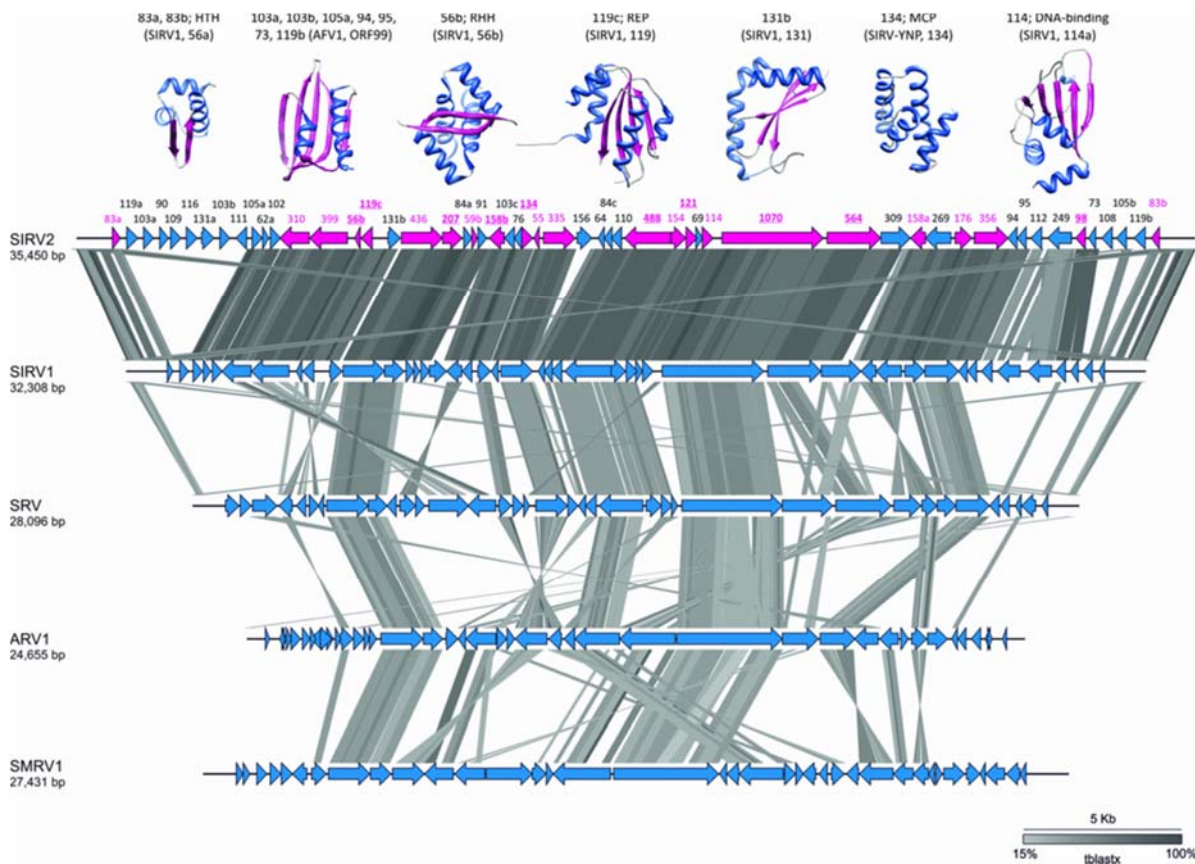


Figure 7 : Relations génomiques entre les membres de la famille des *Rudiviridae*. Les ORFs du virus de SIRV2 pour lesquelles les fonctions ont été démontrées ou déduites *in silico* sont montrées en magenta. Les noms des ORFs, qui codent pour des protéines avec des fonctions vérifiées expérimentalement sont soulignées. Les hits entre les génomes des *rudivirus* sont indiqués par différents nuances de gris (l'échelle d'identité est incluse dans la figure). Les structures des protéines des *Ligamenvirales* (*rudivirus* et *lipothrixvirus*) avec des orthologues dans le génome de SIRV2 sont montrées au-dessus de la carte du génome de SIRV2, avec la source de chaque structure indiquée en parenthèses. Les structures des protéines sont colorées selon la structure secondaire des éléments : en bleu, les hélices α , et en magenta, les feuilletts β (Prangishvil *et al.*, 2013).

b. Organisation du génome

Les génomes des cinq rudivirus séquencés portent un génome ADN db linéaire, aux extrémités fermées de façon covalente par des structures en épingle à cheveux (Blum *et al.*, 2001). Cette structure terminale est impliquée dans la réplication (voir paragraphe suivant). Les génomes viraux portent de longues répétitions terminales (ITR : *inverted terminal repeats*), qui incluent de multiples répétitions directes. La séquence du génome et la composition des rudivirus *Acidianus* (ARV) et *Stygiolobus* (SRV) diffèrent significativement de celles des deux virus *Sulfolobus* (SIRV1 et SIRV2) (Figure 7). Les génomes de SIRV1 (32,3 kb) et SIRV2 (35,8kb), contiennent quarante-cinq et cinquante-quatre ORFs, respectivement, dont quarante-quatre sont homologues entre les deux virus (Tableau 3).

Tableau 3 : Correspondance des annotations des gènes entre les virus SIRV1 et SIRV2. Le nom des ORFs a été défini en fonction de nombre d'acides aminés composant la protéine codée par l'ORF correspondante. Certains gènes sont spécifiques de SIRV1 ou de SIRV2 et la majorité est homologue entre les deux.

SIRV1		SIRV2		SIRV1		SIRV2	
Gène	ORF	Gène	ORF	Gène	ORF	Gène	ORF
<i>gp01/45</i>	55	<i>gp01/54</i>	83a	<i>gp24</i>	110	<i>gp32</i>	110
<i>gp02/44</i>	90	<i>gp52</i>	105b	<i>gp25</i>	510/486	<i>gp33</i>	488
<i>gp03</i>	102	<i>gp53</i>	119b	<i>gp26</i>	154	<i>gp34</i>	154
<i>gp04</i>	76			<i>gp27</i>	121	<i>gp35</i>	121
<i>gp05</i>	105	<i>gp12</i>	102	<i>gp28</i>	74	<i>gp36</i>	69
<i>gp06</i>	306	<i>gp13</i>	310	<i>gp29</i>	114	<i>gp37</i>	114
<i>gp07</i>	399	<i>gp14</i>	399	<i>gp30</i>	1070	<i>gp38</i>	1070
<i>gp08</i>	56 (SvtR)	<i>gp15</i>	56b (SvtR)	<i>gp31</i>	562	<i>gp39</i>	564
<i>gp09</i>	117	<i>gp16</i>	119c	<i>gp32</i>	417	<i>gp40</i>	309
<i>gp10</i>	131	<i>gp17</i>	131b	<i>gp33</i>	158b	<i>gp41</i>	158a
<i>gp11</i>	440	<i>gp18</i>	436	<i>gp34</i>	268	<i>gp42</i>	269
<i>gp12</i>	207	<i>gp19</i>	207	<i>gp35</i>	209	<i>gp43</i>	176
<i>gp13</i>	81	<i>gp20</i>	84a	<i>gp36</i>	356	<i>gp44</i>	356
<i>gp14</i>	59	<i>gp21</i>	59b	<i>gp37</i>	90b	<i>gp03</i>	103a
<i>gp15</i>	91	<i>gp22</i>	91	<i>gp38</i>	95	<i>gp10</i>	105a
<i>gp16</i>	179			<i>gp39</i>	112	<i>gp47</i>	112
<i>gp17</i>	158a			<i>gp40</i>	241	<i>gp48</i>	249
<i>gp18</i>	101	<i>gp24</i>	103c	<i>gp41</i>	252	<i>gp48</i>	249
<i>gp19</i>	134	<i>gp26</i>	134	<i>gp42</i>	98	<i>gp49</i>	98
<i>gp20</i>	70	<i>gp27</i>	55	<i>gp43</i>	75	<i>gp50</i>	73
<i>gp21</i>	335	<i>gp28</i>	335	<i>gp44/02</i>	90c	<i>gp52</i>	105b
<i>gp22</i>	64	<i>gp30</i>	64	<i>gp45/01</i>	56c	<i>gp54/01</i>	83b
<i>gp23</i>	84	<i>gp31</i>	84c			<i>gp04</i>	90

D'après leur homologie à des équivalents bien caractérisés, les fonctions ont été attribuées et confirmées ensuite pour une dUTPase (ORF138a; Prangishvili *et al.*, 1998) et une résolvasse de jonction Holliday (ORF121; Birkenbihl *et al.*, 2011). Le gène codant pour la protéine de capsid, P134 a été identifié par séquençage N-terminal (ORF134; Peng *et al.*, 2001) (Figure 7). De plus, au moins 10% des protéines codées sont prédites pour exposer différents motifs de liaison à l'ADN et comme des régulateurs transcriptionnels putatifs (Prangishvili *et al.*, 2006). D'autres fonctions prédites sont indiquées dans la Figure 11c, et inclut une ARNt transglycosylase, une glucosyl-transférase, une S-méthyltransférase adénosylméthionine-dépendante et une recB exonucléase (Tableau 4).

La particularité du virus SIRV1 réside dans son instabilité génomique. Contrairement au génome du virus SIRV2, très stable (Peng *et al.* 2004), jusqu'à 10% de substitutions ont pu être observées entre deux variants du virus SIRV1, dont certains présentaient des délétions (Peng *et al.* 2004). Le mécanisme qui permet à SIRV1, mais pas à SIRV2, de s'adapter à de nouveaux hôtes en accumulant des mutations ponctuelles conduisant à la production de variants viables n'est pas encore éclairci. A cause de sa stabilité dans beaucoup d'hôtes, SIRV2 est un meilleur candidat pour l'étude des interactions avec l'hôte, tandis que SIRV1 pour des raisons historiques est un modèle utilisé pour étudier la réplication ou la régulation transcriptionnelle.

L'ADN du rudivirus SIRV1 montre des propriétés exceptionnelles. Les isolats viraux contiennent invariablement une population de variants avec des génomes différents mais très proches. Lors de la propagation dans une souche hôte donnée, un ou plusieurs génomes dominant dans la population virale. Cependant, lors du passage dans une nouvelle souche hôte, la population virale subit des changements et d'autres variants dominants sont sélectionnés. Le séquençage et l'analyse des génomes des variants révèlent que les différences majeures apparaissent dans l'ordre, la taille et le contenu des gènes au niveau de sites génomiques localisés (Peng *et al.*, 2001).

c. Réplication de l'ADN

Les deux brins des ADN db linéaires des virus SIRV1/2 sont liés de façon covalente aux deux extrémités du génome, formant une chaîne polynucléotidique continue (Blum *et al.*, 2001). Un tel schéma structural du génome combiné à l'absence d'une ADN polymérase codée par le virus identifiable suggère un mécanisme unique de réplication du génome des rudivirus. Le virus SIRV1 a servi de modèle expérimental pour ces études.

La résolution de la structure de la protéine P119 codée par génome de SIRV1 a été cruciale dans la compréhension de la réplication du génome chez les rudivirus (Figure 7)(ORF119c; Oke *et al.*, 2011). Cette protéine P119 s'est avérée être un membre de la superfamille des initiateurs de la réplication (Rep), qui normalement initie la réplication de type « *rolling-circle* » de divers virus et plasmides. Cependant, dans le cas des virus SIRV1/2, la réplication ne suit pas le modèle « *rolling-circle* ».

Il a été montré que la protéine Rep des virus SIRV1/2, qui fonctionne en dimère, réalise une coupure simple brin site spécifique au niveau de la structure en épingle à cheveux, à l'extrémité du génome du virus, ce qui libère une extrémité 3'-OH, appropriée pour l'amorçage de la réplication (Peng *et al.*, 2001). Cette enzyme catalyse aussi la réaction de raccordement, qui referme l'ADN et termine la réplication. Le dimère de protéine, avec deux sites actifs fonctionnant en tandem, assure le clivage et la jointure de l'ADN, pendant la progression de la réplication

Tableau 4 : Fonctions codées par le génome SIRV2. D'après Krupovic *et al.*, 2013.

Rôle	Protéine	Fonction	Mise en évidence
Structure du virion	P134	Protéine majeur de capsid	Expérimentale
	P488	Protéine structurale	Expérimentale
	P1070	Protéine structurale, fibres terminales	Expérimentale
	P564	Protéine structurale	Expérimentale
Liaison à l'ADN et contrôle transcriptionnel	P56b (SvtR)	Régulateur transcriptionnel	Expérimentale
	P83a/P83b	Liaison à l'ADN et/ou oligomérisation protéique	Similarité similaire
	P59b	Protéine de liaison à l'ADN	Analyse <i>in silico</i>
	P55	Protéine de liaison à l'ADN	Analyse <i>in silico</i>
	P114	Protéine de liaison à l'ADN	Similarité similaire
Réplication du génome	P119C	Initiateur de la réplication	Expérimentale
	P121	Hjr	Expérimentale
Métabolisme des acides nucléiques	P158b	dUTPase	Expérimentale
	P207	Endonucléase ADN sb spécifique	Expérimentale
Sortie du virus	P98	Formation des VAPs	Expérimentale
Modification covalente de substrats variés	P356	GTase	Analyse <i>in silico</i>
	P335	GTase	Analyse <i>in silico</i>
	P176	GTase	Analyse <i>in silico</i>
	P154	Protéine acétyl-transférase	Analyse <i>in silico</i>
	P310	Queueine/archéosine ARNt-ribosyl-transférase	Analyse <i>in silico</i>
	P158a	AdoMet dépendante méthyl-transférase	Analyse <i>in silico</i>
Autre	P399	Transporteur acide aminé	Analyse <i>in silico</i>
	P436	AAA+ (ATPase associée avec différentes activités cellulaires)	Analyse <i>in silico</i>

Ceci est un nouveau mécanisme d'initiation de réplication de l'ADN et de sa terminaison, et est le premier exemple d'une protéine Rep qui n'est pas liée à une réplication « *rolling-circle* » (Oke *et al.*, 2011).

Néanmoins, l'analyse transcriptomique réalisée lors de l'infection de *S. islandicus* par le virus SIRV2 démontrent de façon inattendue une très faible expression de l'ORF119c dont le pic de transcription a lieu à la fin de l'infection. Ceci est en contradiction avec le rôle prédit de ORF119c pour la réplication du génome de SIRV2 (Kessler *et al.*, 2009; Quax *et al.*, 2013).

d. Transcription de l'ADN

Les mécanismes de régulation de l'expression des gènes chez les virus d'archées hyperthermophiles sont très peu étudiés.

Au tout début des recherches sur les virus d'archées, les seules études systématiques de la transcription étaient conduites chez le virus SSV1 de *Sulfolobus* (Reiter *et al.*, 1987; 1988). Ces résultats permirent l'identification des séquences promotrices des archées, similaires à celles retrouvées chez les eucaryotes corroborant les premières observations concernant la ressemblance entre la machinerie de transcription des archées et l'ARN polymérase II eucaryote (Prangishvili *et al.*, 1982; Reiter *et al.*, 1988). Cependant, ces études suivaient une induction de la production de virus SSV1 par UV, plutôt qu'un cycle d'infection (Reiter *et al.*, 1987; 1988).

Plus tard, les profils de transcription des rudivirus SIRV1 et SIRV2 ont été analysés par des expériences d'hybridation Northern, à partir de trente minutes jusqu'à trois heures après l'infection, en utilisant comme sondes des fragments des génomes viraux produits par PCR ou des sondes spécifiques aux ORFs (Figure 8) (Kessler *et al.*, 2004).

Dans le cas de SIRV1/2, il semble que tous les gènes soient transcrits plus au moins simultanément dès le début de l'infection car à trente minutes post-infection le transcrit de presque n'importe quel gène de SIRV1/2 peut être observé. La transcription commence presque simultanément à partir de vingt-et-un sites dans le génome de SIRV1 et à partir de vingt-neuf sites dans le génome de SIRV2, puis s'étend le long du génome. Aucune spécificité de brin n'a été observée dans le début de la transcription : les deux brins sont transcrits avec une efficacité égale. Les transcrits sont souvent assez grands pour couvrir plusieurs ORFs, suggérant que plusieurs ORFs sont regroupées en opérons.

Seules quatre ORFs spécifiques à SIRV2 ne sont apparemment pas transcrits (ORF15, ORF90, ORF116 et ORF76). Aucune transcription n'a été également détectée pour les ORF119c (protéine Rep d'initiation de la réplication), ORF59 (protéine putative de liaison à l'ADN) et ORF91 de SIRV1 et SIRV2, pour l'ORF101 de SIRV1 et pour l'ORF103c de SIRV2 (protéines dont les fonctions sont inconnues) (Kessler *et al.*, 2004). La signification biologique de ces observations n'est toujours pas complètement comprise.

Très peu de transcrits apparaissent tardivement, en plus de ceux observés trente minutes après l'infection. Peu avant l'assemblage des particules des deux virus, en plus des transcrits déjà existants, les transcrits des ORF399 (transporteur putatif d'acides aminés) et ORF134 (MCP) lesquels étaient transcrits en opérons, au début du cycle apparaissent comme monocistroniques. Ils sont détectés proche de la période d'éclipse du cycle de l'infection, deux heures après l'infection dans l'hôte de SIRV2 et trois heures dans l'hôte de SIRV1 (Figure 8) (Kessler *et al.*, 2004).

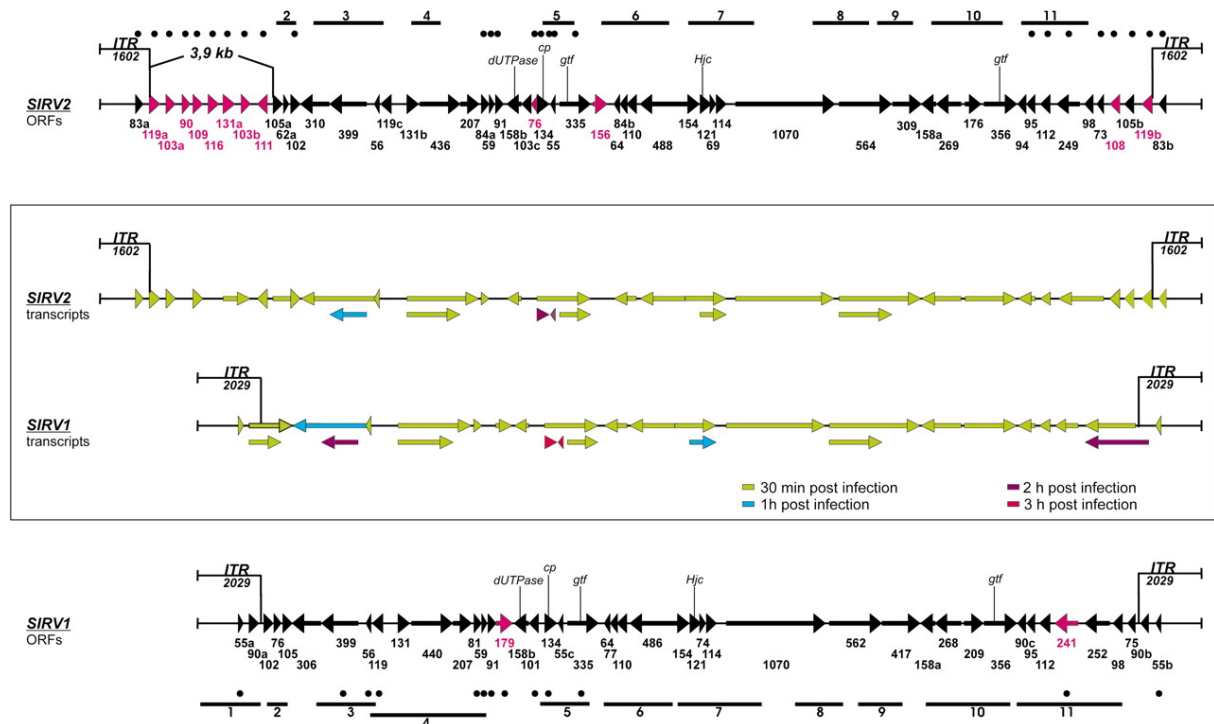


Figure 8 : La carte de transcription des virus SIRV1 et SIRV2. Les nombres utilisés pour identifier les ORFs réfèrent au nombre de codons. Les ORFs spécifiques pour soit le virus SIRV1 (variant VIII) soit SIRV2 sont soulignées en rose. Les gènes avec des fonctions identifiées ou putatives sont indiquées comme tels : *cp*, protéine de capsid; *Hjc*, résolvasse de jonction Holliday; *dUTPase*, déoxyuridine triphosphatase; *gtf*, glycosyl-transférase. En vert = 30minutes post-infection, en bleu = 1heure post-infection, en violet foncé = 2heures post-infection et en rouge = 3heures post-infection (Kessler *et al.*, 2004).

Les sites d'initiation de la transcription de treize transcrits du génome de SIRV1 ont été déterminés par « *primer extension* » et les séquences de leurs promoteurs ont été identifiées. Cela comprend 52% de tous les transcrits identifiés. Dans tous ces cas, des *TATA-box* potentielles ont été identifiées vingt à vingt-cinq nucléotides en amont du site d'initiation ainsi qu'une région adjacente riche en purine, qui correspond probablement à la séquence BRE (pour « *transcription factor B responsive element* »). La même organisation des promoteurs a été caractérisée chez leurs hôtes *Sulfolobus* (Bell *et al.*, 1999). Un trinuécléotide GTC immédiatement en amont des *TATA-box* présumées est retrouvé dans 70% des cas et pourrait jouer un rôle dans la cis-régulation spécifique aux rudivirus. Les transcrits des régions conservées des génomes de SIRV1 et SIRV2 sont très similaires en taille et en contenu. Il est probable que les sites d'initiation de la transcription soient aussi similaires (Kessler *et al.*, 2004) (Figure 9).

```

SIRV1-ORF399  TGCTGACTTTATTTATGTCCTTAGCAATTAGCTAGCTTATG*
SIRV2-ORF399  TTGAGTCTTTATTTATGTCCTTAGTAACTAGCTAGCTTATG

SIRV1-ORF56   AAAAAGCATATAAAAATGTCATAATTCAAAATCTAAAATGAGGAAAAATG*
SIRV2-ORF56   AAAAAGCATATAAAAATGTCATAATCGAAAATCATAAACAGAGAAAAATG

SIRV1-ORF131  AAAAAGCATATAAAAATGTCATAATTAAAAGATATTTT*TAGGGGAAGAAACATG
SIRV2-ORF131b AAAAAGCATATAAAAATGTCATAATTAAAAGATATTTT*TAGAGGCGAAAAACATG

SIRV1-ORF81   AAAAAGCATATAAAAATGTCATAATATAAAAATGTA AAAATAGAGAACATG
SIRV2-ORF84a  AAAAAGCATATAAAAATGTCATAATCAAAATGTA AAAATG

SIRV1-ORF158b TGGGAGATATAAATATGTCAGAAAGTAATAACTAAAATGAGAAAAAGATG*
SIRV2-ORF158b TGGGAAACATAAATATGTCAGAAACTGAAAATAAGACTGATAAAAATATG

SIRV1-ORF134  ACCTGAAAATTTATTAATTTAAAGAATAAAAATGATAT*TGACCAAAAATGAC
SIRV2-ORF134  ACCTGAAATTTATTAATTTGAAGAATAAAAATGATAT*TGACCAAAATATG

SIRV1-ORF55c  AATAGCATTTAAAATGTCAAAATCCAAATTTCTCCTTATG
SIRV2-ORF55   ACTAACATTTAAAATGTCAAAATCAAAATTTCTCCTTATG

SIRV1-ORF335  AGTCAGCTTTATTTTGTGCATAATTTTATTTTATCTCATG
SIRV2-ORF335  ATTTAGCTTTATTTTGTGCATAATTTTATGATTTCAATTATG

SIRV1-ORF486  TTTGTAAGTTTATTTCTACTTGAGATATATAAAT*TGATAGCTATG
SIRV2-ORF488  -----

SIRV1-ORF154  CATAAATGATATAAATGTCATACATTATGATTTAATAAATG
SIRV2-ORF154  GTATATGTATAAATGTCATACAGTGTCAATTTAATCTATG

SIRV1-ORF1070 TAGTAAATGTTTATAATTTGAAAAGAAACATTTT*TAATCAATATG
SIRV2-ORF1070 TAGTAAATGTTTATAATTTGAAAAGAAACATTT*TAATCAATATG

```

Figure 9 : Exemples de séquences nucléotidiques des promoteurs SIRV1 et SIRV2. Les positions relatives de la *TATA-box*, des éléments BRE et GTC sont montrés en rose, orange et vert, respectivement. Les sites d'initiation de la transcription sont indiqués par un astérisque (Kessler *et al.*, 2004).

Chez SIRV1, les deux sites où certaines données permettent de supposer l'existence d'un contrôle transcriptionnel fort sont les régions intergéniques entre l'ORF56b (protéine SvtR avec un domaine de liaison RHH, Guillière *et al.*, 2009) et l'ORF131 et entre l'ORF158b et l'ORF134 (MCP).

Pour promouvoir la transcription de ses gènes, il a été démontré que SIRV1 coopte la protéine activatrice de la transcription Sta1, codée par l'hôte. Cette protéine a la particularité de posséder un domaine HTH de liaison à l'ADN (Kessler *et al.*, 2006). Un autre candidat impliqué dans l'expression des gènes viraux est la protéine SvtR (ORF56b) qui possède un domaine typique RHH de liaison à l'ADN. SvtR réprime et active la transcription de plusieurs promoteurs viraux dans les expériences *in vitro* (Figure 7) (Guillière *et al.*, 2009). Ces deux protéines ont été étudiées au cours de cette thèse.

En plus de la protéine SvtR, plusieurs autres protéines putatives de liaison à l'ADN sont prédites par des approches structurales et de génomique comparative (Prangishvili *et al.*, 2006). Elles incluent P83a/P83b (motif HTH), P59b (motif RHH), P55 (domaine de liaison zinc) et P114 (Prangishvili *et al.*, 2013) (Tableau 4). Bien que la fonction exacte de ces protéines soit à déterminer, la présence de domaines typiques de liaison à l'ADN suggère qu'elles pourraient être impliquées dans la régulation de promoteurs viraux et/ou cellulaires. Enfin, la protéine P114 est d'un intérêt particulier car elle est très conservée, pas seulement chez les ruidivirus et les lipothrixvirus, mais aussi chez d'autres virus de crénaarchées (Larson *et al.*, 2007).

Des études transcriptomiques ont été réalisées récemment par Quax *et al.*, (Quax *et al.*, 2013) chez la souche *S. islandicus* LAL14/1 pour laquelle j'ai établi un système génétique au cours de ma thèse. Les résultats obtenus ont permis de définir une carte générale de l'expression des gènes hôtes et viraux pendant le cycle d'infection du virus lytique SIRV2, chez son hôte *S. islandicus* LAL14/1 et de suivre

les changements dans l'expression des gènes viraux et hôtes pendant le cycle infectieux. Cette approche a confirmé une faible régulation temporelle de transcription du génome viral, comme déjà observé précédemment (Kessler *et al.*, 2004), mais a permis de détecter des changements importants dans l'expression des certains gènes cellulaires à la suite de l'infection par SIRV2.

Les temps sélectionnés pour l'analyse de l'ARN-seq était t=0, 1, 2, 3, 5, 7 et 9h post-infection, selon les informations disponibles sur le cycle du virus (Bize *et al.*, 2009).

Le nombre de « reads » augmente régulièrement au cours de l'infection, atteignant un plateau autour de cinq heures après l'infection. Les « reads » viraux au début de l'infection mappent typiquement aux ORFs des extrémités du génome linéaire, confirmant que la transcription commence simultanément aux deux extrémités du génome de SIRV2.

Quelques minutes après l'infection (t=0h), les gènes les plus exprimés sont les ORFs 83a et 83b (protéine putatives de liaison à l'ADN avec un motif HTH) localisés chacune à une de deux extrémités du génome. Elles ont des séquences nucléotidiques identiques et peuvent être distinguées seulement d'après leurs régions non traduites (Quax *et al.*, 2013). Tous les autres gènes exprimés à ce point sont localisés près des extrémités du génome : ORF119a (fonction inconnue), ORF103a et ORF119b (fonction inconnue).

A chaque point de l'infection, excepté t=0h post-infection, une expression significative de tous les gènes viraux est détectée et l'expression génique semble être distribuée au hasard sur le génome de SIRV2.

A t=1h après l'infection, les gènes les plus abondamment transcrits sont : ORF83a et ORF83b (protéines putatives de liaison à l'ADN avec un motif HTH), ORF56b (régulateur transcriptionnel SvtR contrôlant, entre autres, le gène *gp30* codant la protéine de capsid, (Guillière *et al.*, 2009)), ORF108 et ORF103a (protéines sans fonctions connues).

De façon intéressante, les expériences de double-hybride (Quax *et al.*, 2013) ont révélé une interaction entre la résolvasse de jonction Holliday et les protéines P83 dont les gènes sont fortement exprimés au cours de l'infection du virus (protéines putatives de liaison à l'ADN avec un motif HTH).

Au stade avancé de l'infection, l'expression des gènes codant pour les protéines structurales augmente, de telle sorte que 35% des « reads » viraux mappent au gène codant pour P134, la protéine majeure de capsid et 13% au gène codant pour P98, la protéine composant les structures pyramidales de sortie du virus SIRV2 (voir paragraphe suivant).

Après l'infection par SIRV2, des changements complets dans l'expression génique de *S. islandicus* LAL14/1 sont observés. Plus d'un tiers de tous les gènes sont transcrits différemment, en particulier les gènes impliqués dans la division cellulaire et la défense contre les éléments génétiques étrangers.

L'expression des gènes impliqués dans la division cellulaire diminue, notamment les homologues ESCRT-III et les composants de l'opéron *cdv*, qui comprend *cdvA*, *cdvB* et *cdvC* et qui code pour la machinerie de division cellulaire des *Sulfolobus*, responsable de la constriction des cellules en division (Lindas *et al.*, 2008; Samson *et al.*, 2011). De façon intéressante, l'expression de l'opéron *cdv* diminue dix fois à la suite de l'infection.

Au contraire, l'expression des gènes jouant un rôle dans la défense anti-virale est activée à grande échelle. L'expression des gènes appartenant aux systèmes toxine-antitoxine et CRISPR-Cas est activée de façon remarquable dès t=1h post-infection. Le différent degré d'activation des diverses parties du systèmes CRISPR-Cas (six blocs CRISPR-cas chez *S. islandicus* LAL14/1) souligne les fonctions spécialisées qu'ils réalisent (Quax *et al.*, 2013).

Tableau 5 : Expression génique de SIRV2 pendant l'infection de *S. islandicus* LAL14/1. Adapté de Quax *et al.*, 2013.

Numéro gp	Numéro d'accension GenBank	Nom	t0	t1	t2	t3	t5	t7	t9
1	NP_666535.1	ORF83a	6062	234957	114130	102725	124763	122335	145288
2	NP_666536.1	ORF119a	1844	18097	18415	11469	10876	10111	15075
3	NP_666537.1	ORF103a	1579	55032	31152	22256	19755	17921	21598
14	NP_666548.1	ORF399	10	2309	20762	28828	39014	36691	37579
15	NP_666549.1	ORF56b	506	119860	176877	133456	94962	65015	74368
16	NP_666550.1	ORF119c	0	31	79	79	110	154	249
26	NP_666560.1	ORF134	11	1632	204021	352185	500990	494474	642358
27	NP_666561.1	ORF55	15	6298	9896	8268	8152	9753	14292
38	NP_666572.1	ORF1070	2	480	10110	14126	21404	24211	30762
49	NP_666583.1	ORF98	35	5141	64921	94658	156902	189222	239929
50	NP_666584.1	ORF73	0	291	432	401	448	392	534
51	NP_666585.1	ORF108	332	61453	49118	30234	20815	14979	19408
52	NP_666586.1	ORF105b	840	5937	4530	2846	2818	2835	4261
53	NP_666587.1	ORF119b	1525	19554	17692	10939	10603	11636	17028
54	NP_666588.1	ORF83b	3031	117479	57065	51363	62382	61168	72644

e. Cycle de l'infection

Parmi les quatre rudivirus connus, seuls les virus SIRV1 et SIRV2 sont capables d'infecter les cellules appartenant au genre des *Sulfolobales*, dont les souches *S. islandicus* LAL14/1 et E233S sont utilisées dans ma thèse. Ces deux virus sont génétiquement très proches (Peng *et al.*, 2001). Le virus SIRV2 est un modèle intéressant pour l'étude des interactions hôte-virus.

La période d'éclipse, qui correspond à l'intervalle de temps entre l'infection et l'apparition de particules intracellulaires, est de quatre heures pour SIRV1 et deux heures et demi pour SIRV2. La période de latence, qui correspond à l'intervalle de temps entre l'infection et la libération des premières particules virales, est de huit heures pour SIRV1 et six heures pour SIRV2.

En raison de l'absence d'indications visibles de lyse cellulaire, il a été initialement pensé que les rudivirus n'étaient pas lytiques (Prangishvili *et al.*, 1999). Ce point de vue a été contesté par l'analyse détaillée récente du cycle de SIRV2 (Bize *et al.*, 2009). SIRV2 est un virus lytique agissant de façon inhabituelle sur la cellule hôte. L'infection se développe selon un ordre bien défini supposant une coordination parfaite dans l'expression des gènes viraux impliqués. A partir du stade précoce et jusqu'à environ huit heures après l'infection une dégradation massive du chromosome de l'hôte s'observe en même temps que la synthèse de l'ADN viral. Cela s'accompagne de la formation de structures pyramidales à sept faces à la surface de la cellule hôte qui rompt le S-layer (à partir de dix heures post-infection) (Figure 10). Vers la fin de la phase éclipse, ces structures pyramidales s'ouvrent vers l'extérieur créant des ouvertures à travers lesquelles les virions matures s'échappent de la cellule (à partir de quatorze heures post-infection) (Bize *et al.*, 2009; Quax *et al.*, 2010). Il faut souligner que la DO de la culture infectée ne diminue pas drastiquement comme attendu. La raison de cette anomalie est simple. Le S-layer rigide maintient la structure de la cellule et la transforme en « fantôme » (Bize *et al.* 2009). A part des études microbiologiques des valeurs de CFU, seule la dégradation de l'ADN chromosomique ou l'observation au microscope électronique des cellules permettent d'identifier la mort cellulaire.

Ce mécanisme de libération du virus est original et très différent des mécanismes de lyse et de sortie décrits chez les bactériophages et des virus des eucaryotes. Un mécanisme similaire a été décrit pour le virus STIV qui infecte également les cellules *Sulfolobus* (Snyder *et al.*, 2013). Cela est surprenant que SIRV2 et STIV partagent ce même système de lyse étant donné que SIRV2 et STIV appartiennent à deux familles virales non apparentées; les morphologies de leurs particules ne se ressemblent pas et aucune ressemblance de leurs séquences génomiques n'est détectée.

La protéine P98 codée par le génome de SIRV2 a été montrée comme le seul constituant de ces ultra structures cellulaires pyramidales (Quax *et al.*, 2010) tout comme la protéine C92, codée par le génome de STIV (Snyder *et al.*, 2013). En exprimant la protéine P98 chez *Sulfolobus acidocaldarius* ou encore chez *Escherichia coli*, il a été démontré que cette protéine est capable de s'auto-assembler pour former des structures pyramidales traversant la membrane cellulaire (Quax *et al.*, 2011).

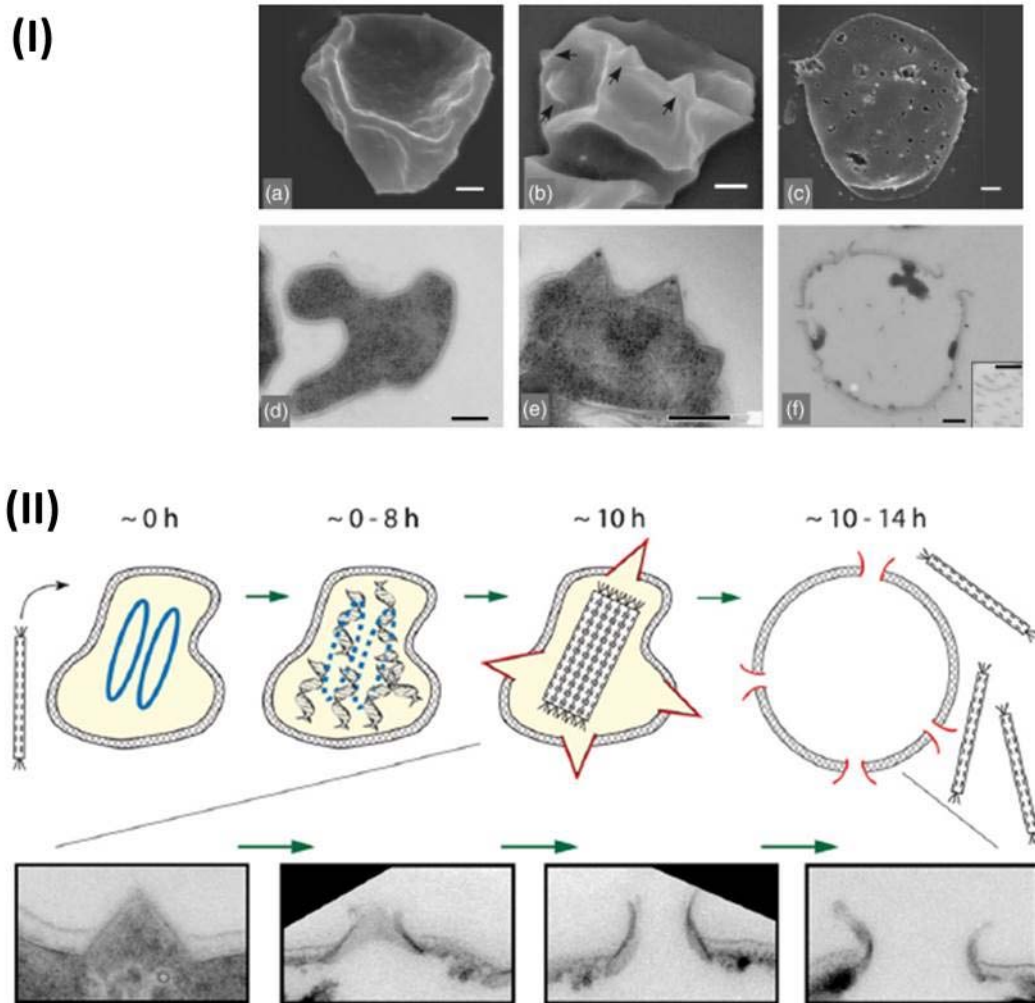


Figure 10 : Cycle de l'infection du *Sulfolobus islandicus rod-shaped virus 2* (SIRV2). (I) Observation des cellules hôtes *Sulfolobus* infectées par SIRV2, sous microscope à balayage électronique (A, B, C) et sous microscope à transmission électronique (D, E, F). (A, D) Cellules non infectées. (B, E) Cellules 10h après l'infection. (C, F) Cellules 10-14h après l'infection. Barre d'échelle : 200nm. (II) Représentation schématique des étapes majeures du cycle d'infection du virus SIRV2 dans les cellules *Sulfolobus*. Le temps après l'infection est indiqué en heures. A 0h, deux chromosomes de *Sulfolobus* sont indiqués en bleu. Plus tard, entre 0 et 8h, ils sont dégradés de manière concomitante avec la synthèse d'ADN viral (hélices grises). A 10h, les VAPs (*virus associated pyramids*) (en rouge) et les groupes intracellulaires de virions assemblés sont montrés. Enfin, à différents points entre 10 et 14h, les VAPs s'ouvrent (restes des VAPs montrés en rouge), la cellule se lyse, et les virions sont sortent (modifié d'après Bize *et al.*, 2009).

2. Famille Lipothrixviridae

La famille *Lipothrixviridae* (du latin « lipo » pour gras, et du grec « thrix » pour cheveux) comprend quatre genres (Tableau 2) dont les membres démontrent des morphologies distinctes de virions.

a. Structure du virion et protéines

Les virions de cette famille sont des filaments enveloppés flexibles qui varient de 900 à 2200 nm en longueur et de 24 à 38 nm de diamètre. Les enveloppes sont constituées de protéines virales et de lipides dérivés de l'hôte. Le noyau hélicoïdal de nucléoprotéines contient une seule molécule d'ADN db linéaire (15.9 à 56 kbp de long) (Janekovic *et al.*, 1983).

Chaque extrémité du virion est fuselée et porte des structures terminales spécifiques. Ces structures peuvent représenter des pinces (AFV1), des pioches (AFV9), des pointes ressemblant à des goupillons (AFV2), des structures en forme de balai (SIFV) dotés de trois (AFV1) ou six (SFV) courts filaments ou (Figure 11a).

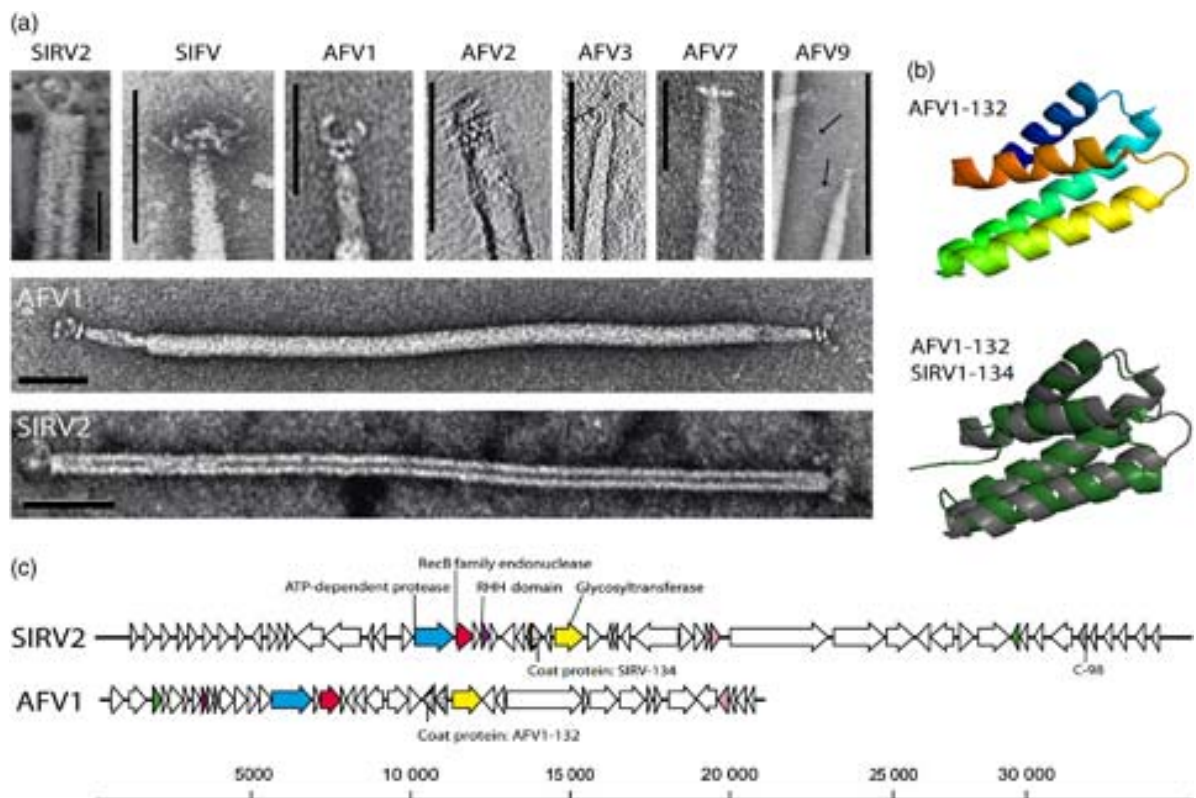


Figure 11 : Les « Ligamenvirales ». (a) Virions représentatifs des *Rudiviridae* et des *Lipothrixviridae* et leurs structures terminales sous microscope à transmission électronique. Les barres d'échelles représentent 100nm. (b) Représentation des structures cristallines des protéines majeures de capsidie des virus AFV1 et SIRV1 (ce dernier est en gris et est superposé à la structure de la protéine d'AFV1 en verte). (c) Alignement graphique des génomes des représentants des *Lipothrixviridae* et des *Rudiviridae*. Les protéines homologues sont de la même couleur et les fonctions identifiées sont indiquées (adapté de Pina *et al.*, 2011).

Les protéines du *Gammalipothrixvirus* AFV1 ont été étudiées plus en détails. Le virion porte deux protéines majeures de capsides, P132 et P140, dont les structures ont été résolues (Goulet *et al.*, 2009). Les deux présentent une nouvelle structure en forme de faisceau à quatre hélices. La protéine P140 portent un domaine C-terminal extra, qui grâce à ses propriétés lipophiles peut interagir avec l'enveloppe du virion. De plus, chacune de ces deux protéines sont capable de lier l'ADN de manière non spécifique et peuvent polymériser le long de l'ADN, en formant des filaments flexibles.

b. Organisation du génome

Les extrémités des génomes linéaires des *Lipothrixvirus* sont modifiées. Les génomes des membres de tous les genres portent de longues répétitions terminales inversées (ITR). Le *gammalipothrixvirus* AFV1 est une exception et porte de très courtes répétitions terminales inversées CG₁₀. Ces courtes répétitions sont précédées par une séquence d'environ 350pb qui contient des groupes de courtes répétitions directes de la séquence TTGTT ou de proches variants, ressemblant donc à l'organisation structurale des extrémités des télomères des chromosomes linéaires eucaryotes (Bettstetter *et al.*, 2003).

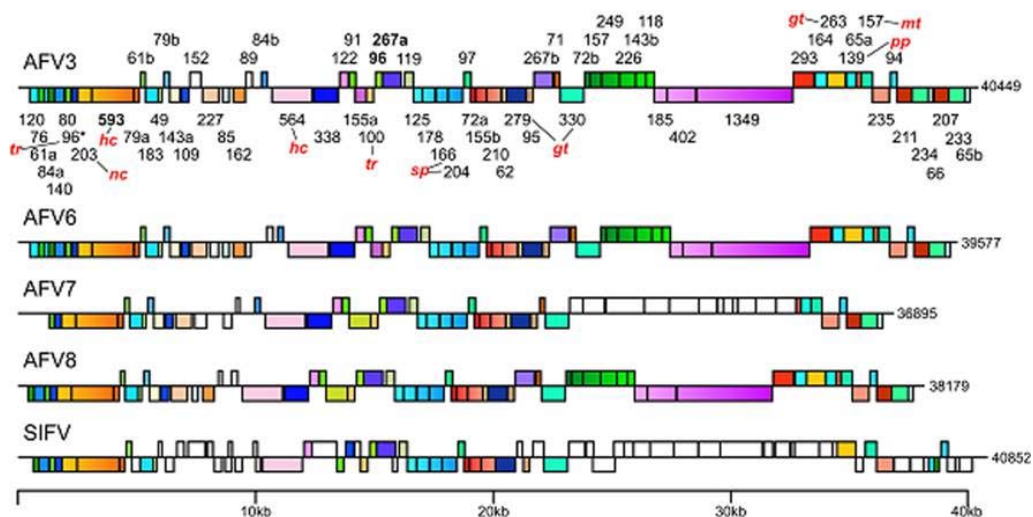


Figure 12 : Alignement des génomes des *Betalipothrixviridae*. La taille relative, la localisation et l'orientation des gènes prédits est montrée. Les gènes homologues sont de la même couleur. Les opérons homologues portant trois gènes ou plus sont montrés avec un dégradé de même couleur. Les boîtes blanches indiquent qu'aucun homologue n'a été retrouvé dans les autres génomes. Les ORFs d'AFV3 sont appelées selon leur taille en acides aminés. Les fonctions prédites des protéines sont abrégées ainsi : hc, hélicase; gt, glycosyltransférase; mt, méthyltransférase SAM-dépendante; nc, nucléase; pp, protéine phosphatase; sp, protéine structurale; et tr, régulateur transcriptionnel. Les ORFs qui donnent des correspondances dans les séquences CRISPR des chromosomes des crénarchées sont indiquées en gras (modifié de Vestergaard *et al.*, 2008).

Les génomes des six membres espèces du genre *Betalipothrixvirus* révèlent un niveau élevé de conservation à la fois, dans le contenu et dans l'ordre des gènes sur de grandes régions (Figure 12). D'autres caractéristiques partagées des génomes sont : 1) la présence de grandes répétitions terminales inversées, exhibant des répétitions directes conservées, régulièrement espacées, 2) l'existence d'un opéron très conservé codant pour deux protéines majeures structurales, et 3) la présence de multiples ORFs chevauchants qui peuvent être l'indicateur de reprogrammation génétique.

Quelques produits de gènes prédits, outre les protéines structurales, peuvent être assignées à des fonctions qui incluent entre autres une putative hélicase, une nucléase (Goulet *et al.*, 2010), une protéine phosphatase, une glycosyltransférase, de liaison à l'ADN (Guillière *et al.*, 2013) et des régulateurs transcriptionnels. Aucune information n'est disponible sur la réplication de l'ADN chez aucun des *Lipothrixvirus*.

3. La transcription chez les Archées

1. ARNs polymérase ADN-dépendantes dans les trois domaines du vivant

La transcription est un processus cellulaire fondamental à travers les trois domaines du vivant comme en témoigne l'ubiquité de l'ARN dans tous les organismes vivants. Cela signifie que toutes les formes de vie dépendent d'enzymes capables de synthétiser des ARNs. Les ARN polymérase ADN-dépendantes (ARNP) sont responsables de la transcription de l'ensemble de l'information génétique présente dans un génome et les gènes codant pour ces enzymes multiprotéiques sont obligatoirement présents dans les génomes de tous les organismes cellulaires.

Ce sont les travaux de B.G. Louis et P.S. Fitt en 1971 (Louis and Fitt, 1971) sur l'halophile extrême *Halobacterium cutirubrum*, puis en 1978 ceux de W. Zillig et C. Stetter (Zillig *et al.*, 1978) sur *Halobacterium halobium* qui ont permis pour la première fois la purification et l'analyse de la composition de l'ARNP d'une archée. De façon surprenante, ils isolèrent une enzyme qui ne ressemblait, ni par son poids moléculaire, ni par le nombre de sous unités la composant, à une ARN polymérase bactérienne. Le manque évident de similitude entre la composition de l'ARNP de cette archée et la composition caractéristique générale ($\beta'\beta\alpha 2\omega$) chez les bactéries était en accord avec les idées de C. Woese concernant la structure de l'arbre universel du vivant (Zillig *et al.*, 1978). L'analyse plus approfondie réalisée sur d'autres espèces d'archées halophiles (Madon and Zillig, 1983) et chez les crénarchées thermoacidophiles (Prangishvili *et al.*, 1982) confirmèrent les premières observations de W. Zillig sur la complexité de la composition de l'ARNP des archées. Les structures cristallines des cinq sous-unités de l'ARNP bactérienne ont été définies en 2001 (Campbell *et al.*, 2001), et celles des douze sous-unités de l'ARNP II eucaryote de *Saccharomyces cerevisiae* en 2005 (Armache *et al.*, 2005). Ces structures ont attesté clairement qu'il existe une relation évolutive rapprochée entre les ARNPs des trois domaines du vivant. En fin de compte, avec la résolution des structures cristallines des ARN polymérase issues de *Sulfolobus solfataricus* et *Sulfolobus shibatae* en 2008 et 2009, respectivement, une image complète de cette enzyme a été dévoilée et le nombre total de sous unités chez les archées a été établi à treize (Kusser *et al.*, 2007; Hirata *et al.*, 2008; Korkhin *et al.*, 2009). La Figure 13 montre la structure cristalline de l'ARN polymérase ADN-dépendante de *S. shibatae*. Pour faciliter la comparaison, les structures de l'ARNP de *Saccharomyces cerevisiae* et celle de l'enzyme bactérienne provenant de *Thermus aquaticus* sont représentées sur la même figure.

Les ARNPs des trois domaines sont liées d'un point de vue évolutif et dérivent d'une enzyme ancestrale commune présente déjà chez LUCA (**L**ast **U**niversal **C**ommon **A**ncestor) - l'hypothétique organisme référé comme le dernier ancêtre commun universel existant la séparation des trois domaines du vivant (Figure 13).

Dans les trois domaines du vivant, la nomenclature des sous-unités des ARNPs est différente (Tableau 6). Les sous-unités de l'ARNP d'archée sont nommées Rpo (**R**NA **p**olymerase) suivi par un nombre (Rpo1, Rpo2, etc.); une classification plus ancienne utilise Rpo suivie par une lettre (RpoA, RpoB, etc.) ou seulement une lettre (A, B, etc.). L'ARNP II des organismes eucaryotes utilise l'acronyme RPB (**R**NA **p**olymerase **B**, soit l'ARNP II) suivi par la même numérotation que les sous-unités archées (par exemple, RBP7 correspond à Rpo7). Les sous-unités bactériennes sont appelées par des lettres grecques (α , β , β' et ω).

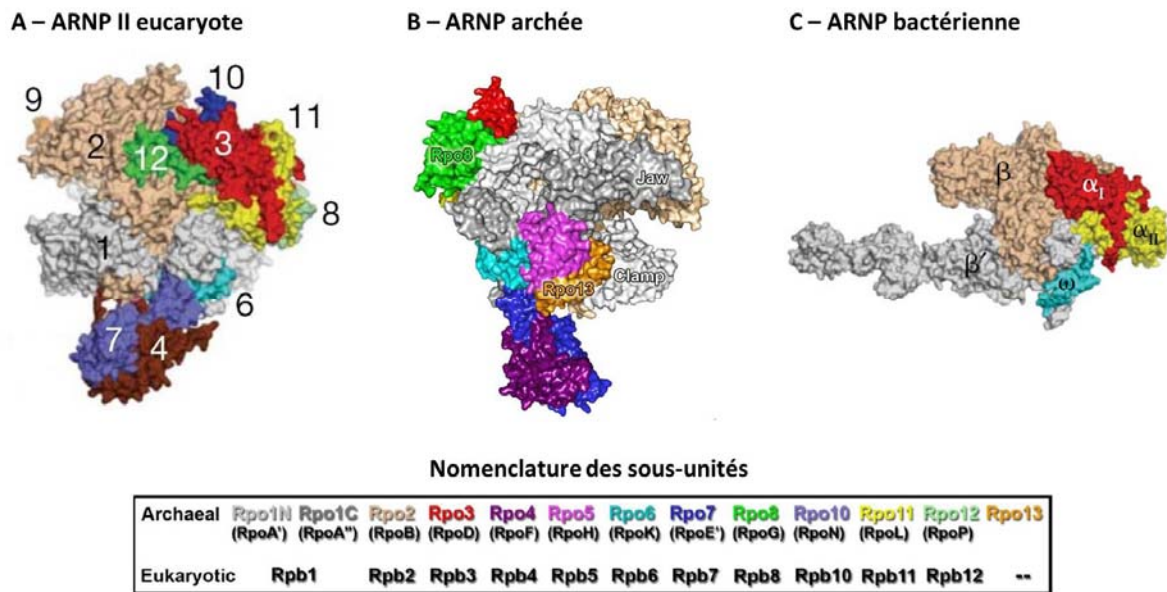


Figure 13 : Structures des ARN polymérases-ADN dépendantes à sous-unités multiples de (A) *Sacharomyces cerevisiae*, (B) *Sulfolobus shibatae* et (C) *Thermus aquaticus*. Les sous-unités orthologues sont représentées dans la même couleur. ARN polymérase est abrégée par ARNP. La nomenclature des sous-unités de l'ARNP archée est donnée avec celle de l'ARNP II eucaryote, pour faciliter la comparaison (modifié de Hirata *et al.*, 2008; Korkhin *et al.*, 2009).

L'organisation des sous-unités de toutes les ARNPs reflète une architecture commune basée sur cinq sous-unités, liées par leurs structures et leurs fonctions et qui forment le noyau universellement conservé de l'ARNP (Werner, 2007) (Tableau 6). Si les orthologues de cinq sous unités décrites chez les bactéries ($\beta'\beta\alpha 2\omega$) existent chez les deux autres domaines, les *Eukarya* et les *Archaea* possèdent en plus un groupe de sous-unités homologues qui leur sont spécifiques et qui n'ont pas d'équivalents chez les bactéries (Tableau 6). Au moins une sous unité (RPB9) est spécifique aux eucaryotes ainsi qu'un certain nombre de facteurs de transcription non identifiés chez les archées étudiées (Tableau 6). Les archées codent pour la sous-unité Rpo13 absente chez les eucaryotes.

La forme globale d'une ARNP ressemble à une « pince de crabe » (*crab claw*), avec de grandes sous-unités formant la partie majeure de chacun des bras de la pince, qui lie l'ADN entre ses pinces et contient le site actif au centre (Werner *et al.*, 2007). L'architecture du site actif de l'enzyme est aussi très conservée dans les trois domaines (Sweetser *et al.*, 1987; Hirata *et al.*, 2008). Il faut noter que la caractéristique la plus marquante qui distingue les ARNPs d'archée et eucaryote de l'ARNP bactérienne est la présence d'une tige dont le rôle sera détaillé plus loin dans le texte (Grohmann and Werner, 2009).

Les sous-unités de l'ARNP bactérienne sont conservées à travers les trois domaines. Le cœur minimal catalytique de l'ARNP possède tous les facteurs déterminants nécessaires pour la transcription, même faible et non spécifique. Chez les bactéries, la présence d'un facteur σ (sigma) dissociable est requis pour adresser l'enzyme aux promoteurs spécifiques, et donc former l'holoenzyme complète, capable d'initier la transcription (Murakami *et al.*, 2003).

Tableau 6 : Les sous-unités conservées des ARN polymérase et des facteurs généraux de transcription dans les trois domaines de vie (modifié de Werner and Grohmann, 2011).

BACTERIES	ARCHAEA	EUKARYA				
		ARNPII	ARNPIII	ARNPI	Plante ARNPIV	Plante ARNPV
Sous-unités ARNP						
Sous-unité β	Rpo1N (RpoA') Rop1C (RpoA'')	RPB1	C160	A190	NRPD1	NRPE1
Sous-unité β	Rpo2 (RpoB)	RPB2	C128	A135	NRPD/E2	NRPD/E2
Sous-unité α	Rpo3 (RpoD)	RPB3	AC40	AC40	RPB3	RPB3
Sous-unité α	Rpo11 (RpoL)	RPB11	AC19	AC19	RPB11	RPB11
Sous-unité ω	Rpo6 (RpoK)	RPB6	RPB6	RPB6	RPB6	RPB6
	Rpo5 (RpoH)	RPB5	RPB5	RPB5	RPB5	NRPE5
	Rpo8 (RpoG)	RPB8	RPB8	RPB8	RPB8	RPB8
	Rpo10 (RpoN)	RPB10	RPB10	RPB10	RPB10	RPB10
	Rpo12 (RpoP)	RPB12	RPB12	RPB12	RPB12	RPB12
	Rpo4 (RpoF)	RPB4	C17	A14	NRPD/E4	NRPD/E4
	Rpo7(RpoE)	RPB7	C25	A43	NRPD2	NRPE7
		RPB9	C11	A12	NRPD96	RPB9
	Rpo13					
Facteurs de transcription						
		TFIIF α (RAP74)	C53(C4)	A49		
		TFIIF β (RAP30)	C37(C5)	A34.5		
	TFE α	TFIIE α	C82			
	TFE β /C34	TFIIE β	C34			
			C31			
	TBP	TBP	TBP	TBP		
	TFB	TFIIB	BRF1			
		TFIIA				
		TFIIH				
	TFS	TFIIS	TFIIS			
	Spt4	SPT4	SPT4			
NusG	Spt5	SPT5	SPT5			
NusA	NusA					
Rho						
Facteurs σ						

Les eucaryotes emploient cinq types d'ARN polymérase nucléaires (de type I à type V) pour transcrire les groupes de gènes distincts (Cramer, 2008), alors que les bactéries et les archées utilisent un seul type d'ARNpol (type II) pour toutes les activités transcriptionnelles. L'ARNP I transcrit l'ARN ribosomal (ARNr 18S et 28S), alors que l'ARNP III se charge des petits ARNs nucléaires (ARNsn), de l'ARNr de la petite sous-unité 5.7S et des ARNt (Werner *et al.*, 2009). L'ARNP II est l'enzyme responsable de la transcription des gènes codant pour les ARNm chez tous les organismes eucaryotes. Il s'agit de la version de l'enzyme la plus étudiée, à la fois sur le plan structural et sur le plan fonctionnel (Armache *et al.*, 2005). L'ARNP II contient douze sous-unités, dont la plupart d'entre elles sont des homologues fonctionnels et structurels de l'ARNP archée, bien que des sous-unités avec des domaines spécifiques existent. Les ARNP des types IV et V sont exclusifs aux plantes (Ream *et al.*, 2009; Marcussen *et al.*, 2010).

L'unique enzyme de transcription des archées est composée de treize sous-unités et montrent une forte conservation avec l'ARNP II eucaryote, d'un point de vue structural. La composition et la fonction de cette ARNP seront discutées avec plus de détails dans le chapitre suivant.

Les principales différences entre ces plusieurs types d'enzymes résident dans le choix des facteurs généraux de transcription qui interagissent avec l'ARNP et modulent les trois phases principales du processus : initiation, élongation et terminaison. Les sous-unités auxiliaires présentes chez les archées et les eucaryotes ne sont pas essentielles mais aident pour les interactions avec soit l'ADN, soit les facteurs de transcription généraux qui forment le complexe de pré-initiation (PIC). Par exemple, les bactéries ont besoin seulement de la présence de facteurs σ pour former le PIC, alors que l'ARNP II eucaryote a besoin d'au moins six facteurs généraux de transcription (TFFIIA, TFIIB, TFIIE, TFIIH et TFIIF) et d'un complexe médiateur pour initier la transcription. L'ARNP d'archées peut être considérée comme une version simplifiée de l'enzyme décrite chez les eucaryotes. Il n'y a pas de complexe médiateur chez les archées et, également, *in vivo* l'ARNP des archées requiert la présence de seulement trois facteurs généraux de transcription : TBP, TFB et TFE (Bell and Jackson, 1998). Néanmoins, le système de transcription *in vitro*, établi dans notre laboratoire, est constitué seulement de l'ARNP et des facteurs TBP et TFB, ce qui suffit pour réaliser une transcription efficace.

Pour conclure, le système de transcription ARNP chez les archées constitue un modèle plus simple pour étudier les détails fonctionnels, mécanistiques, structuraux et enzymatiques de la transcription et en particulier de la transcription chez les non-bactéries. De façon étonnante, les régulateurs transcriptionnels d'archées sont très proches des régulateurs bactériens par leurs mécanismes d'action et par les familles de protéines régulatrices impliquées dans cette régulation (Aravind *et al.*, 1999; Kyrpides *et al.*, 1999). Il peut être constaté qu'en général la machinerie de transcription chez les archées est une mosaïque où se mélangent des caractéristiques bactériennes et eucaryotiques. Comprendre son fonctionnement et son évolution permettra de décrire les mécanismes de transcription basiques partagés par tous les domaines du vivant.

2. ARN polymérase d'archées

La structure et les fonctions de toutes les sous-unités individuelles de l'ARNP et des facteurs généraux de transcription sont détaillées dans la section suivante afin de mieux présenter les particularités de la machinerie de transcription chez les archées.

Les treize sous-unités de l'enzyme archée peuvent être divisées en trois groupes fonctionnels : les sous-unités catalytiques, celles assurant l'assemblage et les sous-unités auxiliaires (Tableau 6). L'architecture de l'ARNP et l'organisation des trois groupes aussi bien que d'autres îlots fonctionnels ou structuraux sont présentées sur la figure ci-dessous (Figure 14).

a. Le noyau minimum de l'ARN polymérase archée

La version minimale de l'enzyme ARNP comprend les sous-unités catalytiques et une plateforme d'assemblage. Ce complexe est suffisant pour initier une transcription dirigée à partir des promoteurs naturels. Ceci s'explique par le fait que ce complexe minimal peut être recruté sur les séquences promotrices par les facteurs généraux de transcription. Il est doté de la capacité d'ouvrir la double hélice d'ADN, de placer le brin matrice dans le site actif, de s'échapper du promoteur et d'assurer une élongation efficace (Werner and Weinzierl, 2002).

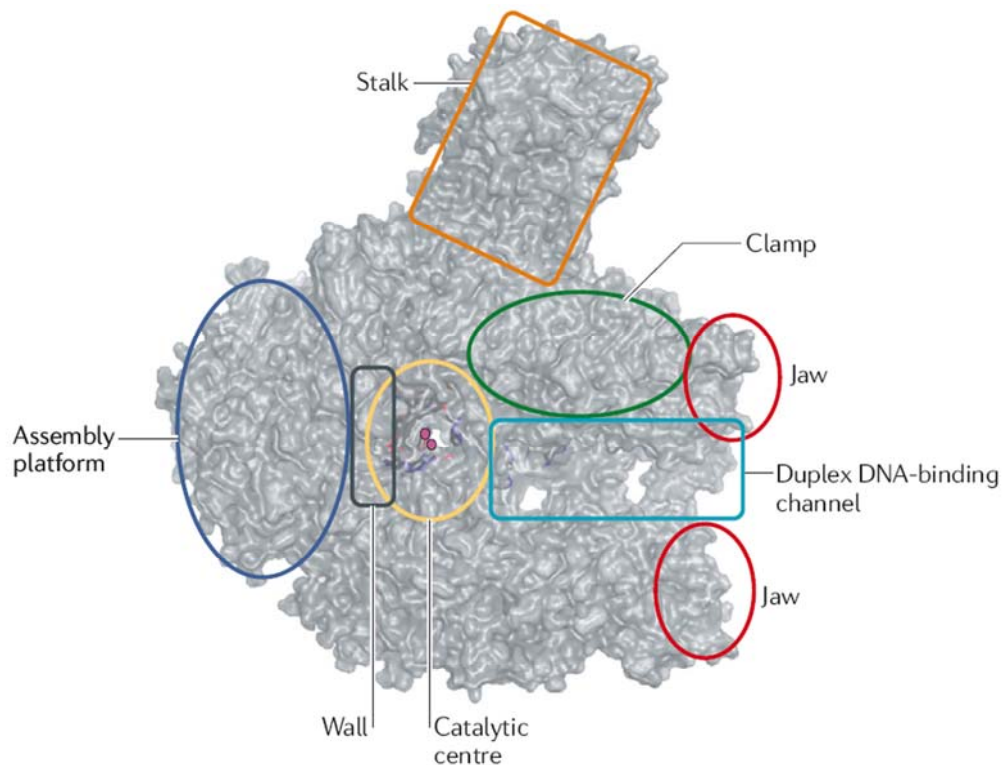


Figure 14 : Organisation et architecture de l'ARN polymérase. Ce schéma simplifié de l'ARN polymérase montre les caractéristiques structurales et fonctionnelles importantes, la plateforme d'assemblage, le site actif, le canal de liaison à l'ADN, les domaines *stalk*, *clamp*, *jaw* et *wall* (modifié de Werner and Grohmann, 2011).

Les sous-unités auxiliaires de l'ARNP qui se lient au complexe minimal contribuent à une meilleure processivité de l'enzyme.

Le noyau catalytique de l'ARNP – Les sous-unités Rpo1 et Rpo2 forment le noyau catalytique de l'enzyme d'archées et sont des homologues des sous-unités bactériennes β '- β et eucaryotes RPB1-2. Ce sont les deux plus grandes sous-unités de l'ARNP et elles forment une « pince de crabe » qui interagit avec le brin matrice ADN db. Ces sous-unités abritent les sites de liaison des NTPs, de l'ADN matrice et de l'hybride ADN-ARN et constituent le site actif de l'enzyme, localisé dans la fente (Werner and Grohmann, 2011).

La plateforme d'assemblage – L'assemblage du noyau catalytique dépend d'une « plateforme » formée par les sous-unités Rpo3-10-11-12, homologues de RPB3-10-11-12 de l'ARNP II eucaryote. Ces sous-unités forment un complexe stable à l'opposé de la « pince de crabe » et du site d'entrée de l'ADN (Werner *et al.*, 2000). Ces quatre sous-unités permettent l'assemblage du noyau et assurent la stabilité de l'ARNP. La comparaison du processus d'assemblage de l'ARNP dans les trois domaines du vivant permet des observations évolutives intéressantes. L'ARNP bactérienne requiert pour l'assemblage, seulement des homo-dimères de la sous-unité α , qui est l'homologue de Rpo3-11. Cela suggère que les sous-unités additionnelles Rpo10-12 ont été acquises après la séparation entre les *Bacteria* et les *Eukarya/Archaea* (Werner *et al.*, 2002).

Rpo10 et Rpo12 remplissent le creux concave sur la sous-unité Rpo2 et agissent comme des adaptateurs structuraux entre Rpo2 et Rpo3, d'où leur rôle dans l'assemblage et la stabilité de

l'ARNP (Werner *et al.*, 2002). Il est possible que ces sous-unités soient impliquées dans d'autres fonctions, comme probablement les interactions avec les facteurs généraux de transcription (Kostrewa *et al.*, 2009).

b. Les sous-unités auxiliaires

La « tige » – La présence d'une tige (*stalk*) est la principale différence structurale entre les ARNP bactérienne et eucaryote/archée. Placée en périphérie de l'ARNP archée, à côté du canal de sortie de l'ARN, la tige est composée des sous-unités Rpo4 et Rpo7. Comme ces sous-unités sont aussi connues sous le nom de RpoF et RpoE, la tige peut être désignée comme « complexe F/E ». La tige est le complexe le mieux caractérisé de l'ARNP des archées et est une pièce polyvalente de l'enzyme, importante pour toutes les phases de la transcription (Grohmann and Werner, 2011). Lors de l'initiation, la tige permet l'ouverture de l'ADN et interagit avec le facteur général de la transcription TFE (Naji *et al.*, 2007; Ouhammouch *et al.*, 2004).

Rpo5 – Rpo5 participe à l'ouverture de l'ADN et aux étapes précoces de l'initiation de la transcription. Elle est en proche contact avec l'extrémité C-terminale de Rpo1, formant la mâchoire qui est impliquée dans les interactions avec l'ADN db en aval. De plus, cette sous-unité subit plusieurs changements conformationnels lors de la transition entre les phases d'initiation et d'élongation de la transcription (Grünberg *et al.*, 2010). Alors que la sous-unité Rpo5 marque la limite en aval des contacts entre le PIC et le brin transcrit par l'intermédiaire de « *crosslinking* », aucun lien entre la sous-unité Rpo5 et le brin transcrit n'est détecté dans le complexe d'élongation.

Rpo8 – Rpo8 est une sous-unité périphérique localisée entre la plateforme d'assemblage et le pore d'entrée des NTPs. Cette sous-unité interagit avec l'extrémité N-terminale de Rpo1. Chez les archées, Rpo8 est seulement présente chez les Crénarchées et les Korarchées (Koonin *et al.*, 2007; Kwapisz *et al.*, 2008; Werner *et al.*, 2011). Ceci reflète la parenté qui existe entre les Eucaryotes et les Crénarchées et non pas entre les Eucaryotes et l'autre phylum principal, les Euryarchées (Cox *et al.*, 2008).

Rpo13 – Rpo13 a été découverte lors de la cristallisation de l'ARNP de *S. shibatae* (Korkhin *et al.*, 2009). C'est la seule sous-unité de l'ARNP spécifique aux archées. Cette sous-unité est détectée chez certaines espèces du phylum des Crénarchées, et aussi chez les Korarchées, mais pas chez les Euryarchées et les Thaumarchées. Elle est spécifique à certains groupes taxonomiques des Crénarchées (ordres des *Sulfolobales* et des *Desulfurococcales*) (Wojtas *et al.*, 2012). Elle occupe un espace entre Rpo5 et Rpo1. A cause de sa topologie (motif de liaison à l'ADN HTH), elle peut remplacer les cofacteurs de transcription additionnels eucaryotiques impliqués dans l'initiation de transcription et absents chez les archées comme ceci a été suggéré récemment (Korkhin *et al.*, 2009). Elle pourrait fonctionner comme l'archétype d'un facteur de transcription modulant les interactions de l'ADN pendant l'initiation et l'élongation de la transcription (Wojtas *et al.*, 2012). Il apparaît que Rpo13 n'a pas la conservation évolutive que les autres facteurs de transcription partagent (comme TBP, TFB, TFE et aSpt5) avec leurs correspondants homologues eucaryotes (TBP, TFIIB, TFIIE et eSpt5) (Werner *et al.*, 2011; Bell *et al.*, 2001; Werner 2012). Cette sous-unité est présente aux stades de l'initiation et de l'élongation de la transcription, donc constitue une partie intégrale de l'ARNP (Wojtas *et al.*, 2012). Des études ChiP ont montré que Rpo13 est associée de façon constitutive à l'ARNP de *S. shibatae in vivo*, lors de l'élongation de la transcription (Wojtas *et al.*, 2012). Cette observation a été renforcée par des expériences de gel retard qui ont démontré que Rpo13 se lie à

l'ADN double brin et non simple brin (Wojtas *et al.*, 2012). Ces résultats suggèrent que Rpo13 posséderait les caractéristiques structurales et fonctionnelles d'une sous-unité de type facteur de transcription.

Bien que les archées possèdent une seule ARNP, cette enzyme n'est pas identique dans sa composition à travers les phyla et les espèces, avec les sous-unités Rpo8 et Rpo13 différenciant les espèces des phyla des Crénarchées et des Korarchées.

En résumé, dans tous les domaines du vivant, le noyau enzymatique de l'ARNP conservé universellement, a assimilé des sous-unités additionnelles, à travers l'évolution, surtout chez les eucaryotes et les archées.

3. L'analyse fonctionnelle de l'ARN polymérase d'archées

La transcription assurée par l'ARNP cellulaire repose, dans tous les domaines de vie, sur trois étapes principales: (i) l'initiation précédée par le recrutement de l'ARNP sur le site du promoteur qui conduit au démarrage de la polymérisation; (ii) l'élongation; et (iii) la terminaison avec la libération de l'enzyme du promoteur. Ces trois étapes sont toutes stimulées par la présence de facteurs de transcription généraux, qui, entre autres fonctions, aident à positionner l'ARNP sur le promoteur, à dérouler l'ADN autour du site d'initiation de la transcription et à positionner l'ADN matrice simple brin près du site actif de l'ARNP.

a. Le complexe de pré-initiation chez les archées

La formation du complexe de pré-initiation (PIC) chez les *Archaea* dépend de deux facteurs généraux: TBP (« *TATA-box binding protein* ») et TFB (« *transcription factor B* »), qui sont respectivement les analogues des protéines TBP et TFIIB du système ARNP eucaryote (Grohmann *et al.*, 2009). Il existe un troisième facteur général, TFE (« *transcription factor E* »), considéré comme non essentiel. Bien que sa présence ne soit pas nécessaire pour l'initiation de la transcription *in vitro* (Qureshi *et al.*, 1997; Hirata *et al.*, 2008), TFE interagirait avec le brin non transcrit et stabiliserait donc le PIC (Bell *et al.*, 2001).

i. Les promoteurs chez les archées

Chez les archées, un promoteur peut contrôler un ou plusieurs gènes constituant un opéron, comme chez les bactéries. Les promoteurs d'archées sont très proches par leur organisation de ceux des eucaryotes. Cependant, chez les archées les régulateurs protéiques qui modulent l'expression des gènes ont la même origine et les mêmes mécanismes d'action que les régulateurs codés par les bactéries.

Les promoteurs d'archées contiennent une séquence conservée riche en nucléotides A-T et appelée la *TATA-box*, située entre 25 et 30pb en amont du site d'initiation de la transcription (Thomm and Wich, 1988). A côté de la *TATA-box*, un second élément conservé appelé l'élément *BRE* est présent, (Bell *et al.*, 1999). Ces deux sites sont également présents dans les promoteurs des eucaryotes (Figure 15). Chez les *Archaea*, ces sites sont reconnus par les facteurs généraux de transcription TBP et TFB (homologues des facteurs TBP et TFB chez les *Eukarya*) discutés dans les paragraphes suivants.

Récemment, Ao *et al.*, ont fourni la preuve de la présence d'un troisième élément dans le promoteur des archées : le *Sulfolobus initiator* (Inr) dont l'action dépend d'une *TATA-box* et d'une séquence BRE (Ao *et al.*, 2013). L'ajout de cette séquence à des promoteurs qui en sont dépourvus a pour l'effet une augmentation de leur force. Il a été proposé que TFB et les sous-unités de l'ARN polymérase se lient à l'Inr pour augmenter la force du promoteur, ce qui suggère que la présence ou l'absence d'un Inr dans les promoteurs basaux est importante pour la régulation génique globale chez *Sulfolobus* (Ao *et al.*, 2013).

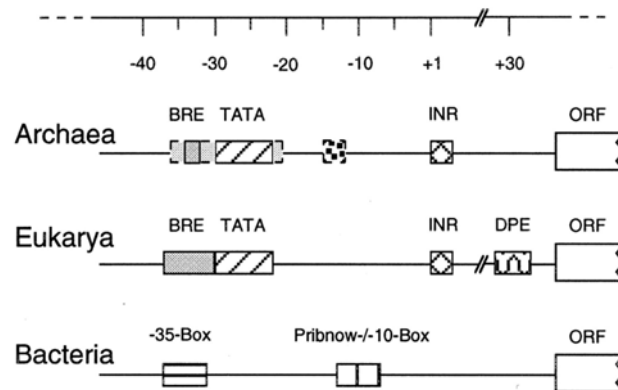


Figure 15 : Comparaison entre l'organisation des promoteurs archée, eucaryote et bactérien. INR = *Initiator element*; DPE = *downstream promoter element* (Soppa, 1999).

ii. La *TATA-box binding* protéine – TBP

La « *TATA-box binding protein* », ou TBP est un facteur général de transcription qui reconnaît la *TATA-box* en amont du site d'initiation de la transcription. La fixation de TBP est le premier pas d'une série d'événements qui conduira au recrutement de l'ARNP et à l'initiation de la transcription.

La protéine TBP comprend un domaine principal composé de deux parties symétriques, chacune codée par une séquence répétée (Ouhammouch *et al.*, 2009). Malgré cette structure inhabituelle, suggérant un événement de duplication de gène, ce gène prototype codant pour une seule de ces parties n'a jamais été trouvé, ni chez les *Archaea*, ni chez les *Eukarya* (Patikoglou *et al.*, 1999). La structure cristalline de la protéine d'archée TBP de *Pyrococcus woesei* (DeDecker, *et al.*, 1996) a permis une meilleure comparaison avec son homologue eucaryote. Elles ont une structure très similaire, bien que la similarité de leurs séquences ne soit comprise qu'entre 36 et 41% (Rowlands *et al.*, 1994).

La protéine TBP d'archée, comme celle des eucaryotes, est constituée d'une sorte de point d'ancrage à l'ADN autour duquel se formera par la suite le complexe d'initiation. En se liant à la séquence de la *TATA-box*, TBP provoque la courbure de l'ADN d'un angle d'environ 50-90°C (Figure 16). La distorsion de l'ADN change l'organisation spatiale autour de la *TATA-box* et va permettre aux facteurs et à l'ARNP d'être plus étroitement associés (Whittington *et al.*, 2008).

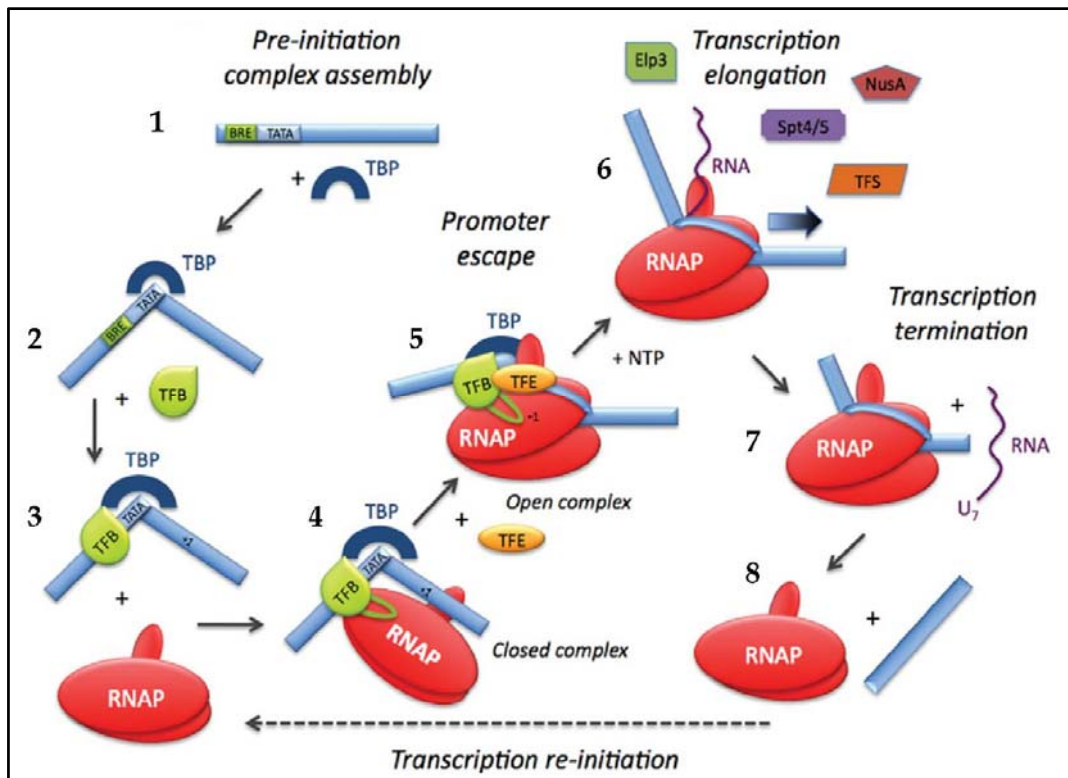


Figure 16 : Le cycle de la transcription chez les *Archaea*. La formation du complexe de pré-initiation dépend de la *TATA-binding* protéine (TBP) et du facteur de transcription B (TFB). (1) TBP reconnaît la *TATA-box* et (2) induit une courbure du promoteur. (3) TFB se lie au complexe *TATA-box*-TBP via son domaine C-terminal qui reconnaît l'élément BRE. (4) Le domaine N-terminal de TFB se lie à l'ARNP, en recrutant et orientant l'enzyme. Le doigt de Zn de TFB (boucle verte) inséré dans le site actif de l'ARNP stimule l'activité catalytique. (5) La transition entre le complexe « fermé » et « ouvert » est accompagnée par une ouverture de l'ADN et l'insertion du brin matrice dans le site actif. La réorganisation de l'ARNP et la topologie de l'ADN matrice sont stabilisées par le facteur de transcription E (TFE). (6) Les interactions entre l'ARNP et les facteurs cessent après la sortie du promoteur. Le brin ARN transcrit est lié à la tige de l'ARNP (la protubérance vue sur la figure). Pendant cette phase, l'ARNP peut être modulée par les facteurs d'élongation (TFS, Spt4/5 et peut-être NusA et Elp3). (7) La transcription de portions de 7U conduit à la terminaison et (8) à la libération du transcrit et de l'ARNP pour ré-initier la transcription (modifié depuis Grohmann, *et al.*, 2009).

iii. Le facteur de transcription B – TFB

Le « *transcription factor B*, ou TFB se lie au complexe TBP-*TATA-box* en reconnaissant spécifiquement la séquence de 6 bp d'un motif conservé RNWAAW (R = purine, W = A ou T), l'élément BRE (Figure 16), présent immédiatement en amont de la *TATA-box*. Cela fixe et stabilise le complexe de pré-initiation (Bell *et al.*, 1999).

La protéine TFB est faite de trois domaines (N-terminal, C-terminal et un doigt de zinc) unis par un connecteur flexible. L'organisation structurale de ce facteur est similaire à celle de la protéine eucaryote TFIIB. Le domaine principal C-terminal de TFB reconnaît l'élément BRE et se lie à TBP. L'ARNP est alors recrutée à ce complexe par les interactions entre le domaine N-terminal en doigt de Zn de TFB et la sous-unité Rpo1 (Grohmann *et al.*, 2009). Cela oriente l'enzyme, assurant que la transcription s'initie dans une direction correcte (Bell *et al.*, 1999).

b. L'initiation de la transcription

Après l'établissement du complexe de pré-initiation (PIC), l'ARNP est à un stade où le complexe est « fermé » (Figures 16 et 17) et n'a que peu de contacts directs avec le promoteur ADN qui est toujours sous forme double brin. La figure 17 illustre le passage de l'initiation de la transcription à l'élongation avec plus de détails.

Lors du passage d'un complexe « fermé » à « ouvert », les topologies de l'enzyme et du promoteur ADN vont subir de nombreux changements importants (Figures 17A) au cours desquels l'ADN db est ouvert et le brin matrice est chargé dans le domaine de liaison à l'ADN et le site actif. Le domaine « doigt de Zn » de TFB est inséré dans le site actif et stimule la catalyse, probablement en interagissant avec le brin matrice et/ou les NTPs (Kostrewa *et al.*, 2009). Ce mécanisme ne requiert pas d'hydrolyse de l'ATP pour conduire l'ouverture du promoteur (Bell *et al.*, 1998), contrairement à l'ARNPII eucaryote où de nombreuses étapes dépendantes de l'ATP existent dans l'initiation de la transcription (Yan and Gralla, 1997). Cela suggère que malgré l'origine commune et une similitude très forte entre les principales protéines constituant l'ARNP comme entre les facteurs (TBP, TFB et TFE), les détails de fonctionnement du complexe de transcription sont sensiblement différents chez les archées et chez les eucaryotes.

La pince se ferme ensuite sur le canal de liaison à l'ADN, avec le brin non transcrit passant en dehors de la pince (Figure 17B). Sans que cette interaction soit indispensable, la formation du complexe ouvert est stimulée par la tige (sous-unités Rpo4 et Rpo7) qui très vraisemblablement participe à la fermeture de la pince.

Un troisième facteur, TFE, présent chez tous les génomes d'archées séquencés, est un homologue de la sous-unité α du facteur général de la transcription eucaryote, TFIIE. TFE est le facteur général le moins compris chez les archées et les eucaryotes. Dans les essais *in vitro*, la présence de ce facteur n'est pas nécessaire ni pour l'initiation, ni pour la transcription (Qureshi *et al.*, 1997; Hirata *et al.*, 2008). Néanmoins, il a été montré qu'il stimulait la transcription de promoteurs faibles, lorsque les interactions entre TBP et la *TATA-box* ne sont pas optimales. En effet, TFE se lierait à TBP pour faciliter ou stabiliser la reconnaissance des éléments du promoteur (Bell *et al.*, 2001; Hanzelka *et al.*, 2001). En outre, la liaison de TFE (et TFIIE) à l'ARNP suggère qu'il promeut l'ouverture de l'ADN d'une manière allostérique en modulant la position de la pince de l'ARNP (Grohmann *et al.*, 2011).

Dans un complexe ouvert, l'enzyme, toujours associée à la séquence promotrice grâce à son interaction avec TFB, n'est toujours pas complètement processive et entre en « cycle abortif ». Elle effectue la synthèse de courts transcrits abortifs de 3-11 nucléotides qu'elle relâche sans suite et sans que l'ARNP ne se sépare du promoteur (Kapanidis *et al.*, 2006). Ainsi, le brin d'ADN matrice serait « froissé » de façon répétée par l'enzyme. L'importance du cycle abortif n'est pas encore comprise complètement mais, il semblerait que l'énergie contenue dans ces transcrits abortifs soit utilisée pour aider à rompre les interactions du complexe PIC et permettre finalement à l'enzyme de s'échapper du promoteur et d'initier la phase d'élongation (Revyakin *et al.*, 2006). Le plus probable est qu'en interagissant avec le brin d'ADN matrice, TFB augmente l'activité catalytique de l'ARNP (Werner and Weinzierl, 2005), permettant enfin à l'enzyme d'échapper au promoteur et d'initier la phase d'élongation (Figure 17C).

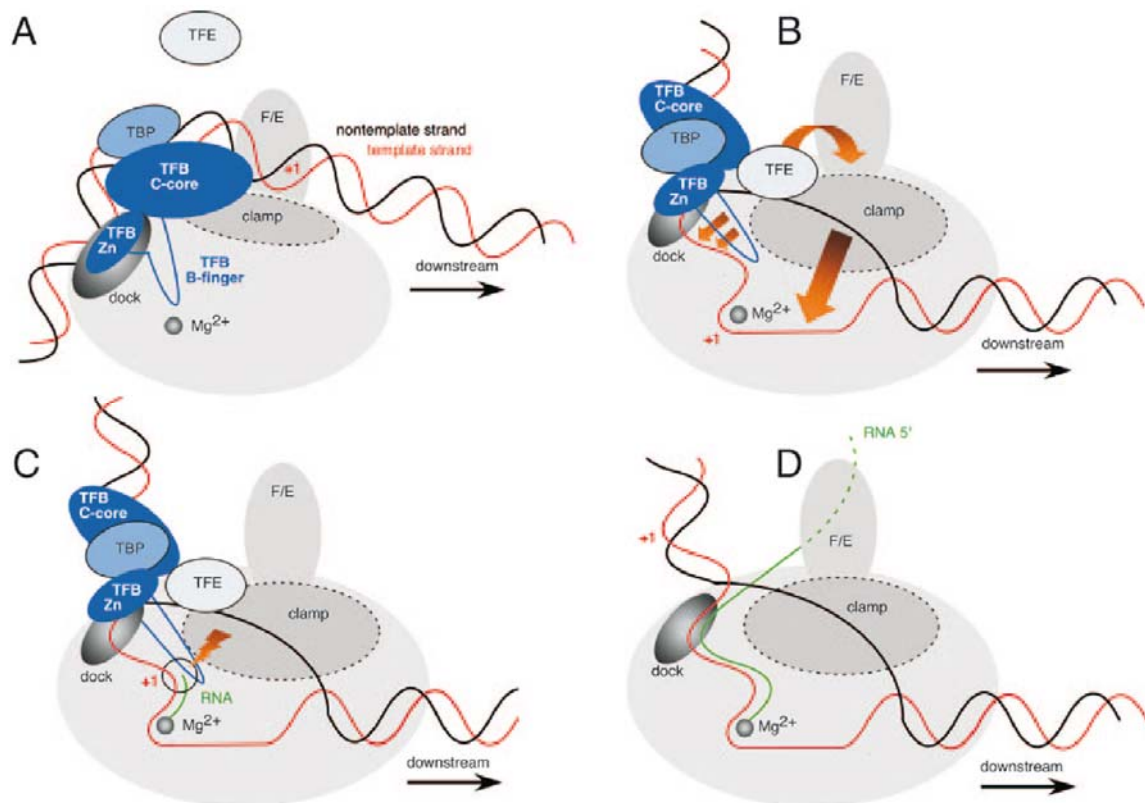


Figure 17 : L'initiation de la transcription chez les *Archaea*. (A) L'ARNP est recrutée au complexe TFB-TBP-ADN par les interactions entre TFB et l'ARNP. Dans ce complexe « fermé », l'ARNP fait seulement des contacts superficiels avec le promoteur ADN. Les sous-unités F et E de l'ARNP sont abrégées F/E. (B) Pendant la formation du complexe « ouvert », les brins ADN sont ouverts et le brin matrice est inséré dans le site actif de l'enzyme où il est stabilisé par le doigt de zinc de TFB. TFE stabilise aussi ce complexe en influençant la position de la pince de l'ARNP par les sous-unités F et E et peut-être en interagissant directement avec le brin non transcrit. (C) La phase d'initiation abortive est caractérisée par une production de courts transcrits ARN. Cette étape est stimulée par TFB probablement parce que le doigt de zinc de TFB facilite une configuration ADN/NTP ou même ADN/ARN/NTP (indiqué par le flash orange). (D) Pendant la fuite de l'ARNP du promoteur, les facteurs généraux de transcription TBP, TFB et TFE sont ôtés et le transcrit ARN est adressé vers les sous-unités F et E de l'ARNP dans le complexe d'élongation (adapté de Werner and Weinzierl, 2005).

c. L'élongation de la transcription

La phase d'élongation de la transcription est dépendante des éventuelles structures secondaires se formant à la fois par les séquences de l'ADN matrice et de l'ARNm produit. Egalement, des facteurs de transcription généraux ou spécifiques aux gènes transcrits interviennent dans la régulation fine de la transcription (Vassylyev, 2009; Selth *et al.*, 2010).

Dans le complexe d'élongation de la transcription (TEC), l'enzyme interagit avec l'ADN matrice db, avec le duplex ADN-ARN et avec le nouveau transcrit ARN (Brueckner *et al.*, 2009). D'après l'analyse des structures observées aux rayons X, la stabilité et l'intégrité du TEC dépend des interactions entre la tige de l'ARNP et l'ARN, (Andrecka *et al.*, 2009; Brueckner *et al.*, 2009). La tige a été démontrée comme la structure liant l'ARN *in vitro* et *in vivo* (Meka *et al.*, 2003), mais d'une manière non spécifique aux séquences et avec une affinité faible (Grohmann and Werner, 2010). Ceci suggère que

la tige pourrait être impliquée plus fortement dans l'élongation que dans l'initiation de la transcription. Malheureusement, comme ces interactions sont dynamiques, il est difficile de répondre par l'approche de cristallographie aux différentes questions concernant le déroulement exact des différentes étapes de transcription.

La vitesse d'élongation n'est pas constante, cette étape est marquée par des pauses et même des arrêts de l'ARNP. Quand cela se produit, l'enzyme revient souvent en arrière, ce qui provoque la dissociation du site actif de l'extrémité 3' de l'ARN naissant. Le remplacement de l'extrémité 3' dans la cavité catalytique est vital pour la reprise de l'activité de l'enzyme. Ceci se réalise par la création de nouveaux points en 3' grâce à la coupure de la molécule de l'ARN déjà synthétisée à l'intérieur du complexe ARNP. Il en résulte par la formation d'une nouvelle extrémité 3' de l'ARN placée dans le site actif qui est ensuite utilisée pour redémarrer la polymérisation (Kettenberger *et al.*, 2003).

Chez les *Archaea*, un facteur général de clivage de la transcription (TFS) stimule la coupure du transcrit. TFS est capable de repositionner l'ion catalytique Mg^{2+} porté par l'ARNP en insérant son domaine de doigt de zinc dans le site actif via le pore d'entrée (Hausner *et al.*, 2000; Lange *et al.*, 2004). TFS a une forte similarité de séquence avec la protéine eucaryote TFIIIS et la sous-unité RPB9. Chez les *Bacteria*, la stimulation du clivage de la transcription est faite par GreA et GreB (Borukhov *et al.*, 2005). Les facteurs bactériens de clivage ne sont pas reliés à TFS ou TFIIIS en séquence et en structure, mais les mécanismes d'action sont très similaires. Tous ces facteurs entrent dans le noyau catalytique de l'ARNP, complètent le site actif en modifiant l'association avec l'ADN, l'ARN, les NTPs ou les ions Mg^{2+} , et stimulent l'activité catalytiques « dormante » de l'ARNP. L'existence de mécanismes aussi similaires entre des protéines non liées d'un point de vue évolutif, est un exemple intéressant d'adaptation convergente vers une même fonction.

D'autres facteurs codés par les génomes des archées susceptibles d'être impliqués dans la régulation de l'élongation ont été identifiés par comparaison *in silico*. Ils ressemblent à certaines protéines des bactéries et des eucaryotes connues pour leur rôle dans la transcription. Parmi ces protéines se trouve l'homologue du facteur d'élongation Elp3 des eucaryotes et la protéine Spt5, homologue de la protéine bactérienne NusG universellement conservée, présente chez les eucaryotes comme le complexe hétérodimère Spt4/5. Cette dernière pourrait également être impliquée dans l'évasion de l'ARNP du promoteur. Le site de liaison de TFS chevauche le facteur d'initiation TFB dans la pince, il est donc possible que le facteur d'élongation pourrait déloger les facteurs d'initiation et libérer l'enzyme pour l'élongation de la transcription (Sevostyanova *et al.*, 2008).

d. La terminaison de la transcription

La terminaison de la transcription chez les archées est moins bien caractérisée comparé aux deux autres étapes de la transcription, l'initiation et l'élongation. Chez les procaryotes, seules les bactéries bénéficient ces dernières années, d'avancées remarquables dans la caractérisation de la terminaison (Epshtein *et al.*, 2007; Ciampi, 2006).

L'ARNP d'archée semble terminer un cycle de transcription via un mécanisme similaire à celui décrit pour l'ARNP II eucaryote (Palangat *et al.*, 2004). Dans ce cas, de courtes séquences poly-dA présentes dans le brin matrice en aval et le codon stop servent de signal de terminaison reconnu directement par l'ARNP *in vivo* et *in vitro* (Santangelo and Reeve, 2006; Santangelo *et al.*, 2009). La terminaison se réalise sans besoin de formation des structures secondaires en épingles à cheveux dans le transcrit

ou de facteurs de terminaisons supplémentaires. L'ARNP ainsi en pause peut subir des changements conformationnels, comme l'ouverture de la pince, facilitant la dissociation de l'ARNP de l'ADN matrice et provoquant probablement la terminaison (Martin *et al.*, 1980; Hirtreiter *et al.*, 2010). Alternativement, des données récentes suggèrent que les ARNP archées sont « recyclées » au promoteur et initient à nouveau la transcription comme l'ARNP II eucaryote (Spitalny and Thomm, 2008). La mutation des séquences en aval des poly-dA provoque une diminution de l'efficacité de transcription. L'équilibre subtil entre la translocation de l'ARNP au-delà des séquences poly-dA et l'arrêt au niveau de la séquence terminaison est important pour la terminaison et le recyclage du système (Spitalny and Thomm, 2008) respectivement.

Cependant, les données biologiques pointent sur l'insuffisance des hypothèses présentées ci-dessus. En effet, beaucoup de gènes chez les archées ne possèdent pas de terminateurs sur leurs extrémités 3'. Dans d'autres cas, des séquences terminatrices sont prédites mais elles se trouvent dans la partie codante de ces gènes. Malgré l'absence de données directes, il semble que les archées pourraient coder pour des protéines qui agissent de la même façon que les terminateurs bactérien et eucaryote, Rho (Campi *et al.*, 2006) et Xrn2 (Nagarajan *et al.*, 2013).

4. La régulation de la transcription chez les archées

Parmi plusieurs types de régulations d'expression des gènes existant dans une cellule vivante (transcriptionnelle, post-transcriptionnelle, traductionnelle), la régulation de l'initiation de la transcription est une des plus courantes. Elle est réalisée par l'action de facteurs de transcription spécifiques.

Dans ce chapitre, plusieurs stratégies de régulation de l'initiation de la transcription découvertes chez les archées sont présentées en mettant l'accent sur l'utilisation de protéines de liaison à l'ADN de type bactérien.

1. Caractérisation générale des mécanismes de régulation de l'expression génique chez les archées

a. Régulation par des facteurs généraux de la transcription

La majorité des génomes séquencés chez les *Archaea* codent pour de multiples orthologues des facteurs généraux de transcription, qui pourraient être responsables de la modification du niveau d'expression de certains gènes. Par exemple, *Halobacterium salinarum* possède six copies du gène *tbp*, *tbpA-F*, dont seul le gène *tbpE* est porté par le chromosome. Les cinq autres gènes sont portés par les plasmides pRNC100 et PRNC200 (200kb) (Teufel *et al.*, 2008). Le nombre de copies alternatives de gènes *tbp* varie selon la souche et dépend du nombre de copies de plasmides porteurs. L'analyse comparative de l'expression de ces gènes au cours de la croissance cellulaire révèle que le gène *tbpE* est le mieux exprimé faisant penser que le facteur TbpE est le facteur principal. En ce qui concerne les autres copies de *tbp*, il a été suggéré qu'elles pourraient jouer un rôle essentiel dans la croissance en phase stationnaire. Une étude récente a révélé qu'en conditions anaérobies, les souches portant un plasmide délété pour ces gènes alternatifs poussent moins vite que les souches sauvages (Teufel *et al.*, 2008). En outre, *H. salinarum* code pour sept copies du gène *tfb* (Basset *et al.*, 2000). Il semble que l'utilisation de différentes combinaisons de TBP et TFB permette une régulation très fine du niveau de transcription d'une majorité de gènes selon les conditions environnementales dans lesquelles se trouve la cellule d'archée (Facciotti *et al.*, 2007).

La plupart des *Archaea* ont deux paralogues TFB, TFB1 et TFB2 (le rôle de TFB1 dans la transcription a déjà été présenté dans le Chapitre 1.3.3). Notamment chez *Thermococcus kodakarensis*, les deux copies de TFB1 semblent être complètement redondantes aussi bien *in vivo* qu'*in vitro* (Santangelo *et al.*, 2007). *Pyrococcus furiosus* encode également pour deux paralogues TFB1, qui sont transcrits *in vivo*. Cependant, il est à noter que la structure de TFB2 de *P. furiosus* semble être altérée car la protéine ne possède pas de motif B-finger reconnaissable (Micorescu *et al.*, 2008). L'importance physiologique des TFB1 multiples chez les *Archaea* n'est pas entièrement comprise. Cependant, certaines données indiquent que le rôle physiologique de deux orthologues de TFB1 n'est pas identique. Ainsi, il a été démontré que chez *P. furiosus*, les niveaux de transcription de *tfb2* augmentaient après un choc thermique tandis que ceux de *tfb1* restaient stables. Ceci suggère l'implication directe (Micorescu *et al.*, 2008) ou indirecte (Shockley *et al.*, 2003) de TFB2 dans la réponse au stress thermique.

b. Régulation par des facteurs liant des séquences ADN spécifiques

Chez les *Archaea*, la grande majorité des protéines susceptibles d'intervenir dans la régulation de la transcription, en se liant à l'ADN et en affectant l'expression génique, sont prédites *in silico* sur la base de leur ressemblance aux différents régulateurs bactériens (Aravind and Koonin, 1999). En effet, l'analyse *in silico* des génomes archées disponibles n'a mis en évidence qu'un très faible nombre de protéines homologues aux protéines eucaryotiques capables de lier l'ADN. Ceci a été une observation surprenante en sachant que les archées transcrivent leurs gènes à l'aide d'une machinerie transcriptionnelle très proche de celle des eucaryotes.

La plupart des régulateurs transcriptionnels d'archées identifiés jusqu'ici possèdent des domaines de liaison à l'ADN typiquement bactériens, comme les motifs RHH (*ribbon-helix-helix*) et HTH (*helix-turn-helix*) (Aravind and Koonin, 1999; Perez-Rueda, Collado-Vides *et al.*, 2004). Il convient de mentionner que certains régulateurs présumés ont un motif en doigt de zinc de type C₂H₂ (Bettstetter *et al.*, 2003), mais ils sont codés essentiellement par des génomes de virus infectant des archées et ces protéines sont très peu étudiées. Une de ces protéines, AFV1p06 du virus AFV1 des *Lipothrixviridae* a été caractérisée en détails au cours de mon travail doctoral (Guillière *et al.*, 2013). Un autre domaine de liaison, le domaine *swapped-hairpin* est aussi bien représenté dans les génomes des *Sulfolobus* (Prangishvili *et al.*, 2006).

i. Les répresseurs transcriptionnels chez les archées

Deux mécanismes principaux de répression ont été décrits chez les *Archaea* et leurs virus. Le premier mécanisme implique la liaison du répresseur sur un site chevauchant l'élément BRE et la séquence *TATA-box* bloquant ainsi par encombrement stérique l'accès au promoteur pour les facteurs TBP et TFB (Lee *et al.*, 2005; Lipscomb *et al.*, 2009). Dans le second mécanisme, le régulateur transcriptionnel se lie à un site en aval de la *TATA-box* (Lie *et al.*, 2010; Karr 2010; Lee *et al.*, 2008.; Kanai *et al.*, 2007). Dans ce cas, la liaison des facteurs TBP et TFB n'est pas empêchée, mais la présence du régulateur inhibe le recrutement de l'ARNP (Peeters *et al.*, 2013). Plusieurs exemples illustrant ces deux mécanismes de répression sont présentés plus loin.

ii. Les activateurs transcriptionnels chez les archées

Les activateurs transcriptionnels chez les *Archaea* sont moins caractérisés que les répresseurs. Le mécanisme d'activation le plus connu consiste à faciliter l'accès des facteurs généraux de la transcription à l'ADN, soit par des changements conformationnels de l'ADN, soit par des interactions directes avec les facteurs généraux de la transcription.

Habituellement, les activateurs transcriptionnels se lient à un site opérateur primaire, localisé en amont ou chevauchant partiellement l'élément BRE, qui est aussi appelé « *upstream activating sequence* » (UAS). Des sites opérateurs auxiliaires sont localisés en amont du premier site, en aval de la *TATA-box* ou en aval du site d'initiation de la transcription. Il a été montré qu'ils amélioreraient ainsi directement le recrutement *in vitro* de TBP à la *TATA-box* (Ouhammouch *et al.*, 2003; 2005; Peng *et al.*, 2009), même si les mécanismes d'activation *in vivo* de ces régulateurs restent à prouver. Plusieurs exemples d'une telle régulation sont discutés plus en détails dans les sections suivantes (Figure 18).

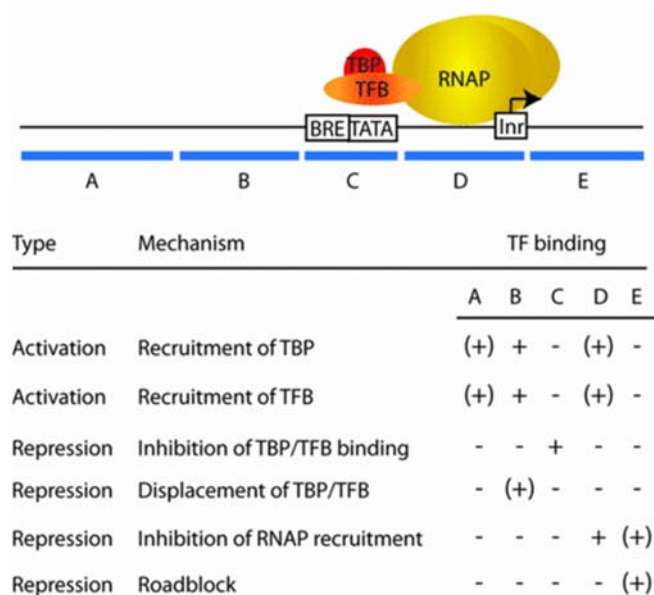


Figure 18 : Vue d'ensemble des mécanismes moléculaires de la régulation de la transcription et des localisations de liaison observées. (A) très en aval du promoteur; (B) en aval mais proche du promoteur; (C) chevauchement du promoteur principal; (D) en amont du promoteur; (E) en amont du site d'initiation de la transcription. Le symbole + indique que le mécanisme a été démontré, le symbole (+) que le mécanisme a été postulé sans preuves concluantes (adapté de Peeters *et al.*, 2013).

c. Régulation post-transcriptionnelle et dégradation des ARNm

Il existe très peu d'informations sur la régulation post-transcriptionnelle chez les *Archaea*. Les principales données décrivant les modifications de l'ARNt, viennent des archées halophiles (Phillips and de Crecy-Lagard, 2011).

La dégradation des ARNm est un processus biologique essentiel très important dans tous les domaines de la vie. Il a été largement étudié sur les modèles bactériens (Kaberdin *et al.*, 2011; Bechhofer 2011; Lalaouna *et al.*, 2013) et viraux, surtout pour le bactériophage T4 (Uzan and Miller, 2010). Ce processus est également décrit en détails chez les eucaryotes (Moore, 2005). La dégradation chez les archées est moins bien comprise mais plusieurs études indiquent son importance dans la régulation. Par exemple, chez les *Sulfolobales*, le facteur d'initiation de la traduction α eIF2, homologue du facteur eucaryote eIF2, possède un rôle supplémentaire. L'une de ses sous-unités peut lier l'extrémité 5' des ARNm et contrecarrer la dégradation de l'ARNm (Hasenohrl *et al.*, 2008). Ce mécanisme ressemble fortement à celui décrit pour le modèle eucaryote (Newbury *et al.*, 2006).

d. La régulation par les histones chez les archées

Dans les cellules eucaryotes, l'ADN chromosomique est structuré par les histones dont l'ensemble forme des complexes nucléoprotéiques appelés les nucléosomes. La présence des histones a été considérée comme une caractéristique marquante et unique des *Eukarya* jusqu'à ce que les histones d'*Archaea* soient découvertes (Sandman *et al.*, 1990).

Chez les *Archaea*, presque toutes les *Euryarchaeota* connues codent pour des homologues des histones et quelques *Crenarchaeota* codent pour des protéines capables de structurer l'ADN de façon semblable aux histones.

Les histones d'archées existent en tant que dimères et par leur structure ressemblent aux histones eucaryotiques H3 et H4 (Reeve *et al.*, 2004). Chez *P. furiosus* (Soares *et al.*, 1998) et *Methanothermobacter thermoautotrophicus* (Xie and Reeve, 2004), il a été démontré que la structuration de l'ADN chromosomique en tétrasomes formés par les dimères d'histones réprime la transcription *in vitro*. Les auteurs expliquent cet effet par l'interférence entre les tétrasomes et le processus de l'assemblage du PIC. Toutefois, les histones des euryarchées n'ont pas les queues N terminales (Cheung *et al.*, 2000) constituant normalement les cibles des modifications chez les eucaryotes. Aucun homologue des enzymes eucaryotes de modification des histones n'a été trouvé *in silico* dans aucun des phyla d'archées.

Chez les crénarchées, une autre protéine de type histone a été identifiée. Cette protéine, Alba (**A**cetylation **L**owers **B**inding **A**ffinity), est capable de former des structures proches de la chromatine (Xue *et al.*, 2000; Bell *et al.*, 2002; Wardleworth *et al.*, 2002). Alba peut être acétylée et désacétylée par des enzymes de modification. En outre, il a été démontré que la forme désacétylée d'Alba réprime la transcription *in vitro* (Bell *et al.*, 2002). Néanmoins, le mécanisme moléculaire de l'action d'Alba demeure hypothétique, même s'il est de plus en plus évident (Soppa *et al.*, 2010) que son état d'acétylation influe sur le degré de compactage de la chromatine chez les *Sulfolobales*.

2. Deux groupes majoritaires de régulateurs transcriptionnels chez les archées – les familles HTH et RHH

Une inspection *in silico* des régulateurs transcriptionnels annotés dans les génomes d'archées et leurs virus montrent que quatre motifs principaux sont employés pour lier l'ADN : *Helix-Turn-Helix* (HTH), *Ribbon-Helix-Helix* (RHH), *swapped-hairpin* et doigt de zinc C2H2.

Dans cette section, une caractérisation détaillée des régulateurs HTH et RHH actuellement décrits chez les *Archaea* est présentée.

a. Les domaines de liaison HTH et wHTH

Le domaine HTH est la structure de liaison à l'ADN la plus abondante décrite jusqu'ici chez les procaryotes (Sauer *et al.*, 1982; Aravind *et al.*, 2005). Ce domaine est composé par deux hélices α . La plupart des facteurs de transcription portant ce motif, placent une hélice α dans le grand sillon de l'ADN pour lier des séquences ADN spécifiques. Le domaine *winged* HTH (wHTH) se distingue par la présence d'une unité en épingle à cheveux d'un feuillet β (l'aile) en C-terminal (Aravind *et al.*, 2005). La superfamille des protéines de liaison à l'ADN wHTH (*winged* HTH), qui est une variante du motif HTH, inclut la majorité des régulateurs transcriptionnels des *Sulfolobus*. Dans ces protéines, l'aile offre une interface additionnelle pour le contact avec l'ADN, en interagissant typiquement avec le petit sillon de l'ADN, à travers des résidus chargés (Gajiwala and Burley, 2000).

b. Les régulateurs HTH de la famille Lrp chez les archées

D'après les analyses *in silico*, il est prédit que la famille Leucine-Responsive Regulatory Protein (Lrp), une famille bien connue de régulateurs chez les *Bacteria*, est également très présente chez les *Archaea* (Brinkman *et al.*, 2003). Les protéines de type Lrp ont été annotées dans tous les génomes d'archées disponibles et sont considérées comme l'une des plus abondantes familles de protéines régulatrices de la transcription dans ce domaine (Wu *et al.*, 2008; Charoensawan, Wilson *et al.*, 2010; Pérez-Rueda and Janga, 2010).

Chez les bactéries, les protéines Lrp sont d'importants régulateurs impliqués dans le contrôle du métabolisme des acides aminés. La famille Lrp, nommée ainsi d'après la protéine prototype de cette famille (le régulateur général Lrp décrite initialement chez *Escherichia coli*), est aussi connue sous le nom de *Feast/Famine Regulatory Proteins* (FFRP) (Calvo and Matthews, 1994), à cause de son rôle joué dans l'adaptation métabolique lors de l'appauvrissement du milieu de croissance.

Ci-dessous, des exemples et des détails fonctionnels des régulateurs négatifs et positifs trouvés chez les *Crenarchaeota* et les *Euryarchaeota* sont donnés ci-dessous avant d'énoncer les caractéristiques plus générales de la famille des protéines Lrp.

i. Mécanismes de contrôle et exemples de régulateurs Lrp archées

La régulation de l'expression génique chez les *Archaea* est souvent obtenue par différentes interactions avec le complexe de pré-initiation de la transcription.

Généralement, les régulateurs archées de type Lrp se lient à proximité de la région promotrice, permettant la formation d'un complexe ternaire TBP-TFB-Lrp. Ces régulateurs ont souvent une double fonction : ils peuvent affecter la formation du PIC de façon positive et négative. Un autre mécanisme de régulation implique le changement de topologie de l'ADN à la suite de la formation d'un filament nucléoprotéique grâce à l'oligomérisation de la protéine Lrp sur l'ADN cible. Ce changement affecte l'efficacité de la transcription (Peeters and Charlier, 2010).

Il apparaît que les stratégies de régulation qui peuvent être employées sont très variables et encore mal comprises.

1. Les répresseurs Lrp chez les archées

FL11 de *Pyrococcus horikoshii* et **Lrp** et **LrpA1** de *Halobacterium salinarum* sont probablement les régulateurs Lrp les mieux étudiés chez les *Euryarchaeota* (Yokoyama *et al.*, 2007; Kawashima *et al.*, 2008; Schwaiger *et al.*, 2010).

Chez *P. horikoshii*, FL11 est prédit pour réprimer l'expression de plus de deux-cent gènes. La structure cristalline du régulateur sous sa forme de dimère, dans un complexe avec de l'ADN db a été résolue (Yokomaya *et al.*, 2007) (Figure 19b) et son analyse démontre que FL11 induit une flexion de 55° au milieu du site de liaison. Lorsque la lysine est présente dans le milieu en concentrations élevées, quatre dimères de FL11 se lient ensemble en octamère (Figure 19d) et couvrent le promoteur du gène *fl11*, autorégulant sa transcription. La transcription du gène de la biosynthèse de la lysine est également réprimée par le même mécanisme. La réduction de la concentration de FL11 diminue la répression de plusieurs gènes sous le contrôle de FL11 activant le mode « festin ». Il s'agit

des gènes codant pour le catabolisme des acides aminés et de la synthèse de l'ATP. En mode « famine », peu de lysine est disponible et les dimères de FL11 ne s'assemblent pas en octamères. La répression du promoteur *fl11* est levée et à la suite de son expression, une grande quantité de dimères de FL11 deviennent disponibles pour réprimer la transcription de ses gènes cibles. Dans ces conditions, la croissance cellulaire s'arrête rapidement. Parmi les gènes régulés par FL11 se trouvent quatre autres gènes codant pour des protéines de type Lrp. Le réseau hiérarchique établi par le régulateur général FL11 et les régulateurs sous son contrôle régule 20 % de gènes de *P. horikoshii* (Yokoyama *et al.*, 2009).

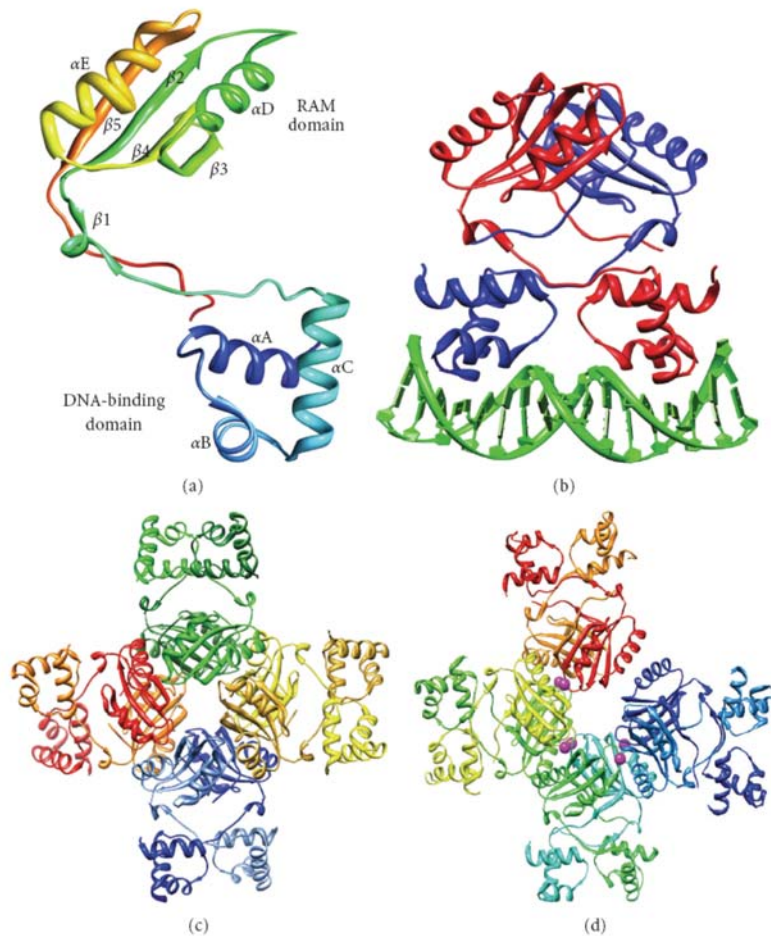


Figure 19 : Structures caractéristiques des protéines de type Lrp chez les *Archaea*. (a) Structure monomérique de Grp de *S. tokodaii*. Les hélices α sont désignées : αA - αE et les feuillets β : $\beta 1$ - $\beta 5$. Le domaine de liaison à l'ADN en N-terminal correspond à la partie basse de la structure, et le domaine de liaison au ligand en C-terminal correspond à la partie haute de la structure. Ils sont appelés, respectivement, *DNA-binding-domain* et *RAM domain*. (b) Structure d'un dimère de FL11 de *Pyrococcus horikoshii* lié à l'ADN. (c) Structure d'un octamère de LrpA de *P. furiosus*. (d) Structure d'un octamère de FL11 de *P. horikoshii*, lié par de l'arginine, dans sa conformation ouverte. Les molécules arginines sont les points violets (Yokoyama *et al.*, 2007; Kumarevel *et al.*, 2008).

Chez *H. salinarum*, le rôle des régulateurs de Lrp et LrpA1 a été évalué en comparant l'expression d'une souche sauvage à des mutants de perte (KO) ou de gain (surexpression) de fonction, à l'aide d'une approche de puces à ADN (Kawashima *et al.*, 2008; Schwaiger *et al.*, 2010). Lrp s'est avéré agir comme un régulateur transcriptionnel général, contrôlant des gènes impliqués dans le métabolisme central, dans la synthèse des acides aminés, dans les processus de transport et de transduction du signal. LrpA1, en revanche, possède un rôle plus spécifique, en contrôlant le gène de l'aspartate transaminase, en présence de L-aspartate.

Comme déjà discuté dans la section 1.4.1.a, *H. salinarum* code pour sept gènes *tfb*. De façon intéressante, Lrp et LrpA1 contrôlent la transcription de différents gènes *tfb*, reliant la régulation de facteurs alternatifs de transcription à une régulation plus spécifique par les Lrp.

Pour les protéines **Lrs14** de la crénarchée *S. solfataricus* et **LrpA** de l'euryarchée *P. furiosus*, il a été démontré qu'elles répriment la transcription des promoteurs de leurs propres gènes. La répression semble indépendante de la présence d'une molécule effectrice. LrpA (Figure 19c) se lie à la région promotrice de son propre gène reconnaissant un site de liaison situé juste en aval de la *TATA-box* (Brinkman *et al.*, 2008). Les facteurs généraux de transcription peuvent continuer à se lier au promoteur, mais la présence du régulateur accroché à son site bloque le recrutement de l'ARNP (Dahkle *et al.*, 2002). Les auteurs ont suggéré que ce type de régulation est bénéfique pour une réponse rapide face aux changements environnementaux, car dès la libération du régulateur négatif, le PIC n'a besoin que de directement recruter l'ARNP pour commencer la transcription. Le site de liaison de Lrs14 sur le promoteur de son gène est localisé de façon à chevaucher avec l'élément BRE et la *TATA-box*, empêchant ainsi les facteurs TBP et TFB de se lier au promoteur et bloquant directement l'initiation de la transcription (Bell and Jackson, 2000; Schwaiger *et al.*, 2010).

Des mécanismes d'action similaires ont été proposés pour d'autres régulateurs, comme **Sa-Lrp** de la crénarchée *S. acidocaldarius* (Enoru-Eta *et al.*, 2000). Sa-Lrp régule négativement la transcription de son propre gène couvrant son promoteur entier. Bien que jamais prouvé expérimentalement, cela semble bloquer la formation du PIC (Vassart *et al.*, 2013).

Il a été possible de cristalliser la protéine **Grp** (Figure 19a) (protéine du récepteur de la glutamine) de *S. tokodaii* sous forme native et en complexe avec son effecteur, la L-glutamine. Un changement conformationnel important se produit dans le domaine de liaison à l'effecteur lors de la liaison de la L-glutamine qui pourrait être essentielle pour la stabilisation de la structure octamérique. Cela pourrait signifier qu'en fonction de la présence/absence de L-glutamine il existe deux conformations distinctes de la forme octamérique de Grp. Selon sa conformation, la protéine agira sur les différents sous-ensembles de gènes cibles (Kumarevel *et al.*, 2008).

2. Les activateurs Lrp chez les archées

Ptr2 a été identifié dans le génome de l'euryarchée *Methanococcus jannaschii* comme un régulateur transcriptionnel putatif, homologue de Lrp. Son site consensus de liaison à l'ADN a été tout d'abord déterminé par l'approche de SELEX (*Systematic Evolution of Ligands by Exponential Enrichment*) et une recherche dans la séquence du génome de *M. jannaschii* des sites de liaison putatifs a révélé plusieurs cibles potentielles (Ouhammouch and Geiduschek, 2001). Ptr2 active la transcription de gènes codant pour la rubredoxin 2 (*rb2*), la ferrédoxine A (*fdxA*) et la rubrerythrine (*rbr*) (Ouhammouch *et al.*, 2003; 2005).

L'effet d'activation a d'abord été observé pour le promoteur *rbr* (Ouhammouch *et al.*, 2003), sur lequel Ptr2 a trois sites de liaison (Figure 20). Les sites 1 et 2 sont localisés juste en amont de la *TATA-box* et sont séparés par trois tours d'hélices de l'ADN.

L'effet d'activation après la liaison de Ptr2 sur le site 1, le site avec l'affinité la plus élevée, provient d'un contact direct et spécifique entre la protéine régulatrice et le facteur général de transcription TBP, améliorant la formation du PIC. Le site 2 contribue à cet effet activateur en promouvant l'occupation du site 1, par le biais de changements de conformation de l'ADN. Ces sites constituent le site d'activation en amont du promoteur (UAS pour « *Upstream Activation Site* »). Un troisième site de liaison, le site X, est situé dans la partie codante du gène *rb2* en aval de la *TATA-box*. Le rôle du site X dans la transcription n'est pas clair, mais il a été suggéré qu'il devrait, très probablement, avoir un rôle fonctionnel *in vivo* (Ouhammouch *et al.*, 2005).

Le module d'activation de la protéine a été identifié en construisant des mutants de Ptr2. Deux classes de mutants présentant un défaut d'activation ont été obtenues. Les mutants du premier groupe ont été affectés dans la capacité de Ptr2 à interagir avec TBP. Les mutants du second groupe ont préservé leur capacité à recruter TBP et TFB, mais ils échouent à assembler des complexes d'initiation fonctionnels. Le dernier phénotype suggère la participation de Ptr2 dans des étapes postérieures à l'initiation de la transcription (Pritchett, *et al.*, 2009).

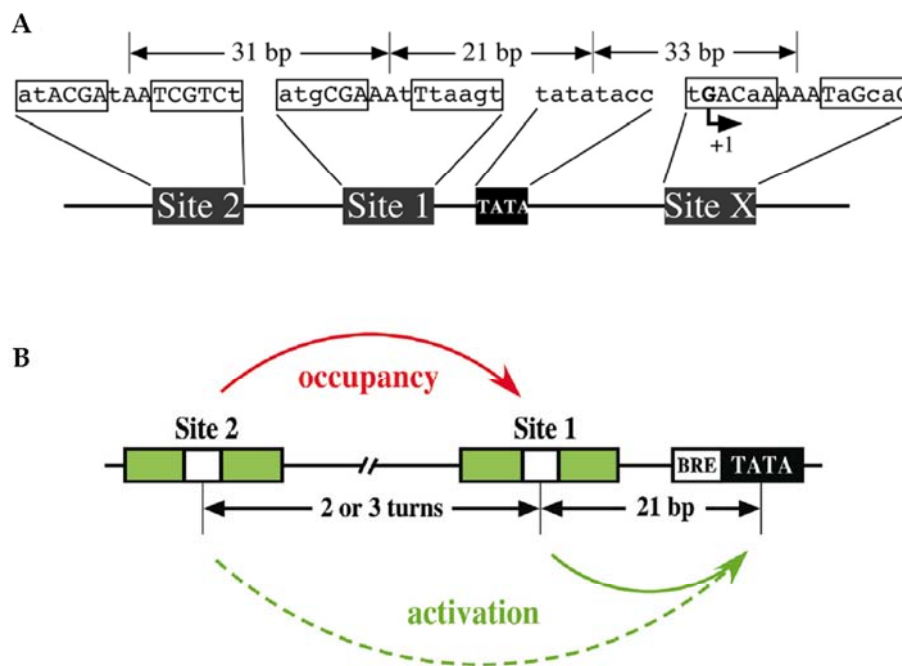


Figure 20 : Mécanisme de régulation transcriptionnel par Ptr2 de *Methanococcus jannaschii*. (A) La région promotrice du gène *rb2* et ses éléments régulateurs. Les sites de liaison 1 et 2 de Ptr2 sont à -48pb et -79pb, respectivement, par rapport au site d'initiation de la transcription (+1) et le site X est à +7pb (les distances centre à centre sont indiquées au-dessus). Les motifs hexamères de liaison sont encadrés et les nucléotides correspondant au site consensus sont en majuscules. (B) Modèle d'architecture et du mécanisme des UAS (*upstream activation sites*) (modifié de Ouhammouch *et al.*, 2005).

Ss-LrpB de la crénarchée *S. solfataricus* est probablement le membre de la famille Lrp le mieux étudié. Cette protéine a été d'abord caractérisée comme un régulateur hypothétique de son propre gène étant donné qu'il était capable de se lier à la région de son promoteur dans la partie chevauchant la *TATA-box* et l'élément BRE (Peeters *et al.*, 2004). Un mécanisme d'autorégulation qui implique le blocage de la formation du PIC a été proposé (Peeters *et al.*, 2004). L'analyse de la région du promoteur révèle que la protéine reconnaît trois sites spécifiques de liaison imparfaits de quinze nucléotides séparés par seize ou dix-sept nucléotides (Peeters *et al.*, 2004).

Comme illustré à la Figure 21, le site le plus éloigné (Site 3) est localisé à cent nucléotides en amont du site d'initiation de la transcription et le site le plus proche (Site 1) est à quarante nucléotides du +1. Les trois sites sont situés sur la même face de l'hélice d'ADN. De façon intéressante, le site intermédiaire (Site 2) a l'affinité pour Ss-Lrp la plus faible. Cette configuration conduit à la formation de trois complexes distincts de manière dépendante de la concentration (Peeters *et al.*, 2004).

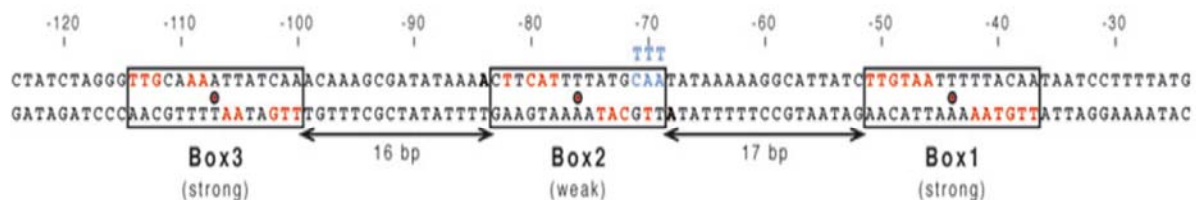


Figure 21 : Séquence de la région promotrice de gène *ss-lrpB* avec les trois sites de reconnaissance et leur distance relative du site d'initiation de la transcription. Les lettres rouges représentent les séquences palindromiques et les points rouges les centres de symétrie 2D (modifié depuis Peeters, *et al.*, 2004).

En analysant la capacité de former le PIC en présence de ce régulateur, il a été démontré que Ss-LrpB n'inhibait pas la liaison des facteurs généraux de la transcription. Ces observations ont induit les auteurs à suggérer que la liaison initiale au Site 1 activerait la transcription à distance de la même façon que décrite pour le régulateur Ptr2. En outre, comme la liaison de Ss-LrpB induit une forte déformation de l'ADN sur chaque site, chaque complexe induira un degré différent de courbure d'ADN (Figure 22).

L'occupation de l'ensemble des trois sites provoque une très forte déformation de l'ADN dans cette région promotrice, conduisant à une répression de la transcription (Peeters *et al.*, 2004). En effet, il a été démontré ultérieurement que lorsque les trois sites sont occupés, il y a formation de complexes nucléoprotéiques globulaires couvrant une région d'environ 100pb (Peeters *et al.*, 2006).

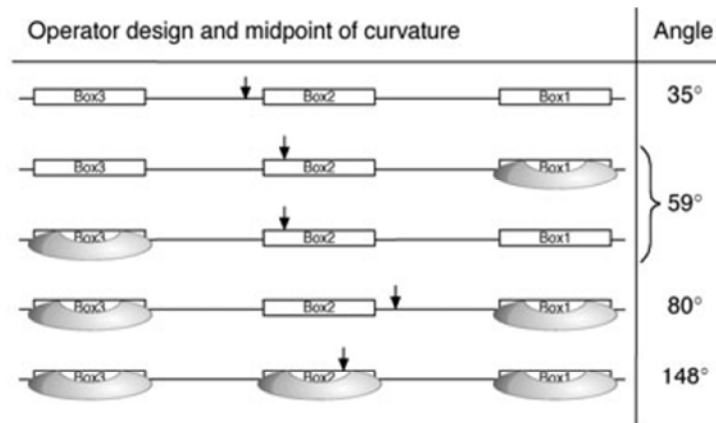


Figure 22 : Représentation des différents complexes formés après la liaison de Ss-LrpB à son propre promoteur et les courbures induites associées. Les flèches indiquent le milieu de la courbure (modifié depuis Peeters *et al.*, 2004)

Les sites de liaison de Ss-LrpB ont été identifiés dans la région promotrice d'un opéron codant pour une pyruvate-ferredoxine-oxidoréductase (*porDAB*) et en amont de deux gènes codant des perméases putatives (Peeters *et al.*, 2009). Une combinaison d'études de liaison *in vitro* et d'analyses d'expression de gènes *in vivo* ont montré que Ss-LrpB agit comme un activateur de la transcription de ces trois gènes comme prédit initialement (Peeters *et al.*, 2009). De façon intéressante, l'organisation des sites de liaison du régulateur dans *por* est le même que dans la région promotrice de Ss-LrpB, ce qui suggère que des mécanismes de régulation similaires, dépendants de la concentration de la protéine régulatrice, s'appliquent également à cet opéron.

Il a été démontré *in vitro* que Ss-LrpB peut fonctionner à la fois comme un activateur et comme un répresseur de la transcription de son propre gène et ceci de façon dépendante de sa concentration (Peeters *et al.*, 2009). La liaison coopérative de dimères et le recouvrement de l'ADN semble être le mécanisme expliquant l'effet positif de Ss-LrpB à faible concentration de la protéine et l'effet négatif qu'elle exerce sur la transcription de son propre gène à forte concentration.

Des informations sont également disponibles concernant un autre régulateur de type Lrp identifié chez les *Sulfolobales*. Ainsi, **LysM** de *S. solfataricus* régule la biosynthèse de la lysine et peut-être de l'arginine, à travers le groupe de gènes *lys*. La protéine LysM se lie à une UAS dans la région d'un promoteur faible *lysW*, juste en amont de l'élément BRE. La liaison est favorisée en l'absence de lysine, lorsque la transcription de ce cluster est maximale (Brinkman *et al.*, 2002). Lorsque de la lysine exogène est présente, elle sert de ligand pour LysM, modifiant son affinité de liaison à l'ADN et inhibant l'effet de l'activation (Brinkman *et al.*, 2002). Récemment, soixante-dix nouveaux sites de liaison pour LysM chez *S. solfataricus* ont été identifiés par des expériences de ChIP-chip (« chromatin immunoprecipitation ») et confirmés *in vitro*. LysM se fixe à des régions intergéniques et codantes, incluant les promoteurs de plusieurs gènes impliqués dans la biosynthèse d'acides aminés et le transport de gènes. LysM apparaît comme un régulateur plus polyvalent que pensé initialement capable d'utiliser une variété d'acides aminés pour détecter la qualité nutritionnelle de l'environnement et moduler en fonction la machinerie métabolique de *Sulfolobus* (Song *et al.*, 2013).

ii. Analyse structurale des protéines Lrp archées

Même si les protéines de type Lrp démontrent une faible conservation de séquences aminoacides, elles ont une structure bien conservée. Les structures cristallines de plusieurs protéines Lrp de plusieurs phyla d'*Archaea* ont été résolues récemment (Peeters and Charlier, 2010). La Figure 19 présente quelques exemples de ces régulateurs.

Ces protéines de petit poids moléculaire sont typiquement des monomères d'environ 15kDa, dont les deux domaines caractéristiques ont été bien définis : le domaine de liaison à l'ADN, en N-terminal, contenant le motif HTH (le plus souvent, il s'agit d'un motif wHTH); et le domaine en C-terminal qui facilite l'oligomérisation et la liaison avec l'effecteur, aussi connu comme le domaine RAM (**R**egion of **A**mino-acid **M**etabolism) (Ettema *et al.* 2002).

Le fait que la plupart des protéines de type Lrp d'archées cristallisent sous forme d'octamères ou de formes d'association plus élevées indique la capacité de ces protéines à s'oligomériser (Koike *et al.*, 2004). L'unité fonctionnelle est un dimère où l'interaction entre deux brins β appartenant à deux domaines RAM différents, $\beta 2$ et $\beta 5$, établit un noyau hydrophobe stable (Enoru-Eta *et al.*, 2002; Ouhammouch *et al.*, 2001; Brinkman *et al.*, 2000). La reconnaissance des séquences spécifiques de liaison à l'ADN est obtenue par l'insertion d'hélices α du domaine de liaison à l'ADN dans le grand sillon de l'ADN (Aravind *et al.*, 2005; Wintjens *et al.*, 1996; Harrison *et al.*, 1990).

iii. Interaction de Lrp avec les séquences cibles

La définition d'une séquence consensus pour une protéine de liaison à l'ADN permet normalement l'identification de toutes ses cibles spécifiques dans le génome. Cette information ouvre la possibilité d'étudier le réseau de régulation entier sous le contrôle d'une protéine.

Dans le cas des protéines de la famille Lrp, la conservation réduite des sites de liaison rend l'identification des cibles plus problématique. Par exemple, les hélices α du domaine de liaison à l'ADN des dimères Lrp d'archées reconnaissent des sites de treize à dix-sept nucléotides de long, avec des répétitions dégénérées inversées (Ouhammouch and Geiduschek, 2001; Peeters *et al.*, 2007; Yokoyama *et al.*, 2009). Il y a différentes techniques disponibles pour identifier le site spécifique de liaison à l'ADN d'un régulateur. Pour beaucoup de protéines d'archées de type Lrp, cela été réalisé par l'approche de SELEX (Dahlke and Thomm, 2002; Yokoyama *et al.*, 2005; Yokoyama *et al.*, 2009). Une autre approche pour mieux définir la séquence conservée d'une cible est de réaliser des substitutions de nucléotides sur un site consensus établi par l'alignement des sites de liaison connus (Peeters *et al.*, 2007).

Dans les génomes d'archées, les cibles des protéines de type Lrp semblent souvent être dans des groupes de sites multiples imparfaits, avec chacun une affinité de liaison différente. Néanmoins, pour les différentes protéines de famille Lrp caractérisées chez les archées, plusieurs profils d'organisation des sites de liaison et de coopération entre des copies multiples de la protéine peuvent être distingués :

(i) la liaison coopérative de dimères séparés à des sites multiples qui interagissent étroitement et pourraient former des oligomères d'ordre plus élevé (Peeters *et al.*, 2004).

(ii) la liaison d'une forme d'oligomère d'ordre plus élevé (tétramères), à un ou plusieurs sites de liaison régulièrement espacés (Figure 19c) (Brinkman *et al.*, 2000; Brinkman *et al.*, 2002).

(iii) une liaison moins spécifique d'oligomères d'ordre plus élevé (octamères) à de longues régions d'ADN de plus de 100pb (Figure 19b et 19d) (Enoru-Eta *et al.*, 2000; Yokoyama *et al.*, 2005).

Dans certains cas, pour le même régulateur de type Lrp, plusieurs profils de liaison peuvent être observés; le degré d'oligomérisation dépend alors de l'interaction de Lrp avec un effecteur (Yokoyama *et al.*, 2006) qui sont le plus souvent des acides aminés correspondant aux voies de biosynthèse régulées par le régulateur Lrp correspondant.

Des exemples détaillés correspondant aux trois profils de liaison mentionnés ci-dessus ont été présentés et discutés dans la section I.4.2.b.i.

c. Les régulateurs HTH de la famille MarR

Les protéines de la famille MarR (**M**ultiple **a**ntibiotic **r**esistance **R**egulators) constituent un groupe divers de régulateurs transcriptionnels qui modulent l'expression de gènes codant pour des protéines impliquées dans une grande variété de mécanismes cellulaires, incluant les voies métaboliques, la virulence, les réponses au stress et l'efflux des molécules nocives pour la cellule, tels que les antibiotiques, les composés phénoliques et les détergents ménagers communs (Perera and Grove, 2010). Toutes les protéines MarR sont capables de lier l'ADN grâce à un motif wHTH conservé et des ligands spécifiques (beaucoup ne sont pas encore identifiés), qui modulent, lors de la liaison, la reconnaissance de l'ADN par la protéine.

Dans le genre *Sulfolobus*, les protéines membres de cette famille semblent être très importantes pour l'adaptation de ces organismes à des habitats ou des modes de vie particuliers. La recherche *in silico* de gènes *marR* (partageant une identité de séquence des acides aminés de 30%) dans les génomes de *S. solfataricus*, *S. islandicus*, *S. tokodaii* et *S. acidocaldarius*, révèle qu'à la différence de la majorité des bactéries, les *Sulfolobus* analysées possèdent toutes au moins deux copies du gène *marR*. Ceci peut indiquer la présence de plusieurs régulons *marR* dans la même cellule. Les rôles de quelques représentants de la famille MarR ont été caractérisés au cours de ces dernières années (Fiorentino *et al.*, 2007; Di Fiore *et al.*, 2009; Fiorentino *et al.*, 2011; Kumarevel *et al.*, 2009) (Tableau 7) mais pour beaucoup d'autres le rôle et les cibles dans le génome ne sont toujours pas connus.

Chez les *Sulfolobus*, cinq protéines issues de cette famille, sont caractérisées comme régulateurs transcriptionnels.

Tableau 7 : Régulateurs transcriptionnels décrits expérimentalement chez les *Sulfolobus*.

Famille	Espèce <i>Sulfolobus</i>	Membre	Forme oligomérique	Fonction	Références
Lrp	<i>S. solfataricus</i> P2	Lrs14	Dimère	Répresseur	Napoli <i>et al.</i> , 1999; Bell <i>et al.</i> , 2001
		SsLrp	Tétramère	Répresseur putatif	Enoru-Eta <i>et al.</i> , 2002
		SsLrpB	Dimère	Activateur	Peeters <i>et al.</i> , 2004; 2006; 2009; 2013
		LysM	N. d.	Activateur putatif	Brinkman <i>et al.</i> , 2002; Song <i>et al.</i> , 2013
		CopT	N. d.	Répresseur putatif	Ettema <i>et al.</i> , 2006
	<i>S. solfataricus</i> 98/2	CopR	N. d.	Activateur	Villafane <i>et al.</i> , 2009
	<i>S. acidocaldarius</i>	Sa-Lrp	Tétramère/ Octamère	Activateur/ répresseur putatif	Enoru-Eta <i>et al.</i> , 2000; Vassart <i>et al.</i> , 2013
<i>S. tokodaii</i>	Grp	Octamère	-	Kumarevel <i>et al.</i> , 2008; Nakano <i>et al.</i> , 2007	
MarR	<i>S. solfataricus</i> P2	BldR	Dimère	Activateur putatif	Di Fiore <i>et al.</i> , 2009; Fiorentino <i>et al.</i> , 2007
		BldR2	Dimère	Répresseur putatif	Fiorentino <i>et al.</i> , 2011
		Sta1	Dimère	Activateur	Kessler <i>et al.</i> , 2006
	<i>S. tokodaii</i>	St1710	Dimère	Répresseur	Kumarevel <i>et al.</i> , 2008; 2009
Autres	<i>S. solfataricus</i> P2	SsMerR	Dimère	Activateur	Schelert <i>et al.</i> , 2004; 2006

En particulier, le groupe de G. Fiorentino a caractérisé récemment deux membres de la famille MarR chez *S. solfataricus*, les protéines **BldR** et **BldR2**. Il a été démontré que la protéine BldR participe à la régulation de la voie de détoxification des composés aromatiques. BldR reconnaît une séquence spécifique composée de répétitions inversées adjacentes en amont de l'élément BRE de la région promotrice du gène de l'alcool déshydrogénase (Fiorentino *et al.*, 2003). BldR active la transcription soit en améliorant le recrutement de TBP et TFB, soit en stabilisant leur liaison. BldR se lie également à sa propre région promotrice, contrôlant un opéron de type *marR*, composé de son propre gène et d'un transporteur putatif. Le benzaldéhyde, le substrat des alcool-déshydrogénases, stimulent la liaison de BldR à ses deux cibles identifiées (Fiorentino *et al.*, 2007; Cannio *et al.*, 1999; Fiorentino *et al.*, 2003). BldR2 est une protéine dimérique qui se lie spécifiquement à son propre promoteur au niveau de deux sites palindromiques chevauchant la *TATA-box*, la séquence BRE et le site d'initiation de la transcription. Selon la localisation des éléments basaux du promoteur et de la région de liaison à BldR2, il a été proposé que, contrairement à BldR, cette protéine peut également agir comme un répresseur (Fiorentino *et al.*, 2011).

La structure cristalline d'un régulateur proche de BldR, **ST1710** chez *S. tokodaii* a aussi été déterminée. L'étude montre que, bien que la liaison du ligand n'affecte pas la structure entière, un changement conformationnel significatif se produit à la suite de la liaison du régulateur à l'ADN cible (Kumarevel *et al.*, 2008; 2009; Miyazono *et al.*, 2007).

TrmB est un régulateur général du transport des sucres chez les euryarchées *Thermococcus litoralis* et *P. furiosus*. Chez chacune de ces deux espèces, cette protéine contrôle au moins deux opérons. L'un, l'opéron *mal*, code pour les transporteurs ABC du maltose et du tréhalose (système TM); le second, l'opéron *mdx*, code pour un transporteur ABC des maltodextrines (système MD). TrmB se lie aux deux régions promotrices dans différents endroits : dans l'opéron *mal* son site de liaison chevauche la *TATA-box* et dans l'opéron *mdx* il couvre le site d'initiation de la transcription. TrmB est un répresseur transcriptionnel. Etant donné les positions du site de liaison de TrmB dans ces deux promoteurs, dans le premier cas, il bloque la formation du PIC en réprimant la liaison de TBP et TFB et dans le second, il empêche le recrutement de l'ARNP. La fixation de différents sucres à TrmB crée des changements de conformation de la protéine, changeant son affinité de liaison à l'ADN. Le maltotriose et le saccharose agissent comme inducteurs du système MD en se liant directement à TrmB (Lee *et al.*, 2005), tandis que le maltose est un inducteur pour le système TM et un répresseur du système MD (Lee *et al.*, 2003). En outre, le glucose peut se lier à TrmB et agir comme un répresseur des systèmes TM et MD, en enlevant les inducteurs, maltose et maltotriose par compétition (Lee *et al.* 2005; 2007). Ce système simple donne à ces archées un instrument moléculaire astucieux leur permettant de transporter préférentiellement un sucre (glucose) au détriment des autres. Ceci par son effet physiologique (mais pas par son mécanisme) ressemble dans les finalités à la répression catabolique décrite chez des bactéries (Lee *et al.*, 2007).

Un des exemples les mieux étudiés est la protéine **Sta1** codée par la crénarchée *S. islandicus*. Sta1 est impliquée dans la régulation du cycle de vie des ruidivirus SIRV1/2. Ce régulateur a été isolé par une expérience de *pull-down* avec un extrait brut de *S. islandicus* en utilisant comme appâts les fragments d'ADN correspondant aux promoteurs de certains gènes viraux (Kessler *et al.*, 2006). L'analyse faite par l'empreinte à la DNase I a montré que Sta1 reconnaît deux sites de liaison de 16pb avec des répétitions inversées. Un de ces sites est localisé 30pb en amont des éléments du promoteur, alors que le second est localisé juste en aval de la *TATA-box* du promoteur, ce qui suggère que l'effet de l'activation pourrait provenir d'une interaction soit avec TBP, soit avec TFB. Sta1 a été décrite comme un activateur de la transcription de trois promoteurs de SIRV1. Il a récemment été proposé qu'en plus de son rôle dans l'activation des promoteurs viraux, Sta1 active également l'expression d'un gène hôte *sso0777*, paralogue du gène *radA*, vraisemblablement impliqué dans la réparation de l'ADN (Abella *et al.*, 2007). Le gène *sta1* fait partie de l'étude réalisée au cours de cette thèse (voir chapitre Résultats).

d. Les autres régulateurs HTH chez les archées

Chez les *Archaea*, la régulation de la transcription par les protéines à domaine de liaison à l'ADN de type HTH n'est pas limitée aux familles Lrp et MarR. Ci-dessous, sont discutés quatre exemples d'autres régulateurs HTH impliqués dans le contrôle de la transcription chez les *Archaea*. Le choix a été fait de se concentrer uniquement sur les archées hyperthermophiles.

Quatre régulateurs transcriptionnels portant un domaine HTH et ne faisant pas partie des familles Lrp et MarR ont été caractérisés jusqu'ici chez les archées hyperthermophiles.

Mdr1 (*Metal Dependant Repressor 1*) de la crénarchée *Archaeoglobus fulgidus* a été le premier régulateur transcriptionnel d'archées caractérisé (Bell *et al.*, 1999). Le gène *mdr1* est situé en amont de trois gènes codant pour un transporteur de métaux ABC. *In vivo*, la transcription de cet opéron comprenant ces quatre gènes est strictement dépendante de la disponibilité des ions métalliques

dans le milieu de croissance. *In vitro*, Mdr1 se lie de façon métallo-dépendante à trois sites sur le promoteur de l'opéron et réprime la transcription par chevauchement avec le site d'initiation de la transcription, en empêchant la fixation de l'ARNP (Bell *et al.*, 1999).

SurR de *Pyrococcus furiosus* a été le premier exemple d'une protéine impliquée dans le contrôle d'activation de la transcription par oxydoréduction chez les *Archaea*. La protéine régule la production d'hydrogène d'une manière sulfo-dépendante.

P. furiosus produit du H₂ pendant la fermentation, mais en présence de S⁰, il y a un changement métabolique immédiat vers la production de H₂S (Fiala and Stetter, 1986). L'analyse de l'expression à l'aide de puces à ADN a révélé que dix minutes après l'ajout de S⁰, la répression de tous les opérons codant pour les hydrogénases et l'activation de l'expression des gènes nécessaires pour le métabolisme de S⁰ sont observées (Schut *et al.*, 2007).

Une séquence consensus de liaison à l'ADN de SurR a été caractérisée par l'approche SELEX. La majorité des gènes dont l'expression est modifiée en présence de S⁰ contiennent ce motif conservé dans leurs régions promotrices. La position des sites de liaison dans les régions promotrices semble discriminante pour déterminer quel effet, activation ou répression, SurR aura sur ce promoteur. Dans le cas d'une activation, SurR se lie à une région située juste en amont de l'élément BRE. Cette position lui permet en principe d'activer la transcription par un mécanisme semblable à celui utilisé par Ptr2. Dans le cas d'une répression, elle semble se produire à travers le mécanisme de blocage des sites de liaison de TBP et TFB ou en inhibant le recrutement l'ARNP (Lipscomb *et al.*, 2009). L'activité de SurR est modulée par oxydoréduction. L'oxydation par S⁰ diminue l'activité de liaison de SurR à des séquences spécifiques, libérant les gènes impliqués dans le métabolisme de S⁰ de la répression de SurR et diminuant la transcription des gènes participant à la production de H₂ qui requièrent l'activation de SurR (Yang *et al.*, 2010).

La protéine **Phr** de *P. furiosus* est un régulateur de la réponse au choc thermique. Ce régulateur est bien conservé parmi les *Euryarchaeota*. Phr réprime sa propre transcription, celle du gène codant pour la protéine de choc thermique Hsp20 et celle d'une AAA+ ATPase à des températures de croissance normale. Des tests de liaison à l'ADN ont révélé que la transcription de ces promoteurs est réprimée par l'inhibition de la liaison de l'ARNP. Aussi, des analyses mutationnelles ont montré qu'il faut un motif TTTA présent à la position -10 pour la liaison de Phr à l'ADN (Vierke *et al.*, 2003). A des températures supérieures à 107° C Phr s'est dissociée des promoteurs permettant le développement de la réponse au choc thermique (Liu *et al.*, 2007). Le mécanisme assurant l'activation des gènes codant pour cette réponse n'est toujours pas décrit.

La protéine **SSMerR** codée par le génome de *S. solfataricus* possède un domaine wHTH. Il est phylogénétiquement éloigné de la majorité de ses équivalents bactériens à cause de sa longueur plus courte et de la présence de seulement deux résidus cystéines généralement requis pour le ligand, au lieu de trois (Schelert *et al.*, 2004). SsMerR agit comme un répresseur de l'opéron *merHAI* en se liant à une répétition inversée qui chevauche la séquence BRE du promoteur *merH*. Hg²⁺ dérprime la transcription en interagissant avec SsMerR (Schelert *et al.*, 2006).

Les exemples présentés montrent comment les *Archaea* utilisent largement des régulateurs transcriptionnels de type bactérien pour interagir avec l'ADN par l'intermédiaire de leur domaine de liaison de HTH. La répression de la transcription implique généralement la compétition entre les

régulateurs et les facteurs généraux de transcription (TBP ou TFB) pour la région du promoteur. Les mécanismes d'activation de la transcription sont moins bien compris chez les *Archaea*. L'utilisation de sites UAS implique probablement des changements complexes de la conformation de l'ADN ainsi que des interactions directes avec le complexe d'initiation de la transcription.

e. La famille *Ribbon-Helix-Helix* des régulateurs transcriptionnels

La résolution des structures cristallines des répresseurs Arc et Mnt codés par le phage P22 (Knight *et al.*, 1989; Breg *et al.*, 1990) et celle du répresseur MetJ d'*E. coli* cristallisé en complexe avec l'ADN (Somers and Phillips, 1992) a clairement établi qu'au point de vue de leur structure et de leur façon d'interagir avec l'ADN ces protéines constituent un groupe à part par rapport à la famille des protéines HTH. Agissant, habituellement, dans la plupart des cas sous forme de dimères, ces protéines reconnaissent spécifiquement des séquences ADN à l'aide d'un feuillet β introduit dans le grand sillon de l'ADN. Leur forme fonctionnelle montre l'ordre conservé suivant : brin β -hélice α -hélice α . Alors, elles ont été nommées : les protéines **Ruban-Hélice-Hélice** (RHH).

Des gènes correspondant à des régulateurs transcriptionnels RHH ont été identifiés *in silico* chez les bactéries et leurs phages aussi bien que chez toutes les archées et leurs virus (Schreiter and Drennan, 2007). A ce jour, aucune protéine portant le domaine RHH n'est retrouvée chez les eucaryotes. Dans une analyse réalisée en 2007 avec les données disponibles sur la base de données InterPro, il a été prédit l'existence jusqu'à 5355 gènes putatifs pouvant coder pour des protéines avec le motif RHH chez les *Bacteria*, 390 chez les *Archaea* et 39 dans les génomes viraux (Schreiter and Drennan, 2007). Actuellement, seulement 21 régulateurs RHH sont bien caractérisés. Il est évident que ce groupe de protéines représente une famille de régulateurs très importants et largement utilisés pour contrôler de multiples fonctions biologiques, résumées dans le Tableau 8.

Tableau 8 : Régulateurs transcriptionnels *ribbon-helix-helix* caractérisés et leurs fonctions (modifié d'après Schreiter and Drennan, 2007).

Nom de la protéine et numéro d'accèsion	Organisme ou plasmide	Fonction	Type de régulateur
Arc, P03050	Bactériophage P22	Contrôle du cycle lytique du bactériophage	Négatif
CopG, P13920	<i>Streptococcus species</i>	Contrôle du nombre de copies de plasmide	Négatif
MetJ, P0A8U6	<i>Escherichia Coli</i> , autres bactéries	Régulation dépendante d'AdoMet de gènes de biosynthèse de méthionine	Négatif
Omega, Q83UM6	Plasmides de la famille Inc18 chez les bactéries Gram positif	Contrôle du nombre de copies de plasmide	Négatif
FitA, Q9RF92	<i>Neisseria gonorrhoeae</i>	Antitoxine	Négatif
NikR, P0A6Z6	<i>E. coli</i> , <i>Helicobacter pylori</i> , <i>Pyrococcus horkoshff</i> , autres bactéries et archées	Répression Ni ²⁺ dépendante de la prise de nickel chez <i>E. coli</i> , induction de l'uréase chez <i>H. pylori</i>	Négatif chez <i>E. coli</i> , négatif et positif chez <i>H. pylori</i>
CcdA, Q46995	Plasmide F	Antitoxine	Négatif
Mnt, P03049	Bactériophage P22	Contrôle du cycle lytique du phage	Négatif
ParG, Q70W75	Plasmide TP228 multi résistant aux antibiotiques	Partionnement de plasmide	Négatif
PutA, P09546	<i>E. coli</i> , autres bactéries	Régulation des gènes qui utilisent la proline, oxydation de la proline en glutamate	Négatif
HP0222, Q9ZMLO	<i>H. pylori</i>	Inconnue	Inconnu
AlgZ, Q9RPY7	<i>Pseudomonas aeruginosa</i>	Régulation de la biosynthèse de l'alginate et motilité	Positif et négatif
ParD, P22995	Plasmide RP4-RK2	Antitoxine	Négatif
7kMk, Q8TH48	<i>Methanopyrus kandleri</i>	Inconnue	Inconnu
TrwA, Q04229	Plasmide R388	Régulation de la conjugaison bactérienne	Négatif
Tray, P06627	Plasmide F	Régulation de la conjugaison bactérienne	Positif

Le rôle des protéines RHH chez les *Archaea* est assez peu connu. Aussi, je commencerai par introduire leurs caractéristiques générales et leurs propriétés biologiques, au travers des exemples bactériens observés. Leur structure typique ainsi que les spécificités de leur liaison à l'ADN et de leurs séquences cibles seront données. Enfin, les trois exemples connus chez les archées seront décrits à la fin de ce chapitre.

i. Présentation générale des protéines de la famille RHH

1. Structure du domaine RHH

Malgré une faible ressemblance entre les séquences des protéines RHH connues, leur structure est bien conservée. Dans toutes les protéines RHH, la structure qui interagit avec l'ADN est composée d'un feuillet β en N-terminal, suivi par deux hélices α , α_1 et α_2 (Rafferty *et al.*, 1989; Breg *et al.*, 1990; Raumann *et al.*, 1994).

D'une part, le feuillet β en N-terminal est composé d'une alternance de chaînes latérales hydrophobes et hydrophiles entre les positions 2 et 7. Les chaînes latérales hydrophobes se dirigent vers le noyau du domaine (les positions 3, 5 et 7, Figure 23). Les chaînes latérales hydrophiles s'éloignent du domaine et servent pour établir des contacts spécifiques avec les nucléotides du site de fixation lors de la liaison à l'ADN (les positions 2, 4 et 6 sur la Figure 23a).

D'autre part, dans la boucle formée par les hélices α_1 et α_2 , un motif G-X-T/S/N est conservé chez tous les membres de la famille RHH. Cette boucle permet le positionnement correct de l'extrémité N-terminale de l'hélice α_2 , qui établit des contacts d'ancrage avec le squelette phosphate de l'ADN (Figure 23c). La protéine **MetJ** est une exception, ayant six acides aminés additionnels entre les deux hélices (Schreiter and Drennan, 2007).

En outre, le noyau hydrophobe est formé non seulement par les aminoacides en positions 3, 5 et 7 du feuillet β , mais également par quatre aminoacides hydrophobes conservés dans les hélices α_1 et α_2 . Ces sept positions hydrophobes conservées constituent le motif le plus conservé parmi les séquences RHH. En effet, même si la longueur de la boucle $\alpha_1 - \alpha_2$ peut varier, modifiant la distance entre ces résidus hydrophobes, maintenir l'intégrité du noyau hydrophobe de ces petites protéines est crucial (Schreiter and Drennan, 2007).

Au niveau de la structure, la petite protéine **CopG** de quarante-cinq acides aminés est l'exemple d'une protéine qui se réduit presque au domaine RHH pur. Elle est considérée comme la structure minimale de liaison à l'ADN de la famille RHH (del Solar *et al.*, 2002). Structurellement, elle est très proche du répresseur **Arc** du bactériophage P22 de *Salmonella*. La principale différence réside dans l'absence chez CopG de la région N-terminale impliquée normalement dans l'établissement des contacts non spécifiques avec l'ADN (Raumann *et al.*, 1994). Les sous-sites adjacents de la région du promoteur sont occupés par des dimères de dimères de CopG qui se lient de manière coopérative, couvrant les éléments du promoteur et réprimant la transcription (Bowie et Sauer, 1989; Gomis-Ruth *et al.*, 1998; del Solar *et al.*, 2002). Un élément central pseudo-symétrique de treize nucléotides de long est la cible primaire de reconnaissance de la protéine CopG (del Solar *et al.*, 1990). Dans cet élément central, deux séquences de 4pb (5' – TGCA – 3') sont séparées par 5pb. Chaque dimère de CopG reconnaîtra et se liera directement à chacune des moitiés de l'élément central (Gomis-Ruth *et al.*, 1998). CopG entre en contact avec environ 50 pb d'ADN dans la région du promoteur, ce qui suppose que les quatre dimères de la protéine lient, de façon coopérative, l'ADN.

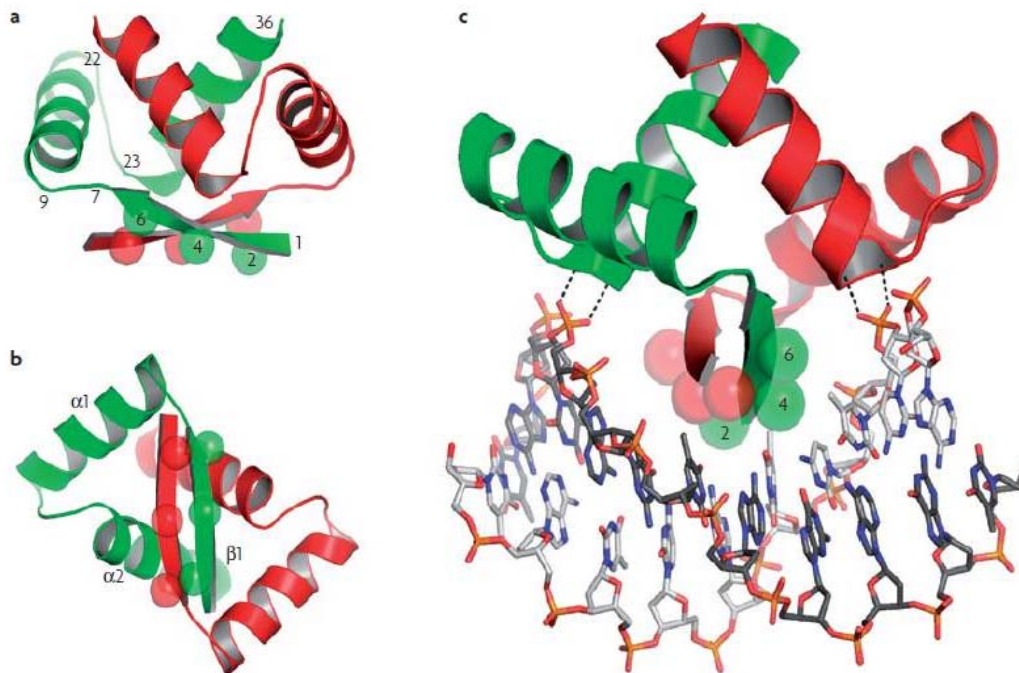


Figure 23 : Représentation des structures conservées et liaison à l'ADN des protéines RHH. Trois vues d'un domaine (RHH₂) en dimère est montré sur ces images et coloré par sous-unité. Les positions des trois acides aminés (2, 4 et 6) de chaque sous-unité qui sont impliqués dans les contacts spécifiques avec les nucléotides sont les sphères mi-transparentes. (a) Les positions références dans le motif RHH sont numérotées sur la sous-unité verte. (b) Les éléments de la structure secondaire sont désignés sur la sous-unité verte. (c) Interaction entre un domaine RHH₂ avec l'ADN opérateur. Les contacts d'ancrage non spécifiques entre la terminaison N de la seconde hélice α et le squelette ADN phosphate sont montrés en lignes noires pointillées. Les contacts spécifiques sont faits par les positions 2, 4 et 6 de chaque sous-unité du feuillet β à l'intérieur du grand sillon de l'AND (modifié d'après Schreiter and Drennan, 2007).

2. Spécificités de liaison à l'ADN des protéines RHH et de leurs séquences cibles

L'unité fonctionnelle de ce domaine de liaison à l'ADN est un dimère (désigné par RHH₂), formé par deux motifs RHH soigneusement entremêlés pour former un domaine stable avec une double symétrie (Schreiter and Drennan, 2007; Bowie *et al.*, 1990)

Les feuillet β des deux monomères sont appariés dans le dimère, formant un ruban β antiparallèle qui est positionné dans le grand sillon de l'ADN et qui crée des contacts spécifiques avec les nucléotides par l'intermédiaire des trois résidus hydrophobes déjà mentionnés (positions 2, 4 et 6) de chaque monomère.

Il existe seulement deux exemples de protéines RHH agissant sans passer par la phase de l'homodimérisation. Dans ces cas, le motif RHH est dupliqué et il est présent en tandem dans la séquence de la protéine. Une de ces deux exceptions est la protéine **TraY** codée par le plasmide F (Bowie etand Sauer, 1990; Lum *et al.*, 2002) qui est impliquée dans la régulation de la conjugaison bactérienne. Sa structure n'est pas encore résolue. L'autre, retrouvée chez les *Archaea*, est la protéine AvtR codée par le virus crénarchée AFV6 infectant le genre *Acidianus* (Peixeiro *et al.*, 2013).

Toutes les protéines RHH caractérisées lient l'ADN en reconnaissant plusieurs sites situés dans les régions promotrices. Ces sites peuvent être répétés en tandem ou être en ordre inversé. En se liant à ces sites les protéines RHH forment des oligomères d'ordre élevé (normalement des dimères de dimères) grâce à l'interaction protéines-protéines existant entre les dimères liés à l'ADN. Par exemple, **MetJ** lie des séquences symétriques qui suivent la double symétrie des dimères. Après l'activation par AdoMet, les dimères MetJ se lient à l'ADN de façon coopérative, comme des dimères de dimères (Figure 24). Plusieurs sous sites peuvent être trouvés dans la région de l'opérateur et l'espace entre eux est intimement connecté aux interactions protéines-protéines entre les dimères. Jusqu'à cinq sous-sites répétés en tandem sont retrouvés dans la région promotrice des gènes régulés par MetJ. Cette observation peut expliquer la différence dans la régulation de différents gènes sous contrôle de MetJ (Augustus *et al.*, 2006).

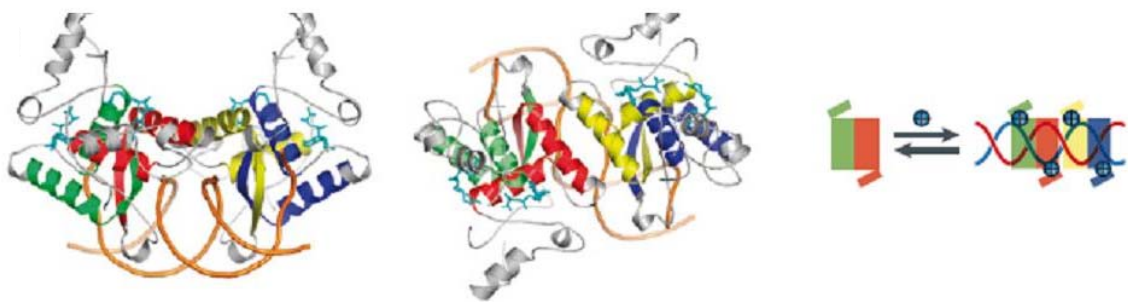


Figure 24 : Mécanisme de liaison à l'ADN de MetJ de la famille RHH. Les vues de côté et au-dessus des structures cristallines du complexe ADN de MetJ. Les représentations dessinées de la portion RHH sont colorées selon la sous-unité; les autres régions de la protéine sont en gris; l'ADN est en orange. Les dimères MetJ sont activés par AdoMet et se lient de façon coopérative à l'ADN (modifié d'après Schreiter and Drennan, 2007).

Dans les régions promotrices reconnues par le répresseur **oméga** codé par les bactéries Gram positives, jusqu'à dix sous-sites en tandem peuvent être trouvés, disposés en trois configurations différentes. Les sous-sites adjacents peuvent être positionnés comme des séquences répétées en tandem ou comme des séquences répétées inversées, faisant face vers l'intérieur ou vers l'extérieur et l'organisation de ces sous-sites influence l'affinité du répresseur pour les différents promoteurs (de la Hoz *et al.*, 2004; Weihofen *et al.*, 2006).

La résolution des structures des régulateurs liés à l'ADN a révélé qu'il n'y a aucune modification conformationnelle dans le motif RHH après la liaison. En outre, quand MetJ (Somers and Phillips, 1992), CopG (Gomis-Ruth *et al.*, 1998; Costa *et al.*, 2001) et Arc (Raumann *et al.*, 1994) lient leurs cibles ADN, ils induisent différents degrés de courbure sur la molécule ADN. La courbure de l'ADN induite par MetJ se produit au centre du sous-site, où se font les contacts spécifiques entre l'ADN et le feuillet β du dimère. Dans le cas de CopG et d'Arc, chaque dimère reconnaît différentes moitiés du sous-site et la courbure a lieu au centre du sous-site entre les deux dimères. Le répresseur Oméga est une exception, comme le complexe ADN-Oméga est pratiquement linéaire (Weihofen *et al.*, 2006).

Après avoir présenté les caractéristiques générales des protéines RHH, la section suivante fera état de la situation actuelle pour les régulateurs transcriptionnels connus chez les archées.

ii. Les régulateurs RHH chez les archées

Comme les bactéries, les archées codent en abondance des protéines RHH. Cependant, seulement trois cas de ces protéines ont été caractérisés en détails, à ce jour.

Chez les *Archaea*, comme les connaissances sur la famille RHH sont encore clairsemées et presque exclusivement limitées aux virus crénarchées, il n'est pas exclu que des motifs adjacents et des fonctions existent et intègrent différents facteurs dans la modulation de l'expression de différents gènes.

L'analyse *in silico* des génomes des virus des crénarchées a révélé que les gènes codant de petites protéines contenant le domaine RHH sont présentés de façon ubiquitaire dans les génomes de tous les virus d'archée décrits (Prangishvili *et al.*, 2006). Par exemple, dans les dix génomes viraux de crénarchées analysés, au moins dix-huit régulateurs RHH putatifs ont été répertoriés (Prangishvili *et al.*, 2006). Malgré leur abondance, très peu est connu sur le rôle physiologique et la spécificité des protéines RHH chez les *Archaea*.

La première protéine RHH archée caractérisée est la protéine ORF56, codé par le plasmide pRN1 de *S. islandicus*, qui régule le nombre de copie de ce plasmide. Ses structures RMN et cristalline ont été publiées en 2009 (Weininger *et al.*, 2009).

ORF56 est une petite protéine dimérique de liaison à l'ADN, très stable thermodynamiquement (Lipps *et al.*, 2001). Le plasmide pRN1 a été séquencé (Keeling *et al.*, 1996) et elle code pour cinq transcrits effectivement détectés (Berkner and Lipps, 2007). Parmi eux, un long transcrit qui inclut le gène *orf56*, qui code pour le régulateur, et le gène *orf904*, qui code pour un initiateur putatif de la réplication du plasmide. *In vitro*, la protéine se lie sous forme d'un tétramère à une répétition inversée de 12pb présente dans la région promotrice de cet opéron (Lipps *et al.*, 2001). Le site de liaison ne chevauche ni l'élément BRE, ni la *TATA-box*, ce qui exclut la possibilité de réprimer la transcription par interaction directe avec TBP ou TFB. Le site de liaison de la protéine ORF56, néanmoins, chevauche le site d'initiation de la transcription, ce qui suggère qu'il pourrait nuire à la liaison de l'ARNP à cette région. En effet, des expériences de gène reporteur couplé au promoteur de l'*orf56* ont démontré que la protéine ORF56 réprime la transcription de cet opéron *in vivo* (Berkner and Lipps, 2007). A ce jour, l'interaction d'ORF56 avec la machinerie de transcription reste incomprise.

La résolution des structures RMN (Weininger *et al.*, 2009) et cristalline (Weininger *et al.*, 2009) du régulateur ORF56 a confirmé la présence d'un motif de liaison à l'ADN RHH et des expériences de titration ont révélé que comme attendu, les feuilletts β interagissaient avec l'ADN, comme pour le reste des membres de la famille RHH (Weininger *et al.*, 2009).

L'analyse structurale et fonctionnelle de la protéine **AvtR** (*Acidianus virus transcriptional regulator*) codée par le virus AFV6 (*Acidianus filamentous virus 6*), qui infecte les archées du genre *Acidianus*, révèle une structure inhabituelle et son rôle dans la régulation transcriptionnelle de plusieurs gènes viraux. La structure cristalline d'AvtR (cent acides aminés) montre qu'elle est constituée d'un motif RHH répété. Des expériences de SELEX, combinées à des mutagénèses systématiques du site prédit, ont prédit huit cibles potentielles de la protéine AvtR dans le génome du virus AFV6. Deux de ces cibles ont été étudiées et le rôle complexe de la protéine AvtR dans la régulation de la transcription

des gènes viraux a été établi. AvtR réprime l'expression de son propre gène *gp29*, mais peut aussi agir comme activateur d'un autre gène *gp30*. Ses sites de liaison sont distants des *TATA-box* des promoteurs de ces gènes et le mécanisme de régulation dépendant d'AvtR semble inclure une oligomérisation débutant au site initial de liaison de la protéine. Plusieurs régulateurs RHH transcriptionnels pourraient partager ce mécanisme régulateur (Peixeiro *et al.*, 2013).

Pour finir, au cours des recherches présentées dans ce mémoire, j'ai approfondi l'étude d'un troisième régulateur RHH transcriptionnel viral, **SvtR** du virus SIRV1 (*Sulfolobus islandicus rod-shaped virus 1*) (Guillière *et al.*, 2009). De nouveaux gènes cibles ont été identifiés et son rôle de régulateur de la machinerie de transcription a été mis en place. Ces données sont présentées dans la partie Résultats.

3. Les protéines aux doigts de zinc, des régulateurs transcriptionnels putatifs chez les archées

A part des régulateurs transcriptionnels RHH et HTH qui sont majoritaires, d'autres familles de régulateurs transcriptionnels, nettement plus minoritaires ont été identifiées par l'analyse *in silico* des génomes des archées et de leurs virus. Notamment, certains génomes présentent des gènes portant la signature de la famille des protéines aux doigts de Zinc, du type C2H2 (Figure 25). Ce type de protéines est presque exclusivement retrouvé chez les *Eukarya*; très fréquemment, elles sont associées aux régulateurs transcriptionnels (Figure 26A).

```

AFV1_YP_003731.1      13-34  CLRCGLTFRTRKKQLIRHLVN-TE
ATV_ORF189_N         117-138 CPYCVSRFPTVRALKIHLKR-RH
ATV_ORF189_C         145-165 CPVCYKPLKNKNAT-NEFRR-FI
SSV1_NP_039793.1     4-25   CLRCGGIFNKRREVVEHLLV-GH
SSV1_NP_039795.1N   77-99  CPLCLMPFSSSVSLKQHIRYTEH
SSV1_NP_039795.1C  103-124 CPVCKKEFTSTDSALDHVCK-KH
STIV_AAS89086.1     13-32  CN-CGFKTNDFIQPFHEFL--NH
STIV_AAS89085.1     10-29  CY-CGYQTYDYLLQLLNHYT--KH
SIRV1_NP_666608.1   23-43  CPICGEIYIKRKSMTHLR--KH
Human_PRD_AAK29075.1 470-490 CPVCGRPLANQGSRLRNEMR--LH
Human_ZNF_AAK72122.1 1266-1286 CKVCAKTFETEALNTHMR--TH

```

Figure 25 : Alignement multiple de putatifs régulateurs transcriptionnels avec des domaines de liaison à l'ADN en doigt de zinc (C2H2) partagés par des virus crénarchées. Les cystéines et histidines conservées sont encadrées en noir. Les alignements ont été construits à l'aide du programme MACAW (Prangishvili *et al.*, 2006).

La présence chez les archées de protéines avec des domaines prédits en doigts de zinc, de type C2H2 ou C2HC est surprenant car le motif de doigt de zinc « classique », nommé ZNF, est considéré comme étant le motif spécifique aux eucaryotes. Initialement découvert dans le facteur de transcription TFIIIA de *Xenopus oocytes* (J. Miller *et al.*, 1985), le domaine ZNF est très abondant chez les eucaryotes. Par exemple, 3% des gènes humains codent pour des protéines contenant le motif ZNF. Sauf quelques rares exceptions (N. Bouhouche *et al.*, 2000) souvent prédites *in silico* mais sans confirmation *in vitro* ou *in vivo*, les gènes portant le domaine ZNF sont pratiquement absents chez les bactéries. Chez les archées, ils sont identifiés essentiellement chez les virus d'archées.

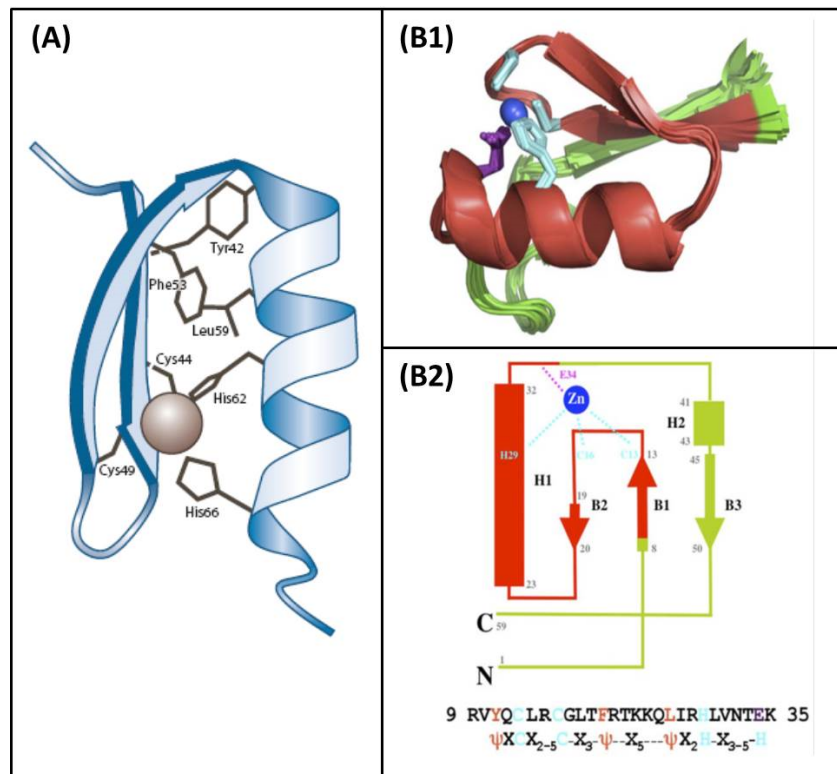


Figure 26 : Structures de protéines portant un domaine en doigt de zinc (C2H2). (A) Le facteur de transcription SWI5, issu de *Saccharomyces cerevisiae* est un doigt de zinc C2H2 typique. (B1) Topologie de la structure de la protéine AFV1p06. La séquence de AFV1p06 dans la région ZNF et la séquence motif de ZNF est montrée en (B2). Les résidus dans la région ZNF sont colorés en rouge, les chaînes latérales qui coordonnent le zinc sont en cyan ou violet et l'atome de zinc est en bleu. Les résidus hydrophobes sont représentés par un Ψ , les hélices par des rectangles et les feuilletts β par des flèches. (Neuhaus *et al.*, 1992; Guillière *et al.*, 2013).

Le motif classique ZNF consiste en une courte séquence (trente résidus) qui utilise deux ou trois cystéines et deux ou une histidine pour coordonner un ion zinc (types C2H2 et C2HC respectivement). Le domaine ZNF forme une structure caractéristique qui consiste en une hélice α et une tête d'épingle β tenues ensemble par l'ion zinc et les interactions hydrophobes entre les résidus hydrophobes à des positions conservées de la séquence (Gamsjaeger *et al.*, 2007; Brown 2005). La plupart des protéines étudiées contenant les domaines ZNF, sont impliquées dans la régulation de la transcription et elles lient l'ADN grâce à leurs domaines ZNF. L'ADN n'est pas leur seule cible. Les domaines ZNF de certaines protéines peuvent aussi participer à des interactions protéine-ARN ou protéine-protéine (Brown 2005; Gamsjaeger *et al.*, 2007) (Figure 26). Les ZNFs se lient à l'ADN en insérant l'hélice α dans le grand sillon de l'ADN et utilisent trois ou quatre résidus exposés de l'hélice pour réaliser des contacts spécifiques avec trois ou quatre bases ADN (Wolfe *et al.*, 2000; A. Klug, 2010). Pour reconnaître leur cible ADN dans la cellule, les ZNFs sont généralement présents en répétitions en tandem séparées par un court connecteur. Chaque répétition ZNF se lie spécifiquement à l'ADN en utilisant l'hélice α et les répétitions s'enroulent autour de l'ADN. Quelques ZNF, comme SW15, ADR1 ou GAGA, cependant, sont présents en seulement une à trois copies et utilisent des extensions du motif ZNF pour plus de contacts avec l'ADN (Omichinski *et al.*, 1997; Bowers *et al.*, 1999; Dutnall *et al.*, 2007).

Aucun régulateur de type ZNF n'a encore été identifié chez les *Archaea* et leurs virus, seulement des prédictions d'après les analyses *in silico* des génomes. Lors de ce travail de thèse, nous nous sommes focalisés sur l'étude de la protéine P06 codée par le génome du virus AFV1. Sa structure en solution a été décrite en collaboration avec le laboratoire Unité de RMN des Biomolécules (Inaki Guijaro, Muriel Delepierre) et ses capacités à lier l'ADN ont été analysées. Il s'agit de la première protéine archée avec un motif eucaryotique ZNF à être caractérisée expérimentalement. Les résultats sont présentés dans la partie Résultats II.5.

En conclusion de cette partie, les régulateurs transcriptionnels d'archées interagissent avec l'ADN selon un mode bactérien, en termes de motifs de liaison et d'emplacements relatifs aux promoteurs (proche ou chevauchant le cœur du promoteur). Le mode d'action des répresseurs transcriptionnels d'archées est basé sur de simples modèles d'encombrement stérique de la *TATA-box*, de la séquence BRE et du site d'initiation de la transcription. A l'inverse, les séquences opératrices localisées en amont de la *TATA-box* ou de la séquence BRE peuvent faciliter le recrutement de TBP et/ou TFB aux promoteurs (Bell *et al.*, 2000). En outre, l'altération de la conformation ADN, telle que la courbure de la région promotrice (Vassart *et al.*, 2013) ou l'utilisation de paires alternatives de facteurs généraux de transcription (TBP-TFB, comme évoqué dans la section précédente I.4.1.a) sont d'autres mécanismes possibles de régulation transcriptionnelle (Tenorio et Salgado, 2010). L'activité des régulateurs transcriptionnels peut être modulée fortement par des agents physico-chimiques, comme la liaison de petites molécules, les changements de température, les variations de pH, ou l'exposition aux UV (Bartolucci *et al.*, 2013).

La nature chimérique de l'appareil transcriptionnel soulève d'intéressantes questions concernant les mécanismes moléculaires de la régulation transcriptionnelle: comment des régulateurs de type bactérien peuvent opérer dans un contexte basal de type eucaryote ? Cela est particulièrement marquant pour les activateurs qui agissent par des contacts directs avec l'ARNP ou les facteurs généraux de transcription. Ainsi, la caractérisation structurale et fonctionnelle des régulateurs transcriptionnels après leur identification dans les génomes séquencés représente une approche magistrale pour étendre les connaissances sur la régulation transcriptionnelle des archées.

Une étude *in silico* récente réalisée par le groupe de Pérez-Rueda a analysé la distribution des régulateurs transcriptionnels identifiés chez les cinquante-deux génomes d'archées et a montré que 53% des régulateurs transcriptionnels ont au moins un homologue dans les génomes bactériens, 45% sont spécifiques aux archées, moins de 2% ont une homologie seulement avec les régulateurs eucaryotes et 6% est commun aux trois domaines (Pérez-Rueda and Janga, 2010). La même étude montre aussi que les régulateurs d'archées sont sous représentés numériquement comparé à ceux des génomes bactériens. Leur structure semble également être plus simple car les régulateurs d'archées ne contiennent assez souvent qu'un module de liaison à l'ADN sans domaines régulateurs additionnels. Une telle différence peut être compensée par l'emploi de stratégies de régulations transcriptionnelles encore inconnues (Pérez-Rueda and Janga, 2010).

L'approche génomique et l'analyse *in silico* restent les approches les plus puissantes pour l'identification des régulateurs transcriptionnels et de leurs gènes cibles (Song *et al.*, 2013; Nguyen-Duc *et al.*, 2013). Les données accumulées par l'approche *in silico* contribuent largement à comprendre la machinerie basale de transcription et son interaction avec les régulateurs généraux et spécifiques des *Sulfolobus*.

Les familles les plus abondantes de régulateurs transcriptionnels dans les génomes des *Sulfolobus* sont les familles Lrp et MarR, qui possèdent un domaine de liaison à l'ADN wHTH. L'abondance des régulateurs de ces familles pourrait être liée à leur polyvalence fonctionnelle et à leur stabilité dans les conditions extrêmes optimales pour les *Sulfolobus*. (Pérez-Rueda and Janga, 2010).

De façon intéressante, les virus et les plasmides de *Sulfolobus* codent en abondance pour des régulateurs transcriptionnels disposant de divers types de domaines de liaison à l'ADN appartenant en majorité aux familles RHH, HTH, *swapped hairpin* et doigt de zinc. De la même façon que la majorité des régulateurs transcriptionnels codés par les cellules d'archées qui ne possèdent pas de domaines additionnels de liaison à des effecteurs en plus du module de liaison à l'ADN, les régulateurs des virus et plasmides de *Sulfolobus* sont en général des protéines petites et compactes constituées seulement par un unique motif de liaison à l'ADN. A l'heure actuelle, la question comment ces petits régulateurs transcriptionnels peuvent interagir et/ou interférer avec les composants de la machinerie basale de transcription reste toujours une question ouverte.

Pour conclure, dans le genre *Sulfolobus*, la dissection exhaustive des réseaux complexes de régulation est fortement limitée par manque d'outils *ad hoc*, surtout des outils génétiques (Geiduschek *et al.*, 2005).

Malgré les efforts significatifs pour comprendre la logique *cis*-régulationnelle des régulateurs transcriptionnels d'archées, seulement un groupe mineur de tous les régulateurs d'archées existants a été caractérisé jusqu'ici. Donc, il ne serait pas surprenant si de futures investigations mettaient en évidence des stratégies et mécanismes de régulation variés, plus compliqués qu'initialement pensés. De plus, comme chez les organismes bactériens et eucaryotes, un réseau de régulation transcriptionnelle existe chez les archées, qui n'est pas encore compris. Dans le futur, les techniques à haut débit, telles que le CHIP-seq, le SELEX-seq, les puces de protéines de liaison (Stormo *et al.*, 2010; Sharon *et al.*, 2012) pourraient accélérer le gain de connaissances des spécificités de liaison à l'ADN des régulateurs d'archées et les règles de *cis*-régulation.

5. Génétique des *Sulfolobus*

Chez les archées, l'approche génétique est disponible pour certains représentants de deux des trois phyla, les *Euryarchaeota* et les *Crenarchaeota*. Cependant, même pour les souches modèles d'archées, comme par exemple *Methanococcus maripaludis*, *Haloferax volcanii*, *Pyrococcus furiosus* ou *Sulfolobus islandicus*, l'efficacité et la puissance des systèmes sont encore loin de ceux d'un grand nombre des eucaryotes et des bactéries.

Chez les *Euryarchaeota*, des systèmes génétiques ont été développés chez plusieurs d'espèces halophiles (*Halobacterium salinarum* NRC-1 (Ng *et al.*, 2000; Peck *et al.*, 2000), R1 (Pfeiffer *et al.*, 2008); *Haloferax volcanii* (Robinson *et al.*, 2005)), méthanogènes (*Methanococcus maripaludis* (Haydock *et al.* 2004; Hendrickson *et al.*, 2008), *Methanosarcina* (Welanders et Metcalf, 2008)) et des membres des *Thermococcales* (*Thermococcus kodakarensis* KU216, KW126 (Sato *et al.*, 2005), *Pyrococcus abyssi* (Lucas *et al.*, 2002), *Pyrococcus furiosus* (Waege *et al.*, 2010)).

Les archées halophiles ont été les premières à être manipulées génétiquement après la découverte de la technique de transformation (Cline and Doolittle, 1987; Charlebois *et al.*, 1987). Leur système génétique est probablement le plus sophistiqué de toutes les archées (Leigh *et al.*, 2011). Le succès de ces modèles halophiles s'explique en partie par la facilité et la bonne vitesse de leur croissance en laboratoire. Concernant les archées méthanogènes mésophiles, étant donné que la composition saline de leur cytoplasme est assez semblable à celle des bactéries, il a été possible de développer leur génétique en s'inspirant directement des techniques et des méthodes préexistantes chez les bactéries (comme par exemple, la sélection par la résistance aux antibiotiques en transformant *M. voltae* en exprimant un marqueur de résistance à la puromycine issu de *Streptomyces* (Gernhardt *et al.*, 1990)).

L'approche génétique appliquée aux euryarchées a permis d'identifier et de mieux caractériser les gènes impliqués dans la méthanogénèse et dans la fixation du carbone inorganique (Meuer *et al.*, 2002). Elle a été également très utile pour décrypter les mécanismes d'adaptation des protéines des archées halophiles aux fortes concentrations en potassium dans leur cytoplasme (Rose *et al.*, 2002; Hutcheon *et al.*, 2003).

Chez les *Crenarchaeota*, seules des espèces du genre *Sulfolobus* bénéficient de la disponibilité d'approches génétiques. Un grand effort a été réalisé pour développer des marqueurs de sélection applicables à ces organismes (pour rappel, par exemple la majorité des antibiotiques est sans effet sur les cellules d'archées faute de cibles cellulaires sensibles). Egalement, il a fallu tenir compte des conditions de croissance de ces espèces. Peu de composés sont stables à la température optimale de croissance des *Sulfolobus* qui est 80-90°C. Depuis de considérables progrès ont été réalisés dans le domaine de la génétique des *Sulfolobales* (She *et al.*, 2009). Des outils génétiques basiques et nécessaires, comme des techniques de *knock-out* de gènes et de vecteurs d'expression navettes ont été développés à un rythme accéléré.

La construction de mutants KO sur mesure chez les *Sulfolobus*, a permis, notamment, la découverte du rôle des gènes *cdv* (codant pour des protéines impliquées dans la formation des vésicules chez les eucaryotes) dans la division cellulaire (Samson *et al.*, 2008), l'identification de voies de réponse au stress oxydatif (Maaty *et al.*, 2009) et de caractériser plusieurs régulateurs transcriptionnels (Schelert *et al.*, 2006; Peeters *et al.*, 2009; Peng *et al.*, 2011; Zhang *et al.*, 2013). Des systèmes d'expression ont

également été mis en place (Albers *et al.*, 2004; 2006; Samson *et al.*, 2008; Wagner *et al.*, 2009; 2012) et ont permis, entre autres, de compléter des mutants de délétion d'opérons impliqués dans la formation des pilis (Zolghadr *et al.*, 2007; Frols *et al.*, 2008) ou de comprendre les interactions hôte-virus (Quax *et al.*, 2011).

La présente section donne une vue détaillée des systèmes génétiques développés chez certaines espèces *Sulfolobus*, en mettant l'accent sur les principes de sélection des transformants (résumés dans le Tableau 9) pour les souches modèles utilisées (Tableau 10) et les différents vecteurs de clonages disponibles (Tableau 11). Un certain nombre d'exemples expliquant les particularités des méthodes de disruption de gènes chez *Sulfolobus* seront également présentés.

1. Marqueurs de sélection

La majeure partie des systèmes de sélections connus et développés chez les bactéries est tout simplement inapplicable pour les archées. Notamment, les archées sont résistantes à la majorité des antibiotiques utilisés en routine pour la sélection chez les bactéries. Le fonctionnement des cellules d'archées, comme indiqué dans la section I.1, est proche à celui des cellules d'eucaryotes. Les cibles d'antibiotiques retrouvés chez les bactéries sont donc absentes (ex. peptidoglycane) ou insensibles (ex. ARNpol) aux antibiotiques habituels. Pour pallier ce problème, des marqueurs de sélections spécifiques ont été développés. Il s'agit essentiellement de marqueurs d'auxotrophie.

Pour résumer, trois marqueurs de sélection sont utilisés chez tous les modèles génétiques des *Sulfolobus*. Il s'agit d'un marqueur d'auxotrophie, *pyrEF*, un marqueur de sélection visuelle et d'auxotrophie, *lacS*, et un marqueur de résistance à la simvastatine, *hmg*. D'autres marqueurs comme le gène *adh* codant pour l'alcool déshydrogénase (Aravalli *et al.*, 1997) et le gène *hph* de résistance à l'hygromycine B ont été proposés par certaines équipes pour la sélection chez *S. solfataricus* (Cannio *et al.*, 1998; Cannio *et al.*, 2001). Cependant, d'autres laboratoires n'ont pas pu confirmer l'utilité et l'efficacité de tels marqueurs (Jonuscheit *et al.*, 2003; Berkner *et al.*, 2008) et ils ne seront pas discutés dans ce manuscrit.

a. Sélection uracile

Cette sélection repose sur le phénotype d'auxotrophie/prototrophie provoqué respectivement, par l'inactivation ou par l'intégration des allèles sauvages des gènes *pyrEF+* codant pour la biosynthèse de l'uracile.

Dans la voie métabolique de la biosynthèse des nucléotides, la synthèse d'uridine monophosphate (UMP) est l'étape clef de la synthèse des pyrimidines. L'UMP est synthétisée en six étapes décrites dans la Figure 27. Les étapes (a) et (b) sont respectivement catalysées par les enzymes orotate phosphoribosyltransferase (PyrE) et orotidine-5'-monophosphate decarboxylase (PyrF). L'enzyme PyrE catalyse la formation d'orotidine 5'-monophosphate à partir d'orotate et de phosphoribosylpyrophosphate (PRPP). L'orotidine 5'-monophosphate formé sera converti en UMP par l'activité de la protéine PyrF (Jund and Lacroute, 1970 ; Grogan and Gunsalus, 1993).

Les souches déficientes pour les gènes *pyrEF* sont incapables de synthétiser l'UMP et sont auxotrophes pour l'uracile. Elles sont donc capables de pousser uniquement dans un milieu complétement par de l'uracile exogène. Les cellules portant l'allèle sauvage *pyrEF+* peuvent être directement sélectionnées sur un milieu dépourvu d'uracile. Etant donné que les souches *pyrEF-* sont

résistantes au 5-FOA (acide 5'-fluoroorotique), la perte des gènes *pyrEF+* peut également être directement sélectionnée sur un milieu riche complété par 5-FOA. L'astuce de cette sélection est que des mutants *pyrEF-* peuvent être facilement et rapidement isolés en tant que colonies résistantes au 5-FOA (Kondo *et al.*, 2011, Grogan 1991). En effet, les enzymes PyrE et PyrF peuvent également convertir l'analogue non toxique, le 5-FOA (acide 5'-fluoroorotique), en un substrat toxique pour les cellules, le 5-FU (5-fluorouracil) (Figure 27). La molécule ainsi produite est un puissant inhibiteur qui bloque de façon irréversible l'enzyme thymidilate synthétase et provoque la mort de la cellule par manque de thymidine (Dolla *et al.* 2000).

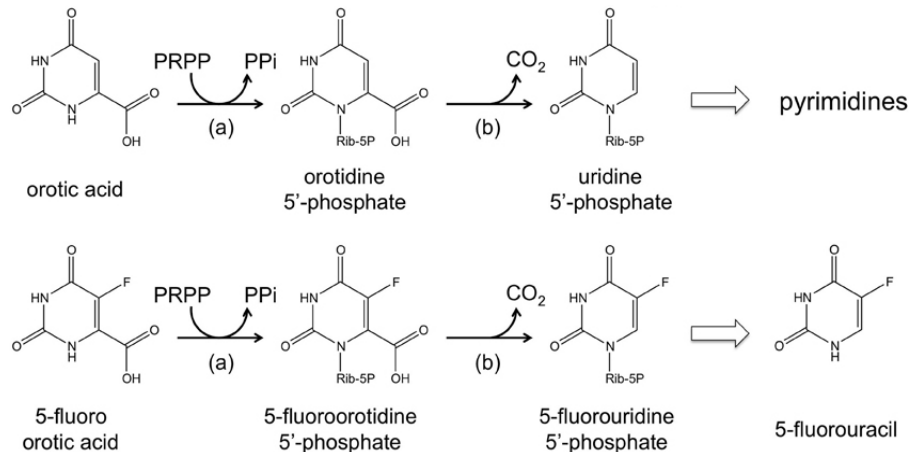


Figure 27 : Voie de la biosynthèse des pyrimidines et de la conversion du 5-FOA en 5-FU. Ces réactions sont catalysées par (a) l'orotate phosphoribosyltransférase (protéine PyrE) et (b) l'orotidine 5'-monophosphate décarboxylase (protéine PyrF) (d'après Leigh *et al.*, 2012).

Ce phénotype est très utile car la contre sélection permet de facilement retirer une cassette portant les gènes *pyrEF* pour obtenir un mutant de délétion propre ou pour recycler ce marqueur génétique. Des applications génétiques utilisant ce marqueur de sélection ont été mises au point chez les eucaryotes (Boeke *et al.* 1984; Fukagawa *et al.* 1999) et les archées (Bitan-Banin *et al.* 2003; Pritchett *et al.* 2004).

Des mutants spontanés stables *pyrEF-* ont obtenus et sont utilisés chez *S. solfataricus* (Jonuscheit *et al.* 2003), *S. acidocaldarius* (Wagner *et al.* 2009) et *S. islandicus* (Deng *et al.* 2009; Zhang and Whitaker 2012) (Tableau 10). Au cours de cette thèse, un mutant *pyrEF-* de *S. islandicus* LAL14/1 a été construit par l'approche de remplacement de gène (discuté dans la partie Résultats).

Tableau 9 : Les outils génétiques disponibles chez les *Sulfolobales* (adapté de Leigh *et al.*, 2011).

	<i>S. solfataricus</i> PBL2025	<i>S. islandicus</i> E233S	<i>S. acidocaldarius</i>
Milieu défini	Oui (Grogan, 1989)	Oui	Oui (Grogan, 1989)
Transformation de l'ADN	Electroporation (Schleper <i>et al.</i> 1992; Worthington <i>et al.</i> , 2003; Albers & Driessen, 2008)	Electroporation (Deng, <i>et al.</i> , 2009)	Electroporation (Kurosawa & Grogan 2005; Wagner <i>et al.</i> , 2009)
Barrière de restriction	Non	Suile coupe à GCwGC (Soilner <i>et al.</i> , 2006)	Sual restreint l'ADN non-méthylé à CCGG (Prangisshvili <i>et al.</i> , 1985)
Vecteur navette de réplication	pEXSs (Cannio <i>et al.</i> , 1998), pKMSD48 (Stedman <i>et al.</i> , 1999), PMSS dérivative (Aucelli <i>et al.</i> , 2006), pJlacS (Berkner <i>et al.</i> , 2007)	Vecteurs basés sur PRN2 (Deng <i>et al.</i> , 2009)	pCSV1 et pAG vecteurs de base (Aagaard <i>et al.</i> 1996; Aravalli & Garrett, 1997), pCmalLacS (Berkner <i>et al.</i> , 2010)
Sélection positive	Lactose (Worthington <i>et al.</i> , 2003), uracile (Jonuscheit <i>et al.</i> , 2003), hygromycine (Cannio <i>et al.</i> , 1998)	Uracile (Deng <i>et al.</i> , 2009)	Uracile (Wagner <i>et al.</i> , 2009), alcools (Aravalli & Garrett, 1997)
Contre sélection	Non	5-FOA (<i>pyrEF</i>) (Deng <i>et al.</i> , 2009)	5-FOA (<i>pyrEF</i>) (Wagner <i>et al.</i> , 2009)
"Markerless gene knock-out" ou remplacement allélique	Utilise <i>lacS</i> (Schelert <i>et al.</i> , 2004; Albert & Driessen, 2008)	Utilise <i>pyrE</i> (Deng <i>et al.</i> , 2009)	Utilise <i>pyrE</i> (Wagner <i>et al.</i> , 2009; Ellen <i>et al.</i> , 2010)
Intégration ectopique			A <i>pyrE</i> et <i>amyA</i> (M. Wagner & S.-V. Albers, unpublished data)
Gènes rapporteurs	<i>lacS</i> (β -galactosidase) (Jonuscheit <i>et al.</i> , 2003)	<i>lacS</i> (β -galactosidase) (Deng <i>et al.</i> , 2009)	<i>lacS</i> (β -galactosidase)
Expression du gène régulé	<i>araS</i> promoteur (Albers <i>et al.</i> , 2006)	<i>araS</i> promoteur (Peng <i>et al.</i> , 2009)	<i>malE</i> promoteur (Berkner <i>et al.</i> , 2010)
Surexpression de la protéine	Vecteurs virus pMJ0503 inductibles avec tags His et Strep, vecteurs basés sur pRN1 (Albers <i>et al.</i> , 2006; Berkner <i>et al.</i> , 2007)	Basés sur PRN2 pHZ2lacS (Peng <i>et al.</i> , 2009)	Expression inductible avec le maltose en utilisant pCmal (Berkner <i>et al.</i> 2010)

b. Sélection lactose

Cette sélection est basée sur l'utilisation du gène *lacS* codant pour une β -galactosidase spécifique aux archées (Cobucci-Ponzano *et al.*, 2010). Les mutants *lacS*⁻ sont incapables de croître sur lactose et de métaboliser les dérivés chromogènes comme le X-gal.

Toutes les espèces étudiées de *Sulfolobus* possèdent le gène *lacS*. La séquence nucléotidique du gène de la β -galactosidase ne montre aucune similarité avec le gène *lacZ* bactérien, ce qui suggère leur évolution convergente (Cubellis *et al.* 1990). La protéine LacS est une β -galactosidase très résistante à haute température (son activité optimale est à 90°C) et elle est résistante aux protéases et aux substances dénaturant habituellement les protéines (Moracci *et al.* 1992). Comme son analogue bactérien, cette enzyme catalyse l'hydrolyse des β -galactosides, parmi lesquels se trouvent le lactose et son analogue le X-Gal.

Les mutants *lacS*⁻ sont incapables de croître sur un milieu contenant du lactose comme seule source de carbone (milieu lactose). La sélection des recombinants intégrant une copie sauvage du gène *lacS*⁺ dans une souche au génotype *lacS*⁻ a été utilisée avec succès chez *S. solfataricus* pour réaliser des mutants *knock-out* et pour la sélection des transformants portant des vecteurs navettes (Worthington *et al.*, 2003; Schelert *et al.*, 2004; Berkner *et al.*, 2007). La sélection avec le marqueur *lacS* est applicable à tous les modèles génétiques de *Sulfolobus* (*S. solfataricus* et *S. islandicus*) capables de croître sur milieu lactose (Worthington *et al.*, 2003; Berkner and Lipps, 2008) sauf *S. acidocaldarius* dépourvue de cette capacité.

Le système de sélection avec *lacS* comporte des défauts. Ainsi, il y a un bruit de fond assez gênant provoqué par la croissance résiduelle des mutants *lacS*⁻ même sur un milieu minimum lactose. Un autre désavantage est le temps nécessaire pour obtenir la croissance des recombinants. Cette procédure prend en général jusqu'à quatorze jours en milieu liquide avant l'étalement (Albers and Driessen 2008).

Sur un milieu contenant du X-Gal, les colonies *lacS*⁺ développent un phénotype très caractéristique et facile à repérer car elles sont bleues (Figure 28) (Schelert *et al.* 2004; Deng *et al.* 2009).

Les mutants spontanés *lacS*⁻ s'obtiennent assez facilement. Ils ont été obtenus et utilisés chez *S. solfataricus* P2 (Jonuscheit *et al.* 2003), *S. islandicus* E233S (Deng *et al.* 2009) et *S. islandicus* M.16.4 (Zhang and Whitaker 2012).

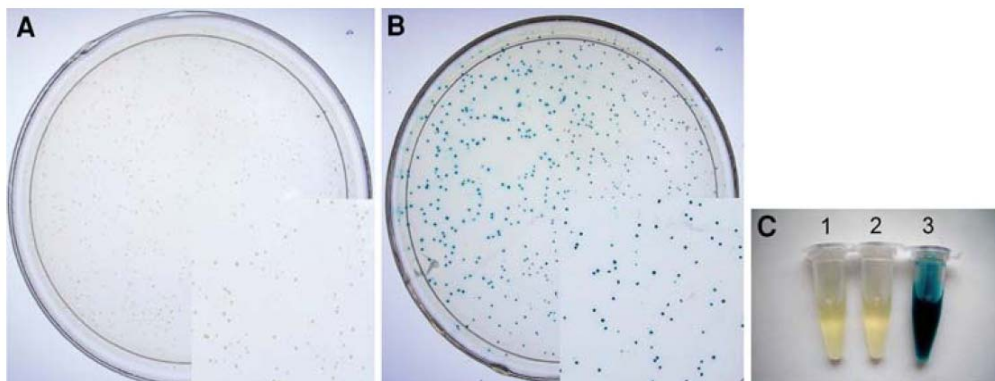


Figure 28 : Transformation du double mutant *S.islandicus* E233S Δ pyrEF Δ lacS par les plasmides pHZ2 et pHZ2lacS. (A) Les colonies E233S *lacS*⁻ transformées par le plasmide pHZ2 restent blanches en présence de X-Gal. (B) Les colonies E233S transformées par le plasmide pHZ2lacS devenues *lacS*⁺ se colorent en bleu en présence de X-Gal. (C) Les cultures liquides de *S. islandicus* E233S non transformées (C1) ou portant le plasmide pHZ2 (C2) ne se colorent pas en présence de X-Gal. A l'inverse, les cultures liquides de *S. islandicus* E233S transformées par le plasmide pHZ2lacS (C3) se colorent en présence de X-Gal. Adapté de Deng *et al.* 2009.

c. Sélection avec l'antibiotique simvastatine

La simvastatine est un des rares antibiotiques capable d'inhiber la croissance des *Sulfolobus* (Noll *et al.* 1997; Zheng *et al.*, 2012). Cet antibiotique inhibe l'action de la 3-hydroxy-3-méthylglutaryl coenzyme A (HMG-CoA) réductase par une compétition avec son substrat naturel, le HMG-CoA. Cette enzyme est impliquée et essentielle dans la voie de biosynthèse de la membrane des archées (Figure 29; (Matsumi *et al.* 2007)).

Utilisé dans un premier temps comme un marqueur de sélection chez deux euryarchées modèles, *T. kodakaraensis* (Matsumi *et al.* 2007) et *P. furiosus* (Waege *et al.* 2010), le gène *hmg* de résistance à la simvastatine a été ensuite employé avec succès chez *S. islandicus* (Zhang and Whitaker 2012; Zheng *et al.* 2012). La surexpression chez *S. islandicus* E233S du gène *hmg* codant pour la HMG-CoA réductase de *S. tokodaii* fusionné au promoteur constitutif fort *Psac7d* du gène *sac7d* de *S. acidocaldarius* (Berkner *et al.* 2010) permet en présence de simvastatine de sélectionner positivement les recombinants exprimant ce gène (Zhang and Whitaker 2012; Zheng *et al.* 2012). La sélection est assez difficile à cause de la croissance résiduelle de cellules non transformées. En outre, elle fonctionne essentiellement sur les milieux solides et est peu applicable aux milieux liquides (Zhang and Whitaker 2012).

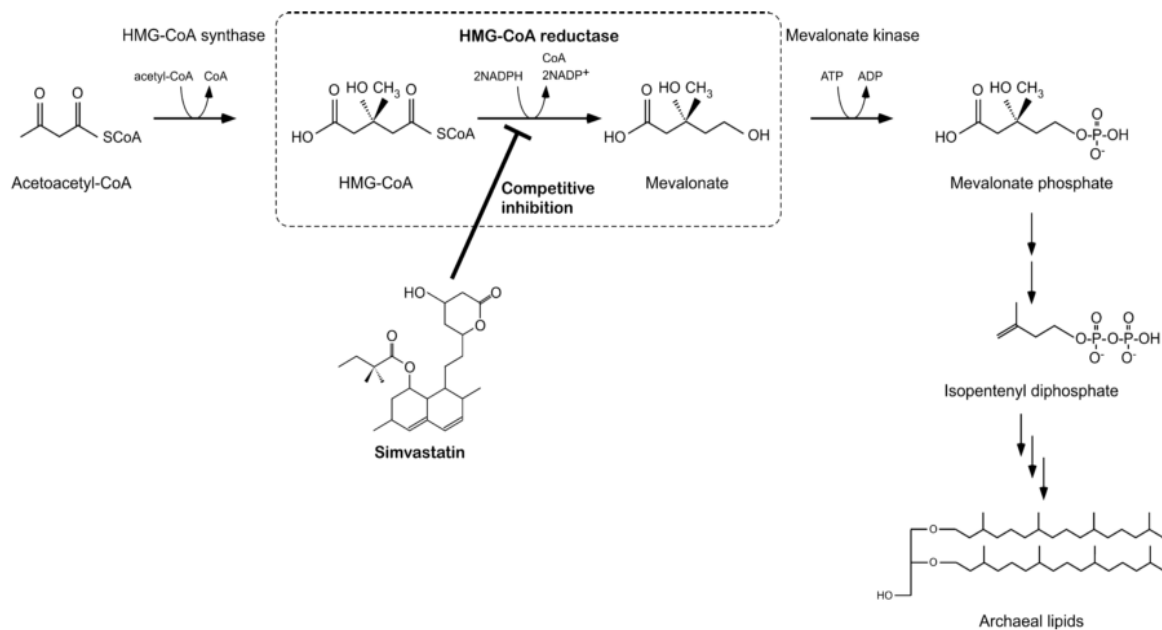


Figure 29 : Cible de la simvastatine dans la voie de biosynthèse de la membrane des archées. L'HMG-CoA est réduit par la HMG-CoA réductase en mévalonate, ensuite transformé après plusieurs étapes intermédiaires en lipides. La simvastatine agit en antagoniste de la HMG-CoA et inhibe sa réduction en mévalonate par la HMG-CoA. (Matsumi *et al.* 2007).

2. Modèles génétiques de *Sulfolobus* - souches réceptrices

a. *Sulfolobus solfataricus*

La première indication de la possibilité de transformer d'une souche de *Sulfolobus* a été décrite en 1999 dans un article montrant la transformabilité de *S. solfataricus* P1 et *S. solfataricus* 98/2 par de l'ADN non modifié du virus SSV1 isolé de souches *E. coli*. Les transformants étaient sélectionnés par leur capacité à induire le cycle viral après le traitement par les UV (Stedman *et al.*, 1999).

La génétique de *Sulfolobus* a commencé avec la souche *S. solfataricus* 98/2 lorsque Worthington *et al.* (2003) ont obtenu un mutant naturel *lacS*- de cette souche. Ce mutant Lac- spontané, **PBL2002**, est apparu à la suite de la transposition de l'*IS1217* dans le gène *lacS* qui code pour la β -galactosidase (Worthington *et al.* 2003). Cependant, son phénotype LacS- était trop instable (la souche génère des

révertants *lacS*⁺ avec une forte fréquence) rendant la sélection des recombinants Lac⁺ très difficile (Worthington *et al.* 2003).

Par la suite, un mutant spontané **PBL2025** portant une délétion de 58kb qui s'étend sur presque cinquante gènes (*Sso3004-Sso3050*), incluant le gène *lacS*, a été sélectionné et utilisé pour les expériences génétiques (Schelert *et al.* 2004) (Tableau 10). Néanmoins, compte tenu de l'étendu de la délétion, ce mutant présente un phénotype particulier. Par exemple, il diffère de la souche sauvage par plusieurs paramètres (taille des cellules, adhérence sur les matériaux) (Zolghadr *et al.* 2010) ce qui peut compromettre son utilisation pour des études générales (transcriptomique, protéomique, etc).

D'autres souches de *S. solfataricus* (*S. solfataricus* P1 et P2) ont également été utilisées pour développer des approches génétiques mais leurs génomes se sont avérés instables de telle façon que de grandes parties du chromosome pouvaient être délétées ou inversées en quelques générations (Redder et Garrett, 2006). Cette instabilité s'explique essentiellement par la transposition d'éléments mobiles fonctionnels et présents en nombre très élevé chez ces souches. Cette situation est fortement défavorable pour les expériences génétiques. Les multiples remaniements survenant à si forte fréquence influencent fortement le contexte génétique des mutants analysés (She *et al.* 2001; Brügger *et al.* 2002; Brügger *et al.* 2004, Schleper *et al.*, 1994; Martusewitsch *et al.*, 2000; Redder and Garrett, 2006) A présent, ces souches ne sont plus ou peu utilisées comme modèles génétiques.

Toutefois, historiquement la souche *S. solfataricus* P1 **PH1-16**, auxotrophe à l'uracile, a été la première souche à avoir été utilisée pour l'expression de protéines chez *Sulfolobus* (Martusewitsch *et al.*, 2000).

b. *Sulfolobus acidocaldarius*

A l'inverse des souches *S. solfataricus*, le génome de la souche *S. acidocaldarius* est très stable avec un taux remarquablement faible de remaniements chromosomiques (Berkner *et al.* 2008). De plus, contrairement aux souches de *S. solfataricus* (Schleper *et al.* 1994; Martusewitsch *et al.* 2000; Redder *et al.* 2006) et *S. islandicus* (Blount *et al.* 2005; Berkner and Lipps 2007; Jaubert *et al.*, 2013), son génome est le seul génome de *Sulfolobus* connu à ne pas posséder d'éléments mobiles actifs (Grogan *et al.* 2001; Chen *et al.* 2005). La disponibilité du génome de *S. acidocaldarius*, une des premières à être séquencée, a grandement facilité le développement rapide des techniques génétique pour cette souche (Chen *et al.* 2005).

En comparant les fréquences de mutations spontanées chez différentes souches de *Sulfolobus*, il a été démontré que la souche *S. acidocaldarius* est une des plus stables. La plus grande fréquence de mutations a été détectée chez *S. solfataricus* P1 (jusqu'à une fraction de 10^{-4} , Martusewitsch *et al.*, 2000) et ses dérivés. Au contraire, les fréquences nettement plus faibles sont retrouvées chez *S. acidocaldarius* et *S. islandicus* (entre 10^{-6} et 10^{-8} , Berkner and Lipps, 2008). En considérant la stabilité génétique, *S. acidocaldarius* et *S. islandicus* sont les souches les mieux adaptées pour les études génétiques.

Plusieurs mutants spontanés *pyrE*⁻ portant soit des mutations ponctuelles (Grogan *et al.* 2001) soit des délétions (Grogan *et al.* 2003) ont été sélectionnés chez *S. acidocaldarius* (Tableau 10). Ces

mutants ont par la suite été utilisés comme modèles pour la génétique car ils permettent la sélection des recombinants *pyrE+*.

Les souches **MR31** et **MW001** des mutants *S. acidocaldarius* auxotrophes pour l'uracile ont été utilisées pour la construction de mutants délétés. Plusieurs outils génétiques utilisant des événements de recombinaison simples ou doubles ont été développés pour obtenir chez *S. acidocaldarius* des mutants de délétion sans laisser le marqueur génétique, tagger des protéines et introduire de larges régions génomiques dans le génome (Wagner *et al.* 2012) (plus de détails dans les sections I.5.3.4 et 5).

Tableau 10 : Pertinence des différentes souches et espèces *Sulfolobus* comme destinataires pour des systèmes génétiques (adapté de Berkner and Lipps, 2008).

Mutants disponibles	<i>S. solfataricus</i>			<i>S. islandicus</i>			<i>S. acidocaldarius</i>
	P1	Gθ	98/2	REY15A	LAL14/1	M16.4	
Mutants <i>pyrEF</i> Mutants ponctuels/élément d'insertion Mutants délétés	PH1-16 ^{ab}			E233 ^f	LAL14/1-CD ^g	RJW001 ^h	MR31 ⁱ MR103 ^j MW001 ^j
Mutants <i>lacS</i>^a Mutants ponctuels/élément d'insertion Mutants délétés	PH1-16 ^{ab}	GθW ^c	PBL2002 ^d PBL2025 ^e	E233S ^f		RJW002 ^h	
Stabilité génétique	-	-	-	+	+	+	++
Pas de barrière de restriction	++	++	++	++	++	++	+

^a peuvent être des doubles mutants *lacS/pyrEF*

^b Martusewitsch *et al.*, 2000

^c Bartolucci *et al.*, 2003

^d Worthington *et al.*, 2003

^e Schelert *et al.*, 2004

^f Deng *et al.*, 2009

^g Jaubert *et al.*, 2013

^h Zhand et Whitaker, 2012

ⁱ Reuilly et Grogan 2001

^j Wagner *et al.*, 2009

++ : évaluation très positive; + : évaluation positive; - : évaluation négative

Grâce à ces mutants, il a été possible d'estimer l'efficacité de la recombinaison homologue entre l'ADN génomique et un fragment d'ADN linéaire selon la longueur de la région d'homologie. Les fragments minimaux suffisants pour permettre la recombinaison homologue restituant le génotype *pyrE+* ont une taille de seulement 18nt (Kurosawa *et al.* 2005). Des résultats proches sont observés chez d'autres souches. Ainsi, 14 et 45nt (selon l'étude) sont nécessaires pour assurer une recombinaison efficace chez *S. acidocaldarius* (Grogan *et al.* 2008; Sakofsky *et al.* 2011).

Récemment, la sélection uracile a été utilisée pour développer des vecteurs navettes plasmidiques chez *S. acidocaldarius* (Berkner *et al.*, 2007). Ce sont des vecteurs navettes *E. coli-Sulfolobus* construits à partir du plasmide multicopie pRN1 de *S. islandicus* (décrits plus loin dans le paragraphe I.5.3).

Un système d'expression induite par le maltose a été développé chez MR31 pour l'expression des gènes recombinants. De tous les promoteurs testés pour la production de protéines chez *S.*

acidocaldarius, le promoteur *saci1165* ou *mbp/mal*, inductible par le maltose s'est révélé le plus stable (Berkner *et al.*, 2010; Meyer *et al.*, 2011; Henche *et al.*, 2012; Lassak *et al.*, 2012).

Bon modèle pour les études génétiques, *S. acidocaldarius* présente toutefois, un certain nombre de limites à son utilisation. L'utilisation de *S. acidocaldarius* comme modèle génétique se heurte aux problèmes liés à la présence dans cette souche d'un système de restriction/modification spécifique. La souche encode pour une enzyme de restriction, *SuaI* clivant la séquence GGCC dans les ADNs non méthylés (Grogan 2003). Un passage par une souche *E. coli* génétiquement modifiée permettant la méthylation spécifique de l'ADN destiné à être introduit dans *S. acidocaldarius* est nécessaire avant toute tentative de transformation de la souche (Prangishvili *et al.* 1985; Grogan 2003). Le faible nombre de protéines assurant le transport des sucres et peptides encodés par ce génome, explique l'incapacité de *S. acidocaldarius* à utiliser efficacement certaines sources de carbone organique et, par conséquent, à croître rapidement sur un grand nombre de milieux (Elferink *et al.* 2001; Albers *et al.* 2004).

Pour conclure, bien qu'adaptée pour les études génétiques, *S. acidocaldarius* ne peut par exemple servir comme modèle pour une problématique très actuelle chez les archées concernant l'étude de l'interaction entre les cellules hôtes des *Sulfolobus* et leurs virus-plasmides conjugatifs. Ceci est lié au fait que *S. acidocaldarius* n'est un hôte pour aucun des éléments extra-chromosomiques connus. Au contraire, les souches *S. islandicus* REY15A et LAL14/1 sont des hôtes pour tous les fusellovirus testés (SSV1, SSV2, pSSVx et pSSVi) et les plasmides conjugatifs (pNOB8 et KEF9). Le développement d'une approche de génétique pour *S. islandicus* est donc une priorité importante dans la génétique des archées.

c. *Sulfolobus islandicus*

Un autre modèle de génétique très prometteur a été développé en utilisant certaines souches de *S. islandicus* (Deng *et al.* 2009). Ces souches présentent des fréquences de mutations spontanées proches de celles observées chez *S. acidocaldarius* (Berkner *et al.* 2008). Contrairement à la souche *S. acidocaldarius*, les souches de *S. islandicus* analysées ne semblent pas contenir de système de restriction/modification, à l'exception de deux souches : *S. islandicus* REN2H1 encodant pour l'enzyme *SuiI* étudiée *in vivo* (Sollner *et al.* 2006) et HVE10/4 portant un opéron de trois gènes codant pour une enzyme de restriction putative analysée *in silico* (SiH_1435-1437; (Guo *et al.* 2011)). Grâce à une forte prévalence de gènes codant pour les transporteurs de type ABC dans leurs génomes, les souches de *S. islandicus* peuvent, contrairement à *S. acidocaldarius*, croître sur une grande variété de milieux synthétiques (Grogan 1989; Guo *et al.* 2011). Le génome de la souche modèle *S. islandicus* **REY15A** est disponible (Guo *et al.* 2011) aussi bien que son mutant stable **E233S** *pyrEF-* et *lacS-* (Deng *et al.* 2009). L'inactivation par délétion des gènes *pyrEF* élimine tout risque de faux positifs dû à la réversion.

Ces avantages ont conduit au développement de l'approche génétique pour *S. islandicus* REY15A. Plusieurs outils et techniques sont maintenant disponibles pour cette souche comme un système rapporteur (Peng *et al.* 2009), différentes modifications de l'approche de *knock-out* (Deng *et al.* 2009; Zhang *et al.* 2010), des vecteurs de clonage plasmidiques permettant la surexpression de protéines (Peng *et al.* 2012) et la production de protéines recombinantes chez *Sulfolobus* (Lintner *et al.* 2011; Snyder *et al.* 2011; Mei *et al.* 2012; Zheng *et al.* 2012), etc. Ces approches ont par exemple permis de confirmer la prédiction *in silico* du rôle de gène *topA* codant pour une TopoIII fonctionnelle chez *S.*

islandicus (Li *et al.* 2011), l'implication de la protéine Cbp1 dans la transcription des CRISPRs de la souche REY15A (Deng *et al.* 2012), ainsi que le caractère essentiel des gènes *pcna* codant pour des protéines impliquées dans la réplication de l'ADN (Kelman 1997; Zhang *et al.* 2010).

Récemment, une nouvelle approche de *knock-out* basée sur la résistance à la simvastatine a été mise au point chez la souche *S. islandicus* **M.16.4**. En plus de démontrer l'utilisation possible du gène *hmg* dans les approches de *knock-out*, ces expériences ont permis la création des mutants *S. islandicus* **RJW001** et **RJW002**, respectivement $\Delta pyrEF$ et $\Delta lacS::sim^R$ pouvant profiter des approches génétiques précédemment citées (Zhang and Whitaker 2012) (Tableau 10).

Une autre souche développée récemment est la souche auxotrophe à l'uracile, *S. islandicus* **E233** qui a été utilisée pour obtenir un mutant $\Delta lacS$ (Deng *et al.*, 2009; She *et al.*, 2009). Un vecteur navette est disponible pour cette souche et est utilisé pour une étude détaillée du promoteur du gène *araS* (Peng *et al.*, 2009).

Les souches *S. islandicus*, de la même façon que les souches de *S. acidocaldarius*, ont le même avantage concernant la possibilité d'utiliser le 5-FOA pour une contre sélection des Ura⁺

En conclusion, plusieurs souches de *Sulfolobus* peuvent être utilisées dans les expériences génétiques. Sauf la souche *S. islandicus* REN2H1 (Söllner *et al.*, 2006), elles sont dépourvues de l'activité de restriction/modification très gênante pour les procédures de clonage et de transformation.

Pour terminer, au cours de mon travail doctoral j'ai travaillé sur le développement d'une approche génétique pour une souche particulière de *S. islandicus*, la souche LAL14/1. La particularité importante de cette souche est sa capacité à servir d'hôte pour plusieurs virus d'archées, notamment le virus SIRV1/2, un modèle sous étude dans notre laboratoire. Les résultats de ce travail (Jaubert *et al.*, 2013) sont discutés en détails dans la partie Résultats.

3. Vecteurs de clonage de *Sulfolobus*

Les virus infectant spécifiquement les espèces de *Sulfolobus* sont fréquents et bien caractérisés (voir Chapitre I.2) (Prangishvili *et al.*, 2001; Prangishvili 2013). Le même constat peut être fait pour les plasmides, conjugatifs ou non, présents dans leurs cellules (Lipps, 2006). La disponibilité de petits réplicons qui peuvent être manipulés *in vitro* crée une situation favorable pour le développement d'outils génétiques performants (Berkner and Lipps, 2008) et elle constitue un des avantages des *Sulfolobus* comme modèle génétique des archées.

Trois classes d'éléments extra-chromosomiques sont utilisées comme vecteurs : virus, plasmides conjugatifs et non conjugatifs (Zillig *et al.*, 1996; Prangishvili *et al.*, 1998 et Peng *et al.*, 2000). Dans cette partie, les exemples les plus utilisés pour les études génétiques illustrant les propriétés de chacun de ces trois groupes d'éléments génétiques sont présentés plus en détails (Tableau 11).

a. Virus SSV1

La disponibilité d'un protocole de transformation efficace était un des prérequis pour le développement de systèmes génétiques chez les archées. Dans le cas des souches *Sulfolobus*, le protocole de leur transformation a été établie en 1992 quand il a été démontré que la souche *S.*

solfataricus pouvait être efficacement transfectée avec l'ADN du virus SSV1 par électroporation (Schleper *et al.*, 1992). Depuis, ce protocole constitue un socle méthodologique permettant l'introduction des ADNs hétérologues chez les *Sulfolobus*. Avec des changements mineurs propres à chaque laboratoire, la technique est universellement utilisée pour la transformation des cellules *Sulfolobus*. Il a été avantageux de développer cette approche avec de l'ADN viral (transfection) permettant de sélectionner des transfectants par la recherche de simples plaques de lyse. Aucun autre marqueur de sélection n'est requis dans ce système. Également, l'éventuelle faible efficacité de transformation ne sera pas un obstacle étant donné que le virus peut s'auto-propager *in vivo* en infectant des milliers de cellules voisines à partir de la seule cellule transfectée.

Le virus SSV1, de la famille des *Fuselloviridae*, a été isolé à partir de la souche japonaise *S. shibatae* (Palm *et al.* 1991) mais il infecte également la souche modèle *S. solfataricus* P2 (Schleper *et al.* 1992). Son virion, en forme de citron, est enveloppé et porte sur une de ses extrémités une courte queue présentant trois fibres adhésives (Zillig *et al.* 1996). Son génome circulaire est représenté par une molécule d'ADN double brin de 15 kb. Le génome de SSV1 est composé de trente-quatre ORFs dont dix-huit codent pour des protéines non conservées chez les *Fuselloviridae* (Figure 30, ORF colorées du rouge au noir) (Tumbula *et al.* 1994; Wiedenheft *et al.* 2004).

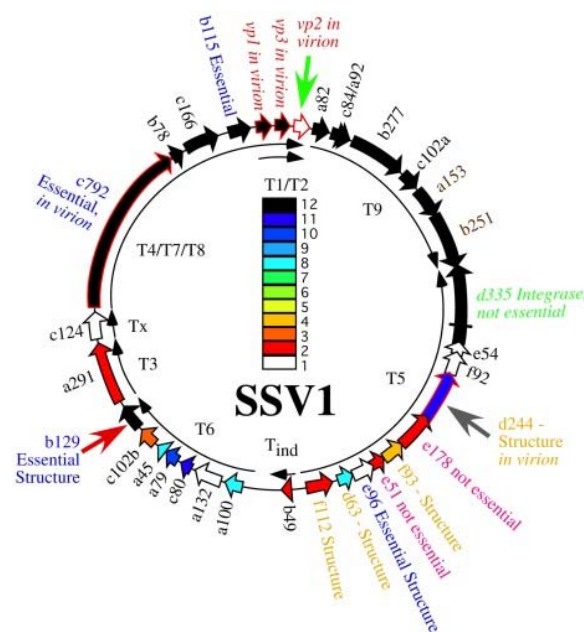


Figure 30 : Carte génétique du virus SSV1. Les gènes codant pour des protéines structurales ou retrouvées dans les virions sont symbolisés par une bordure rouge. La coloration des flèches indique la conservation de l'ORF dans les 12 génomes de fusellovirus analysés (SSV1, SSV2, SSV3, SSV4, SSV5, SSVRH, SSVK1, SSVL, SSVKM1, SSVKU1, SSVL2 et SSVG1) selon le diagramme indiqué au centre (du noir, ubiquitaire au blanc, spécifique à SSV1) (Iverson *et al.* 2012).

C'est un des virus des *Sulfolobus* les mieux étudiés et sa caractérisation moléculaire et génétique détaillée (Frols *et al.* 2007; Qureshi 2007; Menon *et al.* 2008; Iverson *et al.* 2012) a facilité la description et l'étude d'autres membres de cette même famille virale comme SSV2 et SSV-K1 (Wiedenheft *et al.* 2004) ou comme les quatre nouveaux fusellovirus récemment isolés (Redder *et al.* 2009) à partir de cultures de *S. solfataricus* P2 (SSV5), *Acidianus brierleyi* (ASV) ou de systèmes hôte-virus purs (*S. islandicus* G4T-1/SSV6 et *S. islandicus* G4T-T-11/SSV7).

SSV1 est un virus tempéré capable de se propager efficacement dans une cellule infectée ou transfectée. Il s'intègre dans le génome de son hôte de façon site spécifique au niveau du gène de l'ARNt-Arg. Son site d'attachement est composé de quarante-quatre nucléotides (*attA*) couvrant la partie 3' du gène de l'ARNt-Arg dont la séquence est identique au site viral d'attachement *attP* (Reiter *et al.* 1989; Muskhelishvili *et al.* 1993). La position du site d'attachement dans le gène de l'intégrase provoque son inactivation lors de l'intégration du génome viral. Le rôle de l'intégrase dans l'excision du provirus SSV1 est intrigante comme le site d'attachement viral, *attP* est localisé dans le gène de l'intégrase. Donc, l'intégration du virus sépare le gène *int* en deux parties de l'intégrase (Clare et Stedman, 2007). L'excision du provirus suivie par une forte production des particules virales peut être facilement induite par l'irradiation aux UV ou par le traitement des cellules lysogènes à la mitomycine C (Reiter *et al.* 1987; Frols *et al.* 2007). Même non induite, une souche lysogène produit spontanément une quantité non négligeable de particules virales ce qui indique un état relativement instable du provirus. L'excision est propre et elle restitue le site d'attachement chromosomique *attA* intact, sans aucune trace de provirus (site « vide ») (Muskhelishvili *et al.* 1993).

La construction de dérivés de SSV1 portant des mutations affectant certains de ses gènes a permis de localiser les gènes indispensables pour le cycle lytique. Ces données démontrent que parmi les protéines essentielles à l'établissement du cycle viral lytique, se trouvent l'intégrase (*d335*), la protéine structurale VP2 et la protéine codée par l'ORF *d244* (Iverson and Stedman, 2012). De façon surprenante, l'inactivation de l'ORF *d244* produit un dérivé de SSV1 avec une forte capacité d'inhiber la croissance des cellules de *Sulfolobus* infectées. Les raisons de ce phénotype restent inconnues, mais, de façon intéressante, des retards de croissance similaires sont observés lors de l'infection de *S. solfataricus* par le fusellovirus SSVK1 qui ne possède pas d'homologue du gène *d244* (Iverson and Stedman, 2012).

L'ORF ubiquitaire *b129* retrouvée chez tous les fusellovirus séquencés (Stedman *et al.* 1999; Iverson *et al.* 2012) code pour une protéine pouvant théoriquement se lier à l'ADN grâce à deux structures en doigt de zinc (Lawrence *et al.* 2009). Par ailleurs, la capacité de la protéine en doigt de zinc AFV1p06, codée par un autre virus, AFV1, à se lier à l'ADN de façon spécifique a récemment été confirmée pour la première fois chez les archées (Guilliere *et al.* 2013) et fait partie de mon travail de thèse. L'ORF *b129* est essentielle pour le cycle viral chez SSV1 et sa présence chez tous les fusellovirus indique que la protéine qu'elle code joue vraisemblablement un rôle important dans le cycle lytique des fusellovirus.

La possibilité de délivrer par une simple transfection le génome de SSV1 dans des cellules de *Sulfolobus* a ouvert la voie à l'étude génétique de ce virus et son utilisation comme un vecteur de clonage très efficace. Des vecteurs navettes *E. coli-Sulfolobus* pouvant se répliquer dans les deux hôtes ont donc été construits (Cannio *et al.*, 1998; 2001; Stedman *et al.* 1999; Jonuscheit *et al.* 2003; Albers *et al.* 2006; Clare and Stedman, 2006). Ces vecteurs présentent l'avantage de se propager par l'infection dans une culture initialement transfectée ce qui, au final, augmente grandement l'efficacité de la transfection. Les vecteurs construits sur la base de SSV1 présentent les avantages des vecteurs multicopies car ils se maintiennent à l'intérieur des cellules en nombre de copies relativement élevé sous forme d'ADN circulaire et fermé de façon covalente (Yeats *et al.*, 1982). A noter que SSV1 et beaucoup de ses dérivés se trouvent également sous forme d'une copie intégrée dans le génome. Le virus est donc maintenu de façon stable, mais les cellules lysogènes le contenant peuvent être difficilement « guéries » de SSV1 (Reiter and Palm, 1990).

En 1998, Cannio *et al.* décrivent un vecteur navette *Sulfolobus-E. coli*. Ce vecteur **pEXSs** est construit sur une partie du génome du virus SSV1 cloné dans le plasmide pGEM5Zf (-) et contient un gène marqueur hétérologue codant pour une version thermostable de la hygromycine phosphotransférase (Hph) de *E. coli* (Cannio *et al.*, 2001) (Tableau 11). Une version encore plus sophistiquée de ce vecteur portant le gène *lacS* (**pEXSlacS**) a été utilisée dans les expériences génétiques avec comme hôte cellulaire le mutant $\Delta lacS$ de *S. solfataricus* G θ W (Bartolucci *et al.*, 2003).

Tableau 11 : Liste chronologique de toutes les constructions basiques de vecteurs navettes *Sulfolobus-E. coli* (adapté de Berkner and Lipps, 2008).

Construction	Taille (kb)	Partie vecteur <i>E. coli</i>	Réplicon <i>Sulfolobus</i>	Auto-propagation	Marqueurs et phénotypes sélectifs	Destinataire	Références
pCSV1	6.1	pUC19	pGT5	Non	-	<i>S. acidocaldarius</i>	(1)
pAG1	4.7	pUC19	pGT5	Non	-	<i>S. acidocaldarius</i>	(2)
pAG2	5.3	pUC19 + <i>rom/op</i>			-		
pAG21	6.5	pUC19 + <i>rom/op</i>			<i>adh</i> ^a , <i>S.so.P2</i>		
pEXSs	6.4	pGEM5Zf(-)	SSV1	Non	<i>hph</i> ^b , <i>E. coli</i>	<i>S. solfataricus</i> G θ	(3)
pEXSlacS	8.4		(en partie)		Gènes <i>lacS</i> ^d , <i>S. so.MT-4</i>		(4)
pKMSD48	18.5	pBluescriptII SK+	SSV1 (complet)	Oui	-	<i>S. solfataricus</i> P1 <i>S. solfataricus</i> PH1 <i>S. solfataricus</i> P2 <i>S. solfataricus</i> G θ	(5)
pMJ03	21.8	pUC18	SSV1 (complet)	Oui	Gènes <i>pyrEF</i> ^c , <i>S. so.P1</i> Gènes <i>lacS</i> ^d , <i>S. so.P2</i>	<i>S. solfataricus</i> PH1-16	(6)
pMJ05							(7)
pAJC96	14.5	Topo PCR Blunt	SSV1 (sans intégrase)			<i>S. solfataricus</i> P2	(8)
pSSVrt	8.4	pUC19			-	<i>S. solfataricus</i> G θ W	(9)
pMSSV	7.6	pUC18 (en partie)	pSSVx/SSV2	Oui	-	<i>S. solfataricus</i> P2	
pMSSVlacS	9.6						
pA-pN	9	pBluescript (en partie)	pRN1	Non	Gènes <i>pyrEF</i> ^c , <i>S.so.P2</i> Gènes <i>lacS</i> ^d , <i>S. so.P2</i>	<i>S. acidocaldarius</i>	(10)
pJlacS	11					<i>S. solfataricus</i> PBL2025	
pHZ2	11.2	pGEM3Z	pRN2	Non	Gènes <i>pyrEF</i> ^c , <i>S. so.P2</i> Gènes <i>pyrEF</i> ^c et <i>lacS</i> ^d , <i>S. so.P2</i>	<i>S. islandicus</i> E223S	(11)
pHZ2lacS	13.2						

^a code pour une alcool déshydrogénase, sélection avec butanol ou benzylalcool

^b code pour une hygromycine phosphotransférase, une protéine thermostabilisée de *E. coli* qui confère la résistance à l'hygromycine

^c code pour une orotate phosphoribosyltransferase et orotidine-5'-monophosphatdécarboxylase, mutants complémentaires à la prototrophie d'uracile

^d code pour une β -glycosidase

(1) : Aagaard *et al.*, 1996; (2) : Aravalli and Garrett, 1997; (3) : Cannio *et al.*, 2001; (4) : Bartolucci *et al.*, 2006; (5) : Stedman *et al.*, 1999; (6) : Jonuscheit *et al.*, 2003; (7) : Albers *et al.*, 2006; (8) : Clore and Stedman, 2006; (9) : Aucelli *et al.*, 2006; (10) : Berkner *et al.*, 2007; (11) : Deng *et al.*, 2009.

Stedman *et al.*, en 1999, ont développé un des premiers vecteurs navettes *Sulfolobus-E. coli*, basé sur le génome complet du virus SSV1 et le vecteur pBluescript (**pKMSD48**). Le vecteur viral navette fonctionnel se propage efficacement à travers une culture infectée de *S. solfataricus*. Son excision et une forte réplication peuvent être provoquées par irradiation UV et ainsi permettre d'obtenir à la fin des particules virales en titre élevé. Plus tard, les gènes *pyrEF* de *S. solfataricus* P1 ont été ajoutés comme marqueurs sélectifs à cette construction donnant le vecteur navette **pSSV64** (Jonuscheit *et al.*, 2003). L'expression de ces gènes marqueurs permet la complémentation de destinataires auxotrophes à l'uracile (Tableau 10 et 11).

Le vecteur de cette génération le plus performant est **pMJ03** (Figure 31). C'est un dérivé du vecteur pSSV64 portant en plus le gène rapporteur *lacS* de *S. Solfataricus* P1 fusionné au promoteur de *tf55α* (Jonuscheit *et al.*, 2003). *Ptf55α* correspond au promoteur du gène *tf55α* qui code pour une chaperonine de choc thermique chez *S. solfataricus* et qui est induit dans les conditions de choc thermique. Pour cette construction, une expression forte et inductible par la chaleur du gène *lacS* a été démontrée. Il a été utilisé pour l'étude de la régulation de l'expression du promoteur arabinose (Lubelska *et al.* 2006) et de la réplication du plasmide pRN1 (Berkner and Lipps 2007), pour effectuer la surexpression de la protéine ABCE1 impliquée dans l'initiation de la traduction chez *S. solfataricus* (Barthelme *et al.* 2007). Dans les transformants obtenus, ce vecteur a été retrouvé intégré dans le génome de *S. solfataricus* PH1-16 au locus du gène ARNt-Arg (Martusewitsch *et al.*, 2000) exactement comme le virus sauvage (Schleper *et al.*, 1992). Le vecteur pMJ03 a été amélioré par Albers *et al.* (2006) créant une série de vecteurs d'expression qui permettent l'expression des gènes *in trans* chez *S. solfataricus*. Les souches recombinantes produisent des protéines correspondantes. Le système a été validé par la surproduction des protéines tagguées, aussi bien cytoplasmiques que membranaires. Les multiples modifications de ce vecteur sont résumées dans Tableau 9.

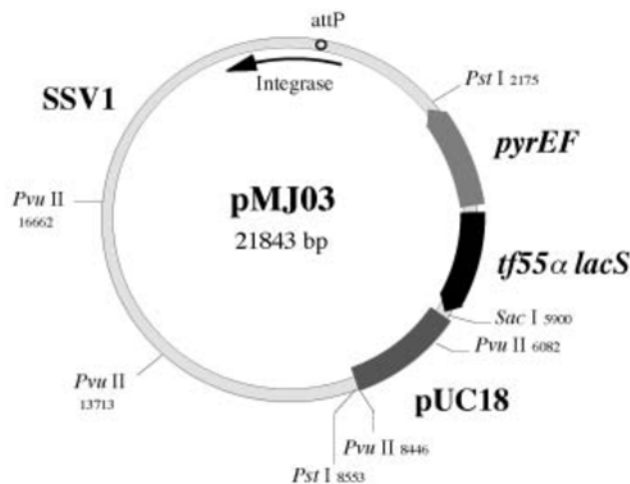


Figure 31 : Carte génétique du plasmide pMJ03. Le plasmide bactérien pUC18 et les marqueurs de sélection *pyrEF* et *lacS* placé sous le contrôle du promoteur *Ptf55α* sont indiqués par les flèches correspondantes (Jonuscheit *et al.*, 2003).

Par exemple, dans le plasmide **pSVA5**, le promoteur *tf55α* a été remplacé par le promoteur du gène *araS* de *S. solfataricus* qui code pour une sous-unité du transporteur de l'arabinose (Elferink *et al.*, 2001). Ce promoteur permet une induction forte en présence d'arabinose. Le promoteur *araS*

montre une activité basale plus faible que le promoteur $tf55\alpha$ en l'absence d'un inducteur. L'ajout d'arabinose dans le milieu permet à l'activité d'augmenter environ treize fois (Albers *et al.*, 2006).

b. Vecteurs plasmidiques

i. Vecteurs basés sur le plasmide conjugatif pNOB8

Les plasmides conjugatifs codent pour des fonctions leur permettant de franchir des barrières entre les deux cellules, donatrice et réceptrice. Ces plasmides sont très bien caractérisés et largement utilisés dans la génétique bactérienne, le facteur F d'*E. coli* en premier lieu (Lawley *et al.* 2003). Néanmoins, chez les archées, leur utilisation n'a pas été très répandue. Aussi, je présenterai les caractéristiques des plasmides conjugatifs à travers un seul exemple, pNOB8.

Le plasmide **pNOB8** a été le premier plasmide conjugatif des *Sulfolobus*, isolé chez *Sulfolobus japonicus*. Chez son hôte naturel, ce plasmide se maintient dans les cellules hôtes stablement avec un faible nombre de copies (~5 copies/génome). Son variant, **pNOB8-33**, obtenu après le passage de pNOB8 dans la souche *S. solfataricus* P1, a perdu les gènes codant pour les protéines ParA et ParB, homologues des protéines impliquées dans la formation du segrosome chez les plasmides dotés d'un système *par* de ségrégation (Gerdes *et al.* 2000; Hayes *et al.* 2006) (Figure 32). Les plasmides conjugatifs en général, et pNOB8 en particulier, encodent quarante à cinquante protéines dont dix sont conservées dans tous les plasmides conjugatifs d'archées (Greve *et al.* 2005; Erauso *et al.* 2006). Deux de ces gènes encodent des protéines homologues aux protéines bactériennes TraG et TrbE impliquées dans le transport de l'ADN double brin au travers des membranes bactériennes. Chez les plasmides conjugatifs de *Sulfolobus*, ces protéines pourraient être impliquées dans la formation d'un pore traversant les membranes d'archées (Lipps 2006). D'autres gènes codent pour des homologues des protéines CopG, Int (intégrase) et PrIA (ORF80, Figure 32) retrouvée également chez les plasmides non conjugatifs (discutés ci-dessous). Contrairement au virus SSV1, le site d'attachement reconnu par l'intégrase de pNOB8 est situé hors du gène, et par conséquent ce dernier n'est pas inactivé après l'intégration du plasmide dans le génome archéen (Lipps 2006).

Un dérivé de ce plasmide pNOB8 a été construit contenant une copie de *lacS* en aval d'un promoteur fort et a servi à transformer une souche *LacS*- de *S. solfataricus* (Elferink *et al.*, 1996).

Tout comme SSV1, pNOB8 se propage efficacement dans les cultures par transfert conjugatif. Le plasmide envahit la totalité des cellules à partir de quelques cellules initiales transformées par électroporation. Dans cette situation, une forte efficacité de transformation n'est pas requise. Ce plasmide se maintient de façon stable dans les cellules infectées mais sa présence provoque un fort retard de croissance chez son hôte (Schleper *et al.* 1995; Grogan 1996).

Les plasmides conjugatifs **pSOG1** et **pSOG2**, deux autres représentants de ce groupe, ont été isolés à partir de la même souche *S. islandicus* SOG2/4 (Erauso *et al.* 2006). Le premier (pSOG1) présente un profil d'expression et de stabilité comparable à celui de pNOB8 (Figure 32) qui peut s'expliquer par présence dans ce plasmide des gènes *parA* et *parB* (fort nombre de copies, ne se maintient pas dans la cellule). Le plasmide pSOG2 existe en un faible nombre de copies et affiche une forte stabilité structurale ainsi qu'une forte stabilité de maintien dans les cellules hôtes. L'absence d'homologues des gènes *parA* et *parB* sur le plasmide pSOG2, associé à son faible nombre de copies, suggère

l'existence d'un autre mécanisme de contrôle (Erauso *et al.* 2006). Malgré leur potentiel, aucun vecteur construit sur la base de ces plasmides n'est encore publié.

Pour conclure, le phénotype qui accompagne le processus de conjugaison (retard de croissance des cellules réceptrices; apparition d'une zone claire (plage) autour de la cellule donatrice) facilite l'identification des plasmides conjugatifs. Leur transfert dans les souches réceptrices est très efficace cependant leur utilisation en tant que vecteurs de clonage se complique par leur grande taille. Finalement, avec les progrès des techniques de transformation par électroporation, les plasmides non conjugatifs, décrits dans la section suivante, sont plus utilisés dans la génétique chez les archées que les plasmides conjugatifs.

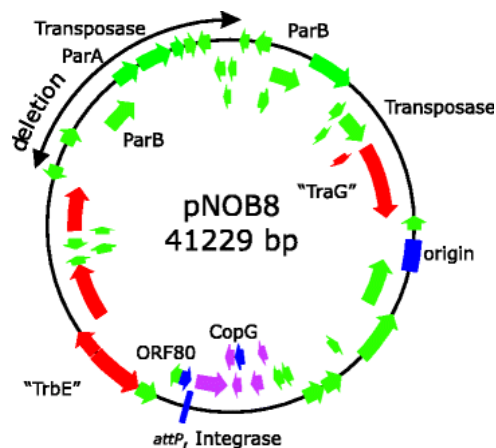


Figure 32 : Carte génétique du plasmide conjugatif pNOB8. En rouge sont indiqués les six gènes extrêmement conservés parmi les plasmides conjugatifs. En violet, sont représentés cinq gènes conservés, codant pour des protéines potentiellement impliquées dans la répllication du plasmide. Les ORFs des gènes *copG* et *ORF80/plrA*, aussi retrouvés dans les génomes des plasmides non-conjugatifs, sont colorés en bleu. Les gènes n'entrant dans aucune des catégories décrites ci-dessus sont colorés en vert. La large délétion provoquant la perte du plasmide pNOB8 est indiquée par une flèche noire (Lipps 2006).

ii. Plasmides non conjugatifs

Les plasmides non conjugatifs de *Sulfolobus* connus sont nombreux et largement utilisés pour le clonage et l'expression. Parmi les différentes familles de ces plasmides, celle de pRN est probablement la mieux étudiée (Peng *et al.*, 2000). Cette famille comprend plusieurs plasmides notamment pRN1, pRN2, pSSVx, PHEN7 et PLD10 (Keeling *et al.*, 1996; 1998; Arnold *et al.*, 1999). Le plasmide pRN1 a été le sujet d'études très détaillées et peut être considéré comme le plasmide le mieux étudié des archées. Ce plasmide a été isolé à partir de la souche *S. islandicus* REN1H1 (Zillig *et al.*, 1996) portant également le plasmide pRN2. Ces plasmides sont maintenus dans la même souche de façon stable et se répliquent de façon indépendante ce qui indique qu'ils sont compatibles entre eux (Zillig *et al.* 1996).

Le plasmide pRN1 a été utilisé comme point de départ pour construire des vecteurs navettes non conjugatifs pour les souches *Sulfolobus*. Ce plasmide se distingue par sa relativement petite taille (5.4kb), son nombre de copies entre dix et vingt par cellule en phase exponentielle et les trois gènes (*copG*, *plrA* et *repA*) sont conservés chez beaucoup d'autres plasmides *Sulfolobus* (Figure 33).

En ce qui concerne les fonctions codées par ces trois gènes, la protéine CopG serait un répresseur impliqué dans la régulation du nombre de copies du plasmide (Berkner *et al.*, 2007). PrlA est aussi une protéine de liaison séquence-spécifique à l'ADN, mais sa fonction reste inconnue. Malgré la forte conservation de sa séquence, elle n'est pas essentielle pour la réplication du plasmide mais son absence diminue la stabilité du plasmide et fragilise son maintien dans les cellules hôtes. La protéine PlrA pourrait être impliquée dans la distribution entre les cellules sœurs des copies fraîchement répliquées ou avoir un rôle accessoire dans la réplication du plasmide (Lipps 2009). Le gène *repA* essentiel pour la réplication des plasmides (Lipps *et al.* 2003; Lipps 2004), code pour la protéine RepA qui a les activités primase, polymérase et hélicase. Son activité primase est séquence-spécifique et les *primers* ne sont efficacement synthétisés qu'en présence du motif GTG. Cette protéine initie la réplication du plasmide en ouvrant la matrice double brin sur l'origine de réplication et en synthétisant les premiers *primers* dans la bulle de réplication (Lipps 2009). Cependant, ni le mécanisme de réplication, ni l'origine de réplication ne sont connus pour pRN1.

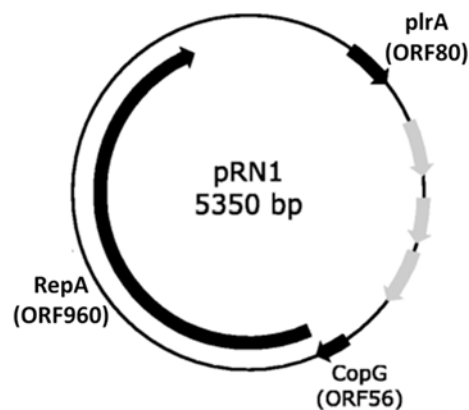


Figure 33 : Carte génétique du plasmide pRN1. Les flèches noires indiquent les trois ORFs ubiquitaires de la famille des plasmides pRN. Les flèches grises indiquent les trois ORFs spécifiques à pRN1 (Lipps 2006).

Le plasmide pSSVx appartient également à la famille des pRN. Il a été co-découvert avec le virus SSV2 chez *S. islandicus* REY15A et il présente un phénotype unique. Ce plasmide ne peut se diffuser dans la culture sans la coinfection avec les *fusellovirus* helper SSV1 ou SSV2 (Arnold *et al.* 1999; Stedman *et al.* 2003). Un vecteur navette construit à partir de cet élément génétique a donc les mêmes avantages au niveau de la sélection des transformants que celui basé sur le virus SSV1.

Les vecteurs navettes construits sur la base des plasmides cryptiques pRN1 (Berkner *et al.* 2007), pRN2 (Deng *et al.* 2009) ou pSSVx/SSV2 (Aucelli *et al.*, 2006) présentent plusieurs avantages absents des vecteurs basés sur les plasmides conjugatifs et les *fusellovirus*. Ces vecteurs sont non-intégratifs (par opposition aux vecteurs basés sur les *fusellovirus*) et ne modifient pas la vitesse de croissance des cellules hôtes (par opposition aux vecteurs basés sur les plasmides conjugatifs).

Voici quelques exemples des vecteurs construits à partir des plasmides décrits.

Les vecteurs **pC** (dérivé du plasmide pRN1) possèdent le gène fonctionnel *lacS*. En plus de son rôle de marqueur de sélection, le gène *lacS* a été utilisé dans ce vecteur comme rapporteur pour isoler des promoteurs inductibles chez *S. acidocaldarius*. Le promoteur du gène *mbp/mal* ou *saci_1165* (codant pour une protéine liant le maltose et le maltotriose) s'est avéré être fortement inductible par des

polymères du glucose (maltose, dextrine) (Berkner *et al.* 2010). C'est l'un des premiers exemples de plasmide d'expression chez les crénaarchées où un promoteur fort inductible utilisé pour la surexpression de gènes dans le but de surproduire la protéine codée. Il a permis, entre autres, de dévoiler le rôle de la protéine P98 dans la formation des structures pyramidales lors de l'infection du virus SIRV2 (Quax *et al.*, 2011).

Les plasmides d'expression **pHZ2** et **pHZ2lacS** construits sur la base du replicon pRN2 de *Sulfolobus* et du plasmide *E. coli* pGEM3Z (Tableau 11, Figure 34) assurent une expression stable des gènes clonés sur ces vecteurs chez *S. islandicus* E233S. Le plasmide pHZ2 est construit en clonant la cassette *pyrEF* de *S. solfataricus* P2 et le réplicon de pRN2 dans le plasmide *E. coli* pGEM3Z. Le plasmide pHZ2lacS porte également le gène rapporteur *lacS* issu de *S. solfataricus* P2 (Peng *et al.* 2009). Le promoteur constitutif *sac7d*, de *S. acidocaldarius*, est employé pour la surexpression du marqueur de sélection *hmg* (Berkner *et al.*, 2010) permettant la résistance à l'antibiotique simvastatine (voir la section « Simvastatine ») (Blaseio *et al.* 1990; Lam *et al.* 1992; Matsumi *et al.* 2007; Waege *et al.* 2010). Le remplacement de la cassette *pyrEF* dans le plasmide pHZ2 par la cassette *hmg* issu de *Sulfolobus tokodaii* sous contrôle de *sac7d* a donné une nouvelle série de plasmides, les plasmides navette **pSSR** (Zheng *et al.*, 2012). Enfin, la présence du promoteur *araS*, déjà utilisé dans les vecteurs navettes chez *S. solfataricus* (Jonuscheit *et al.* 2003; Albers *et al.* 2006) permet l'expression inductible du gène cloné en aval (Zheng *et al.*, 2012). L'équipe de Q. She a complété un mutant $\Delta herA$ en exprimant le gène *herA* (qui code pour une hélicase) sous contrôle du promoteur *araS* à l'aide de ce type de plasmide (Zheng *et al.*, 2012).

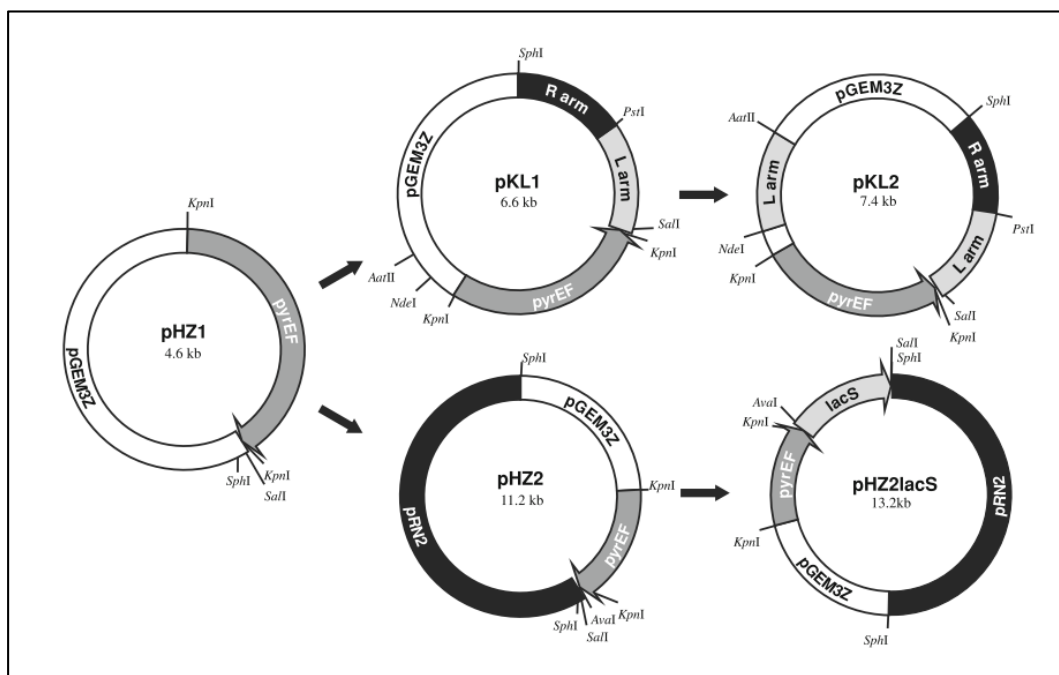


Figure 34 : Cartes génomiques de plasmides dérivés du plasmide pHZ1 de *Sulfolobus*. Les dérivés du plasmide pHZ1 (vecteur bactérien porteur de l'opéron *pyrEF*) s'utilisent pour obtenir des mutants KO (partie haute, plasmides pKL1 et pKL2) ou comme vecteurs d'expression (partie basse, pHZ2 et pHZ2lacS). Le réplicon permettant leur réplication chez les archées est celui pRN2. D'après Deng *et al.* 2009.

D'autres vecteurs navettes ont été basés sur l'hybride virus-plasmide pSSVx et le virus SSV2 (Arnold *et al.*, 1999). Le plasmide pMSSV est un vecteur de petite taille construit à partir de la fusion de pSSVx avec une partie du vecteur *E. coli* pUC18 (Aucelli *et al.*, 2006) (Tableau 11). Aucun marqueur sélectif n'est nécessaire pour une répllication stable chez *Sulfolobus*. Il se réplique et se propage sans réarrangement, recombinaison ou intégration dans le chromosome de l'hôte (Aucelli *et al.*, 2006). Il a néanmoins été montré que l'addition dans le vecteur pMSSV de la cassette du gène *lacS* du plasmide pMJ03 (Tableau 11 et paragraphe sur SSV1) était suffisant pour compléter le mutant délété pour *lacS* *S. solfataricus* GθW (plasmide **pMSSVlacS**) (Aucelli *et al.*, 2006). Dans cette cassette, le gène *lacS* de *S. solfataricus* est sous le contrôle du promoteur *tf55α*.

Il est à noter que les plasmides apparentés à la famille des pRN ont été retrouvés sous forme de séquences intégrées dans les génomes de *S. solfataricus* et *S. tokodaii* (She, Peng, *et al.* 2001; She *et al.* 2002; She *et al.* 2004). Leur intégration est probablement due à l'activité d'intégrases site-spécifiques.

Les plasmides de la famille pRN servent à présent à large échelle pour construire des vecteurs navettes *E. coli-Sulfolobus* et présentent de nombreux avantages, comme leur petite taille, leur nombre élevé de copies et la gamme de leurs hôtes.

4. etConstruction de mutants KO chez les *Sulfolobus*

Pour déterminer la fonction codée par un gène, la création d'un mutant KO et l'analyse de son phénotype est une voie d'analyse classique. Chez les *Sulfolobus*, les mutants KO peuvent être obtenus à l'aide de la recombinaison homologue entre l'allèle sauvage du gène chromosomique et l'allèle mutant construit sur mesure et amplifié, après transformation d'un vecteur « suicide » incapable de se répliquer chez les archées. En général, ce vecteur contient des régions homologues correspondant aux séquences en aval et en amont du gène cible permettant d'obtenir la recombinaison homologue dans une région choisie du chromosome de l'archée.

Les différentes techniques de *knock-out* seront données pour les trois espèces modèles *Sulfolobus* mais surtout détaillées pour l'espèce *S. islandicus*.

a. Génétique de *Sulfolobus solfataricus* et *Sulfolobus acidocaldarius* : construction de mutants KO

Pour les *Sulfolobus*, le premier système de disruption des gènes a été développé par P. Blum et ses collègues pour clarifier le rôle du gène de l' α -amylase (*amyA*) (Worthington *et al.*, 2003). Leur système utilisait une sélection Lac⁺ et un mutant avec un fond génétique *lacS::ISC1217* comme souche réceptrice (*S. solfataricus* PBL2002 dérivé de *S. solfataricus* 98/2, Worthington *et al.*, 2003). L'allèle permettant de muter la souche *amyA*⁺ a d'abord été construit *in vitro*. Il s'agit d'un fragment d'ADN portant le gène *lacS*⁺ de *S. solfataricus* P2 sous contrôle de son promoteur natif et entouré en aval et en amont de deux séquences de 700-800pb (Worthington *et al.*, 2003). Dans le chromosome de *S. solfataricus*, ces séquences se trouvent également en aval et en amont du gène cible, *amyA*. Le fragment d'ADN ainsi construit est intégré dans le vecteur pNEB193 et amplifié chez *E. coli*. La souche réceptrice *S. solfataricus lacS*⁻ est transformée par ce plasmide sous forme circulaire. Par un événement de double *crossing-over*, le gène *lacS* est intégré dans le chromosome hôte remplaçant le gène ou la région à enlever en restaurant le phénotype *lacS*⁺. La région est déterminée par le choix

des séquences de 700-800 pb placées en aval et en amont de *lacS*⁺. Après l'électroporation et la régénération, la sélection est d'abord réalisée dans un milieu minimum lactose liquide où seules les cellules *lacS*⁺ peuvent pousser. Ensuite, les cellules sont étalées sur des boîtes contenant le même milieu afin d'isoler des colonies mutants *amyA*::*lacS* (Worthington *et al.*, 2003) (Figure 35). Ces expériences ont mis en évidence que le gène *amyA* était essentiel pour la croissance sur glycogène, amidon et pullulane (polymères de glucose) comme seule source de carbone et d'énergie.

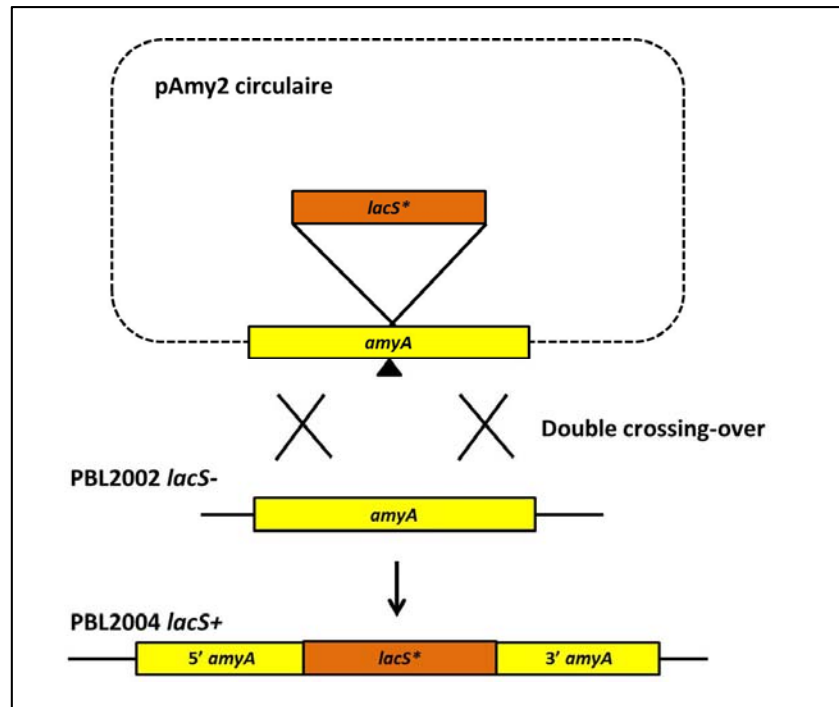


Figure 35 : Schéma de la stratégie de délétion pour le gène *amyA* chez *S. solfataricus*. La souche PBL2002 utilisée est *lacS*::*ISC1217*. L'événement de double *crossing-over* entre le plasmide pAmyA2 et le gène *amyA* conduisant à la délétion du gène *amyA* est représenté par deux croix. La sélection des recombinants *LacS*⁺ a lieu sur un milieu minimum lactose.

Plus tard, cette souche hôte a été remplacée par un mutant spontané délété pour le gène *lacS* (*S. solfataricus* PBL2025) utilisé comme souche réceptrice. Ce mutant porte une délétion de 58kb qui s'étale des ORFs SSO3004 à SSO3050. La souche PBL2025 a été utile pour la création de plusieurs mutants KO chez *S. solfataricus* notamment pour les gènes *merR* et *merA* impliqués dans la résistance au mercure (Schelert *et al.*, 2004). Les disruptions de gène (*marked gene disruption*) chez *S. solfataricus* PBL2025 ont été d'abord obtenues en utilisant des plasmides circulaires et par une approche visant un seul événement de *crossing-over*. Le marqueur de sélection utilisé était le gène *lacS*⁺ de *S. solfataricus* P2. Pour l'intégration d'ADN étranger dans le chromosome, un plasmide circulaire portant les régions flanquantes (en amont et en aval) du gène cible et le marqueur *lacS* sous le contrôle de son propre promoteur est transformé dans *S. solfataricus* PBL2025. L'intégration du plasmide ADN via un unique *crossing-over* conduit à une intégration en tandem (Wagner *et al.*, 2009). Les colonies positives sont d'abord sélectionnées sur un milieu minimum lactose et après étalement, les colonies bleues (*lacS*⁺) sont ensuite analysées (Figure 36). Wagner *et al.* ont optimisé ce protocole et des résultats plus stables ont été obtenus par la suite en utilisant des plasmides linéarisés ou des produits de PCR (Wagner *et al.*, 2009) (portant les régions flanquantes en amont et

en aval du gène cible et la cassette *lacS*). Dans ce cas, les recombinants apparaissent à la suite d'un événement de double *crossing-over* (Wagner *et al.*, 2009).

Grâce à cette approche, des mutants délétés pour les gènes codant pour l' α -amylase (Worthington *et al.*, 2003), pour des protéines membranaires (Kouril *et al.*, 2013), pour une protéine structurale de l'ATPase (Kouril *et al.*, 2013), pour une protéine pseudo-piline (Kouril *et al.*, 2013) ont été obtenus avec succès chez *S. solfataricus* et sélectionnés dans un milieu contenant du lactose comme seule source de carbone et d'énergie (Albers and Driessen 2007; Zolghadr *et al.*, 2007).

Cependant, de nombreux *rounds* de sélection sur milieu solide lactose minimum sont nécessaires pour obtenir les mutants recherchés. La technique produit des faux positifs avec une assez forte fréquence ce qui constitue une des principales limites de ce système (Zolghadr *et al.*, 2007; Albers *et al.* 2007).

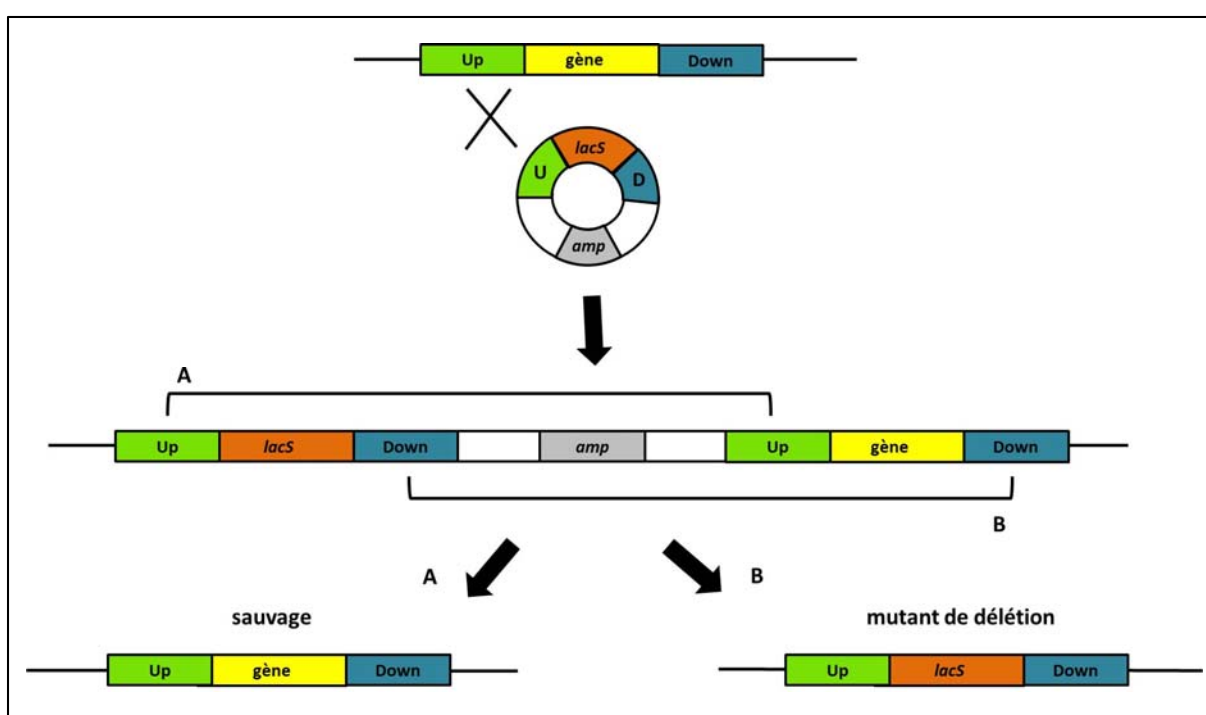


Figure 36 : Recombinaison du plasmide ADN via un unique événement de *crossing-over* avec le chromosome de *S. solfataricus* PBL2025. Le marqueur de sélection utilisé est le gène *lacS*⁺ de *S. solfataricus* P2. Pour l'intégration d'ADN étranger dans le chromosome, un plasmide circulaire portant les régions flanquantes (en amont et en aval) du gène cible et le marqueur *lacS* sous le contrôle de son propre promoteur est transformé dans *S. solfataricus* PBL2025 qui est $\Delta lacS$. L'intégration du plasmide ADN via un unique *crossing-over* conduit à une intégration en tandem. Le deuxième événement de recombinaison génèrent deux types de recombinants : (A) un recombinant sauvage, s'il a lieu entre les séquences homologues UP (B) un recombinant délété pour le gène cible, s'il a lieu entre les séquences homologues DOWN. Les mutants délétés sont d'abord sélectionnés sur un milieu minimum lactose et après étalement, les colonies bleues (*lacS*⁺) sont ensuite analysées. Adapté de Wagner *et al.* 2009.

Chez *S. acidocaldarius*, après les premières expériences démontrant la faisabilité de l'approche de mutagenèse par la recombinaison homologue (Grogan *et al.*, 2001), un système génétique réellement efficace n'a été établi que huit ans plus tard (Wagner *et al.*, 2009). Il est fondé sur

l'utilisation d'un mutant Ura⁻ qui contient une délétion de 18pb dans l'opéron *pyrEF*. Le plasmide pΔ2*pyrEF* utilisé pour réaliser les KO est un vecteur de suicide composé de la séquence du vecteur pBluescript d'*E. coli* et de la cassette *pyrEF*⁺ cloné dans ce vecteur. Le mutant Ura⁻ n'est pas un mutant de délétion complet de la séquence *pyrEF*. Pour éviter la recombinaison non désirée de deux copies de *pyrEF*, les auteurs de ce travail utilisent astucieusement la séquence *pyrEF*⁺ de *S. solfataricus* dont la séquence diffère assez significativement de *pyrEF* de *S. acidocaldarius* (Wagner *et al.*, 2009).

Ce vecteur a été utilisé pour construire des mutants délétés par simple et double événement de *crossing-over* (Figure 37) (Wagner *et al.*, 2009). Après la transformation, les cellules sont directement sélectionnées sur un milieu solide sans uracile ou après deux ou trois rounds dans un milieu liquide. De cette manière, deux mutants ont été construits : un mutant auxotrophe pour le tryptophane Δ*trpA* et un mutant Δ*Saci_1494* (Wagner *et al.*, 2009). Pour enlever le gène cible du chromosome, le plasmide doit être retiré par un second événement de recombinaison, conduisant soit au génotype sauvage, soit au génotype de délétion. Ce processus n'apparaît pas spontanément mais doit être provoqué par la contre-sélection en présence de 5-FOA. Le gène *Saci_1494* fait partie d'un opéron qui code pour des protéines de pili qui après induction par les UV sont responsables de l'agrégation des cellules (Ajon *et al.*, 2011). Il a été confirmé que les mutants Δ*Saci_1494* ne pouvaient en effet plus agréger après traitement aux UV.

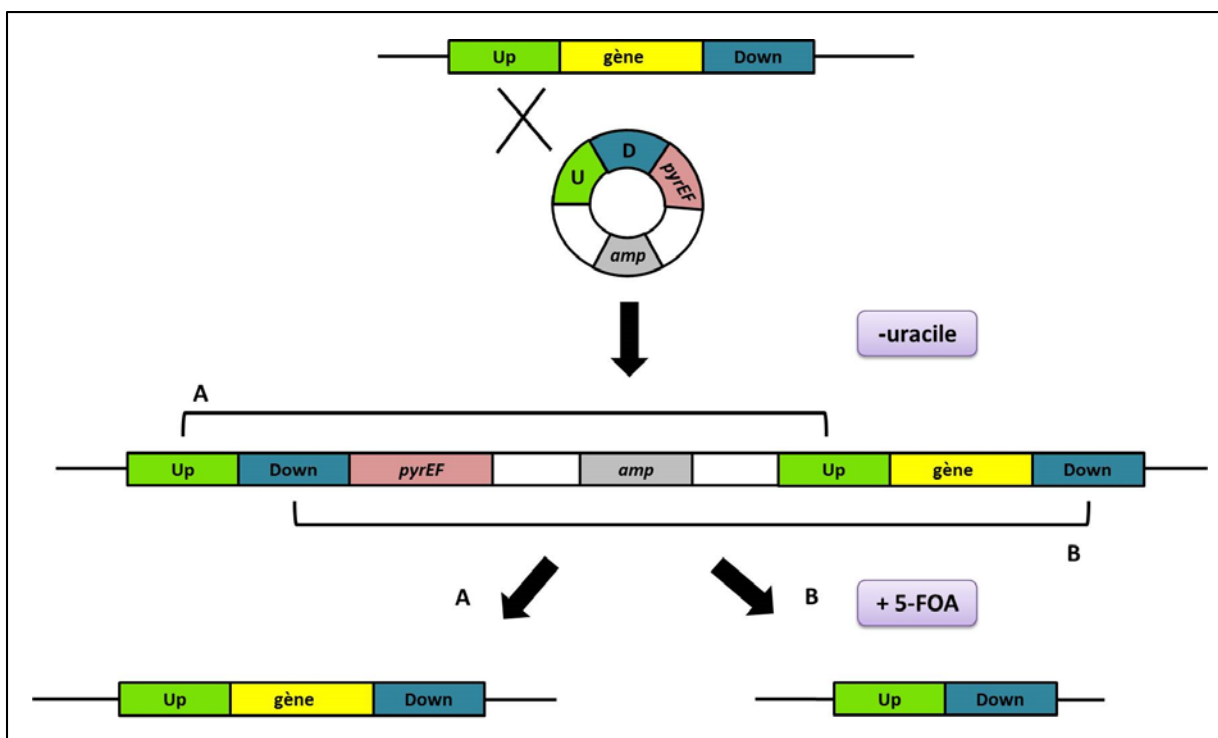


Figure 37 : Schéma de construction d'un mutant de délétion chez *S. acidocaldarius* utilisant le plasmide pΔ2*pyrEF*. Après intégration dans le chromosome par un unique *crossing-over*, une seconde recombinaison conduit soit au génotype *wild-type* (A) soit au génotype de délétion voulu (B). La cassette *pyrEF* est issue de *S. solfataricus*. La souche réceptrice utilisée est la souche *Sulfolobus acidocaldarius* MR31 *pyrEF*⁻. La sélection des cellules *pyrEF*⁺ se fait sur un milieu minimum sans uracile et la contre-sélection des cellules *pyrEF*⁻ se fait en présence de 5-FOA. Adapté de Wagner *et al.* 2009.

La souche utilisée, *S. acidocaldarius* MR31, auxotrophe pour l'uracile a été obtenue après mutagenèse sous UV (Grogan, 1991). Donc, il est possible que ces mutants Ura⁻ aient un fond génétique complexe et qu'ils portent des mutations supplémentaires. C'est pour cette raison qu'un nouveau mutant de délétion pour le gène *pyrE* a été construit chez *S. acidocaldarius* (Wagner *et al.*, 2012). Il s'agit du mutant MW001 qui contient une délétion de 322pb dans le gène *pyrE*. A partir de cette souche, trois méthodes de KO ont été développées. La première repose sur un schéma classique de *pop-in/pop-out* (Figure 38A) utilisant un simple événement de recombinaison. La deuxième est améliorée par l'ajout du gène *lacS* dans la cassette *pyrEF* qui permet de sélectionner visuellement (coloration blanc/bleu) les colonies qui ont inséré la cassette des marqueurs génétiques (Figure 38B). Enfin, la dernière donne des mutants par un double événement de *crossing-over* en utilisant une cassette linéaire (Figure 38C) et empêche de retrouver un génotype sauvage avec la contre-sélection avec du 5-FOA (Wagner *et al.*, 2012). Les techniques très similaires ont également été développées pour le modèle génétique *S. islandicus*. Pour éviter les redits, ces techniques seront décrites plus en détails dans la section dédiée à la génétique de *S. islandicus*.

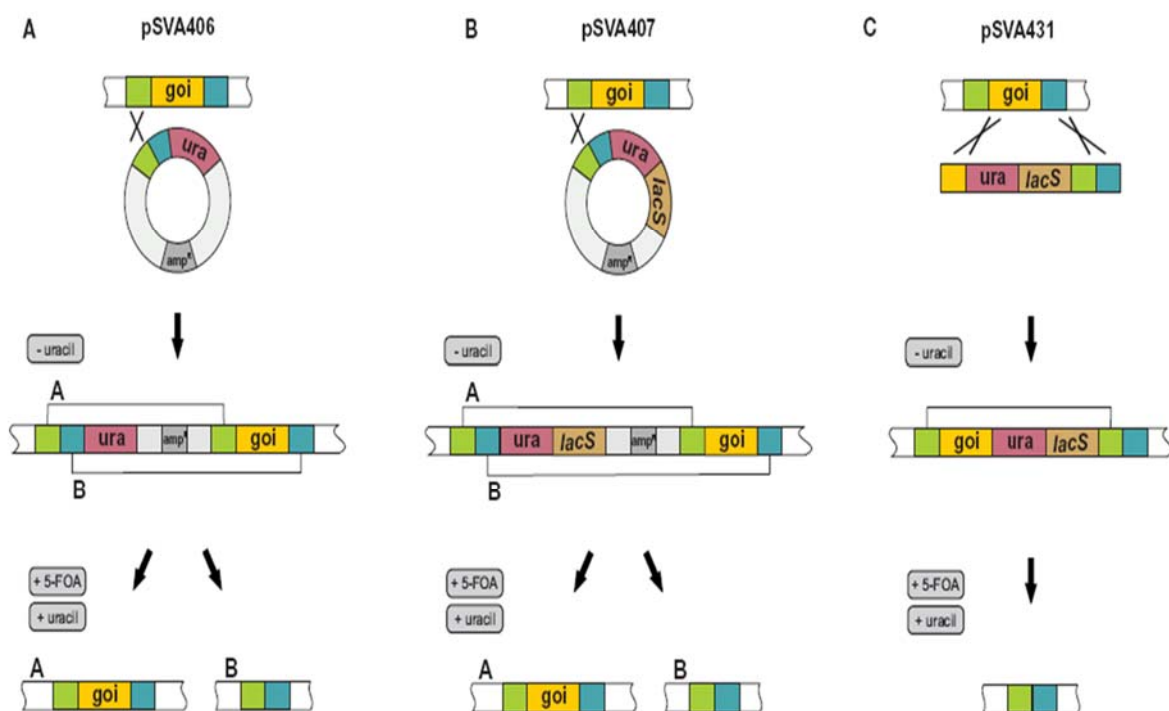


Figure 38 : Description de trois différentes méthodes pour obtenir des mutants de délétions chez *S. acidocaldarius*. (A) le plasmide pSVA406 s'intègre via un événement unique de *crossing-over* dans le génome et par ajout de 5-FOA, le second événement unique de *crossing-over* qui mène soit à un génotype sauvage, soit au génotype mutant souhaité peut être sélectionné; (B) en utilisant pSVA407 les mutants de délétions peuvent être obtenus via la même méthode que celle employée par pSVA406, seulement pSVA407 contient le gène *lacS* dans la cassette de sélection qui permet une identification facile des transformants au premier niveau de sélection; (C) pSVA431 permet une insertion par double *crossing-over* via la séquence du gène et la région en aval du gène cible. Donc les colonies positives doivent se colorer en bleu. Sous la sélection avec du 5-FOA, la cassette de sélection va être excisée et les mutants positifs de délétion doivent rester blancs. Goi désigne le gène cible (*gene of interest*), *ura* désigne la cassette *pyrEF* issue de *S. solfataricus* et *amp^r*, la cassette de résistance à l'ampicilline pour la sélection chez *E. coli*. Adapté de Wagner *et al.* 2012.

Les mutants délétés construits grâce à ces systèmes ont permis l'étude de gènes qui codent pour des protéines impliquées dans l'assemblage du pili (Ajon *et al.*, 2011; Henche *et al.*, 2012), ou du flagelle chez *S. acidocaldarius* (Lassak *et al.*, 2012) ou encore du gène *alg3* qui code pour une sulfoquinovose synthétase (Meyer *et al.*, 2011).

Le problème majeur du modèle *S. acidocaldarius* est la présence de l'endonucléase de restriction *SuaI* dans cette espèce. En conséquence, la préparation de tous les ADNs inclut un passage obligatoire dans une souche spécifique d'*E. coli* exprimant l'enzyme de méthylation correspondante.

b. Génétique de *Sulfolobus islandicus* : construction de mutants KO

Concernant un autre modèle génétique chez *Sulfolobus*, *S. islandicus*, l'équipe de Q. She à Copenhague a développé au cours de ces dernières années plusieurs approches très efficaces permettant la création des mutants KO pour cette espèce. L'originalité de leur approche est liée à l'utilisation d'ADN linearisé ce qui exige (au minimum) deux événements de *crossing-over* dans la cellule transformée. La liste des mutants obtenus chez *S. islandicus* E233S est présentée dans le Tableau 12 et détaillée ci-après. La section ci-dessous détaille les différentes modifications de la technique de construction de ces mutants KO.

Trois méthodes de *knock-out* conventionnelles ont été établies pour *S. islandicus*, incluant : (i) la méthode **AR** (*allelic replacement*) d'un gène cible avec les gènes marqueurs *pyrEF*, (ii) les méthodes **PIS** (*plasmid integration and segregation*), aussi appelée *pop-in/pop-out* et **MRL** (*marker replacement and looping out*). Ces deux dernières approches produisent des mutants avec une délétion des gènes cibles. Leur avantage est la perte au cours de la sélection du marqueur génétique qui a servi pour la sélection des transformants primaires. C'est un atout considérable car il permet de « recycler » le même marqueur plusieurs fois et de construire des souches portant de multiples KO par des cycles successifs de mutagénèses ciblées. En appliquant ces méthodes, des mutants *knock-out* de plus d'une dizaine de gènes ont été construits. Dans plusieurs des cas, la construction des mutants KO a échoué. Malheureusement, les approches PIS et MRL ne permettent pas de distinguer si l'échec est lié au fait que le gène cible est un gène essentiel ou il s'agit d'un simple problème technique. En génétique, les résultats négatifs sont toujours difficiles à interpréter ! Le cas des gènes *pcna*, pour lesquels la construction de mutants *knock-out* par les approches PIS et MRL a échoué, peut servir d'exemple (Zhang *et al.*, 2010). Une nouvelle méthode de *knock-out* a alors été développée : (iii) la méthode **MID** (*marker insertion and target gene deletion*), qui est une version améliorée des méthodes PIS et MRL. Cette méthode a permis de démontrer le caractère essentiel des gènes *pcna* (Zhang *et al.*, 2010).

i. Remplacement Allélique (AR)

La méthode de remplacement allélique consiste à remplacer tout ou partie d'un gène ou d'une région cible par une cassette contenant les gènes *pyrEF* et/ou *lacS* (Figure 39 et 40, module I). Au cours de mon travail doctoral, nous avons appliqué cette approche pour construire un mutant $\Delta pyrEF$ (voir la partie Résultats). Zhang, Dalgaard *et al.* ont également utilisé avec succès cette approche pour inactiver les quatre gènes impliqués dans un mécanisme de réparation de l'ADN par BER (*Base excision Repair*) (Zhang *et al.*, 2013). Seul un de ces gènes (*udg2*) a pu être enlevé par l'approche AR (Tableau 12). Les autres sont probablement des gènes essentiels ou l'approche n'a pas été suffisamment efficace pour permettre l'obtention des mutants recherchés.

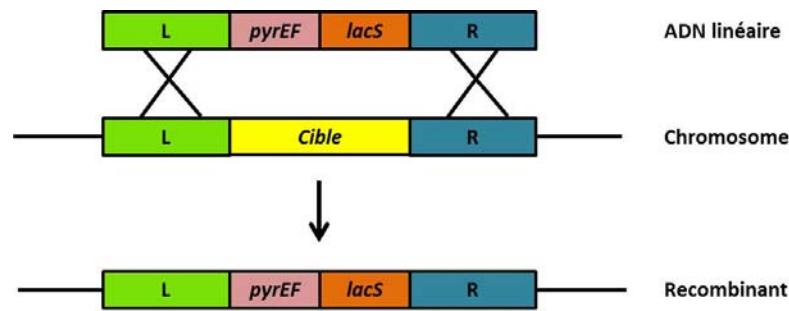


Figure 39 : Mécanisme de recombinaison mis en œuvre par l'approche de remplacement allélique. Les régions L et R recombinent indépendamment pour intégrer la cassette à la place de la région ciblée. Adapté de Zhang *et al.* 2013.

Le principal défaut de cette approche réside dans l'impossibilité d'enlever les marqueurs de sélection une fois substitués au gène cible. Compte tenu du faible choix de marqueurs de sélection chez les archées, ceci empêche la construction de mutants doubles ou multiples et les expériences de complémentation pour confirmer le rôle du gène inactivé dans les phénotypes observés chez le mutant KO. L'introduction récente dans la génétique des archées d'un nouveau marqueur de sélection, le gène *hmg* (présenté dans la partie I.5.1.c, Zhang and Whitaker 2012; Zheng *et al.* 2012), donne désormais la possibilité de construire des doubles mutants sans nécessairement être obligé d'enlever le premier marqueur .

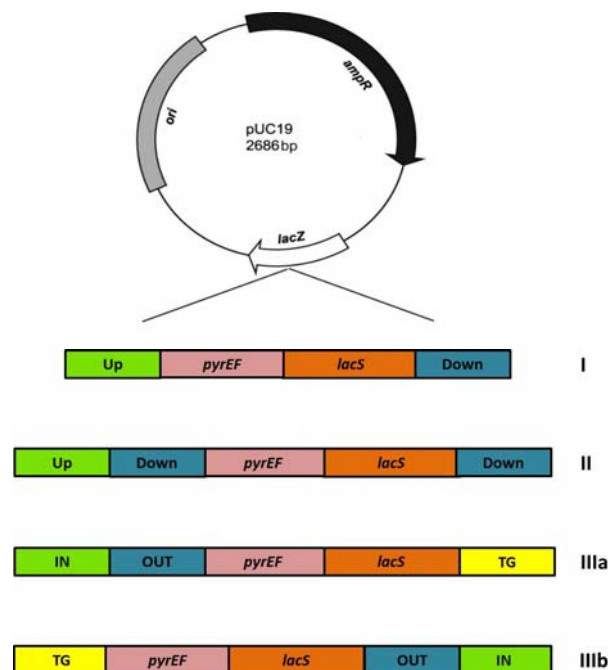


Figure 40 : Modules de recombinaison de chaque plasmide *knock-out*. Trois séries de plasmides ont été construits. Les plasmides pAR et pMRL contiennent les module de recombinaison I et II, respectivement. Les plasmides pMID contiennent soit le module de recombinaison IIIa ou IIIb, ce qui conduit l'insertion de la cassette marqueurs en amont ou en aval du gène cible. Les rectangles de même couleur montre des séquences homologues : le gène cible est jaune, alors que les séquences en amont et en aval sont en vert et bleu, respectivement (adapté de Zhang *et al.*, 2010).

ii. Remplacement du marqueur et recircularisation (MRL) et intégration du plasmide et ségrégation (PIS)

Il faut savoir que de nombreux gènes sont présents en multiples copies dans les génomes des *Sulfolobus*. Par exemple, il y a quatre gènes *udg* (codent pour une uracile-ADN-glycolase) dans tous les génomes de *Sulfolobus* séquencés (Dionne *et al.*, 2005). Donc, pour étudier le rôle de ce type de paralogues, il est indispensable de pouvoir créer des mutants multiples. A présent, le seul moyen de le faire chez les *Sulfolobus* est de réutiliser le même marqueur et de conduire plusieurs *rounds* de délétions.

Ainsi, pour réutiliser le marqueur génétique *pyrEF* chez *S. islandicus*, ce marqueur doit être enlevé du chromosome mutant à la fin des manipulations génétiques. Ce protocole est aussi appelé « *markerless gene knockout* ».

Deux méthodes « *markerless gene knockout* » ont été développées : les approches MRL et PIS. Ces approches permettent de restaurer le génotype *pyrEF* et ainsi ouvrent la voie pour recycler le marqueur ce que ne permet pas l'approche AR. Ces approches ont démontré leur potentiel dans plusieurs études, notamment pour l'inactivation du gène non-essentiel *phrB* et l'étude de l'essentialité du gène *topR1* (Liu *et al.*, 2009). Ces gènes codent respectivement pour une protéine impliquée la réparation de l'ADN suite à une exposition aux UV (PhrB) (Zhang *et al.*, non publié) et une reverse gyrase (TopR1) (Liu *et al.*, 2009; Garnier *et al.*, 2008) (Tableau 12).

Tableau 12 : Approches *knock-out* pour l'étude des gènes de *S. islandicus* REY15A impliqués dans la réparation et la réplication de l'ADN. AR : Remplacement allélique (*Allelic Replacement*); MID : Insertion de marqueur et délétion du gène cible (*Marker Insertion and unmarked target gene Deletion*); MRL : Remplacement du marqueur et recircularisation (*marker replacement and looping out*); PIS : Recombinaison *Pop-in/Pop-Out* (*plasmid integration and segregation*). Adapté de (Zhang *et al.* 2013).

Gène ID	Protéine	Méthode de KO	Phénotype	Sources
SiRe_1280	XPF	MID	Réplication de l'ADN altérée, redémarrage des fourches de réplication	(Zhang <i>et al.</i> 2010; 2013)
SiRe_1685	XPD	MID	Non essentielle, pas de phénotype apparent	C. Zhang, non publié
SiRe_1128	XPB1	MID	Non essentielle, pas de phénotype apparent	(Peng <i>et al.</i> 2012)
SiRe_1526	XPB2	MID	Non essentielle, pas de phénotype apparent	(Peng <i>et al.</i> 2012)
SiRe_1524	Bax1	MID	Non essentielle, pas de phénotype apparent	C. Zhang, non publié
SiRe_0261	PhrB	PIS	Non essentielle, pas de phénotype apparent	C. Zhang, non publié
SiRe_0235	Dpo4	AR	Croissance normale, faible taux de mutation spontanée	S. Gökce, non publié
SiRe_1747	RadA	MID	Essentielle	(Zhang <i>et al.</i> , 2013)
SiRe_1224	RadC	MID	Non essentielle, pas de phénotype apparent	B. Tian, non publié
SiRe_2356	UDG1	AR	Sans-succès	K. Dalgaard, non publié
SiRe_0084	UDG2	AR	Faible vitesse de croissance en phase exponentielle précoce	K. Dalgaard, non publié
SiRe_0278	UDG3	AR	Sans-succès	K. Dalgaard, non publié
SiRe_1884	UDG4	AR/MID	Sans succès avec AR; Retard de croissance observé avec MID	K. Dalgaard, non publié
SiRe_1830	XPG/Fen1	MID	Essentielle	(Zhang <i>et al.</i> , 2013)
SiRe_0061	NurA	MID	Essentielle	(Zhang <i>et al.</i> , 2013)
SiRe_0062	Rad50	MID	Essentielle	(Zhang <i>et al.</i> , 2013)
SiRe_0063	Mre11	MID	Essentielle	(Zhang <i>et al.</i> , 2013)
SiRe_0064	HerA	MID	Essentielle	(Zheng <i>et al.</i> 2012)
SiRe_0250	Hjm	MID	Essentielle	(Hong <i>et al.</i> 2012)
SiRe_1431	Hjc	MID	Non essentielle	(Hong <i>et al.</i> 2012)
SiRe_1124	TopR1	AR/MRL	Essentielle	(Liu <i>et al.</i> 2009)
SiRe_1581	TopR2	MID	Essentielle	X. Ao, non publié
SiRe_1176	TopIII (TopA)	AR	Trouble de croissance	(Li <i>et al.</i> 2011)
SiRe_1048				
SiRe_1594	Pcna	MID	Essentielle	Zhang <i>et al.</i> , 2010
SiRe_1048				

L'approche MRL est une évolution de l'approche AR. Comme dans cette dernière, le plasmide linéarisé (ici illustrée avec le vecteur pKL2, Figure 40, module II et Figure 41A) s'intègre dans le génome par un double événement de recombinaison remplaçant le gène cible par un marqueur de sélection (Figure 41A). L'ajout d'une seconde copie de la région Up (ou Down, comme le montre la figure 40, module II) permet, en recombinant avec la première région Up (ou Down) via une structure appelée *looping-out*, la délétion du marqueur *pyrEF+*. Cet événement de recombinaison restaure le phénotype *PyrEF-*, ouvrant la possibilité de réutiliser le même marqueur génétique pour obtenir, par la même approche, d'autres mutations ou pour introduire un plasmide portant le marqueur *pyrEF+*. L'absence de transformant ne permet pas de conclure sur le caractère essentiel du gène et peut toujours être expliquée par des problèmes d'efficacité de transformation. Ce problème a pu être contourné avec le développement de l'approche PIS qui permet, en séparant l'étape de transformation de l'étape de recombinaison et expression du phénotype KO, de confirmer le caractère essentiel du gène.

Dans l'approche PIS illustrée sur l'exemple du plasmide pKL1 (Figure 41), ce plasmide en forme circulaire s'intègre dans le chromosome d'une souche *lacS+ pyrEF-* par une première étape de recombinaison DownxDown (Figure 41B). L'approche est expliquée ici sur l'exemple d'obtention du mutant propre $\Delta lacS$ (Deng *et al.*, 2009). Après la première recombinaison, et contrairement aux approches AR et MRL, le gène ciblé (*lacS* dans la Figure 41B) est conservé intact sur le chromosome. La présence du deuxième marqueur, *pyrEF+*, facilite grandement la sélection des transformants. Aucun gène n'est inactivé à ce stade. Que le gène cible soit essentiel ou non, les recombinants *Ura⁺* seront capables de croître sur un milieu minimum. Ils seront sélectionnés pour être utilisés dans l'étape de sélection suivante. Ces recombinants sont par la suite mis en contact avec le 5-FOA dans un milieu contenant de l'uracile. En présence de 5-FOA, les produits des deux événements de recombinaison sont sélectionnés, l'un correspondant à la recombinaison UpxUp qui génère le mutant KO, l'autre correspondant à la recombinaison DownxDown qui restaure le phénotype WT. En théorie, ces deux événements de recombinaison se produisent avec une même fréquence si les régions Up et Down sont de taille similaire. Dans le cas d'un gène essentiel, les seules cellules qui pourront croître en présence de 5-FOA seront les cellules *pyrEF-* qui portent un allèle fonctionnel de ce gène (sauvage et muté) car elles n'ont subi aucun de ces deux événements de recombinaison. Les mutants spontanés *pyrEF-*, sont suffisamment fréquents pour être observés en assez grand nombre. A l'inverse, dans le cas d'un gène non essentiel, les deux types de descendants sont viables. Cependant, cette méthode présente un certain nombre de limitations. Tout d'abord, en l'absence de gène rapporteur, la différenciation entre ces deux cas de figure nécessite l'analyse moléculaire individuelle pour plusieurs dizaines voir plus de colonies candidates avant que le mutant désiré ne soit identifié.

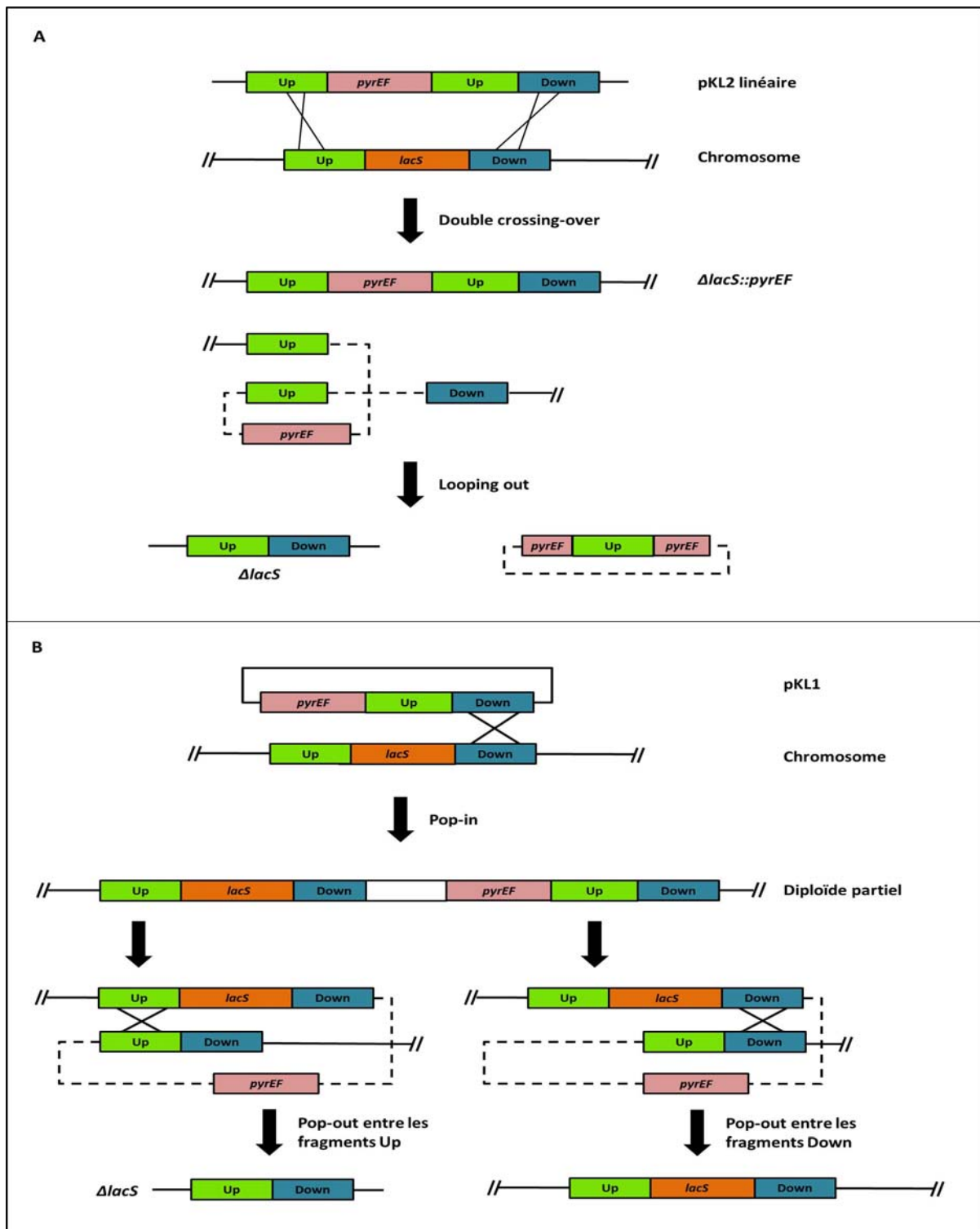


Figure 41 : Construction de mutants de délétion *lacS* par l'approche MRL (A) et PIS (B). (A) En utilisant le plasmide linéarisé pKL2. *lacS* est remplacé par *pyrEF* au niveau du chromosome par un évènement de recombinaison « double *crossing-over* » aux séquences Up et Down. Le marqueur est enlevé par une boucle au niveau de la région Up. (B) En utilisant le plasmide pKL1. Il y a intégration au niveau de la région Down. Les diploïdes *lacS* partiels se séparent soit au niveau de la région Up, soit au niveau de la région Down, en donnant le mutant *knock-out* ou l'hôte original, respectivement (Deng *et al.* 2009).

iii. Insertion de marqueur et délétion du gène cible (MID)

Aucune des méthodes présentées ci-dessus, ne convient parfaitement pour l'étude de l'essentialité d'un gène. Une nouvelle approche, appelée MID proposée récemment est certainement la plus adaptée pour étudier ce type de thématique.

L'approche MID, qui est décrite dans cette section, permet, comme l'approche PIS, de dissocier l'étape de transformation de l'étape de sélection des mutants délétés. Le principe de cette stratégie expérimentale est résumé dans la figure 42 et permet de statuer plus facilement sur la question d'essentialité d'un gène.

Le plasmide linéarisé pMID (Figure 40, modules IIIa et b) utilisé dans cette approche contient trois séquences chromosomiques d'environ 1kb : les régions IN (située en amont de la région cible) et OUT (située en aval de la région) mais aussi la région TARGET (correspondant à la fin ou au début du gène cible) (Figure 42).

Comme pour l'approche PIS, lors du premier évènement de recombinaison, les marqueurs génétiques *pyrEF+* et *lacS+* sont intégrés à proximité du gène cible, sans perturber ni sa séquence, ni son expression. Les transformants *pyrEF+* sont sélectionnés sur un milieu minimum et leur obtention dépend principalement de l'efficacité de transformation et de recombinaison. Ensuite, un deuxième évènement de recombinaison entre les séquences OUT (quand TARGET correspond au début de la région cible) ou IN (quand TARGET correspond à la fin de la région cible) provoque l'excision du génome de la région cible et des marqueurs génétiques *pyrEF+* et *lacS+*. La sélection des mutants est réalisée dans un milieu riche en présence d'uracile et de 5-FOA. La majorité ses colonies *pyrEF-* apparues sont des produits de recombinaisons qui enlèvent le locus *pyrEF+* (les autres sont des mutants spontanés) Dans le cas d'un gène ou d'une région essentiels, cette dernière étape ne permet pas d'obtenir de recombinants viables, et les seules cellules sélectionnées seront les recombinants initiaux qui accumulent des mutations ponctuelles inactivant les gènes *pyrEF+*. Ces mutants survivent en présence de 5-FOA, mais leur gène *lacS+* restera intact. En présence de X-Gal, ces colonies seront toutes bleues. Dans le cas où le gène cible n'est pas essentiel, un mélange de colonies bleues et blanches est observé sur les boîtes 5-FOA-X-Gal (Zhang *et al.* 2010).

Pour conclure, contrairement à l'approche PIS, la présence du gène rapporteur *lacS* permet d'identifier aisément le caractère essentiel du gène de la région ciblée.

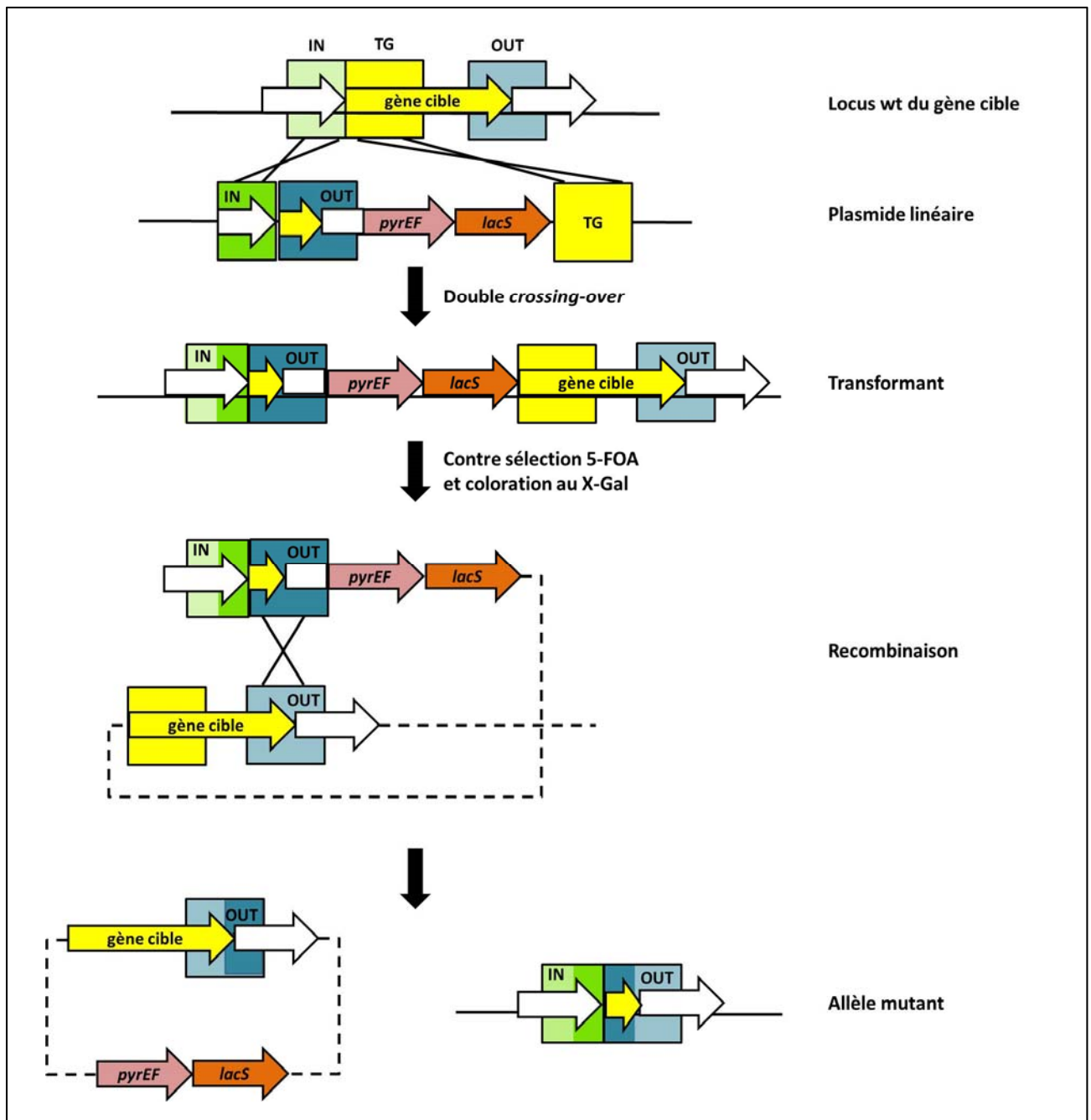


Figure 42 : Schéma de construction de mutant via la méthode MID. Les localisations des trois séquences homologues, IN, OUT et TG (TARGET), sont illustrées au niveau du locus du gène cible sauvage du chromosome hôte. Après la transformation, une recombinaison « double *crossing-over* » entre le plasmide linéarisé pMID et le chromosome de l'hôte, au niveau des régions TG et IN, conduit à l'insertion de la cassette marqueur dans le locus du gène cible. La recombinaison entre les régions OUT enlève la cassette marqueur et le gène cible, ce qui conduit à un mutant *knock-out*, sans marqueur, contre-sélectionnable sur un milieu avec du 5-FOA (Zhang *et al.* 2010).

5. Système du gène rapporteur, surexpression et expression contrôlée

La construction des mutants KO n'est pas la seule approche génétique disponible chez les archées. Depuis peu de temps, des plasmides d'expression sont disponibles pour les études génétiques chez les *Sulfolobus*. Cette question est traitée dans la section ci-dessous.

Le gène *lacZ* est communément utilisé chez les bactéries comme gène rapporteur dans les fusions transcriptionnelles pour les études de régulation (Hughes *et al.*, 2007). Le gène *lacS* des archées sert pour la même chose car son activité peut être facilement visualisée sur un milieu solide en présence de X-Gal et mesurée en milieu liquide en présence du substrat ONPG (Peng *et al.*, 2009). C'est à présent le seul marqueur d'expression utilisé dans tous les vecteurs plasmidiques construits.

Les plasmides pMJ0503, issu du virus SSV1 (Jonuscheit *et al.*, 2003), mais également pA-pK issu de pRN1 (Berkner *et al.*, 2007) ou pHZ2lacS, issu de PRN2 (Deng *et al.*, 2009) permettent la fusion transcriptionnelle de *lacS* (utilisé comme un gène rapporteur) avec les promoteurs sous études. Ainsi, Peng *et al.* ont caractérisé le promoteur *araS* et ses éléments constitutifs chez *S. islandicus* (Peng *et al.*, 2009) (le gène *lacS* et le promoteur *araS* issus de *S. solfataricus* ont été fusionnés et insérés dans le vecteur navette pZC1, issu de pHZ2 (voir paragraphe I.5.3.b.ii pour la description de ce plasmide).

Des vecteurs d'expression ont été développés récemment en utilisant soit des plasmides soit des génomes viraux. Ils ont été employés pour surproduire des protéines recombinantes chez les *Sulfolobus*. Dans ces constructions, les gènes choisis pour être surexprimés ont été placés sous contrôle des promoteurs hétérogènes. Parmi ces promoteurs, sont retrouvés ceux des gènes *araS*, *mbp/mal* et *tf55 α* codant respectivement pour des protéines de liaison à l'arabinose, au maltose et une protéine de choc thermique (Albers *et al.*, 2006; Berkner *et al.*, 2010; Mei *et al.*, 2012; Zheng *et al.*, 2012). Le promoteur *mbp/mal* ou *saci_1165* est utilisé en remplacement du promoteur *araS* chez *S. acidocaldarius*, qui présente une activité très faible (Berkner *et al.*, 2010). Le promoteur *mbp/mal* a permis de dévoiler le rôle de la protéine P98 dans la formation des structures pyramidales lors de l'infection du virus SIRV2 chez *S. acidocaldarius* en exprimant le gène *gp98* sous son contrôle (Quax *et al.*, 2010). Ce travail a permis de conclure que ces structures sont composées d'une seule protéine virale.

Précédemment chez *S. solfataricus*, il a été démontré que le promoteur *araS* est inductible par l'arabinose (Jonuscheit *et al.*, 2003; Albers *et al.*, 2006). Utilisé comme promoteur hétérologue, il permet une induction de la transcription du gène mis sous son contrôle et assure un bon niveau d'expression. Il s'agit d'un promoteur avec une activité basale très faible, en l'absence d'un inducteur. L'ajout d'arabinose dans le milieu permet à l'activité d'augmenter environ treize fois (Albers *et al.*, 2006). Son utilisation a permis, entre autres, d'étudier les fonctions des protéines de la sulfo-oxygénase réductase Sor et de l'ATPase Flal liée au flagelle (Albers *et al.*, 2006). Néanmoins, dans certains cas, le niveau d'expression assuré par le promoteur *araS* n'était pas suffisante et ce problème a été résolu par la construction d'un promoteur synthétique dérivant de *araS* (Peng *et al.*, 2012).

Ces vecteurs d'expression ont également été utilisés pour compléter des mutants délétés pour les gènes *basABC* et *basEF* (Zolghadr *et al.*, 2007), *ups* (Frols *et al.*, 2008) qui codent des protéines formant le pili chez *S. solfataricus* ou pour le gène *herA* qui code pour une hélicase (Zheng *et al.*, 2012).

6. Découvertes et progrès récents

Les espèces *Sulfolobus* sont les seuls membres des crénarchées pour lesquels l'approche génétique est développée et divers outils génétiques sont à présent disponibles pour les modèles biologiques appartenant à ce genre. Le recours aux approches génétiques a été très productif pour avancer plusieurs sujets très prometteurs chez les archées. Par exemple, il a été utilisé avec succès pour comprendre le rôle des gènes *cdvA*, *cdvB*, *cdvC* dans la division cellulaire chez les crénarchées et ont permis de conclure sur la proximité de cet appareil avec celui de formation des vésicules chez les eucaryotes (Samson *et al.*, 2008; Lindås *et al.*, 2008). D'autres exemples comprennent l'identification de nouvelles structures cellulaires de surface impliquées dans la formation du pili (Zolghadr *et al.*, 2007; Frols *et al.*, 2008) mais aussi la caractérisation de plusieurs régulateurs transcriptionnels notamment des gènes *merA* (Schelert *et al.*, 2006) ou *ss-lrpB* Peeters *et al.*, 2009).

Ces dernières années, de nombreux outils génétiques ont été développés pour les souches modèles de *Sulfolobus*. Cependant le système actuel est limité par un nombre de marqueurs de sélection insuffisant et un choix restreint de vecteurs. L'établissement d'un système génétique de la souche *S. islandicus* LAL14/1 constitue une partie principale de l'étude de cette thèse

Partie II – Résultats

1. Objectifs de la thèse

Au cours de ma thèse, j'ai travaillé sur deux types de modèles biologiques : un modèle cellulaire, une archée hyperthermophile *Sulfolobus islandicus* (souche LAL14/1) et des modèles viraux représentés par deux virus non apparentés, SIRV1 et AFV1. Les propriétés marquantes de chacun de ces modèles biologiques ont été détaillées dans l'Introduction.

Les questions que nous avons adressées à chacun de ces modèles ne sont forcément pas les mêmes, compte tenu de la connaissance actuelle des cellules d'archées et de leurs virus. En ce qui concerne le modèle cellulaire, *S. islandicus* LAL14/1, notre intérêt principal a été le développement d'une approche génétique pour cette souche. Il va de soi que la disponibilité d'une telle approche ouvre de très larges possibilités pour beaucoup de sujets (actuellement en développement ou à venir dans notre laboratoire) concernant le fonctionnement des cellules d'archées. Les modèles viraux quant à eux ouvrent la possibilité de mieux comprendre non seulement la régulation de leur cycle d'infection, question intéressante en soi, mais également une question fondamentale concernant les particularités de la régulation transcriptionnelle chez les archées. Chez ces organismes, l'activité d'une machine enzymatique typiquement eucaryotique (ARNP) est contrôlée par des protéines régulatrices typiquement bactériennes. Les virus codent en général des protéines régulatrices en assez grand nombre et ils constituent d'excellents modèles pour mieux comprendre cette régulation.

L'étude de l'interaction hôte-virus sur le modèle des virus SIRV1-2 est un des sujets principaux de notre laboratoire. Etant donné que la souche LAL14/1 est l'unique hôte naturel du virus SIRV2 (un article très récent décrit un mutant de *S. solfataricus* P2 sensible à SIRV2 (Okutan *et al.*, 2013) mais la souche est trop instable génétiquement pour être utilisée comme un bon modèle génétique), il est évident que le développement d'une approche génétique pour cette souche et la connaissance accrue de la régulation du cycle des virus SIRV1-2 apporteront une grande aide à l'étude de cette interaction.

Les objectifs de ma thèse ont été les suivants :

- Etablir un système génétique pour la souche *Sulfolobus islandicus* LAL14/1, hôte de plusieurs virus et modèle reconnu pour l'étude de l'interaction hôte-virus chez les archées.
- Réaliser une étude approfondie de plusieurs régulateurs transcriptionnels provenant à la fois des cellules d'archées et de leurs virus.

Les résultats obtenus font objet de deux publications parues et d'un article en cours d'écriture.

2. Développement d'un système génétique pour la souche modèle *S. islandicus* LAL14/1

1. Contexte de l'étude

Grâce à leur maintien facile au laboratoire, plusieurs espèces de *Sulfolobus* servent comme modèles d'études au laboratoire et représentent les espèces d'archées les mieux étudiées. Depuis plusieurs dizaines d'années, elles ont servi de modèles d'études pour comprendre les particularités de traduction (La Teana *et al.*, 2013), de transcription (Peerters *et al.*, 2013), de réplication (Samson *et al.*, 2013), de réparation (Zhang *et al.*, 2013a), de division cellulaire (Ettema *et al.*, 2009) et d'autres mécanismes cellulaires fondamentaux chez les archées hyperthermophiles.

Une de ces espèces, *Sulfolobus islandicus* étudiée dans le cadre de ma thèse, est très présente dans les différents écosystèmes à travers le monde (Russie, Etats-Unis, Islande) ce qui l'a également rendu intéressante pour l'étude de l'évolution (Zhang *et al.*, 2013b). Contrairement à *S. acidocaldarius*, elle est capable de croître sur une grande variété de sources de matières organiques telles que des peptides ou des mélanges d'acides aminés et elle utilise plusieurs sucres comme l'arabinose, le glucose, le sucrose, le tréhalose et bien d'autres (De Rosa *et al.*, 1984) En outre, à la différence de *S. solfataricus*, un autre modèle d'étude chez *Sulfolobus*, elle est très stable génétiquement.

La souche *S. islandicus* LAL14/1 est un modèle riche et unique pour l'étude de l'interaction hôte-virus chez les archées. LA14/1 peut être infectée par le virus SIRV2 qui code pour un mécanisme inhabituel de libération des virions à l'aide de structures pyramidales (Prangishvili and Quax, 2011; Bize *et al.*, 2009; Quax *et al.*, 2010 et 2011). La souche modèle de notre laboratoire, *S. islandicus* LAL14/1 est également apte pour étudier le rôle des CRISPRs chez les archées. Très riche en *spacers* visant des plasmides et des virus connus et disponibles chez les *Sulfolobus*, les CRISPRs de *S. islandicus* LAL14/1 semblent participer au phénotype de résistance de la souche au rudivirus SIRV1 compte tenu de la composition de ses séquences CRISPRs.

Lors de mon travail de thèse, le génome de LAL14/1 a été séquencé au laboratoire (travail de C. Jaubert) ouvrant la possibilité de démarrer les essais génétiques avec cette souche. J'ai pu établir une approche génétique pour ce modèle d'archée permettant la création efficace de mutants KO. Les résultats obtenus font partie d'une publication parue en 2013 dont je suis le premier co-auteur. Le pdf de cet article est inclus dans le texte de ce manuscrit. Ci-dessus, un court résumé des principaux résultats concernant l'établissement du modèle génétique.

2. Etablissement de l'approche génétique pour la souche *S. islandicus* LAL14/1

a. Isolation de mutants *pyrEF*-

Au début de ce travail, il n'était pas connu si la souche LAL14/1 était facilement transformable, aucun mutant n'était disponible et la souche ne portait aucun marqueur de sélection. Pour établir un système génétique pour la souche *S. islandicus* LAL14/1, la première étape était d'obtenir un mutant rendant possible l'utilisation des marqueurs génétiques déjà connus chez les archées.

Comme indiqué dans l'Introduction, le choix de marqueurs génétiques chez les archées est assez limité puisque pour ce domaine il existe très peu de sélection efficace à l'aide d'antibiotiques et nous

avons choisi de tirer profit de la double sélection des prototrophes-auxotrophes pour l'uracile (Ura⁺/Ura⁻).

Tout d'abord, j'ai cherché à isoler des mutants spontanés stables Ura⁻ à l'aide de la sélection en présence de 5-FOA. Ces mutants apparaissent sur les boîtes de sélection avec une fréquence comprise entre 1×10^{-5} et 1×10^{-6} . Une centaine de ces colonies résistantes au 5-FOA a été étudiée plus en détails. L'état du locus *pyrEF* de ces cent mutants Ura⁻ a été analysé par PCR. Aucun de ces mutants ne porte une délétion identifiable par l'approche PCR laissant penser qu'il s'agit de mutants ponctuels, ce qui a été confirmé par séquençage pour une dizaine de colonies (données non présentées).

Le taux apparent de mutants spontanés chez *S. islandicus* LAL14/1 a donc été estimé à 10^{-6} , qui est semblable au taux observé chez *S. islandicus* E233S (Deng *et al.*, 2009), un modèle de génétique de *Sulfolobales* stable et confirmé. Cette observation indique, entre autre, que les mutations liées aux remaniements chromosomiques ou aux insertions des éléments transposables ne sont pas fréquents chez LAL14/1.

Ces données confirment directement la stabilité génétique de LAL14/1 prédite par l'analyse génomique *in silico* (voir Introduction et Jaubert *et al.*, 2013).

b. Construction d'un mutant *S. islandicus* LAL14/1 Δ *pyrEF* par l'approche de remplacement allélique (AR)

La meilleure solution pour éviter un bruit de fond des révertants Ura⁺ est d'obtenir un mutant KO des gènes *pyrEF*. Une large délétion de cette région évitera également l'intégration des vecteurs utilisés dans le locus *pyrEF* chromosomique à la place de l'intégration dans un gène cible.

Pour construire ce mutant KO des gènes *pyrEF* chez *S. islandicus* LAL14/1, l'allèle sauvage *pyrEF*⁺ a été remplacé par l'allèle muté Δ *pyrEF* par un remplacement allélique direct. L'allèle KO des gènes *pyrEF* provenant d'une autre souche de *S. islandicus*, la souche E233S (dérivée de la souche *S. islandicus* REY15A) (Deng *et al.*, 2009) a été amplifié par PCR. Il s'agit d'une délétion de 1107pb touchant partiellement le gène *pyrE* et entièrement le gène *pyrF*. Une cassette *knock-out* linéaire de 2233pb contenant l'allèle Δ *pyrEF* et les régions en amont et en aval du locus *pyrEF* (d'une longueur de 1kb chacune) obtenues par PCR ont été introduites par transformation dans les cellules *S. islandicus* LAL14/1. Les mutants Δ *pyrEF* résultants du remplacement de la copie sauvage des gènes *pyrEF* du chromosome hôte par un double *crossing-over* ont été sélectionnés sur un milieu en présence d'uracile et de 5-FOA. La sélection a été laborieuse en raison du nombre élevé de mutants ponctuels *pyrEF*⁻ résistants au 5-FOA. Environ cent colonies ont été observées par boîte. Néanmoins, cette approche s'est révélée efficace puisque sur une vingtaine de colonies analysées, environ quinze comportaient l'allèle Δ *pyrEF*. Au final, l'efficacité de transformation de cette technique est de 75 mutants délétés/ μ g d'ADN transformé ce qui est retrouvé dans la littérature pour des transformations avec de l'ADN linéaire (entre 0-200 colonies/ μ g d'ADN linéaire contre 0-5 colonies/ μ g d'ADN circulaire, Deng *et al.*, 2009).

Ce résultat indique clairement que le protocole de transformation établi pour d'autres modèles génétiques de *Sulfolobus* est applicable à la souche LAL14/1 et son système de recombinaison

homologue est suffisamment efficace pour produire les recombinants recherchés avec une fréquence exploitable.

En résumé, nous avons obtenu le mutant LAL14/1 délété pour les gènes *pyrEF* indispensable pour débiter les expériences génétiques sur cette souche. La souche obtenue a été baptisée LAL14/1-CD et ses potentialités ont été exploitées pour réaliser la suite du projet.

3. Complémentation du phénotype Ura- par le plasmide pHZ2 chez *S. islandicus* LAL14/1-CD

Avant d'entamer chez LAL14/1-CD la construction des mutants KO qui nous intéressaient, il était indispensable de tester l'efficacité de transformation de cette souche mutante et démontrer la complémentation du phénotype Ura- en présence de la copie sauvage du locus *pyrEF*. Pour cela, nous avons transformé les cellules avec le plasmide d'expression pHZ2 (réplicon pRN2, voir Chapitre I.5.3.b.ii) qui se maintient de façon autonome chez *S. islandicus* et contient les allèles *pyrEF* sauvages. L'opéron *pyrEF+* dans cette construction est exprimé de façon constitutive sous contrôle de son propre promoteur.

L'efficacité de transformation par pHZ2 était d'environ 10^2 - 10^3 colonies/ μ g d'ADN, comparable à celle observée pour d'autres souches de *S. islandicus* (Figure 43) (Deng *et al.*, 2009. Berkner *et al.*, 2007). Les transformants sont, comme attendu, auxotrophes pour Ura+.

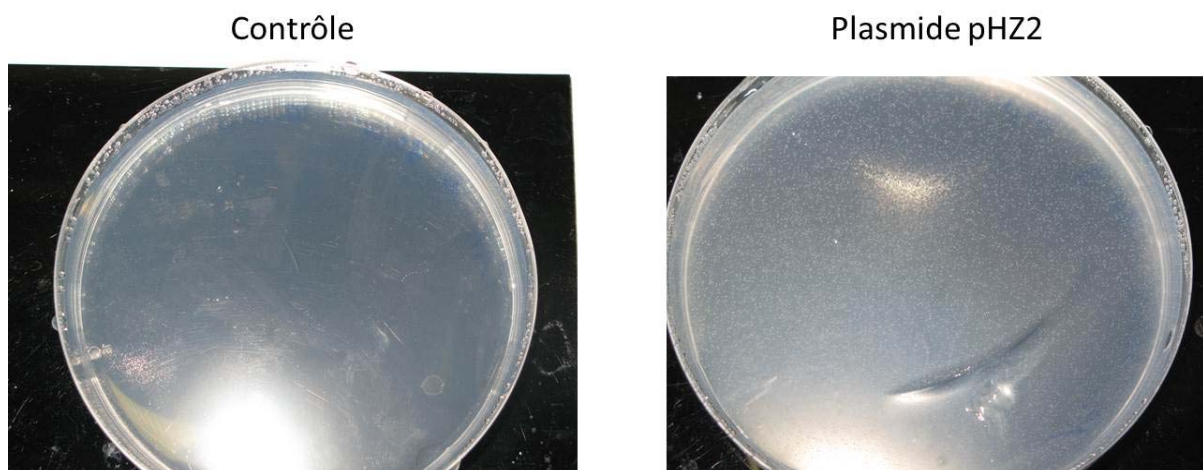


Figure 43 : Efficacité de transformation de la souche *S. islandicus* LAL14/1-CD par le plasmide d'expression pHZ2. La sélection a lieu sur un milieu minimum sans uracile où seules les cellules PyrEF+ peuvent pousser. Les cellules *S. islandicus* LAL14/1-CD sont PyrEF- et ne poussent pas sur ce milieu (à gauche). Les cellules *S. islandicus* LAL14/1-CD transformées par le plasmide pHZ2, qui contient le gène *pyrEF+* sont PyrEF+ et poussent sur ce milieu. L'efficacité de transformation est de 3300 colonies par μ g d'ADN.

4. Construction d'un mutant Δ CRISPR_1

Les systèmes CRISPR/Cas représentent une ligne de défense cellulaire contre les éléments génétiques extra chromosomiques (plasmides conjugatifs, virus). Appelé « mécanisme de l'immunité acquise », le système CRISPR/Cas permet le développement d'une réponse très spécifique et ciblée contre ces « envahisseurs », réponse comparable par sa spécificité aux réponses immunologiques chez les eucaryotes (Marrafini and Sontheimer, 2008; Deveau *et al.*, 2010; Horvath and Barrangou, 2010).

Toutes les séquences et les gènes liés au système CRISPR présents chez *S. islandicus* LAL14/1 se situent dans la région génomique variable. Le génome de LAL14/1 porte cinq loci CRISPR, trois cassettes de gènes *cas* associées au sous-type I-A et deux cassettes de gènes *cmr* associées au sous-type III-B. D'après la composition des séquences *leader* et répétées, les cinq CRISPRs de *S. islandicus* LAL14/1 peuvent être divisés en deux familles, I et III (Lillestol *et al.*, 2009).

Le locus CRISPR de la famille I comprend deux blocs orientés de façon opposée composés de répétitions-*spacers* séparés par le premier module de gènes *cas*. Le deuxième module est localisé à l'extrémité du deuxième bloc de répétitions-*spacers*. L'analyse des 285 *spacers* formant les séquences CRISPRs de LAL14/1 révèle la présence d'un nombre étonnamment élevé de *spacers* qui matchent parfaitement (dans trois cas) ou imparfaitement aux génomes des rudivirus, fusellovirus et plasmides conjugatifs, décrits dans les sources d'eau chaudes d'Islande (Peng *et al.*, 2004; Blum *et al.*, 2001). Les deux *spacers* qui matchent parfaitement, portés par les CRISPR_1 et CRISPR_5, ciblent le génome du rudivirus SIRV1 (Zillig *et al.*, 1996). De façon intéressante, le *spacer* dans le CRISPR_1 matche avec le gène de SIRV1 codant pour la protéine P98, responsable de la formation de structures pyramidales impliquées dans la sortie du virion (Bize *et al.*, 2009; Quax *et al.*, 2010; 2011).

Une approche génétique puissante de *pop-in/pop-out* a été développée précédemment pour un autre modèle génétique, *S. islandicus* REY15A (Deng *et al.*, 2009). Elle permet la création rapide et efficace de mutants *knock-out*. Pour montrer que cette approche est efficace chez LAL14/1, nous avons choisi de déléter un des loci CRISPR (CRISPR_1) parce que les fonctions codées par les CRISPR ne sont habituellement pas essentielles pour les cellules en absence de virus et leurs mutants de délétion sont censés être viables. Aussi, un des *spacers* du CRISPR_1 matche parfaitement avec le virus SIRV1 contre lequel LAL14/1 est résistante. Si la résistance est liée à l'activité CRISPR, son inactivation peut diminuer le niveau de résistance donnant un phénotype détectable à ce mutant.

Le plasmide *knock-out* utilisé pour inactiver le CRISPR_1 et les positions des régions IN (867pb), OUT (919pb) et TARGET (776pb) dans le plasmide pSEF décrit par Deng *et al.* (Deng *et al.*, 2009) sont présentés dans la Figure 44. Il est basé sur la technique MID présentée dans le chapitre 1.5.4. Deux *rounds* successifs de recombinaison suppriment le fragment chromosomique situé entre les régions IN et OUT produisant un dérivé *pyrEF+* à partir duquel la région CRISPR_1/*cas* peut être délétée (Δ CRISPR_1/ Δ *csx1* Δ *cas4*; *csx1* est annoté comme Sil_0393) (Figure 44).

Après la transformation des cellules LAL14/1 avec ce plasmide *knock-out*, une première étape de sélection des transformants *pyrEF+* (qui ont intégré la cassette dans leur génome) a été réalisée sur un milieu sans uracile. Après dix jours d'incubation à 78°C, environ 150 colonies/ μ g d'ADN ont été observées, ce qui témoigne d'une très bonne efficacité de transformation pour *S. islandicus*. Les données fournies par la littérature sont entre 10 à 100 colonies/ μ g d'ADN (Deng *et al.*, 2009). Ensuite, ces colonies ont été lancées dans des cultures liquides en milieu sans uracile sur quatre cycles de purification-dilution. Puis, les cellules ont été étalées sur un milieu riche contenant du 5-FOA et de l'uracile, pour sélectionner le deuxième événement de recombinaison et donc les mutants délétés. Contrairement à la technique initiale qui utilise un double mutant Δ *pyrEF* Δ *lacs*, la souche LAL14/1-CD est seulement Δ *pyrEF* ce qui empêche la sélection visuelle blanc/bleu des mutants Δ CRISPR_1 en présence de X-Gal, lors de la dernière étape de cette approche. Un des avantages de la technique MID est perdu, contrairement au cas de la construction du *knock-out* du gène *sta1* qui permettait de discriminer facilement avec la coloration blanc/bleu.

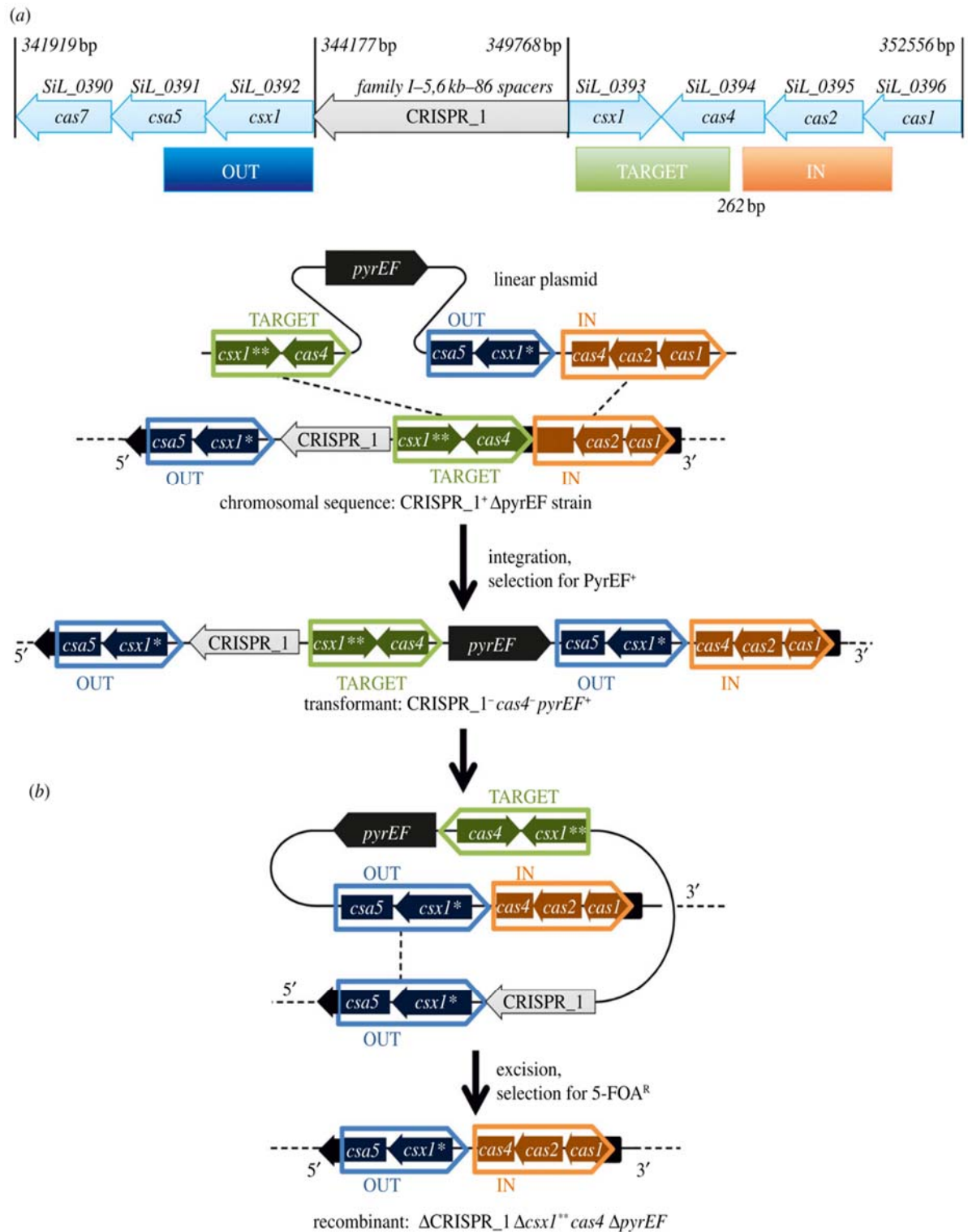


Figure 44 : Carte génétique de la région CRISPR_1 délétée par l'approche *pop-in/pop-out*. (a) Carte génétique de la région CRISPR_1. La région cible est située entre les régions OUT et TARGET. (b) Schéma représentant les deux étapes d'évènements de recombinaison générant le mutant ΔCRISPR_1/ Δcsx1Δcas4. Deux paralogues de csx1 sont présents dans cette région. Le gène csx1* correspond au gène SiL_0392 et csx1** au gène SiL_0393.**

Une centaine de colonies sont apparues par boîte et la région CRISPR_1 a été analysée par PCR pour une dizaine d'entre elles. La délétion a été confirmée pour chacune des colonies testées (Figure 45). Cela a par ailleurs permis de démontrer qu'en utilisant cette approche même des régions relativement longues du génome (dans ce cas 6.6kb) peuvent être supprimées efficacement.

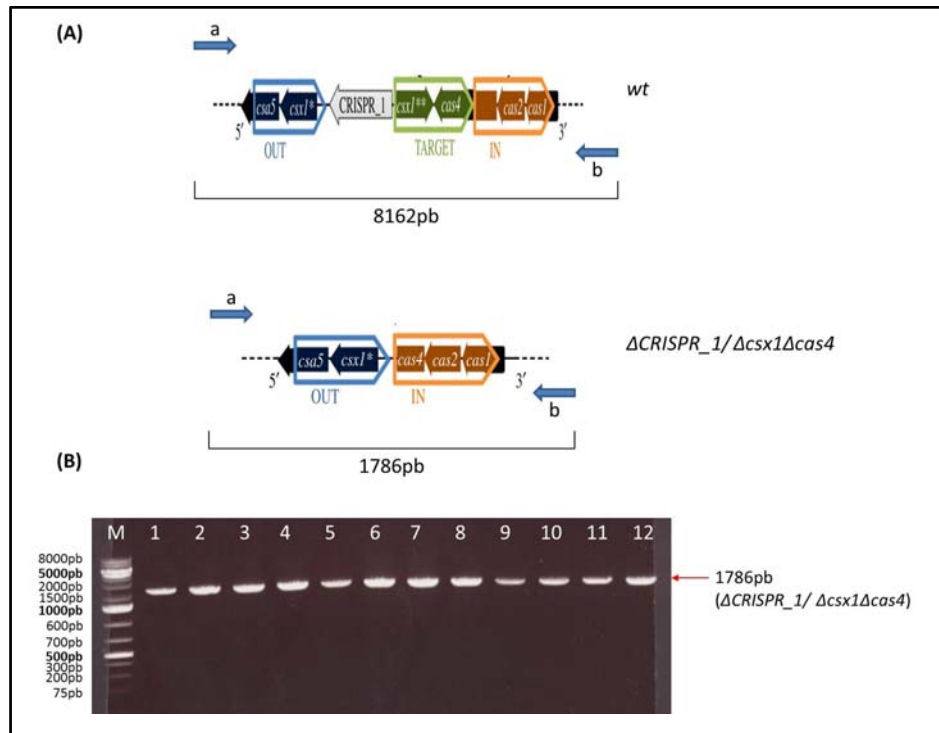


Figure 45 : Analyse par PCR des mutants obtenus. (A) Le couple d'amorces flanquantes est représenté et les tailles attendues des fragments amplifiés sont données en pb. (B) Les fragments correspondant à la région du locus cible *CRISPR_1* ont été amplifiés par PCR sur les cultures de transformants (puits 1 à 12) après plusieurs cycles de purification. La taille observée est la taille attendue : 1786pb pour $\Delta CRISPR_1$.

La dernière étape consistait à étudier le phénotype de ce mutant (travail fait en collaboration avec Carole Jaubert) afin de déterminer si la délétion du *CRISPR_1* avait un effet sur la résistance de la souche LAL14/1 au virus SIRV1. Ce phénotype de résistance/sensibilité contre le virus SIRV1 a été caractérisé par des techniques de spot-test semi quantitatif, qui permettent une estimation assez précise du nombre de plages de lyse formées.

De façon intéressante, le mutant *CRISPR_1/Δcsx1Δcas4* est résistant à l'infection par SIRV1 comme *S. islandicus* LAL14/1-CD. La présence d'un second *spacer* et *CRISPR_5* fonctionnels peut expliquer ce résultat.

Une construction d'un double mutant $\Delta CRISPR_1 \Delta CRISPR_5$ aiderait à mieux caractériser une éventuelle implication de l'immunité générée par les CRISPR dans la résistance de LAL14/1 à SIRV1.

Les approches génétiques ont été développées pour un nombre limité d'espèces *Sulfolobus* (Allers *et al.*, 2005; Leigh *et al.*, 2011), incluant *S. solfataricus* P1 et 98/2 (Jonuscheit *et al.*, 2003; Worthington *et al.*, 2003), *S. acidocaldarius* (Wagner *et al.*, 2009; 2012), *S. islandicus* REY15A (Deng *et al.*, 20069;

She *et al.*, 2009) et M.16.4 (Zhang and Whitaker, 2012). Cependant, aucune approche génétique n'avait été disponible précédemment pour LAL14/1.

Dans ce travail de thèse, une approche génétique a été établie pour cette souche en créant deux mutants $\Delta pyrEF$ et $\Delta CRISPR_1$.

L'efficacité de transformation observée pour LAL14/1 ainsi que la facilité de création et de sélection de ces mutants de délétions indique que LAL14/1 représente un excellent modèle génétique en général. Plus particulièrement, elle est une souche de choix pour l'étude de tous les aspects de l'interaction hôte-virus, du premier contact entre la cellule et le virus et l'entrée de l'ADN viral dans la cellule à la réponse immunitaire de cette dernière qu'elle développe grâce aux systèmes CRISPR-*cas* codés par le chromosome.

5. Construction d'un mutant $\Delta topR2$

Les topoisomérases sont des protéines ubiquitaires présentes dans tous les organismes vivants : Eucaryotes, Bactéries et Archées. Elles sont classées en deux types différents selon leur activité enzymatique: type I et type II, réalisant des cassures transitoires simple et double brin, respectivement (Wang 2002). Une topoisomérase IA particulière, nommée réverse gyrase a été découverte chez les archées hyperthermophiles (Kikuchi et Asai, 1984; Nadal 2007) et plus tard chez les bactéries hyperthermophiles (Nadal 2007). Cette topoisomérase introduit des supertours positifs dans l'ADN (Kikuchi and Asai, 1984; Forterre *et al.*, 1985; Nadal *et al.*, 1988). Parce que le gène codant pour la réverse gyrase est restreint aux organismes hyperthermophiles (Archées ou Bactéries), cette protéine a été considérée comme un marqueur moléculaire de la thermophilie (Forterre 2002).

L'archée *S. islandicus* code deux réverse gyrases, TopR1 et TopR2 ayant des propriétés différentes (Bizard *et al.*, 2011). Cette duplication semble être importante car la plupart des Crénarchées présentent deux copies de réverse gyrase (Brochier-Armanet and Forterre, 2007).

Pour mieux comprendre le rôle de la protéine TopR2, nous avons décidé de construire un mutant $\Delta topR2$ chez *S. islandicus*. Des expériences précédentes conduites chez *S. islandicus* E233S ont confirmé que le gène *topR1* était essentiel (Liu *et al.*, 2009; Zhang *et al.*, 2013). C'est pour cela que nous avons décidé de construire ce mutant par la technique MID, décrite en détails dans la partie Introduction I.5.4.b.ii. Seule la construction du plasmide utilisé avec les trois fragments pour la recombinaison dans le génome est présentée ici (Figure 46).

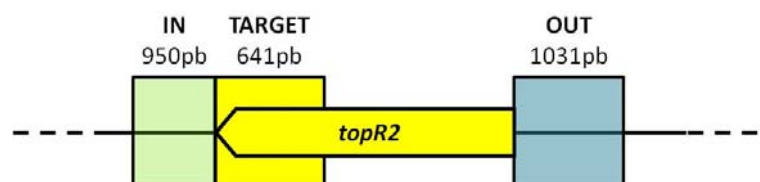


Figure 46 : Région du locus du gène *topR2* chez *Sulfolobus islandicus*. La figure indique les positions des fragments OUT, IN et TARGET clonés dans un vecteur permettant l'utilisation de l'approche MID (décrite dans l'Introduction, Deng *et al.*, 2009) pour réaliser le *knock-out* du gène *topR2*.

Un premier essai réalisé à 68°C a été infructueux, laissant penser que le gène peut être essentiel.

Les organismes hyperthermophiles sont soumis à un risque permanent de dénaturation de l'ADN due à la haute température de leur environnement. Par conséquent, pour maintenir la topologie intacte de l'ADN génomique ils ont besoin d'une réverse gyrase capable d'introduire des super tours positifs. De façon intéressante, avec l'augmentation de la température, l'activité de la réverse gyrase augmente également (Lopez-Garcia and Forterre 1999). Cette observation nous a conduit à penser qu'il serait intéressant de réaliser l'expérience *de knock-out* à une température plus basse avant de conclure définitivement à l'essentialité du gène *topR2*.

Par manque de temps, nous n'avons pas pu réaliser la deuxième étape de cette expérience. Par ailleurs, une autre équipe a démontré très récemment qu'à la température optimale de croissance de *S. islandicus* (78°C) le gène *topR2* est essentiel (Zhang *et al.*, 2013). Cependant, l'article ne répond pas à la question de l'essentialité de ce gène à plus basse température (60-65°C).



Cite this article: Jaubert C, Danioux C, Oberto J, Cortez D, Bize A, Krupovic M, She Q, Forterre P, Prangishvili D, Sezonov G. 2013 Genomics and genetics of *Sulfolobus islandicus* LAL14/1, a model hyperthermophilic archaeon. *Open Biol* 3: 130010. <http://dx.doi.org/10.1098/rsob.130010>

Received: 10 January 2013

Accepted: 28 March 2013

Subject Area:

genetics/genomics/microbiology/bioinformatics

Keywords:

Archaea, *Sulfolobus islandicus* LAL14/1, genome analysis, genetics, CRISPR

Author for correspondence:

Guennadi Sezonov
e-mail: sezonov@pasteur.fr

[†]These authors contributed equally to this study.

[‡]Present address: Centre Intégré de Génétique, Université de Lausanne, Le Génopode, Quartier UNIL-Sorge, Lausanne, Switzerland.

Electronic supplementary material is available at <http://dx.doi.org/10.1098/rsob.130010>.

Genomics and genetics of *Sulfolobus islandicus* LAL14/1, a model hyperthermophilic archaeon

Carole Jaubert^{1,†}, Chloë Danioux^{1,†}, Jacques Oberto², Diego Cortez^{1,‡}, Ariane Bize³, Mart Krupovic¹, Qunxin She⁴, Patrick Forterre^{1,2}, David Prangishvili¹ and Guennadi Sezonov^{1,5}

¹Département de Microbiologie, Unité Biologie Moléculaire du Gène chez les Extrémophiles, Institut Pasteur, Paris, France

²CNRS, UMR8621, Institut de Génétique et Microbiologie, Université Paris-Sud 11, 91405 Orsay Cedex, France

³Irstea, UR HBAN, 92761 Antony, France

⁴Danish Archaea Centre, Department of Biology, University of Copenhagen, Ole Maaløes Vej 5, 2200 Copenhagen N, Denmark

⁵UMR 7138 'Systématique, Adaptation, Evolution', Université Pierre et Marie Curie, Paris, France

1. Summary

The 2 465 177 bp genome of *Sulfolobus islandicus* LAL14/1, host of the model rudivirus SIRV2, was sequenced. Exhaustive comparative genomic analysis of *S. islandicus* LAL14/1 and the nine other completely sequenced *S. islandicus* strains isolated from Iceland, Russia and USA revealed a highly syntenic common core genome of approximately 2 Mb and a long hyperplastic region containing most of the strain-specific genes. In LAL14/1, the latter region is enriched in insertion sequences, CRISPR (clustered regularly interspaced short palindromic repeats), glycosyl transferase genes, toxin–antitoxin genes and MITE (miniature inverted-repeat transposable elements). The tRNA genes of LAL14/1 are preferential targets for the integration of mobile elements but clusters of atypical genes (CAG) are also integrated elsewhere in the genome. LAL14/1 carries five CRISPR loci with 10 per cent of spacers matching perfectly or imperfectly the genomes of archaeal viruses and plasmids found in the Icelandic hot springs. Strikingly, the CRISPR_2 region of LAL14/1 carries an unusually long 1.9 kb spacer interspersed between two repeat regions and displays a high similarity to pING1-like conjugative plasmids. Finally, we have developed a genetic system for *S. islandicus* LAL14/1 and created Δ *pyrEF* and Δ *CRISPR_1* mutants using double cross-over and pop-in/pop-out approaches, respectively. Thus, LAL14/1 is a promising model to study virus–host interactions and the CRISPR/Cas defence mechanism in Archaea.

2. Introduction

The genus *Sulfolobus* was first described by Brock *et al.* in 1972 [1] and includes thermoacidophilic Archaea that grow at 70–85°C and pH 2–3 under aerobic conditions either chemolithotrophically by oxidizing elementary sulfur/hydrogen sulfide or

heterotrophically [2]. *Sulfolobus* strains have been isolated from various acidic thermal habitats (in the USA, Italy, Iceland, Russia and elsewhere). They are easily maintained under laboratory conditions, making them convenient models to study the molecular organization of the archaeal cell [3].

The sequences of 12 *Sulfolobus* genomes are currently available. They include *Sulfolobus solfataricus* P2 [4], *Sulfolobus tokodaii* [5], *Sulfolobus acidocaldarius* [6] and nine strains of *Sulfolobus islandicus*: HV10/4 and REY15A [7] (isolated from hot springs in Iceland) [8]; M.14.25, M.16.27 and M.16.4 (from hot springs at the Mutnovsky Volcano, Kamchatka, Russia); Y.N.15.51 and Y.G.57.14 (from hot springs in Yellowstone National Park, USA); and L.D.8.5 and L.S.2.15 (from Lassen National Park, USA) [9].

The strain *S. islandicus* LAL14/1 was isolated in 1995 from a solfataric field in Iceland by the group of Zillig [8]. Its geographical origin, growth requirements and physiology indicate that LAL14/1 is a close relative of two *S. islandicus* strains also isolated from Iceland, HVE10/4 and REY15A. However, LAL14/1 has a particular pattern of sensitivity to various archaeal viruses. LAL14/1 is resistant to the rudivirus SIRV1 but can be efficiently infected by its close relative SIRV2 [10], which has a complex cycle of development in the host cells. At the end of the infection cycle, that lasts about 14 h, specific pyramid-like structures are formed on the cell surface facilitating release of virus particles [11–13]. These unique characteristics make *S. islandicus* LAL14/1 an interesting model to study virus–host interactions in Archaea.

Effective genetic tools have been developed for a limited number of *Sulfolobus* species [14], including *S. solfataricus* P1 and 98/2 [15,16], *S. acidocaldarius* [17,18], *S. islandicus* REY15A [19–21] and *S. islandicus* M.16.4 [22]. However, genetic approaches have not previously been available for LAL14/1.

In this study, we report the results of the *in silico* analysis of the genome sequence of *S. islandicus* LAL14/1 and detailed comparisons with other available *S. islandicus* strains, and in particular the closely related strains, REY15A and HVE10/4. We also have established genetic tools for this strain by creating both Δ *pyrEF* and Δ *CRISPR_1* mutants. This work has made substantial progress towards the possibility of applying powerful global approaches (for example, transcriptome, RNAseq and proteome analyses) to elucidate the interplay between host and viral genes and proteins during the viral infection cycle.

3. Material and methods

3.1. Strains growth

Sulfolobus islandicus strains were grown aerobically at 80°C and under constant agitation in rich medium containing 0.2 g l⁻¹ of Tryptone Peptone, 2 g l⁻¹ of sucrose and 1 g l⁻¹ of yeast extract. The minimal medium used to select Ura⁺ variant isolates was as described previously [2]. Ura⁻ mutants were selected on rich solid medium in the presence of 50 mg l⁻¹ of 5'-fluoro-orotic acid.

3.2. Genetic experiment

3.2.1. PCR amplification

pyrEF mutant. The following primers were used to amplify the locus, including the *pyrEF* operon and the upstream (1 kb) and

downstream (1 kb) situated regions of the *S. islandicus* E233S chromosome: oligoUP, CAGTAGCTAAAACAATTGAAAGA GTAGGTG; oligoDOWN, CTAATGATGCTTGATAGAAGTA TTTAGCGT. The PCR amplification was performed in 50 µl of reaction mixture containing 10 µM of each primer, 1 µl template, 10 µl 5× HF Phusion Buffer (Finzyme), 10 nM dNTPs and 0.5 µl *pfu* DNA polymerase (Finzyme) with the following conditions: 30 s at 98°C, 60 s at 55°C and 90 s at 72°C for 35 cycles.

CRISPR mutant. The following primers were used to amplify the DNA fragments IN (oligoup1, AAAAAACCATG GTACGATCCGCTTAAGCC; oligodown2, AAAAAAGGA TCCGTAATGAGAGCTTGTTT); OUT (oligoup3, AAAAA GTCGACTACTACCGTGTACTTCCCC; oligodown4, AAAA ACCATGGTGCCTTAATGAGGCAAGGT) and TARGET (oligoup5, AAAAAAGCATGCTTCTGCTCAAAAAGGAGGA; oligodown6, AAAAACTGCAGTAGAAGAAGATAGCCC AC). The positions of these fragments is indicated in figure 8.

Transformation. Electroporation of *S. islandicus* LAL14/1 was performed as described by Deng *et al.* [19].

3.3. Genome sequencing

Total DNA was extracted from the cells using phenol–chloroform and ethanol precipitation. The sequencing was done by Fidelity System Inc. using Illumina technology and assembled using the software VELVET v. 1.2 [23]. The genome was automatically annotated and refined manually. Open reading frames (ORFs) were predicted using *phred*, *phrap* and *consed* software [24–26] and tRNA with tRNAscan-SE [27]. Putative insertion sequence (IS) elements were identified by BLASTn search against the IS Finder Database (<http://www.is-biotoul.fr/>). Annotations were manually curated using UGENE software [28].

3.4. Phylogenetic analysis

The genome sequences of nine *S. islandicus* strains were downloaded as Genbank files from the NCBI database (NC_012588, NC_012623, NC_017275, NC_013769, NC_012589, NC_012632, NC_012726, NC_017276 and NC_012622).

The DNA sequences of all genomes were aligned using the *progressiveMauve* algorithm with default parameters and analysed with *stripSubsetLCBs*. The *S. islandicus* core genome sequence was used for phylogenetic dating based on the standard rate of accumulation of random mutations in hyperthermophilic Archaea (4.66×10^{-9} substitution per site per year; [9]). Clonal genealogy was inferred using the *ClonalFrame* algorithm three times.

3.5. Dot-Plot

The UGENE Dot-Plot algorithm with the option ‘search for inverted repeats’ enabled was used to generate dot-plots from pairs of sequences. Sequences from the 10 *S. islandicus* strains were aligned with the *progressiveMauve* algorithm with scoring parameters divided by four to ensure the recognition of large conserved regions.

3.6. Replication origins

ZPLOTTER APPLET v. 2.0 was used to calculate Zcurves (<http://tubic.tju.edu.cn/zcurve/>).

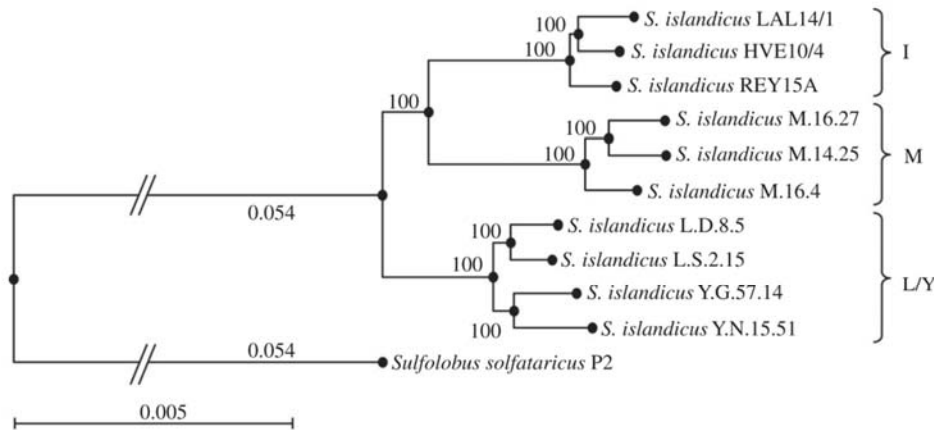


Figure 1. Phylogenetic position of *S. islandicus* LAL14/1 among other sequenced *S. islandicus* strains. *Sulfolobus solfataricus* P2 is used as an external group. The length of branches is proportional to the phylogenetic distance between the strains. Bootstrap values are indicated.

OriC1 and OriC2 structures were determined by alignment of OriC1 and OriC2 sequences in *S. solfataricus* P2 with *S. islandicus* LAL14/1 sequences using the BLASTn algorithm with default parameters. OriC3 structures were determined by mapping UCM (UnCharacterized Motif) sequences in the genome of LAL14/1 using the UGENE search tool with default parameters.

3.7. Exceptional motifs

R'MES software [29] was used to evaluate the significance of motif frequency in *S. islandicus* genomes. This statistical method compares the observed count of each motif to the count predicted by a reference probabilistic Markovian model. Exceptional motifs of lengths two to eight were analysed choosing models according to the authors' guidelines (<http://migale.jouy.inra.fr/?q=method>) and using the highest possible order each time (usually $I-2$, where I is the length of the studied motif).

3.8. Spacer data

All available CRISPR spacers from *S. islandicus* LAL14/1 were determined using CRISPRFINDER (<http://crispr.u-psud.fr/Server/>) [30].

3.9. PAM motifs

To find PAM motifs, protospacers corresponding to the selected spacers listed in the electronic supplementary material, table S15 were aligned and visualized with WEBLOGO (<http://weblogo.berkeley.edu/logo.cgi>). For each protospacer, the regions analysed were 10 nucleotide-long sequences immediately upstream and downstream from the region identical or similar to the spacer.

3.10. *Sulfolobus islandicus* pan-genome characterization

The pan-genome of the 10 available *S. islandicus* strains was obtained with an in-house program [31]. Briefly, the proteins encoded by the 10 genomes were compared using two-way

Table 1. General properties and composition of the genome of *S. islandicus* LAL14/1.

genome size	2 465 177 bp
protein-coding genes	2601 (85.6%)
CDS average size	815 bp
average size of intergenic regions	190 bp
tRNA-coding genes	45 functional tRNA

BLASTp analysis and ranked in families of orthologous proteins according to the criteria defined by Lerat *et al.* [32].

4. Results

4.1. General features of the *Sulfolobus islandicus* LAL14/1 genome

The genome of *S. islandicus* strain LAL14/1 (NCBI accession no. CP003928) was sequenced at 104-fold coverage by Fidelity System (<http://fidelitysystems.com/>) using Illumina technology. The protein-coding genes were annotated using UGENE software [28]. The *S. islandicus* LAL14/1 genome consists of a single circular chromosome of 2 465 177 bp; 85.6 per cent of the genome is coding. The chromosome carries 2601 protein-coding genes and has a GC content of 35 per cent. The general properties of the *S. islandicus* LAL14/1 genome composition are summarized in table 1.

4.2. Phylogenetic position of *Sulfolobus islandicus* LAL14/1

The phylogenetic tree of available *S. islandicus* strains was established by comparison of chromosomal DNA sequences shared by all 10 strains (figure 1) as described in §3. The strain LAL14/1 is phylogenetically very close to the strains HVE10/4 and REY15A, also isolated from Iceland. On the basis of the standard rate of random mutation accumulation in *S. islandicus* [9], we estimate that strain REY15A separated from the clade LAL + HVE about 460 000 years ago and strain LAL14/1 diverged from the clade LAL + HVE 60 000

years later. The tree is divided into three main clades corresponding to the geographical origins of the strains: clade I from Iceland; clade M from Kamchatka (Russia) and clade L/Y from Lassen and Yellowstone (USA) [9,33].

The general features of the 10 *S. islandicus* genomes are reported in table 2. LAL14/1 has the smallest genome, apparently due to the small number of horizontally transferred CAG regions (see below). It carries a remarkably complex CRISPR system, with more spacers than the other *S. islandicus* genomes, some matching the virus SIRV1 perfectly.

4.3. The *Sulfolobus islandicus* pan-genome and specific genomic pattern of the strain LAL14/1

The first version of the *S. islandicus* pan-genome was published in 2009, based on the analysis of seven strains; it includes 20 610 proteins [9]. Three additional *S. islandicus* genome sequences have since become available (REY15A, HVE10/4 [7] and LAL14/1 (present work)). The new updated version of the *S. islandicus* pan-genome has 27 578 proteins (*in silico* prediction) that can be divided into 3492 families of orthologous proteins (see §3). The statistics of the family distribution is presented in the electronic supplementary material, figure S1. There are 1892 ubiquitous families, present in at least one copy in all of the *S. islandicus* genomes. This group, indicated in the annotation by *arCOG* + *number*, constitutes the *S. islandicus* core-genome.

There are 1030 families present in two or more (but not all) strains. The best-represented families of *S. islandicus* include various transposases, ABC transporters and CoA pathway genes (see the electronic supplementary material, table S1). The majority of these families are present in the LAL14/1 genome, and some are overrepresented (more frequent than predicted from the average pan-genome statistics), for example, the transposases belonging to the *IS1* family. Others, for example transposases of families *ISH3* and *IS110*, are clearly underrepresented.

There are 570 families classified as *singletons*. They are strictly strain-specific and present in only one copy per genome (see the electronic supplementary material, figure S2 and table S2). The *S. islandicus* LAL14/1 genome contains 65 singletons, and specific functions could be predicted for 14 of them. They include a putative transcription regulator of the MarR family (SiL_0405), a putative acyl-coenzyme A synthetase/AMP fatty acid ligase (SiL_0481), a small subunit of the methyltransferase (SiL_0587), a putative glycosyltransferase (SiL_0818), a membrane protein involved in the export of O-antigen and teichoic acid (SiL_0839), a putative secreted endonuclease distantly related to the archaeal Holliday junction resolvase (SiL_1319), an Fe-S oxidoreductase (SiL_1473), and seven families of Cmr proteins related to CRISPRs: Cmr3 (SiL_0600), Cas10 (SiL_0601), Cmr6 (SiL_0602), Cmr5 (SiL_0603), Cmr1 (SiL_0604), Cmr4 (SiL_0605) and Csm6-like protein (SiL_0630).

The taxonomical distribution of the individual genes and singletons of *S. islandicus* LAL14/1 is summarized in table 3. Of the 2601 annotated protein-coding genes, only 4.7 per cent are exclusive to this strain; 10 per cent are only found in *S. islandicus* species and 37.6 per cent are specific to Sulfolobales. The 51 *S. islandicus* LAL14/1-specific singletons, and the seven *S. islandicus*-specific and seven Crenarchaeota-

Table 2. Comparison of genomic patterns of 10 *S. islandicus* strains (compilation of our own data and those from [7,9]. CAG are discussed elsewhere in the text. SIRV1 and SIRV2 are rudiviruses described in [10,34,35]. R, resistant; S, sensitive; nd, no data. CRISPR and Cmr families are indicated following established classification [36,37].

	LAL14/1	REY15A	HVE10/4	L.D.8.5	L.S.2.15	M.16.4	M.16.27	M.14.25	Y.G.57.14	Y.N.15.51
genome size (Mb)	2.4	2.7	2.5	2.7	2.7	2.6	2.7	2.6	2.7	2.8
gene no.	2601	2666	2745	2939	2760	2759	2680	2632	2928	2868
CAG no.	3	4	7	16	13	5	5	11	16	22
CRISPR/Cas families (no. of spacers)	I (199) III (84)	I (208)	I (215) III (47)	I (223) III (42)	I (225) II (7)	I (139)	I (164) II (72)	I (143) II (39)	I (57)	I (99)
Cmr families	B,F	B	2 × B	B	B	2 × B	B	B,D	3 × B	E
perfect spacers against SIRV1 and phenotype	2 (R)	0 (R)	0 (R)	0 nd	0 nd	0 nd	0 nd	0 nd	0 nd	0 nd
perfect spacers against SIRV2 and phenotype	0 (S)	2 (R)	3 (R)	0 nd	0 nd	0 nd	0 nd	0 nd	0 nd	0 nd

Table 3. Taxonomic specificity of the protein-coding genes of *S. islandicus* LAL14/1.

gene specificity	<i>S. islandicus</i> LAL14/1	
	total gene distribution	singleton distribution
<i>S. islandicus</i> LAL14/1-exclusive	123	51
<i>S. islandicus</i> -exclusive (10 strains)	134	0
<i>Sulfolobus</i> -specific	719	7
Crenarchaeota-specific	238	7
Archaea-specific	409	0
Archaea + Bacteria-specific	535	0
Archaea + Eukarya specific	72	0
universal	371	0
total	2601	65

specific singletons are listed in the electronic supplementary material, tables S3A–C.

The specific genomic pattern of LAL14/1 is presented in table 4. Like all the other studied *S. islandicus* genomes, LAL14/1 codes for a large number of transposases representing families composed of multiple paralogues. In this strain, the most represented family is the protein OrfB encoded by IS200/IS605. The transposon ISC1048 of the family IS607 is also overrepresented. However, other transposase families, such as families IS110 and ISH3, are clearly underrepresented in LAL14/1.

The families of orthologous proteins were compared between the *S. islandicus* strains LAL14/1, REY15A and HVE10/4, all isolated from the same geographical location. This revealed substantial similarity between the genomes of these strains: of the 2770 families analysed, 2130 (77%) are shared by these three *S. islandicus* genomes. A surprisingly large number of families are unique for each of these strains (figure 2), constituting a specific genomic signature for each of the strains.

4.4. Exceptional motifs in *Sulfolobus islandicus* LAL14/1 genome

Many non-coding motifs have specific biological functions in genomes and statistical analyses of oligomer frequencies in genome sequences can identify possibly significant motifs (e.g. reviewed in [38] for bacteria). Similar analyses have also been very useful for studying the evolution of genome functions and regulation [39].

Systematic analyses of short oligonucleotides of fixed composition (usually called *words*) were conducted with the LAL14/1 and other *S. islandicus* genomes to identify non-coding functional motif candidates. The *word* frequency and preference patterns are very similar in the three closely related *S. islandicus* strains, suggesting that their functional motifs and associated mechanisms are generally similar. Some more pronounced differences were observed for

longer *words*, and this was mainly associated with the different genomic content of the large variable regions.

As commonly observed in archaeal and bacterial genomes, palindromic motifs are generally avoided (e.g. electronic supplementary material, table S4 for LAL14/1 and electronic supplementary material, table S5), possibly as a consequence of the presence of restriction–modification systems encoded in the genome [40] or of other biological phenomena such as the control of chromosome replication [38]. Among the analysed words (see the electronic supplementary material, tables S4 and S5), some are clearly overrepresented because of their presence in the repeats of the CRISPR sequences. For example, the overrepresented *word* ACTATAGA, included in the CRISPR repeats, is repeated 196 times (see the electronic supplementary material, figure S3).

For most of the other candidates, no obvious biological function could directly be inferred. There is evidence for eukaryotes that the non-random pattern of short words (two to four letters) may be due to evolutionary changes in informational processes such as DNA replication and repair, and the pattern of long *words* (eight letters) may reflect evolutionary changes in gene regulatory machinery. The presence of such long *words* may reflect a non-random frequency of the DNA-binding sites specific for transcription factors [39]. This observation might indicate the presence of eukaryotic-like transcriptional regulation in Archaea.

4.5. Structure and dynamics of the *Sulfolobus islandicus* genomes

The structural comparison of two closely related genomes, HVE10/4 and REY15A, was published in 2011 [7]. It revealed a high level of synteny of gene content for these genomes and the presence of two variable regions, one of about 0.5–0.7 Mb and a second corresponding to a 200 kb inversion. The structure of the LAL14/1 genome was compared with those of the HVE10/4 and REY15A genomes by dot-plot analyses (see the electronic supplementary material, figure S4). This revealed a well-conserved 2 Mb core region common to all three genomes, with substantial synteny conservation. Many rearrangements were detected in the variable regions of each genome. The localization, length and genetic context of all major differences detected by the dot-plot approach are listed in the electronic supplementary material, table S6. Previous analysis revealed a large inversion of 0.5 Mb as one of the major differences between HVE10/4 and REY15A genomes [7]. Genome sequencing of the LAL14/1 in combination with the phylogenetic analysis (figure 1) allowed us to infer that the inversion occurred in the ancestor of the HVE10/4, while LAL14/1 retained the ancestral genome organization.

The program *progressiveMauve* [41] was used for alignments of the 10 *S. islandicus* genomes and visualization of genome rearrangements (figure 3). All the genomes have a common general organization, with a well-preserved part covering about 75 per cent of the genome and a long variable region. Small strain-specific variable regions are scattered throughout the conserved regions of all of the 10 genomes analysed; many correspond to the insertion of heterologous genes transferred horizontally (see below).

Each genome contains several relatively small specific regions and all of them have a unique long variable region situated in the same segment of the genome (see the

Table 4. Major groups of paralogues in *S. islandicus* LAL14/1 and their representation in 10 *S. islandicus* genomes.

<i>S. islandicus</i> LAL14/1			all <i>S. islandicus</i>	
range	protein no.	protein families	protein no.	range
1	16	IS200/605 families protein OrfB ^a	129	3
2	13	oligopeptide/dipeptide ABC transporter, ATPase subunit ^a	113	4
3	10	IS630 family transposase ^a	67	8
4	9	IS1 family transposase	53	7
5	6	IS110 family transposase ^a	151	1
6	6	3-hydroxyacyl-CoA dehydrogenase NAD-binding protein ^a	78	5
7	5	inosine/uridine nucleoside hydrolase	32	30
8	4	high-affinity nickel-transporter	21	68
9	4	dTDP-4-dehydrorhamnose 3,5-epimerase	18	145

^aIncludes several *nodes*.

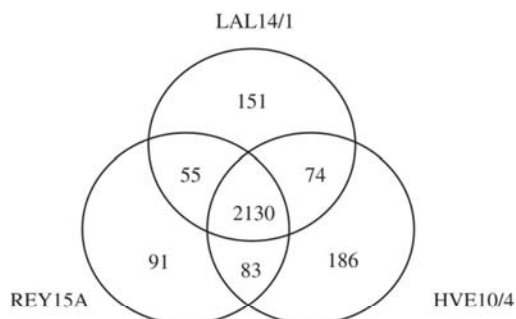


Figure 2. Conservation of protein *nodes* in three closely related *S. islandicus* strains. A total of 2130 families are shared by three strains. Each of the strain is characterized by the presence of a specific set of families: 151 for LAL14/1; 91 for REY15A and 186 for HVE10/4.

electronic supplementary material, table S7). These variable regions range in size from 587 to 802 kb and have very heterogeneous genetic contexts. The largest variable region in *S. islandicus* LAL14/1 is 608 kb (between positions 282 and 890 kb) and represents 24.7 per cent of the genome. It does not contain any known essential genes, such as those for tRNA, rRNA or ribosomal proteins, or the replication origin sites (*ori*). Variable regions are usually preferential sites for integration and accumulation of non-essential genes in the genome [7] and the LAL14/1 genome is not an exception. The variable region of LAL14/1 carries most of its integrative elements, including both functional and inactivated IS, MITEs (miniature inverted-repeat transposable elements), the two largest CAG regions, all of the identified CRISPR/*cas* and *cmr* modules, half of the toxin/antitoxin genes, and many of the putative glycosylase genes (figure 4).

4.6. Analysis of tRNA genes, integration events and horizontal gene transfer

The pattern of tRNA genes in *S. islandicus* LAL14/1 is the same as those in *S. islandicus* REY15A and HVE10/4 [7] and very similar to those in other sequenced *S. islandicus*

strains. The LAL14/1 chromosome carries 45 functional tRNA genes (see the electronic supplementary material, table S8) all located in the conserved regions. Sixteen of the tRNA genes include intron sequences (12–65 nt long) immediately downstream from the anticodon triplet. The genes for the tRNA^{Glu}[CTC] and tRNA^{Glu}[TTC] each have an insertion in their D-loop.

In Sulfolobales, the tRNA genes are preferential sites of integration of conjugative plasmids and fuselloviruses [4,42,43]. The mechanism of integration usually involves site-specific recombination between the tRNA gene target and the integrase gene (*int*) carried by an extrachromosomal element [44]. The presence of remnants of the corresponding *int* gene, overlapping the sequence of the tRNA gene target, often serves as a strong indication for an ancestral integrative event. The sequences of the remnants of the integrative elements are often incomplete or extensively degenerated, making their *in silico* identification challenging.

To identify the potential integrated extrachromosomal elements in the LAL14/1 genome, we set out to locate gene clusters enriched in homologues of proteins encoded by archaeal plasmids and viruses. For this purpose, all LAL14/1 proteins were compared (BLASTp) against the local protein database containing sequences of publicly available archaeal viruses (fifty-five) and crenarchaeal plasmids (twenty-five). Genomic loci containing at least five plasmid/viral homologues per 20 kb region were retained and manually inspected. This approach led to identification of four putative integrated elements. Notably, none of them appears to be functional, as judged from their incomplete gene complements when compared with ‘autonomous’ elements and lack of identifiable attachment sites. Three elements (SiL-E1 [SiL_0398..SiL_0402], SiL-E2 [SiL_1310..SiL_1321], SiL-E3 [SiL_1467..SiL_1481]) are likely to be remnants of conjugative plasmids, while the fourth one (SiL_2367..SiL_2371) is related to SSV-like fuselloviruses. SiL-E2, SiL-E3 and the SSV-like element are located in the proximity of different tRNA genes, which probably served as their respective integration targets, while SiL-E1 was found within the CRISPR_2 locus (see below).

In order to uncover the potentially more ancient integration events at the tRNA genes, which could have eluded identification using the criteria detailed above, we have inspected

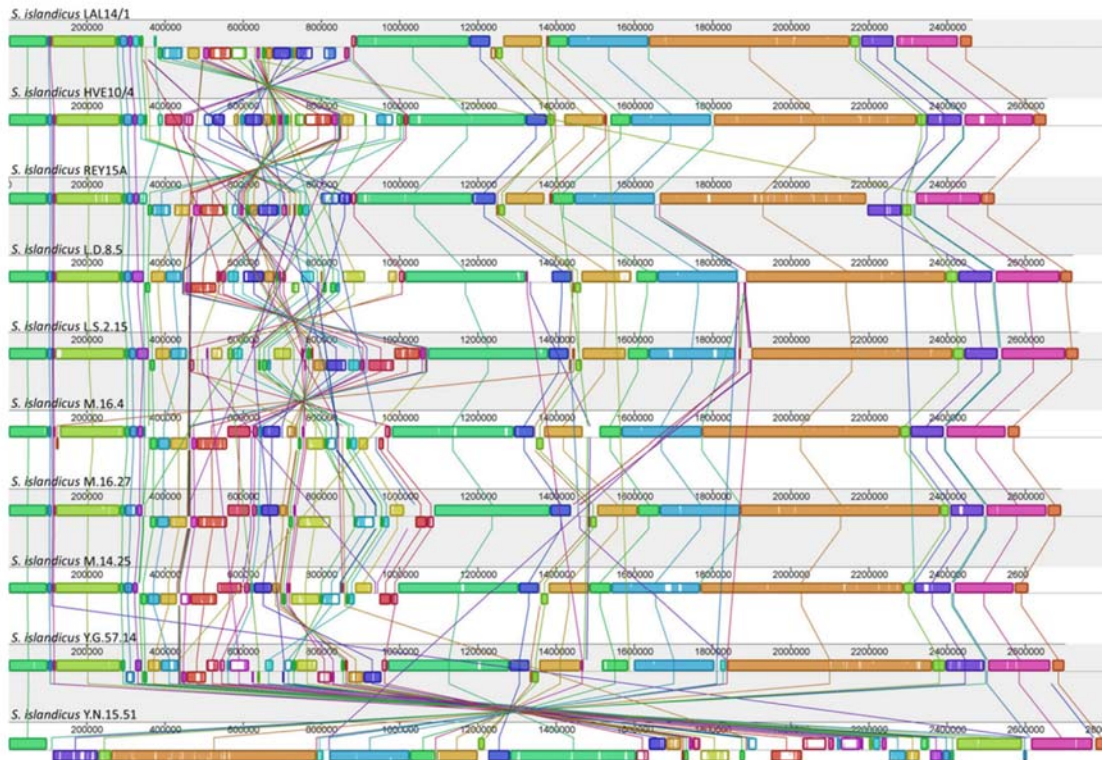


Figure 3. Alignment of 10 *S. islandicus* genomes by *progressiveMauve* approach. The blocs of the same colours indicate the regions of synteny. The unique regions are indicated by the absence of blocs.

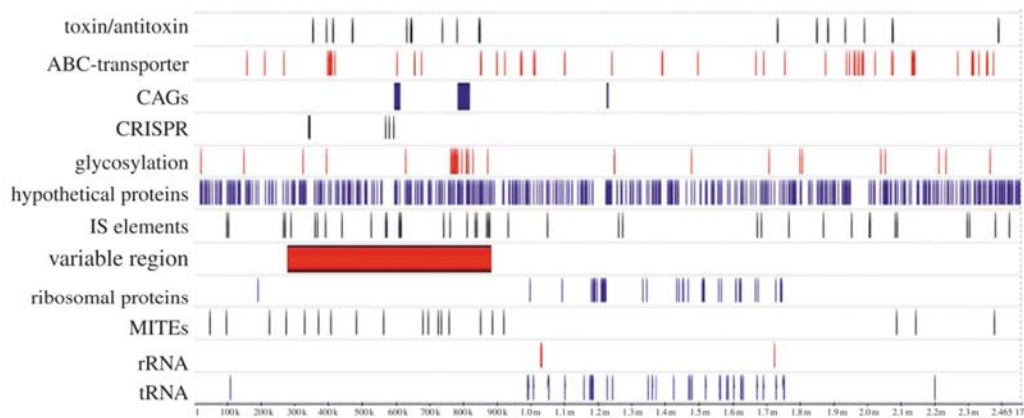


Figure 4. Distribution of some families of genes in the *S. islandicus* LAL14/1 genome compared with the position of the large variable region.

the genomic regions proximal to other LAL14/1 tRNA genes for the presence of plasmid/viral homologues and compared the obtained patterns with those of HVE10/4 and REY15A.

In general, the pattern of insertions linked to the tRNA genes in the LAL14/1 strain is similar, but not identical, to those in REY15A and HVE10/4 (table 5).

Some of the insertions are clearly strain-specific (insertions into the tRNA^{Ala}[CGC], tRNA^{Val}[GAC], tRNA^{Pro}[TGG], tRNA^{Leu}[CAG], tRNA^{Gly}[GCC], tRNA^{Arg}[CCT] and tRNA^{Arg}[GCG] genes). Other are present in two (insertions in tRNA^{Lys}[CTT] and tRNA^{His}[GTG]) or all three strains (tRNA^{Leu}[TAA] and tRNA^{Ser}[GGA] genes). In the case of 7 tRNA genes (tRNA^{Phe}[GAA], tRNA^{Glu}[TTC], tRNA^{Ala}[GGC], tRNA^{Thr}[GGT],

tRNA^{Pro}[GGG], tRNA^{Leu}[TAA] and tRNA^{Ser}[GGA]), the three strains carry nearly identical remnants of the same integrated elements. This suggests that the respective integration events occurred in the common ancestor of REY15A, HVE10/4 and LAL14/1.

Proviruses are common companions of archaeal genomes [45–47]. Thus, it was somewhat surprising not to find potentially functional proviruses in the LAL14/1 genome. The only virus-derived element of LAL14/1 integrated in the tRNA^{Thr}[GGT] gene is a highly degenerated remnant of an SSV-like fusellovirus. Notably, the element does not appear to be closely related to any particular fusellovirus, since different genes display affinities to distinct fuselloviruses.

Table 5. Overview of integration events targeting the tRNA genes in three *S. islandicus* strains, LAL14/1 (present work) and REY15A and HVE10/4 [7]. For the corresponding lanes of the table, the sequences found in three strains are given in footnotes a–c. In the CAG column the main digit gives the number of genes and the upper small digit indicates the CAG score rated from 1 (most atypical) to 10 (not atypical). CAGs written in bold are detailed in table 6.

integration events									
tRNA	intron	REY15A			HVE10/4			LAL14/1	
		BLASTp	CAG	BLASTp	CAG	BLASTp	CAG	BLASTp	CAG
Val-TAC	no	SiRe1242–1247 conj. plasmid	5 ²⁻⁶	SiH1814	—	1 ¹	—	—	—
Phe-GAA ^a	no	SiRe1321–SiRe1323 conj. plasmid	9 ¹⁻⁷	SiH1399–1402 conj. plasmid	—	4 ⁵⁻⁷	SiL1310–1321 SiI-E2 conj. plasmid	—	6 ¹⁻⁵ CAG3
Met-CAT	yes	SiRe1465–SiRe1479	15¹⁻⁷ CAG4	SiH1557–1560	—	4 ²⁻⁷	SiL1467–1481 SiI-E3 conj. plasmid	—	13 ¹⁻⁵
Glu-TTC ^b	no	intN fragment ^c SiRe1484–1490 conj. plasmid	7 ³⁻⁷	intN fragment ^c SiH1561–1574 conj. plasmid	—	12 ¹⁻⁷	intN fragment SiL1484–1485 conj. plasmid	—	5 ¹⁻³
Ala-GCC ^a	no	intN fragment	1 ⁶	intN fragment	—	1 ⁷	intN fragment	—	1 ⁵
Thr-GGT	no	SiRe2413–2417 SSV-like	3 ¹	SiH2464–2472 SSV-like	—	5^{1 to 2} CAG 7	SiL2367–2371 SSV-like	—	3 ¹
Pro-GGG ^a	yes	intN fragment	1 ³	intN fragment	—	1 ³	intN fragment	—	1 ³
His-GTG ^b	no	SiRe1787–1792	12 ¹⁻⁷	SiH1866–SiH1871	—	5 ¹⁻⁶	—	—	—
Leu-TAA ^a	yes	SiRe1255; SiRe1257	2 ^{5 and 4}	SiH1333; SiH1335	—	2 ^{5 and 4}	SiL1246–SiL1248	—	3 ²⁻⁷
Arg-GCG	no	—	—	SiH1544 transposase	—	1 ⁵	—	—	—
Ser-GGA ^a	no	SiRe1778–1779	2 ^{3 and 5}	SiH1858–1859	—	2 ^{3 and 5}	SiL1773–SiL1774	—	2 ^{3 and 5}
Lys-CTT ^b	yes	—	—	SiH1917 transposase	—	1 ¹	SiL1826; SiL1828	—	2 ^{2 and 4}
Ala-CGC	no	SiRe1038 hyp. protein	1 ¹	—	—	—	—	—	—
Val-GAC	no	SiRe1734 hyp. protein	1 ¹	—	—	—	—	—	—
Pro-TGG	no	SiRe1936–1939 transposase and hyp proteins	4 ¹⁻⁵	—	—	—	—	—	—
Leu-CAG	no	—	—	—	—	—	—	—	1 ¹
Gly-GCC	no	—	—	—	—	—	—	—	1 ¹
Arg-CCT	yes	—	—	—	—	—	—	—	1 ¹

^aNearly identical.

^bDistantly related sequences.

^cNot mentioned by [7].

Table 6. Major CAG regions and corresponding predicted functions in *S. islandicus* LAL14/1, HVE10/4 and REY15A.

strain/CAG	position	no. of atypical genes ^a	description
<i>S. islandicus</i> HVE10/4			
CAG 1	497317–519235	15 (5)	transposase IS200/IS605
CAG2	554497–563952	8 (4)	gene <i>orfB</i>
CAG3	725905–750622	17 (4)	genes <i>vapBC</i> , CRISPR_3 of family III and six genes <i>cas</i>
CAG4-1	895122–912372	20 (6)	genes of hydrogenases and ABC transporter; HVE10/4 specific
CAG4-2	921672–938729	16 (9)	genes of hydrogenases; HVE10/4 specific
CAG5	967655–977019	11 (9)	HVE10/4 specific
CAG6	1399248–1420774	14 (7)	restriction–modification system of type I found HVE10/4 specific
CAG7	2380543–2386271	6 (6)	insertion in tRNA[Thr] ^{GGT}
<i>S. islandicus</i> REY15A			
CAG 1	555326–562352	7 (4)	partially similar to CAG1 of LAL14/1 and to CAG3 of HVE10/4; <i>vapBC</i>
CAG2	722610–725192	5 (4)	REY15A specific
CAG3-1	790437–804374	17 (13)	transposase IS5
CAG3-2	829169–837249	6 (4)	glycosyl transferase gene
CAG3-3	845351–852630	7 (1)	IS200/605; <i>vapBC</i> ; genes <i>cas</i>
CAG4	1372809–1383519	12 (6)	insertion in tRNA[Met] ^{CAT} ; <i>vapBC</i>
<i>S. islandicus</i> LAL14/1			
CAG1	600528–615132	10 (4)	<i>vapBC</i> ; gene <i>csm6-like</i>
CAG2	789437–821915	26 (16)	genes of methyltransferases and glycosyltransferases
CAG3	1232632–1235126	7 (5)	inserts in tRNA[Phe] ^{GAA}

^aThe first digit indicates the total number of genes in the CAG region and the digit shown in the parenthesis indicates those for which the function could not be predicted.

To gain an insight into the timeframe of this viral integration event, we analysed the equivalent loci in all available *S. islandicus* genomes. The traces of SSV integration were found in all *S. islandicus* strains, except for the M.16.27, which contained a gene for the pNOB8-type integrase at the equivalent position [48]. Interestingly, in all cases the elements were severely degenerated; a selection of genomic alignments can be found in the electronic supplementary material, figure S5. The most parsimonious scenario for the observed distribution of SSV-like remnants in *S. islandicus* genomes involves a single event of SSV-like virus genome integration into the tRNA^{Thr}[GGT] gene, followed by gradual deterioration of the provirus along the evolutionary history of *S. islandicus* species. The integration has probably occurred following the divergence of *S. islandicus* and *S. solfataricus* from their common ancestor, since *S. solfataricus* lacks a detectable SSV-like element at the equivalent genomic locus.

Identification of insertions by BLASTp analysis may be hampered by the insufficient conservation of the inserted genes or by limited coverage of the diversity of archaeal mobile genetic elements. Furthermore, some of the insertions could occur in loci other than the tRNA genes. To overcome these caveats, we have applied a BLASTp-independent approach based on the search of CAGs [17,49]. Following this approach, the putative integrated elements could be identified following their atypical codon usage compared with that of the conserved part of the host chromosome. For LAL14/1, HVE10/4 and REY15A the results obtained by this approach

are summarized in table 6, and a brief general comparison of CAG distribution in 10 *S. islandicus* genomes is present in the electronic supplementary material, table S9. Notably, nearly all insertions detected in LAL14/1, HVE10/4 and REY15A by the BLASTp analyses were confirmed by the CAG approach; in addition, some of the integrative events were only predicted by the CAG search. LAL14/1 has three CAG regions (CAG1–3) of 14.6, 32.5 and 2.5 kb that carry 43 genes with atypical codon usage. CAG3 was found to correspond to SiL-E2 element integrated into tRNA^{Phe}[GAA] identified by BLASTp analysis, while CAG1 and CAG2 could not be predicted by other approaches.

Some of the functions identified as being associated with the CAG loci are: the restriction–modification system I characteristic of HVE10/4; some elements of CRISPR-based immunity in HVE10/4 and LAL14/1; and various enzyme families (methyl- and glycosyltransferases, hydrogenases). The genes transferred horizontally and integrated into the chromosomes of these *S. islandicus* strains include many transposons and toxin/antitoxin gene pairs of the *vapBC* family.

To summarize, LAL14/1 carries the remnants of 13 insertion events into the tRNA genes, and three additional elements (SiL-E1, CAG1 and CAG2) are integrated into other loci; the same or a similar number of insertions is found in HVE10/4 and REY15A. These results further illustrate the fact that tRNA genes are frequently attacked by various mobile genetic elements. The observation that all (or at least the majority) of the integrated elements appear to be

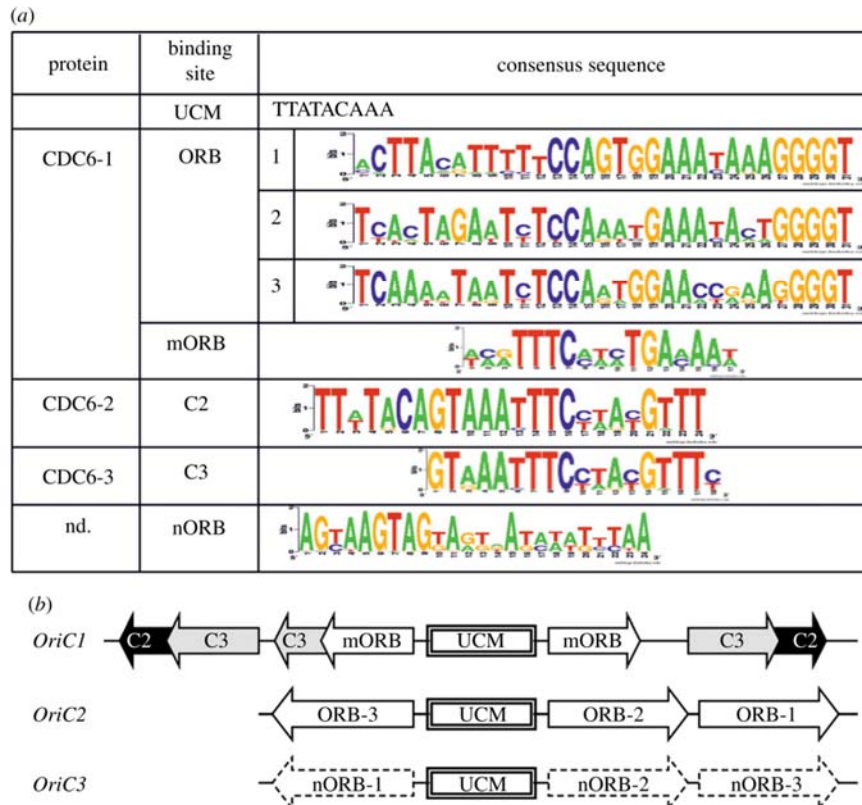


Figure 5. Conserved sequences and structural organization of replication origins *oriC* in *S. islandicus*. (a) Consensus DNA sequences present in *oriC* sites. The size of the letters is directly proportional to the residue conservation. (b) Structural organization of three *oriC* in *S. islandicus*.

non-functional suggests that LAL14/1 possesses an efficient mechanism of purging its genome of unwelcomed insertions.

4.7. Replication origins, *oriC*

The positions of the replication origins in the chromosome of LAL14/1 were predicted by two independent approaches, Z-curve [50,51] and ACCA-plot [52], that produced very similar results (see the electronic supplementary material, figure S6). Consistent with all other Sulfolobales genomes analysed *in silico* [7,53] or *in vivo* [54–56], three *oriC* origins of replication were detected. Their positions and genomic contexts are well conserved with respect to other *Sulfolobus* genomes (see the electronic supplementary material, table S10 and [7]). The *oriC1* site (mapped at position 1.59 Mb) is linked to the *cdc6-1* gene (SiL_0002), *oriC2* (position 800 bp) is linked to the *cdc6-3* gene (SiL_1733), and *oriC3* (position 1.15 Mb) to the *cdc6-2/whiP* genes (SiL_1228/SiL_1206).

The structures of the *oriC1* and *oriC2* sites in *S. solfataricus* P2 are well characterized [56,57], and these two replication origins are organized similarly in *S. islandicus* LAL14/1 (figure 5). Their central AT-rich UCM sequence (uncharacterized motif) is surrounded by the characteristic ORB sequences (origin recognition box). The *oriC1* site also contains additional specific palindromic sequences, called C2 and C3, that are recognized by the replication initiation proteins Cdc6-1, Cdc6-2 and Cdc6-3 [56].

In Sulfolobales, the *oriC3* site is usually linked to the *whiP* gene. A comparative analysis of *oriC3* in the 10 *S. islandicus* strains and in *S. solfataricus* P2 provided new insights into the

organization of this region: we identified three conserved ORB-like sites (nORB) upstream and downstream from the typical UCM site (figure 5). The *oriC2* and *oriC3* sites are organized similarly and, unlike *oriC1*, do not contain the C sequences.

4.8. Toxin–antitoxin systems

A family II (VapBC) toxin–antitoxin (TA) system is present in many Archaea and is very abundant in Sulfolobales [7,58,59]. All *S. islandicus* strains carry many TA gene pairs of the VapBC family as well as genes of another family considered to play a TA role [60]: HEPN-NT (*Higher Eukaryotes and Prokaryotes Nucleotide-binding-Nucleotidyl Transferase*; electronic supplementary material, table S11).

Eight of the 15 *vapBC* gene pairs in *S. islandicus* LAL14/1 map in the variable region of the genome. The other seven *vapBC* gene pairs map in the conserved part of the genome and all share a similar genetic context in the three strains analysed. For four of these loci (SiL_2040/2041, SiL_2042/2043, SiL_2080/2081, SiL_2253/2254), the genomic context is particularly well preserved. All *vapBC* loci, except SiL_2575/2576, are flanked by degenerated copies of *IS* elements, probably involved in the transposition of *vapBC*.

As observed in HVE10/4 and REY15A [7], the *vapB* (toxin) and *vapC* (antitoxin) genes of different subtypes were found in *S. islandicus* LAL14/1 in various combinations giving different variants of the *vapBC* operon (data not shown). This combinatorial diversity of *vapBC* gene pairs may indicate the existence of several types of the toxin/antitoxin mechanisms. All *vapB* and *vapC* gene combinations found in *S. islandicus* LAL14/1

are also found in HVE10/4 and REY15A, except for SiL_0413/0414 and SiL_0631/0632 combinations present in LAL14/1 and HVE10/4 but not in REY15A.

Members of the toxin/antitoxin family HEPN-NT were detected in all *S. islandicus* strains, with multiple copies of the corresponding genes (see the electronic supplementary material, table S12). Unlike the *vapBC* system, HEPN-NT gene pairs are stable and each HEPN gene type is strictly associated with its specific NT gene type. HEPN-NT operons are classified into two subfamilies, I and II. Subfamily I is ubiquitous and all of its representatives in the 10 *S. islandicus* genomes analysed both occupy the same genetic regions and are always localized in conserved parts of the genomes. Subfamily II is much more diverse. Its representatives in *S. islandicus* map in both conserved and variable regions of the chromosome. Note that many copies of HEPN-HT family II pairs include only truncated forms of the HEPN gene and are not functional.

The production of sulfolobocins, a type of toxin that inhibits the growth of sensitive *Sulfolobus* strains, is characteristic of two other well-studied Sulfolobales, *S. acidocaldarius* and *S. tokodaii* [61–63]. No sulfolobocin-encoding genes, such as *sulA*, *sulB* and *sulC*, were found in any of the 10 *S. islandicus* genomes, indicating the absence of this toxin system from these species. Nevertheless, a truncated copy of the *sulA* gene, which obviously cannot code for a functional toxin, is present in *S. islandicus* REY15A [61].

4.9. UV-inducible type IV pili

Many Sulfolobales (ex. *S. solfataricus*, *S. tokodaii* and *S. acidocaldarius*) code for a UV-inducible type IV pilus system that promotes cellular aggregation and efficient exchange of chromosomal markers [64,65]. The formation of pili is controlled by the UV-inducible *ups* operon which comprises five genes: *upsX*, *upsE*, *upsF*, *upsA* and *upsB* [4–6]. This operon is present in all 10 *S. islandicus* strains (and all other sequenced species of Sulfolobales, electronic supplementary material, table S13) and is in all cases in the conserved part of the genome. The Ups proteins encoded by all *S. islandicus* are very similar and form a specific phylogenetic group within the Ups family in Sulfolobales (see the electronic supplementary material, figure S7).

The *in silico* data strongly suggest that the *ups* locus of *S. islandicus* is functional *in vivo*, and it very probably plays the same biological role as in *S. solfataricus*, *S. tokodaii* and *S. acidocaldarius*.

4.10. Insertion sequence elements and miniature inverted-repeat transposable elements

Sulfolobus islandicus LAL14/1, as HVE10/4 and REY15A, contains several families of IS elements with members present in multiple copies [7] (see the electronic supplementary material, table S14).

The *orfB*-containing IS (families IS605 and IS200/605) considered to be ancestral for the archaeal domain [66,67] are overrepresented in all three *S. islandicus* strains analysed: 95% of the copies of the *orfB* gene not linked to *orfA* in HVE10/4, REY15 and LAL14/1 were predicted to be functional.

Only seven of the 53 structurally valid IS (IS with intact inverted terminal repeats (ITRs)) in the LAL14/1 genome are predicted to code for functional transposases; transposase

genes in the remaining 46 IS are truncated. The IS patterns of the two other strains, HVE10/4 and REY15A, are very different. Most of the IS in these genomes (45/76 in HVE10/4 and 52/96 in REY15A) code for a full-length transposase and are therefore predicted to be functional. However, many of the mutated IS may be mobilized by transposases of the same family encoded in *trans* [68]. Thus, 62 of the 76 IS (81.6%) detected in HVE10/4 and 87 of the 96 IS (90.6%) in REY15A could, in theory, be mobile (see the electronic supplementary material, table S14); the proportion is lower for LAL14/1, for which 31 of the 53 IS (58.4%) are potentially active.

This may indicate greater genetic stability of *S. islandicus* LAL14/1 than of either HVE10/4 or REY15A. Were this the case, strain LAL14/1 would be the most attractive model for genetic manipulations.

Another group of mobile elements in *S. islandicus* is MITEs, believed to correspond to truncated derivatives of autonomous DNA transposons [69–73]. MITEs exhibit the structural features of DNA transposons, containing terminal inverted repeats flanked by small direct repeats. The internal sequences of MITEs are short and devoid of ORFs. As non-autonomous elements, the transposition of MITEs is totally dependent on trans-acting transposases [68,74,75]. Only two classes of MITEs, SMN1 (320 bp) and SM3A (164 bp), were detected in the 10 *S. islandicus* genomes (see the electronic supplementary material, table S15). The SM3A family is more numerous in LAL14/1 than any of the other *S. islandicus* strains. The three *S. islandicus* strains from Iceland share two identical SM3A, but LAL14/1 also carries nine extra copies of SM3A that share only 95% similarity with other two SM3A copies. SMN1 transposition is dependent on the presence of a functional ISC1733 transposase and SM3A transposition on ISC1058 [68,76,77]. The MITEs of the SMN1 type in LAL14/1, HVE10/4 and REY15A could be mobilized by the ISC1733 type transposase [76] predicted to be functional in these strains. The observation of the mobilization of SMN1 in *S. islandicus* REN1H1 is consistent with this prediction [76]. None of the three *S. islandicus* strains analysed codes for a functional ISC1058 transposase, suggesting that SM3A, although present, cannot transpose in HVE10/4, REY15A and LAL14/1.

4.11. CRISPRs: structure, targets and phenotype

All sequences and genes related to the CRISPR system present in *S. islandicus* LAL14/1 (CRISPR arrays, *cas* and *cmr* gene cassettes) map within the large variable genomic region. It carries five CRISPR loci, three *cas* gene cassettes associated with subtype I-A and two *cmr* gene cassettes associated with subtype III-B [78,79]. Following the leader and repeat sequence compositions, the five CRISPRs of *S. islandicus* LAL14/1 could be divided into two families, I and III [36], while the two *cmr* modules belong to families B and F ([37]; figure 6 and table 7).

The family I CRISPR locus comprises two oppositely oriented blocks of repeat-spacer arrays separated by the first module of Cas genes. The second module is situated at the end of one of the repeat-spacer arrays (figure 6a). The additional Cmr module is not linked to this part and maps several hundred kilobases downstream in the genome. The family III CRISPR (figure 6a) comprises three clusters of spacers/repeats associated with two gene modules, one including the *cas* gene and another *cmr*. The *cmr* gene order in this module is the same as that in the Cmr module associated with the family I CRISPR.

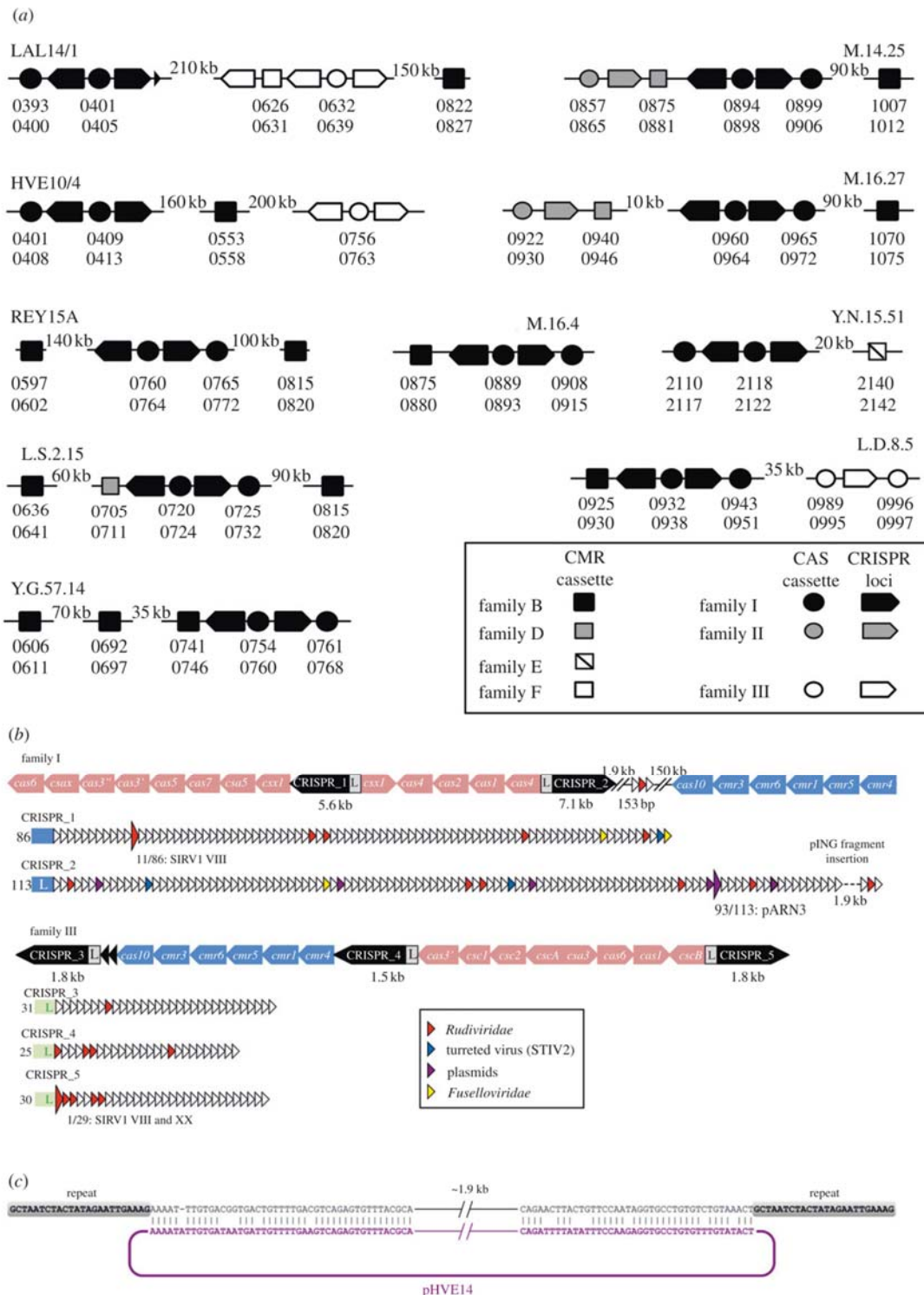


Figure 6. CRISPR organization in *S. islandicus* LAL14/1. (a) Comparison of the CRISPR structures and composition in 10 *S. islandicus* strains. (b) Structure and composition of CRISPRs in *S. islandicus* LAL14/1. Large-scale presentation: the orientation, position and size of CRISPR arrays (black arrows) as well as *cas* (red arrows) and *cmr* (blue arrows) gene modules are indicated. Detailed presentation: all perfectly matching spacers (large arrows) as well as selected imperfect spacers are indicated by small coloured arrows specified in the legend included in the figure body. (c) The pING-like insertion in the CRISPR_2 locus is interspersed precisely between two complete 24 bp repeats, resembling by its position typical CRISPR spacers.

The analysis of 285 spacers forming the CRISPR array of LAL14/1 revealed the presence of a surprisingly high number of spacers that perfectly (in three cases) or imperfectly

(30 cases; electronic supplementary material, table S16) match the genomes of rudiviruses, fuselloviruses and conjugative plasmids previously described in the Icelandic hot spring

Table 7. Composition of CRISPRs in *S. islandicus* LAL14/1 and their putative targets. SIRV1 VIII and SIRV1 XX are different subtypes of the virus SIRV1 [34].

CRISPRs and their families	repetition	position and direction	no. of spacers	spacers with 100% identity to the indicated putative targets
CRISPR_1 I	GCTAATCTACTATAGAATTGAAAG	344177..349768 ←L	86	11/86 SIRV1 VIII
CRISPR_2 I	GCTAATCTACTATAGAATTGAAAG	353820..363037 L→	113	93/113 pARN3
CRISPR_3 III	GTAACAACACAAAGAACTAAAAC	573686..575526 ←L	31	—
CRISPR_4 III	GTAACAACACAAAGAACTAAAAC	584662..586213 ←L	25	—
CRISPR_5 III	GTAACAACACAAAGAACTAAAAC	597345..599148 L→	30	1/29 SIRV1 VIII and SIRV1 XX

environments [34,35]. The two perfectly matching spacers carried by CRISPR_1 and CRISPR_5 target the genome of the ruidivirus SIRV1 [80]. Interestingly, the spacer in CRISPR_1 matches a SIRV1 gene encoding a protein, P98, responsible for formation of pyramidal structures involved in virion egress [11–13]. No spacers with 100 per cent identity to a closely related virus, SIRV2, were detected. The third perfectly matching spacer is identical to a sequence in the conjugative plasmid pARN3 [81]. Unexpectedly, we have identified an about 2 kb insertion, SiL-E1, in the CRISPR_2 locus. SiL-E1 resembles the typical CRISPR spacers in that it is interspersed between two identical repeats and is followed by additional spacer-repeat units (figure 6). This pseudo-spacer encompasses five ORFs and displays high sequence similarity to and collinearity with pING1-like conjugative plasmids of *S. islandicus* [81–83]. More specifically, SiL-E1 shares overall 82 per cent identity with plasmids pING1 [83] and pHVE14 [81]. SiL-E1 is not present in other *S. islandicus* strains. Notably, sequence similarity between SiL-E1 and pING1-like plasmids extends throughout the length of SiL-E1, leaving no unaccounted positions between the inserted sequence and the repeat regions (figure 6c). This suggests that SiL-E1 is unlikely to be a result of illegitimate recombination.

Some of the spacers imperfectly match DNA regions present in the genomes of other *S. islandicus* strains (M.14.25, M.16.4, Y.N.15.51, Y.G.57.14, L.D.8.5, L.S.2.15) and in *S. solfataricus* 98/2 and P2. The functions of the corresponding genes are unknown, and even the biological significance of this observation is unclear. Possibly, the genomic loci matched by these spacers represent the remnants of unknown viruses or plasmids integrated into the corresponding genomes.

The analysis of the protospacer corresponding to the spacers listed in the electronic supplementary material, table S15 allows the identification of the PAM sequence (protospacer adjacent motif). These sequences situated at the proximity of protospacers are crucial for two essential steps of CRISPR-based immunity: adaptation [84] and interference [37,85,86]. They also play an important role in the mechanism of target discrimination that prevents the recognition of chromosomal spacers as valid targets [87].

For the LAL14/1 CRISPRs of the family I, we found the same PAM motif, CC, in the position (−3, −2) at the 5′ end as was already described by Gudbergstottir *et al.* [88]

(figure 7a). No specific PAM motif was detected at the 3′ end of these protospacers (figure 7b). Little is known about the protospacers corresponding to the CRISPRs of the family III. Our data indicate the existence of a conserved motif [T/A]GT occupying the position (−4, −3, −2) at the 5′ end of the protospacer (figure 7c). The 3′ end of these protospacers is very rich in A/T nucleotides (figure 7d).

4.12. Development of a genetic model

Sulfolobus islandicus LAL14/1 is a promising model for studying virus–host interaction in Archaea. LAL14/1 cells can be infected by SIRV2, a model rod-shaped virus [89] that codes for a unique mechanism of virion release: pyramidal structures form on the host cell surface, breaking the S-layer and allowing the virions to escape from the cells [11–13,90]. Investigations on SIRV2 cycle regulation and SIRV2–host interaction will require genetic tools for strain LAL14/1 as no available genetic models of *Sulfolobus* can be infected by SIRV2.

One of the most commonly used genetic markers in Archaea is the *pyrEF* operon. Pyrimidine prototrophs (Pyr+) can be easily selected, on minimal medium without uracil, after transformation of a *pyrEF*–host strain by a plasmid or viral vector carrying the wild-type *pyrEF* operon [19,91–94]. A non-reversing spontaneous *pyrEF*–deletion mutant of *S. islandicus* REY15A is widely used as a host for genetic manipulations [20]. No such mutant of *S. islandicus* LAL14/1 was available, so we constructed a $\Delta pyrEF$ mutant via allelic replacement approach.

To generate a *pyrEF* disruption mutant, a knockout cassette containing the $\Delta pyrEF$ allele from *S. islandicus* REY15A, strain E233S [19] and 1 kb regions situated downstream and upstream from *pyrEF* was obtained by PCR amplification (see §3). The *S. islandicus* LAL14/1 cells were transformed by this linear DNA fragment of 2233 bp and the $\Delta pyrEF$ mutants resulting from the replacement of the wild-type copy of the *pyrEF* operon on the host chromosome by a double cross-over were selected on 5′FOA (5′-fluoro-orotic acid). Twenty transformants were selected and the *pyrEF* operon was analysed by PCR and sequencing; 15 of the analysed colonies carried the expected $\Delta pyrEF$ deletion. This mutant strain, called *S. islandicus* LAL14/1-CD, showed the same virus

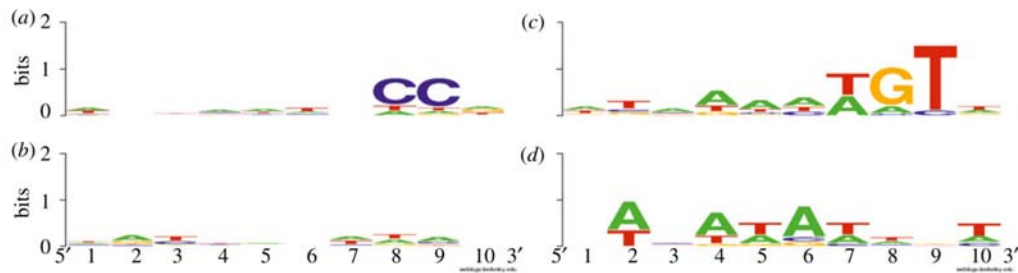


Figure 7. Conserved motifs in the protospacer sequences for CRISPRs I ((a) 5' and (b) 3') and CRISPRs III ((c) 5' and (d) 3').

resistance/sensitivity phenotype as the parental strain (data not shown). We confirmed that this mutant is indeed derived from strain LAL14/1 by sequencing the gene coding for the A subunit of the cytochrome b558/566 (Sil_2350; the sequence of this gene is not identical in REY15A, HV10/4 and LAL14/1).

Sulfolobus islandicus LAL14/1-CD could be efficiently transformed with the pHZZ2 (pRN2 replicon) [19] autonomously replicating in *S. islandicus* and carrying a wild-type copy of the *pyrEF* operon. The transformation efficiency was 10^2 – 10^3 colonies/ μ g of DNA.

A powerful genetic *pop-in/pop-out* approach was previously developed for another genetic model, *S. islandicus* REY15A [19]. It allows rapid and efficient creation of knock-out mutants. To show that this approach is efficient in LAL14/1, we have chosen to delete one of the CRISPR loci (CRISPR_1), because CRISPR-coded functions are usually not essential for the cells in the absence of viruses and their deletion mutants are expected to be viable. Also, one of the spacers of the CRISPR_1 matches perfectly the SIRV1 virus for which LAL14/1 is resistant. If the resistance is linked to the CRISPR activity, its inactivation could decrease the level of resistance giving a detectable phenotype to this mutant.

The recombinant plasmid used to inactivate the CRISPR_1 and the positions of the regions IN (867 bp), OUT (919 bp) and TARGET (776 bp) in the vector pSEF described by Deng *et al.* [19] are indicated in figure 8.

Two successive rounds of recombination (figure 8) deleted the chromosomal fragment situated between the IN and OUT regions producing a *pyrEF*⁺ derivative from which the CRISPR_1/*cas* region has been deleted (Δ CRISPR_1/ Δ *csx1* Δ *cas4*; *csx1* is annotated as Sil_0393). The deletion was confirmed by PCR analysis (data not shown). Interestingly, Δ CRISPR_1/ Δ *csx1* Δ *cas4* was as resistant to infection by SIRV1 as *S. islandicus* LAL14/1-CD. The presence of a second spacer and functional CRISPR_3 may explain this result.

The observed efficient transformation of LAL14/1 as well as the ease of creation and selection of its deletion mutants indicate that LAL14/1 represents an excellent genetic model.

5. Discussion

Sulfolobus islandicus LAL14/1 is a promising model for studies on virus–host interaction and CRISPR/*cas*-based acquired immunity in hyperthermophilic Archaea. We report an extensive comparative *in silico* analysis of its genome and established this strain as genetic model.

The genome of LAL14/1 is the 10th of the species *S. islandicus* to be sequenced [9], and the third of an *S. islandicus*

strain isolated in Iceland [8]. Strain LAL14/1 has the smallest known *S. islandicus* genome: it has only 2601 genes carried by a 2.47 Mb chromosome. With other sequenced *S. islandicus* strains LAL14/1 shares the same major groups of paralogous genes, most of which are transposases of various families (table 4). Its genome also encodes a large number of diverse ABC transporters, including the oligopeptide transporter. This is consistent with the high frequency of isolation of *S. islandicus* strains from enrichment cultures using rich organic media [7]. Indeed, *S. islandicus* LAL14/1 grows heterotrophically on standard laboratory liquid media with yeast extract as main carbon source.

The *S. islandicus* pan-genome contains 570 singletons representing a strain-specific set of proteins; 65 are only found in *S. islandicus* LAL14/1. As has been widely documented for other prokaryotic virus/host models (see review [95]), it is possible that some of the particular features of LAL14/1, for example, virus–host range, could be linked to the presence of specific genes or gene repertoires absent from other *S. islandicus* strains. A large proportion of the strain-specific genes map in a large variable region. In each of the *S. islandicus* genomes analysed this region covers more than 25 per cent of the chromosome length, and contains most of the transposons promoting the horizontal gene transfer and all CRISPR sequences (figure 4).

Our *in silico* analysis of the relics of mobile elements in the genomes of the three closely related *S. islandicus* strains confirms that mobile genetic elements preferentially integrate into tRNA genes. Nevertheless, searches for CAG revealed several previously undescribed long DNA segments of heterologous origin that were located in loci other than tRNA genes [49]. These clusters, most probably the consequences of genetic transfer via conjugative plasmids or other mobile genetic elements, contain 1.6 per cent of the genes in LAL14/1. Biological functions can be predicted for only a small fraction of these genes. For example, the CAGs carry several copies of *vapBC* genes of toxin/antitoxin systems, some CRISPR-related genes and genes coding for methyl- and glycosyltransferases.

A comparative analysis of the genome structure and composition of three closely related strains (HIVE10/4, REY15A and LAL14/1) confirms that they have a very similar genomic pattern with a strong conservation of synteny. Nevertheless, the presence in each of these genomes of multiple local rearrangements raised the issue of the stability of the LAL14/1 genome. All three strains carry many copies of IS elements of various families. The transposition of an IS or transposon is an important source of genome instability and rearrangements in any cell [96]. Such instability is well documented in the case of REY15A, for which a relatively

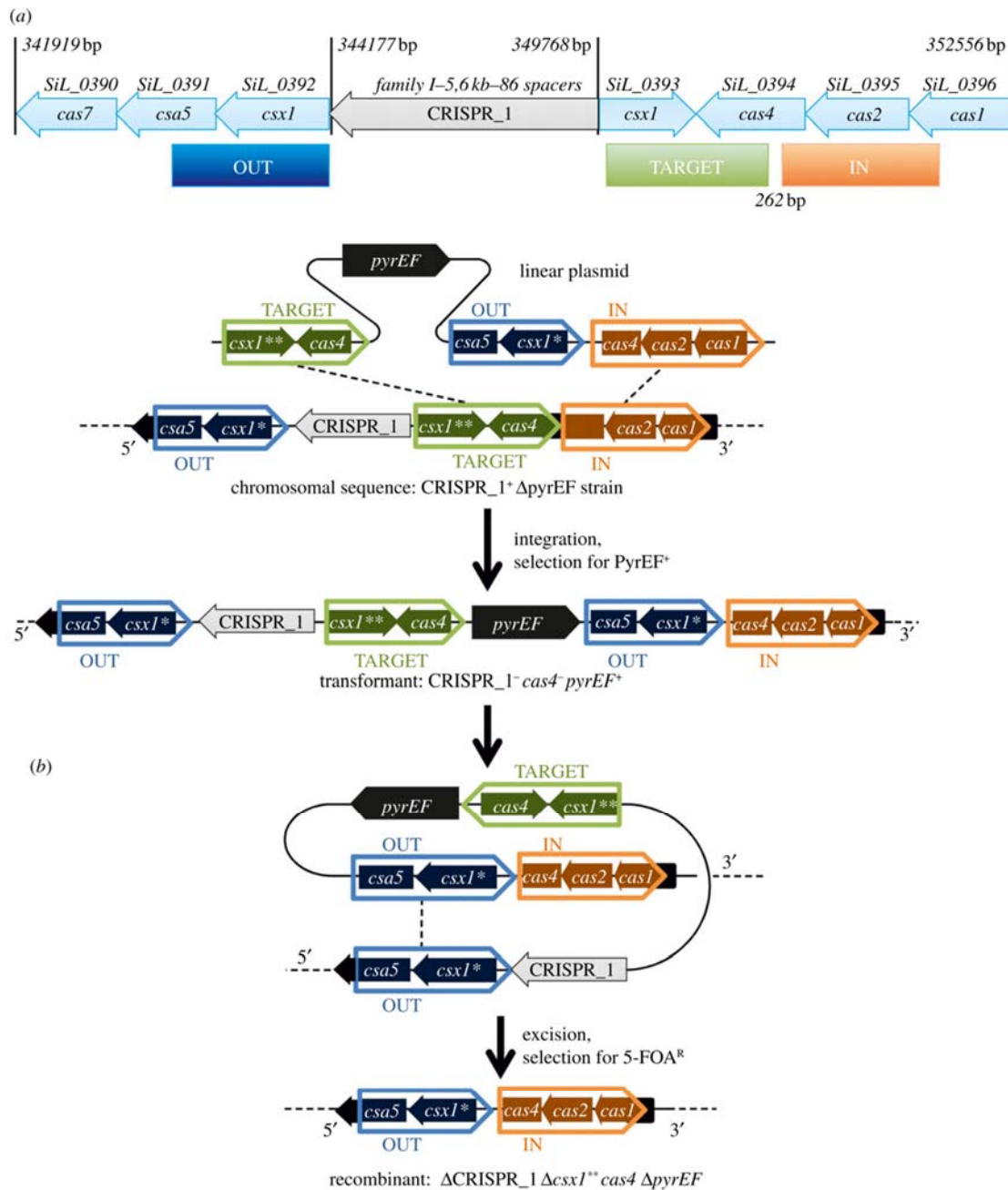


Figure 8. Genetic map of the CRISPR_1 region deleted by the pop-in/pop-out approach. (a) Genetic map of the CRISPR_1 region. The deleted region is situated between the regions OUT and TARGET. (b) A scheme representing two stages of recombination events generating the mutant Δ CRISPR_1 Δ csx1^{**} Δ cas4 Δ pyrEF. Two paralogues of *csx1* are present in this region. The gene *csx1** corresponds to the gene *SiL_0392* and *csx1*** to *SiL_0393*.

high incidence of the *pyrEF*-deletion mutants (one from 50 analysed *PyrEF*-colonies) is observed [7,19].

Sulfolobus islandicus LAL14/1 seems to be genetically more stable as no deletions in the *pyrEF* locus were detected by PCR analysis of 100 colonies of spontaneous LAL14/1 *pyrEF* mutants resistant to FOA (C. Jaubert, C. Danioux, G. Sezonov 2013, unpublished data). This could be due to a lower transposition activity in LAL14/1 and consequently lower frequency of genome rearrangements.

Thus, *in vivo* and *in silico* indications concerning the stability of the genome of strain LAL14/1 suggest that it would

be a useful model for genetic studies. Strain LAL14/1 is the host of the model rudivirus SIRV2 [10,89,97] and has been used to study virus-host interactions in Archaea [11–13]. The availability of the sequence of the LAL14/1 genome makes global genomic analysis of the interaction between viral and host genomes during the infection cycle possible. A genetic approach would facilitate investigations of the role of particular host genes involved in this interaction, as well as the host immune response dependent on the activity of CRISPRs. We successfully inactivated two genetic loci in LAL14/1 of different sizes, *pyrEF* (2.2 kb) and CRISPR_1/

cas (6.6 kb), using allelic replacement and in-frame markerless genetic exchange approaches. We thereby demonstrated the potential of LAL14/1 for genetic experimentation.

More than 90 per cent of the archaeal genomes analysed code for an adaptive immunity system called the CRISPR/*cas* system [98–100]. In LAL14/1, CRISPRs are represented by five loci associated with three *cas* (subtype I-A) and two *cmr* (subtype III-B) gene cassettes. The presence in the CRISPR arrays of LAL14/1 spacers matching extrachromosomal elements (SIRV1 virus, pARN1) make this strain an interesting model for studying the biological and functional role of CRISPRs in Archaea. The presence of two spacers perfectly matching the genome of SIRV1 virus allows speculation about a connexion between the CRISPR composition and the SIRV1-resistance phenotype of *S. islandicus* LAL14/1. Deletion of the CRISPR_1 carrying one of two SIRV1-specific spacers did not change the phenotype of the obtained mutant; it remained as resistant to SIRV1 as the initial strain. A construction of a double mutant Δ CRISPR_1 Δ CRISPR_5 will help to better characterize the eventual involvement of CRISPR generated immunity in resistance of LAL14/1 to SIRV1.

Recent publications report spacer acquisition under laboratory conditions in several bacterial [84,101–104] and archaeal (*S. solfataricus*) [105] models. However, analysis of the CRISPR content of 12 independent SIRV2-resistant mutants of LAL14/1 did not detect any new insertions in the CRISPR sequences (data not shown). Consequently, the acquired resistance does not appear to be related to CRISPRs and presumably involves a different mechanism of resistance [106].

LAL14/1 also carries a unique pseudo-spacer, SiL-E1, which represents an insertion of an approximately 2 kb region from a pING1-like plasmid into the CRISPR_2 array. To our knowledge, such large spacers have not been previously described in archaeal or bacterial CRISPR loci. The acquisition mechanism of this pseudo-spacer as well as its role in LAL14/1 immunity against conjugative plasmids is

unclear. However, the fact that SiL-E1 is flanked by perfect repeats of CRISPR_2 (figure 6c) argues against the possibility of a random integration event. Plausible acquisition scenarios include faulty protospacer processing by the Cas machinery or homologous recombination between the episomal plasmid and the pre-existing CRISPR_2 spacer(s) matching the plasmid. Future studies should provide important additional information regarding spacer acquisition mechanisms and reveal whether such atypical spacers are competent in conferring immunity against mobile genetic elements in Archaea.

Physical isolation of the geothermal hot spring in which *S. islandicus* thrives makes this species a very valuable model to study microbial speciation and evolution [107]. Such studies were recently conducted on *S. islandicus* strains isolated from the hot spring located in Russia ('M') and USA ('Y/N') [9,33], and have already provided important insights into the population dynamics of hyperthermophilic Archaea. Our in-depth comparative genomics analysis clearly indicates divergence of the 'Icelandic trinity'—LAL14/1, HVE10/4 and REY15A—from the groups 'M' and 'L/Y', supporting previously suggested biogeographical patterns of differentiation of *S. islandicus* species. Genetic tools developed in this study and those available for REY15A will help to experimentally tackle questions regarding the evolution and divergence of these Icelandic strains and compare the elucidated patterns with those available for *S. islandicus* strains isolated from other continents.

6. Acknowledgements

This work was supported by PhD fellowships from the 'Ministère de l'enseignement supérieur et de la recherche' (C.J. and C.D.) and by Pasteur-Weizmann (C.J.) allocations. We thank Nuno Peixeiro and Sophie Schbates for helpful discussions.

References

1. Brock TD, Brock KM, Belly RT, Weiss RL. 1972 *Sulfolobus*: a new genus of sulfur-oxidizing bacteria living at low pH and high temperature. *Arch. Mikrobiol.* **84**, 54–68. (doi:10.1007/BF00408082)
2. Redder P, Garrett RA. 2006 Mutations and rearrangements in the genome of *Sulfolobus solfataricus* P2. *J. Bacteriol.* **188**, 4198–4206. (doi:10.1128/JB.00061-06)
3. Ciaramella M, Pisani FM, Rossi M. 2002 Molecular biology of extremophiles: recent progress on the hyperthermophilic archaeon *Sulfolobus*. *Antonie Van Leeuwenhoek* **81**, 85–97. (doi:10.1023/A:1020577510469)
4. She Q *et al.* 2001 The complete genome of the crenarchaeon *Sulfolobus solfataricus* P2. *Proc. Natl Acad. Sci. USA* **98**, 7835–7840. (doi:10.1073/pnas.141222098)
5. Kawarabayasi Y *et al.* 2001 Complete genome sequence of an aerobic thermoacidophilic crenarchaeon, *Sulfolobus tokodaii* strain 7. *DNA Res* **8**, 123–140. (doi:10.1093/dnares/8.4.123)
6. Chen L *et al.* 2005 The genome of *Sulfolobus acidocaldarius*, a model organism of the Crenarchaeota. *J. Bacteriol.* **187**, 4992–4999. (doi:10.1128/JB.187.14.4992-4999.2005)
7. Guo L *et al.* 2011 Genome analyses of Icelandic strains of *Sulfolobus islandicus*, model organisms for genetic and virus–host interaction studies. *J. Bacteriol.* **193**, 1672–1680. (doi:10.1128/JB.01487-10)
8. Zillig W *et al.* 1998 Genetic elements in the extremely thermophilic archaeon *Sulfolobus*. *Extremophiles* **2**, 131–140. (doi:10.1007/s007920050052)
9. Reno ML, Held NL, Fields CJ, Burke PV, Whitaker RJ. 2009 Biogeography of the *Sulfolobus islandicus* pan-genome. *Proc. Natl Acad. Sci. USA* **106**, 8605–8610. (doi:10.1073/pnas.0808945106)
10. Prangishvili D, Arnold HP, Gotz D, Ziese U, Holz I, Kristjansson JK, Zillig W. 1999 A novel virus family, the Rudiviridae: structure, virus–host interactions and genome variability of the *Sulfolobus* viruses SIRV1 and SIRV2. *Genetics* **152**, 1387–1396.
11. Bize A *et al.* 2009 A unique virus release mechanism in the Archaea. *Proc. Natl Acad. Sci. USA* **106**, 11 306–11 311. (doi:10.1073/pnas.0901238106)
12. Quax TE, Krupovic M, Lucas S, Forterre P, Prangishvili D. 2010 The *Sulfolobus* rod-shaped virus 2 encodes a prominent structural component of the unique virion release system in Archaea. *Virology* **404**, 1–4. (doi:10.1016/j.virol.2010.04.020)
13. Quax TE, Lucas S, Reimann J, Pehau-Arnaudet G, Prevost MC, Forterre P, Albers SV, Prangishvili D. 2011 Simple and elegant design of a virion egress structure in Archaea. *Proc. Natl Acad. Sci. USA* **108**, 3354–3359. (doi:10.1073/pnas.1018052108)
14. Allers T, Mevarech M. 2005 Archaeal genetics: the third way. *Nat. Rev. Genet.* **6**, 58–73. (doi:10.1038/nrg1504)
15. Jonuscheit M, Martusewitsch E, Stedman KM, Schleper C. 2003 A reporter gene system for the hyperthermophilic archaeon *Sulfolobus solfataricus* based on a selectable and integrative shuttle vector. *Mol. Microbiol.* **48**, 1241–1252. (doi:10.1046/j.1365-2958.2003.03509.x)

16. Worthington P, Hoang V, Perez-Pomares F, Blum P. 2003 Targeted disruption of the alpha-amylase gene in the hyperthermophilic archaeon *Sulfolobus solfataricus*. *J. Bacteriol.* **185**, 482–488. (doi:10.1128/JB.185.2.482-488.2003)
17. Wagner M, Berkner S, Ajon M, Driessen AJ, Lipps G, Albers SV. 2009 Expanding and understanding the genetic toolbox of the hyperthermophilic genus *Sulfolobus*. *Biochem. Soc. Trans.* **37**, 97–101. (doi:10.1042/BST0370097)
18. Wagner M, van Wolferen M, Wagner A, Lassak K, Meyer BH, Reimann J, Albers SV. 2012 Versatile Genetic Tool Box for the Crenarchaeote *Sulfolobus acidocaldarius*. *Front Microbiol.* **3**, 214. (doi:10.3389/fmicb.2012.00214)
19. Deng L, Zhu H, Chen Z, Liang YX, She Q. 2009 Unmarked gene deletion and host-vector system for the hyperthermophilic crenarchaeon *Sulfolobus islandicus*. *Extremophiles* **13**, 735–746. (doi:10.1007/s00792-009-0254-2)
20. Leigh JA, Albers SV, Atomi H, Allers T. 2011 Model organisms for genetics in the domain Archaea: methanogens, halophiles, Thermococcales and Sulfolobales. *FEMS Microbiol. Rev.* **35**, 577–608. (doi:10.1111/j.1574-6976.2011.00265.x)
21. She Q, Zhang C, Deng L, Peng N, Chen Z, Liang YX. 2009 Genetic analyses in the hyperthermophilic archaeon *Sulfolobus islandicus*. *Biochem. Soc. Trans.* **37**, 92–96. (doi:10.1042/BST0370092)
22. Zhang C, Whitaker RJ. 2012 A broadly applicable gene knockout system for the thermoacidophilic archaeon *Sulfolobus islandicus* based on simvastatin selection. *Microbiology* **158**, 1513–1522. (doi:10.1099/mic.0.058289-0)
23. Zerbino DR, Birney E. 2008 Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **18**, 821–829. (doi:10.1101/gr.074492.107)
24. Ewing B, Green P. 1998 Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **8**, 186–194.
25. Ewing B, Hillier L, Wendt MC, Green P. 1998 Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **8**, 175–185.
26. Gordon D, Abajian C, Green P. 1998 Consed: a graphical tool for sequence finishing. *Genome Res.* **8**, 195–202.
27. Lowe TM, Eddy SR. 1997 tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955–964.
28. Okonechnikov K, Golosova O, Fursov M. 2012 Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics* **28**, 1166–1167. (doi:10.1093/bioinformatics/bts091)
29. Schbath S. a. H. M. 2011 R'MES: a tool to find motifs with a significantly unexpected frequency in biological sequences. In *Advances in genomic sequence analysis and pattern discovery*, vol. 7 (eds L Elnitski, O Piontkivska, L Welch). Science, Engineering, and Biology Informatics. Singapore: World Scientific.
30. Grissa I, Vergnaud G, Pourcel C. 2007 CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res.* **35**, W52–W57. (doi:10.1093/nar/gkm360)
31. Oberto J, Forterre P. In preparation. Comparative genomics of Thermococcales.
32. Lerat E, Daubin V, Moran NA. 2003 From gene trees to organismal phylogeny in prokaryotes: the case of the gamma-Proteobacteria. *PLoS Biol.* **1**, E19. (doi:10.1371/journal.pbio.0000019)
33. Cadillo-Quiroz H, Didelot X, Held NL, Herrera A, Darling A, Reno ML, Krause DJ, Whitaker RJ. 2012 Patterns of gene flow define species of thermophilic Archaea. *PLoS Biol.* **10**, e1001265. (doi:10.1371/journal.pbio.1001265)
34. Peng X, Kessler A, Phan H, Garrett RA, Prangishvili D. 2004 Multiple variants of the archaeal DNA ruvivirus SIRV1 in a single host and a novel mechanism of genomic variation. *Mol. Microbiol.* **54**, 366–375. (doi:10.1111/j.1365-2958.2004.04287.x)
35. Blum H, Zillig W, Mallok S, Domdey H, Prangishvili D. 2001 The genome of the archaeal virus SIRV1 has features in common with genomes of eukaryal viruses. *Virology* **281**, 6–9. (doi:10.1006/viro.2000.0776)
36. Lillestol RK, Shah SA, Brugger K, Redder P, Phan H, Christiansen J, Garrett RA. 2009 CRISPR families of the crenarchaeal genus *Sulfolobus*: bidirectional transcription and dynamic properties. *Mol. Microbiol.* **72**, 259–272. (doi:10.1111/j.1365-2958.2009.06641.x)
37. Shah SA, Garrett RA. 2011 CRISPR/Cas and Cmr modules, mobility and evolution of adaptive immune systems. *Res. Microbiol.* **162**, 27–38. (doi:10.1016/j.resmic.2010.09.001)
38. Touzain F, Petit MA, Schbath S, El Karoui M. 2011 DNA motifs that sculpt the bacterial chromosome. *Nat. Rev. Microbiol.* **9**, 15–26. (doi:10.1038/nrmicro2477)
39. Bush EC, Lahn BT. 2006 The evolution of word composition in metazoan promoter sequence. *PLoS Comput. Biol.* **2**, e150. (doi:10.1371/journal.pcbi.0020150)
40. Rocha EP, Danchin A, Viari A. 2001 Evolutionary role of restriction/modification systems as revealed by comparative genome analysis. *Genome Res.* **11**, 946–958. (doi:10.1101/gr.GR-1531RR)
41. Darling AE, Mau B, Perna NT. 2010 progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS ONE* **5**, e11147. (doi:10.1371/journal.pone.0011147)
42. She Q, Peng X, Zillig W, Garrett RA. 2001 Gene capture in archaeal chromosomes. *Nature* **409**, 478. (doi:10.1038/35054138)
43. She Q, Shen B, Chen L. 2004 Archaeal integrases and mechanisms of gene capture. *Biochem. Soc. Trans.* **32**, 222–226. (doi:10.1042/BST0320222)
44. Muskhelishvili G, Palm P, Zillig W. 1993 SSV1-encoded site-specific recombination system in *Sulfolobus shibatae*. *Mol. Gen. Genet.* **237**, 334–342.
45. Held NL, Whitaker RJ. 2009 Viral biogeography revealed by signatures in *Sulfolobus islandicus* genomes. *Environ. Microbiol.* **11**, 457–466. (doi:10.1111/j.1462-2920.2008.01784.x)
46. Krupovic M, Prangishvili D, Hendrix RW, Bamford DH. 2011 Genomics of bacterial and archaeal viruses: dynamics within the prokaryotic virosphere. *Microbiol. Mol. Biol. Rev.* **75**, 610–635. (doi:10.1128/MMBR.00011-11)
47. Mochizuki T, Sako Y, Prangishvili D. 2011 Provirus induction in hyperthermophilic archaea: characterization of *Aeropyrum pernix* spindle-shaped virus 1 and *Aeropyrum pernix* ovoid virus 1. *J. Bacteriol.* **193**, 5412–5419. (doi:10.1128/JB.05101-11)
48. She Q, Phan H, Garrett RA, Albers SV, Stedman KM, Zillig W. 1998 Genetic profile of pNOB8 from *Sulfolobus*: the first conjugative plasmid from an archaeon. *Extremophiles* **2**, 417–425. (doi:10.1007/s007920050087)
49. Cortez D, Forterre P, Gribaldo S. 2009 A hidden reservoir of integrative elements is the major source of recently acquired foreign genes and ORFans in archaeal and bacterial genomes. *Genome Biol.* **10**, R65. (doi:10.1186/gb-2009-10-6-r65)
50. Zhang CT, Zhang R, Ou HY. 2003 The Z curve database: a graphic representation of genome sequences. *Bioinformatics* **19**, 593–599. (doi:10.1093/bioinformatics/btg041)
51. Zhang R, Zhang CT. 2005 Identification of replication origins in archaeal genomes based on the Z-curve method. *Archaea* **1**, 335–346. (doi:10.1155/2005/509646)
52. Cortez D, Quevillon-Cheruel S, Gribaldo S, Desnoues N, Sezonov G, Forterre P, Serre MC. 2010 Evidence for a Xer/dif system for chromosome resolution in archaea. *PLoS Genet.* **6**, e1001166. (doi:10.1371/journal.pgen.1001166)
53. Flynn KM, Vohr SH, Hatcher PJ, Cooper VS. 2010 Evolutionary rates and gene dispensability associate with replication timing in the archaeon *Sulfolobus islandicus*. *Genome Biol. Evol.* **2**, 859–869. (doi:10.1093/gbe/evq068)
54. Dueber EC, Costa A, Corn JE, Bell SD, Berger JM. 2011 Molecular determinants of origin discrimination by Orc1 initiators in archaea. *Nucleic Acids Res.* **39**, 3621–3631. (doi:10.1093/nar/gkq1308)
55. Lundgren M, Andersson A, Chen L, Nilsson P, Bernander R. 2004 Three replication origins in *Sulfolobus* species: synchronous initiation of chromosome replication and asynchronous termination. *Proc. Natl Acad. Sci. USA* **101**, 7046–7051. (doi:10.1073/pnas.0400656101)
56. Robinson NP, Dionne I, Lundgren M, Marsh VL, Bernander R, Bell SD. 2004 Identification of two origins of replication in the single chromosome of the archaeon *Sulfolobus solfataricus*. *Cell* **116**, 25–38. (doi:10.1016/S0092-8674(03)01034-1)
57. Robinson NP, Bell SD. 2007 Extrachromosomal element capture and the evolution of multiple replication origins in archaeal chromosomes. *Proc. Natl Acad. Sci. USA* **104**, 5806–5811. (doi:10.1073/pnas.0700206104)
58. Pandey DP, Gerdes K. 2005 Toxin-antitoxin loci are highly abundant in free-living but lost from host-associated prokaryotes. *Nucleic Acids Res.* **33**, 966–976. (doi:10.1093/nar/gki201)

59. Makarova KS, Wolf YI, Koonin EV. 2009 Comprehensive comparative-genomic analysis of type 2 toxin-antitoxin systems and related mobile stress response systems in prokaryotes. *Biol. Direct.* **4**, 19. (doi:10.1186/1745-6150-4-19)
60. Grynberg M, Erlandsen H, Godzik A. 2003 HEPN: a common domain in bacterial drug resistance and human neurodegenerative proteins. *Trends Biochem. Sci.* **28**, 224–226. (doi:10.1016/S0968-0004(03)00060-4)
61. Ellen AF, Rohulya OV, Fusetti F, Wagner M, Albers SV, Driessen AJ. 2011 The sulfolobin genes of *Sulfolobus acidocaldarius* encode novel antimicrobial proteins. *J. Bacteriol.* **193**, 4380–4387. (doi:10.1128/JB.05028-11)
62. O'Connor EM, Shand RF. 2002 Halocins and sulfolobins: the emerging story of archaeal protein and peptide antibiotics. *J. Ind. Microbiol. Biotechnol.* **28**, 23–31.
63. Prangishvili D, Holz I, Stieger E, Nickell S, Kristjansson JK, Zillig W. 2000 Sulfolobins, specific proteinaceous toxins produced by strains of the extremely thermophilic archaeal genus *Sulfolobus*. *J. Bacteriol.* **182**, 2985–2988. (doi:10.1128/JB.182.10.2985-2988.2000)
64. Ajon M, Frols S, van Wolferen M, Stoecker K, Teichmann D, Driessen AJ, Grogan DW, Albers SV, Schleper C. 2011 UV-inducible DNA exchange in hyperthermophilic archaea mediated by type IV pili. *Mol. Microbiol.* **82**, 807–817. (doi:10.1111/j.1365-2958.2011.07861.x)
65. Frols S *et al.* 2008 UV-inducible cellular aggregation of the hyperthermophilic archaeon *Sulfolobus solfataricus* is mediated by pili formation. *Mol. Microbiol.* **70**, 938–952. (doi:10.1111/j.1365-2958.2008.06459.x)
66. Brugger K, Redder P, She Q, Confalonieri F, Zivanovic Y, Garrett RA. 2002 Mobile elements in archaeal genomes. *FEMS Microbiol. Lett.* **206**, 131–141. (doi:10.1016/S0378-1097(01)00504-3)
67. Filee J, Siguier P, Chandler M. 2007 Insertion sequence diversity in archaea. *Microbiol. Mol. Biol. Rev.* **71**, 121–157. (doi:10.1128/MMBR.00031-06)
68. Redder P, She Q, Garrett RA. 2001 Non-autonomous mobile elements in the crenarchaeon *Sulfolobus solfataricus*. *J. Mol. Biol.* **306**, 1–6. (doi:10.1006/jmbi.2000.4377)
69. Feschotte C, Mouches C. 2000 Evidence that a family of miniature inverted-repeat transposable elements (MITEs) from the *Arabidopsis thaliana* genome has arisen from a pogo-like DNA transposon. *Mol. Biol. Evol.* **17**, 730–737. (doi:10.1093/oxfordjournals.molbev.a026351)
70. Feschotte C, Swamy L, Wessler SR. 2003 Genome-wide analysis of mariner-like transposable elements in rice reveals complex relationships with stowaway miniature inverted repeat transposable elements (MITEs). *Genetics* **163**, 747–758.
71. Yang G, Dong J, Chandrasekharan MB, Hall TC. 2001 Kiddo, a new transposable element family closely associated with rice genes. *Mol. Genet. Genomics* **266**, 417–424. (doi:10.1007/s004380100530)
72. Yang G, Hall TC. 2003 MDM-1 and MDM-2: two mutator-derived MITE families in rice. *J. Mol. Evol.* **56**, 255–264. (doi:10.1007/s00239-002-2397-y)
73. Zhang Q, Arbuckle J, Wessler SR. 2000 Recent, extensive, and preferential insertion of members of the miniature inverted-repeat transposable element family Heartbreaker into genic regions of maize. *Proc. Natl Acad. Sci. USA* **97**, 1160–1165. (doi:10.1073/pnas.97.3.1160)
74. Jiang N, Bao Z, Zhang X, Hirochika H, Eddy SR, McCouch SR, Wessler SR. 2003 An active DNA transposon family in rice. *Nature* **421**, 163–167. (doi:10.1038/nature01214)
75. Yang G, Nagel DH, Feschotte C, Hancock CN, Wessler SR. 2009 Tuned for transposition: molecular determinants underlying the hyperactivity of a Stowaway MITE. *Science* **325**, 1391–1394. (doi:10.1126/science.1175688)
76. Berkner S, Lipps G. 2007 An active nonautonomous mobile element in *Sulfolobus islandicus* REN1H1. *J. Bacteriol.* **189**, 2145–2149. (doi:10.1128/JB.01567-06)
77. Blount ZD, Grogan DW. 2005 New insertion sequences of *Sulfolobus*: functional properties and implications for genome evolution in hyperthermophilic archaea. *Mol. Microbiol.* **55**, 312–325. (doi:10.1111/j.1365-2958.2004.04391.x)
78. Haft DH, Selengut J, Mongodin EF, Nelson KE. 2005 A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. *PLoS Comput. Biol.* **1**, e60. (doi:10.1371/journal.pcbi.0010060)
79. Makarova KS *et al.* 2011 Evolution and classification of the CRISPR-Cas systems. *Nat. Rev. Microbiol.* **9**, 467–477. (doi:10.1038/nrmicro2577)
80. Zillig W, Prangishvili D, Schleper C, Elferink M, Holz I, Albers S, Janekovic D, Gatz D. 1996 Viruses, plasmids and other genetic elements of thermophilic and hyperthermophilic Archaea. *FEMS Microbiol. Rev.* **18**, 225–236. (doi:10.1111/j.1574-6976.1996.tb00239.x)
81. Greve B, Jensen S, Brugger K, Zillig W, Garrett RA. 2004 Genomic comparison of archaeal conjugative plasmids from *Sulfolobus*. *Archaea* **1**, 231–239. (doi:10.1155/2004/151926)
82. Prangishvili D *et al.* 1998 Conjugation in archaea: frequent occurrence of conjugative plasmids in *Sulfolobus*. *Plasmid* **40**, 190–202. (doi:10.1006/plas.1998.1363)
83. Stedman KM, She Q, Phan H, Holz I, Singh H, Prangishvili D, Garrett R, Zillig W. 2000 pING family of conjugative plasmids from the extremely thermophilic archaeon *Sulfolobus islandicus*: insights into recombination and conjugation in Crenarchaeota. *J. Bacteriol.* **182**, 7014–7020. (doi:10.1128/JB.182.24.7014-7020.2000)
84. Datsenko KA, Pougach K, Tikhonov A, Wanner BL, Severinov K, Semenova E. 2012 Molecular memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system. *Nat. Commun.* **3**, 945. (doi:10.1038/ncomms1937)
85. Marchfelder A *et al.* 2012 Small RNAs for defence and regulation in archaea. *Extremophiles* **16**, 685–696. (doi:10.1007/s00792-012-0469-5)
86. Marraffini LA, Sontheimer EJ. 2010 CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. *Nat. Rev. Genet.* **11**, 181–190. (doi:10.1038/nrg2749)
87. Marraffini LA, Sontheimer EJ. 2010 Self versus non-self discrimination during CRISPR RNA-directed immunity. *Nature* **463**, 568–571. (doi:10.1038/nature08703)
88. Gudbergstottir S, Deng L, Chen Z, Jensen JV, Jensen LR, She Q, Garrett RA. 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. (doi:10.1111/j.1365-2958.2010.07452.x)
89. Prangishvili D, Koonin EV, Krupovic M. 2013 Genomics and biology of Rudiviruses, a model for the study of virus–host interactions in Archaea. *Biochem. Soc. Trans.* **41**, 443–450. (doi:10.1042/BST20120313)
90. Prangishvili D, Quax TE. 2011 Exceptional virion release mechanism: one more surprise from archaeal viruses. *Curr. Opin. Microbiol.* **14**, 315–320. (doi:10.1016/j.mib.2011.04.006)
91. Palm P, Schleper C, Grampp B, Yeats S, McWilliam P, Reiter WD, Zillig W. 1991 Complete nucleotide sequence of the virus SSV1 of the archaeobacterium *Sulfolobus shibatae*. *Virology* **185**, 242–250. (doi:10.1016/0042-6822(91)90771-3)
92. Reiter WD, Palm P, Yeats S, Zillig W. 1987 Gene expression in archaeobacteria: physical mapping of constitutive and UV-inducible transcripts from the *Sulfolobus* virus-like particle SSV1. *Mol. Gen. Genet.* **209**, 270–275. (doi:10.1007/BF00329653)
93. Stedman KM, Schleper C, Rumpf E, Zillig W. 1999 Genetic requirements for the function of the archaeal virus SSV1 in *Sulfolobus solfataricus*: construction and testing of viral shuttle vectors. *Genetics* **152**, 1397–1405.
94. Zheng T, Huang Q, Zhang C, Ni J, She Q, Shen Y. 2012 Development of a simvastatin selection marker for a hyperthermophilic acidophile, *Sulfolobus islandicus*. *Appl. Environ. Microbiol.* **78**, 568–574. (doi:10.1128/AEM.06095-11)
95. Kirzinger MW, Stavrinides J. 2012 Host specificity determinants as a genetic continuum. *Trends Microbiol.* **20**, 88–93. (doi:10.1016/j.tim.2011.11.006)
96. Bennett PM. 2004 Genome plasticity: insertion sequence elements, transposons and integrons, and DNA rearrangement. *Methods Mol. Biol.* **266**, 71–113.
97. Peng X, Blum H, She Q, Mallok S, Brugger K, Garrett RA, Zillig W, Prangishvili D. 2001 Sequences and replication of genomes of the archaeal rudiviruses SIRV1 and SIRV2: relationships to the archaeal lipothrixvirus SIFV and some eukaryal viruses. *Virology* **291**, 226–234. (doi:10.1006/viro.2001.1190)
98. Deveau H, Garneau JE, Moineau S. 2010 CRISPR/Cas system and its role in phage-bacteria interactions.

- Annu. Rev. Microbiol.* **64**, 475–493. (doi:10.1146/annurev.micro.112408.134123)
99. Horvath P, Barrangou R. 2010 CRISPR/Cas, the immune system of bacteria and archaea. *Science* **327**, 167–170. (doi:10.1126/science.1179555)
100. Marraffini LA, Sontheimer EJ. 2008 CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. *Science* **322**, 1843–1845. (doi:10.1126/science.1165771)
101. Cady KC, Bondy-Denomy J, Heussler GE, Davidson AR, O'Toole GA. 2012 The CRISPR/Cas adaptive immune system of *Pseudomonas aeruginosa* mediates resistance to naturally occurring and engineered phages. *J. Bacteriol.* **194**, 5728–5738. (doi:10.1128/JB.01184-12)
102. Lopez-Sanchez MJ, Sauvage E, Da Cunha V, Clermont D, Ratsima Hariniaina E, Gonzalez-Zorn B, Poyart C, Rosinski-Chupin I, Glaser P. 2012 The highly dynamic CRISPR1 system of *Streptococcus agalactiae* controls the diversity of its mobilome. *Mol. Microbiol.* **85**, 1057–1071. (doi:10.1111/j.1365-2958.2012.08172.x)
103. Swarts DC, Mosterd C, van Passel MW, Brouns SJ. 2012 CRISPR interference directs strand specific spacer acquisition. *PLoS ONE* **7**, e35888. (doi:10.1371/journal.pone.0035888)
104. Yosef I, Goren MG, Qimron U. 2012 Proteins and DNA elements essential for the CRISPR adaptation process in *Escherichia coli*. *Nucleic Acids Res.* **40**, 5569–5576. (doi:10.1093/nar/gks216)
105. Erdmann S, Garrett RA. 2012 Selective and hyperactive uptake of foreign DNA by adaptive immune systems of an archaeon via two distinct mechanisms. *Mol. Microbiol.* **85**, 1044–1056. (doi:10.1111/j.1365-2958.2012.08171.x)
106. Bikard D, Marraffini LA. 2012 Innate and adaptive immunity in bacteria: mechanisms of programmed genetic variation to fight bacteriophages. *Curr. Opin. Immunol.* **24**, 15–20. (doi:10.1016/j.coi.2011.10.005)
107. Zhang C, Krause DJ, Whitaker RJ. 2013 *Sulfolobus islandicus*: a model system for evolutionary genomics. *Biochem. Soc. Trans.* **41**, 458–462. (doi:10.1042/BST20120338)

3. Sta1, un régulateur transcriptionnel codé par *Sulfolobus islandicus*

1. Contexte de l'étude

Le régulateur transcriptionnel Sta1 (*Sulfolobus* transcriptional activator 1) a été identifié, lors de l'étude de l'expression génique du rudivirus SIRV1 dans les cellules de son hôte, *Sulfolobus islandicus* REN2H1 (Kessler *et al.*, 2006).

Trois promoteurs viraux différents immobilisés sur des billes magnétiques ont été utilisés pour conduire des expériences de purification d'affinité, à partir d'extraits bruts de cellules hôtes dans le but d'isoler des protéines impliquées dans la transcription de gènes du virus SIRV1.

Les promoteurs des gènes de SIRV1, *ORF56* (codant pour le régulateur transcriptionnel SvtR), *ORF399* (codant pour un transporteur putatif d'acides aminés) et *ORF134* (codant pour la protéine majeure de capsid), ont été choisis pour la recherche de protéines impliquées dans la régulation de la transcription du génome viral, comme la carte de transcription de SIRV1 avait indiqué que l'expression des trois gènes pouvait être sous contrôle transcriptionnel (Kessler *et al.*, 2004).

A l'issue de ces expériences de *pull-down*, Sta1 est la seule protéine à avoir été récupérée. Cette protéine de 14kDa est codée par le chromosome hôte et contient, d'après les observations sous RMN, un motif WHTH (le plus probablement impliqué dans la liaison à l'ADN) et lie l'ADN sous forme de dimère (Kessler *et al.*, 2006).

Son effet activateur sur l'initiation de la transcription de ces promoteurs viraux a été démontré dans des expériences de transcription *in vitro*, en utilisant un système reconstitué contenant l'ARN polymérase et les protéines TBP et TFB. Sta1 est capable de lier les promoteurs viraux indépendamment de tout composant du complexe de pré-initiation de l'hôte. Deux sites de liaison ont été révélés par *footprinting* : le premier localisé au cœur de la région promotrice et le second, environ 30pb en amont. La proximité apparente d'un des sites de liaison de Sta1 avec la *Tata-box* et l'élément BRE suggère que l'effet activateur de Sta1 peut influencer la liaison des facteurs généraux de transcription, TBP et/ou TFB (Kessler *et al.*, 2006).

Des homologues de Sta1 composés de 120-130 acides aminés sont présents dans tous les génomes séquencés des espèces *Sulfolobus*. Plusieurs copies du gène sont retrouvées dans chaque génome analysé. Le nombre minimal d'orthologues de *sta1* par génome est cinq (estimation basée sur un seuil d'E-value de 0.01). En appliquant ce même seuil, des homologues de Sta1 ont été aussi identifiés dans plusieurs autres génomes d'archées (Kessler *et al.*, 2006).

Sta1 est le premier régulateur transcriptionnel d'archées isolé *in vitro* par l'approche de *pull-down* (Gabielsen *et al.*, 1989). Tous les autres régulateurs transcriptionnels chez les archées ont été identifiés par des analyses *in silico*.

La découverte de Sta1, une protéine cellulaire, recrutée par le virus, indique que les virus d'archées, comme les virus des eucaryotes et les bactériophages, utilisent des régulateurs de l'hôte pour activer la transcription de leurs propres gènes. La compréhension du rôle de Sta1 passe bien évidemment par l'étude de son rôle primaire dans les cellules d'archées. Au moment de la découverte de cette protéine, aucune information sur ce point n'était disponible dans la littérature.

Par la suite, il y a été démontré que parmi les gènes du chromosome, Sta1 régule de façon positive l'expression du gène *ss0777*, un paralogue du gène *radA* impliqué dans la réparation de l'ADN (Abella *et al.*, 2007).

La réparation de l'ADN chez les archées n'est pas décrite en détails. La régulation des gènes impliqués dans ce processus est encore moins connue. L'analyse *in silico* indique que la plupart des protéines de réparation de l'ADN sont similaires à celles des eucaryotes et non celles des bactéries et que beaucoup de gènes « classiques » de réparation sont absents des génomes d'archées (Kelman and White, 2005).

La famille de protéines RecA, qui inclut RadA chez les Archées et Rad51 chez les Eucaryotes, est une des plus connues des protéines de réparation de l'ADN et elle joue un rôle fondamental dans l'échange de brin d'ADN lors de la recombinaison homologue et la réparation des cassures doubles brins. Chez *S. solfataricus* et les espèces proches, il y a deux homologues de RecA, la très bien caractérisée protéine RadA (Sso0250) (Ariza *et al.*, 2005; Sandler *et al.*, 1996; Naoe *et al.*, 2004) et un paralogue de RadA, moins conservé (Sso0777) dont la fonction est inconnue (Guy *et al.*, 2006).

En conclusion, Sta1 serait le premier régulateur transcriptionnel directement impliqué dans la régulation d'une des voies de réparation de l'ADN. En outre, Sta1 représente le premier régulateur transcriptionnel cellulaire d'archées recruté par un virus pour la régulation de ses propres gènes. Son rôle dans le contrôle de certains gènes de SIRV1 est mieux compris que son implication dans la régulation des gènes chromosomiques. En raison de toutes ces données, notre objectif dans l'étude du gène *sta1* a été d'estimer son importance dans le fonctionnement de la cellule de *S. islandicus* et de tenter d'inactiver ce gène. La comparaison du phénotype de mutant *vs* sauvage (résistance/sensibilité à SIRV1; résistance/sensibilité aux UV) servirait à mieux comprendre le rôle de *sta1*.

2. Le régulon hypothétique de Sta1

En plus d'activer la transcription de plusieurs gènes viraux, Sta1 contrôlerait un régulon d'environ 75 gènes dans la cellule *S. islandicus* (Sezonov, Peixeiro, données non publiées, présentées dans l'annexe) dont quelques-uns semblent être impliqués dans la réponse au stress. L'ensemble de ces données suggère que Sta1 pourrait représenter l'exemple d'un régulateur général important pour la survie de la cellule placée en conditions de stress permanent (pH acide, haute température). Pour vérifier les hypothèses concernant le rôle et l'importance de *sta1*, il était indispensable de construire un mutant *knock-out* de ce gène.

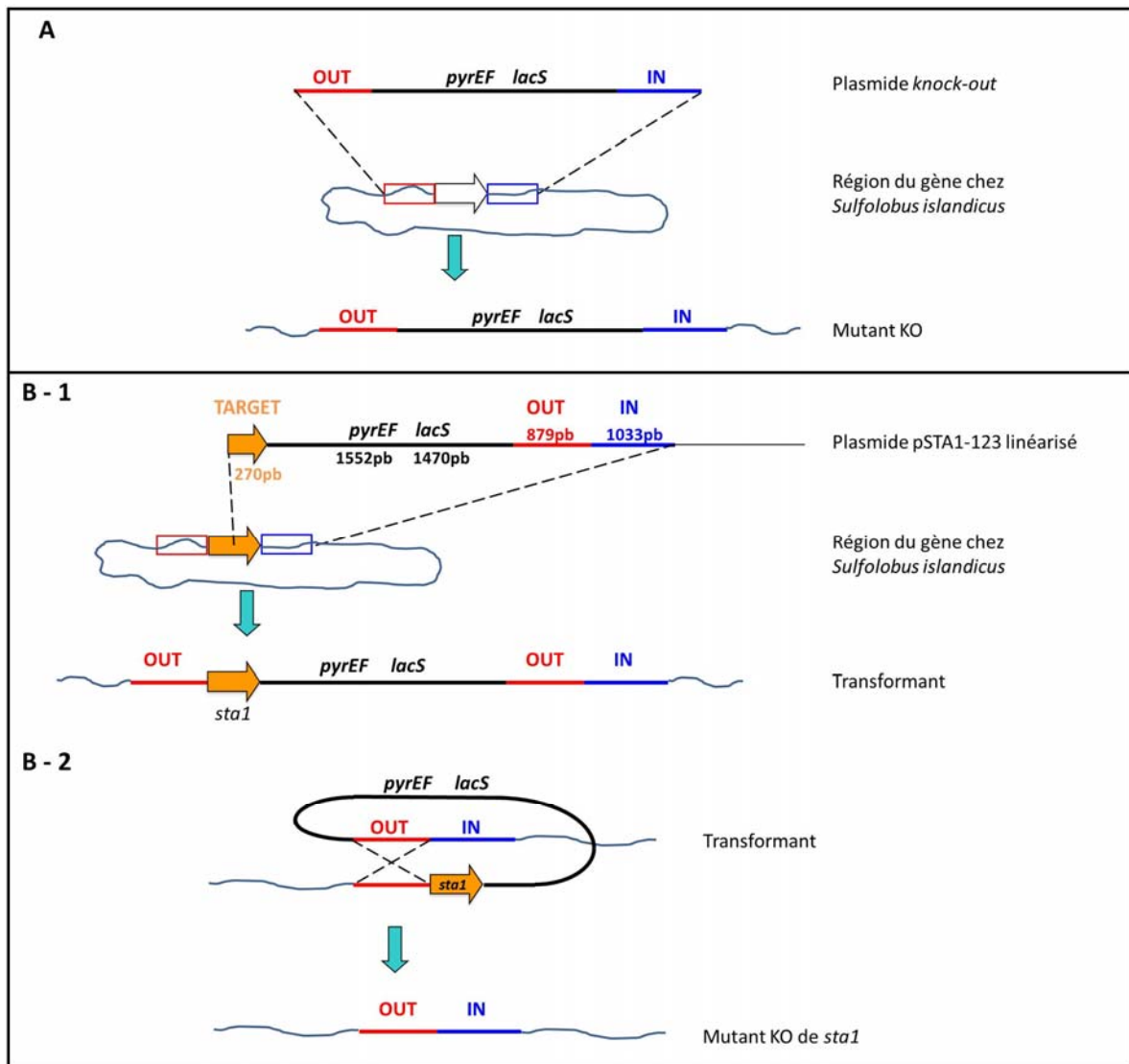


Figure 47 : Les principales étapes du *knock-out* du gène *sta1*. (A) *Knock-out* « classique » de gène avec remplacement du gène cible par les marqueurs *pyrEF* et *lacS*, (B) *Knock-out* en deux étapes (pour les gènes essentiels, seule la première étape peut être mise en évidence). B1 : Intégration de la cassette de recombinaison conduisant à un gène fonctionnel. Même si le gène cible est essentiel, ce *knock-out* en deux étapes permet d'obtenir des transformants primaires qui ont intégré la cassette de recombinaison contenant les gènes *pyrEF* et *lacS* et les séquences homologues du locus du gène *sta1*, sans altérer la séquence du gène cible (*sta1*). B2 : Recombinaison entre les régions OUT qui inactive le gène s'il n'est pas essentiel. Ces transformants peuvent être sélectionnés comme des prototrophes Pyr⁺ et ils doivent également avoir le phénotype Lac⁺ (ou plutôt X-gal⁺). La seconde phase de sélection est réalisée dans un milieu riche complété par de l'uracile et du 5-FOA. Les cellules *pyrEF*⁺ convertissent le 5-FOA en un dérivé toxique, le 5-FU (5-fluoro-uracile), et ne survivent pas. Si le gène à inactiver n'est pas essentiel, la cassette de recombinaison *sta1-pyrEF-lacS* peut être facilement perdue. En conséquence, étalées sur un milieu uracile-5-FOA-Xgal, les colonies des survivants sont majoritairement blanches. Dans le cas d'un gène essentiel, les colonies observées sur le même milieu sont, le plus souvent, bleues. Elles gardent la cassette de recombinaison, et donc le gène *lacS*, mais la présence de 5-FOA ne permet pas la survie de celles qui expriment les gènes *pyrEF*. Dans ce cas, les colonies observées sont issues de mutants spontanés Pyr⁻. De plus, l'excision de la cassette de recombinaison avec les marqueurs génétiques du chromosome permet leur réutilisation, si la construction d'un double mutant est envisagée.

3. Raisonnement génétique

Etant donné l'incertitude sur l'importance du réseau de régulation de Sta1 dans la cellule, nul ne peut ignorer que son inactivation peut fortement influencer le fonctionnement de la machine cellulaire. Le gène *sta1* peut être essentiel et son inactivation provoquerait un phénotype létal. Dans cette situation, la meilleure approche pour tenter d'obtenir un mutant KO sera sans doute l'approche MID décrite dans l'Introduction (I.5.4.b.iii). En absence de double mutant *pyrEF- lacS-* chez *S. islandicus* LAL14/1 cette approche est inapplicable pour le moment à notre souche modèle. Dans ces circonstances, nous avons décidé d'utiliser la souche *S. islandicus* REY15A (mise à notre disposition par l'équipe de Q. She; Danish Archaea Centre) qui est adaptée pour l'approche MID et pour laquelle plusieurs outils génétiques sont déjà disponibles (She *et al.*, 2009). La construction du mutant KO *sta1-* dans cette souche aurait utile pour étudier comprendre le rôle cellulaire de Sta1. Cependant, il était évident dès le départ que ce mutant ne serait pas utilisable pour l'étude du rôle de Sta1 dans l'infection par le virus SIRV1 car la souche REY15A et ses dérivées sont résistantes à SIRV1/2.

Les étapes essentielles de l'approche MID sont illustrées dans la Figure 47 et discutées en détails dans l'Introduction et dans la légende de la Figure. La deuxième étape, donnant lieu au mutant haploïde $\Delta sta1$, n'est possible que si le gène *sta1* est un gène non essentiel. Les mutants $\Delta sta1$ devraient avoir un phénotype particulier, Lac-.

4. Construction du plasmide pSTA1-123 pour réaliser le KO du gène *sta1*

Le vecteur d'inactivation pSTA1-123 a été construit par trois clonages successifs.

Le gène *sta1* chez *S. islandicus* a été identifié par son homologie avec le gène *sso0048 (sta1)* de *S. solfataricus* (la similitude nucléotidique est de 100%). Le plasmide suicide initial, **pSEF** (Zhang *et al.*, 2010), capable de se répliquer uniquement chez *E. coli* (réplicon pUC19) et non chez *S. islandicus*, portant les marqueurs génétiques *pyrEF+* et *lacS+* a été utilisé pour réaliser le clonage de trois fragments d'ADN correspondant au locus *sta1* et aux régions avoisinantes chez *S. islandicus*. Ces trois fragments ont été amplifiés par PCR et appelés **OUT**, **IN** et **TARGET** puis, clonés successivement dans le vecteur. Les positions respectives de ces fragments par rapport au gène *sta1* et leurs longueurs sont indiquées dans la Figure 48. Le plasmide résultant nommé **pSTA1-123** (Figure 49) a servi par la suite sous sa forme linéarisée à transformer les cellules *Sulfolobus islandicus* E233S ($\Delta pyrEF \Delta lacS$), en vue de la réalisation du mutant *knock-out* du gène *sta1*.

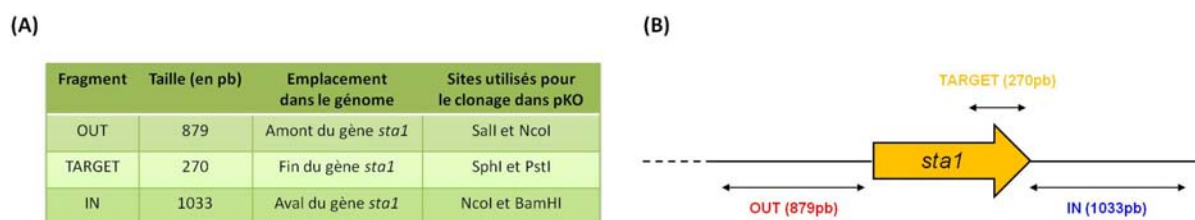


Figure 48 : Construction du plasmide pSTA1-123. (A) Détails du clonage des fragments OUT, IN et TARGET dans le plasmide pKO. (B) Représentation du gène *S. islandicus sta1* et des régions correspondantes aux fragments clonés. La position du gène *sta1* dans le génome de *Sulfolobus islandicus* est entre le 2014950^{ème} et 2014561^{ème} nucléotide.

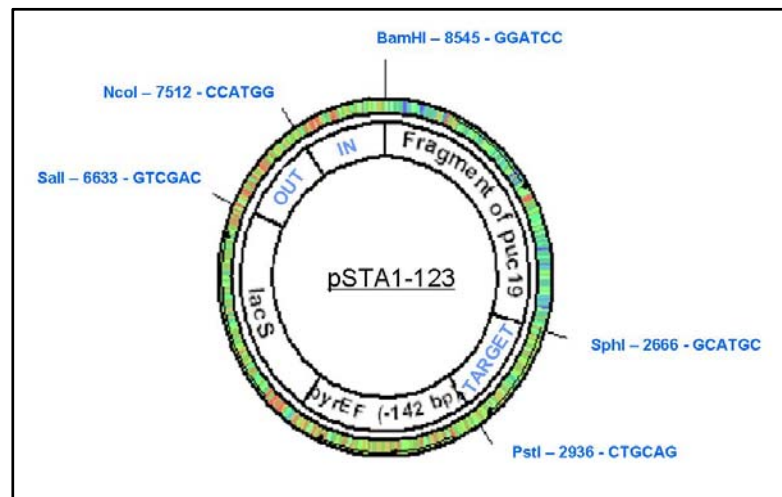


Figure 49 : Carte génétique et physique du plasmide pSTA1-123. Ce plasmide suicide est utilisé pour réaliser le *knock-out*. Il est construit à partir du vecteur *E. coli* pUC19. Les gènes *pyrEF* et *lacS* de *S. islandicus* sont utilisés comme marqueurs génétiques permettant la sélection et la visualisation des transformants chez les archées. Les fragments marqués comme OUT et IN correspondent, respectivement, aux régions en amont et en aval du gène, et le fragment TARGET à la fin du gène cible. Les séquences homologues OUT, TARGET et IN ont été amplifiées par PCR à partir du chromosome hôte.

La figure 47.B1 détaille la position relative des fragments d'ADN clonés et leur implication dans les différentes étapes de recombinaison. La localisation des fragments clonés est telle que les remaniements provoqués par l'intégration de la cassette n'auront aucune incidence sur l'intégrité du gène cible, *sta1*. Le premier événement de recombinaison entre les régions TARGET et IN du plasmide et du chromosome intègre la cassette et une deuxième copie de la région OUT. Cette recombinaison n'inactive pas le gène *sta1* (Figure 47.B1). Le deuxième événement de recombinaison se réalise entre les deux copies des séquences « OUT » et excise la cassette et le gène *sta1* (Figure 47.B2).

Après la linéarisation du plasmide pSTA1-123, les cellules de la souche réceptrice *S. islandicus* E233S $\Delta pyrEF \Delta lacS$ (dérivée de la souche REY15A) ont été transformées par électroporation (1.5kV, 600 Ω , 25 μ F). Avec 1 μ g d'ADN plasmidique électroporé et après dix jours d'incubation à 78°C, environ soixante colonies de transformants ont été observées sur les boîtes de sélection (milieu pauvre en uracile). L'efficacité de transformation observée correspond à celle décrite dans la littérature (Berkner *et al.*, 2007). Les colonies possèdent les locus *pyrEF* et *lacS* fonctionnels, puisque que la croissance a lieu dans un milieu sélectif sans uracile et que le test de coloration avec du X-gal s'est avéré positif pour 100% des colonies testées. A ce stade, nous avons conclu que les deux phénotypes observés (Ura⁺ et Lac⁺) confirmaient l'intégration de la cassette dans le génome de *S. islandicus* E233S.

A partir des colonies uniques de ces transformants, des cultures liquides ont été lancées en vue de la deuxième étape de sélection, la recherche des recombinants Ura⁻, qui devraient majoritairement correspondre aux mutants KO de *sta1* recherchés. La sélection de ces recombinants en présence de 5-FOA se réalise en milieu liquide et demande plusieurs cycles successifs de croissances-dilutions.

Au cours de la croissance dans le milieu toxique pour les cellules Ura⁺, l'état de la région *sta1* a été analysé par PCR. Comme l'explique la Figure 50, le choix des amorces utilisées pour cette analyse est tel que la présence et l'accumulation des mutants $\Delta sta1$ devraient se manifester par l'apparition d'un

amplicon d'une longueur plus petite (2100 pb) par rapport aux amplicons correspondant à la souche initiale *sta1+* (2402 pb) et au transformant (6500pb).

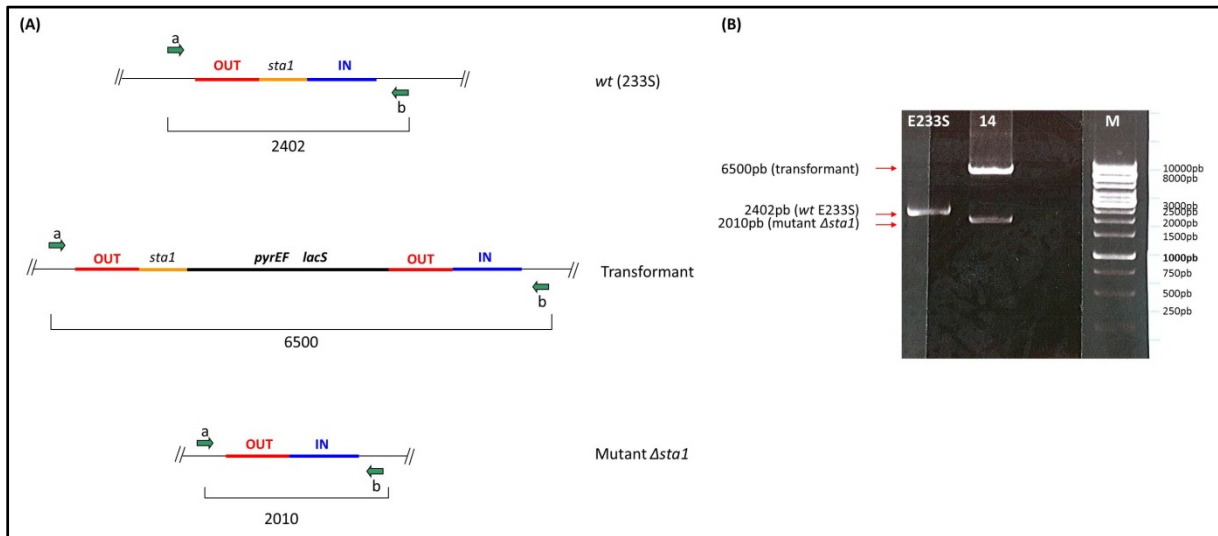


Figure 50 : Vérification par PCR des colonies initiales de transformants. (A) Le couple d'amorces flanquantes est représenté et les tailles attendues des fragments amplifiés sont données en pb, (B) Les fragments correspondant à la région du gène cible *sta1* ont été amplifiés par PCR sur la souche *Sulfolobus islandicus* E233S (puits E233S) et une culture de transformant (puits 14) après plusieurs cycles de purification. Les tailles observées sont celles attendues : 2402pb pour le wt, 6500pb pour le transformant et 2010pb pour le mutant $\Delta sta1$.

Les résultats obtenus (Figure 50B) démontrent clairement que les recombinants issus du deuxième événement de recombinaison apparaissent comme attendu. Il est difficile d'estimer la fréquence de cet événement mais l'intensité de l'amplicon de 2100 pb indique, certes de façon indirecte, qu'il ne s'agit pas d'un événement exceptionnellement rare. Etant donné que les mutants $\Delta sta1$ sont par défaut *pyrEF*⁻, quelle que soit la fréquence de cet événement de recombinaison, ces mutants devraient s'accumuler au cours de croissance suite à la pression de sélection imposée par la présence de 5-FOA. Si le gène *sta1* n'est pas un gène essentiel, l'augmentation après plusieurs jours d'incubation du signal de l'amplicon correspondant et la possibilité d'obtenir la croissance de ces mutants *pyrEF*⁻ $\Delta sta1$ *lacS*⁻ en milieu solide en tant que colonies isolées pourraient être attendues. Nous avons observé le contraire. Tout d'abord, le signal de PCR restait assez faible indiquant un problème évident dans la viabilité des cellules $\Delta sta1$. Nous n'avons également jamais pu obtenir la croissance des colonies *pyrEF*⁻ $\Delta sta1$ sur un milieu solide. Pour obtenir ces colonies, à l'issue d'une trentaine de cycles de purification et de dilutions successives des cultures, les cellules ont été étalées sur un milieu solide complété par de l'uracile (20 mg.L⁻¹) et du 5-FOA (50 mg.L⁻¹). En présence de X-Gal, toutes les colonies Ura⁻ observées étaient bleues (Figure 51). Cela indique que les cellules présentent un phénotype Lac⁺ et Ura⁻. L'apparition des mutants Ura⁻ Lac⁺ n'est pas étonnante car le phénotype Ura⁻ peut apparaître à la suite de la recombinaison homologue entre les régions OUT-OUT (dans ce cas, ils sont forcément Lac⁻) mais aussi à la suite de mutations spontanées dans les gènes *pyrEF*⁺. Dans ce cas, les mutants sont Ura⁻, Lac⁺ et *sta1*⁺.

Seul ce type de mutant est viable après cette sélection. Quinze colonies Ura- LacS+ ont été vérifiées par PCR. Aucune ne porte allèle $\Delta sta1$ sur son chromosome. Les mutants KO du gène *sta1*, dont le signal correspondant était observé par PCR ne sont pas viables

L'ensemble de ces résultats suggère très fortement que le gène *sta1* est un gène essentiel.

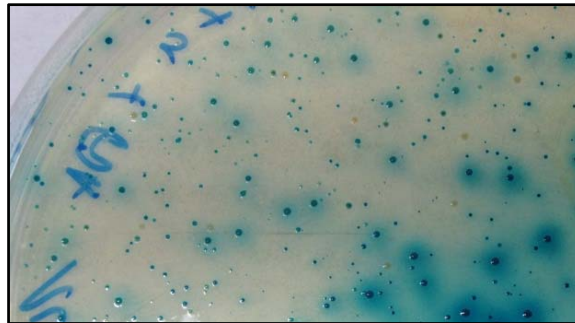


Figure 51 : Coloration au X-Gal des colonies obtenues à partir des cultures de transformants. Les colonies se colorent en bleues sur un milieu riche avec du 5-FOA et de l'uracile, en présence de X-Gal. Elles sont *lacS+* et *pyrEF-*. A partir des transformants *lacS+ pyrEF+*, seulement des mutants ponctuels *lacS+ pyrEF-* ont été obtenus.

Nous avons tenté de répéter cette expérience en présence de la deuxième copie du gène *sta1+* portée par un plasmide capable de se maintenir chez *S. islandicus*.

Dans ce vecteur (pSSR-Asta1), le gène *sta1* est exprimé sous le contrôle d'un promoteur inducible par l'arabinose (promoteur *araS* de *S. islandicus*). Pour pouvoir sélectionner les transformants portant ce vecteur, le gène de résistance à la simvastatine, *hmg* (issu de *S. islandicus*) a été cloné sur le vecteur pSSR-Asta1.

Le but était de démontrer l'apparition en présence d'arabinose, des mutants $\Delta sta1$ chromosomiques viables car la délétion de la copie résidente du gène *sta1* serait complétée par la présence d'une copie plasmidique et fonctionnelle du même gène. Ces mutants devraient également fortement ralentir sa croissance (voir mourir) en l'absence d'arabinose.

Parallèlement, en absence d'arabinose, il devrait toujours être impossible d'obtenir les mutants $\Delta sta1$.

La sélection des cellules transformées portant ce plasmide d'expression reposait sur la résistance à la simvastatine. Cette sélection est décrite comme très incertaine, dépendante de la souche utilisée (Whitaker and Zhang, 2010) et même de l'état du milieu, liquide ou solide. Malheureusement, nous n'avons pas trouvé la dose utilisable de simvastatine permettant de sélectionner les transformants portant le plasmide *ad hoc*. En conséquence, l'approche du « gène de sauvetage » n'a pu être utilisée pour élucider l'essentialité du gène *sta1*.

D'après les indications que nous disposons, le rôle de Sta1 semble être essentiel au sein de la cellule. Mais, la question concernant le rôle biologique de Sta1 et sur l'étendue de son régulon reste ouverte.

4. Nouvelles cibles et mécanisme inédit de régulation transcriptionnelle par une protéine de la famille RHH codée par le virus d'archées hyperthermophiles SIRV1

C. Danioux, N. Peixeiro, A. Joubert, O. Piétrement, P. Forterre, D. Prangishvili, E. Le Cam, G. Sezonov

1. Résumé

L'analyse structurale et fonctionnelle de la protéine SvtR, codé par le virus *Sulfolobus islandicus rod-shaped 1* (SIRV1), qui infecte la crénarchée *S. islandicus*, révèle un motif de liaison à l'ADN de type RHH. SvtR est impliquée dans la régulation transcriptionnel de deux gènes viraux : son propre gène *gp08* et le gène *gp30* codant pour une protéine structurale (Guillière *et al.*, 2009). Cependant, son mécanisme d'action sur les promoteurs sous son contrôle ainsi que la liste complète de ses cibles dans le génome viral restaient à établir.

Tout d'abord, le site consensus reconnu par la protéine SvtR a été défini à l'aide d'une mutagénèse systématique d'un de ses sites déjà caractérisé et de l'analyse de fixation de SvtR à 78 versions mutées de ce site. La recherche *in silico* du site consensus déterminé par cette approche a permis d'identifier douze sites candidats pouvant fixer SvtR, dont quatre étaient déjà connus. La liaison de SvtR à six nouveaux candidats a été confirmée par les expériences d'AG-EMSA et pour quatre d'entre eux, le rôle prédit de SvtR a été confirmé par des expériences de transcription *in vitro*. Selon la cible, SvtR agit soit comme un activateur, soit comme un répresseur transcriptionnel.

Pour certains des gènes régulés par SvtR, le site de liaison est situé entre soixante et quatre-vingt nucléotides en amont des promoteurs correspondants. En prenant comme modèle le promoteur du gène *gp30*, nous avons pu démontrer indirectement, par l'analyse par *footprinting* à la DNase, ainsi que directement, par l'observation sous microscope AFM/EM, la polymérisation de la protéine SvtR depuis son site de *binding* initial jusqu'au promoteur. Une certaine concentration intracellulaire de SvtR sera nécessaire pour déclencher sa polymérisation indispensable pour atteindre les promoteurs régulés à distance, faisant de SvtR « une horloge moléculaire » du cycle viral.

A notre connaissance, ce mécanisme peu commun de régulation à distance n'a jamais été décrit chez les *Archaea*.

2. Présentation de l'étude - introduction

Des publications récentes décrivant plusieurs membres de la famille *Rudiviridae* ont établi ces virus comme des modèles appropriés pour l'étude approfondie de la biologie des virus des Archées. Ces modèles ont déjà été utilisés pour décrire le mécanisme d'infection, les interactions hôte-virus (Bize *et al.*, 2009, Quax *et al.*, 2010; 2011; Snyder *et al.*, 2013; Queminn *et al.*, 2013), la réplication du génome viral (Peng *et al.*, 2010) ainsi que la régulation de la transcription (Guillière *et al.*, 2009; Okutan *et al.*, 2013; Quax *et al.*, 2013; Peixeiro *et al.*, 2013). Malgré de fortes avancées, les aspects moléculaires de la biologie des virus des crénarchées sont encore peu connus.

Les analyses *in silico* révèlent la diversité des régulateurs transcriptionnels codés par les différents virus d'archées (Prangishvili *et al.*, 2006). Les motifs structuraux les plus répandus dans ces protéines sont les motifs HTH (*Helix-Turn-Helix*) et RHH (*Ribbon-Helix-Helix*). Les protéines avec ce dernier

domaine sont codées par presque tous les virus de crénarchées (Prangishvili *et al.*, 2006) et par certains des virus d'euryarchées (Geslin *et al.*, 2007). En outre, un certain nombre de virus d'archées codent également pour des protéines contenant des domaines doigts de zinc et *swapped-hairpin* (Prangishvili *et al.*, 2006) comme cela a été développé dans la partie Introduction (1.4.2.e).

Les protéines RHH sont des régulateurs transcriptionnels bien connus, présents dans les domaines *Bacteria* et *Archaea* et leurs virus respectifs, où ce motif de liaison à l'ADN apparaît être une structure très fréquente (Schreiter and Drennan, 2007). De façon remarquable, les protéines RHH semblent être absentes chez les *Eukarya* (Schreiter and Drennan, 2007). Bien que plus de quatre mille protéines RHH soient actuellement prédites par les analyses *in silico*, seules quelques-unes ont été étudiées expérimentalement (Schreiter and Drennan, 2007).

Les structures de vingt-unes protéines RHH sont actuellement disponibles (Guillière *et al.*, 2009; Peixeiro *et al.*, 2013; Gomis-Ruth *et al.*, 1998; Somers and Phillips, 1992). Parmi elles, le répresseur méthionine MetJ (Somers and Phillips, 1992), le régulateur du nombre de copies de plasmides CopG (Gomis-Ruth *et al.*, 1998; Costa *et al.*, 2001; del Solar *et al.*, 2002; Hernandez-Arriaga *et al.*, 2009) et les répresseurs Mnt et Arc du bactériophage P22 (Raumann *et al.*, 1994; He *et al.*, 1992) peuvent être considérées comme les protéines RHH prototypiques. Le motif RHH consiste en un brin β N-terminal et deux hélices α connectées par un court tour. Les protéines RHH forment des dimères fonctionnels entrelacés à travers la formation d'un feuillet β des brins N-terminaux. Toutes les protéines RHH lient l'ADN comme des oligomères d'ordre élevé et établissent des contacts au niveau de répétitions inversées ou en tandem dans les séquences opératrices. Une présentation détaillée des protéines RHH est faite dans le Chapitre Introduction (1.4.2.5).

L'abondance des gènes codant pour des protéines appartenant à la famille RHH souligne leur importance dans la régulation transcriptionnelle des gènes viraux et chromosomiques (Krupovic *et al.*, 2012). Malgré leur abondance dans les génomes des virus d'archées, seuls deux régulateurs transcriptionnels de virus d'archées, les protéines SvtR codée par le virus SIRV1 (Guillière *et al.*, 2009) et AvtR codée par le virus AFV6 (Peixeiro *et al.*, 2013) ont été étudiés en détails jusqu'ici.

Chez les archées, la transcription est effectuée par des protéines spécialisées qui ressemblent à celles retrouvées chez les eucaryotes. Si le fonctionnement de la machinerie basale de transcription est assez bien compris, il n'en est pas de même pour la régulation de son activité. Nous avons été particulièrement intéressés d'étudier le mécanisme de régulation de la transcription dans un modèle biologique où l'activité d'une version simplifiée de l'appareil de transcription eucaryote est contrôlée par des protéines régulatrices de type « bactérien ».

La protéine SvtR (*Sulfolobus viral transcriptional Repressor*) codée par le rudivirus SIRV1, qui infecte la crénarchée hyperthermophile *S. islandicus* a été caractérisée comme un régulateur transcriptionnel (Guillière *et al.*, 2009). La structure RMN de la protéine révèle un domaine RHH typique (Figure 52). La protéine SvtR forme un dimère et lie l'ADN par son feuillet β . Certaines de ces cibles ont été identifiées dans le génome de SIRV1. Il a été démontré que la protéine se lie aux régions promotrices de deux opérons localisés aux extrémités du génome et qui codent pour des protéines sans fonctions connues. SvtR autorégule l'expression de son propre gène *gp08* et réprime la transcription du gène *gp30* qui code pour une protéine structurale de la capsid virale. Récemment, l'étude par l'approche de l'ARN-seq (Quax *et al.*, 2013) a donné une vision globale du

transcriptome d'une souche de *S. islandicus* infectée par SIRV2 permettant de comparer les données *in silico* et *in vitro* déjà disponibles.

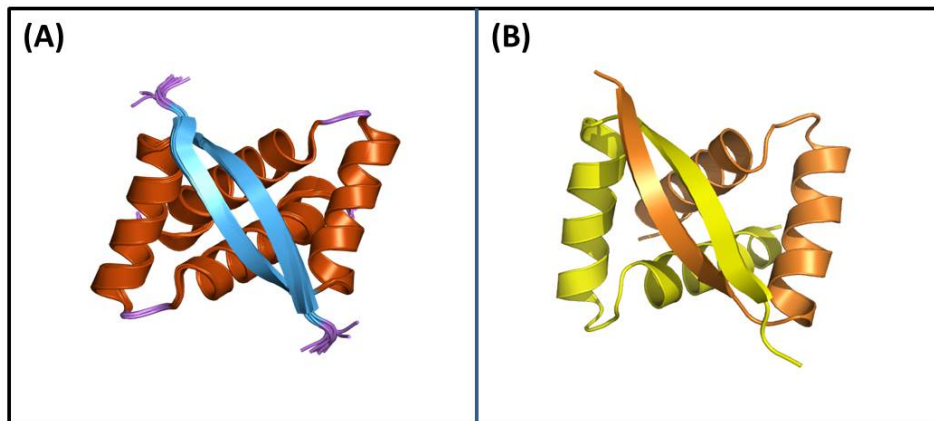


Figure 52 : Structures des dimères des protéines SvtR (A) et CopG (B) de la famille des protéines RHH. (A) SvtR est codée par le virus SIRV1 (Guillière *et al.*, 2009) et (B) CopG est codée par le plasmide streptococal pMV158 (Gomis-Rüth *et al.*, 1998). Ces deux protéines présentent un motif de liaison à l'ADN *Ribbon-Helix-Helix* (RHH).

Dans cette étude, nous nous sommes focalisés sur la protéine SvtR de 6.6kDa, codée par le ruidivirus SIRV1. Cette protéine est également présente chez SIRV2 (Prangishvili *et al.*, 2013).

Bien que la structure de cette protéine et sa fonction aient été précédemment décrites en détails (Guillière *et al.*, 2009), son mécanisme d'action et la liste complète de ses cibles dans le génome du virus restaient à établir au début de mon projet de thèse.

Au cours de ce travail, le site consensus reconnu par la protéine SvtR a été défini par une analyse mutationnelle systématique. Au moins dix gènes viraux (22% de tous les gènes) ont été révélés sous le contrôle de SvtR et pour quatre d'entre eux, le rôle prédit de SvtR a été confirmé par EMSA et par transcription *in vitro*.

Nous avons également démontré que pour certaines de ses cibles la régulation par SvtR implique la polymérisation de la protéine sur une longue région qui sépare son site principal et le promoteur régulé. Ce mode de régulation rappelle celui de AvtR déjà décrit par notre équipe (Peixeiro *et al.*, 2013). Cela indique qu'il ne s'agit peut-être pas d'un mécanisme isolé.

La caractérisation détaillée de cibles de SvtR permettra de mieux comprendre les mécanismes de régulation du cycle d'infection du virus modèle d'archée SIRV1/2.

3. Matériels et Méthodes

a. Clonage, expression et purification de la protéine

Le protocole pour purifier la protéine SvtR (gène *gp08* ou ORF56b chez le virus SIRV1) est décrit par Guillière *et al.*, 2009.

b. Oligonucléotides

Les oligonucléotides ont été achetés auprès de Sigma-Aldrich. L'ADN double brin a été obtenu en hybridant les oligonucléotides simple brin correspondants. Pour les expériences PAGE, les oligonucléotides ont été marqués au γ -P³² à l'aide de la T4 kinase (Fermentas®) et de suite hybridés avec leur oligonucléotide complémentaire non-marqué pour obtenir de l'ADN double brin.

c. Expériences d'EMSA (*Electrophoretic gel mobility shift assays*)

Des fragments d'ADN double brin de 39pb issus de la région promotrice du gène *gp30* (position 16903-16941 de la séquence du variant VIII du virus SIRV1; NP_666596.1) ont été obtenus en utilisant les oligonucléotides commerciaux simple brin. Chaque nucléotide du site de liaison précédemment identifié (Guillière *et al.*, 2009) en amont du gène *gp30* a été substitué exhaustivement à partir de la séquence initiale (Tableau 13). Des transitions (T vers C et A vers G, et vice versa) et des transversions (T vers G et A vers C, et vice versa) de nucléotides ont été réalisées. Comme témoin négatif pour les expériences de *binding* avec SvtR, nous avons utilisé un fragment d'ADN de trente-neuf nucléotides ne provenant pas de la séquence de SIRV1 (contrôle non spécifique).

Des fragments d'ADN double brin de 60pb issus de la région promotrice du gène *gp30* ont été obtenus en utilisant les oligonucléotides commerciaux simples brin. La séquence entre le site de liaison de SvtR et la région promotrice du gène *gp30* a été mutée pour étudier l'implication de sites secondaires hypothétiques de la protéine SvtR dans l'oligomérisation de la protéine.

Chaque fragment double brin marqué au γ -P³²(1.5pmol, ~75ng) a été incubé en présence de 2 μ l de 50% (v/v) glycérine et 1 μ g d'ADN non-spécifique (sperme de saumon) dans un volume final de 20 μ l de tampon (50 mM Tris-HCl, pH 8.0, 75 mM KCl, 25 mM MgCl₂ et 1 mM DTT) pendant 15 minutes à 68°C, avec des concentrations différentes de SvtR (de 0 à 15 μ M). Puis, les complexes ADN-protéines ont été migrés dans un gel non dénaturant 8% 37.5:1 acrylamide/bisacrylamide. La migration a eu lieu dans du tampon TBE (89mM Tris-Borate, 2mM NaEDTA, pH 8.3). Après la migration, le gel a été séché par un vacuum, exposé sur des Amersham Biosciences Hyperfilm™ MP et développé avec une machine Kodak X-OMAT 2000.

d. Agarose Gel EMSA (AG-EMSA)

Les fragments d'ADN ont été générés par PCR en utilisant les *primers* décrits dans le Tableau 13.

La mobilité des fragments d'ADN générés est analysée dans des gels d'agarose. Les fragments de chaque groupe ont été incubés dans 20 μ l de tampon de *binding* (10 mM Tris-HCl, pH 8.0, 10 mM HEPES, 1 mM EDTA, 1 mM dithiothreitol, 50 mM KCl, 50 μ g/ml sérum albumine bovine) pendant 15 minutes à 68°C avec des quantités croissantes de SvtR (de 0 à 2 μ M). 5 μ l de tampon de charge (20

mm Tris-HCl, 10 mm EDTA, 50% glycérol) a été ajouté et 20µl de l'échantillon a été analysé dans des gels d'agarose 1.2%. Les échantillons ont été migrés à 135V pendant 10 minutes, suivi d'une heure de migration à 100V et visualisés sous lumière UV après coloration au bromure d'éthidium.

e. Expériences de transcription *in vitro* (IVT)

L'ARN polymérase de *S. solfataricus* et les facteurs de transcription TBP et TFB ont été purifiés comme décrit précédemment (Abella *et al.*, 2007). Des essais de transcription *in vitro* (IVT) ont été réalisés pour évaluer l'influence de la protéine SvtR sur la transcription des promoteurs des gènes *gp03*, *gp07*, *gp19*, *gp20* et *gp30*. Les produits de PCR des promoteurs des gènes *gp03* (318pb, position 1910-2227 dans la séquence du variant VIII de SIRV1; NP_666596.1) *gp07* (310pb, position 4977-5286), *gp19* (222pb, position 10964-11185), *gp20* (321pb, position 11610-11930) et *gp30* (363pb, position 16746-17108) (Tableau 13) sont produits à partir de l'ADN génomique de SIRV1 et clonés directement dans le vecteur de clonage pDrive (Qiagen) par T/A clonage. Un plasmide portant le promoteur T6 du virus SSV, utilisé comme contrôle dans les expériences d'IVT, a été généré comme décrit par Qureshi *et al.*, 1997.

Tableau 13 : Oligonucléotides utilisés pour les expériences. En rouge souligné : substitutions de nucléotides par rapport à la séquence initiale.

Nom	Séquence (5'-3')	Application
EMSA – définition du site consensus de SvtR		
SvtR_39mers	ATTACAATTTTTCAACAGCTAAATGCAAAAAATTAACAG	Contrôle positif
39mers	CAGTGTTTTGAGGTATGGATAGGTATGATGACGCCATGC	Contrôle négatif
SvtR_1	<u>T</u> TTACAATTTTTCAACAGCTAAATGCAAAAAATTAACAG	Transition
SvtR_2	A <u>C</u> TACAATTTTTCAACAGCTAAATGCAAAAAATTAACAG	Transition
SvtR_3	AT <u>C</u> ACAATTTTTCAACAGCTAAATGCAAAAAATTAACAG	Transition
SvtR_4	ATT <u>G</u> CAATTTTTCAACAGCTAAATGCAAAAAATTAACAG	Transition
SvtR_5	ATTAT <u>T</u> AATTTTTCAACAGCTAAATGCAAAAAATTAACAG	Transition
SvtR_6	ATTAC <u>G</u> ATTTTTCAACAGCTAAATGCAAAAAATTAACAG	Transition
SvtR_7	ATTACAG <u>T</u> TTTTCAACAGCTAAATGCAAAAAATTAACAG	Transition
SvtR_8	ATTACAAC <u>C</u> TTTTCAACAGCTAAATGCAAAAAATTAACAG	Transition
SvtR_9	ATTACAAT <u>C</u> TTTTCAACAGCTAAATGCAAAAAATTAACAG	Transition
SvtR_10	ATTACAATT <u>C</u> TTTCAACAGCTAAATGCAAAAAATTAACAG	Transition
SvtR_11	ATTACAATTT <u>C</u> TCAACAGCTAAATGCAAAAAATTAACAG	Transition
SvtR_12	ATTACAATTTTT <u>C</u> CAACAGCTAAATGCAAAAAATTAACAG	Transition
SvtR_13	ATTACAATTTTTT <u>T</u> AACAGCTAAATGCAAAAAATTAACAG	Transition
SvtR_14	ATTACAATTTTTT <u>C</u> GACAGCTAAATGCAAAAAATTAACAG	Transition
SvtR_15	ATTACAATTTTTT <u>C</u> A <u>G</u> CAGCTAAATGCAAAAAATTAACAG	Transition
SvtR_16	ATTACAATTTTTTCAAT <u>T</u> AGCTAAATGCAAAAAATTAACAG	Transition
SvtR_17	ATTACAATTTTTTCAAC <u>G</u> GCTAAATGCAAAAAATTAACAG	Transition
SvtR_18	ATTACAATTTTTTCAACA <u>A</u> CTAAATGCAAAAAATTAACAG	Transition
SvtR_19	ATTACAATTTTTTCAACAG <u>T</u> TAAATGCAAAAAATTAACAG	Transition
SvtR_20	ATTACAATTTTTTCAACAGC <u>C</u> AAATGCAAAAAATTAACAG	Transition
SvtR_21	ATTACAATTTTTTCAACAGCT <u>G</u> AATGCAAAAAATTAACAG	Transition
SvtR_22	ATTACAATTTTTTCAACAGCTA <u>G</u> ATGCAAAAAATTAACAG	Transition
SvtR_23	ATTACAATTTTTTCAACAGCTAA <u>G</u> TGCAAAAAATTAACAG	Transition
SvtR_24	ATTACAATTTTTTCAACAGCTAAA <u>C</u> GCAAAAAATTAACAG	Transition
SvtR_25	ATTACAATTTTTTCAACAGCTAAAT <u>A</u> CAAAAAATTAACAG	Transition
SvtR_26	ATTACAATTTTTTCAACAGCTAAATG <u>T</u> AAAAAATTAACAG	Transition

SvtR_27	ATTACAATTTTTCAACAGCTAAATGC <u>G</u> AAAAATTAACAG	Transition
SvtR_28	ATTACAATTTTTCAACAGCTAAATGCA <u>G</u> AAAAATTAACAG	Transition
SvtR_29	ATTACAATTTTTCAACAGCTAAATGCAA <u>G</u> AAATTAACAG	Transition
SvtR_30	ATTACAATTTTTCAACAGCTAAATGCAAAA <u>G</u> AATTAACAG	Transition
SvtR_31	ATTACAATTTTTCAACAGCTAAATGCAAAA <u>G</u> ATTAACAG	Transition
SvtR_32	ATTACAATTTTTCAACAGCTAAATGCAAAAA <u>G</u> TTAACAG	Transition
SvtR_33	ATTACAATTTTTCAACAGCTAAATGCAAAAAA <u>C</u> TAACAG	Transition
SvtR_34	ATTACAATTTTTCAACAGCTAAATGCAAAAAA <u>T</u> CAACAG	Transition
SvtR_35	ATTACAATTTTTCAACAGCTAAATGCAAAAAA <u>T</u> <u>G</u> ACAG	Transition
SvtR_36	ATTACAATTTTTCAACAGCTAAATGCAAAAAA <u>T</u> <u>T</u> <u>G</u> CAG	Transition
SvtR_37	ATTACAATTTTTCAACAGCTAAATGCAAAAAA <u>T</u> <u>T</u> <u>A</u> <u>T</u> AG	Transition
SvtR_38	ATTACAATTTTTCAACAGCTAAATGCAAAAAA <u>T</u> <u>T</u> <u>A</u> <u>C</u> <u>G</u> G	Transition
SvtR_39	ATTACAATTTTTCAACAGCTAAATGCAAAAAA <u>T</u> <u>T</u> <u>A</u> <u>C</u> <u>A</u>	Transition
SvtR_1bis	<u>A</u> TTACAATTTTTCAACAGCTAAATGCAAAAAA <u>T</u> <u>T</u> <u>A</u> <u>C</u> <u>A</u> G	Transversion
SvtR_2bis	<u>A</u> <u>G</u> TACAATTTTTCAACAGCTAAATGCAAAAAA <u>T</u> <u>T</u> <u>A</u> <u>C</u> <u>A</u> G	Transversion
SvtR_3bis	<u>A</u> <u>T</u> <u>G</u> ACAATTTTTCAACAGCTAAATGCAAAAAA <u>T</u> <u>T</u> <u>A</u> <u>C</u> <u>A</u> G	Transversion
SvtR_4bis	<u>A</u> <u>T</u> <u>T</u> <u>C</u> CAATTTTTCAACAGCTAAATGCAAAAAA <u>T</u> <u>T</u> <u>A</u> <u>C</u> <u>A</u> G	Transversion
SvtR_5bis	<u>A</u> <u>T</u> <u>T</u> <u>A</u> <u>A</u> ATTTTTCAACAGCTAAATGCAAAAAA <u>T</u> <u>T</u> <u>A</u> <u>C</u> <u>A</u> G	Transversion
SvtR_6bis	<u>A</u> <u>T</u> <u>T</u> <u>A</u> <u>C</u> <u>C</u> ATTTTTCAACAGCTAAATGCAAAAAA <u>T</u> <u>T</u> <u>A</u> <u>C</u> <u>A</u> G	Transversion
SvtR_7bis	<u>A</u> <u>T</u> <u>T</u> <u>A</u> <u>C</u> <u>A</u> <u>C</u> TTTTTCAACAGCTAAATGCAAAAAA <u>T</u> <u>T</u> <u>A</u> <u>C</u> <u>A</u> G	Transversion
SvtR_8bis	<u>A</u> <u>T</u> <u>T</u> <u>A</u> <u>C</u> <u>A</u> <u>A</u> <u>G</u> TTTTTCAACAGCTAAATGCAAAAAA <u>T</u> <u>T</u> <u>A</u> <u>C</u> <u>A</u> G	Transversion
SvtR_9bis	<u>A</u> <u>T</u> <u>T</u> <u>A</u> <u>C</u> <u>A</u> <u>A</u> <u>T</u> <u>G</u> TTTTTCAACAGCTAAATGCAAAAAA <u>T</u> <u>T</u> <u>A</u> <u>C</u> <u>A</u> G	Transversion
SvtR_10bis	<u>A</u> <u>T</u> <u>T</u> <u>A</u> <u>C</u> <u>A</u> <u>A</u> <u>T</u> <u>T</u> <u>G</u> TTTTTCAACAGCTAAATGCAAAAAA <u>T</u> <u>T</u> <u>A</u> <u>C</u> <u>A</u> G	Transversion
SvtR_11bis	<u>A</u> <u>T</u> <u>T</u> <u>A</u> <u>C</u> <u>A</u> <u>A</u> <u>T</u> <u>T</u> <u>T</u> <u>G</u> TCAACAGCTAAATGCAAAAAA <u>T</u> <u>T</u> <u>A</u> <u>C</u> <u>A</u> G	Transversion
SvtR_12bis	<u>A</u> <u>T</u> <u>T</u> <u>A</u> <u>C</u> <u>A</u> <u>A</u> <u>T</u> <u>T</u> <u>T</u> <u>G</u> CAACAGCTAAATGCAAAAAA <u>T</u> <u>T</u> <u>A</u> <u>C</u> <u>A</u> G	Transversion
SvtR_13bis	<u>A</u> <u>T</u> <u>T</u> <u>A</u> <u>C</u> <u>A</u> <u>A</u> <u>T</u> <u>T</u> <u>T</u> <u>A</u> ACAGCTAAATGCAAAAAA <u>T</u> <u>T</u> <u>A</u> <u>C</u> <u>A</u> G	Transversion
SvtR_14bis	<u>A</u> <u>T</u> <u>T</u> <u>A</u> <u>C</u> <u>A</u> <u>A</u> <u>T</u> <u>T</u> <u>T</u> <u>C</u> CACAGCTAAATGCAAAAAA <u>T</u> <u>T</u> <u>A</u> <u>C</u> <u>A</u> G	Transversion
SvtR_15bis	<u>A</u> <u>T</u> <u>T</u> <u>A</u> <u>C</u> <u>A</u> <u>A</u> <u>T</u> <u>T</u> <u>T</u> <u>C</u> ACAGCTAAATGCAAAAAA <u>T</u> <u>T</u> <u>A</u> <u>C</u> <u>A</u> G	Transversion
SvtR_16	<u>A</u> <u>T</u> <u>T</u> <u>A</u> <u>C</u> <u>A</u> <u>A</u> <u>T</u> <u>T</u> <u>T</u> <u>C</u> A <u>A</u> AGCTAAATGCAAAAAA <u>T</u> <u>T</u> <u>A</u> <u>C</u> <u>A</u> G	Transversion
SvtR_17	<u>A</u> <u>T</u> <u>T</u> <u>A</u> <u>C</u> <u>A</u> <u>A</u> <u>T</u> <u>T</u> <u>T</u> <u>C</u> A <u>C</u> GCTAAATGCAAAAAA <u>T</u> <u>T</u> <u>A</u> <u>C</u> <u>A</u> G	Transversion
SvtR_18bis	<u>A</u> <u>T</u> <u>T</u> <u>A</u> <u>C</u> <u>A</u> <u>A</u> <u>T</u> <u>T</u> <u>T</u> <u>C</u> A <u>C</u> A <u>T</u> CTAAATGCAAAAAA <u>T</u> <u>T</u> <u>A</u> <u>C</u> <u>A</u> G	Transversion
SvtR_19bis	<u>A</u> <u>T</u> <u>T</u> <u>A</u> <u>C</u> <u>A</u> <u>A</u> <u>T</u> <u>T</u> <u>T</u> <u>C</u> A <u>C</u> A <u>A</u> TAAATGCAAAAAA <u>T</u> <u>T</u> <u>A</u> <u>C</u> <u>A</u> G	Transversion
SvtR_20bis	<u>A</u> <u>T</u> <u>T</u> <u>A</u> <u>C</u> <u>A</u> <u>A</u> <u>T</u> <u>T</u> <u>T</u> <u>C</u> A <u>C</u> A <u>G</u> GAATGCAAAAAA <u>T</u> <u>T</u> <u>A</u> <u>C</u> <u>A</u> G	Transversion
SvtR_21bis	<u>A</u> <u>T</u> <u>T</u> <u>A</u> <u>C</u> <u>A</u> <u>A</u> <u>T</u> <u>T</u> <u>T</u> <u>C</u> A <u>C</u> A <u>G</u> CTAAATGCAAAAAA <u>T</u> <u>T</u> <u>A</u> <u>C</u> <u>A</u> G	Transversion
SvtR_22bis	<u>A</u> <u>T</u> <u>T</u> <u>A</u> <u>C</u> <u>A</u> <u>A</u> <u>T</u> <u>T</u> <u>T</u> <u>C</u> A <u>C</u> A <u>G</u> CTA <u>C</u> ATGCAAAAAA <u>T</u> <u>T</u> <u>A</u> <u>C</u> <u>A</u> G	Transversion
SvtR_23bis	<u>A</u> <u>T</u> <u>T</u> <u>A</u> <u>C</u> <u>A</u> <u>A</u> <u>T</u> <u>T</u> <u>T</u> <u>C</u> A <u>C</u> A <u>G</u> CTAA <u>C</u> TGCAAAAAA <u>T</u> <u>T</u> <u>A</u> <u>C</u> <u>A</u> G	Transversion
SvtR_24bis	<u>A</u> <u>T</u> <u>T</u> <u>A</u> <u>C</u> <u>A</u> <u>A</u> <u>T</u> <u>T</u> <u>T</u> <u>C</u> A <u>C</u> A <u>G</u> CTAA <u>A</u> <u>G</u> GCAAAAAA <u>T</u> <u>T</u> <u>A</u> <u>C</u> <u>A</u> G	Transversion
SvtR_25bis	<u>A</u> <u>T</u> <u>T</u> <u>A</u> <u>C</u> <u>A</u> <u>A</u> <u>T</u> <u>T</u> <u>T</u> <u>C</u> A <u>C</u> A <u>G</u> CTAAAT <u>A</u> CAAAAAA <u>T</u> <u>T</u> <u>A</u> <u>C</u> <u>A</u> G	Transversion
SvtR_26bis	<u>A</u> <u>T</u> <u>T</u> <u>A</u> <u>C</u> <u>A</u> <u>A</u> <u>T</u> <u>T</u> <u>T</u> <u>C</u> A <u>C</u> A <u>G</u> CTAAAT <u>G</u> AAAAAA <u>T</u> <u>T</u> <u>A</u> <u>C</u> <u>A</u> G	Transversion
SvtR_27bis	<u>A</u> <u>T</u> <u>T</u> <u>A</u> <u>C</u> <u>A</u> <u>A</u> <u>T</u> <u>T</u> <u>T</u> <u>C</u> A <u>C</u> A <u>G</u> CTAAAT <u>G</u> <u>C</u> AAAAA <u>T</u> <u>T</u> <u>A</u> <u>C</u> <u>A</u> G	Transversion
SvtR_28bis	<u>A</u> <u>T</u> <u>T</u> <u>A</u> <u>C</u> <u>A</u> <u>A</u> <u>T</u> <u>T</u> <u>T</u> <u>C</u> A <u>C</u> A <u>G</u> CTAAAT <u>G</u> <u>C</u> A <u>A</u> AATTAACAG	Transversion
SvtR_29bis	<u>A</u> <u>T</u> <u>T</u> <u>A</u> <u>C</u> <u>A</u> <u>A</u> <u>T</u> <u>T</u> <u>T</u> <u>C</u> A <u>C</u> A <u>G</u> CTAAAT <u>G</u> <u>C</u> A <u>A</u> AATTAACAG	Transversion
SvtR_30bis	<u>A</u> <u>T</u> <u>T</u> <u>A</u> <u>C</u> <u>A</u> <u>A</u> <u>T</u> <u>T</u> <u>T</u> <u>C</u> A <u>C</u> A <u>G</u> CTAAAT <u>G</u> <u>C</u> A <u>A</u> AATTAACAG	Transversion
SvtR_31bis	<u>A</u> <u>T</u> <u>T</u> <u>A</u> <u>C</u> <u>A</u> <u>A</u> <u>T</u> <u>T</u> <u>T</u> <u>C</u> A <u>C</u> A <u>G</u> CTAAAT <u>G</u> <u>C</u> A <u>A</u> AATTAACAG	Transversion
SvtR_32bis	<u>A</u> <u>T</u> <u>T</u> <u>A</u> <u>C</u> <u>A</u> <u>A</u> <u>T</u> <u>T</u> <u>T</u> <u>C</u> A <u>C</u> A <u>G</u> CTAAAT <u>G</u> <u>C</u> A <u>A</u> AATTAACAG	Transversion
SvtR_33bis	<u>A</u> <u>T</u> <u>T</u> <u>A</u> <u>C</u> <u>A</u> <u>A</u> <u>T</u> <u>T</u> <u>T</u> <u>C</u> A <u>C</u> A <u>G</u> CTAAAT <u>G</u> <u>C</u> A <u>A</u> AATTAACAG	Transversion
SvtR_34bis	<u>A</u> <u>T</u> <u>T</u> <u>A</u> <u>C</u> <u>A</u> <u>A</u> <u>T</u> <u>T</u> <u>T</u> <u>C</u> A <u>C</u> A <u>G</u> CTAAAT <u>G</u> <u>C</u> A <u>A</u> AATTAACAG	Transversion
SvtR_35bis	<u>A</u> <u>T</u> <u>T</u> <u>A</u> <u>C</u> <u>A</u> <u>A</u> <u>T</u> <u>T</u> <u>T</u> <u>C</u> A <u>C</u> A <u>G</u> CTAAAT <u>G</u> <u>C</u> A <u>A</u> AATTAACAG	Transversion

SvtR_36bis	ATTACAATTTTTCAACAGCTAAATGCAAAAAATTA <u>C</u> CAG	Transversion
SvtR_37bis	ATTACAATTTTTCAACAGCTAAATGCAAAAAATTA <u>A</u> AG	Transversion
SvtR_38bis	ATTACAATTTTTCAACAGCTAAATGCAAAAAATTAAC <u>A</u> G	Transversion
SvtR_39bis	ATTACAATTTTTCAACAGCTAAATGCAAAAAATTAACA <u>C</u>	Transversion
EMSA – Etudes des sites secondaires hypothétiques dans la région du gène gp30		
SvtR-60-wt	ACAATTTTTCAACAGCTAAATGCAAAAAATTAACAGATGTAGAATTGAAAGAACTTTTAG	Contrôle
SvtR-60-hyp3	ACAATTTTTCAACAGCTAAATGCAAA <u>CGATT</u> AACAGATGTAG <u>GCGTT</u> GAAAG <u>GACG</u> TTTAG	Muté
Footprinting à la DNase I – Etude de l'oligomérisation de SvtR		
gp30_363F gp30_363R	AGTTTAGGTTATCCATTTCTGCT AAACTTTACTAATTGGATATCTAG	Amplification PCR – Région du gène gp30 chez SIRV1 (363pb)
gp30_80	CGATAAAAAAGCTAATTTTTTTTATATAGTTGTCATTACA AAAAAATTAACAGATGTAGATTTTTTAGTAATGTTTATAAT	Site de SvtR absent
Microscopie – Etude de l'oligomérisation de SvtR		
gp30_504F gp30_504R	TACAAATTTTCACAGAGATTAGAAG TATCCAATTAGTAAAGTTTTCCAA	Amplification PCR Région du gène gp30 chez SIRV1 (504pb)
AG – EMSA et IVT – Amplification des régions promotrices des gènes cibles		
gp0304_318F gp0304_318R	GAAAAACAAGGCAATAAAT'TAAA AAACCGCTCTCAGTAGTTTCAAAT	Amplification PCR Promoteur des gènes gp03-04 (318pb)
gp07_310F gp07_310R	AGTGGAATTTCAAATACTGCTCCA TGGCTTAGCAATTTAGCTAGCTTA	Amplification PCR Promoteur du gène gp07 (310pb)
gp12_244F gp12_244R	TATGCAGATTGTGGGAATTTGGA ATACCAACTTCAGATGGAAAAAT	Amplification PCR Promoteur du gène gp12 (244pb)
gp14_268F gp14_268R	GGTTAACGGTAGACATTATCGTTT ACTGTTGGCTCTTGGTCTCAACTC	Amplification PCR Promoteur du gène gp14 (268pb)
gp19_222F gp19_222R	AGCAGATATGACAATTTAATAGTT TATGATCTTCCTTTGGCTAGCGTA	Amplification PCR Promoteur du gène gp19 (222pb)
gp20_321F gp20_321R	CTATATATTTCCCCAGTTCTCAAAC TTCAGAAACATTAGATATAGAAGA	Amplification PCR Promoteur du gène gp20 (321pb)
gp26_304F gp26_304R	GAAAATGCAGGGAAAGATTTTCTA CAGCTCCACCGAATTTCTACTCTA	Amplification PCR Promoteur du gène gp26 (304pb)
gp30_363F gp30_363R	AGTTTAGGTTATCCATTTCTGCT AAACTTTACTAATTGGATATCTAG	Amplification PCR Promoteur du gène gp30 (363pb)
gp33_315F gp33_315R	CAAATTCGAATTCCTCCAATAACA TATCTCCTCCAACTATAGAAGGTTT	Amplification PCR Promoteur du gène gp33 (315pb)
2160R	CAAAAAATACATATAGAACTACATC	Primer extension, gp03-04, transcrit de 42nt
5080R	CCATATAGAGAAGATTTATTGAAAA	Primer extension, gp07, transcrit de 80nt
11190R	ATTTTGCGTACCTTTGGCTATATGA	Primer extension, gp19, transcrit de 57nt
11720R	GACATTCATATCCTAAGAGTGCCAT	Primer extension, gp20, transcrit de 60nt
17068R	AATTAGGAACATTCTTTGCAGATGA	Primer extension, gp30, transcrit de 84nt

Les réactions de transcription *in vitro* sont réalisées avec 100ng du plasmide ADN correspondant, 0.2mm de NTP, 5ng de TBP, 5ng de TFB, 125ng d'ARN polymérase et des concentrations variantes de SvtR (0 à 15µM). Les réactions sont effectuées pendant 20 minutes à 68°C dans 50µl de tampon de transcription (50 mm Tris-HCl, pH 8.0, 75 mm KCl, 25 mm MgCl₂, and 1 mm dithiothreitol). Les réactions sont stoppées en ajoutant 250µl de tampon NEW (10 mm Tris-HCl, pH 8.0, 750 mm NaCl,

10 mM EDTA, 0.5% SDS, and 40 mg/ml de glycogène). L'ARN synthétisé *in vitro* est isolé par une extraction phénol-chloroforme suivie par une précipitation éthanol. Les produits de transcription sont détectés par *primer extension* en utilisant des *primers* spécifiques aux séquences pour les échantillons de promoteurs viraux décrits précédemment (Kessler *et al.*, 2004 et Qureshi *et al.*, 1997). Après l'ajout de 20µl de tampon de charge avec 50% de formamide, 20µl de l'échantillon dénaturé a été analysé dans un gel polyacrylamide 10% dénaturant. Les gels ont été fixés dans 10% acide acétique-10% éthanol, avant l'exposition et le développement.

f. Footprinting à la DNase I

Pour les expériences de *footprinting* à la DNase I, nous avons confectionné de longs fragments d'ADN double brin (100 ou 363pb), comportant ou non le site de fixation de la protéine SvtR et les régions en amont et en aval (Tableau 13).

Les échantillons ADN ont été incubés avec la protéine SvtR pendant 15 minutes à 68°C, dans du tampon 50 mM Tris-HCl, pH 8.0, 75 mM KCl, 25 mM MgCl₂ et 1 mM DTT. Les échantillons ont été traités 5 minutes avec 1U de DNase I (Roche). Les réactions ont été arrêtées en ajoutant 200µL de solution *stop* (10 mM Tris-HCl, pH 8.0, 750 mM NaCl, 10 mM EDTA, 0.5% SDS et 40µg/ml de glycogène). Puis, l'ADN a été précipité en présence de 300µL d'éthanol 100% et 8µL de ARNt. Les culots ont été resuspendus dans 20µL de bleu de charge (50% formamide). Les échantillons dénaturés ont été analysés dans un gel dénaturant 8% 19:1 acrylamide/bisacrylamide.

Après la migration, le gel a été séché par un vacuum, exposé sur des Amersham Biosciences Hyperfilm™ MP et développé avec une machine Kodak X-OMAT 2000.

g. Méthodes bio-informatiques – Recherche *in silico* des cibles SvtR

Un script « Perl » qui utilise une matrice de pondération pour la recherche de séquences similaires dans le génome de SIRV1 et des Sulfolobales a été écrit. La matrice a été construite avec les données sorties de la définition de la séquence consensus. Le script a analysé les deux brins du génome en utilisant une fenêtre de vingt-cinq nucléotides qui se déplace le long du génome. Pour la suite, seules les séquences détectées par le programme qui se trouvaient entre -50 et +100 de la position du premier codon du gène et qui avaient au moins 75% de la note maximale de la matrice ont été gardées.

h. Microscopie

i. Complexes ADN-protéines

Les études de fixation (liaison) de SvtR ont été réalisées en mélangeant 1,5 nM (concentration en molécules ou sites) d'ADN (construction d'ADN de 502pb avec le site de fixation de SVTR ou construction d'ADN de 534 pb sans le site à partir du pET15b, Tableau 13) avec 400 nM de protéine SVTR (soit un ratio 266 protéines par site/molécule d'ADN) dans un tampon Tris-HCl 10 mM, pH 8, NaCl 30 mM dans un volume final de réaction de 20 µL. Le temps d'incubation est de 10 min à température ambiante.

ii. Microscopie électronique à transmission (MET)

5 μ L de la réaction sont déposés sur des grilles de cuivre de 600 mesh recouvertes d'un film mince de carbone (moins de 5 nm) et fonctionnalisées avec un plasma de pentylamine (Dubochet *et al.*, 1971). Les grilles sont lavées avec une solution aqueuse à 2 % (w/v) d'acétate d'uranyl puis séchées. Les observations de microscopie électronique sont faites sur un microscope LEO Zeiss 902, en fond noir annulaire, avec élimination des électrons inélastiques en utilisant le spectromètre de perte d'énergie (mode *zero loss*) pour augmenter le contraste (Le Cam and Delain, 1995). Les images ont été enregistrées à des grossissements de 50000X, 85000X and 140000X avec une caméra CCD VELETA 2kx2k et le logiciel d'acquisition iTEM (Olympus Soft Imaging Solution).

iii. Microscopie à force atomique (AFM)

5 μ L de la réaction sont déposés sur une surface de mica, prétraitée au préalable avec 50 μ M de spermidine (Pastré *et al.*, 2006), pendant 1 minute. La surface de mica est ensuite rincée avec une solution d'acétate d'uranyle 2%, puis séchée sur papier filtre. L'analyse se fait à l'aide d'un AFM Multimode (Bruker) équipé d'une électronique de contrôle Nanoscope V. Les images sont obtenues en mode contact intermittent, afin d'éviter la dégradation des biomolécules, à l'aide de pointes AC160TS (Olympus). Les images sont enregistrées à la fréquence de 1 Hz et avec une résolution de 1024 \times 1024 pixels. Les images sont ensuite analysées à l'aide du logiciel d'acquisition du Nanoscope V et seul un filtre polynomial du 3^{ème} ordre est appliqué sur toutes les images pour soustraire le bruit.

4. Résultats

a. Identification d'un site de liaison consensus de SvtR.

Dans l'étude précédente publiée par notre équipe (Guillière *et al.*, 2009), quatre cibles de SvtR dans le génome de SIRV1 ont été mises en évidence. Il s'agit des promoteurs des gènes *gp03/04*, *gp08*, *gp30* et *gp43/44* dont *gp08* code pour le régulateur lui-même et *gp30* code pour la protéine majeure de la capsid (Figure 53). La régulation de ce gène, intervenant tard dans la formation des capsides devrait représenter une étape cruciale dans le cycle de SIRV1. Cependant, l'approche utilisée dans cette première étude pour identifier les cibles de SvtR n'était pas sans défauts. Les cibles ont été identifiées par la technique du retard sur gel en testant, en présence de SvtR, la mobilité d'une dizaine de fragments d'ADN représentant dans leur ensemble la totalité du génome de SIRV1. Il n'est pas exclu qu'un certain nombre de cibles de SvtR ait échappé à la détection par cette approche dont l'efficacité dépend de la taille des fragments d'ADN et des positions des sites putatifs à l'intérieur du fragment.

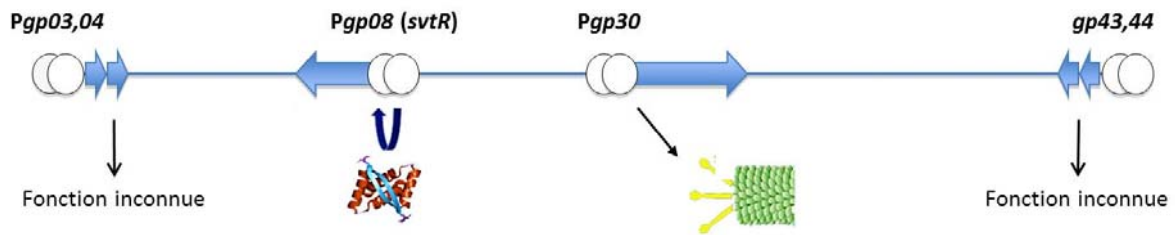


Figure 53 : Le régulon contrôlé par la protéine SvtR dans le génome de SIRV1 connu au départ de cette étude. La protéine se lie sur quatre régions du génome de SIRV1, en amont des gènes *gp03/gp04*, *gp08*, *gp30* et *gp43/44*. SvtR autorégule l’expression de son propre gène *gp08* et réprime la transcription du gène *gp30* qui code pour une protéine mineure de capside (Guillière *et al.*, 2009).

Pour pallier à ce problème, nous avons décidé de rechercher de nouveaux sites candidats par l’approche *in silico*. La séquence du génome de SIRV1 est disponible (Peng *et al.*, 2001) mais pas le site consensus de SvtR. Les deux sites de liaison de SvtR caractérisés dans l’étude précédente (*Pgp08* et *Pgp30*) ne permettaient pas d’établir la séquence consensus étant donné la faible conservation du site. Comme illustre la Figure 54, il peut être supposé que le site est composé de la suite des parties conservées et variables, mais il est impossible de se prononcer sur la conservation des nucléotides en comparant les séquences de seulement deux sites.

ATGATGTC ATAATTCAA AATCTTAA AAATG AGGAAAAATG	<i>Pg08</i>
ATTACA ATTTTCAAC AGCT AAATG CAAAAAATTAACAG	<i>Pg30</i>

Figure 54 : Comparaison de séquences entre les sites reconnus par SvtR des promoteurs des gènes *gp30* et *gp08*. Les nucléotides identiques entre chaque séquence sont en gras.

Dans le but d’établir la séquence consensus reconnue par SvtR, nous avons déterminé par l’analyse mutationnelle systématique les nucléotides fortement conservés dans un site d’une longueur de 39 pb (Figure 54). Etant donné la forte affinité de SvtR au site de liaison présent dans le promoteur du gène *gp30*, ce site a été choisi comme la séquence de référence pour réaliser sur sa base les 78 versions mutées et tester leurs affinités à SvtR. Chacun des trente-neuf nucléotides a été muté individuellement dans les deux sens, transition (T en C et C en T; A en G et G en A) et transversion (T en G et G en T; A en C et C en A) et la mobilité de chacun des soixante-dix-huit oligonucléotides ainsi générés a été testée individuellement en présence de SvtR (Figure 55). Pour chacun des nucléotides testés, les deux types de mutations avaient le même effet sur la mobilité des oligonucleotides correspondants. Les résultats de cette étude mutationnelle pour les mutations de type transition sont présentés dans la Figure 55. Pour simplifier la présentation seule la partie centrale de 25 pb de cette séquence est reproduite sur la Figure 55. Pour 6-7 nucléotides situés à l’extrémité du site, nous n’avons jamais vu le moindre effet de mutations sur la fixation de SvtR. Ces résultats donc été omis de la Figure 55.

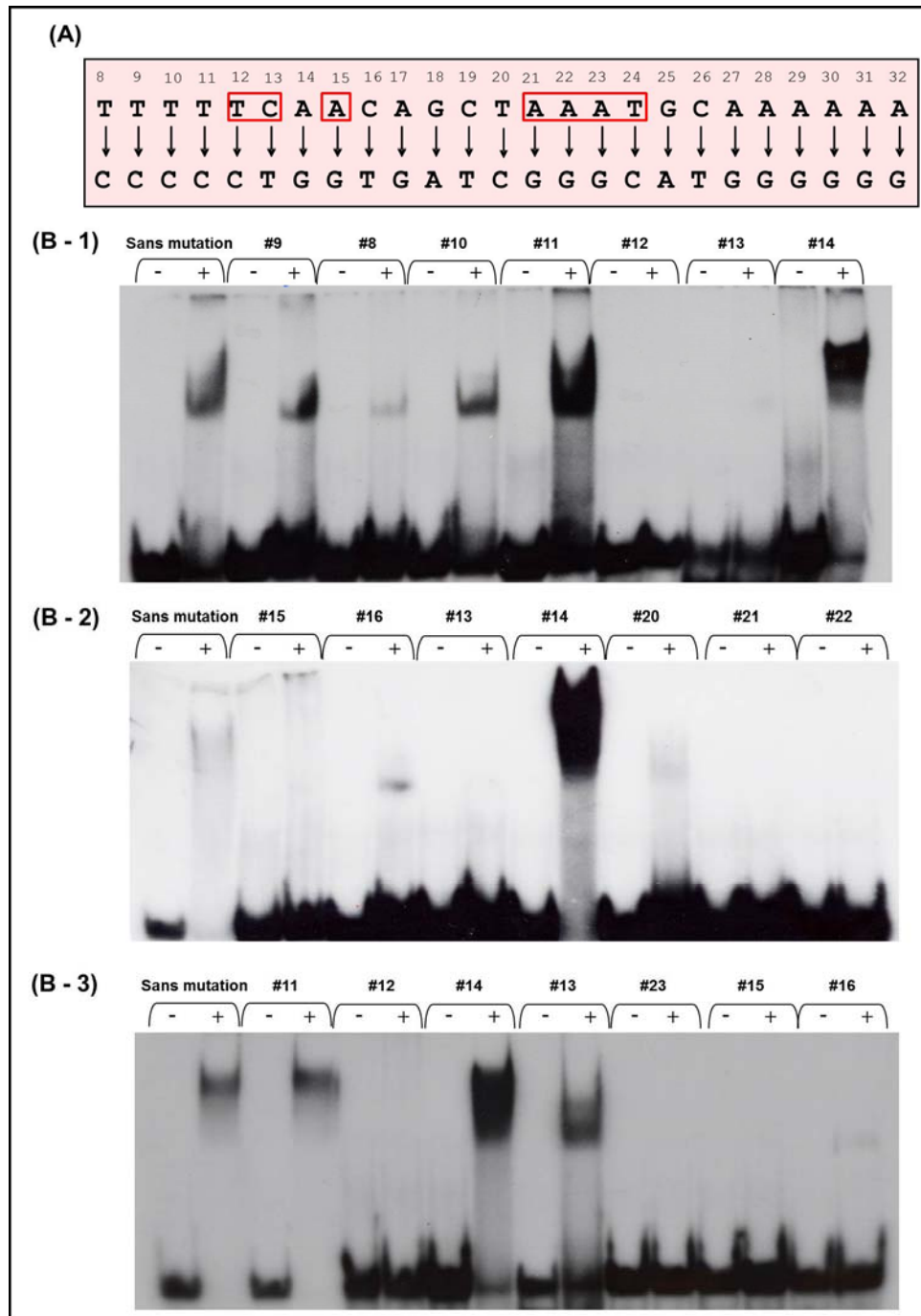


Figure 55 : Analyse mutationnelle des nucléotides du site de liaison de SvtR par transitions. (A) Représentation schématique de la stratégie de mutagenèse de nucléotide. Ligne supérieure – site wt de 39pb de liaison de SvtR de *gp30* (Guillière *et al.*, 2009). Par souci de clarté, seule la partie centrale de la séquence est représentée sur la figure. Aucun n’effet n’avait été constaté pour les nucléotides de chaque côté de cette partie centrale. Ligne inférieure – substitutions présentées des trente-neuf dérivés mutés générés. Les nucléotides conservés révélés par cette analyse sont encadrés. (B) Profils de migration de 17 des 78 oligonucléotides γ -P³²-radio marqués et mutés, analysés par PAGE, en absence (-) et en présence (+) de 2.5 μ M (300ng) de SvtR. Les positions des substitutions de type transition sont indiquées au-dessus des parties B1, B2 et B3. Les PAGE-EMSA ont été effectués en présence d’un fort excès (13X) d’ADN non spécifique compétiteur, poly[dI,dC]. L’absence de retard de migration indique que la substitution introduite affecte le *binding* de SvtR. En conséquence, le nucléotide correspondant était considéré comme essentiel pour la liaison de la protéine.

Les expériences de gels-retard ont permis d'estimer si un nucléotide est essentiel pour le *binding* de SvtR. En effet, comme le montre la Figure 55, le fragment témoin de 39pb non muté est retardé en présence de la protéine SvtR. Si le nucléotide muté est non essentiel pour le *binding* de la protéine, le fragment ADN sera retardé de la même façon que le témoin. A l'inverse, si le nucléotide muté est essentiel pour le *binding* de SvtR, la protéine ne sera plus capable de se lier au fragment avec la même efficacité et celui-ci ne sera plus retardé. Cet effet a pu être observé pour sept sur trente-neuf nucléotides testés indiquant que les nucléotides en positions 12, 13, 15, 21, 22, 23 et 24 étaient indispensables pour la liaison de la protéine SvtR à son site.

Cette information nous a permis de définir le site conservé minimal de SvtR comme la séquence de 13 nt TCnAnnnnnAAAT (n – non essentiel). De façon intéressante, ce site consensus ressemble par sa structure et sa longueur à celui du régulateur CopG (13pb – TGCAnnnnnTGCA) (considéré comme une protéine RHH prototype) avec deux parties de 4pb séparées par un *linker* de 5pb.

b. Identification *in silico* de tous les gènes candidats susceptibles d'être régulés par SvtR

A l'aide de la séquence de liaison consensus de 13pb définie précédemment, nous avons pu rechercher dans le génome de SIRV1 tous les gènes appartenant au régulon de SvtR.

Pour identifier les cibles putatives de SvtR, un script spécifique a été créé par D. Cortez (BMGE, Institut Pasteur) pour rechercher les séquences similaires à la séquence consensus dans le génome du virus SIRV1, en utilisant une fenêtre coulissante de seize nucléotides. Le script utilise une matrice pondérée, construite en affectant différents poids aux différents nucléotides, basés sur les résultats de l'analyse mutationnelle. Seules les séquences localisées à l'intérieur des régions promotrices prédites par l'étude transcriptionnelle de SIRV1 (Kessler *et al.*, 2004) et situées entre les positions -150 et +50 par rapport au codon ATG présumé de chaque gène ont été considérées. De même, seuls les sites candidats avec au moins 75% de la note maximale de la matrice ont été gardés. Au final, douze sites remplissant ces critères ont été retenus pour des analyses plus approfondies (Tableau 14). Les onze gènes cibles putatifs trouvés en amont des sites hypothétiques de SvtR sont répertoriés dans le Tableau 14.

Tous les sites de liaison prédits sont bien conservés dans les régions promotrices des gènes homologues correspondants chez le virus SIRV2, l'autre membre de la famille des *Rudiviridae*.

Cette liste comprend quatre cibles précédemment identifiées (Guillière *et al.*, 2009) *gp03/04* (transcrits en opéron), *gp08*, *gp30* et *gp43* mais elle comprend aussi sept nouveaux gènes candidats *gp07*, *gp12*, *gp14*, *gp19*, *gp20*, *gp26* et *gp33*. Il faut noter que le site localisé dans le promoteur *gp08* (le propre gène de SvtR) a le score le plus bas, comparé aux autres cibles retenues (Tableau 14). Cette observation suggère qu'il s'agit du site présentant la plus faible affinité à SvtR parmi tous les sites identifiés. Cependant, il ne s'agit en aucun cas d'un artefact car compte tenu des données de l'analyse fonctionnelle (Guillière *et al.*, 2009) son importance biologique ne peut pas être contestée.

Tableau 14 : La liste de l'ensemble des sites de liaison de SvtR dans le génome de SIRV1 (prédits ou confirmés). Les cibles sont prédites à partir de l'analyse *in silico* basée sur la séquence consensus du site de liaison de SvtR. Les nucléotides conservés sont en caractères gras. Seuls les sites localisés à l'intérieur des régions promotrices, entre -150 et +50 de chaque gène considéré. Ceux dont le score est plus que 75% de la note maximale de la matrice (soit >8) ont été retenus comme significatifs biologiquement. Les gènes *gp03* et *gp04* sont transcrits en opéron.

* la position de la *TATA-box* est hypothétique et n'a pas été prouvée *in vitro*, donc la position du site par rapport à la *TATA-box* est hypothétique.

** la distance est le nombre de nucléotides incluant le site et le *spacer* entre le site et la *TATA-box*.

Site	Gène le plus proche	Score	Distance à la <i>TATA-box</i> (en pb)**	Séquence	Fonction prédite
S1	<i>gp03/gp04</i>	8.2	-47*	TCATTTTCCAAATA	Protéine hypothétique
S2		8	-47*	TACGGCTGGAAAAGA	Protéine hypothétique
S3	<i>gp07</i>	8	-56	TACTTATAAAAAGA	Transporteur d'acides aminés
S4	<i>gp08</i>	8	+22	TCAAAATCTTAAAATG	Protéine SvtR
S5	<i>gp12</i>	8.5	+110	TAGATCTGCAAATT	Nucléase de type RecB; Protéine de type Cas4
S6	<i>gp14</i>	8.5	+70*	TCACITTTTAAAATT	Régulateur putatif, motif RHH
S7	<i>gp19</i>	8.05	+18	TAAAGAATAAAAATT	Protéine majeure de capsid
S8	<i>gp20</i>	8.2	-48*	TAAAAATAAAAATT	Protéine avec un domaine de liaison à l'ADN, <i>Zinc-finger</i>
S9	<i>gp26</i>	8.6	Entre +20 et +30*	TAAACTTACAAATT	Acétyl-transférase
S10	<i>gp30</i>	8.85	-73	TCAACAGCTAAATG	Protéine structurale mineure putative
S11	<i>gp33</i>	8.5	-28*	TCACGTTAGAAATC	Méthyl-transférase
S12	<i>gp43</i>	8	+20	TATTTTTTTTAAAATT	Protéine hypothétique

c. Les expériences EMSA révèlent une reconnaissance spécifique de certaines des nouvelles cibles prédites par la protéine SvtR

Pour vérifier les prédictions de l'analyse *in silico*, nous avons testé *in vitro* le *binding* de SvtR à huit (*gp03/gp04*, *gp07*, *gp12*, *gp14*, *gp19*, *gp20*, *gp26* et *gp33*) de onze de ces sites. Les cibles correspondant aux gènes *gp03/gp04*, *gp08*, *gp30* et *gp43*, ayant déjà été mises en évidence précédemment (Guillière *et al.*, 2009), seule la cible *gp03/gp04* a été testée dans ces nouvelles expériences. Pour pouvoir analyser l'interaction de ces sites putatifs avec SvtR, les régions du génome de SIRV1 les comportant ont été amplifiées par PCR donnant des fragments d'ADN de taille comprise entre 200 et 300 pb. Dans chacun de ces fragments le site conservé présumé de 13 nt se situe au milieu du fragment.

Les mobilités électrophorétiques de ces huit fragments ADN ont été étudiées en présence de la protéine SvtR. Nous avons utilisé l'approche d'AG-EMSA (**A**garose-**E**lectromobility **S**hift **A**ssay) publiée antérieurement (Guillière *et al.*, 2009) pour comparer dans le même gel, la mobilité des fragments ADN non marqués, en présence de concentration croissante de SvtR. Le promoteur de *gp30* identifié précédemment (Guillière *et al.*, 2009) porté par le fragment de 363 pb et ayant une forte affinité à SvtR a été utilisé comme un contrôle positif. Les résultats des expériences EMSA pour les cibles présumées, excepté *gp12*, sont présentés dans la Figure 56.

Sept (*gp03/gp04*, *gp07*, *gp14*, *gp19*, *gp20*, *gp26* et *gp33*) sur les huit fragments testés sont spécifiquement reconnus par SvtR, donnant une signification biologique aux prédictions de l'analyse *in silico*.

Les profils de retards observés ont permis de diviser les cibles testées en deux groupes. Le groupe I inclut les promoteurs *gp03*, *gp07*, *gp19* et *gp20* pour lesquels le retard est observé avec une quantité minimale de SvtR dans le mélange. Dans ces cas, l'efficacité de l'effet retard est similaire à celle observée pour *Pgp30*. L'effet de retard sur la mobilité des fragments d'ADN correspondants commence à être détectable à partir de 0.75-0.9 μ M (100-125ng) de SvtR. Une inspection visuelle des résultats des EMSA obtenus montre clairement que les sites situés dans les régions promotrices des gènes *gp07* et *gp19* sont reconnus de façon la plus forte comparé à toutes les autres cibles identifiées (Figure 56).

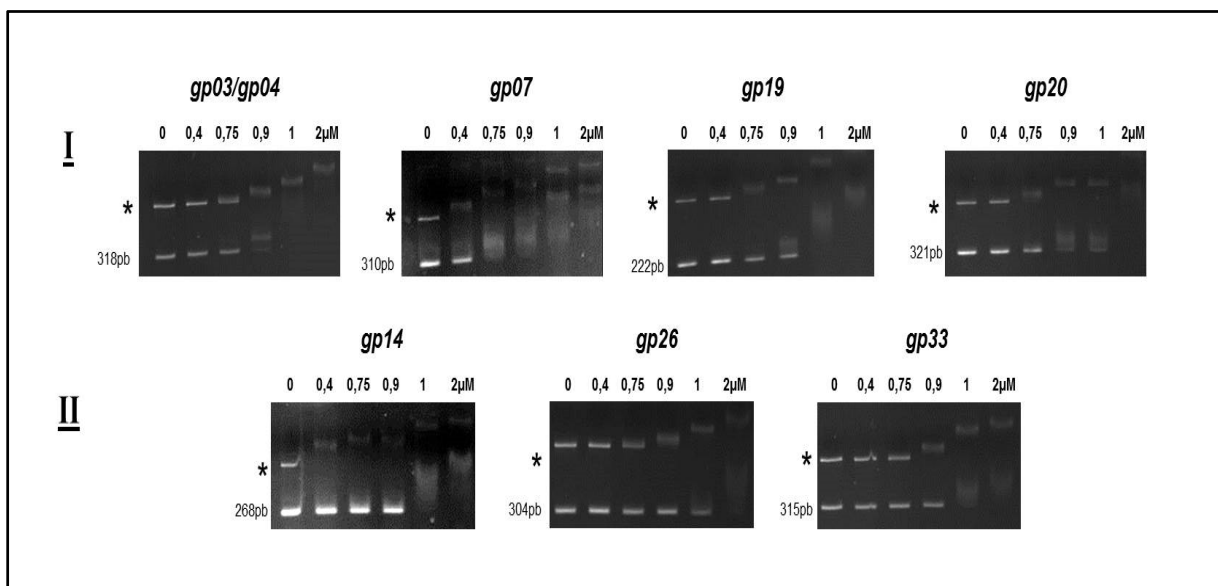


Figure 56 : Deux groupes de cibles de SvtR dans le génome de SIRV1 révélés par l'approche EMSA.

Le profil de migration des fragments d'ADN portant le site de liaison putatif de SvtR a été analysé par l'approche AG-EMSA (Guillière *et al.* 2009), en présence de concentration croissante de SvtR. Brièvement, les fragments ADN correspondants générés par PCR sont incubés avec SvtR, et ensuite migrés en gel d'agarose 1,2%. Dans ces conditions, le complexe SvtR-ADN migre plus lentement que l'ADN non lié et est visible comme un *smear* au lieu d'une bande ADN bien nette. Dans chacun des gels présentés dans la figure, le fragment contrôle représentant *Pgp30* a été migré simultanément avec la cible testée. La bande supérieure correspond au fragment de 363pb portant le promoteur *gp30* (contrôle positif) et la bande inférieure est toujours la cible testée de la taille indiquée sur la gauche de chaque gel. Le groupe I inclut quatre cibles (*gp03*, *gp07*, *gp19* and *gp20*) avec une affinité similaire à celle de SvtR pour *Pgp30*. Le groupe II inclut trois cibles (*gp14*, *gp26* and *gp33*) avec moins d'affinité. N'ayant donné aucun résultat, le gel correspondant à la cible *gp12* n'est pas montré.

Toutes les autres cibles prédites, excepté *gp12*, forment le groupe II (*gp33*, *gp14* et *gp26*) et requièrent plus de SvtR, entre 1 et 2 μ M (150-250ng) pour générer un signal retardé (Figure 56). Aucun *binding* n'a été constaté pour le promoteur *Pgp12* (données non montrées) dans les conditions utilisées. Selon nos critères d'analyse, cette cible a été considérée comme un faux-positif. En conséquence, elle a été exclue de l'analyse.

Afin de concentrer nos efforts sur les cibles candidates les plus prometteuses, nous avons décidé de restreindre dans un premier temps notre travail uniquement aux promoteurs du groupe I.

Les résultats observés valident partiellement les prédictions de l'analyse *in silico* et il était intéressant de mieux comprendre l'effet biologique de l'interaction des promoteurs concernés du groupe I avec le régulateur SvtR à l'aide de l'analyse fonctionnelle par l'approche de transcription *in vitro*.

d. Analyse fonctionnelle des nouvelles cibles SvtR par l'approche de transcription *in vitro*

Les facteurs transcriptionnels de la superfamille des RHH sont connus pour être capables de réguler positivement et négativement la transcription des gènes sous leur contrôle (Schreiter and Drennan, 2007). La position des sites de liaison de SvtR dans les régions promotrices ne donnent pas toujours d'indications claires sur le sens de régulation.

Il était intéressant de compléter l'analyse par l'approche EMSA réalisée sur quatre promoteurs cibles de SvtR du groupe I par l'analyse fonctionnelle. Pour cela, l'activité transcriptionnelle des promoteurs de ces gènes a été étudiée en utilisant un système de transcription *in vitro* (IVT) spécifique aux crénarchées établi dans notre laboratoire en collaboration avec le groupe de S. Bell (Qureshi *et al.*, 1997; Kessler *et al.*, 2004). La transcription *in vitro* est efficace en présence de trois éléments : les facteurs TBP et TFB et l'ARN polymérase, et son principe est illustré dans la Figure 57.

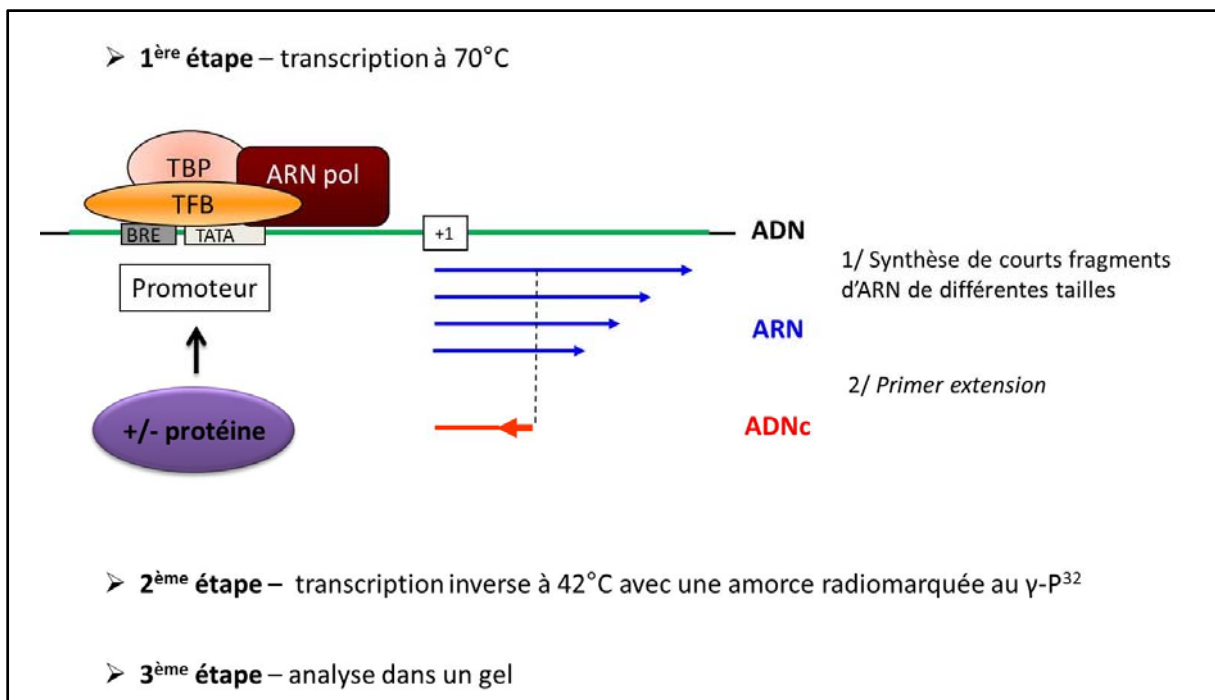


Figure 57 : Schéma d'explication de l'approche de transcription *in vitro*. Les promoteurs des gènes à tester sont clonés un vecteur plasmidique et sont incubés en présence de l'ARN polymérase de *S. solfataricus*, des facteurs de transcription TBP et TFB et du régulateur transcriptionnel en quantité variable. A partir des ARNs synthétisés, les ADNc radiomarqués sont obtenus par *primer extension* et ensuite quantifiés dans un gel (Qureshi *et al.*, 1997; Kessler *et al.*, 2004).

Comme décrit, plus en détails dans la partie Matériels et Méthodes, les régions promotrices de toutes les cibles présumées du groupe I ont été clonées dans un vecteur plasmidique *E. coli* et l'efficacité de transcription de ces promoteurs a été estimée en présence et en l'absence de la protéine SvtR. Les résultats de cette expérience sont présentés dans la Figure 58.

SvtR agit comme un répresseur fort sur le promoteur de l'opéron *gp03/gp04* et ceci de la même manière que pour la cible *gp08* (Guillière *et al.*, 2009). Les gènes *gp03/gp04* sont transcrits avec une faible efficacité en l'absence du régulateur mais leur transcription est considérablement diminuée, même en présence de faible concentration de SvtR (Figure 58).

L'effet d'action de SvtR sur le promoteur du gène *gp19* est toute à fait inverse. Dans ce cas, aucune transcription ne peut être détectée en l'absence de SvtR. A faible concentration, la présence de SvtR n'a pas d'effet tandis qu'à partir de 7.5µM, SvtR active la transcription de *Pgp19* et à 15µM de protéine l'activation est remarquablement forte (Figure 58).

Deux autres promoteurs testés du groupe I, *gp07* et *gp20*, ne changent pas leur activité en présence de SvtR. Aucun changement notable, ni vers l'activation, ni vers la répression ne sont détectés à aucune concentration de SvtR (Figure 58). Il peut être conclu que la protéine SvtR seule ne peut pas réguler le niveau d'activité de ces promoteurs. Cependant, il ne peut être exclu que la régulation de l'expression de *gp07* et *gp20* nécessite, en plus de SvtR, l'implication d'autres régulateurs.

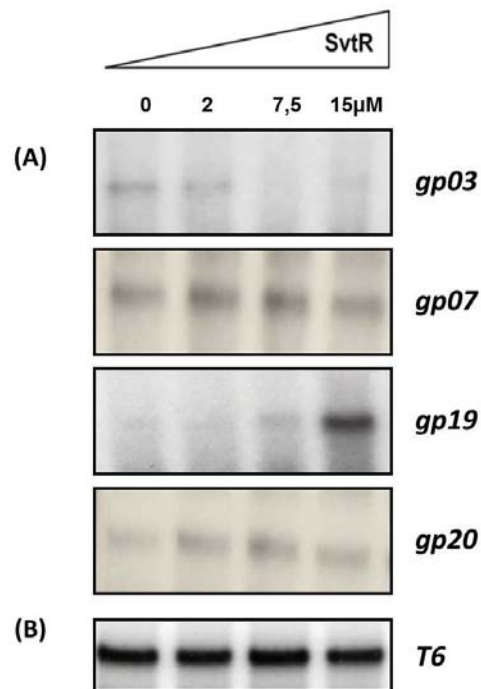


Figure 58 : Transcription *in vitro* de l'activité de SvtR sur les gènes cibles identifiés. L'expérience de transcription *in vitro* s'est déroulée en présence de concentration croissante de SvtR de 0 à 15µM. Tous les mélanges réactionnels contiennent des concentrations optimales d'ARN polymérase hautement purifiée à partir de *S. solfataricus* et des facteurs de transcription TBP et TFB. Les ADNc radiomarqués au γ - P³² ont été analysés dans un gel d'acrylamide 10%. (A) Analyse IVT des promoteurs *gp03*, *gp07*, *gp19* et *gp20*. (B) L'interaction de SvtR avec les cibles testées est spécifique puisqu'aucun effet sur le promoteur hétérologue T6 issu du virus non apparenté SSV1 n'est observé.

Des données transcriptomiques publiées récemment confirment globalement les résultats de l'analyse fonctionnelle sur toutes les cibles de SvtR (Quax *et al.*, 2013). Par exemple, comme montre la Figure 59, l'expression du gène *gp08* augmente fortement pour atteindre un maximum à t=2h post-infection, puis diminue jusqu'à t=7h p.i. Cela est en accord avec nos conclusions (Guillière *et al.*, 2009) sur le rôle de SvtR considéré comme répresseur de la transcription de son propre gène *gp08* (Guillière *et al.*, 2009). Quand une quantité suffisante de SvtR est atteinte dans la cellule, l'expression du gène *gp08* est réprimée. Les mêmes données transcriptomiques indiquent que l'expression du gène *gp30* augmente progressivement tout au long du cycle viral. D'après les résultats publiés en 2009 (Guillière *et al.*, 2009), il était attendu (compte tenu que SvtR a été décrit comme un répresseur transcriptionnel du gène *gp30*) que sa transcription devrait plutôt être réprimée mais ce n'est pas le cas. Les résultats de l'analyse transcriptomique sont donc en contradiction avec les résultats de l'analyse fonctionnelle ce que nous a ramené à refaire l'analyse fonctionnelle pour cette cible pour vérifier les conclusions concernant *gp30*.

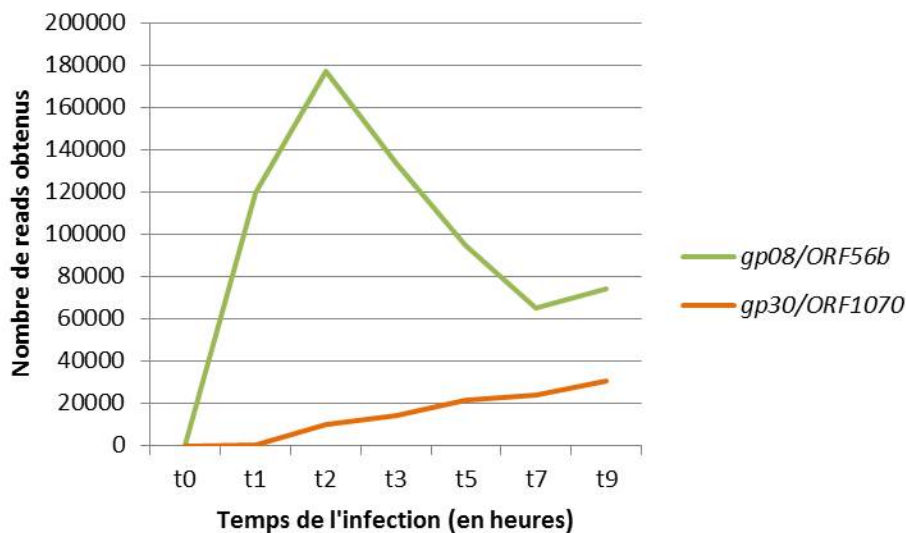


Figure 59 : Expression des gènes *gp08* et *gp30* au cours de l'infection de *S. islandicus* LAL14/1 par le virus SIRV2. Les données sont issues des expériences de ARN-seq réalisées par Quax *et al.*, 2013. Le nombre de *reads* est indiqué à différents temps (t=0, 1, 2, 3, 5, 7 et 9 heures) post-infection par SIRV2. Le régulateur SvtR autorégule sa propre expression et *gp30* est une de ses cibles dans le génome (Guillière *et al.*, 2009).

La transcription du gène *gp30* a été réétudiée en utilisant le système de transcription *in vitro* (IVT). Pour optimiser le système, nous avons synthétisé une nouvelle amorce utilisée pendant l'étape de transcription inverse. Située plus loin du site +1 d'initiation de la transcription par rapport à l'ancienne amorce utilisée par Guillière *et al.*, 2009, cette nouvelle amorce permet d'obtenir un fragment d'ADNc plus long (76 pb contre 39pb précédemment) ce qui facilite son identification et analyse. D'après l'analyse *in silico*, la région correspondant à cette amorce est moins susceptible de former des structures secondaires par rapport à l'ancienne amorce ce qui permet d'espérer moins d'artefacts pendant l'analyse.

Le résultat de cette nouvelle analyse fonctionnelle est présenté dans la Figure 60. Sans SvtR, la transcription de *gp30* est quasi inexistante. En présence de concentration croissante de protéine SvtR

(de 2 à 15 μM), l'intensité du signal augmente fortement ce qui permet de conclure sans ambiguïté que SvtR active la transcription du promoteur du gène *gp30*.

Pour conclure, l'approche IVT a permis de déterminer le rôle fonctionnel de SvtR dans le contrôle transcriptionnel d'au moins deux nouvelles cibles prédites dans l'étude *in silico*. La liste actualisée des gènes composant le régulon SvtR dans le génome du virus SIRV1 pour lesquels la fonction de SvtR est caractérisée comprend maintenant les quatre gènes suivants : *gp03/gp04*, *gp08*, *gp19* et *gp30* (Tableau 15). La protéine SvtR agit comme répresseur de la transcription pour les gènes *gp03* et *gp08* mais active la transcription des gènes *gp19* et *gp30*.

Ces données sont en bonne adéquation avec les résultats de l'analyse transcriptomique réalisée sur une culture de *S. islandicus* LAL14/1 infectée par SIRV2 (Quax *et al.*, 2013) et du transcriptome de mutant de *S. solfataricus* infectée par le même virus (Okutan *et al.*, 2013). Les deux types de données sont comparés dans la partie Discussion et Perspectives.

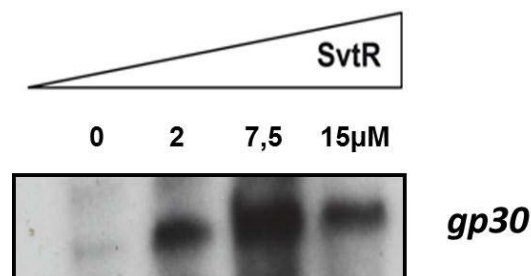


Figure 60 : Transcription *in vitro* de l'activité de SvtR sur le promoteur du gène *gp30*. L'expérience de transcription *in vitro* s'est déroulée en présence de concentration croissante de SvtR de 0 à 15 μM . Tous les mélanges réactionnels contiennent des concentrations optimales d'ARN polymérase hautement purifiée à partir de *S. solfataricus* et des facteurs de transcription TBP et TFB. Les ADNc radiomarqués au $\gamma\text{-P}^{32}$ ont été analysés dans un gel d'acrylamide 10%. La protéine SvtR semble avoir un effet activateur spécifique sur la transcription de *Pgp30* (à l'inverse des résultats précédents, Guillière *et al.*, 2009).

Tableau 15 : Cibles reconnues par la protéine SvtR dans le génome de SIRV1. Le *binding* de la protéine SvtR aux cibles putatives obtenues par l'analyse *in silico* a été testé. Celles reconnues par la protéine ont été résumées dans ce tableau. Les nucléotides conservés de la séquence consensus de liaison sont en gras. Les gènes *gp03* et *gp04* sont transcrits en opéron. ¹ Guillière *et al.*, 2009.* la position de la *TATA-box* est hypothétique et n'a pas été prouvée *in vitro*, donc la position du site par rapport à la *TATA-box* est hypothétique.** la distance est le nombre de nucléotides incluant le site et le *spacer* entre le site et la *TATA-box*.

Gène	Site de liaison de SvtR	Distance à la <i>TATA-box</i> (en pb)**	<i>Binding in vitro</i>	Activité <i>IVT</i>	Fonction
<i>gp03/gp04</i>	TCATTTTCC AAATA	-47*	++ ¹	Négative	Protéine hypothétique
<i>gp03/gp04</i>	TACGGCTGG AAAGA	-47*	++	Négative	Protéine hypothétique
<i>gp07</i>	TACTTATA AAAGA	-56	++	N. d.	Transporteur d'acides aminés
<i>gp08</i>	TC AAAT NNNN AAATG	+22	+ ¹	Négative ¹	Protéine SvtR
<i>gp14</i>	TCACTTTT AAATT	+70*	+	N. d.	Régulateur putatif, motif RHH
<i>gp19</i>	T AAAGAATAAAATT	+18	++	Positive	Protéine majeure de capsid
<i>gp20</i>	T AAAAATAAAATT	-48*	++	N. d.	Protéine avec un domaine de liaison à l'ADN, <i>Zinc-finger</i>
<i>gp26</i>	T AAACTTACAAATT	Entre +20 et +30*	+	N. d.	Acétyl-transférase
<i>gp30</i>	TC ACAGCTAAATG	-73	+++ ¹	Positive	Protéine structurale mineure putative
<i>gp33</i>	TC ACGTTAGAAATC	-28*	+	N. d.	Méthyl-transférase
<i>gp43</i>	T ATTTTTTTAAATT	+20	+ ¹	N. d.	Protéine hypothétique

e. Le mécanisme d'action de SvtR inclut l'oligomérisation sur l'ADN

Comprendre le mode d'action du régulateur SvtR sur certaines cibles était l'étape suivante de notre étude.

i. Positions des sites de liaison de la protéine SvtR comparé aux régions promotrices de ses gènes cibles

Le site de fixation de SvtR dans le promoteur de son propre gène est situé à proximité immédiate (position +7) du site +1 d'initiation de la transcription. La position du site très proche du +1 permet de supposer que la fixation de SvtR sur son site crée un obstacle stérique empêchant soit la formation du complexe de pré-initiation de transcription soit l'avancement de l'ARNP sur la matrice d'ADN le démarrage de la synthèse des transcrits courts. Ce mécanisme de régulation a déjà été décrit pour d'autres systèmes étudiés chez les archées et les exemples sont détaillés dans la partie Introduction-Régulation de la transcription.

Comme l'indique le tableau 15, sur onze sites reconnus par SvtR, cinq sont situés plutôt loin en amont du site +1 des gènes correspondants. Dans ces cas, la présence de SvtR à priori ne peut pas avoir d'influence directe, ni sur la formation du complexe de pré-initiation de transcription ni sur la progression de l'ARNP sur la matrice.

Pour mieux comprendre le mécanisme de régulation pour les gènes contenant le site de fixation de SvtR loin du promoteur, nous avons choisi comme modèle le promoteur du gène *gp30* pour lequel la distance entre le site de SvtR et la *TATA-box* est 58 nucléotides. Le choix de ce site s'explique par les résultats concluants de l'analyse fonctionnelle et la grande distance entre le site et le promoteur (58 nt). Ce gène code pour une protéine mineure de la capsid.

ii. La polymérisation de SvtR sur l'ADN cible lui permet de couvrir la région entre son site de fixation jusqu'à la séquence promotrice du gène *gp30*

La protéine type de la famille RHH, CopG, est capable de former des oligomères sur son ADN cible, son propre gène codé par le plasmide pMV158 (Costa *et al.*, 2001). Des expériences de *footprinting* à la DNase révèlent plusieurs régions protégées successives de 4-5nt séparées par des régions non protégées, couvrant quatre tours d'hélices sur la même double hélice ADN (Costa *et al.*, 2001). Nous avons émis l'hypothèse que SvtR est dotée de la même capacité et que son action à distance sur certains promoteurs peut s'expliquer par son oligomérisation lui permettant d'atteindre le promoteur cible.

Pour vérifier cette hypothèse nous avons réalisé des expériences de *footprinting* à la DNase I sur un fragment d'ADN de 363 pb représentant le promoteur du gène *gp30* incluant une longue région de 68 nt située en aval de son promoteur (les deux brins ont été observés).

Les résultats de cette analyse sont présentés sur la Figure 61.

Les résultats de l'analyse par le *footprinting* révèlent l'existence d'une longue région d'ADN protégée en présence de concentration croissante de SvtR sur le brin transcrit (Figure 61A) et également sur le brin non transcrit (Figure 61B).

L'oligomérisation observée est spécifique et strictement dépendante de la présence du site principal. En effet, les mêmes expériences réalisées avec un oligonucléotide de 80pb, contenant la *TATA-box* du gène *gp30* et la région en amont (environ 60pb) mais sans le site de liaison de la protéine SvtR, ne révèlent aucune oligomérisation même à forte concentration de SvtR (Figure 61C). Cette région s'étend sur cinquante-huit nucléotides, situés entre le site de SvtR et la *TATA-box* du promoteur de *gp30*. Etant une petite protéine de seulement 6.6kDa, SvtR peut protéger un fragment ADN de cette taille seulement si plusieurs sous-unités de la protéine forment un oligomère protégeant l'ADN cible. L'oligomérisation se propage des deux cotés du site principal (par souci d'esthétique, seules les parties entre le site de SvtR et les régions promotrices sont montrées dans les figures 61AB). Nous avons aussi réalisé les expériences de *footprinting* avec un autre oligonucléotide de 80pb, contenant le site de liaison de SvtR cette fois, et la région protégée a été observée comme pour les plus grands fragments.

La présence de nucléotides hypersensibles à l'action de la DNase apparaissent au niveau du site de liaison initial (marqués par deux points sur la figure 61AB) suggère que la partie centrale du site est libre de SvtR laissant l'ADN non protégé et flexible. De façon intéressante, le même phénomène de flexibilité a été observé dans le cas de la protéine CopG également capable d'oligomériser à partir de son site principal. Le modèle expliquant l'existence des nucléotides hypersensibles au centre du site principal de CopG et SvtR est présenté sur la Figure 61D.

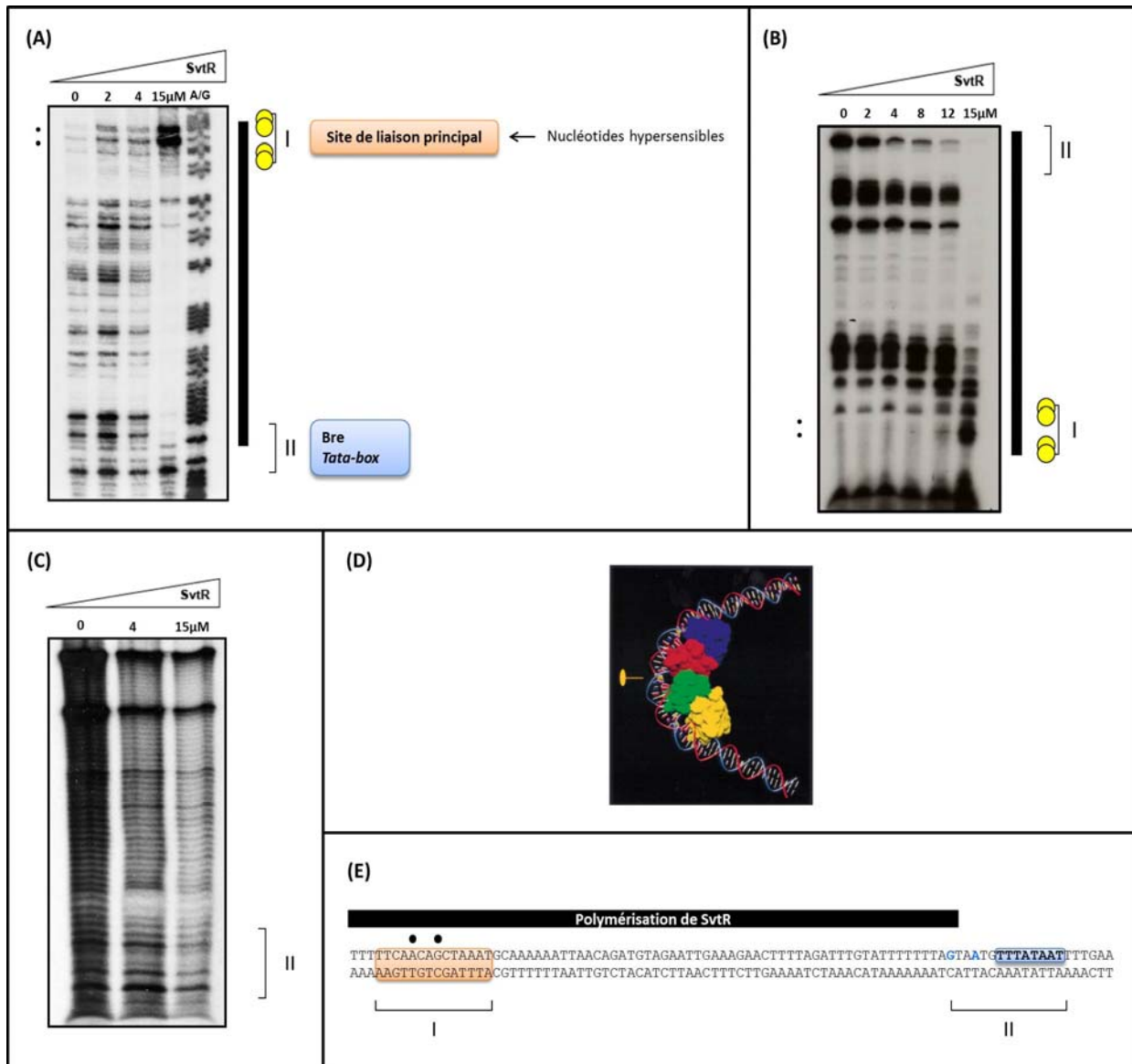


Figure 61 : Polymérisation de la protéine SvtR sur l'ADN, pour le gène *gp30* de SIRV1. (A, B, C) *Footprinting* à la DNase I du *binding* de la protéine SvtR à la région promotrice de *gp30*. Les concentrations appliquées de SvtR sont indiquées en μM . A+G représente la séquence correspondante. Le fragment ADN marqué utilisé contient la région promotrice du gène *gp30* et la région en amont avec (A) ou sans (C) le site de liaison de SvtR. Un fragment d'ADN de 363pb contenant le site de liaison de la protéine SvtR et la région en amont de la région promotrice du gène *gp30* a été amplifié par PCR et radiomarqué par du $\gamma\text{-P}^{32}$ pour l'expérience. Les brins transcrit (A) et non transcrit (B) ont été observés. (C) Un oligonucléotide de 80 pb a été conçu contenant la région en amont de la région promotrice du gène *gp30* sans le site de liaison de SvtR. (D) Modèle hypothétique de CopG en interaction avec un plasmide ADN double brin *in vitro* (Costa *et al.*, 2001). (E) Description résumée des données du *footprinting* à la DNase. La séquence ADN du site de liaison de SvtR au promoteur *gp30* est donnée. La TATA-box et la séquence BRE sont indiquées en caractères gras sur le brin codant; et la TATA-box est surlignée en bleu. Le site de liaison de SvtR est encadré en orange. Les régions protégées du clivage de l'ADNase I sont indiquées par une ligne noire; les sites hypersensibles sont signalés avec des points noirs. La position du site de liaison de SvtR (I) et les signaux des promoteurs (TATA-box et séquence BRE) (II) sont indiquées.

Le SvtR ne protège plus l'ADN matrice situé en aval du site BRE vers la séquence codante (bas du gel Figure 61A) ce qui indique que l'étendue de la polymérisation est contrôlée et le processus ne dépasse pas la séquence BRE (Figure 61A et 61E).

La Figure 61D résume les données présentées et montre la position exacte et l'étendue de la région protégée par SvtR polymérisée. Il peut être conclut, au moins pour le cas étudié, que l'action régulatrice de SvtR sur le promoteur de *gp30* inclut une étape d'oligomérisation permettant à SvtR d'atteindre le promoteur cible.

Dans le cas de CopG, la polymérisation est guidée par la présence de courts sites de liaison dégénérés placés en aval et en amont du site principal (Costa *et al.*, 2011) (Figure 62A). L'analyse de la séquence protégée par SvtR suggère également l'existence de sites de liaison dégénérés qui pourraient jouer le même rôle que dans le cas de CopG (Figure 62B). Un oligonucléotide de 60pb comportant le site principal de liaison de la protéine, les trois premiers sites dégénérés (*aat*, *gat* et *aact*) et la région promotrice du gène *gp30* a servi de témoin. Ce même oligonucléotide avec les sites de liaison secondaires hypothétiques mutés (voir Matériel et Méthodes) a été testé dans des expériences de gel-retard.

Le retard du fragment portant le site principal et les sites secondaires lié à l'oligomérisation de SvtR est observé. Cependant, dans le cas de la matrice d'ADN avec les sites secondaires mutés, le retard est observé en présence de quantité plus élevée de protéines (Figure 62C). Ceci suggère fortement l'implication de ces sites de liaison secondaires dans la polymérisation de la protéine SvtR mais également dans la stabilisation du lien entre SvtR et le site primaire.

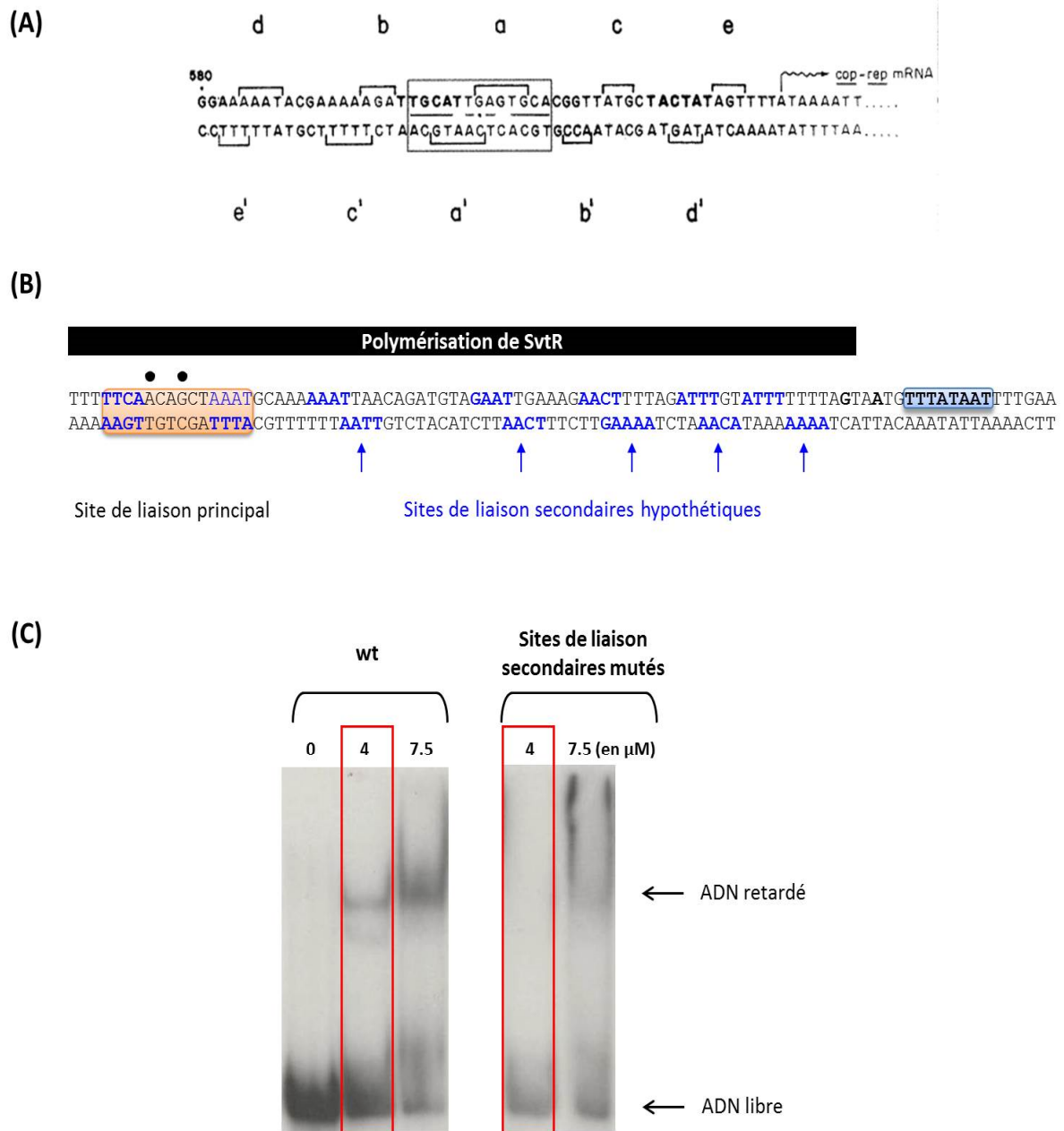


Figure 62 : Sites de liaison secondaires impliqués dans la polymérisation de la protéine SvtR dans le cas de la cible *gp30*. (A) Séquence nucléotidique de la région du plasmide ADN *wild-type* avec laquelle interagit CopG. Les crochets désignent les bases protégées par CopG du clivage lors du *footprinting*. Le site de liaison est encadré (Costa *et al.*, 2001). (B) Séquence nucléotidique du site de liaison de SvtR et de la région promotrice du gène *gp30*. Le site de liaison principal de SvtR est encadré. La *TATA-box* est surlignée en bleu. La région correspondante à la polymérisation de la protéine SvtR est indiquée par une ligne noire; les sites hypersensibles sont signalés avec des points noirs. Les sites de liaison secondaires hypothétiques sont indiqués en bleu. (C) *Binding* de la protéine SvtR à des oligonucléotides de 60 pb avec des sites de liaison secondaires *wt* et mutés. Les PAGE-EMSA ont été réalisés en présence d'un fort excès (13X) d'ADN compétiteur non spécifique poly[dl,c]. Les quantités de SvtR utilisées sont indiquées en μM . Le *binding* est observé à partir de 4 μM de SvtR pour l'oligonucléotide *wt*. Le *binding* est très faible pour l'oligonucléotide comportant les sites de liaison secondaires mutés.

iii. L'approche de microscopie confirme l'hypothèse d'oligomérisation de SvtR

Bien que la technique d'empreinte à la DNase I conforte l'hypothèse d'oligomérisation de SvtR sur l'ADN, nous avons cherché à obtenir une preuve directe de l'existence de ce phénomène.

Nous avons tenté d'observer le complexe SvtR-ADN par les approches de microscopie à force atomique (AFM) et électronique. Un fragment de 504pb (Tableau 13) plus long que celui utilisé pour les expériences de *footprinting*, mais ciblant la même région du gène *gp30* a été mis en contact avec un fort excès de SvtR à un ratio de 266 protéines/site/molécule d'ADN et les complexes ont été visualisés en microscopie électronique à transmission (Figure 63ABCDE) et à force atomique (Figure 63FGH).

Les images typiques sont présentées dans la Figure 63. En présence de SvtR (Figure 63C), le fragment d'ADN est plus épais qu'en l'absence de protéine (Figure 63A). L'ADN est clairement recouvert par la protéine oligomérisée ce qui le rend rigide (Figure 63CDEGH). De façon très intéressante, la région contenant le site principal de liaison de SvtR située aux 2/5 de la longueur totale du fragment correspond à l'endroit où la flexibilité de la molécule est préservée (cette région se situe à 171pb d'une des extrémités du fragment de 504pb). La position de cette région flexible correspond à la position des nucléotides hypersensibles à la DNase révélés par l'approche de *footprinting*.

L'approche de microscopie électronique (EM) appliquée à ce fragment d'ADN mis en présence de SvtR a permis d'observer une interaction intermoléculaire inattendue entre deux filaments (Figure 63CDE). Des structures caractéristiques en forme de croix, formées par des couples de filaments réunis par un point de nucléation sont observées (Figure 63DE). Il peut être supposé, compte tenu de la position de ce point de croisement, qu'il s'agit d'une interaction entre deux complexes ADN-protéines qui interagissent très probablement dans des régions correspondant au site principal de fixation de SvtR. L'interaction intermoléculaire détectée *in vitro* pourrait signifier que ces points de nucléation se forment également *in vivo* entre les différents sites de fixation de SvtR et sur la même molécule d'ADN qui représente le génome entier de SIRV1. Ceci suggère l'existence d'un autre niveau de régulation d'activité des gènes de SIRV1 qui passe par la formation des grandes boucles d'ADN et par le changement global de la conformation de l'ADN du génome viral.

La formation de ces structures est fortement spécifique et nécessite la présence à la fois de la protéine SvtR et de l'ADN contenant son site de fixation. En l'absence de SvtR, les structures en croix ne sont pas observées pour le fragment d'ADN de SIRV1 de 504pb contenant un site de SvtR (Figure 63A). Elles n'apparaissent pas dans le cas où SvtR est mélangée avec des fragments d'ADN hétérologue de 534pb provenant du plasmide pBR322 d'*E. coli* sans aucun site reconnu par SvtR. (Figure 63B).

Le même fragment d'ADN de 504 pb a été mis en contact avec la protéine SvtR et observé par l'approche de microscopie à force atomique (AFM) (Figure 63GH). Comparé au témoin sans protéine (Figure 63F), le fragment apparaît plus épais et plus « brillant » (Figure 63GH) témoignant que la protéine polymérise sur l'ADN. Comme dans le cas de l'étude au microscope électronique, la présence d'une région plus flexible comparée au reste de l'ADN est également observée. La position de cette zone de flexibilité correspond à la position du site principal de fixation de SvtR.

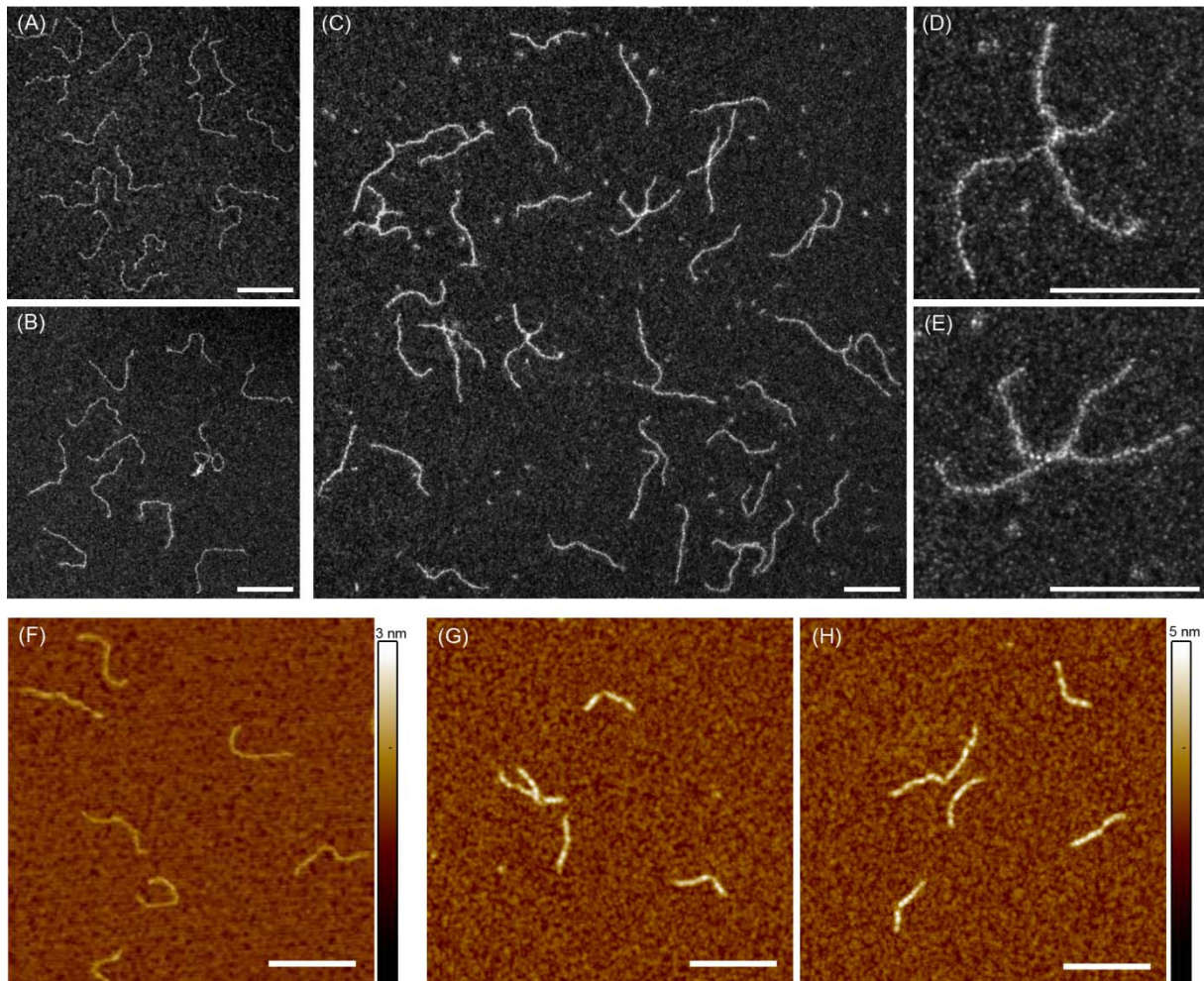


Figure 63 : Complexes SvtR-ADN sous microscope électronique (ABCDE) et à force atomique (FGH). (A) Témoin ADN 502 pb avec le site SvtR (à 1,5nM en sites/molécules), grandissement 50000X. (B) Témoin ADN 534 pb sans le site SvtR (1,5nM en molécules) avec SvtR (400nM), grandissement 50000X. (C) Complexes ADN 502pb 1,5nM en sites/molécules, avec SvtR (400nM), grandissement 50000X. (D,E) Zoom sur des complexes ADN-SVTR formant des croix, grandissement 140000X. Toutes les barres d'échelle représentent 100 nm. (F) Témoin ADN 502 pb avec le site SVTR (à 1,5nM en sites/molécules). (G-H) Complexes ADN 502pb (1,5nM en sites/molécules) avec SVTR (400nM). Toutes les barres d'échelle représentent 200 nm.

Les deux approches indépendantes, *footprinting* et microscopie, confirment et valident notre hypothèse de polymérisation de SvtR sur au moins une de ses cibles dans le génome de SIRV1. Ce mécanisme explique comment à partir d'un site situé à distance du promoteur régulé SvtR influence l'activité des certains promoteurs sous son contrôle.

f. Conclusions

Ce travail a mis en évidence l'existence de plusieurs nouvelles cibles de la protéine régulatrice SvtR dans le génome du virus d'archées SIRV1. Onze sites de liaison de SvtR sont prédits *in silico* et l'analyse fonctionnelle a permis d'établir le rôle exact de SvtR pour quatre de ces sites (Figure 64). SvtR semble réguler les gènes codant pour des protéines de structures (*gp19* et *gp30*), pour le régulateur transcriptionnel SvtR (*gp08*) et pour une protéine hypothétique (*gp03/gp04*). La validation d'autres sites présumés de fixation de SvtR prédits *in silico* nécessite des recherches

supplémentaires. Il n'est pas exclu que plusieurs protéines peuvent être impliquées simultanément dans la régulation des promoteurs concernés.

Vu le nombre de cibles putatives de SvtR présentes dans le génome de SIRV1, il paraît évident qu'il s'agit d'un des régulateurs principal du cycle de ce virus.

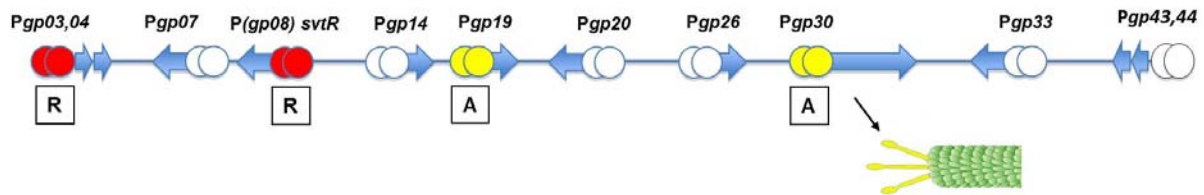


Figure 64 : Régulon de SvtR dans le génome de SIRV1. Le régulateur transcriptionnel SvtR active la transcription des gènes *gp19* et *gp30* et réprime la transcription des gènes *gp03/gp04* et *gp08*, son propre gène. Les autres cibles sont prédites *in silico* mais leur régulation par SvtR n'a pas été démontrée *in vitro*. Au total, SvtR pourrait régulerait 22% des gènes du génome de SIRV1.

Le mécanisme de régulation par SvtR est original. Pour certaines de ses cibles (*gp08* et *gp19*), SvtR devrait agir par une simple fixation de ses deux dimères sur le site principal dans les cas où le site de fixation se trouve proche du promoteur du gène régulé. La présence de ce complexe protéique pourrait directement influencer les étapes sensibles de l'initiation de transcription (fixation des facteurs TBP et/ou TFB; formation du complexe d'initiation; démarrage de la transcription par l'ARNP).

Pour d'autres cibles (*gp03/gp04* et *gp30*) l'hypothèse d'une interaction directe est inapplicable compte tenu de la distance entre le site et la cible de régulation. Sur l'exemple du promoteur du gène *gp30* nous avons démontré une oligomérisation de SvtR qui débute à partir du site principal et permet d'atteindre la *TATA-box*. Les premières indications montrent que cette oligomérisation pourrait être guidée par des sites secondaires situés entre le site principal et la *TATA-box*.

Le phénomène d'oligomérisation a été confirmé de façon indirecte (*footprinting*) et visualisé directement à l'aide de l'observation par l'AFM-EM. Il semble qu'une interaction entre deux sites principaux de SvtR est envisageable ce qui prédit la possibilité de formation de structures « en huit » rapprochant différentes parties du génome de SIRV1. Cependant, seule l'interaction intermoléculaire est à présent mise en évidence et l'interaction intramoléculaire reste à confirmer. Si ce type d'interaction est confirmé, il s'agit du premier exemple de régulation chez les archées impliquant le changement global de conformation de tout un génome viral.

Les génomes des virus d'archées sont remarquablement riches en protéines RHH. Une analyse *in silico* récente de dix génomes de crénarchées révèle la présence d'au moins dix-huit régulateurs putatifs RHH dans les génomes analysés (Prangishvili *et al.*, 2006). Leur rôle dans la régulation du cycle viral reste largement inconnu. Le mécanisme de régulation découvert sur l'exemple de SvtR et impliquant l'oligomérisation de la protéine régulatrice et éventuellement le changement global de conformation de l'ADN cible pourrait être un nouveau mécanisme inconnu assez général pour beaucoup de protéines de la famille RHH.

5. P06, premier régulateur transcriptionnel putatif au doigt de zinc chez les archées

1. Contexte de l'étude

Au début de ce travail, seuls les régulateurs transcriptionnels proches par leurs structures des régulateurs connus chez les bactéries (essentiellement des protéines des familles HTH et RHH) ont été identifiés et caractérisés chez les archées. Cette situation est surprenante étant donné que les archées possèdent une ARN polymérase de treize sous-unités, très proche structurellement et fonctionnellement de celle des eucaryotes (Kusser *et al.*, 2007; Hirata *et al.*, 2008; Korkhin *et al.*, 2009). Les eucaryotes ont le plus souvent recours à des familles de régulateurs spécifiques généralement absentes chez les archées et les bactéries. Par exemple, les protéines formant la structure « doigt de zinc » décrite pour la première fois en 1985 chez *Xenopus oocytes* (Miller *et al.*, 1985) sont très répandues chez les eucaryotes en tant que régulateurs transcriptionnels (Razin *et al.*, 2012). Quelques exceptions proviennent des prédictions *in silico* sans qu'elles soient confirmées *in vivo* (Bouhouche *et al.*, 2000).

L'analyse *in silico* des protéines codées par des virus d'archées publiée en 2006 (Prangishvili *et al.*, 2006) a prédit que la majorité des régulateurs putatifs portant des domaines de liaisons à l'ADN reconnaissables appartiennent aux protéines des familles HTH et RHH. Cependant, des protéines putatives portant un domaine doigt de zinc ont été suggérées. Dans un cas, il s'agit de la protéine P06 codée par le génome du virus AFV1 (Figure 65) qui infecte la crénarchée *Acidianus hospitalis* (Prangishvili *et al.*, 2004; You *et al.*, 2011). Compte tenu de l'expérience de caractérisation des facteurs de transcription chez les archées que possède notre équipe, nous avons décidé d'étudier la structure et les fonctions de cette protéine *in vivo*.

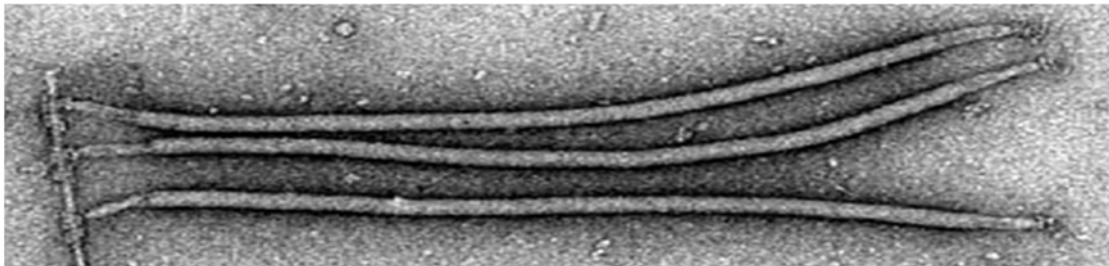


Figure 65 : Virus AFV1 de la famille des *Lipothrixviridae*, sous microscopie électronique à transmission (Prangishvili *et al.*, 2006).

Tout d'abord, la structure de la protéine P06 a été déterminée par l'équipe de M. Delpierre et d'I. Guijaro (Institut Pasteur, unité de RMN des Biomolécules) par l'approche de RMN. Les résultats obtenus par ce groupe ont démontré clairement la présence d'un domaine de type doigt de zinc dans la structure de P06 (Figure 66). Comme attendu, la région avec le motif ZNF (résidus neuf à trente-cinq) montre une structure typique en doigt de zinc avec une structure β antiparallèle en épingle à cheveux emballée contre une hélice α avec l'ion zinc coincé entre les deux.

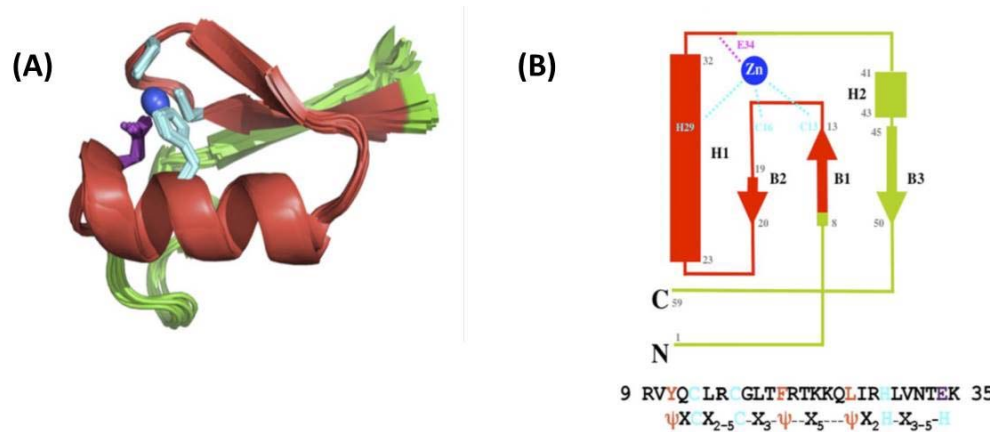


Figure 66 : Structure de la protéine AFV1p06. (A) Topologie de la structure de la protéine AFV1p06. La séquence d'AFV1p06 dans la région ZNF et la séquence motif du ZNF est montrée en (B). Les résidus dans la région ZNF sont colorés en rouge, les chaînes latérales qui coordonnent le zinc sont en cyan ou violet et l'atome de zinc est en bleu. Les résidus hydrophobes sont représentés par un Ψ , les hélices par des rectangles et les feuilletts β par des flèches.

Ma tâche a été de réaliser l'étude de sa fonctionnalité. Deux questions principales ont été soulevées : premièrement, démontrer la capacité de cette protéine à se lier à l'ADN et si cette capacité est spécifique ou non; deuxièmement, rechercher sa ou ses cible(s) putative(s) en choisissant des promoteurs candidats dans le génome d'AFV1.

2. La protéine AFV1p06 est capable de lier l'ADN

La structure et les propriétés de liaison au zinc de la protéine P06, comme le montrent les données obtenues par l'équipe de structuralistes avec laquelle nous avons coopéré indiquent que cette protéine porte la structure typique de doigt de zinc (Figure 66). En tenant compte que la plupart des protéines ZNF étudiées auparavant (Isalan *et al.*, 1997) lient spécifiquement l'ADN, nous avons commencé l'analyse fonctionnelle de P06 par l'étude de sa capacité à se fixer à l'ADN.

Pour vérifier si la protéine P06 était capable de se lier à l'ADN, nous avons testé la mobilité de courts oligonucléotides, double brin et également simple brin, de 24pb et 24nt respectivement, dans des expériences de retard sur gel en présence de concentration croissante de la protéine. Deux types d'oligonucléotides ont été testés, polyA sb (polyA/T en db) et polyG sb (polyG/C en db). Chacune des expériences a été réalisée en présence d'ADN non spécifique.

Les résultats de ces expériences montrent que la protéine AFV1p06 peut se lier à la forme double brin de l'oligonucléotide polyG mais pas à la forme simple brin (Figure 67). De plus, l'oligonucléotide polyG/C était retardé en présence de quantité plus faible de protéine comparé à polyA/T (Figure 68). Cela suggère que la protéine AFV1p06 se fixe préférentiellement sur des régions riches en GC et que sa liaison avec l'ADN peut dépendre des bases azotées.

Différentes compositions de tampons, variant surtout en concentrations de sels ont aussi été testés, ainsi que l'ajout d'ions zinc dans le mélange réactif. Le *binding* est meilleur en présence d'une concentration de 150nM NaCl. L'ajout supplémentaire d'ions de zinc, en plus du zinc retenu très fortement par la protéine P06 avant sa purification ne change aucune de ses propriétés biochimiques testées.

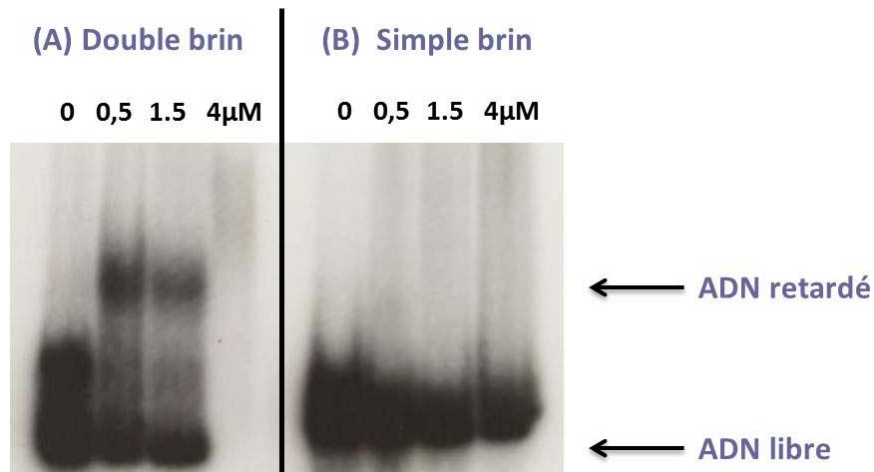


Figure 67 : La protéine AFV1p06 peut se lier à l'ADN double brin et non simple brin. Les expériences PAGE-EMSA ont été réalisées en présence de concentration d'oligonucléotide fixe et de concentration croissante d'AFV1p06 (de 0 à 4 μ M). Le *binding* de la protéine AFV1p06 a été testé en présence de la forme double brin (A) et simple brin (B) du même oligonucléotide polyG/C double brin ou polyG simple brin marqué au γ -P³². Les expériences PAGE-EMSA ont été réalisées en présence d'un excès d'ADN compétiteur non spécifique, poly[dl,dC]. Les quantités sont indiquées μ M.

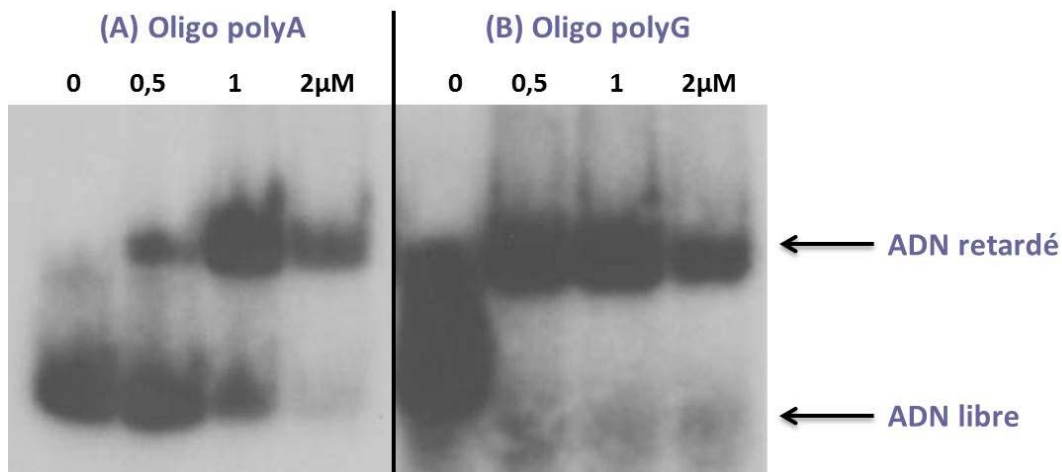


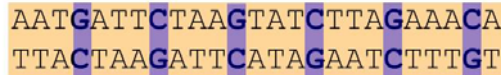
Figure 68 : Comparaison du binding à l'ADN ne contenant aucun nucléotide G (polyA) (A) et les contenant en excès (polyG) (B) de la protéine AFV1p06 par PAGE-EMSA. La concentration de l'oligonucléotide marqué au γ -P³² est fixe (75nM) en présence de concentration croissante d'AFV1p06 (de 0 à 4 μ M). Les expériences PAGE-EMSA ont été réalisées en présence d'un excès d'ADN compétiteur non spécifique, poly[dl,dC]. Les quantités sont indiquées en μ M.

3. La protéine AFV1p06 lie l'ADN de façon spécifique

Pour la suite, deux nouveaux oligonucléotides de 25pb ont été pensés, afin de mieux caractériser l'activité de liaison à l'ADN de la protéine AFV1p06 (Figure 69). La composition de ces oligonucléotides était inspirée par la structure cristalline du complexe du domaine de liaison à l'ADN du facteur de transcription Zif268 (isolé chez la souris) et de son site de liaison (Pavletich *et al.*, 1991; Isalan *et al.*, 1997; 1998). Dans ce complexe, chacun des trois ZNFs de Zif268 établit des contacts spécifiques avec trois bases d'un sous-site sur un brin de l'ADN (Pavletich *et al.*, 1991; Isalan *et al.*, 1997; 1998) (Figure 70). Parce que la protéine AFV1p06 possède un seul domaine ZNF, nous avons

émis l’hypothèse que son hélice α pourrait interagir avec un court site ADN de trois nucléotides. Les oligonucléotides « dsATcomb » et « dsGCcomb » ont donc été dessinés pour porter des répétitions régulièrement espacées de différentes combinaisons de triplets (six sur les huit possibles) composés de A et/ou T pour l’oligonucléotide « dsATcomb » et de G et/ou C pour l’oligonucléotide « dsGCcomb ». Avec cette approche combinatoire, nous avons essayé de créer une ou plusieurs courtes sous-régions ADN qui pourraient être mieux reconnues par AFV1p06 pour tester si la protéine est capable de discriminer entre les différentes séquences ADN.

Oligo dsATcomb (25pb)



Oligo dsGCcomb (25pb)



Figure 69 : Les oligonucléotides « dsATcomb » et « dsGCcomb ». Les oligonucléotides « dsATcomb » et « dsGCcomb » portent différentes combinaisons de triplets composés de A et/ou T et de G et/ou C, respectivement.

Nous avons comparé la mobilité de ces deux oligonucléotides « antagonistes » par leur composition et par leur richesse en GC dans des expériences de gel retard en présence de P06. Une nette baisse du signal correspondant à la forme retardée de l’oligonucléotide « dsATcomb » marqué au γ -P³² est observée en présence de l’oligonucléotide « dsGCcomb » « froid ».

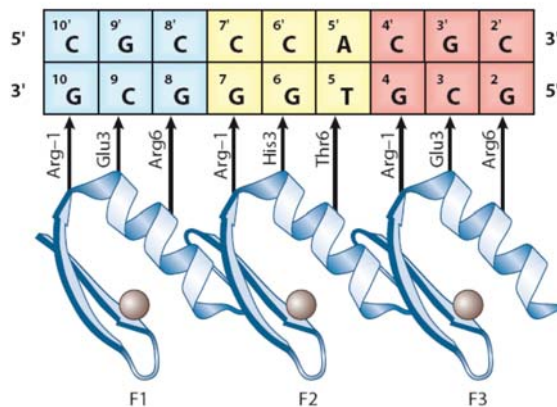


Figure 70 : Schéma du premier modèle de reconnaissance de l’ADN par un peptide à trois doigts de zinc. D’après les résultats de la détermination de la première structure cristalline du complexe entre le domaine de liaison à l’ADN du facteur de transcription Zif268 et son site de liaison. Chaque doigt de zinc interagit avec un sous-site de 3pb sur un brin de l’ADN, utilisant les résidus aminoacides des positions -1, 3 et 6. (Klug 2010).

Pour observer la même efficacité, cinq fois plus d'oligonucléotide « dsATcomb » froid a dû être ajouté à l'oligonucléotide « dsGCcomb » radiomarqué (la figure correspondante est présentée dans le pdf de l'article joint à ce manuscrit). Ce résultat suggère fortement que la protéine AFV1p06 a une préférence pour des motifs d'ADN GC riches (par rapport aux motifs AT riches) et de ce fait elle devrait être capable de reconnaître et de se lier à certaines séquences d'ADN de façon spécifique. En tenant compte de ce que nous connaissons sur l'interaction des protéines à doigt de zinc chez les eucaryotes, il peut être supposé que la protéine P06 a une préférence pour les triplets GC riches comme site(s) de liaison.

4. Rôle biologique de la protéine AFV1p06

La capacité de P06 à se lier de façon spécifique à l'ADN double brin suggère fortement son implication dans la régulation transcriptionnelle de certains gènes dans le génome d'AFV1. Cependant, rien n'est connu à ce sujet et nous avons tenté de mettre en évidence les promoteurs cibles de P06 dans le génome de ce virus.

Cela n'a pas été une tâche évidente car, d'après l'analyse structurale de P06, la spécificité de liaison à l'ADN de la protéine P06 seule ne peut pas, par défaut, être très forte. P06 ne possède qu'un seul doigt de zinc et pour cette raison ne peut reconnaître qu'un site ADN très limité, de l'ordre de trois nucléotides. Il est probable qu'*in vivo* la spécificité de P06 soit renforcée par l'interaction de cette protéine avec d'autres protéines participantes à la régulation du promoteur sous contrôle de P06 et dotées de la capacité d'interagir avec l'ADN. Cependant, à ce stade d'étude aucune protéine « *helper* » d'AFV1 si elle existe, n'est connue.

Le site reconnu par P06 seul est donc par défaut trop petit pour espérer l'identifier par les analyses conventionnelles (retard sur gel par exemple). Malgré ce constat, nous avons quand même tenté de tester *in vitro* la cible la plus probable, par comparaison avec d'autres régulateurs viraux décrits dans la littérature, à savoir le propre promoteur du gène codant pour P06. Il est connu que très fréquemment les régulateurs transcriptionnels, surtout chez les virus, contrôlent l'expression de leurs propres gènes (Guillière *et al.*, 2009; Peixeiro *et al.*, 2013). En partant de cette hypothèse, nous avons décidé de tester le *binding* de la protéine P06 sur son propre promoteur à l'aide d'expériences de retard sur gel. Dans le génome d'AFV1, l'ORF59a qui code pour AFV1p06, fait partie du même opéron que les ORFs 99, 72 et 157 (Figure 71A). Nous avons testé le *binding* d'AFV1p06 au promoteur de cet opéron (il sera appelé opéron 1).

Une autre piste pour la recherche de la cible de P06 est suggérée par la préférence de cette protéine aux régions d'ADN GC riches. Le génome du virus AFV1 a une composition générale de 36% en GC et nous avons choisi comme cibles potentielles de P06 les régions promotrices où le pourcentage en GC était plus élevé que dans le reste du génome d'AFV1 (>36%). Par exemple, c'est le cas du promoteur qui contrôle l'expression des ORFs 52 et 80 (appelé ici opéron 2) (Figure 71A).

Aucun retard spécifique n'a été mis en évidence témoignant que la protéine AFV1p06 ne se lie pas dans les conditions employées à ces deux régions du génome. (Figure 71B). La ou les cible(s) de P06 dans le génome d'AFV1 restent à découvrir.

Les résultats obtenus confirment la conclusion de l'analyse structurale de P06 qui la caractérise comme une protéine à doigt de zinc. Chez les archées, c'est la première protéine de cette famille pour laquelle le *binding* spécifique à l'ADN riche en GC est démontré.

La découverte et l'analyse de la première protéine à doigt de zinc chez les archées ébranlent la vision communément acceptée sur la nature typiquement « bactérienne » de la régulation de transcription chez les archées. D'après nos premiers résultats, il est temps de réviser ce dogme et de chercher d'autres exemples impliquant des protéines au doigt de zinc dans la régulation dans le troisième domaine du vivant.

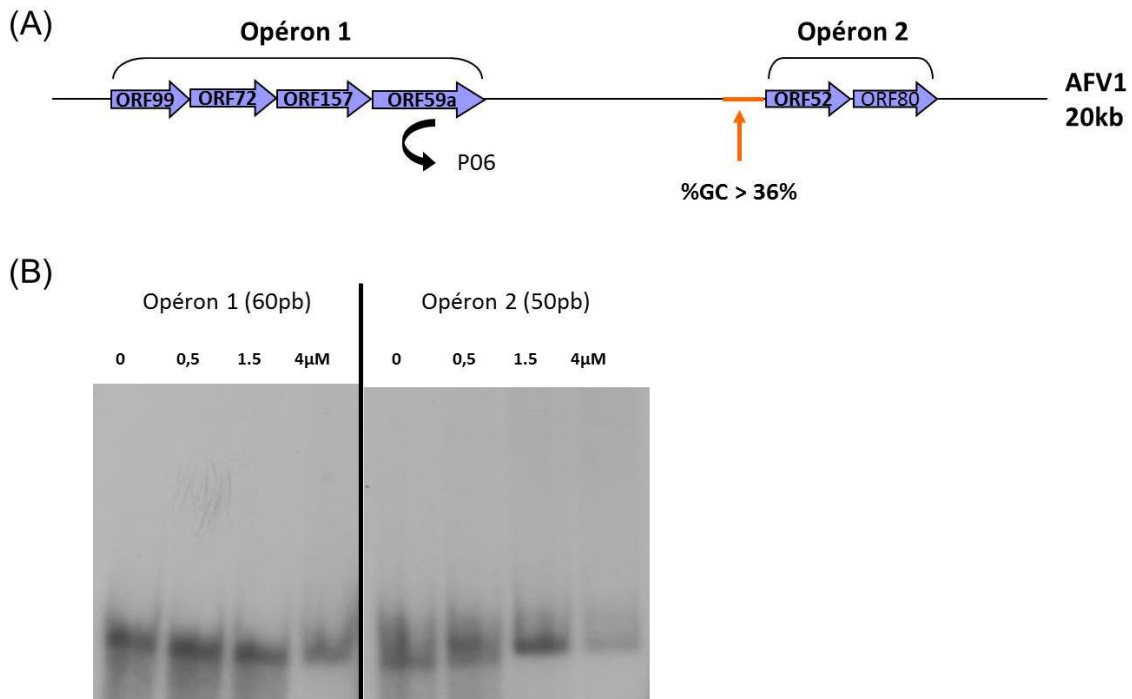


Figure 71 : Recherche de cibles potentielles de la protéine AFV1p06 dans le génome du virus AFV1. (A) Organisation de deux opérons testés dans le génome du virus AFV1. L'opéron 1 porte les ORF99, ORF72, ORF157 et ORF59a qui code pour la protéine AFV1p06 et l'opéron 2, les ORF52 et ORF80 et possède une région promotrice plus riche en GC que le reste du génome d'AFV1. (B) Aucun *binding* de la protéine AFV1p06 aux oligonucléotides marqué au γ - P^{32} contenant les régions promotrices de deux opérons de AFV1. Les expériences PAGE-EMSA ont été réalisées en présence d'un excès d'ADN compétiteur non spécifique, poly[dI,dC]. Les quantités sont indiquées en μ M.

Solution Structure of an Archaeal DNA Binding Protein with an Eukaryotic Zinc Finger Fold

Florence Guillière^{1,2,3,9}, Chloé Danioux^{4,5,9}, Carole Jaubert^{4,5}, Nicole Desnoues⁴, Muriel Delepierre¹, David Prangishvili⁴, Guennadi Sezonov^{4,5*}, J. Iñaki Guijarro^{1,2*}

1 Institut Pasteur, Unité de RMN des Biomolécules, Département de Biologie Structurale et Chimie, Paris, France, **2** CNRS UMR 3528, Paris, France, **3** Université Paris 7 Denis Diderot, Paris, France, **4** Institut Pasteur, Unité de Biologie Moléculaire du Gène chez les Extrémophiles, Département de Microbiologie, Paris, France, **5** Université Pierre et Marie Curie, Paris, France

Abstract

While the basal transcription machinery in archaea is eukaryal-like, transcription factors in archaea and their viruses are usually related to bacterial transcription factors. Nevertheless, some of these organisms show predicted classical zinc fingers motifs of the C2H2 type, which are almost exclusively found in proteins of eukaryotes and most often associated with transcription regulators. In this work, we focused on the protein AFV1p06 from the hyperthermophilic archaeal virus AFV1. The sequence of the protein consists of the classical eukaryotic C2H2 motif with the fourth histidine coordinating zinc missing, as well as of N- and C-terminal extensions. We showed that the protein AFV1p06 binds zinc and solved its solution structure by NMR. AFV1p06 displays a zinc finger fold with a novel structure extension and disordered N- and C-termini. Structure calculations show that a glutamic acid residue that coordinates zinc replaces the fourth histidine of the C2H2 motif. Electromobility gel shift assays indicate that the protein binds to DNA with different affinities depending on the DNA sequence. AFV1p06 is the first experimentally characterised archaeal zinc finger protein with a DNA binding activity. The AFV1p06 protein family has homologues in diverse viruses of hyperthermophilic archaea. A phylogenetic analysis points out a common origin of archaeal and eukaryotic C2H2 zinc fingers.

Citation: Guillière F, Danioux C, Jaubert C, Desnoues N, Delepierre M, et al. (2013) Solution Structure of an Archaeal DNA Binding Protein with an Eukaryotic Zinc Finger Fold. PLoS ONE 8(1): e52908. doi:10.1371/journal.pone.0052908

Editor: Ramón Campos-Olivas, Spanish National Cancer Center, Spain

Received: September 14, 2012; **Accepted:** November 23, 2012; **Published:** January 9, 2013

Copyright: © 2013 Guillière et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work was supported by the Institut Pasteur and the Centre National de la Recherche Scientifique (CNRS UMR 3528). FG, CD and CJ were supported by the French «Ministère de l'Enseignement Supérieur et de la Recherche». The 600 MHz spectrometer was funded by the Région Ile de France and the Institut Pasteur. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: inaki.guijarro@pasteur.fr (JIG); guennadi.sezonov@pasteur.fr (GS)

⁹ These authors contributed equally to this work.

[□] Current address: Université de Lyon, Université Claude Bernard Lyon 1, Institut des Sciences Analytiques, CNRS UMR 5280, Villeurbanne, France

Introduction

It is now well established that transcription in archaea, one of the three domains of life, displays characteristics of both eukaryal and bacterial transcription [1,2]. The minimal basal machinery in archaea consists of an RNA polymerase and the general transcription factors TBP (TATA-box-binding protein) and TFB (transcription factor B), required for transcription initiation. These proteins are homologues of the eukaryal RNA polymerase II (RNAPII), TBP and TFIIB proteins, respectively. In particular, the eukaryal and archaeal RNA polymerases show a striking structural similarity [3,4]. The archaeal basal machinery is thus homologous structurally and functionally to the core components of the eukaryal RNAPII machinery. In contrast, non-general transcription factors (TF) in archaea are often bacterial-like, and only a few are predicted to be of eukaryal type [1,2]. For instance, a recent *in silico* analysis based on 52 archaeal genomes suggested that over 50% of the predicted transcription factors show at least one homologue in bacteria, about 43% are specific to archaea and less than 2% have homologues in eukaryotic organisms [5]. Though some transcription factors in archaea have been analysed

in detail [6–8], transcription regulation in archaea is still poorly documented.

The presence in archaea of proteins with predicted zinc finger domains of the C2H2 or C2HC type is intriguing as the so-called “classical” zinc finger, hereafter named ZNF, is considered to be an eukaryal-specific motif. Initially discovered in the transcription factor TFIIIA from *Xenopus* oocytes [9], the ZNF domain has been shown to be very abundant in eukaryotes (*e.g.* 3% of human genes encode ZNF-containing proteins), practically absent in bacteria with some exceptions as in plant pathogens [10] and scarce, but represented in archaea and their viruses. The classical ZNF motif consists of a short (~30 residue-long) sequence that uses two or three cysteines and two or one histidines to coordinate a zinc ion (C2H2 or C2HC types, respectively). The ZNF domains fold into a characteristic structure consisting of an α -helix and a β -hairpin held together by the zinc ion and hydrophobic interactions between hydrophobic residues at conserved positions of the sequence. Most of the proteins containing ZNF domains that have been characterised are involved in transcription regulation and bind DNA through their ZNF domains, although ZNF domains can also mediate protein-RNA or protein-protein interactions [11,12]. ZNFs bind to DNA by inserting the α -helix

into the major groove and use three or four exposed residues of the helix to make specific contacts with three or four DNA bases [13,14]. To recognise their target DNA in a cellular context, ZNFs are usually present in tandem repeats separated by a short linker. Each ZNF repeat binds specifically to DNA using the α -helix and the repeats wrap around the DNA. Some ZNFs like SW15, ADR1 or GAGA, however, are present in only one to three copies and use extensions of the ZNF motif to further contact DNA [15–17].

Hyperthermophilic archaea that thrive in hot springs ($>80^{\circ}\text{C}$) are infected by viruses that show unique morphological and genomic properties that distinguish them from bacteriophages and eukaryal viruses [18]. The majority of the proteins of these viruses does not have detectable homologues in the databases, however, a relatively high proportion is predicted to carry transcription-factor associated folds (up to $\sim 10\%$ of proteins encoded in genomes with about 50 putative genes) [19]. The abundance of putative TFs in the genomes of hyperthermophilic archaeal viruses probably reflects the importance of transcription regulation in the life cycle of the viruses. As in the case of their hosts, the majority of the predicted TFs are bacterial-like and display a ribbon-helix-helix (RHH) or a helix-turn-helix (HTH) fold. One viral predicted TF, the SvtR protein from virus SIRV1, has been characterised and shown, indeed, to display a RHH structure and to repress transcription of viral genes [20]. Structural analysis of another viral protein (E73) coded by the SSV-like virus SSV-RH, also revealed the presence of a RHH motif involved in DNA recognition [21]. In addition to bacterial-like TFs, archaeal viruses from the *Rudiviridae*, *Lipothrixviridae*, *Fuselloviridae* and the *Bicaudaviridae* families as well as the unclassified viruses STSV1 and STIV typically present one or two sequences with ZNF motifs.

In this work, we focused on the protein AFV1p06 coded by the gene *gp06* of the virus AFV1 (*Acidianus* filamentous virus 1 [22]), which infects the hyperthermophilic crenarchaeon *Acidianus hospitalis*. The protein has 59 residues and displays a single ZNF motif with the second zinc-binding histidine of the motif missing. The ZNF motif (28 residues) is flanked by N- and C-terminal regions of unknown structure. AFV1p06 has homologues in crenarchaeal spindle-shaped viruses from the *Fuselloviridae* family (SSV1, SSV2, SSV4, SSV5, SSV6 and SSVK-1), and is distantly related to eukaryal ZNF containing proteins [19]. Here, we describe the solution structure of AFV1p06 and analyse its DNA binding capabilities.

Materials and Methods

Cloning, Protein Expression and Purification

The gene *AFV1p06* of AFV1 (NC_005830.1, also called ORF59a) was amplified by PCR using primers AFV1p06NdeI (5'-ATGCCATATGATTGAGGTTTCTAGTATGG-3') and AFV1p06XhoI (5'-ATTTCTCGAGTCAGATAATCTTGTTTACAT-3'). The PCR product was digested with *NdeI* and *XhoI* and ligated with *NdeI* and *XhoI* digested pET-30a (Novagen) plasmid vector.

Recombinant AFV1p06 was expressed without any tag or cloning-derived additional residues using *Escherichia coli* RosettaTM (BL21 DE3) pLysS (Novagen) cells. Cultures at 37°C in rich (Luria-Bertani broth) or in minimal M9 media for ^{15}N or $^{15}\text{N}/^{13}\text{C}$ labelling, induction with 1 mM isopropyl- β -thio-galactopyranoside, cell harvesting after four hours of induction and cell freezing at -80°C were performed as described [20].

AFV1p06 was purified from inclusion bodies. Frozen cells were thawed, suspended in 50 mM HEPES pH 7.4, 150 mM NaCl, 3 mM dithiothreitol (DTT, buffer A) and lysed with a French press at 4°C adding phenyl-methane-sulphonyl fluoride. The cell lysate

was centrifuged 20 min at 7000 g and 4°C and the supernatant was discarded. The cell pellet was suspended in buffer A supplemented with DNase and RNase to eliminate nucleic acids and centrifuged at 7000 g for 20 min at 4°C . The cell pellet was then suspended in buffer A containing 1% Triton to eliminate hydrophobic compounds, centrifuged and washed twice with buffer A by means of suspension and centrifugation cycles. The washed pellet was solubilised in buffer A containing 6 M urea (buffer B) and loaded into a size exclusion chromatography column (Sephacryl HR100, GE Healthcare) pre-equilibrated with buffer B. The sample was eluted with buffer B and the AFV1p06 containing fractions were pooled and dialysed at low concentration ($\sim 25\ \mu\text{M}$) and temperature (4°C) against buffer A containing 500 mM arginine and 50 μM ZnCl_2 (buffer C) to renature the unfolded protein. After renaturation, arginine was eliminated by dialysis against buffer C prepared without arginine (buffer D) and loaded on an ion-exchange column (SP Sepharose, GE Healthcare) previously equilibrated with the loading buffer. Proteins were eluted using a linear gradient of NaCl from 150 mM to 1 M in buffer D. The AFV1p06 containing fractions, which eluted at ca. 650 mM NaCl, were pooled, dialysed against the desired buffer (typically 50 mM HEPES pH 7.4, 100 or 150 mM NaCl, 50 μM ZnCl_2 , 3 mM DTT) and concentrated by centrifugation using Vivaspin (Sartorius) tubes with a 3 kDa cut-off. Protein preparations were aliquoted and kept at -80°C or used directly for NMR experiments.

Protein preparations were homogeneous as assessed by SDS-PAGE and NMR; protein integrity and identity were checked by SELDI-TOF mass spectrometry (Jacques d'Alayer, Microsequencing Facilities, Institut Pasteur). The concentration of the protein was determined using a molar extinction coefficient of $5960\ \text{M}^{-1}\cdot\text{cm}^{-1}$ calculated from its sequence [23].

Flame Atomic Emission Spectrophotometry

Experiments were carried out at the Ecole Polytechnique (Palaiseau, France) on a Varian AA220 spectrophotometer equipped with an air-acetylene burner. Readings were performed at 213.9 nm in the peak height mode. Two samples in buffer A were analysed: one was prepared as described above and extensively dialysed to eliminate free zinc ions from the sample; the second one was obtained without adding ZnCl_2 during renaturation or the following purification steps and adding a forty fold excess of NaEDTA relative to the protein during the refolding step.

Oligonucleotides

Oligonucleotides were purchased from Proligo (Sigma-Aldrich). Double-stranded DNA was obtained by annealing the corresponding single strand oligonucleotides following standard techniques. For PAGE experiments, oligonucleotides were ^{32}P radiolabelled using the T4-polynucleotide kinase (Fermentas).

DNA Binding

Two 25-bp duplex DNA oligonucleotides, called dsATcomb (top strand sequence 5'-AATGATTCTAAGTATCTTAGAAACA-3') and dsGCcomb (top strand sequence 5'-AGGGTGGCAGCGTCGGAGCCTCGCA-3') were obtained by annealing the corresponding single strand complementary oligonucleotides. Prior to annealing, one strand of each oligonucleotide was ^{32}P -radiolabelled. Each double-stranded labelled oligonucleotide (75 nM) was incubated with increasing amounts of AFV1p06 (from 0 to 2 μM) for 15 min at 48°C in 20 μl of binding buffer: 50 mM HEPES, 10 μM ZnCl_2 , 150 mM NaCl, 5% (v/v) glycerol, 0.02% Tween, 3 mM DTT, pH 7.4. The binding buffer

was supplemented with 50 ng/μL of unspecific salmon sperm DNA. The DNA-protein mixtures were deposited in a non-denaturing 6% 37.5:1 acrylamide/bisacrylamide gel. PAGE was run in TBE buffer (89 mM Tris-borate, 2 mM NaEDTA, pH 8.3). After migration, the gel was vacuum-dried, exposed with Amersham Biosciences Hyperfilm™ MP and developed with a Kodak X-OMAT 2000 processor.

Binding of dsATcomb and dsGCcomb to AFV1p06 was also tested by competition experiments in which labelled dsATcomb or dsGCcomb at a fixed concentration (75 nM) were used as probes in electromobility gel shift assays (EMSA) and unlabelled dsGCcomb or dsATcomb at varying concentrations (0, 0.5, 1, 2 and 5-fold molar ratio of unlabelled/labelled oligonucleotide), were used as competitors. Oligonucleotides and AFV1p06 (0.5 μM) were incubated 15 min at 48°C in 20 μL of binding buffer. PAGE, gel drying and development were performed as described above.

NMR Samples

Samples were prepared in buffer E: 50 mM HEPES pH 7.4, 100 mM NaCl, 44 μM ZnCl₂, 4.5 mM DTT, 12% D₂O. Protein concentration typically ranged between 0.4 and 1.0 mM for ¹⁵N labelled and ¹³C/¹⁵N labelled samples.

NMR

Experiments were performed on a Varian NMR System 600 spectrometer (Agilent Technologies, Santa Clara) with a proton resonating frequency of 599.4 MHz. The spectrometer was equipped with a cryogenic probe. Spectra were recorded at 25°C and referenced to sodium 4,4-dimethyl-4-silapentane-1-sulphonate following IUPAC recommendations. Data were collected using Vnmj 2.3A (Agilent Technologies), processed with NMRPipe [24] and analysed with NMRView 5.2.2 [25].

Standard two- and three-dimensional experiments were recorded to assign chemical shifts to the protein ¹H, ¹³C and ¹⁵N nuclei: ¹³C or ¹⁵N HSQC (heteronuclear single quantum coherence) [26]), HNCOC, HNCACB, CBCA(CO)NH [27], H(CC-TOCSY)NNH, C(CC-TOCSY)NNH [28,29], (HB)CB(CGCD)HD and (HB)CB(CGCDCE)HE [30].

AFV1p06 backbone dynamics analysis was based on ¹⁵N relaxation experiments [31] used to calculate the longitudinal (R₁) and transverse ¹⁵N (R₂) relaxation rates.

NMR and structure calculations—Distance constraints for structure calculations were obtained from 3D ¹³C-edited (aromatic and aliphatic regions) and ¹⁵N-edited NOESY-HSQC (nuclear Overhauser effect spectroscopy - HSQC) experiments recorded with 120 ms mixing times [26,32]. Proton J_{HN-HA} scalar couplings were calculated from a HNHA experiment [33,34] and transformed into dihedral φ angle constraints as follows: -120°±25° for ³J_{HN-Hα} ≥ 8.0 Hz, -65°±25° for ³J_{HN-Hα} ≤ 5.5 Hz. Further dihedral φ and ψ constraints were obtained with Talos [35]. A backbone hydrogen bond in regions of secondary structure was added as distance constraint if the chemical shift data, the nOe pattern and the amide hydrogen exchange data were in agreement with a hydrogen bond, and if it was present in at least 75% of the structures calculated without any hydrogen bond. Hydrogen exchange was analysed using the HET-SOFAST experiment [36]: two spectra with (saturation) or without (reference) inversion of the water signal were acquired to evaluate the protection against exchange from the saturation transfer between water and amide protons. The residues with a ratio of intensities higher than 0.75 between the saturation and reference experiments were considered to be exchange protected.

NOESY spectra assignments and structure calculations were performed with ARIA 2.2 [37,38] coupled to CNS 1.2 [39] following ARIA's standard protocols with spin diffusion correction.

Spin diffusion was corrected using an isotropic rotation correlation time of 6.3 ns (±0.6 ns), which was determined from ¹⁵N relaxation data as described in [40]. Chemical shift tolerances were set to 0.03 and 0.04 ppm for protons in the direct and indirect dimensions, respectively, 0.5 ppm for ¹³C and 0.35 ppm for ¹⁵N. For structure calculations and nOe (nuclear Overhauser enhancement) assignments, the zinc atom was coordinated with a tetrahedral geometry by the Sγ atoms of cysteines 13 and 16 and the Nε2 atom of histidine 29 (see Results section). Histidine was protonated. The zinc ion was attached to the Sγ atom of residue 16, and the tetrahedral geometry was maintained by modifying the force field topology and parameter files. Once the nOes were assigned and the distance constraints were obtained, two different final structure ensembles were calculated using either the full-length protein (residues 1–59) or only the structured region (residues 7–51). The final structures were obtained by calculating 200 structures with ARIA 2.2/CNS 1.2 and refining the lower energy 150 structures in explicit water using the PARALLHDG 5.3 force field [41]. The 10 lowest-total-energy structures were selected. The quality of the structures was analysed with Procheck 3.5.4 [42], What_check [43], Molmol 2K.2 [44] and Pymol (Schrödinger LLC).

Phylogenetic Analysis

To gather the amino acid sequences for phylogenetic analysis, we searched the non-redundant protein sequence database (nr) at NCBI for homologues of AFV1 virus p06-ORF59a (GI: 82056192) using the PSI-BLAST algorithm 2.2.26+ [45] in ten iterative steps with default parameters. Whenever the algorithm ran out of new proteins to include in the iteration, the protein with the best E-value and with a conserved C2H2 motif was manually picked. Sequences were aligned using the CLC Sequence Viewer software (CLC Bio, Denmark) with default parameters. The tree was calculated using the Neighbour-Joining method and a 100 replicate bootstrap analysis.

The protein knowledgebase UniProtKB database was questioned using the query “zinc AND finger AND C2H2” to obtain the sequences of proteins with predicted ZNF motifs in the three domains of life.

Accession Codes

The structure and chemical shifts of AFV1p06 have been deposited in the PDB protein data bank (<http://www.pdb.org>) and the BMRB database (<http://www.bmrwisc.edu>) under the accession numbers 2LVH and 18570, respectively.

Results

Zinc Chelation by AFV1p06

As the homology of AFV1p06 to C2H2 and C2HC zinc fingers suggested that the protein could bind zinc, we performed flame photometry experiments on samples that had been carefully depleted of free zinc. These experiments confirmed that AFV1p06 binds zinc and showed that one mole of protein binds one mole of zinc. In addition, sedimentation-diffusion equilibrium ultracentrifugation experiments performed at a 50 μM concentration (Bertrand Raynal, Plate-forme de Biophysique, Institut Pasteur), and NMR ¹H-¹⁵N HSQC spectra, which were invariant for AFV1p06 concentrations between 50 μM and 1.0 mM, indicated

that the protein is monomeric up to millimolar concentrations. Thus, one AFV1p06 monomer binds one zinc ion.

In ZNF proteins, zinc is tetra-coordinated by two Cys and two His residues (C2H2) or three Cys and one His residue (C2HC). AFV1p06 contains two Cys residues (C13 and C16) and a single His residue (H29) that are part of the ZNF motif and that could be involved in zinc coordination. Nevertheless, the protein lacks the fourth zinc ligand, which could be either a water molecule or the side chain of residue E34 that in the sequence alignments with ZNF proteins is positioned close to the fourth zinc ligand (H or C). We performed a NMR chemical shift analysis to verify if residues C13, C16 and H29 could be involved in metal chelation. On the one hand, the C β and C α chemical shifts of residues C13 and C16 were in agreement with those of metalloproteases [46], indicating that both cysteine S γ atoms bind zinc; on the other hand, the comparison of the aromatic ring carbon chemical shifts (C δ 2 and C ϵ 1) of H29 with that of histidine residues (deposited in the BMRB database) that bind or do not bind zinc, indicated that H29 binds zinc and that it ligates zinc through its N ϵ 2 atom. This analysis was corroborated by a recently published method to determine the coordination of zinc by His residues based on the difference of the aromatic C δ 2 and C ϵ 1 chemical shifts [47]: in the case of AFV1p06, this difference is 12.96 ppm, which corresponds well to the value observed for N ϵ 2 coordination 12.32 \pm 0.86 ppm. Based on this experimental data, we calculated AFV1p06 structures considering that zinc was coordinated by residues C13 (S γ), C16(S γ) and H29 (N ϵ 2) and we used the structures to determine the fourth ligand of zinc. Importantly, no bias that could influence the determination of the fourth ligand was introduced in the calculations because the nOe assignments for distance constraints were performed automatically.

Resonance Assignments of AFV1p06

The ^1H , ^{15}N and ^{13}C resonance frequencies of most backbone and side chain atoms were assigned (92%). Missing assignments mainly corresponded to exchangeable protons of lysine, arginine, asparagine and glutamine side chains, as well as to the backbone amide protons of residues S6, M7 and K23 (the assigned ^{15}N - ^1H HSQC spectrum of AFV1p06 is shown in Figure S1).

Structure of AFV1p06

The structure ensemble of AFV1p06 shows a compact and convergent region between residues 8–50 and disordered N (1–7) and C (51–59) termini (Figure 1, Table 1). The structure consists of a three-stranded antiparallel β -sheet (residues 8–13, 19–20 and 45–50) packed against an α -helix (23–32), as well as of a short 3_{10} helix (41–43) located at the end of a long loop between the α -helix and the third strand of the β -sheet. As expected, the region with the ZNF sequence motif (residues 9–35) shows a typical zinc finger fold with an antiparallel β -hairpin packed against an α -helix and with the zinc ion sandwiched between the latter structural elements. Indeed, a search for structural homologues in the DALI database (<http://www.dali.server.org>) with the structure of AFV1p06 between residues 9–35 produces over 150 ZNF structures with statistically significant scores and low root mean square deviations (RMSD \leq 1.8 Å over \sim 25 CA atoms). When the coordinates of the structured region between residues 8 and 50 were used to find structural homologues, only the ZNF region gave significant hits, indicating that AFV1p06 shows a novel extension of the ZNF fold (loop with a 3_{10} helix+3rd strand of ZNF β -sheet).

The lack of convergence observed for the N- and C-termini of the protein correlates with a very low number of nOes shown by residues 1–7 and 51–59 and more specifically, with the absence of medium or long range nOes. This disorder is due to the dynamics

of the protein as assessed by the ^{15}N relaxation characteristics of the backbone amide groups. For instance, most of the N and C-termini amide groups showed low ^{15}N transverse relaxation rates (R_2) values relative to those observed for the rest of the protein, indicating high amplitude motions in the nanosecond-picosecond time scale (Figure S2). Also, residue S5 showed a very high R_2 rate, and amide resonances of residues S6 and M7 were not observed, presumably due to exchange broadening (high R_2 rates), suggesting that the latter residues exchange between different conformations in the microsecond-millisecond time scale. Thus, the N- and C-termini of AFV1p06 are highly dynamic.

The fourth ligand of the zinc ion was identified using the structure ensemble of AFV1p06: in all the structures, a side-chain oxygen atom of the carboxylic group of residue E34 is close to the zinc ion at a distance (1.99 \pm 0.04 Å) that is in agreement with those observed for zinc coordinated by a glutamic acid residue [1.95 \pm 0.08 Å, [48]]. This observation indicates that residue E34 is the fourth residue implicated in zinc coordination. Although a glutamic acid residue is not commonly observed as a zinc ligand, it coordinates zinc in some proteins in which the latter ion plays a structural role [48]. In AFV1p06, the zinc ion is tightly bound. Indeed, the protein retains zinc in the presence of a 10 fold excess of NaEDTA, as evidenced by the lack of changes in the NMR ^1H - ^{15}N HSQC spectrum of the protein in the presence of the latter chelating agent.

AFV1p06 Binds Preferentially to GC Rich DNA

The structure and zinc binding properties of AFV1p06 indicate that this protein is a classical zinc finger. As most of the ZNF proteins that have been characterised have been shown to bind double stranded DNA [14], we tested the DNA binding capabilities of AFV1p06. Because the putative binding site for AFV1p06 was not known, we initially performed EMSA experiments in the presence of unspecific DNA and high concentrations of the protein. We chose two DNA oligonucleotides that were extremely different in their nucleotide composition: the oligonucleotides, either single (ssDNA) or double stranded (dsDNA), were 24 nt long and exclusively composed of the succession of AT (polyAT) or CG (polyCG) pairs. The EMSA experiments indicated that AFV1p06 could bind dsDNA at micromolar concentrations and did not bind the corresponding single strand DNAs, and this independently of their DNA composition. Interestingly, the dsDNA polyCG oligonucleotide was clearly better recognised by AFV1p06 than the polyAT one (not shown). At high salt concentration (500 mM), AFV1p06 was also able to bind dsDNA and recognised better the polyCG oligonucleotide, suggesting that the interaction of this protein with DNA is not only based on protein – DNA-backbone electrostatic interactions but involves DNA bases.

Following these observations and in order to better characterise the DNA binding activity of AFV1p06, we designed two additional double strand oligonucleotides of 25 bp called “dsATcomb” (5'-AATGATTCTAAGTATCTTAGAAACA-3') and “dsGCcomb” (5'-AGGGTGGCAGCGTCGGAGCCTCGCA-3'). The composition of these oligonucleotides was inspired by the crystal structure of the complex of the DNA-binding domain of the transcription factor Zif268 and its binding site. In the latter complex, each of the three ZNFs of Zif268 establishes specific contacts with 3 bases on one strand of the DNA [49–51]. Because the protein AFV1p06 has a single ZNF domain we hypothesized that its α -helix would interact with a short 3 nt DNA core site. The oligonucleotides “dsATcomb” and “dsGCcomb” were hence designed to carry regularly interspaced repetitions of different combinations of triplets (6 from 8 possible) composed of either A or/and T for the

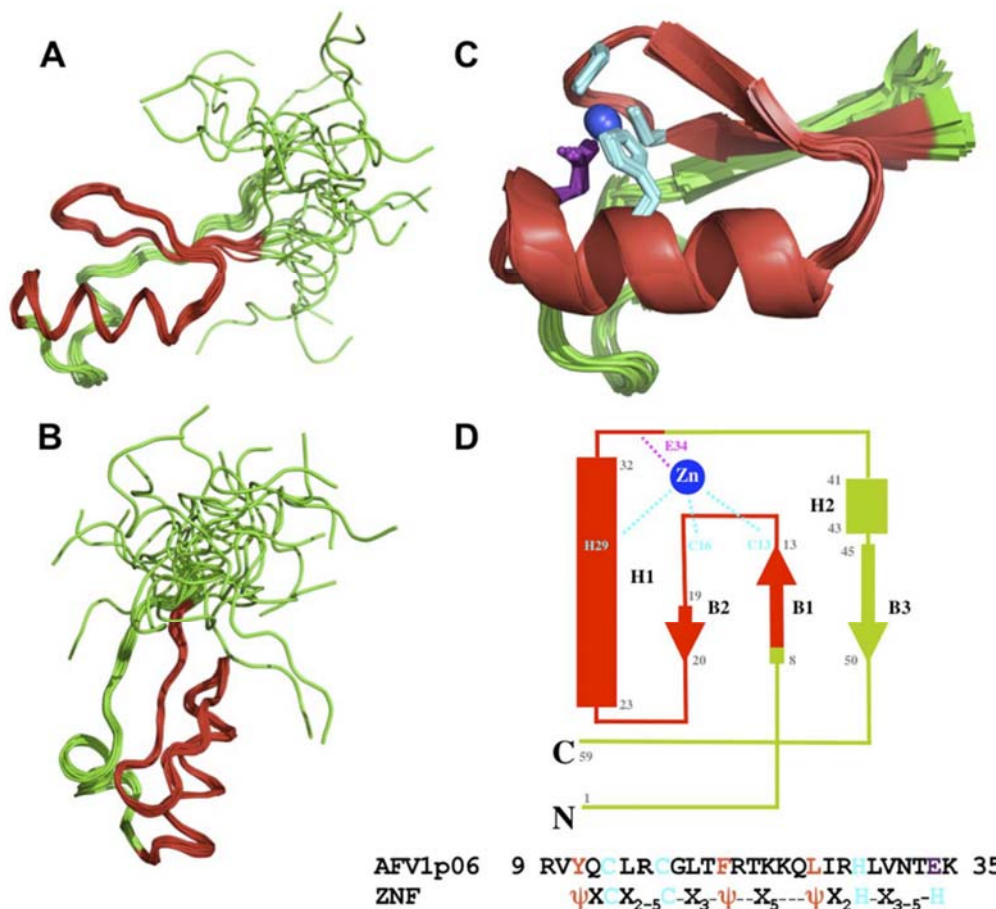


Figure 1. Structure of AFV1p06. The backbone superposition of the 10 structures calculated for the full-length protein is shown in two different orientations (A and B) and on a main-chain cartoon representation for residues 7–51 (C). A topology diagram of the structure, the sequence of AFV1p06 in the ZNF region (residues 9–35) and the ZNF sequence motif are shown in (D). Residues in the ZNF region are coloured in red. In (C), the side-chains of the residues that coordinate zinc are displayed in cyan (C13, C16 and H29) or violet (E34) and the zinc atom in blue. In (D), ψ stands for a hydrophobic residue. Helices are represented by rectangles and β -strands by arrows.
 doi:10.1371/journal.pone.0052908.g001

“dsATcomb” and G or/and C for the dsGCcomb oligonucleotides. With this combinatorial approach we tried to create one or several short DNA sub-regions in the analysed oligonucleotides that could be better recognised by AFV1p06 to test if the protein is able to discriminate between different DNA sequences.

PAGE-EMSA experiments were performed with radioactively labelled dsATcomb and dsGCcomb in the presence of an excess of non-specific unlabelled dsDNA (Figure 2A). Both oligonucleotides show a retard in migration in the presence of AFV1p06, indicating that the protein binds dsDNA on the μ M concentration range. The binding of AFV1p06 to the GC-rich dsDNA oligonucleotide is at least twice more efficient than that observed for the AT-rich one. We also compared the efficiency of retardation of each oligonucleotide in the presence of the second one as a “cold” competitor. Even at a 1:0.5 ratio between 32 P labelled dsATcomb and unlabelled dsGCcomb a clear decrease of the signal corresponding to the shifted form of dsATcomb is observed, indicating that dsGCcomb can efficiently displace dsATcomb (Figure 2B). To observe a similar efficiency, a five-fold excess of “cold” dsATcomb has to be added to “hot” dsGCcomb. These

results suggest that AFV1p06 shows a preference for GC motifs and thus can sense different bases and display some specificity in dsDNA recognition.

In an attempt to identify its presumed DNA target sequence, we followed a target candidate approach testing the binding of the protein to the region of its own promoter, as very often transcription regulators show *cis*-regulation. However, even if the promoter region of the *gp06* gene is unusually GC rich compared to the generally low GC content of the AFV1 genome (36%), under the *in vitro* conditions used, the efficiency of AFV1p06 binding to this region was not significantly different from that of a “non-specific” AT rich DNA from a non related virus (data not shown).

Phylogenetic Studies

The AFV1p06-related proteins identified by the PSI-BLAST approach are divided into two clearly separated clusters of archaeal and eukaryal proteins that show a common origin (Figure S3). The archaeal proteins are grouped into two sub-

Table 1. Statistics for the ensemble of 10 structures calculated for AFV1p06 calculated with residues 7–51.

Constraints (residues 7–51)		Energies (kcal/mol)	
Unambiguous restraints	826	Total	-1682±21
Ambiguous distance restraints	133	Van der Waals	-170±13
Total number of distance restraints ^a	959	Electrostatic	-1867±35
Intra-residue j-i =0	380	Mean of pairwise RMSD (Å) (8–50) ^b	
Sequential j-i =1	202	Backbone atoms N, Ca, C'	0.60±0.13
Medium range 2≤ j-i ≤4	157	Heavy atoms	1.69±0.20
Long range j-i >4	220	Ensemble Ramachandran plot (8–50) ^b	
Backbone dihedral φ angle restraints	40	Residues in most favoured regions	90.8%
Backbone dihedral ψ angle restraints	37	additionally allowed	9.2%
Total backbone dihedral angle restraints	77	Structure Z scores (8–50) ^b	
Total number of hydrogen bonds	19	Second generation packing quality	-0.46±0.44
Residual distance constraint violations		Ramachandran plot appearance	-1.62±0.70
Number ≥0.3 Å	6	Chi1/Chi2 rotamer normality	-2.38±0.96
Number ≥0.1 Å	68	Backbone conformation	-7.29±3.30
RMS deviation from nOes (Å) ^c	0.0194±0.0037	Unsatisfied H-bond donors per molecule ^b	3.9
Residual dihedral angle constraint violations		Unsatisfied H-bond acceptors per molecule ^b	0
Number ≥5.0°	1	Bumps (8–50) ^b	0
RMS deviation from dihedrals (°)	0.512±0.136		

^aDistance constraints used for structure calculations, which excluded fixed intra-residue distances.

^bValues for the structured region (between residues 8 and 50).

^cIncludes nOe and hydrogen bond data.

doi:10.1371/journal.pone.0052908.t001

clusters representing the two major archaeal phyla, *Cren-* and *Euryarchaeota*. No homologue could be identified in the domain of bacteria or in the third phylum of archaea, the *Thaumarchaeota*.

The group affiliated to *Crenarchaeota* includes 8 representatives forming the “AFV1p06 family”. All of them are coded either by crenarchaeal viruses (SVS-K1 [52], SSV1 [53], SSV2 [54], SSV4 and SSV5 [55], SSV6 [56] and AFV1 [22] or by proviruses integrated into the chromosome. In the case of *S. islandicus*

M.14.25 the AFV1p06 homologue (M1425–1829) is annotated as being coded by a chromosomal gene but a more detailed analysis of this region, which shows typical viral *att*-like sites, clearly indicates the viral origin of the locus. The alignment of the predicted ZNF regions of these proteins indicates the presence of the seven highly conserved amino acids of the ZNF motif (two Cys, two His as well as three hydrophobic amino acids indicated by squares in Figure 3) except in the case of AFV1p06 in which the

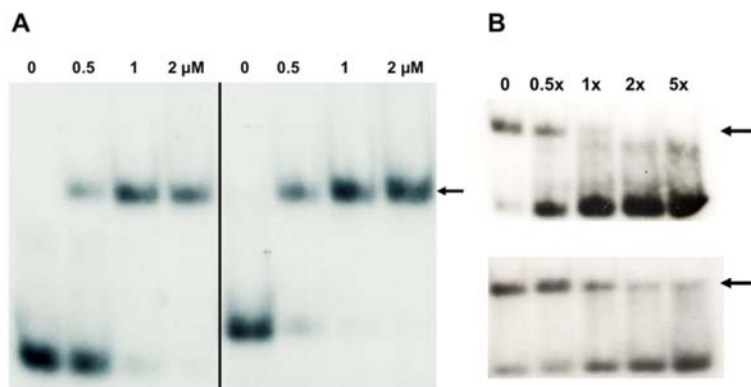


Figure 2. DNA binding of AFV1p06 monitored by PAGE-EMSA. (A) Binding to dsATcomb (left) and dsGCcomb (right) at a fixed concentration (75 nM) with increasing concentrations of AFV1p06 (0 to 2 μM). (B) Competition assays: experiments were performed in the presence of 0.5 μM AFV1p06 using “hot” radiolabelled dsATcomb and increasing amounts of dsGCcomb as a “cold” competitor (top), or radiolabelled dsGCcomb and increasing amounts dsATcomb as a “cold” competitor (bottom). The ratios between “hot” and “cold” oligonucleotides are indicated. Arrows show the position of the shifted DNA band.
doi:10.1371/journal.pone.0052908.g002

last His residue is replaced by a Glu residue. Thirteen additional amino acids (Figure 3) are well conserved in crenarchaeal C2H2-like proteins and six of them (indicated by asterisks) are localised in the loop, helix and third β -strand situated downstream to the ZNF fold. The latter six residues are exposed to the solvent in the structure of AFV1p06, suggesting that these residues are conserved because of their functional importance rather than their role in structure maintenance. The conservation pattern of the proteins in the AFV1p06 family and the nature of the amino residues, strongly suggest that the crenarchaeal C2H2-like proteins show the same structure extension of the ZNF fold as AFV1p06.

A distant group of putative ZNF proteins found in the *Euryarchaeota* (20 representatives) is very similar to the crenarchaeal viral AFV1p06 family in the ZNF motif region but does not show any conservation in the ZNF downstream extension. The origin, cellular or viral, of the genes belonging to this sister of the AFV1p06 group of euryarchaeal ZNF proteins is unclear.

Noteworthy, although AFV1p06 is the only protein in the alignment shown in Figure 3 that displays a Glu residue at the position of the second histidine of the C2H2 motif, it should be mentioned that in eukaryotic ZNFs, the 4th ligand in the motif is also not conserved in a number of variant ZNFs. Conservation of a histidine seems thus less important at the fourth position, an observation that could be explained by the fact the 4th ligand is not crucial to retain zinc binding capabilities as shown in a mutation/folding and stability analysis [57] or by the fact that it can be replaced by a water molecule [58].

Discussion

The results described in this paper indicate that the archaeal virus protein AFV1p06 has a classical ZNF structure composed of an α -helix and a β -hairpin, a novel extension to this fold and disordered N and C terminal ends. In addition, the EMSA

experiments show that the protein can bind to DNA at sub-micromolar concentrations and discriminate between different DNA sequences. Although the presumed biological target(s) of AFV1p06 on the AFV1 virus and/or its host (*Acidianus sp.*) genome remains unknown, these results suggest that AFV1p06 could potentially be a transcription regulator.

Classical zinc fingers usually bind to DNA using exposed residues at positions -1, +2, +3 and +6 of the α -helix that make specific contacts with DNA bases and establish other non-specific contacts with DNA as well. The electrostatic potential of AFV1p06 on the α -helix face is positive (Figure 4) and seems thus favourable for interacting with the negatively charged DNA poly-anion. Moreover, residues that occupy the DNA-contacting positions in AFV1p06 [T (-1), K (+2), Q (+3) and L (+6)] have been observed in ZNF-DNA complexes and could in principle establish specific contacts with DNA bases [13]. Manual docking of AFV1p06 into ZNF-DNA complex structures [PDB codes 2JPA and 2GLI, [59,60]] suggests that AFV1p06 may also interact with DNA using the α -helix: the superimposition of the structure of AFV1p06 with that of ZNFs in complex with DNA indicates that residues at key positions of the helix could indeed make contacts with DNA and that the basic residues R8, R21 (-2) and K23 (+1), would be close to the DNA phosphate groups.

To recognise its cognate DNA in a cellular context, more than three or four specific DNA nucleotide bases/ α -helix residue protein contacts must be established. To this end, eukaryal TFs usually show tandem repeats of ZNF motifs, or like in the case of the GAGA protein, make use of another module that also binds specifically to DNA [16]. The manual docking of AFV1p06 shows that the novel extension of the ZNF motif (the loop and third strand of the β -sheet) is far from the ZNF-DNA contact region (not shown), suggesting that this extension cannot directly contribute to the interaction without major conformational changes. Hence, two possibilities can be envisioned for specific

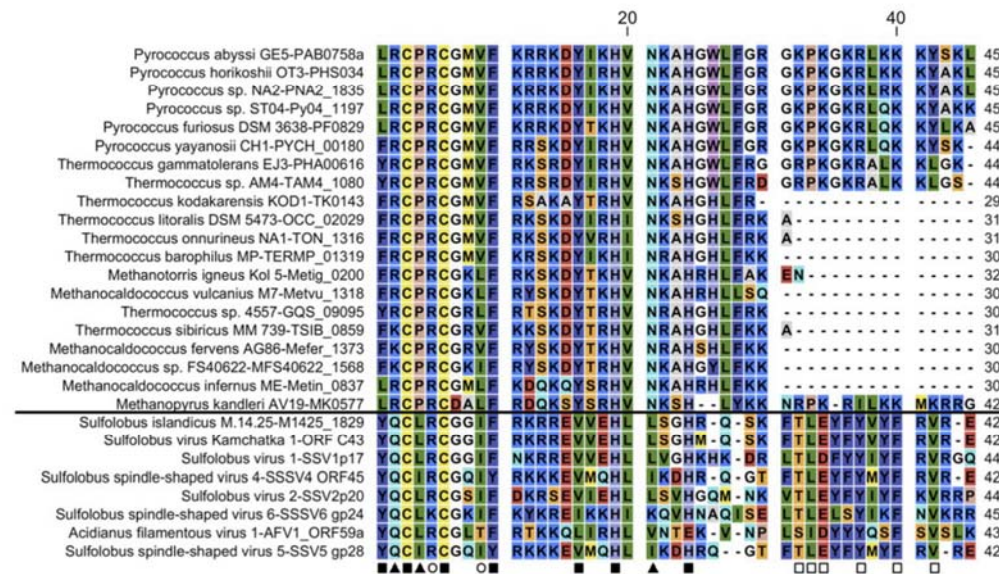


Figure 3. The AFV1p06 family of ZNF proteins in archaea. The figure shows the alignment of 27 hits corresponding to archaeal zinc finger proteins bearing an AFV1p06-like motif. Squares: position of the seven idiosyncratic residues of the ZNF fold; open circles: amino acids conserved in archaea but not in eukaryotes; triangles: amino acids specific to cren- or euryarchaea; open squares: amino acids conserved only in crenarchaea in the ZNF fold extension observed in AFV1p06 (loop+helix+3rd strand of the β -sheet). The horizontal line separates the archaeal viral and cellular proteins. doi:10.1371/journal.pone.0052908.g003

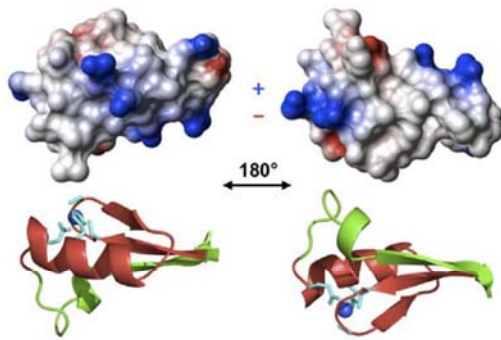


Figure 4. Surface electrostatic potential (top) and main-chain cartoon representations (bottom) of the structure of AFV1p06. A representative conformer was used. Positive charges are represented in blue and negative ones in red. The left and right views are rotated by 180° on the x-axis. The side chains of the residues that coordinate zinc are shown in cyan and the zinc ion as a blue sphere.
doi:10.1371/journal.pone.0052908.g004

binding of AFV1p06 to DNA: (i) the N- and/or C- disordered termini could participate in the interaction; (ii) another DNA binding protein could bind to AFV1p06 on the β -sheet face. Indeed, the hydrophobic character of the exposed β -sheet face and the conservation of its hydrophobicity in AFV1p06 homologues in SSV crenarchaeal viruses, make this face of the protein a good candidate for protein-protein interactions. In this respect, it is interesting to note that Pérez-Rueda and Janga have observed that although bacterial-like, predicted TFs in archaea are statistically smaller (shorter sequence) than in bacteria and specific ligand-binding modules are under-represented [5]. These authors have suggested that protein-protein interactions in archaeal TFs could mediate regulatory feedback. Similarly, it can be hypothesised that archaeal ZNF proteins could also need a protein partner for specific DNA recognition. In this sense, predicted archaeal and archaeal virus ZNFs appear in only one (~74%) or two copies (~20%) per protein in relatively short proteins (most often less than 100 residues). This situation is very rare in eukaryotes, in which very often ZNFs are present in tandem repeats. Although we cannot exclude that the ZNF fold in archaea may be preferentially used for protein-protein interactions or RNA-protein interactions, the fact that AFV1p06, which only contains one ZNF motif, does interact with non-specific DNA with relatively high affinities *in vitro* suggests to us that at least some of these proteins may be TFs and may use either other modules within the same protein or may interact with other proteins to control gene expression. Despite its small size, in the case of the N-terminal

References

1. Aravind L, Koonin E (1999) DNA-binding proteins and evolution of transcription regulation in the archaea. *Nucleic Acids Res* 27: 4658–4670.
2. Bell SD, Jackson SP (2001) Mechanism and regulation of transcription in archaea. *Curr Opin Microbiol* 4: 208–213.
3. Armache KJ, Mitterweger S, Meinhardt A, Cramer P (2005) Structures of complete RNA polymerase II and its subcomplex, Rpb4/7. *J Biol Chem* 280: 7131–7134.
4. Hirata A, Klein BJ, Murakami KS (2008) The X-ray crystal structure of RNA polymerase from Archaea. *Nature* 451: 851–854.
5. Perez-Rueda E, Janga SC (2010) Identification and genomic analysis of transcription factors in archaeal genomes exemplifies their functional architecture and evolutionary origin. *Mol Biol Evol* 27: 1449–1459.
6. Bell SD (2005) Archaeal transcriptional regulation—variation on a bacterial theme? *Trends Microbiol* 13: 262–265.

GATA-1 ZNF and its FOG ZNF partner, it has been observed that the ZNF fold can cope with simultaneous specific protein–DNA and protein–protein interactions or that two different ZNFs can bind to form heterodimers that bind DNA specifically [61–63]. Also, fungal GATA proteins involved in gene regulation display only one ZNF motif, bind to specific DNA sequences and can mediate protein–protein interactions that are important to regulate gene expression [64].

It should be underlined that proteins bearing the C2H2 zinc finger motif are essentially known and characterised in eukaryotes. In this domain of life, ZNFs are predicted to be coded by more than 1% of the genes compared to 0.07% for the archaea (278 examples in the Uniprot database) and only 0.003% (489 examples) for the bacteria. The phylogenetic analysis described here, clearly indicated a common origin of the AFV1p06-like ZNF domain for archaea and eukaryotes, and its absence in bacteria. In crenarchaea all the known genes have a virus related origin.

AFV1p06 is the first archaeal protein with an eukaryal ZNF fold to be characterised experimentally and the first for which the DNA binding and sequence preference capabilities have been demonstrated. It would be interesting in the future to identify its presumed targets on the AFV1 and/or *Acidianus* genomes and understand its role in the virus infection cycle.

Supporting Information

Figure S1 Assigned ^1H - ^{15}N HSQC spectrum of AFV1p06. (PDF)

Figure S2 Structural (number of nOe-derived distance constraints and structure backbone RMSD) and dynamics (amide ^{15}N transverse relaxation rate) data of AFV1p06 on a per residue basis. (PDF)

Figure S3 Phylogenetic tree of the distribution of the AFV1p06-like ZNF fold in Eukarya and Archaea. (PDF)

Acknowledgments

We thank Bertrand Raynal (PFBMI, Institut Pasteur) for analytical centrifugation experiments and Michel Fromant for access to and help with the flame-photometer.

Author Contributions

Conceived and designed the experiments: JIG GS DP MD. Performed the experiments: FG CD JIG CJ ND. Analyzed the data: FG CD JIG GS CJ. Contributed reagents/materials/analysis tools: ND. Wrote the paper: JIG GS FG CD CJ DP MD.

7. Kessler A, Sezonov G, Gujjarro JI, Desnoves N, Rose T, et al. (2006) A novel archaeal regulatory protein, Sta1, activates transcription from viral promoters. *Nuc Ac Res* 34: 4837–4845.
8. Abella M, Rodriguez S, Paytubi S, Campoy S, White MF, et al. (2007) The *Sulfolobus solfataricus* radA paralogue sso0777 is DNA damage inducible and positively regulated by the Sta1 protein. *Nucleic Acids Res* 35: 6788–6797.
9. Miller J, McLachlan AD, Klug A (1985) Repetitive zinc-binding domains in the protein transcription factor IIIA from *Xenopus* oocytes. *Embo J* 4: 1609–1614.
10. Bouhouche N, Syvanen M, Kado CI (2000) The origin of prokaryotic C2H2 zinc finger regulators. *Trends Microbiol* 8: 77–81.
11. Brown RS (2005) Zinc finger proteins: getting a grip on RNA. *Curr Opin Struct Biol* 15: 94–98.
12. Gamsjaeger R, Liew CK, Loughlin FE, Crossley M, Mackay JP (2007) Sticky fingers: zinc-fingers as protein-recognition motifs. *Trends Biochem Sci* 32: 63–70.

13. Wolfe SA, Nekudova L, Pabo CO (2000) DNA recognition by Cys2His2 zinc finger proteins. *Annu Rev Biophys Biomol Struct* 29: 183–212.
14. Klug A (2010) The discovery of zinc fingers and their applications in gene regulation and genome manipulation. *Annu Rev Biochem* 79: 213–231.
15. Duttall RN, Neuhaus D, Rhodes D (1996) The solution structure of the first zinc finger domain of SWI5: a novel structural extension to a common fold. *Structure* 4: 599–611.
16. Omichinski JG, Pedone PV, Felsenfeld G, Gronenborn AM, Clore GM (1997) The solution structure of a specific GAGA factor-DNA complex reveals a modular binding mode. *Nat Struct Biol* 4: 122–132.
17. Bowers PM, Schaeffer LE, Klevit RE (1999) A folding transition and novel zinc finger accessory domain in the transcription factor ADR1. *Nat Struct Biol* 6: 478–485.
18. Prangishvili D, Garrett RA (2004) Exceptionally diverse morphotypes and genomes of crenarchaeal hyperthermophilic viruses. *Biochem Soc Trans* 32: 204–208.
19. Prangishvili D, Garrett R, Koonin E (2006) Evolutionary genomics of archaeal viruses: unique viral genomes in the third domain of life. *Virus Res* 117: 52–67.
20. Guilliere F, Peixeiro N, Kessler A, Raynal B, Desnoves N, et al. (2009) Structure, function, and targets of the transcriptional regulator SvtR from the hyperthermophilic archaeal virus SIRV1. *J Biol Chem* 284: 22222–22237.
21. Schlenker C, Goel A, Tripet BP, Menon S, Willi T, et al. (2012) Structural studies of E73 from a hyperthermophilic archaeal virus identify the “RH3” domain, an elaborated ribbon-helix-helix motif involved in DNA recognition. *Biochemistry* 51: 2899–2910.
22. Bettstetter M, Peng X, Garrett RA, Prangishvili D (2003) AFV1, a novel virus infecting hyperthermophilic archaea of the genus acidianus. *Virology* 315: 68–79.
23. Pace CN, Vajdos F, Fee L, Grimsley G, Gray T (1995) How to measure and predict the molar absorption coefficient of a protein. *Protein Sci* 4: 2411–2423.
24. Delaglio F, Grzesiek S, Vuister GW, Zhu G, Pfeifer J, et al. (1995) NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J Biomol NMR* 6: 277–293.
25. Johnson BA, Blevins RA (1994) NMRView: a computer program for the visualization and analysis of NMR data. *J Biomol NMR* 4: 603–614.
26. Kay LE, Keifer P, Saarinen T (1992) Pure absorption gradient enhanced heteronuclear single quantum correlation spectroscopy with improved sensitivity. *J Am Chem Soc* 114: 10663–10665.
27. Muhandiram DR, Kay LE (1994) Gradient-enhanced triple-resonance three-dimensional NMR experiments with improved sensitivity. *J Magn Reson Series B* 103: 203–216.
28. Grzesiek S, Anglister J, Bax A (1993) Correlation of backbone amide and aliphatic side-chain resonances in ¹³C/¹⁵N-enriched proteins by isotropic mixing of ¹³C magnetization. *J Magn Reson Series B* 101: 114–119.
29. Logan TM, Olejniczak ET, Xu RX, Fesik SW (1993) A general method for assigning NMR spectra of denatured proteins using 3D HC(CO)NH-TOCSY triple resonance experiments. *J Biomol NMR* 3: 225–231.
30. Yamazaki T, Forman-Kay JD, Kay LE (1993) Two-dimensional NMR experiments for correlating ¹³Cβ and ¹Hδ/ε chemical shifts of aromatic residues in ¹³C-labeled proteins via scalar couplings. *J Am Chem Soc* 115: 11054–11055.
31. Farrow NA, Muhandiram R, Singer AU, Pascal SM, Kay CM, et al. (1994) Backbone dynamics of a free and phosphopeptide-complexed Src homology 2 domain studied by ¹⁵N NMR relaxation. *Biochemistry* 33: 5984–6003.
32. Zhang O, Kay LE, Olivier JP, Forman-Kay D (1994) Backbone ¹H and ¹⁵N resonance assignments of the N-terminal SH3 domain of drk in folded and unfolded states using enhanced-sensitivity pulsed field gradient NMR techniques. *J Biomol NMR* 4: 845–858.
33. Vuister GW, Bax A (1993) Quantitative J correlation: a new approach for measuring homonuclear three-bond J H^αH^β coupling constants in ¹⁵N enriched proteins. *J Am Chem Soc* 115: 7772–7777.
34. Grzesiek S, Kuboniwa H, Hinck AP, Bax A (1995) Multiple-quantum line narrowing for measurement of H^α-H^β J couplings in isotopically enriched proteins. *J Am Chem Soc* 117: 5312–5315.
35. Cornilescu G, Delaglio F, Bax A (1999) Protein backbone angle restraints from searching a database for chemical shift and sequence homology. *J Biomol NMR* 13: 289–302.
36. Schanda P, Forge V, Brutscher B (2006) HET-SOFAST NMR for fast detection of structural compactness and heterogeneity along polypeptide chains. *Magn Reson Chem* 44 Spec No: S177–184.
37. Nilges M, Macias MJ, O'Donoghue SI, Oschkinat H (1997) Automated NOESY interpretation with ambiguous distance restraints: the refined NMR solution structure of the pleckstrin homology domain from β-spectrin. *J Mol Biol* 269: 408–422.
38. Rieping W, Habeck M, Bardiaux B, Bernard A, Malliavin TE, et al. (2007) ARIA2: automated NOE assignment and data integration in NMR structure calculation. *Bioinformatics* 23: 381–382.
39. Brünger AT, Adams PD, Clore GM, DeLano WL, Gros P, et al. (1998) *Crystallography & NMR system: a new software suite for macromolecular structure determination*. *Acta Cryst D* 54: 905–921.
40. Mandel AM, Akke M, Palmer AG (1995) Backbone dynamics of *Escherichia coli* ribonuclease HI: correlations with structure and function in an active enzyme. *J Mol Biol* 246: 144–163.
41. Linge JP, Habeck M, Rieping M, Nilges M (2003) Refinement of protein structures in explicit solvent. *Prot Struct Funct & Genet* 50: 496–506.
42. Laskowski RA, MacArthur MW, Moss DS, Thornton JM (1993) PROCHECK: a program to check the stereochemical quality of protein structures. *J Appl Crystallogr* 26: 283–291.
43. Hoof RW, Vriend G, Sander C, Abola EE (1996) Errors in protein structures. *Nature* 381: 272.
44. Koradi R, Billeter M, Wüthrich K (1996) MOLMOL: a program for display and analysis of macromolecular structures. *J Mol Graphics* 14: 51–55.
45. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25: 3389–3402.
46. Sharma D, Rajarathnam K (2000) ¹³C NMR chemical shifts can predict disulfide bond formation. *J Biomol NMR* 18: 165–171.
47. Barraud P, Schubert M, Allain FH (2012) A strong (1)3C chemical shift signature provides the coordination mode of histidines in zinc-binding proteins. *J Biomol NMR* 53: 93–101.
48. Alberts JL, Nadassy K, Wodak SJ (1998) Analysis of zinc binding sites in protein crystal structures. *Protein Sci* 7: 1700–1716.
49. Pavletich NP, Pabo CO (1991) Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 Å. *Science* 252: 809–817.
50. Isalan M, Choo Y, Klug A (1997) Synergy between adjacent zinc fingers in sequence-specific DNA recognition. *Proc Natl Acad Sci U S A* 94: 5617–5621.
51. Isalan M, Klug A, Choo Y (1998) Comprehensive DNA recognition through concerted interactions from adjacent zinc fingers. *Biochemistry* 37: 12026–12033.
52. Wiedenheft B, Stedman K, Roberto F, Willits D, Gleske AK, et al. (2004) Comparative genomic analysis of hyperthermophilic archaeal Fuselloviridae viruses. *J Virol* 78: 1954–1961.
53. Palm P, Schleper C, Grampp B, Yeats S, McWilliam P, et al. (1991) Complete nucleotide sequence of the virus SSV1 of the archaeobacterium Sulfolobus shibatae. *Virology* 185: 242–250.
54. Stedman KM, She Q, Phan H, Arnold HP, Holz I, et al. (2003) Relationships between fuselloviruses infecting the extremely thermophilic archaeon Sulfolobus: SSV1 and SSV2. *Res Microbiol* 154: 295–302.
55. Peng X (2008) Evidence for the horizontal transfer of an integrase gene from a fusellovirus to a pRN-like plasmid within a single strain of Sulfolobus and the implications for plasmid survival. *Microbiology* 154: 383–391.
56. Redder P, Peng X, Brugger K, Shah SA, Roesch F, et al. (2009) Four newly isolated fuselloviruses from extreme geothermal environments reveal unusual morphologies and a possible interviral recombination mechanism. *Environ Microbiol* 11: 2849–2862.
57. Simpson RJ, Cram ED, Czolij R, Matthews JM, Crossley M, et al. (2003) CCHX zinc finger derivatives retain the ability to bind Zn(II) and mediate protein-DNA interactions. *J Biol Chem* 278: 28011–28018.
58. Cordier F, Vinolo E, Veron M, Delepierre M, Agou F (2008) Solution structure of NEMO zinc finger and impact of an anhidrotic ectodermal dysplasia with immunodeficiency-related point mutation. *J Mol Biol* 377: 1419–1432.
59. Pavletich NP, Pabo CO (1993) Crystal structure of a five-finger GLI-DNA complex: new perspectives on zinc fingers. *Science* 261: 1701–1707.
60. Stoll R, Lee BM, Debler EW, Laity JH, Wilson IA, et al. (2007) Structure of the Wilms tumor suppressor protein zinc finger domain bound to DNA. *J Mol Biol* 372: 1227–1245.
61. Mackay JP, Kowalski K, Fox AH, Czolij R, King GF, et al. (1998) Involvement of the N-finger in the self-association of GATA-1. *J Biol Chem* 273: 30560–30567.
62. Newton A, Mackay J, Crossley M (2001) The N-terminal zinc finger of the erythroid transcription factor GATA-1 binds GATC motifs in DNA. *J Biol Chem* 276: 35794–35801.
63. Liew CK, Simpson RJ, Kwan AH, Crofts LA, Loughlin FE, et al. (2005) Zinc fingers as protein recognition motifs: structural basis for the GATA-1/friend of GATA interaction. *Proc Natl Acad Sci U S A* 102: 583–588.
64. Feng B, Marzluf GA (1998) Interaction between major nitrogen regulatory protein NIT2 and pathway-specific regulatory factor NIT4 is required for their synergistic activation of gene expression in *Neurospora crassa*. *Mol Cell Biol* 18: 3983–3990.

Partie III - Discussion et perspectives

Nos connaissances de la biologie des eucaryotes, des bactéries et des archées proviennent d'études physiologiques, génétiques et biochimiques d'un petit groupe d'organismes devenus des organismes modèles.

Le choix de l'organisme modèle relève d'une importance capitale. Il doit être facile à cultiver, avoir un temps de génération court et se prêter à des manipulations expérimentales. A l'époque génomique actuelle, la disponibilité du génome d'une souche modèle est devenue indispensable. Parmi d'autres avantages, les données *in silico* facilitent le développement de l'approche génétique permettant dans certaines conditions de réaliser une mutagenèse dirigée et d'obtenir facilement des mutants KO pour des gènes choisis. Ces conditions incluent la disponibilité de méthodes de transformation et l'existence de marqueurs de sélection permettant l'obtention de mutants KO par recombinaison homologue.

Concernant les archées, le développement des approches génétiques pour les modèles représentant ce domaine (espèces méthanogènes, halophiles et thermophiles) a été lent à décoller, parce que la majorité des archées sont très difficiles à cultiver. La découverte d'espèces cultivables et l'établissement des conditions de culture ont permis d'énormes progrès dans le développement des outils de génétiques pour les deux principaux phyla d'archées, *Euryarchaeota* et *Crenarchaeota*. A présent, plusieurs organismes modèles d'archées sont disponibles pour les méthanogènes, les halophiles et les thermophiles. Les *Sulfolobales*, pour lesquelles la génétique a fait un grand progrès au cours de ces quinze dernières années, servent comme modèles d'étude des crénarchées hyperthermophiles.

Les souches les plus couramment utilisées dans les laboratoires sont *S. solfataricus* (Zillig *et al.*, 1980), *S. acidocaldarius* (Brock *et al.*, 1972), *S. tokodaii* (Suzuki *et al.*, 2002) et *S. islandicus* (Zillig *et al.*, 1998). Récemment, sept génomes de *S. islandicus* ont été séquencés à partir d'une variété de sources chaudes acides aux Etats-Unis, l'Islande et la Russie (Whitaker *et al.*, 2005).

Les séquences complètes des génomes, qui varient de 2 à 3 Mpb et sont très riches en A/T (entre 63 et 67%) ont été déterminées pour toutes les principales souches de *Sulfolobus*. De fréquents remaniements génomiques sont signalés pour *S. solfataricus* (Redder et Garrett, 2006) mais pas pour les souches modèles de *S. islandicus* (Blount *et al.* 2005; Berkner and Lipps 2007; Jaubert *et al.*, 2013).

Un grand nombre d'éléments génétiques tels que les virus, les plasmides et les plasmides conjugatifs ont été isolés à partir de plusieurs espèces de *Sulfolobus* (Zillig *et al.*, 1996; Prangishvili *et al.*, 1998, 2006; Peng *et al.*, 2000; Greve *et al.*, 2004). Ces éléments génétiques sont étudiés de façon intensive : certains d'entre eux sont utilisés pour mieux comprendre l'interaction hôte-virus tandis que d'autres sont utilisés essentiellement pour le développement d'outils génétiques (Berkner and Lipps, 2008).

Différents outils génétiques sont à présent disponibles. Leur utilisation a déjà conduit à des découvertes importantes, y compris concernant le mécanisme de division cellulaire de type eucaryotique chez les crénarchées (Samson *et al.*, 2008), l'identification d'une réponse générale au stress (Maaty *et al.*, 2009), l'identification de plusieurs régulateurs transcriptionnels (Schelert *et al.*, 2006; Peeters *et al.*, 2009), etc.

1. Etablissement du nouveau modèle cellulaire chez les *Sulfolobales* – *Sulfolobus islandicus* LAL14/1.

L'existence de plusieurs souches modèles de *Sulfolobus* établies depuis plusieurs années pose une question stratégique : pourquoi chercher à établir d'autres modèles génétiques pour ce groupe de crénarchées hyperthermophiles? La nécessité de l'établissement de ce nouveau modèle est étroitement liée aux avancées importantes dans les études du virus SIRV2 infectant cet hôte. A la fin du cycle, les virions matures se libèrent de la cellule hôte en utilisant des structures protéiques pyramidales ancrées dans l'enveloppe cellulaire (Bize *et al.* 2008; Quax *et al.* 2010; Quax *et al.* 2011). Ces structures sont codées par le génome viral et l'étude de l'interaction hôte-virus au cours de l'infection est l'un des sujets principaux de notre laboratoire. Jusqu'à 2013, aucun autre modèle d'interaction entre un virus lytique et son hôte cellulaire n'était connu chez les crénarchées et le modèle *S. solfataricus* (Okutan *et al.*, 2013) apparu récemment souffre malheureusement de tous les défauts d'un hôte non naturel et en plus fortement mutagénisé.

La propriété de LAL14/1 d'être la souche naturelle unique pour SIRV2 rend ce modèle cellulaire très intéressant et indispensable pour avancer dans l'étude de l'interaction hôte-virus et un des objectifs de ma thèse a été d'établir l'approche de génétique pour cette souche.

1. Le génome de *S. islandicus* LAL14/1

Le séquençage du génome de la souche LAL14/1 et son annotation a permis de mieux caractériser les particularités de l'organisation de son chromosome et donné des indices sur la stabilité génétique de cette souche. Son génome de 2,47Mb codant pour seulement 2689 gènes est le plus petit génome de *S. islandicus* disponible.

La comparaison des dix génomes de *S. islandicus* a mis en évidence la présence dans chaque génome d'une large région variable occupant jusqu'à 29% de la séquence totale. Chez LAL14/1, cette large région variable portant la séquence peu conservée par rapport aux souches proches, représente 25% du génome. Il est constaté chez LAL14/1 l'absence dans la région variable de loci essentiels (origines de réplication, gènes d'ARNt, gènes d'ARNr, gènes codant pour les protéines ribosomales), accompagnée d'une grande richesse en gènes non-essentiels. Ces résultats concordent avec les observations faites précédemment pour les génomes de *S. islandicus* HVE10/4, REY15A (Guo *et al.* 2011), L.D.8.5, L.S.2.15, M.14.25, M.16.4, M.16.27, Y.G.57.14 et Y.N.15.51 (Reno *et al.* 2009).

Le génome de la souche LAL14/1 est très similaire par sa structure et par sa composition aux génomes des souches phylogénétiquement proches, HVE10/4 et REY15A (Guo *et al.*, 2011). Néanmoins, la présence dans chacun de ces génomes de multiples réarrangements locaux soulève la question de la stabilité du génome LAL14/1, une question cruciale pour un modèle de génétique. Comme les deux souches mentionnées, LAL14/1 porte de nombreuses copies d'*IS* éléments de différentes familles. La transposition des *IS* est une source importante de l'instabilité du génome et des réarrangements chromosomiques (Bennett *et al.*, 2004).

Notre analyse *in silico* du génome de LAL14/1 prédit une faible activité des éléments *IS* chez LAL14/1 car la majorité des familles de ces éléments génétiques mobiles est représentée par des copies incomplètes et non fonctionnelles. En concordance avec cette prédiction, nous n'avons pas pu obtenir de mutants spontanés portant des délétions dans le locus *pyrEF*. Ainsi, la souche LAL14/1

semble être génétiquement plus stable comparée à *S. islandicus* REY15A, un autre modèle génétique chez les *Sulfolobales*, pour qui ce type de mutation est observé assez fréquemment (2% des mutants PyrEF- spontanés) (Guo *et al.*, 2011). De plus, les mutants ponctuels Ura- apparaissent chez LAL14/1 avec une fréquence entre 10^{-6} et 10^{-7} , ce qui correspond aux valeurs classiques pour des mutations spontanées chez les bactéries et les archées.

En conclusion, l'ensemble de résultats obtenus, aussi bien *in silico* qu'*in vivo*, souligne une bonne adéquation de la souche LAL14/1 aux critères imposés pour une souche modèle, essentiellement la faible présence d'éléments mobiles fonctionnels et la stabilité générale du génome.

2. Etablissement du modèle de génétique et les premières retombées.

Après la phase de séquençage et d'analyse *in silico*, l'étape suivante a consisté à valider LAL14/1 comme modèle de génétique par des manipulations *in vivo* et par l'obtention des mutants recherchés.

Le dérivé KO *pyrEF* de LAL14/1 indispensable pour utiliser la sélection Ura+/Ura- a été obtenu par l'approche de remplacement allélique. Nous avons obtenu la preuve que la souche LAL14/1 ne présente pas d'obstacles pour la transformation et que le système de recombinaison homologue est aussi efficace que chez d'autres modèles de *Sulfolobus*.

Un autre marqueur de sélection, le gène *lacS+*, est largement utilisé dans les manipulations génétiques chez les *Sulfolobales*. Nous avons cherché à construire le mutant Lac- pour la souche LAL14/1 *pyrEF-* (ou obtenir un mutant spontané stable) mais malheureusement ce travail n'a pas pu être achevé dans le cadre de cette thèse.

Pour valider le système, nous avons mis comme objectif la construction de trois mutants : le premier correspond à l'inactivation d'un gène unique de 390 bp, *sta1*, le second est le mutant $\Delta topR2$ et le troisième représente la délétion d'une région de 6,6 kbp comprenant de la majeure partie du CRISPR_1 de LAL14/1.

Le gène *sta1* a été mis en évidence par notre laboratoire en 2006 en étudiant les facteurs transcriptionnels de l'hôte intervenant dans la régulation du cycle d'infection des rudivirus SIRV1/2 (Kessler *et al.*, 2006). *Sta1* a été décrite comme un activateur de la transcription de trois promoteurs de SIRV1. Un autre groupe a établi par la suite sa première cible chromosomique, le promoteur du gène *sso0777*, paralogue du gène *radA*, prédit comme impliqué dans la réparation de l'ADN (Abella *et al.*, 2007). Obtenir le mutant KO pour le gène *sta1* avait pour but de vérifier expérimentalement la prédiction selon laquelle un tel mutant a de fortes chances de ne plus être infecté efficacement (ou pas du tout) par le virus SIRV1/2 à cause de l'absence de l'activateur pouvant être crucial pour le cycle du virus. Le même mutant pourrait permettre de vérifier l'implication de *Sta1* dans la régulation du processus de réparation de l'ADN de l'hôte. Le mutant *sta1-* pourrait par exemple, devenir hypersensible aux UV montrant l'importance de la voie de réparation régulée par *Sta1*.

Comme indiqué dans la partie Résultats, l'approche MID utilisée pour générer ce mutant a été appliquée sur deux souches, LAL14/1 et E233S. Etant donné que l'approche est pleinement puissante seulement sur les doubles mutants *lacS- pyrEF-*, toutes les conclusions discutées ci-dessous découlent des résultats obtenus avec la souche E233S.

Pour les deux souches, LAL14/1 et E233S, l'insertion des marqueurs *lacS+* et *pyrEF+* portant par la cassette d'inactivation s'est produit efficacement. Pour rappel, ces transformants *lacS+ pyrEF+* portent toujours une copie fonctionnelle du gène *sta1*. La suite, l'excision de la cassette, a été étudiée avec la souche E233S uniquement. Les données physiologiques (phénotypes des colonies sélectionnées sur le milieu 5-FOA) et les données moléculaires (PCR sur le mélange des transformants cultivés dans un milieu liquide en présence de 5-FOA) indiquent clairement que les mutants recherchés apparaissent au cours de la croissance (données de PCR) mais ne se multiplient pas par la suite, ni dans le milieu liquide, ni sur le milieu solide. Ce comportement a déjà été observé pour certains gènes de *S. islandicus* E233S dans les essais d'obtention de leurs mutants KO et les auteurs de ces études ont conclu qu'il s'agissait de gènes essentiels (par exemple, les gènes *pcna* (Zhang *et al.*, 2010), *mre11*, *rad50*, *xpg*, *herA* (Zhang *et al.*, 2013)).

Le gène *sta1* semble donc être également un gène essentiel impliqué dans la survie de la souche dans les conditions classiques de croissance (80°C, pH3). Pour confirmer cette hypothèse, les expériences de délétion de l'allèle sauvage chromosomique de *sta1* devraient être répétées en présence du second allèle du même gène exprimé à partir d'un promoteur inductible (promoteur du gène *araS* du vecteur pSSR). Seuls les recombinants obtenus et maintenus en présence de l'inducteur (arabinose) devraient survivre et, inversement, leur transfert dans un milieu sans inducteur devrait provoquer un fort ralentissement de la croissance ou la mort cellulaire.

Dans tous les cas, le rôle de Sta1 reste énigmatique, aussi bien pour la survie de la cellule que pour le cycle d'infection de SIRV. Compte tenu de sa très probable essentialité, il n'est pas exclu qu'il régule ou participe à un processus cellulaire vital non décrit jusqu'à maintenant.

Le deuxième gène qui nous a intéressé pour construire un mutant KO est le gène *topR2*. Ce gène semble également être un gène essentiel, à la température de sélection utilisée et son mutant n'a pas pu être obtenu. En perspectives, nous aimerions tenter d'obtenir ce mutant à une température plus basse avant de conclure sur son essentialité.

Le troisième locus pour lequel nous avons souhaité obtenir une délétion correspond à une région chromosomique de 6,6 kbp comprenant de la majeure partie du CRISPR_1 de LAL14/1.

Le génome de LAL14/1 possède cinq loci CRISPRs associés à trois cassettes de gènes *cas* (sous-type I-A) et deux cassettes de gènes *cmr* (sous-type III-B) (Figure 72 (Makarova *et al.* 2011)). Suivant les séquences des régions *leader*, des répétitions et des gènes *cas*, les CRISPRs de LAL14/1 peuvent être divisés en deux familles I et III (Lillestol *et al.* 2009). Les CRISPRs de LAL14/1 rattachés à la famille I présentent une organisation habituelle de deux loci CRISPR_1 et CRISPR_2, liés et orientés en sens contraire, accompagnés de deux cassettes de gènes *cas*, l'une interne et l'autre externe. L'organisation des CRISPRs de LAL14/1 de la famille III est unique chez les *S. islandicus*. Les trois loci CRISPR_3, CRISPR_4 et CRISPR_5 encadrent parfaitement une cassette de gènes *cas* et une cassette de gènes *cmr*.

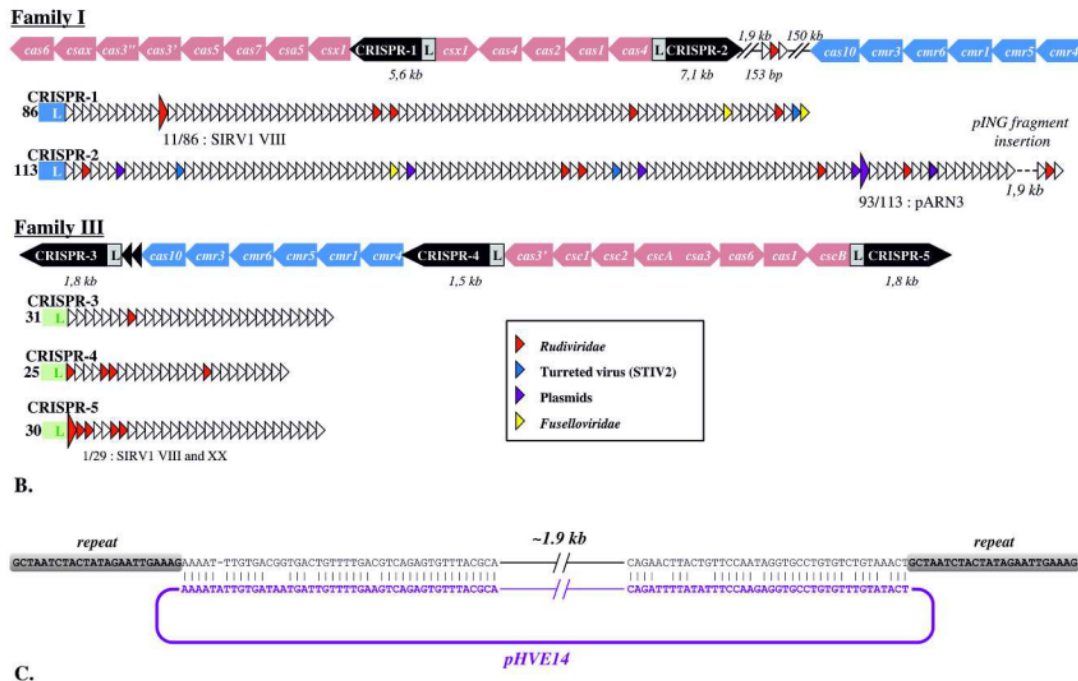


Figure 72 : Organisation des séquences CRISPRs chez *S. islandicus* LAL14/1. Les cinq séquences CRISPRs de LAL14/1 s'organisent en deux régions associées aux familles I et III des séquences CRISPRs de *S. islandicus*. L'insertion étonnante du pseudo-spacer SiL-E1 et son homologie avec le plasmide pHVE14 est aussi illustrée.

L'analyse des 285 *spacers* des CRISPRs de LAL14/1 a révélé trente-trois signatures ciblant parfaitement (trois signatures) ou imparfaitement (trente signatures) de *Rudivirus*, de *Fusellovirus* ou de plasmides conjugatifs déjà décrits dans l'environnement naturel de LAL14/1. Les signatures parfaites se répartissent comme suit : deux signatures ciblant parfaitement le virus SIRV1 ont été mises en évidence, l'une dans le CRISPR_1, et l'autre dans le CRISPR_5. La signature ciblant parfaitement le plasmide pARN3 est retrouvée dans le CRISPR_2. Aucune signature virale parfaite contre le virus SIRV2 n'a été mise en évidence.

Contrairement aux souches préalablement analysées, seule la souche *S. islandicus* LAL14/1 présente une cohérence parfaite entre son phénotype de résistance/sensibilité aux virus SIRV1 et SIRV2 et le profil des signatures virales retrouvées dans ses séquences CRISPRs. Cette cohérence laisse suggérer l'implication des systèmes CRISPR/Cas de LAL14/1 dans la résistance de la souche au virus SIRV1. Cette hypothèse pourrait être confirmée par l'inactivation des deux signatures et l'abolition du phénotype de résistance au virus SIRV1 VIII.

Nous avons tout d'abord réalisé l'inactivation du locus CRISPR_1 portant une des signatures parfaites contre SIRV1. Les détails de ce travail sont discutés dans l'article joint et ici, je me contente de reproduire seulement les constructions utilisées et le schéma des recombinaisons *in vivo* menant vers le mutant Δ CRISPR_1 Δ csx1 Δ cas4 (Résultats II.2.4, Figure 44).

La construction de ce mutant a permis tout d'abord de démontrer que l'approche utilisée est également efficace quand il s'agit de longues délétions de plusieurs kb d'ADN. Le second objectif a été de tenter d'inactiver le CRISPR_1, porteur d'une des signatures contre le virus SIRV1 (séquences des répétitions/*spacers* et les gènes *csx1* et *cas4*) et, comme discuté précédemment, susceptible de jouer un rôle déterminant dans le phénotype de résistance de LAL14/1 contre le virus SIRV1. Le

maintien du phénotype de résistance observé chez le mutant Δ CRISPR_1 Δ csx1 Δ cas4 construit peut être expliqué par la présence d'une seconde signature contre le virus SIRV1 dans le CRISPR_5 ou par l'existence chez LAL14/1 d'autres mécanismes de résistance indépendants du système CRISPR/Cas (Bikard *et al.* 2012).

En perspectives, la construction d'un double-mutant de délétion pour les CRISPR_1 et CRISPR_5 pourrait permettre de mieux caractériser l'implication potentielle du système CRISPR/Cas dans l'immunité de la souche LAL14/1 contre le virus SIRV1. Il serait encore plus intéressant d'obtenir, par plusieurs cycles d'inactivation par la recombinaison spécifique, un dérivé de LAL14/1 dépourvu complètement de toutes les cassettes CRISPRs et des gènes *cas* correspondants. Ce type d'approche réalisée différemment (mutation spontanée sous forme d'une longue délétion) pour *S. solfataricus* P2 a permis de fortement élargir la sensibilité de cette souche pour des virus connus des *Sulfolobales* (Okutan *et al.*, 2013). Ainsi le mutant de *S. solfataricus* « presque sans CRISPRs » peut maintenant servir comme hôte pour l'infection par SIRV2 contre lequel la souche initiale est résistante. Le mécanisme de cet élargissement du spectre d'hôte n'est pas clair surtout dans la situation où aucun *spacer* parfait contre SIRV2 n'est répertorié dans les CRISPRs de la souche sauvage de *S. solfataricus* P2.

Dans le cas de LAL14/1 la construction de ce type de mutants pourrait permettre d'obtenir une souche « hôte universel » capable, si l'hypothèse de départ est correcte, d'être infectée par un plus grand nombre des virus des crénarchées notamment par tous les virus de l'ordre *Ligamenvirales* (Prangishvili et Krupovic, 2012) incluant SIRV1, SIRV2, ARV1 (*Acidianus rod-shaped virus 1*), SRV (*Stygiolobus rod-shaped virus*) et SMRV1 (*Sulfolobales Mexican rudivirus 1*), (Prangishvili *et al.*, 1999; Vestergaard *et al.*, 2005; 2008).

Les CRISPRs de LAL14/1 ouvrent également d'autres perspectives d'études. Ainsi, le génome de LAL14/1 contient un *spacer* inhabituellement long de 2 kb retrouvé dans le CRISPR_2 (Figure 72). Ce *spacer* représente une insertion de près de 2kb d'ADN d'un plasmide partageant 82% de similarité avec les plasmides pHVE14 et pING1. Le mécanisme d'insertion de ce *spacer* et sa fonctionnalité sont des sujets de discussions. Etant donné que ce *spacer* est entouré par deux répétitions parfaites, il n'est pas exclu que son intégration se soit produite grâce à un mécanisme tout à fait classique d'insertion de nouveaux *spacers*. Dans ce cas, cette découverte du premier cas de *spacers* d'aussi grande taille élargira notre connaissance générale des CRISPRs. Quelque soit son mécanisme d'acquisition, il est intéressant de s'interroger sur la fonction de ce pseudo-*spacer*.

Pour répondre à cette question la construction d'un mutant dépourvu de ce *spacer* s'impose. Si ce *spacer* présente une quelconque fonctionnalité, l'efficacité de transformation par le plasmide pHVE14 des souches sauvages et du mutant « Δ *spacer* géant » ne sera pas la même. L'approche génétique ne se limite certainement pas aux seules séquences CRISPRs. Tous les sujets s'adressant à l'hôte seul ou à son interaction avec ses virus peuvent bénéficier de l'approche génétique. Par exemple, les mécanismes d'entrée et de sortie du virus SIRV2 sont des sujets activement développés dans notre laboratoire (Bize *et al.*, 2009; Quax *et al.*, 2010; 2011; Prangishvili *et al.*, 2011). La protéine réceptrice présente sur la surface cellulaire et servant du point d'ancrage du virus au début d'infection est en train d'être identifiée par le groupe de D. Prangishvili (Quemin *et al.*, 2013). Il sera indispensable par la suite d'inactiver le gène correspondant pour démontrer le rôle indispensable de cette protéine dans l'adsorption de SIRV2. Le mutant devrait a priori devenir résistant à SIRV2.

2. Etablissement du modèle viral infectant les *Sulfolobales* – virus SIRV1/2

Les études des virus d'archées hyperthermophiles ont permis l'identification de nombreuses nouvelles espèces virales montrant la richesse insoupçonnée des morphotypes des particules virales (Pina *et al.*, 2011; Prangishvili, 2013). Cependant, notre connaissance du fonctionnement de ces virus reste largement insuffisante. Au cours des dernières années, des efforts considérables ont été faites pour déchiffrer les fonctions des protéines codées par ces virus d'archées et pour caractériser les différentes étapes de leurs cycles d'infection. En conséquence, plusieurs systèmes virus-hôte chez les *Crenarchaota* (Fu et Johnson, 2012; Iverson et Stedman, 2012; Wirth *et al.*, 2011) et chez les *Euryarchaeota* (Klein *et al.*, 2012; Roine *et al.*, 2011; Sencilo *et al.*, 2012; Porter et Dyll-Smith, 2008) sont maintenant disponibles pour étudier plus en détails les interactions hôte-virus d'archées.

Rudiviridae est l'une des neuf familles reconnues de virus de crénarchées (Pina *et al.*, 2011; Prangishvili et Krupovic, 2012; Prangishvili 2013). Ils comprennent des virus non enveloppés linéaires à ADN double brin (Prangishvili *et al.*, 1999). Parmi les cinq isolats de rudivirus, SIRV1 et SIRV2 sont les virus le plus étroitement liés avec 73% d'identité dans leurs séquences génomiques. Dans la nature, ils infectent uniquement certaines souches de *S. islandicus* même si récemment la propagation de SIRV2 dans un mutant de *S. solfataricus* P2 a été décrite (Okutan *et al.*, 2013). La plupart des connaissances actuelles sur le *Rudiviridae* proviennent d'études réalisées avec SIRV1 et SIRV2. Étant donné la forte similitude des protéines codées par chacun de ces virus et une ressemblance générale des séquences et des structures de leurs génomes, les données obtenues pour un de ces virus sont pratiquement toujours directement transférable à l'autre.

Les virus SIRV1/2 s'imposent donc comme un modèle viral très riche pour étudier un large spectre des questions de biologie des virus d'archées. Parmi ces questions – posées, en cours d'étude ou résolues – sont retrouvées l'établissement de la composition et de la structure des virions, étudiées par des approches biochimiques et par microscopie électronique; l'analyse de la structure des génomes des rudivirus et de leurs mécanismes de réplication. Le plus grand progrès a été atteint dans l'étude des étapes du cycle viral et des mécanismes de sortie des virions de la cellule. Le cycle lytique de SIRV2 se termine par l'assemblage des virions dans le cytoplasme, sous la forme de plusieurs paquets contenant environ cinquante virions. Parallèlement à l'assemblage du virion, les structures VAP (appelés associée au pyramides associées aux virus) sont formées à la surface de la cellule provoquant la rupture de la couche S (Bize *et al.*, 2009). Le cycle se termine avec l'ouverture des VAP, permettant la libération du cytoplasme et des virions matures.

Il semble qu'une protéine, SvtR (ou P56B) joue un rôle important dans la régulation transcriptionnelle chez SIRV1. La première étude publiée en 2009 (Guillière *et al.*, 2009) identifie ce régulateur mais seules deux cibles de SvtR avaient été caractérisées en détails, le promoteur de son propre gène et le promoteur du gène codant pour une protéine structurale P1070 (Guillière *et al.*, 2009).

En plus de SvtR, plusieurs autres protéines portant des motifs de liaison à l'ADN (HTH, RHH) ont été prédites par l'approche *in silico* (Prangishvili *et al.*, 2006). Il s'agit notamment de P83a/83b (HTH motif), P59b (RHH motif), P55 (domaine de liaison au zinc) et P114. Bien que la fonction exacte de ces protéines n'ait pas encore été déterminée, la présence de domaines de liaison d'ADN typiques suggère qu'ils pourraient être impliqués dans la régulation des promoteurs viraux et/ou cellulaires.

Identifier tous les régulateurs codés par SIRV1/2, décrire les détails de la régulation transcriptionnelle du génome viral, comprendre les mécanismes d'interaction de ces facteurs avec le complexe d'initiation de la transcription est indispensable pour construire SIRV1/2 en un vrai modèle viral des crénarchées.

Mon travail doctoral a permis, entre autres, d'élargir la connaissance des cibles de SvtR dans le génome de SIRV1. Parmi six cibles identifiées par l'étude *in silico* et EMSA, pour quatre, le rôle de SvtR comme activateur ou répresseur transcriptionnel a été confirmé par l'analyse fonctionnelle. Au total, il a été démontré que SvtR est capable de se lier sur onze sites dans le génome de SIRV1. Le nombre assez élevé de gènes (pour un génome viral de quarante-cinq gènes) inclus dans le régulon de SvtR et le fait que SvtR contrôle deux gènes codant des protéines de capsid (dont la protéine majeure de capsid P134) permet de supposer que SvtR est un régulateur clef du cycle viral.

1. Vision globale de l'activité transcriptionnelle chez SIRV1/2 à l'aide de transcriptome

Malgré la bonne connaissance du rôle et des cibles de SvtR, la régulation transcriptionnelle du cycle viral de SIRV1/2 n'est pas suffisamment comprise dans son ensemble. L'analyse initiale de l'expression des gènes des virus SIRV1 et SIRV2 par hybridation *Northern blot*, de trente minutes à trois heures après l'infection, a donné l'impression d'une quasi absence de régulation temporelle de l'expression des gènes viraux (Kessler *et al.*, 2004). De nombreux gènes sont regroupés et semblent être transcrits en ARNm polycistroniques. Cependant, l'approche de transcriptome beaucoup plus précise et réalisée avec des points de mesures rapprochés donne une image plus nuancée (Quax *et al.*, 2013; Okutan *et al.*, 2013) (Figure 73). Par leur ordre de transcription au cours des trois premières heures après l'infection, les gènes viraux se divisent clairement en trois groupes : précoces, moyens et tardifs. Les gènes précoces occupent les extrémités du génome suivies par les gènes intermédiaires et finalement les gènes tardifs se concentrent dans la partie centrale du génome (Okutan *et al.*, 2013; Quax *et al.*, 2013).

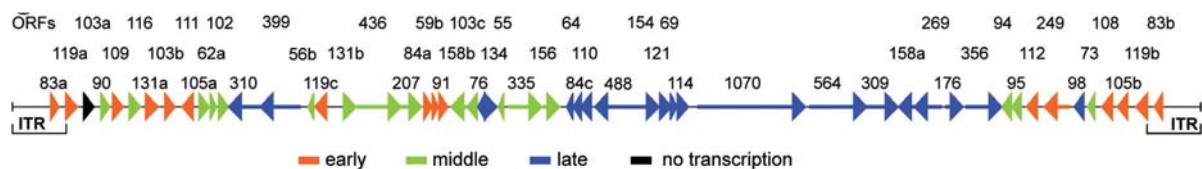


Figure 73 : Carte du génome de SIRV2 montrant l'expression temporelle des gènes viraux. Les gènes sont classés en gènes précoces, intermédiaires et tardifs, selon les résultats d'analyses de puces à ADN (Okutan *et al.*, 2013).

Les transcrits des gènes précoces sont observés à 15-30 minutes post-infection (p.i.) et ils culminent à l'intérieur des 60 minutes p.i. Les transcrits des gènes intermédiaires sont détectés à partir de 30-60 minutes p.i. avec un pic à deux heures p.i. Enfin, les gènes restants culminent tous à trois-quatre heures p.i. et sont classés comme gènes tardifs (Figure 73). Les résultats de Quax *et al.*, 2013 concordent sur l'observation que les gènes situés aux extrémités du génome (ORF83a et b, 119a et 2) sont transcrits en premier et ceux correspondant à la partie centrale du génome (ORF134 et ORF1070 par exemple) ensuite. La position de beaucoup de gènes dans le génome de SIRV est en accord avec les fonctions prédites des protéines correspondantes. Par exemple, les protéines impliquées dans la structure et l'assemblage du virion, telles que les protéines P488, P564, P134

(protéine majeure de capsid) et P1070 (protéine mineure de capsid) et la protéine P98 (formation des VAPs) sont codées par des gènes tardifs (Pangishvili *et al.*, 2013; Okutan *et al.*, 2013) (Figure 73).

Nous nous sommes alors posé la question de comment s'accordaient les données du transcriptome avec nos connaissances du régulon SvtR. Nous avons spécifiquement analysé les profils d'expression des gènes cibles de SvtR pour lesquels le rôle de SvtR, activateur ou répresseur de la transcription, est décrit (*gp03*, *gp08*, *gp19* et *gp30*) mais aussi les gènes *gp07* et *gp20* dont les promoteurs portent des sites fortement reconnus par SvtR. Les résultats de cette analyse sont résumés dans la Figure 74.

Les résultats des puces à ADN (Figure 74A) sont donnés en échelle de Log2 tandis que ceux de l'ARN-seq sont présentés en valeurs absolues (Figure 74B). Les valeurs d'expression déterminées par ces deux transcriptomes sont difficilement comparables et le but de cette analyse a été de comparer de façon qualitative les tendances des profils d'expression pour chaque cible. Pour chacune des cibles observées, les deux transcriptomes donnent des profils d'expression très semblables.

Le gène *gp08* qui code pour la protéine SvtR n'est pas transcrit pendant les premières 15 minutes post-infection. Le gène commence à être transcrit entre 15 et 30 min p.i. jusqu'à atteindre un maximum vers deux heures p.i., puis le nombre de transcrits diminue. Ces données confirment le rôle de SvtR comme répresseur de son propre gène. Le site de fixation de SvtR dans le promoteur de son propre gène est parmi ses sites les plus faibles. Il n'est pas étonnant que l'effet de répression ne soit visible qu'après une forte expression du gène *gp08* et, certainement, une forte accumulation de la protéine SvtR dans le cytoplasme de la cellule infectée. *In vitro* la protéine SvtR est très stable; si elle est aussi stable *in vivo* la répression de *gp08* instaurée vers 2h p.i. va persister pendant le reste du cycle. Une faible augmentation du nombre de transcrits de *gp08* à la fin du cycle, à partir de sept heures, peut être due à la dégradation de la protéine SvtR et la reprise de la transcription de *gp08*, mais l'effet observé est marginal.

SvtR réprime le promoteur du gène *gp03*. Le profil de transcription de *gp03* confirme aussi les résultats. En effet, la transcription du gène augmente rapidement jusqu'à atteindre un pic à une heure p.i. puis redescend progressivement. Cette baisse d'expression est consécutive à l'apparition de transcrits du gène *gp08*, ce qui va fortement dans le sens d'une répression de SvtR sur l'expression du gène *gp03*. Un effet répressif de SvtR sur *gp03* apparaît beaucoup plus précocement par rapport à celui de SvtR sur *gp08* et ceci est en accord avec la meilleure affinité de SvtR au site sur le promoteur de *gp03* que celle pour celui de *gp08*.

En ce qui concerne le gène *gp19*, son expression augmente très fortement à partir d'une heure p.i. lorsque l'expression du gène qui code pour la protéine SvtR est maximale. Cette observation est en accord avec les résultats de mon travail qui montre que SvtR est un activateur transcriptionnel de *gp19*. Cette expression tardive de *gp19* correspond au rôle biologique de la protéine pour laquelle il code : le virus n'a besoin de cette protéine majeure de capsid qu'en fin de cycle d'infection.

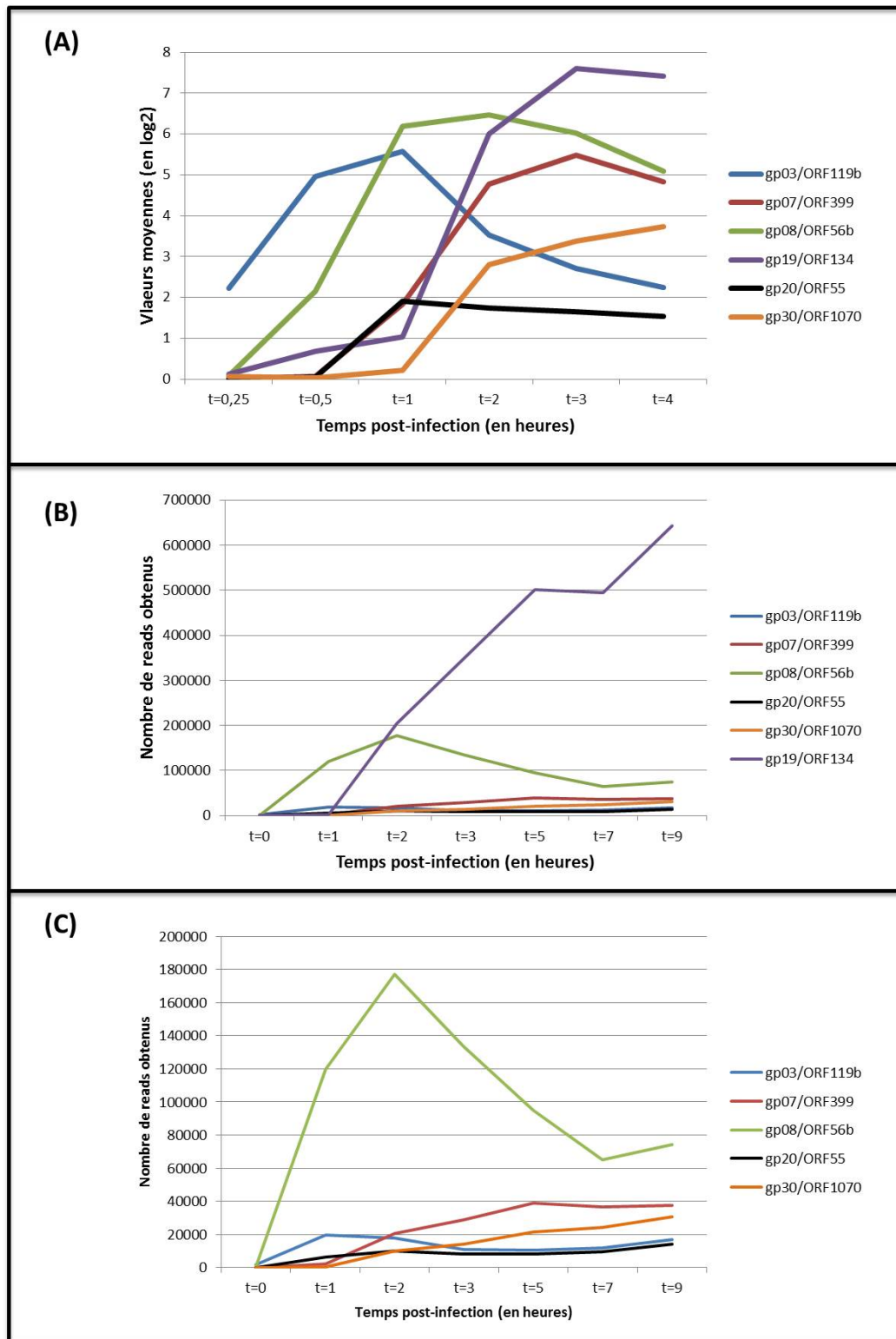


Figure 74 : Profils d'expression du gène *gp08* codant pour le régulateur SvtR et de ses gènes cibles. Données récoltées à partir d'Okutan *et al.*, 2013 et Quax *et al.*, 2013. Sur la partie A, t=0,25 correspond à 15 minutes post-infection et t=0,5 à 30 min. Les valeurs suivantes sont données en heures. Pour les parties B et C, toutes les valeurs sont indiquées en heures post-infection.

Enfin, quant au gène *gp30*, sa transcription est assez faible jusqu'à 60 min p.i. puis elle augmente très fortement pendant le reste du cycle quand la quantité de SvtR accumulée dans la cellule est forte. Ceci est en accord avec les données obtenues au cours de mon travail.

Les transcrits des gènes *gp30* et *gp19* codant respectivement pour les protéines de structure P1070 et P134 sont détectés en quantité importante à partir de quatre-cinq heures. Il peut être conclu que la formation des particules virales peut théoriquement démarrer à partir de cinq heures p.i.

Nous avons démontré que SvtR reconnaît les sites situés dans les promoteurs des gènes *gp07* et *gp20*. Cependant, le rôle fonctionnel de SvtR pour ces cibles est resté non déterminé car l'approche d'IVT n'a pas permis de statuer sur ce point. D'après les données transcriptomiques, l'expression de *gp07* augmente dès le début du cycle, ce qui n'est pas en contradiction avec l'hypothèse d'activation du promoteur *gp07* par SvtR. Quant à *gp20*, sa transcription diminue quand l'expression de SvtR est à son maximum. Ces données suggèrent que SvtR pourrait réprimer la transcription de *gp20*.

Pour conclure, l'ensemble de résultats transcriptomiques présentés confirment les prédictions de l'analyse fonctionnelle concernant le rôle de SvtR sur les cibles analysées.

2. L'originalité du mécanisme de régulation par SvtR

La régulation de l'activité des gènes viraux par SvtR a plusieurs aspects originaux. Tout d'abord, selon la cible SvtR peut agir soit comme répresseur, soit comme activateur transcriptionnel. Ce type de situation a déjà été rencontré pour plusieurs autres régulateurs transcriptionnels chez les archées. Ainsi, les protéines SurR et TrmBL1 codées par *P. furiosus* peuvent agir à la fois comme activateurs ou comme répresseurs de la transcription des gènes concernés selon la localisation de leurs sites de liaison (Lipscomb *et al.*, 2009; Lee *et al.*, 2008). Egalement, le régulateur Ss-LrpB de *S. solfataricus*, qui porte un domaine de liaison à l'ADN de type HTH, agit dans les deux sens et cela de façon dépendante de la concentration (Peeters *et al.*, 2004).

Le sens de régulation par un régulateur transcriptionnel ne semble pas dépendre réellement de mécanismes spécifiques de liaison à l'ADN mais dépend plutôt de la localisation du site de liaison par rapport au promoteur (décrit dans l'Introduction, Figure 18).

Pour certaines des cibles de SvtR, il est compris assez bien comment la protéine influence la transcription, pour d'autres, le mécanisme reste à déterminer. Par exemple, dans le cas du gène *gp08*, le mécanisme de sa répression par SvtR semble être assez clair. Dans ce cas, le site de liaison de la protéine SvtR recouvre le site d'initiation de la transcription, laissant suggérer que la liaison de SvtR n'empêche pas la fixation des facteurs TBP et TFB mais l'avancement de l'ARNP, également recrutée, vers le site +1 (Figure 75A). Un des avantages de ce type de répression peut être la possibilité de réactiver rapidement la transcription de *gp08* sans repasser par la phase de formation du complexe de pré-initiation.

Pour deux autres cibles de SvtR, *gp03* et *gp30*, pour lesquelles SvtR est soit un répresseur (*gp03*) soit un activateur (*gp30*) nous avons constaté que les sites de fixations correspondants se trouvaient à une telle distance du promoteur qu'aucune interaction directe de SvtR avec le promoteur n'était envisageable. Pour *gp03*, le site de SvtR est à quarante-sept nucléotides en amont du site d'initiation de la transcription et dans le cas de *gp30* il est à soixante-treize nucléotides. SvtR agit dans ces deux cas à distance et nous avons pu décrire le mécanisme de son action sur le modèle de *gp30*.

Le travail réalisé au cours de cette thèse a mis en évidence que la protéine SvtR se liait sur un site principal et polymérisait pour atteindre la région promotrice du gène *gp30*. L'oligomère de SvtR atteint et recouvre partiellement la séquence BRE du promoteur de *gp30*. Les détails du mécanisme utilisé par SvtR pour activer la transcription du gène *gp30* ne sont pas connus mais il peut être supposé qu'il facilite le recrutement d'un des trois éléments du complexe de pré-initiation, soit des facteurs TBP et/ou TFB, soit directement de l'ARNP (Figure 75). Les séquences UAS (« *upstream activating sequence* ») localisées en amont ou chevauchant partiellement l'élément BRE sont connues chez les archées et elles peuvent être situées aussi bien en amont qu'en aval de la *TATA-box*. Il a été démontré *in vitro* qu'elles amélioreraient ainsi directement le recrutement *in vitro* de TBP ou TFB à la *TATA-box* (Ouhammouch *et al.*, 2003; 2005; Peng *et al.*, 2009).

De façon courante, les activateurs transcriptionnels se lient à un site opérateur, localisé en amont ou chevauchant partiellement l'élément BRE et appelé « *upstream activating sequence* » (UAS). Des sites opérateurs auxiliaires sont localisés en amont du premier site, en aval de la *TATA-box* ou en aval du site d'initiation de la transcription. Chez *Methanococcus jannaschii*, le régulateur Ptr2, l'activateur transcriptionnel le mieux étudié chez les archées, active la transcription en recrutant directement TBP au promoteur (Ouhammouch *et al.*, 2003; 2005). Un mécanisme similaire est utilisé par GvpE (Teufel *et al.*, 2010). Récemment, Ochs *et al.*, ont mis en évidence l'existence du régulateur PF1088, chez *P. furiosus* dont le site de liaison est localisé juste en amont de la séquence BRE (Ochs *et al.*, 2012). Il s'agit du premier exemple de participation d'un régulateur transcriptionnel dans le recrutement direct de TFB.

Il peut être supposé que SvtR active la transcription de *gp30* de la même manière, en interagissant avec les facteurs TBP et/ou TFB (Figure 75), mais au préalable la protéine devrait atteindre sa cible en se polymérisant sur la matrice d'ADN. Par deux approches indépendantes, l'approche de *footprinting* et l'observation directe en AFM et EM, nos travaux ont démontré l'existence réelle de ce processus, sa spécificité et l'étendue de la région d'ADN couverte par l'oligomère de SvtR. Nous avons démontré que la présence du site principal est indispensable pour initier l'oligomérisation. Le processus d'oligomérisation semble être guidé par de courts sites dégénérés détectés entre le site principal et la *TATA-box*. L'analyse mutationnelle confirme l'importance de ces sites car en leur absence la polymérisation de SvtR est très peu efficace. Quant à l'effet biologique sur l'initiation de transcription il n'est pas exclu qu'il puisse aussi être lié au changement global de la conformation de l'ADN dans la région promotrice conduisant au recrutement plus efficace des facteurs TBP et/ou TFB.

L'action par polymérisation a déjà été décrite pour d'autres protéines de la famille RHH. Ainsi la protéine CopG bactérienne se polymérise sur l'ADN à partir de son site principal grâce à la présence de sites secondaires de liaison présents dans l'opérateur du gène régulé par CopG (Costa *et al.*, 2001).

De la même façon, la protéine AvtR codée par AFV1, un virus d'archées hyperthermophiles, réprime l'expression de son propre gène *gp29* et agit comme activateur d'un autre gène *gp30* grâce à la polymérisation sur l'ADN débutant aux sites principaux de liaison de AvtR (Peixeiro *et al.*, 2013).

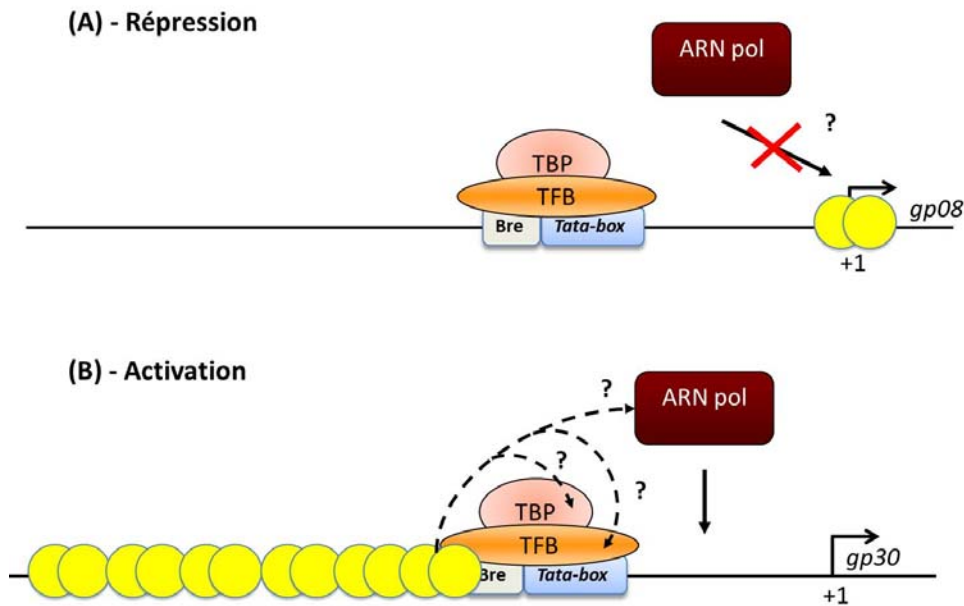


Figure 75 : Modèles de régulation de SvtR pour son propre gène (agit comme un répresseur) et un gène codant une protéine de capsid (agit comme un activateur). (A) En se liant au niveau du site d'initiation de la transcription, le répresseur SvtR empêche peut-être le recrutement de l'ARNP et donc l'initiation de la transcription de son propre gène. (B) La protéine SvtR se lie à distance des régions promotrices du gène *gp30* et polymérise jusqu'à recouvrir partiellement la séquence BRE. Elle active la transcription du gène par un mécanisme encore indéterminé.

La signification biologique de ce mécanisme inhabituel de régulation décrit pour SvtR reste incertaine. L'efficacité de polymérisation de SvtR devrait être assez sensible à la concentration intracellulaire de la protéine. Les données transcriptomiques montrent la forte présence du transcrit de *gp08* jusqu'aux phases tardives du cycle viral. Il peut être supposé que la protéine SvtR, par ailleurs très stable dans les conditions *in vitro*, s'accumule progressivement et sa concentration intracellulaire pourrait signaler au système le temps écoulé depuis le début d'infection. Dans ce cas, SvtR peut jouer le rôle d'une « horloge moléculaire » synchronisant par l'activation ou la répression de l'activité de ces promoteurs cibles différents stades du cycle viral.

La suite de ce travail serait la mise en évidence de l'interaction supposée de SvtR et des facteurs TFB et TBP. Pour cela, nous disposons à présent d'anticorps anti-TBP permettant de réaliser des complexes *super-shifts*, en présence de TBP, TFB, SvtR et même de l'ARNP. A l'aide des anticorps, nous pourrions identifier comment et dans quel ordre la présence de SvtR favorise la formation du complexe de pré-initiation complet. Une autre idée est de varier les concentrations des facteurs en présence de SvtR pour observer l'effet sur la transcription. Si SvtR recrute le facteur TBP par exemple, l'efficacité de transcription en présence de faible quantité de TBP devrait être augmentée, tandis qu'en présence de quantité excessive de TBP, cet effet ne sera plus aussi visible. Enfin, des expériences de transcription *in vitro* pourraient être conduites sur des promoteurs mutés.

De façon plus générale et à plus long terme, il serait intéressant d'identifier les sites réellement occupés par SvtR dans les conditions *in vivo*. La question se pose également sur la participation de SvtR dans la régulation de l'activité des gènes de l'hôte au cours du cycle d'infection du virus SIRV. Le ChIP-on-chip [ChIP (immuno-précipitation de la chromatine) utilisant des puces à ADN] ou le ChIP-seq (ChIP suivi par du séquençage haut débit) qui sont parmi les techniques les plus puissantes pour

mapper les sites de liaison à l'ADN d'un régulateur transcriptionnel à l'échelle du génome dans un contexte *in vivo*, n'ont que rarement été appliquées aux archées (Schmid *et al.*, 2009; Duc *et al.*, 2012; Song *et al.*, 2013; Nguyen *et al.*, 2013), malgré leur succès prouvé chez les systèmes eucaryotes et bactériens.

Récemment, Nguyen-Duc *et al.* ont développé une approche de CHIP-on-Chip avec des *nanobodies*[®] comme réactifs d'affinité (Nguyen-Duc *et al.*, 2012). Les *nanobodies*[®] sont des fragments d'anticorps avec un domaine unique dérivé des chaînes lourdes des camélidés (Nguyen-Duc *et al.*, 2012). Cette nouvelle technique a permis de révéler des sites opérateurs auxiliaires dans l'ORF de Ss-IrpB (Nguyen-Duc *et al.*, 2012) et d'identifier soixante-dix nouveaux sites de liaison pour le régulateur LysM dans le génome *S. solfataricus*, révélant que le régulateur LysM était un régulateur beaucoup plus polyvalent qu'initialement établi (Song *et al.*, 2013).

Il pourrait être envisagé de réaliser des expériences similaires chez *S. islandicus* infectée par le virus SIRV1/2 et identifier l'ensemble de cibles de SvtR. La même approche globale peut être également utilisée pour l'étude d'autres régulateurs viraux et cellulaires, comme Sta1.

L'interaction entre SvtR et l'ADN peut être plus complexe que nous avons pensé initialement. Les observations réalisées sous microscopie électronique ont révélé la présence de structures en forme de croix impliquant deux molécules d'ADN reliées par SvtR. Le centre de la nucléation observée se trouve très probablement dans la région d'ADN correspondant au site principal de fixation de SvtR. Il semble que SvtR possède la capacité de relier au moins deux molécules d'ADN dans un complexe stable *in vitro*. Il reste à démontrer l'existence de ces structures inter- ou intramoléculaires *in vivo*. Si c'est le cas, cela signifie l'existence d'une autre fonction très importante de SvtR, la participation à la structuration du génome viral au cours du cycle. La Figure 76 présente le modèle hypothétique expliquant comment SvtR pourrait participer à la compaction du génome viral en une structure plus réduite. Cette hypothèse peut être vérifiée en observant sous microscopie de l'ADN cible linéaire et circulaire contenant non pas un mais deux ou plus sites de liaison de SvtR. Si les deux sites de nucléations interagissent même en étant sur la même molécule, des structures en « γ » pour l'ADN linéaire (Figures 76A1) et en « 8 » pour l'ADN circulaire (Figures 76A2) devraient être observées. Un tel rôle des protéines de la famille RHH n'a jamais été démontré auparavant et le mécanisme de structuration du génome viral supposé par notre hypothèse est totalement inédit (Figure 76B).

Les gènes codant pour des protéines putatives de liaison à l'ADN, portant le motif RHH sont largement représentés dans les génomes d'archées et de leurs virus. Le mécanisme de contrôle impliquant la polymérisation de la protéine régulatrice sur la matrice d'ADN aussi bien que sa participation à la structuration de l'ADN pourrait être des mécanismes généraux répandus chez les archées.

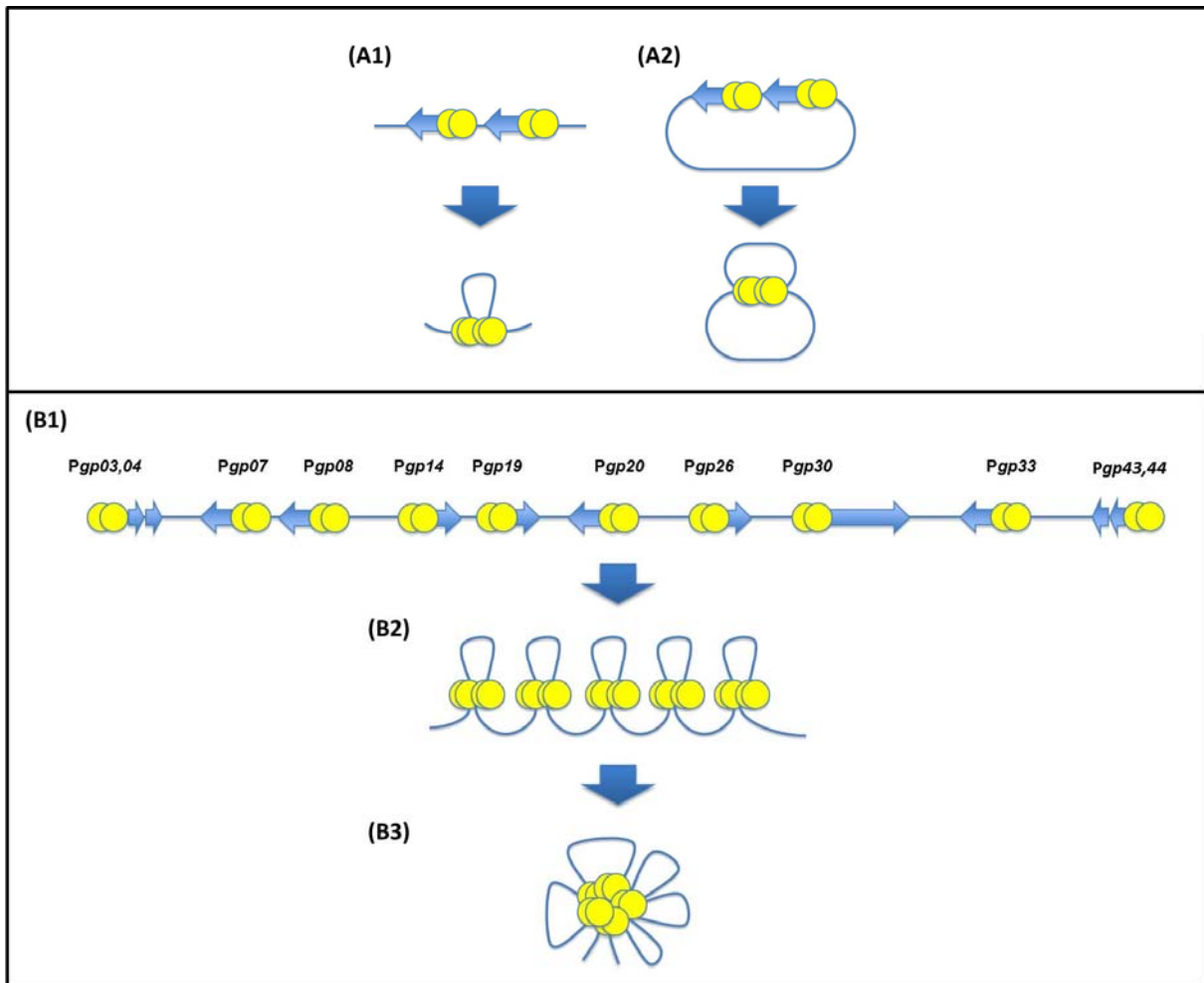


Figure 76 : Modèle de compaction de l'ADN par la protéine SvtR. (A) Structures hypothétiques attendues sous microscopie avec une molécule linéaire (A1) ou circulaire (A2) contenant deux sites de liaison de la protéine SvtR. (B) Modèle de compaction hypothétique du génome viral de SIRV1 par la protéine SvtR. L'ADN génomique linéaire du virus SIRV1 contient de nombreux sites de liaison de SvtR (B1). Si les sites de nucléation peuvent interagir entre eux (B2), il peut être supposé SvtR participe à la transformation de l'ADN linéaire en une structure plus compacte (B3).

Partie IV – Conclusions

La transcription chez les Archées est réalisée et régulée par des protéines à la fois eucaryotiques et bactériennes qui forment une mosaïque complexe. D'un côté, la transcription est initiée à l'aide de facteurs généraux de transcription de type eucaryote associés à une ARNP qui ressemble beaucoup à l'ARNPII eucaryote. De l'autre côté, la transcription est contrôlée par des régulateurs transcriptionnels dont la majorité a des motifs de liaison RHH ou HTH partagés avec les Bactéries. Jusqu'à présent l'existence de régulateurs de transcription spécifiques aux Archées n'a jamais été confirmée.

Au cours de ce travail, nous avons démontré que la protéine AFV1p06, codée par le virus d'archées, AFV1 et possédant un domaine de liaison à l'ADN de type doigt de zinc C2H2 était capable de lier l'ADN. Les domaines de liaison à l'ADN de type doigt de zinc sont habituellement assimilés aux protéines de liaison à l'ADN chez les eucaryotes. La caractérisation du premier exemple de ce type de régulateurs chez les archées indique que la régulation de transcription n'est absolument pas confinée aux régulateurs de type bactérien et des efforts supplémentaires devraient être entrepris pour en découvrir et caractériser d'autres.

En étudiant la protéine régulatrice SvtR, codée par un autre virus d'archées, SIRV1, servant comme modèle pour l'étude de l'interaction hôte-virus, nous avons identifié et caractérisé son régulon dans le génome viral et démontré qu'il s'agit d'un régulateur transcriptionnel très important dans le cycle de SIRV1 contrôlant l'activité d'au moins quatre promoteurs. Ce régulateur transcriptionnel SvtR à double facette, à la fois activateur et répresseur selon le promoteur cible, agit sur certaines de ses cibles par un mécanisme original de polymérisation lui permettant d'atteindre le promoteur régulé. L'avantage de ce mécanisme reste incertain. Il peut être supposé que ce nouveau mécanisme permet une expression hiérarchisée des gènes viraux durant le cycle d'infection. L'accumulation graduelle des protéines SvtR pendant le développement du virus régulerait d'après cette hypothèse certains gènes viraux de façon dépendante de sa concentration, agissant ainsi comme une « horloge » interne.

Au cours de ma thèse, j'ai également développé des approches génétiques pour une souche modèle pour l'étude des interactions hôtes-virus chez les crénaarchées, *S. islandicus* LAL14/1. Le génome de cette souche est séquencé et annoté et il est désormais possible d'obtenir des mutants KO pour des gènes candidats choisis. Ainsi, le mutant Δ CRISPR_1, dont l'obtention ne représente que la première étape dans la création d'une souche de *S. islandicus* LAL14/1 totalement dépourvue de CRISPRs permettra de mieux comprendre le rôle des CRISPRs chez les archées. L'étude des gènes essentiels nécessite le développement d'un système génétique plus sophistiqué qui peut maintenant être réalisé sur la base des résultats, des souches et des approches méthodologiques développés dans cette thèse.

Références

- Abe, H., M. Seki, F. Ohbayashi, N. Tanaka, J. Yamashita, T. Fujii, T. Yokoyama, M. Takahashi, Y. Banno, K. Sahara, A. Yoshido, J. Ihara, Y. Yasukochi, K. Mita, M. Ajimura, M. G. Suzuki, T. Oshiki and T. Shimada (2005). "Partial deletions of the W chromosome due to reciprocal translocation in the silkworm *Bombyx mori*." *Insect Molecular Biology* **14**(4): 339-352.
- Abella, M., S. Rodriguez, S. Paytubi, S. Campoy, M. F. White and J. Barbe (2007). "The *Sulfolobus solfataricus* radA paralogue sso0777 is DNA damage inducible and positively regulated by the Sta1 protein." *Nucleic acids research* **35**(20): 6788-6797.
- Ackermann, H. W. and D. Prangishvili (2012). "Prokaryote viruses studied by electron microscopy." *Archives of Virology* **157**(10): 1843-1849.
- Ajon, M., S. Frols, M. van Wolferen, K. Stoecker, D. Teichmann, A. J. Driessen, D. W. Grogan, S. V. Albers and C. Schleper (2011). "UV-inducible DNA exchange in hyperthermophilic archaea mediated by type IV pili." *Molecular Microbiology* **82**(4): 807-817.
- Albers, S. V. and A. J. Driessen (2008). "Conditions for gene disruption by homologous recombination of exogenous DNA into the *Sulfolobus solfataricus* genome." *Archaea (Vancouver, B.C.)* **2**(3): 145-149.
- Albers, S. V., P. Forterre, D. Prangishvili and C. Schleper (2013). "The legacy of Carl Woese and Wolfram Zillig: from phylogeny to landmark discoveries." *Nature reviews. Microbiology* **11**(10): 713-719.
- Albers, S. V., M. Jonuscheit, S. Dinkelaker, T. Urich, A. Kletzin, R. Tampe, A. J. Driessen and C. Schleper (2006). "Production of recombinant and tagged proteins in the hyperthermophilic archaeon *Sulfolobus solfataricus*." *Applied and Environmental Microbiology* **72**(1): 102-111.
- Albers, S. V., S. M. Koning, W. N. Konings and A. J. Driessen (2004). "Insights into ABC transport in archaea." *Journal of Bioenergetics and Biomembranes* **36**(1): 5-15.
- Albers, S. V. and B. H. Meyer (2011). "The archaeal cell envelope." *Nature reviews. Microbiology* **9**(6): 414-426.
- Allers, T. and M. Mevarech (2005). "Archaeal genetics - the third way." *Nature reviews. Genetics* **6**(1): 58-73.
- Anderson, I. J., H. Sun, A. Lapidus, A. Copeland, T. Glavina Del Rio, H. Tice, E. Dalin, S. Lucas, K. Barry, M. Land, P. Richardson, H. Huber and N. C. Kyrpides (2009). "Complete genome sequence of *Staphylothermus marinus* Stetter and Fiala 1986 type strain F1." *Standards in genomic sciences* **1**(2): 183-188.
- Andrecka, J., B. Treutlein, M. A. Arcusa, A. Muschielok, R. Lewis, A. C. Cheung, P. Cramer and J. Michaelis (2009). "Nano positioning system reveals the course of upstream and nontemplate DNA within the RNA polymerase II elongation complex." *Nucleic acids research* **37**(17): 5803-5809.
- Ao, X., Y. Li, F. Wang, M. Feng, Y. Lin, S. Zhao, Y. Liang and N. Peng (2013). "The *sulfolobus* initiator element is an important contributor to promoter strength." *Journal of Bacteriology* **195**(22): 5216-5222.
- Ao, X., Y. Li, F. Wang, M. Feng, Y. Lin, S. Zhao, Y. Liang and N. Peng (2013). "Sulfolobus Initiator Element: An Important Contributor to Promoter Strength." *Journal of Bacteriology*.
- Aravalli, R. N. and R. A. Garrett (1997). "Shuttle vectors for hyperthermophilic archaea." *Extremophiles : life under extreme conditions* **1**(4): 183-191.

- Aravind, L., V. Anantharaman, S. Balaji, M. M. Babu and L. M. Iyer (2005). "The many faces of the helix-turn-helix domain: transcription regulation and beyond." FEMS Microbiology Reviews **29**(2): 231-262.
- Aravind, L. and E. V. Koonin (1999). "DNA-binding proteins and evolution of transcription regulation in the archaea." Nucleic acids research **27**(23): 4658-4670.
- Ariza, A., D. J. Richard, M. F. White and C. S. Bond (2005). "Conformational flexibility revealed by the crystal structure of a crenarchaeal RadA." Nucleic acids research **33**(5): 1465-1473.
- Armache, K. J., S. Mitterweger, A. Meinhart and P. Cramer (2005). "Structures of complete RNA polymerase II and its subcomplex, Rpb4/7." The Journal of biological chemistry **280**(8): 7131-7134.
- Arnold, H. P., Q. She, H. Phan, K. Stedman, D. Prangishvili, I. Holz, J. K. Kristjansson, R. Garrett and W. Zillig (1999). "The genetic element pSSVx of the extremely thermophilic crenarchaeon *Sulfolobus* is a hybrid between a plasmid and a virus." Molecular Microbiology **34**(2): 217-226.
- Atanasova, N. S., E. Roine, A. Oren, D. H. Bamford and H. M. Oksanen (2012). "Global network of specific virus-host interactions in hypersaline environments." Environmental Microbiology **14**(2): 426-440.
- Atomi, H., T. Imanaka and T. Fukui (2012). "Overview of the genetic tools in the Archaea." Frontiers in microbiology **3**: 337.
- Aucelli, T., P. Contursi, M. Girfoglio, M. Rossi and R. Cannio (2006). "A spreadable, non-integrative and high copy number shuttle vector for *Sulfolobus solfataricus* based on the genetic element pSSVx from *Sulfolobus islandicus*." Nucleic acids research **34**(17): e114.
- Auernik, K. S., Y. Maezato, P. H. Blum and R. M. Kelly (2008). "The genome sequence of the metal-mobilizing, extremely thermoacidophilic archaeon *Metallosphaera sedula* provides insights into bioleaching-associated metabolism." Applied and Environmental Microbiology **74**(3): 682-692.
- Augustus, A. M., P. N. Reardon, W. T. Heller and L. D. Spicer (2006). "Structural basis for the differential regulation of DNA by the methionine repressor MetJ." The Journal of biological chemistry **281**(45): 34269-34276.
- Baliga, N. S., Y. A. Goo, W. V. Ng, L. Hood, C. J. Daniels and S. DasSarma (2000). "Is gene expression in *Halobacterium* NRC-1 regulated by multiple TBP and TFB transcription factors?" Molecular Microbiology **36**(5): 1184-1185.
- Banerjee, A., A. Ghosh, D. J. Mills, J. Kahnt, J. Vonck and S. V. Albers (2012). "FlaX, a unique component of the crenarchaeal archaellum, forms oligomeric ring-shaped structures and interacts with the motor ATPase FlaI." The Journal of biological chemistry **287**(52): 43322-43330.
- Barns, S. M., C. F. Delwiche, J. D. Palmer and N. R. Pace (1996). "Perspectives on archaeal diversity, thermophily and monophyly from environmental rRNA sequences." Proceedings of the National Academy of Sciences of the United States of America **93**(17): 9188-9193.
- Barns, S. M., R. E. Fundyga, M. W. Jeffries and N. R. Pace (1994). "Remarkable archaeal diversity detected in a Yellowstone National Park hot spring environment." Proceedings of the National Academy of Sciences of the United States of America **91**(5): 1609-1613.
- Barry, E. R. and S. D. Bell (2006). "DNA replication in the archaea." Microbiology and molecular biology reviews : MMBR **70**(4): 876-887.

- Barthelme, D., U. Scheele, S. Dinkelaker, A. Janoschka, F. Macmillan, S. V. Albers, A. J. Driessen, M. S. Stagni, E. Bill, W. Meyer-Klaucke, V. Schunemann and R. Tampe (2007). "Structural organization of essential iron-sulfur clusters in the evolutionarily highly conserved ATP-binding cassette protein ABCE1." The Journal of biological chemistry **282**(19): 14598-14607.
- Bartolucci, S., P. Contursi, G. Fiorentino, D. Limauro and E. Pedone (2013). "Responding to toxic compounds: a genomic and functional overview of Archaea." Frontiers in bioscience (Landmark edition) **18**: 165-189.
- Bartolucci, S., M. Rossi and R. Cannio (2003). "Characterization and functional complementation of a nonlethal deletion in the chromosome of a beta-glycosidase mutant of *Sulfolobus solfataricus*." Journal of Bacteriology **185**(13): 3948-3957.
- Bartossek, R., A. Spang, G. Weidler, A. Lanzen and C. Schleper (2012). "Metagenomic analysis of ammonia-oxidizing archaea affiliated with the soil group." Frontiers in microbiology **3**: 208.
- Bechhofer, D. H. (2011). "Bacillus subtilis mRNA decay: new parts in the toolkit." Wiley interdisciplinary reviews. RNA **2**(3): 387-394.
- Bell, S. D. (2005). "Archaeal transcriptional regulation--variation on a bacterial theme?" Trends in Microbiology **13**(6): 262-265.
- Bell, S. D., C. H. Botting, B. N. Wardleworth, S. P. Jackson and M. F. White (2002). "The interaction of Alba, a conserved archaeal chromatin protein, with Sir2 and its regulation by acetylation." Science (New York, N.Y.) **296**(5565): 148-151.
- Bell, S. D., A. B. Brinkman, J. van der Oost and S. P. Jackson (2001). "The archaeal TFIIIEalpha homologue facilitates transcription initiation by enhancing TATA-box recognition." EMBO reports **2**(2): 133-138.
- Bell, S. D. and S. P. Jackson (1998). "Transcription and translation in Archaea: a mosaic of eukaryal and bacterial features." Trends in Microbiology **6**(6): 222-228.
- Bell, S. D. and S. P. Jackson (1998). "Transcription in Archaea." Cold Spring Harbor Symposia on Quantitative Biology **63**: 41-51.
- Bell, S. D. and S. P. Jackson (2000). "Mechanism of autoregulation by an archaeal transcriptional repressor." The Journal of biological chemistry **275**(41): 31624-31629.
- Bell, S. D. and S. P. Jackson (2001). "Mechanism and regulation of transcription in archaea." Current Opinion in Microbiology **4**(2): 208-213.
- Bell, S. D., C. Jaxel, M. Nadal, P. F. Kosa and S. P. Jackson (1998). "Temperature, template topology, and factor requirements of archaeal transcription." Proceedings of the National Academy of Sciences of the United States of America **95**(26): 15218-15222.
- Bell, S. D., P. L. Kosa, P. B. Sigler and S. P. Jackson (1999). "Orientation of the transcription preinitiation complex in archaea." Proceedings of the National Academy of Sciences of the United States of America **96**(24): 13662-13667.
- Bell, S. D., C. P. Magill and S. P. Jackson (2001). "Basal and regulated transcription in Archaea." Biochemical Society Transactions **29**(Pt 4): 392-395.
- Bennett, P. M. (2004). "Genome plasticity: insertion sequence elements, transposons and integrons, and DNA rearrangement." Methods in molecular biology (Clifton, N.J.) **266**: 71-113.
- Benson, L. J., J. A. Phillips, Y. Gu, M. R. Parthun, C. S. Hoffman and A. T. Annunziato (2007). "Properties of the type B histone acetyltransferase Hat1: H4 tail interaction, site preference, and involvement in DNA repair." The Journal of biological chemistry **282**(2): 836-842.

- Berkner, S., D. Grogan, S. V. Albers and G. Lipps (2007). "Small multicopy, non-integrative shuttle vectors based on the plasmid pRN1 for *Sulfolobus acidocaldarius* and *Sulfolobus solfataricus*, model organisms of the (cren-)archaea." Nucleic acids research **35**(12): e88.
- Berkner, S. and G. Lipps (2007). "An active nonautonomous mobile element in *Sulfolobus islandicus* REN1H1." Journal of Bacteriology **189**(5): 2145-2149.
- Berkner, S. and G. Lipps (2007). "Characterization of the transcriptional activity of the cryptic plasmid pRN1 from *Sulfolobus islandicus* REN1H1 and regulation of its replication operon." Journal of Bacteriology **189**(5): 1711-1721.
- Berkner, S. and G. Lipps (2008). "Genetic tools for *Sulfolobus* spp.: vectors and first applications." Archives of Microbiology **190**(3): 217-230.
- Berkner, S. and G. Lipps (2008). "Mutation and reversion frequencies of different *Sulfolobus* species and strains." Extremophiles : life under extreme conditions **12**(2): 263-270.
- Berkner, S., A. Wlodkowski, S. V. Albers and G. Lipps (2010). "Inducible and constitutive promoters for genetic systems in *Sulfolobus acidocaldarius*." Extremophiles : life under extreme conditions **14**(3): 249-259.
- Bettstetter, M., X. Peng, R. A. Garrett and D. Prangishvili (2003). "AFV1, a novel virus infecting hyperthermophilic archaea of the genus *acidianus*." Virology **315**(1): 68-79.
- Bikard, D., A. Hatoum-Aslan, D. Mucida and L. A. Marraffini (2012). "CRISPR interference can prevent natural transformation and virulence acquisition during in vivo bacterial infection." Cell host & microbe **12**(2): 177-186.
- Birkenbihl, R. P., K. Neef, D. Prangishvili and B. Kemper (2001). "Holliday junction resolving enzymes of archaeal viruses SIRV1 and SIRV2." Journal of Molecular Biology **309**(5): 1067-1076.
- Bitan-Banin, G., R. Ortenberg and M. Mevarech (2003). "Development of a gene knockout system for the halophilic archaeon *Haloferax volcanii* by use of the *pyrE* gene." Journal of Bacteriology **185**(3): 772-778.
- Bizard, A., F. Garnier and M. Nadal (2011). "TopR2, the second reverse gyrase of *Sulfolobus solfataricus*, exhibits unusual properties." Journal of Molecular Biology **408**(5): 839-849.
- Bize, A., E. A. Karlsson, K. Ekefjard, T. E. Quax, M. Pina, M. C. Prevost, P. Forterre, O. Tenaillon, R. Bernander and D. Prangishvili (2009). "A unique virus release mechanism in the Archaea." Proceedings of the National Academy of Sciences of the United States of America **106**(27): 11306-11311.
- Blaseio, U. and F. Pfeifer (1990). "Transformation of *Halobacterium halobium*: development of vectors and investigation of gas vesicle synthesis." Proceedings of the National Academy of Sciences of the United States of America **87**(17): 6772-6776.
- Blount, Z. D. and D. W. Grogan (2005). "New insertion sequences of *Sulfolobus*: functional properties and implications for genome evolution in hyperthermophilic archaea." Molecular Microbiology **55**(1): 312-325.
- Blum, H., W. Zillig, S. Mallok, H. Domdey and D. Prangishvili (2001). "The genome of the archaeal virus SIRV1 has features in common with genomes of eukaryal viruses." Virology **281**(1): 6-9.
- Boeke, J. D., F. LaCroute and G. R. Fink (1984). "A positive selection for mutants lacking orotidine-5'-phosphate decarboxylase activity in yeast: 5-fluoro-orotic acid resistance." Molecular & general genetics : MGG **197**(2): 345-346.
- Borukhov, S., J. Lee and O. Laptenko (2005). "Bacterial transcription elongation factors: new insights into molecular mechanism of action." Molecular Microbiology **55**(5): 1315-1324.

- Bouhouche, N., M. Syvanen and C. I. Kado (2000). "The origin of prokaryotic C2H2 zinc finger regulators." Trends in Microbiology **8**(2): 77-81.
- Bowers, P. M., L. E. Schaufler and R. E. Klevit (1999). "A folding transition and novel zinc finger accessory domain in the transcription factor ADR1." Nature Structural Biology **6**(5): 478-485.
- Bowie, J. U. and R. T. Sauer (1989). "Identifying determinants of folding and activity for a protein of unknown structure." Proceedings of the National Academy of Sciences of the United States of America **86**(7): 2152-2156.
- Bowie, J. U. and R. T. Sauer (1990). "TraY proteins of F and related episomes are members of the Arc and Mnt repressor family." Journal of Molecular Biology **211**(1): 5-6.
- Breg, J. N., J. H. van Opheusden, M. J. Burgering, R. Boelens and R. Kaptein (1990). "Structure of Arc repressor in solution: evidence for a family of beta-sheet DNA-binding proteins." Nature **346**(6284): 586-589.
- Bricogne, G., C. Vonrhein, C. Flensburg, M. Schiltz and W. Paciorek (2003). "Generation, representation and flow of phase information in structure determination: recent developments in and around SHARP 2.0." Acta Crystallogr D Biol Crystallogr **59**(Pt 11): 2023-2030.
- Brinkman, A. B., S. D. Bell, R. J. Lebbink, W. M. de Vos and J. van der Oost (2002). "The *Sulfolobus solfataricus* Lrp-like protein LysM regulates lysine biosynthesis in response to lysine availability." The Journal of biological chemistry **277**(33): 29537-29549.
- Brinkman, A. B., I. Dahlke, J. E. Tuininga, T. Lammers, V. Dumay, E. de Heus, J. H. Lebbink, M. Thomm, W. M. de Vos and J. van Der Oost (2000). "An Lrp-like transcriptional regulator from the archaeon *Pyrococcus furiosus* is negatively autoregulated." The Journal of biological chemistry **275**(49): 38160-38169.
- Brinkman, A. B., T. J. Ettema, W. M. de Vos and J. van der Oost (2003). "The Lrp family of transcriptional regulators." Molecular Microbiology **48**(2): 287-294.
- Brochier-Armanet, C., B. Boussau, S. Gribaldo and P. Forterre (2008). "Mesophilic Crenarchaeota: proposal for a third archaeal phylum, the Thaumarchaeota." Nature reviews. Microbiology **6**(3): 245-252.
- Brochier-Armanet, C. and P. Forterre (2007). "Widespread distribution of archaeal reverse gyrase in thermophilic bacteria suggests a complex history of vertical inheritance and lateral gene transfers." Archaea (Vancouver, B.C.) **2**(2): 83-93.
- Brochier-Armanet, C., P. Forterre and S. Gribaldo (2011). "Phylogeny and evolution of the Archaea: one hundred genomes later." Current Opinion in Microbiology **14**(3): 274-281.
- Brochier, C., S. Gribaldo, Y. Zivanovic, F. Confalonieri and P. Forterre (2005). "Nanoarchaea: representatives of a novel archaeal phylum or a fast-evolving euryarchaeal lineage related to Thermococcales?" Genome biology **6**(5): R42.
- Brock, T. D., K. M. Brock, R. T. Belly and R. L. Weiss (1972). "Sulfolobus: a new genus of sulfur-oxidizing bacteria living at low pH and high temperature." Archiv fur Mikrobiologie **84**(1): 54-68.
- Brown, R. S. (2005). "Zinc finger proteins: getting a grip on RNA." Current Opinion in Structural Biology **15**(1): 94-98.
- Brueckner, F., K. J. Armache, A. Cheung, G. E. Damsma, H. Kettenberger, E. Lehmann, J. Sydow and P. Cramer (2009). "Structure-function studies of the RNA polymerase II elongation complex." Acta crystallographica. Section D, Biological crystallography **65**(Pt 2): 112-120.

- Brugger, K., P. Redder, Q. She, F. Confalonieri, Y. Zivanovic and R. A. Garrett (2002). "Mobile elements in archaeal genomes." FEMS Microbiology Letters **206**(2): 131-141.
- Brugger, K., E. Torarinsson, P. Redder, L. Chen and R. A. Garrett (2004). "Shuffling of *Sulfolobus* genomes by autonomous and non-autonomous mobile elements." Biochemical Society Transactions **32**(Pt 2): 179-183.
- Burgering, M. J., R. Boelens, D. E. Gilbert, J. N. Breg, K. L. Knight, R. T. Sauer and R. Kaptein (1994). "Solution structure of dimeric Mnt repressor (1-76)." Biochemistry **33**(50): 15036-15045.
- Calvo, J. M. and R. G. Matthews (1994). "The leucine-responsive regulatory protein, a global regulator of metabolism in *Escherichia coli*." Microbiological Reviews **58**(3): 466-490.
- Campbell, E. A., N. Korzheva, A. Mustaev, K. Murakami, S. Nair, A. Goldfarb and S. A. Darst (2001). "Structural mechanism for rifampicin inhibition of bacterial rna polymerase." Cell **104**(6): 901-912.
- Cannio, R., P. Contursi, M. Rossi and S. Bartolucci (1998). "An autonomously replicating transforming vector for *Sulfolobus solfataricus*." Journal of Bacteriology **180**(12): 3237-3240.
- Cannio, R., P. Contursi, M. Rossi and S. Bartolucci (2001). "Thermoadaptation of a mesophilic hygromycin B phosphotransferase by directed evolution in hyperthermophilic Archaea: selection of a stable genetic marker for DNA transfer into *Sulfolobus solfataricus*." Extremophiles : life under extreme conditions **5**(3): 153-159.
- Cannio, R., G. Fiorentino, M. Rossi and S. Bartolucci (1999). "The alcohol dehydrogenase gene: distribution among Sulfolobales and regulation in *Sulfolobus solfataricus*." FEMS Microbiology Letters **170**(1): 31-39.
- Carey, M. F., C. L. Peterson and S. T. Smale (2012). "Experimental strategies for the identification of DNA-binding proteins." Cold Spring Harb Protoc **2012**(1): 18-33.
- Charlebois, R. L., W. L. Lam, S. W. Cline and W. F. Doolittle (1987). "Characterization of pHV2 from *Halobacterium volcanii* and its use in demonstrating transformation of an archaeobacterium." Proceedings of the National Academy of Sciences of the United States of America **84**(23): 8530-8534.
- Charoensawan, V., D. Wilson and S. A. Teichmann (2010). "Genomic repertoires of DNA-binding transcription factors across the tree of life." Nucleic acids research **38**(21): 7364-7377.
- Chen, L., K. Brugger, M. Skovgaard, P. Redder, Q. She, E. Torarinsson, B. Greve, M. Awayez, A. Zibat, H. P. Klenk and R. A. Garrett (2005). "The genome of *Sulfolobus acidocaldarius*, a model organism of the Crenarchaeota." Journal of Bacteriology **187**(14): 4992-4999.
- Cheung, P., C. D. Allis and P. Sassone-Corsi (2000). "Signaling to chromatin through histone modifications." Cell **103**(2): 263-271.
- Chivers, P. T. and R. T. Sauer (2000). "Regulation of high affinity nickel uptake in bacteria. Ni²⁺-Dependent interaction of NikR with wild-type and mutant operator sites." J Biol Chem **275**(26): 19735-19741.
- Chopra, N., S. Agarwal, S. Verma, S. Bhatnagar and R. Bhatnagar (2011). "Modeling of the structure and interactions of the *B. anthracis* antitoxin, MoxX: deletion mutant studies highlight its modular structure and repressor function." Journal of computer-aided molecular design **25**(3): 275-291.
- Ciampi, M. S. (2006). "Rho-dependent terminators and transcription termination." Microbiology (Reading, England) **152**(Pt 9): 2515-2528.

- Cline, S. W. and W. F. Doolittle (1987). "Efficient transfection of the archaeobacterium *Halobacterium halobium*." Journal of Bacteriology **169**(3): 1341-1344.
- Clore, A. J. and K. M. Stedman (2007). "The SSV1 viral integrase is not essential." Virology **361**(1): 103-111.
- Cobucci-Ponzano, B., C. Zorzetti, A. Strazzulli, S. Carillo, E. Bedini, M. M. Corsaro, D. A. Comfort, R. M. Kelly, M. Rossi and M. Moracci (2011). "A novel alpha-D-galactosynthase from *Thermotoga maritima* converts beta-D-galactopyranosyl azide to alpha-galacto-oligosaccharides." Glycobiology **21**(4): 448-456.
- Comeau, A. M., G. F. Hatfull, H. M. Krisch, D. Lindell, N. H. Mann and D. Prangishvili (2008). "Exploring the prokaryotic virosphere." Research in Microbiology **159**(5): 306-313.
- Contursi, P., R. Cannio, S. Prato, Q. She, M. Rossi and S. Bartolucci (2007). "Transcriptional analysis of the genetic element pSSVx: differential and temporal regulation of gene expression reveals correlation between transcription and replication." Journal of Bacteriology **189**(17): 6339-6350.
- Contursi, P., S. Fusco, D. Limauro and G. Fiorentino (2013). "Host and viral transcriptional regulators in *Sulfolobus*: an overview." Extremophiles : life under extreme conditions.
- Costa, M., M. Sola, G. del Solar, R. Eritja, A. M. Hernandez-Arriaga, M. Espinosa, F. X. Gomis-Ruth and M. Coll (2001). "Plasmid transcriptional repressor CopG oligomerises to render helical superstructures unbound and in complexes with oligonucleotides." Journal of Molecular Biology **310**(2): 403-417.
- Cox, C. J., P. G. Foster, R. P. Hirt, S. R. Harris and T. M. Embley (2008). "The archaeobacterial origin of eukaryotes." Proceedings of the National Academy of Sciences of the United States of America **105**(51): 20356-20361.
- Crooks, G. E., G. Hon, J. M. Chandonia and S. E. Brenner (2004). "WebLogo: a sequence logo generator." Genome Res **14**(6): 1188-1190.
- Cubellis, M. V., C. Rozzo, P. Montecucchi and M. Rossi (1990). "Isolation and sequencing of a new beta-galactosidase-encoding archaeobacterial gene." Gene **94**(1): 89-94.
- Dahlke, I. and M. Thomm (2002). "A *Pyrococcus* homolog of the leucine-responsive regulatory protein, LrpA, inhibits transcription by abrogating RNA polymerase recruitment." Nucleic acids research **30**(3): 701-710.
- de Jong, G., W. Hazeu, P. Bos and G. Kuenen (1997). "Polythionate degradation by tetrathionate hydrolase of *Thiobacillus ferrooxidans*." Microbiology **143**(2): 499-504.
- de Jong, G. A. H., W. Hazeu, P. Bos and J. G. Kuenen (1997). "Polythionate degradation by tetrathionate hydrolase of *Thiobacillus ferrooxidans*." Microbiology **143**(2): 499-504.
- de Koning, B., F. Blombach, S. J. Brouns and J. van der Oost (2010). "Fidelity in archaeal information processing." Archaea (Vancouver, B.C.) **2010**.
- de la Hoz, A. B., F. Pratto, R. Misselwitz, C. Speck, W. Weihofen, K. Welfle, W. Saenger, H. Welfle and J. C. Alonso (2004). "Recognition of DNA by omega protein from the broad-host range *Streptococcus pyogenes* plasmid pSM19035: analysis of binding to operator DNA with one to four heptad repeats." Nucleic acids research **32**(10): 3136-3147.
- De Rosa, M., A. Gambacorta, B. Nicolaus, P. Giardina, E. Poerio and V. Buonocore (1984). "Glucose metabolism in the extreme thermoacidophilic archaeobacterium *Sulfolobus solfataricus*." The Biochemical journal **224**(2): 407-414.

- DeDecker, B. S., R. O'Brien, P. J. Fleming, J. H. Geiger, S. P. Jackson and P. B. Sigler (1996). "The crystal structure of a hyperthermophilic archaeal TATA-box binding protein." Journal of Molecular Biology **264**(5): 1072-1084.
- del Solar, G., A. M. Hernandez-Arriaga, F. X. Gomis-Ruth, M. Coll and M. Espinosa (2002). "A genetically economical family of plasmid-encoded transcriptional repressors involved in control of plasmid copy number." J Bacteriol **184**(18): 4943-4951.
- Deng, L., C. S. Kenchappa, X. Peng, Q. She and R. A. Garrett (2012). "Modulation of CRISPR locus transcription by the repeat-binding protein Cbp1 in *Sulfolobus*." Nucleic acids research **40**(6): 2470-2480.
- Deng, L., H. Zhu, Z. Chen, Y. X. Liang and Q. She (2009). "Unmarked gene deletion and host-vector system for the hyperthermophilic crenarchaeon *Sulfolobus islandicus*." Extremophiles : life under extreme conditions **13**(4): 735-746.
- Deveau, H., J. E. Garneau and S. Moineau (2010). "CRISPR/Cas system and its role in phage-bacteria interactions." Annual Review of Microbiology **64**: 475-493.
- Di Fiore, A., G. Fiorentino, R. M. Vitale, R. Ronca, P. Amodeo, C. Pedone, S. Bartolucci and G. De Simone (2009). "Structural analysis of BldR from *Sulfolobus solfataricus* provides insights into the molecular basis of transcriptional activation in Archaea by MarR family proteins." Journal of Molecular Biology **388**(3): 559-569.
- Dionne, I. and S. D. Bell (2005). "Characterization of an archaeal family 4 uracil DNA glycosylase and its interaction with PCNA and chromatin proteins." The Biochemical journal **387**(Pt 3): 859-863.
- Dubochet, J., M. Ducommun, M. Zollinger and E. Kellenberger (1971). "A new preparation method for dark-field electron microscopy of biomacromolecules." Journal of Ultrastructure Research **35**(1): 147-167.
- Duc, T. N., G. Hassanzadeh-Ghassabeh, D. Saerens, E. Peeters, D. Charlier and S. Muyldermans (2012). "Nanobody-based chromatin immunoprecipitation." Methods in molecular biology (Clifton, N.J.) **911**: 491-505.
- Elferink, M. G., S. V. Albers, W. N. Konings and A. J. Driessen (2001). "Sugar transport in *Sulfolobus solfataricus* is mediated by two families of binding protein-dependent ABC transporters." Molecular Microbiology **39**(6): 1494-1503.
- Elferink, M. G., J. G. de Wit, R. Demel, A. J. Driessen and W. N. Konings (1992). "Functional reconstitution of membrane proteins in monolayer liposomes from bipolar lipids of *Sulfolobus acidocaldarius*." The Journal of biological chemistry **267**(2): 1375-1381.
- Elferink, M. G., C. Schleper and W. Zillig (1996). "Transformation of the extremely thermoacidophilic archaeon *Sulfolobus solfataricus* via a self-spreading vector." FEMS Microbiology Letters **137**(1): 31-35.
- Elkins, J. G., M. Podar, D. E. Graham, K. S. Makarova, Y. Wolf, L. Randau, B. P. Hedlund, C. Brochier-Armanet, V. Kunin, I. Anderson, A. Lapidus, E. Goltsman, K. Barry, E. V. Koonin, P. Hugenholtz, N. Kyrpides, G. Wanner, P. Richardson, M. Keller and K. O. Stetter (2008). "A korarchaeal genome reveals insights into the evolution of the Archaea." Proceedings of the National Academy of Sciences of the United States of America **105**(23): 8102-8107.
- Ellen, A. F., S. V. Albers, W. Huibers, A. Pitcher, C. F. Hobel, H. Schwarz, M. Folea, S. Schouten, E. J. Boekema, B. Poolman and A. J. Driessen (2009). "Proteomic analysis of secreted membrane vesicles of archaeal *Sulfolobus* species reveals the presence of endosome sorting complex components." Extremophiles : life under extreme conditions **13**(1): 67-79.

- Ellen, A. F., B. Zolghadr, A. M. Driessen and S. V. Albers (2010). "Shaping the archaeal cell envelope." Archaea (Vancouver, B.C.) **2010**: 608243.
- Enoru-Eta, J., D. Gigot, N. Glansdorff and D. Charlier (2002). "High resolution contact probing of the Lrp-like DNA-binding protein Ss-Lrp from the hyperthermoacidophilic crenarchaeote *Sulfolobus solfataricus* P2." Molecular Microbiology **45**(6): 1541-1555.
- Enoru-Eta, J., D. Gigot, T. L. Thia-Toong, N. Glansdorff and D. Charlier (2000). "Purification and characterization of Sa-Lrp, a DNA-binding protein from the extreme thermoacidophilic archaeon *Sulfolobus acidocaldarius* homologous to the bacterial global transcriptional regulator Lrp." Journal of Bacteriology **182**(13): 3661-3672.
- Epshtein, V., C. J. Cardinale, A. E. Ruckenstein, S. Borukhov and E. Nudler (2007). "An allosteric path to transcription termination." Molecular Cell **28**(6): 991-1001.
- Erauso, G., K. M. Stedman, H. J. van de Werken, W. Zillig and J. van der Oost (2006). "Two novel conjugative plasmids from a single strain of *Sulfolobus*." Microbiology (Reading, England) **152**(Pt 7): 1951-1968.
- Ettema, T. J. and R. Bernander (2009). "Cell division and the ESCRT complex: A surprise from the archaea." Communicative & integrative biology **2**(2): 86-88.
- Ettema, T. J., A. B. Brinkman, T. H. Tani, J. B. Rafferty and J. Van Der Oost (2002). "A novel ligand-binding domain involved in regulation of amino acid metabolism in prokaryotes." The Journal of biological chemistry **277**(40): 37464-37468.
- Evans, P. (2006). "Scaling and assessment of data quality." Acta Crystallogr D Biol Crystallogr **62**(Pt 1): 72-82.
- Facciotti, M. T., D. J. Reiss, M. Pan, A. Kaur, M. Vuthoori, R. Bonneau, P. Shannon, A. Srivastava, S. M. Donohoe, L. E. Hood and N. S. Baliga (2007). "General transcription factor specified global gene regulation in archaea." Proceedings of the National Academy of Sciences of the United States of America **104**(11): 4630-4635.
- Fiorentino, G., R. Cannio, M. Rossi and S. Bartolucci (2003). "Transcriptional regulation of the gene encoding an alcohol dehydrogenase in the archaeon *Sulfolobus solfataricus* involves multiple factors and control elements." Journal of Bacteriology **185**(13): 3926-3934.
- Fiorentino, G., I. Del Giudice, S. Bartolucci, L. Durante, L. Martino and P. Del Vecchio (2011). "Identification and physicochemical characterization of BldR2 from *Sulfolobus solfataricus*, a novel archaeal member of the MarR transcription factor family." Biochemistry **50**(31): 6607-6621.
- Fiorentino, G., R. Ronca, R. Cannio, M. Rossi and S. Bartolucci (2007). "MarR-like transcriptional regulator involved in detoxification of aromatic compounds in *Sulfolobus solfataricus*." Journal of Bacteriology **189**(20): 7351-7360.
- Forterre, P. (2002). "A hot story from comparative genomics: reverse gyrase is the only hyperthermophile-specific protein." Trends in genetics : TIG **18**(5): 236-237.
- Forterre, P. (2006). "The origin of viruses and their possible roles in major evolutionary transitions." Virus Research **117**(1): 5-16.
- Forterre, P., G. Mirambeau, C. Jaxel, M. Nadal and M. Duguet (1985). "High positive supercoiling in vitro catalyzed by an ATP and polyethylene glycol-stimulated topoisomerase from *Sulfolobus acidocaldarius*." The EMBO journal **4**(8): 2123-2128.
- Forterre, P. and D. Prangishvili (2009). "The origin of viruses." Research in Microbiology **160**(7): 466-472.

- Forterre, P. and D. Prangishvili (2013). "The major role of viruses in cellular evolution: facts and hypotheses." Current opinion in virology.
- French, S. L., T. J. Santangelo, A. L. Beyer and J. N. Reeve (2007). "Transcription and translation are coupled in Archaea." Molecular Biology and Evolution **24**(4): 893-895.
- Frols, S., M. Ajon, M. Wagner, D. Teichmann, B. Zolghadr, M. Folea, E. J. Boekema, A. J. Driessen, C. Schleper and S. V. Albers (2008). "UV-inducible cellular aggregation of the hyperthermophilic archaeon *Sulfolobus solfataricus* is mediated by pili formation." Molecular Microbiology **70**(4): 938-952.
- Frols, S., P. M. Gordon, M. A. Panlilio, I. G. Duggin, S. D. Bell, C. W. Sensen and C. Schleper (2007). "Response of the hyperthermophilic archaeon *Sulfolobus solfataricus* to UV damage." Journal of Bacteriology **189**(23): 8708-8718.
- Fu, C. Y. and J. E. Johnson (2012). "Structure and cell biology of archaeal virus STIV." Current opinion in virology **2**(2): 122-127.
- Fujikane, R., S. Ishino, Y. Ishino and P. Forterre (2010). "Genetic analysis of DNA repair in the hyperthermophilic archaeon, *Thermococcus kodakaraensis*." Genes & Genetic Systems **85**(4): 243-257.
- Gabrielsen, O. S., E. Hornes, L. Korsnes, A. Ruet and T. B. Oyen (1989). "Magnetic DNA affinity purification of yeast transcription factor tau--a new purification principle for the ultrarapid isolation of near homogeneous factor." Nucleic acids research **17**(15): 6253-6267.
- Gajiwala, K. S. and S. K. Burley (2000). "Winged helix proteins." Current Opinion in Structural Biology **10**(1): 110-116.
- Gallo, M., E. Ferrari, T. Eliseo, I. Amata, T. Pertinhez, A. Katsuyama, M. Paci, C. Farah, A. Spisni and D. Cicero (2010). "A new member of the ribbon-helix-helix transcription factor superfamily from the plant pathogen *Xanthomonas axonopodis* pv. citri." Journal of structural biology **170**(1): 21-31.
- Gamsjaeger, R., C. K. Liew, F. E. Loughlin, M. Crossley and J. P. Mackay (2007). "Sticky fingers: zinc-fingers as protein-recognition motifs." Trends in Biochemical Sciences **32**(2): 63-70.
- Garnier, F. and M. Nadal (2008). "Transcriptional analysis of the two reverse gyrase encoding genes of *Sulfolobus solfataricus* P2 in relation to the growth phases and temperature conditions." Extremophiles : life under extreme conditions **12**(6): 799-809.
- Geiduschek, E. P. and M. Ouhammouch (2005). "Archaeal transcription and its regulators." Molecular Microbiology **56**(6): 1397-1407.
- Gerdes, K., J. Moller-Jensen and R. Bugge Jensen (2000). "Plasmid and chromosome partitioning: surprises from phylogeny." Molecular Microbiology **37**(3): 455-466.
- Gernhardt, P., O. Possot, M. Foglino, L. Sibold and A. Klein (1990). "Construction of an integration vector for use in the archaeobacterium *Methanococcus voltae* and expression of a eubacterial resistance gene." Molecular & general genetics : MGG **221**(2): 273-279.
- Geslin, C., M. Gaillard, D. Flament, K. Rouault, M. Le Romancer, D. Prieur and G. Erauso (2007). "Analysis of the first genome of a hyperthermophilic marine virus-like particle, PAV1, isolated from *Pyrococcus abyssi*." J Bacteriol **189**(12): 4510-4519.
- Gill, S. R., M. Pop, R. T. Deboy, P. B. Eckburg, P. J. Turnbaugh, B. S. Samuel, J. I. Gordon, D. A. Relman, C. M. Fraser-Liggett and K. E. Nelson (2006). "Metagenomic analysis of the human distal gut microbiome." Science (New York, N.Y.) **312**(5778): 1355-1359.

- Gilmour, M. W., J. E. Gunton, T. D. Lawley and D. E. Taylor (2003). "Interaction between the IncHI1 plasmid R27 coupling protein and type IV secretion system: TraG associates with the coiled-coil mating pair formation protein TrhB." *Molecular Microbiology* **49**(1): 105-116.
- Golovanov, A. P., D. Barilla, M. Golovanova, F. Hayes and L. Y. Lian (2003). "ParG, a protein required for active partition of bacterial plasmids, has a dimeric ribbon-helix-helix structure." *Mol Microbiol* **50**(4): 1141-1153.
- Gomis-Ruth, F. X., M. Sola, P. Acebo, A. Parraga, A. Guasch, R. Eritja, A. Gonzalez, M. Espinosa, G. del Solar and M. Coll (1998). "The structure of plasmid-encoded transcriptional repressor CopG unliganded and bound to its operator." *The EMBO journal* **17**(24): 7404-7415.
- Gordon, J., S. Xiao, B. Hughes, 3rd, D. M. Su, S. P. Navarre, B. G. Condie and N. R. Manley (2007). "Specific expression of lacZ and cre recombinase in fetal thymic epithelial cells by multiplex gene targeting at the Foxn1 locus." *BMC Developmental Biology* **7**: 69.
- Goulet, A., M. Pina, P. Redder, D. Prangishvili, L. Vera, J. Lichiere, N. Leulliot, H. van Tilbeurgh, M. Ortiz-Lombardia, V. Campanacci and C. Cambillau (2010). "ORF157 from the archaeal virus Acidianus filamentous virus 1 defines a new class of nuclease." *Journal of Virology* **84**(10): 5025-5031.
- Greve, B., S. Jensen, K. Brugger, W. Zillig and R. A. Garrett (2004). "Genomic comparison of archaeal conjugative plasmids from Sulfolobus." *Archaea (Vancouver, B.C.)* **1**(4): 231-239.
- Greve, B., S. Jensen, H. Phan, K. Brugger, W. Zillig, Q. She and R. A. Garrett (2005). "Novel RepA-MCM proteins encoded in plasmids pTAU4, pORA1 and pTIK4 from Sulfolobus neozealandicus." *Archaea (Vancouver, B.C.)* **1**(5): 319-325.
- Gribaldo, S. and C. Brochier-Armanet (2006). "The origin and evolution of Archaea: a state of the art." *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **361**(1470): 1007-1022.
- Gribaldo, S. and C. Brochier-Armanet (2012). "Time for order in microbial systematics." *Trends in Microbiology* **20**(5): 209-210.
- Grogan, D. W. (1989). "Phenotypic characterization of the archaeobacterial genus Sulfolobus: comparison of five wild-type strains." *Journal of Bacteriology* **171**(12): 6710-6719.
- Grogan, D. W. (1991). "Selectable mutant phenotypes of the extremely thermophilic archaeobacterium Sulfolobus acidocaldarius." *Journal of Bacteriology* **173**(23): 7725-7727.
- Grogan, D. W. (1996). "Exchange of genetic markers at extremely high temperatures in the archaeon Sulfolobus acidocaldarius." *Journal of Bacteriology* **178**(11): 3207-3211.
- Grogan, D. W. (2003). "Cytosine methylation by the Sual restriction-modification system: implications for genetic fidelity in a hyperthermophilic archaeon." *Journal of Bacteriology* **185**(15): 4657-4661.
- Grogan, D. W., G. T. Carver and J. W. Drake (2001). "Genetic fidelity under harsh conditions: analysis of spontaneous mutation in the thermoacidophilic archaeon Sulfolobus acidocaldarius." *Proceedings of the National Academy of Sciences of the United States of America* **98**(14): 7928-7933.
- Grogan, D. W. and R. P. Gunsalus (1993). "Sulfolobus acidocaldarius synthesizes UMP via a standard de novo pathway: results of biochemical-genetic study." *Journal of Bacteriology* **175**(5): 1500-1507.
- Grogan, D. W. and J. E. Hansen (2003). "Molecular characteristics of spontaneous deletions in the hyperthermophilic archaeon Sulfolobus acidocaldarius." *Journal of Bacteriology* **185**(4): 1266-1272.

- Grogan, D. W. and K. R. Stengel (2008). "Recombination of synthetic oligonucleotides with prokaryotic chromosomes: substrate requirements of the Escherichia coli/lambdaRed and Sulfolobus acidocaldarius recombination systems." Molecular Microbiology **69**(5): 1255-1265.
- Grohmann, D., A. Hirtreiter and F. Werner (2009). "Molecular mechanisms of archaeal RNA polymerase." Biochemical Society Transactions **37**(Pt 1): 12-17.
- Grohmann, D., A. Hirtreiter and F. Werner (2009). "RNAP subunits F/E (RPB4/7) are stably associated with archaeal RNA polymerase: using fluorescence anisotropy to monitor RNAP assembly in vitro." The Biochemical journal **421**(3): 339-343.
- Grohmann, D., J. Nagy, A. Chakraborty, D. Klose, D. Fielden, R. H. Ebricht, J. Michaelis and F. Werner (2011). "The initiation factor TFE and the elongation factor Spt4/5 compete for the RNAP clamp during transcription initiation and elongation." Molecular Cell **43**(2): 263-274.
- Grohmann, D. and F. Werner (2010). "Hold on!: RNA polymerase interactions with the nascent RNA modulate transcription elongation and termination." RNA biology **7**(3): 310-315.
- Grohmann, D. and F. Werner (2011). "Cycling through transcription with the RNA polymerase F/E (RPB4/7) complex: structure, function and evolution of archaeal RNA polymerase." Research in Microbiology **162**(1): 10-18.
- Grohmann, D. and F. Werner (2011). "Recent advances in the understanding of archaeal transcription." Current Opinion in Microbiology **14**(3): 328-334.
- Grosse-Kunstleve, R. W. and P. D. Adams (2003). "Substructure search procedures for macromolecular structures." Acta Crystallogr D Biol Crystallogr **59**(Pt 11): 1966-1973.
- Grunberg, S., C. Reich, M. E. Zeller, M. S. Bartlett and M. Thomm (2010). "Rearrangement of the RNA polymerase subunit H and the lower jaw in archaeal elongation complexes." Nucleic acids research **38**(6): 1950-1963.
- Guilliere, F., C. Danioux, C. Jaubert, N. Desnoues, M. Delepierre, D. Prangishvili, G. Sezonov and J. I. Guijarro (2013). "Solution structure of an archaeal DNA binding protein with an eukaryotic zinc finger fold." PloS one **8**(1): e52908.
- Guilliere, F., N. Peixeiro, A. Kessler, B. Raynal, N. Desnoues, J. Keller, M. Delepierre, D. Prangishvili, G. Sezonov and J. I. Guijarro (2009). "Structure, function, and targets of the transcriptional regulator SvtR from the hyperthermophilic archaeal virus SIRV1." The Journal of biological chemistry **284**(33): 22222-22237.
- Guo, L., K. Brugger, C. Liu, S. A. Shah, H. Zheng, Y. Zhu, S. Wang, R. K. Liljestol, L. Chen, J. Frank, D. Prangishvili, L. Paulin, Q. She, L. Huang and R. A. Garrett (2011). "Genome analyses of Icelandic strains of Sulfolobus islandicus, model organisms for genetic and virus-host interaction studies." Journal of Bacteriology **193**(7): 1672-1680.
- Guy, C. P., S. Haldenby, A. Brindley, D. A. Walsh, G. S. Briggs, M. J. Warren, T. Allers and E. L. Bolt (2006). "Interactions of RadB, a DNA repair protein in archaea, with DNA and ATP." Journal of Molecular Biology **358**(1): 46-56.
- Hanzelka, B. L., T. J. Darcy and J. N. Reeve (2001). "TFE, an archaeal transcription factor in Methanobacterium thermoautotrophicum related to eucaryal transcription factor TFIIIEalpha." Journal of Bacteriology **183**(5): 1813-1818.
- Haring, M., R. Rachel, X. Peng, R. A. Garrett and D. Prangishvili (2005). "Viral diversity in hot springs of Pozzuoli, Italy, and characterization of a unique archaeal virus, Acidianus bottle-shaped virus, from a new family, the Ampullaviridae." Journal of Virology **79**(15): 9904-9911.

- Harris, R. L. and P. M. Silverman (2004). "Tra proteins characteristic of F-like type IV secretion systems constitute an interaction group by yeast two-hybrid analysis." Journal of Bacteriology **186**(16): 5480-5485.
- Hartman, A. L., C. Norais, J. H. Badger, S. Delmas, S. Haldenby, R. Madupu, J. Robinson, H. Khouri, Q. Ren, T. M. Lowe, J. Maupin-Furlow, M. Pohlschroder, C. Daniels, F. Pfeiffer, T. Allers and J. A. Eisen (2010). "The complete genome sequence of Haloferax volcanii DS2, a model archaeon." PloS one **5**(3): e9605.
- Hasenohrl, D., T. Lombo, V. Kaberdin, P. Londei and U. Blasi (2008). "Translation initiation factor a/eIF2(-gamma) counteracts 5' to 3' mRNA decay in the archaeon Sulfolobus solfataricus." Proceedings of the National Academy of Sciences of the United States of America **105**(6): 2146-2150.
- Hausner, W., U. Lange and M. Musfeldt (2000). "Transcription factor S, a cleavage induction factor of the archaeal RNA polymerase." The Journal of biological chemistry **275**(17): 12393-12399.
- Haydock, A. K., I. Porat, W. B. Whitman and J. A. Leigh (2004). "Continuous culture of Methanococcus maripaludis under defined nutrient conditions." FEMS Microbiology Letters **238**(1): 85-91.
- Hayward, I., L. H. Malcoe, L. A. Cleathero, P. A. Janssen, B. P. Lanphear, M. V. Hayes, A. Mattman, R. Pampalon and S. A. Venners (2012). "Investigating maternal risk factors as potential targets of intervention to reduce socioeconomic inequality in small for gestational age: a population-based study." BMC Public Health **12**: 333.
- He, Y. Y., T. McNally, I. Manfield, O. Navratil, I. G. Old, S. E. Phillips, I. Saint-Girons and P. G. Stockley (1992). "Probing met repressor-operator recognition in solution." Nature **359**(6394): 431-433.
- Held, N. L. and R. J. Whitaker (2009). "Viral biogeography revealed by signatures in Sulfolobus islandicus genomes." Environmental Microbiology **11**(2): 457-466.
- Henche, A. L., A. Ghosh, X. Yu, T. Jeske, E. Egelman and S. V. Albers (2012). "Structure and function of the adhesive type IV pilus of Sulfolobus acidocaldarius." Environmental Microbiology **14**(12): 3188-3202.
- Hendrickson, E. L., R. Kaul, Y. Zhou, D. Bovee, P. Chapman, J. Chung, E. Conway de Macario, J. A. Dodsworth, W. Gillett, D. E. Graham, M. Hackett, A. K. Haydock, A. Kang, M. L. Land, R. Levy, T. J. Lie, T. A. Major, B. C. Moore, I. Porat, A. Palmeiri, G. Rouse, C. Saenphimmachak, D. Soll, S. Van Dien, T. Wang, W. B. Whitman, Q. Xia, Y. Zhang, F. W. Larimer, M. V. Olson and J. A. Leigh (2004). "Complete genome sequence of the genetically tractable hydrogenotrophic methanogen Methanococcus maripaludis." Journal of Bacteriology **186**(20): 6956-6969.
- Hernandez-Arriaga, A. M., T. S. Rubio-Lepe, M. Espinosa and G. del Solar (2009). "Repressor CopG prevents access of RNA polymerase to promoter and actively dissociates open complexes." Nucleic acids research **37**(14): 4799-4811.
- Hirata, A., B. J. Klein and K. S. Murakami (2008). "The X-ray crystal structure of RNA polymerase from Archaea." Nature **451**(7180): 851-854.
- Hirata, A. and K. S. Murakami (2009). "Archaeal RNA polymerase." Current Opinion in Structural Biology **19**(6): 724-731.
- Hirtreiter, A., D. Grohmann and F. Werner (2010). "Molecular mechanisms of RNA polymerase--the F/E (RPB4/7) complex is required for high processivity in vitro." Nucleic acids research **38**(2): 585-596.

- Hjorleifsdottir, S., S. Skirnisdottir, G. O. Hreggvidsson, O. Holst and J. K. Kristjansson (2001). "Species Composition of Cultivated and Noncultivated Bacteria from Short Filaments in an Icelandic Hot Spring at 88 degrees C." *Microbial Ecology* **42**(2): 117-125.
- Hohn, M. J., B. P. Hedlund and H. Huber (2002). "Detection of 16S rDNA sequences representing the novel phylum "Nanoarchaeota": indication for a wide distribution in high temperature biotopes." *Systematic and Applied Microbiology* **25**(4): 551-554.
- Horvath, P. and R. Barrangou (2010). "CRISPR/Cas, the immune system of bacteria and archaea." *Science (New York, N.Y.)* **327**(5962): 167-170.
- Huang, L., P. Yin, X. Zhu, Y. Zhang and K. Ye (2011). "Crystal structure and centromere binding of the plasmid segregation protein ParB from pCXC100." *Nucleic acids research* **39**(7): 2954-2968.
- Huber, H., M. J. Hohn, R. Rachel, T. Fuchs, V. C. Wimmer and K. O. Stetter (2002). "A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont." *Nature* **417**(6884): 63-67.
- Huffman, J. L. and R. G. Brennan (2002). "Prokaryotic transcription regulators: more than just the helix-turn-helix motif." *Current Opinion in Structural Biology* **12**(1): 98-106.
- Hutcheon, G. W. and A. Bolhuis (2003). "The archaeal twin-arginine translocation pathway." *Biochemical Society Transactions* **31**(Pt 3): 686-689.
- Iverson, E. and K. Stedman (2012). "A genetic study of SSV1, the prototypical fusellovirus." *Frontiers in microbiology* **3**: 200.
- Jaubert, C., C. Danioux, J. Oberto, D. Cortez, A. Bize, M. Krupovic, Q. She, P. Forterre, D. Prangishvili and G. Sezonov (2013). "Genomics and genetics of *Sulfolobus islandicus* LAL14/1, a model hyperthermophilic archaeon." *Open biology* **3**(4): 130010.
- Jia, Z., J. Weng, X. Lin and R. Conrad (2010). "[Microbial ecology of archaeal ammonia oxidation--a review]." *Wei sheng wu xue bao = Acta microbiologica Sinica* **50**(4): 431-437.
- Jolma, A., T. Kivioja, J. Toivonen, L. Cheng, G. Wei, M. Enge, M. Taipale, J. M. Vaquerizas, J. Yan, M. J. Sillanpaa, M. Bonke, K. Palin, S. Talukder, T. R. Hughes, N. M. Luscombe, E. Ukkonen and J. Taipale (2010). "Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities." *Genome Res* **20**(6): 861-873.
- Jones, T. A. and M. Kjeldgaard (1991). "Improved methods for building protein models in electron density maps and the location of errors in these models. ." *Acta Cryst A* **47**: 110-119
- Jonuscheit, M., E. Martusewitsch, K. M. Stedman and C. Schleper (2003). "A reporter gene system for the hyperthermophilic archaeon *Sulfolobus solfataricus* based on a selectable and integrative shuttle vector." *Molecular Microbiology* **48**(5): 1241-1252.
- Jund, R. and F. Lacroute (1970). "Genetic and physiological aspects of resistance to 5-fluoropyrimidines in *Saccharomyces cerevisiae*." *Journal of Bacteriology* **102**(3): 607-615.
- Kaberdin, V. R., D. Singh and S. Lin-Chao (2011). "Composition and conservation of the mRNA-degrading machinery in bacteria." *Journal of Biomedical Science* **18**: 23.
- Kanai, T., J. Akerboom, S. Takedomi, H. J. van de Werken, F. Blombach, J. van der Oost, T. Murakami, H. Atomi and T. Imanaka (2007). "A global transcriptional regulator in *Thermococcus kodakaraensis* controls the expression levels of both glycolytic and gluconeogenic enzyme-encoding genes." *The Journal of biological chemistry* **282**(46): 33659-33670.
- Kandler, O. and H. König (1978). "Chemical composition of the peptidoglycan-free cell walls of methanogenic bacteria." *Archives of Microbiology* **118**(2): 141-152.
- Kandler, O. and H. König (1998). "Cell wall polymers in Archaea (Archaeobacteria)." *Cellular and molecular life sciences : CMLS* **54**(4): 305-308.

- Kapanidis, A. N., E. Margeat, S. O. Ho, E. Kortkhonjia, S. Weiss and R. H. Ebright (2006). "Initial transcription by RNA polymerase proceeds through a DNA-scrunching mechanism." Science (New York, N.Y.) **314**(5802): 1144-1147.
- Karr, E. A. (2010). "The methanogen-specific transcription factor MsvR regulates the fpaA-rlp-rub oxidative stress operon adjacent to msvR in *Methanothermobacter thermautotrophicus*." Journal of Bacteriology **192**(22): 5914-5922.
- Karr, E. A., K. Sandman, R. Lurz and J. N. Reeve (2008). "TrpY regulation of trpB2 transcription in *Methanothermobacter thermautotrophicus*." Journal of Bacteriology **190**(7): 2637-2641.
- Kawarabayasi, Y., Y. Hino, H. Horikawa, K. Jin-no, M. Takahashi, M. Sekine, S. Baba, A. Ankai, H. Kosugi, A. Hosoyama, S. Fukui, Y. Nagai, K. Nishijima, R. Otsuka, H. Nakazawa, M. Takamiya, Y. Kato, T. Yoshizawa, T. Tanaka, Y. Kudoh, J. Yamazaki, N. Kushida, A. Oguchi, K. Aoki, S. Masuda, M. Yanagii, M. Nishimura, A. Yamagishi, T. Oshima and H. Kikuchi (2001). "Complete genome sequence of an aerobic thermoacidophilic crenarchaeon, *Sulfolobus tokodaii* strain 7." DNA research : an international journal for rapid publication of reports on genes and genomes **8**(4): 123-140.
- Kawashima, T., H. Aramaki, T. Oyamada, K. Makino, M. Yamada, H. Okamura, K. Yokoyama, S. A. Ishijima and M. Suzuki (2008). "Transcription regulation by feast/famine regulatory proteins, FFRPs, in archaea and eubacteria." Biological & Pharmaceutical Bulletin **31**(2): 173-186.
- Keeling, P. J., H. P. Klenk, R. K. Singh, O. Feeley, C. Schleper, W. Zillig, W. F. Doolittle and C. W. Sensen (1996). "Complete nucleotide sequence of the *Sulfolobus islandicus* multicopy plasmid pRN1." Plasmid **35**(2): 141-144.
- Keeling, P. J., H. P. Klenk, R. K. Singh, M. E. Schenk, C. W. Sensen, W. Zillig and W. F. Doolittle (1998). "Sulfolobus islandicus plasmids pRN1 and pRN2 share distant but common evolutionary ancestry." Extremophiles : life under extreme conditions **2**(4): 391-393.
- Keller, J., N. Leulliot, C. Cambillau, V. Campanacci, S. Porciero, D. Prangishvili, P. Forterre, D. Cortez, S. Quevillon-Cheruel and H. van Tilbeurgh (2007). "Crystal structure of AFV3-109, a highly conserved protein from crenarchaeal viruses." Virology journal **4**: 12.
- Kelman, Z., J. K. Lee and J. Hurwitz (1999). "The single minichromosome maintenance protein of *Methanobacterium thermoautotrophicum* DeltaH contains DNA helicase activity." Proceedings of the National Academy of Sciences of the United States of America **96**(26): 14783-14788.
- Kelman, Z. and M. F. White (2005). "Archaeal DNA replication and repair." Current Opinion in Microbiology **8**(6): 669-676.
- Kessler, A., A. B. Brinkman, J. van der Oost and D. Prangishvili (2004). "Transcription of the rod-shaped viruses SIRV1 and SIRV2 of the hyperthermophilic archaeon *sulfolobus*." J Bacteriol **186**(22): 7745-7753.
- Kessler, A., G. Sezonov, J. I. Guijarro, N. Desnoves, T. Rose, M. Delepierre, S. D. Bell and D. Prangishvili (2006). "A novel archaeal regulatory protein, Sta1, activates transcription from viral promoters." Nucleic acids research **34**(17): 4837-4845.
- Kettenberger, H., K. J. Armache and P. Cramer (2003). "Architecture of the RNA polymerase II-TFIIS complex and implications for mRNA cleavage." Cell **114**(3): 347-357.
- Kikuchi, A. and K. Asai (1984). "Reverse gyrase--a topoisomerase which introduces positive superhelical turns into DNA." Nature **309**(5970): 677-681.

- Klein, R., N. Rossler, M. Iro, H. Scholz and A. Witte (2012). "Haloarchaeal myovirus phiCh1 harbours a phase variation system for the production of protein variants with distinct cell surface adhesion specificities." Molecular Microbiology **83**(1): 137-150.
- Knight, K. L., J. U. Bowie, A. K. Vershon, R. D. Kelley and R. T. Sauer (1989). "The Arc and Mnt repressors. A new class of sequence-specific DNA-binding protein." The Journal of biological chemistry **264**(7): 3639-3642.
- Koerdt, A., S. Jachlewski, A. Ghosh, J. Wingender, B. Siebers and S. V. Albers (2012). "Complementation of *Sulfolobus solfataricus* PBL2025 with an alpha-mannosidase: effects on surface attachment and biofilm formation." Extremophiles : life under extreme conditions **16**(1): 115-125.
- Koonin, E. V., K. S. Makarova and J. G. Elkins (2007). "Orthologs of the small RPB8 subunit of the eukaryotic RNA polymerases are conserved in hyperthermophilic Crenarchaeota and "Korarchaeota". " Biology direct **2**: 38.
- Korkhin, Y., U. M. Unligil, O. Littlefield, P. J. Nelson, D. I. Stuart, P. B. Sigler, S. D. Bell and N. G. Abrescia (2009). "Evolution of complex RNA polymerases: the complete archaeal RNA polymerase structure." PLoS Biology **7**(5): e1000102.
- Kostrewa, D., M. E. Zeller, K. J. Armache, M. Seizl, K. Leike, M. Thomm and P. Cramer (2009). "RNA polymerase II-TFIIB structure and mechanism of transcription initiation." Nature **462**(7271): 323-330.
- Kouril, T., P. Wieloch, J. Reimann, M. Wagner, M. Zaparty, S. V. Albers, D. Schomburg, P. Ruoff and B. Siebers (2013). "Unraveling the function of the two Entner-Doudoroff branches in the thermoacidophilic Crenarchaeon *Sulfolobus solfataricus* P2." The FEBS journal **280**(4): 1126-1138.
- Krupovic, M., M. F. White, P. Forterre and D. Prangishvili (2012). "Postcards from the edge: structural genomics of archaeal viruses." Advances in Virus Research **82**: 33-62.
- Kumarevel, T., N. Nakano, K. Ponnuraj, S. C. Gopinath, K. Sakamoto, A. Shinkai, P. K. Kumar and S. Yokoyama (2008). "Crystal structure of glutamine receptor protein from *Sulfolobus tokodaii* strain 7 in complex with its effector L-glutamine: implications of effector binding in molecular association and DNA binding." Nucleic acids research **36**(14): 4808-4820.
- Kumarevel, T., T. Tanaka, T. Umehara and S. Yokoyama (2009). "ST1710-DNA complex crystal structure reveals the DNA binding mechanism of the MarR family of regulators." Nucleic acids research **37**(14): 4723-4735.
- Kurosawa, N. and D. W. Grogan (2005). "Homologous recombination of exogenous DNA with the *Sulfolobus acidocaldarius* genome: properties and uses." FEMS Microbiology Letters **253**(1): 141-149.
- Kusser, A. G., M. G. Bertero, S. Naji, T. Becker, M. Thomm, R. Beckmann and P. Cramer (2008). "Structure of an archaeal RNA polymerase." Journal of Molecular Biology **376**(2): 303-307.
- Kwapisz, M., F. Beckouet and P. Thuriaux (2008). "Early evolution of eukaryotic DNA-dependent RNA polymerases." Trends in genetics : TIG **24**(5): 211-215.
- Kyrpides, N. C. and C. A. Ouzounis (1999). "Transcription in archaea." Proceedings of the National Academy of Sciences of the United States of America **96**(15): 8545-8550.
- La Teana, A., D. Benelli, P. Londei and U. Blasi (2013). "Translation initiation in the crenarchaeon *Sulfolobus solfataricus*: eukaryotic features but bacterial route." Biochemical Society Transactions **41**(1): 350-355.

- Lalaouna, D., M. Simoneau-Roy, D. Lafontaine and E. Masse (2013). "Regulatory RNAs and target mRNA decay in prokaryotes." *Biochimica et Biophysica Acta* **1829**(6-7): 742-747.
- Lam, W. L. and W. F. Doolittle (1992). "Mevinolin-resistant mutations identify a promoter and the gene for a eukaryote-like 3-hydroxy-3-methylglutaryl-coenzyme A reductase in the archaeobacterium *Haloferax volcanii*." *The Journal of biological chemistry* **267**(9): 5829-5834.
- Lange, U. and W. Hausner (2004). "Transcriptional fidelity and proofreading in Archaea and implications for the mechanism of TFS-induced RNA cleavage." *Molecular Microbiology* **52**(4): 1133-1143.
- Larson, E. T., B. J. Eilers, D. Reiter, A. C. Ortmann, M. J. Young and C. M. Lawrence (2007). "A new DNA binding protein highly conserved in diverse crenarchaeal viruses." *Virology* **363**(2): 387-396.
- Larson, J. D., J. L. Jenkins, J. P. Schuermann, Y. Zhou, D. F. Becker and J. J. Tanner (2006). "Crystal structures of the DNA-binding domain of *Escherichia coli* proline utilization A flavoprotein and analysis of the role of Lys9 in DNA recognition." *Protein Sci* **15**(11): 2630-2641.
- Lassak, K., T. Neiner, A. Ghosh, A. Klingl, R. Wirth and S. V. Albers (2012). "Molecular analysis of the crenarchaeal flagellum." *Molecular Microbiology* **83**(1): 110-124.
- Lawley, T. D., M. W. Gilmour, J. E. Gunton, D. M. Tracz and D. E. Taylor (2003). "Functional and mutational analysis of conjugative transfer region 2 (Tra2) from the IncHI1 plasmid R27." *Journal of Bacteriology* **185**(2): 581-591.
- Lawley, T. D., W. A. Klimke, M. J. Gubbins and L. S. Frost (2003). "F factor conjugation is a true type IV secretion system." *FEMS Microbiology Letters* **224**(1): 1-15.
- Lawley, T. D. and D. E. Taylor (2003). "Characterization of the double-partitioning modules of R27: correlating plasmid stability with plasmid localization." *Journal of Bacteriology* **185**(10): 3060-3067.
- Lawson, F. S., R. L. Charlebois and J. A. Dillon (1996). "Phylogenetic analysis of carbamoylphosphate synthetase genes: complex evolutionary history includes an internal duplication within a gene which can root the tree of life." *Molecular Biology and Evolution* **13**(7): 970-977.
- Lee, S. J., A. Engelmann, R. Horlacher, Q. Qu, G. Vierke, C. Hebbeln, M. Thomm and W. Boos (2003). "TrmB, a sugar-specific transcriptional regulator of the trehalose/maltose ABC transporter from the hyperthermophilic archaeon *Thermococcus litoralis*." *The Journal of biological chemistry* **278**(2): 983-990.
- Lee, S. J., C. Moulakakis, S. M. Koning, W. Hausner, M. Thomm and W. Boos (2005). "TrmB, a sugar sensing regulator of ABC transporter genes in *Pyrococcus furiosus* exhibits dual promoter specificity and is controlled by different inducers." *Molecular Microbiology* **57**(6): 1797-1807.
- Lee, S. J., M. Surma, W. Hausner, M. Thomm and W. Boos (2008). "The role of TrmB and TrmB-like transcriptional regulators for sugar transport and metabolism in the hyperthermophilic archaeon *Pyrococcus furiosus*." *Archives of Microbiology* **190**(3): 247-256.
- Lee, S. J., M. Surma, S. Seitz, W. Hausner, M. Thomm and W. Boos (2007). "Differential signal transduction via TrmB, a sugar sensing transcriptional repressor of *Pyrococcus furiosus*." *Molecular Microbiology* **64**(6): 1499-1505.
- Leigh, J. A., S. V. Albers, H. Atomi and T. Allers (2011). "Model organisms for genetics in the domain Archaea: methanogens, halophiles, Thermococcales and Sulfolobales." *FEMS Microbiology Reviews* **35**(4): 577-608.

- Leininger, S., T. Urich, M. Schloter, L. Schwark, J. Qi, G. W. Nicol, J. I. Prosser, S. C. Schuster and C. Schleper (2006). "Archaea predominate among ammonia-oxidizing prokaryotes in soils." Nature **442**(7104): 806-809.
- Leslie, A. G. W. (1992). Joint CCP4 and EACMB Newsletter Protein Crystallography, Daresbury Laboratory, Warrington, UK. **26**.
- Li, G.-Y., Y. Zhang, M. Inouye and M. Ikura (2008). "Structural Mechanism of Transcriptional Autorepression of the Escherichia coli RelB/RelE Antitoxin/Toxin Module." Journal of Molecular Biology **380**(1): 107-119.
- Li, X., L. Guo, L. Deng, D. Feng, Y. Ren, Y. Chu, Q. She and L. Huang (2011). "Deletion of the topoisomerase III gene in the hyperthermophilic archaeon Sulfolobus islandicus results in slow growth and defects in cell cycle control." Journal of genetics and genomics = Yi chuan xue bao **38**(6): 253-259.
- Lillestol, R. K., S. A. Shah, K. Brugger, P. Redder, H. Phan, J. Christiansen and R. A. Garrett (2009). "CRISPR families of the crenarchaeal genus Sulfolobus: bidirectional transcription and dynamic properties." Molecular Microbiology **72**(1): 259-272.
- Lindas, A. C., E. A. Karlsson, M. T. Lindgren, T. J. Ettema and R. Bernander (2008). "A unique cell division machinery in the Archaea." Proceedings of the National Academy of Sciences of the United States of America **105**(48): 18942-18946.
- Lintner, N. G., M. Kerou, S. K. Brumfield, S. Graham, H. Liu, J. H. Naismith, M. Sdano, N. Peng, Q. She, V. Copie, M. J. Young, M. F. White and C. M. Lawrence (2011). "Structural and functional characterization of an archaeal clustered regularly interspaced short palindromic repeat (CRISPR)-associated complex for antiviral defense (CASCADE)." The Journal of biological chemistry **286**(24): 21643-21656.
- Lipps, G. (2004). "The replication protein of the Sulfolobus islandicus plasmid pRN1." Biochem Soc Trans **32**(Pt 2): 240-244.
- Lipps, G. (2006). "Plasmids and viruses of the thermoacidophilic crenarchaeote Sulfolobus." Extremophiles : life under extreme conditions **10**(1): 17-28.
- Lipps, G. (2009). "Molecular biology of the pRN1 plasmid from Sulfolobus islandicus." Biochemical Society Transactions **37**(Pt 1): 42-45.
- Lipps, G., S. Rother, C. Hart and G. Krauss (2003). "A novel type of replicative enzyme harbouring ATPase, primase and DNA polymerase activity." The EMBO journal **22**(10): 2516-2525.
- Lipps, G., M. Stegert and G. Krauss (2001). "Thermostable and site-specific DNA binding of the gene product ORF56 from the Sulfolobus islandicus plasmid pRN1, a putative archaeal plasmid copy control protein." Nucleic Acids Res **29**(4): 904-913.
- Lipscomb, G. L., A. M. Keese, D. M. Cowart, G. J. Schut, M. Thomm, M. W. Adams and R. A. Scott (2009). "SurR: a transcriptional activator and repressor controlling hydrogen and elemental sulphur metabolism in Pyrococcus furiosus." Molecular Microbiology **71**(2): 332-349.
- Liu, B., L. Wu, T. Liu, Y. Hong, Y. Shen and J. Ni (2009). "A MOFRL family glycerate kinase from the thermophilic crenarchaeon, Sulfolobus tokodaii, with unique enzymatic properties." Biotechnology Letters **31**(12): 1937-1941.
- Liu, J. J., W. X. Wu, Y. Ding, D. Z. Shi and Y. X. Chen (2010). "[Ammonia-oxidizing archaea and their important roles in nitrogen biogeochemical cycling: a review]." Ying yong sheng tai xue bao = The journal of applied ecology / Zhongguo sheng tai xue xue hui, Zhongguo ke xue yuan Shenyang ying yong sheng tai yan jiu suo zhu ban **21**(8): 2154-2160.

- Liu, W., G. Vierke, A. K. Wenke, M. Thomm and R. Ladenstein (2007). "Crystal structure of the archaeal heat shock regulator from *Pyrococcus furiosus*: a molecular chimera representing eukaryal and bacterial features." Journal of Molecular Biology **369**(2): 474-488.
- Liu, X., D. L. Brutlag and J. S. Liu (2001). "BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes." Pac Symp Biocomput: 127-138.
- Londei, P. (2005). "Evolution of translational initiation: new insights from the archaea." FEMS Microbiology Reviews **29**(2): 185-200.
- Lopez-Garcia, P. and P. Forterre (1999). "Control of DNA topology during thermal stress in hyperthermophilic archaea: DNA topoisomerase levels, activities and induced thermotolerance during heat and cold shock in *Sulfolobus*." Molecular Microbiology **33**(4): 766-777.
- Louis, B. G. and P. S. Fitt (1971). "Nucleic acid enzymology of extremely halophilic bacteria. *Halobacterium cutirubrum* deoxyribonucleic acid-dependent ribonucleic acid polymerase." The Biochemical journal **121**(4): 621-627.
- Lu, J., A. den Dulk-Ras, P. Hooykaas and M. Glover (2009). "Agrobacterium tumefaciens VirC2 enhances T-DNA transfer and virulence through its C-terminal ribbon-helix-helix DNA-binding fold." Proceedings of the National Academy of Sciences of the United States of America **106**(24): 9643-9648.
- Lubelska, J. M., M. Jonuscheit, C. Schleper, S. V. Albers and A. J. Driessen (2006). "Regulation of expression of the arabinose and glucose transporter genes in the thermophilic archaeon *Sulfolobus solfataricus*." Extremophiles : life under extreme conditions **10**(5): 383-391.
- Lucas, S., L. Toffin, Y. Zivanovic, D. Charlier, H. Moussard, P. Forterre, D. Prieur and G. Erauso (2002). "Construction of a shuttle vector for, and spheroplast transformation of, the hyperthermophilic archaeon *Pyrococcus abyssi*." Applied and Environmental Microbiology **68**(11): 5528-5536.
- Lum, P. L., M. E. Rodgers and J. F. Schildbach (2002). "TraY DNA recognition of its two F factor binding sites." J Mol Biol **321**(4): 563-578.
- Maaty, W. S., B. Wiedenheft, P. Tarlykov, N. Schaff, J. Heinemann, J. Robison-Cox, J. Valenzuela, A. Dougherty, P. Blum, C. M. Lawrence, T. Douglas, M. J. Young and B. Bothner (2009). "Something old, something new, something borrowed; how the thermoacidophilic archaeon *Sulfolobus solfataricus* responds to oxidative stress." PloS one **4**(9): e6964.
- Madl, T., L. Van Melderen, N. Mine, M. Respondek, M. Oberer, W. Keller, L. Khatai and K. Zangger (2006). "Structural basis for nucleic acid and toxin recognition of the bacterial antitoxin CcdA." J Mol Biol **364**(2): 170-185.
- Madon, J. and W. Zillig (1983). "A form of the DNA-dependent RNA polymerase of *Halobacterium halobium*, containing an additional component, is able to transcribe native DNA." European journal of biochemistry / FEBS **133**(2): 471-474.
- Magrum, L. J., K. R. Luehrsen and C. R. Woese (1978). "Are extreme halophiles actually "bacteria"?" Journal of Molecular Evolution **11**(1): 1-8.
- Makarova, K. S., D. H. Haft, R. Barrangou, S. J. Brouns, E. Charpentier, P. Horvath, S. Moineau, F. J. Mojica, Y. I. Wolf, A. F. Yakunin, J. van der Oost and E. V. Koonin (2011). "Evolution and classification of the CRISPR-Cas systems." Nature reviews. Microbiology **9**(6): 467-477.
- Marcussen, T., B. Oxelman, A. Skog and K. S. Jakobsen (2010). "Evolution of plant RNA polymerase IV/V genes: evidence of subneofunctionalization of duplicated NRPD2/NRPE2-like paralogs in *Viola* (Violaceae)." BMC Evolutionary Biology **10**: 45.

- Marraffini, L. A. and E. J. Sontheimer (2008). "CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA." *Science (New York, N.Y.)* **322**(5909): 1843-1845.
- Martin, A., S. Yeats, D. Janekovic, W. D. Reiter, W. Aicher and W. Zillig (1984). "SAV 1, a temperate u.v.-inducible DNA virus-like particle from the archaeobacterium *Sulfolobus acidocaldarius* isolate B12." *The EMBO journal* **3**(9): 2165-2168.
- Martin, F. H. and I. Tinoco, Jr. (1980). "DNA-RNA hybrid duplexes containing oligo(dA:rU) sequences are exceptionally unstable and may facilitate termination of transcription." *Nucleic acids research* **8**(10): 2295-2299.
- Martusewitsch, E., C. W. Sensen and C. Schleper (2000). "High spontaneous mutation rate in the hyperthermophilic archaeon *Sulfolobus solfataricus* is mediated by transposable elements." *Journal of Bacteriology* **182**(9): 2574-2581.
- Matsumi, R., K. Manabe, T. Fukui, H. Atomi and T. Imanaka (2007). "Disruption of a sugar transporter gene cluster in a hyperthermophilic archaeon using a host-marker system based on antibiotic resistance." *Journal of Bacteriology* **189**(7): 2683-2691.
- Mattison, K., J. S. Wilbur, M. So and R. G. Brennan (2006). "Structure of FitAB from *Neisseria gonorrhoeae* bound to DNA reveals a tetramer of toxin-antitoxin heterodimers containing pin domains and ribbon-helix-helix motifs." *J Biol Chem* **281**(49): 37942-37951.
- Mei, Y., N. Peng, S. Zhao, Y. Hu, H. Wang, Y. Liang and Q. She (2012). "Exceptional thermal stability and organic solvent tolerance of an esterase expressed from a thermophilic host." *Applied Microbiology and Biotechnology* **93**(5): 1965-1974.
- Meka, H., G. Daoust, K. B. Arnvig, F. Werner, P. Brick and S. Onesti (2003). "Structural and functional homology between the RNAP(I) subunits A14/A43 and the archaeal RNAP subunits E/F." *Nucleic acids research* **31**(15): 4391-4400.
- Menon, S. K., W. S. Maaty, G. J. Corn, S. C. Kwok, B. J. Eilers, P. Kraft, E. Gillitzer, M. J. Young, B. Bothner and C. M. Lawrence (2008). "Cysteine usage in *Sulfolobus* spindle-shaped virus 1 and extension to hyperthermophilic viruses in general." *Virology* **376**(2): 270-278.
- Meuer, J., H. C. Kuettnner, J. K. Zhang, R. Hedderich and W. W. Metcalf (2002). "Genetic analysis of the archaeon *Methanosarcina barkeri* Fusaro reveals a central role for Ech hydrogenase and ferredoxin in methanogenesis and carbon fixation." *Proceedings of the National Academy of Sciences of the United States of America* **99**(8): 5632-5637.
- Meyer, B. H., B. Zolghadr, E. Peyfoon, M. Pabst, M. Panico, H. R. Morris, S. M. Haslam, P. Messner, C. Schaffer, A. Dell and S. V. Albers (2011). "Sulfoquinovose synthase - an important enzyme in the N-glycosylation pathway of *Sulfolobus acidocaldarius*." *Molecular Microbiology* **82**(5): 1150-1163.
- Micorescu, M., S. Grunberg, A. Franke, P. Cramer, M. Thomm and M. Bartlett (2008). "Archaeal transcription: function of an alternative transcription factor B from *Pyrococcus furiosus*." *Journal of Bacteriology* **190**(1): 157-167.
- Miller, J., A. D. McLachlan and A. Klug (1985). "Repetitive zinc-binding domains in the protein transcription factor IIIA from *Xenopus oocytes*." *The EMBO journal* **4**(6): 1609-1614.
- Miyazono, K., M. Tsujimura, Y. Kawarabayasi and M. Tanokura (2007). "Crystal structure of an archaeal homologue of multidrug resistance repressor protein, EmrR, from hyperthermophilic archaea *Sulfolobus tokodaii* strain 7." *Proteins* **67**(4): 1138-1146.
- Mochizuki, T., M. Krupovic, G. Pehau-Arnaudet, Y. Sako, P. Forterre and D. Prangishvili (2012). "Archaeal virus with exceptional virion architecture and the largest single-stranded DNA

- genome." Proceedings of the National Academy of Sciences of the United States of America **109**(33): 13386-13391.
- Moore, M. J. (2005). "From birth to death: the complex lives of eukaryotic mRNAs." Science (New York, N.Y.) **309**(5740): 1514-1518.
- Moracci, M., A. La Volpe, J. F. Pulitzer, M. Rossi and M. Ciaramella (1992). "Expression of the thermostable beta-galactosidase gene from the archaeobacterium *Sulfolobus solfataricus* in *Saccharomyces cerevisiae* and characterization of a new inducible promoter for heterologous expression." Journal of Bacteriology **174**(3): 873-882.
- Murakami, K. S., S. Masuda and S. A. Darst (2003). "Crystallographic analysis of *Thermus aquaticus* RNA polymerase holoenzyme and a holoenzyme/promoter DNA complex." Methods in Enzymology **370**: 42-53.
- Murayama, K., P. Orth, A. B. de la Hoz, J. C. Alonso and W. Saenger (2001). "Crystal structure of omega transcriptional repressor encoded by *Streptococcus pyogenes* plasmid pSM19035 at 1.5 Å resolution." J Mol Biol **314**(4): 789-796.
- Murshudov, G. N., A. A. Vagin and E. J. Dodson (1997). "Refinement of Macromolecular Structures by the Maximum-Likelihood Method." Acta Crystallogr D Biol Crystallogr **53**: 240-255.
- Muskhelishvili, G., P. Palm and W. Zillig (1993). "SSV1-encoded site-specific recombination system in *Sulfolobus shibatae*." Molecular & general genetics : MGG **237**(3): 334-342.
- Nadal, M. (2007). "Reverse gyrase: an insight into the role of DNA-topoisomerases." Biochimie **89**(4): 447-455.
- Nadal, M., C. Jaxel, C. Portemer, P. Forterre, G. Mirambeau and M. Duguet (1988). "Reverse gyrase of *Sulfolobus*: purification to homogeneity and characterization." Biochemistry **27**(26): 9102-9108.
- Nagarajan, V. K., C. I. Jones, S. F. Newbury and P. J. Green (2013). "XRN 5'→3' exoribonucleases: structure, mechanisms and functions." Biochimica et Biophysica Acta **1829**(6-7): 590-603.
- Naji, S., S. Grunberg and M. Thomm (2007). "The RPB7 orthologue E' is required for transcriptional activity of a reconstituted archaeal core enzyme at low temperatures and stimulates open complex formation." The Journal of biological chemistry **282**(15): 11047-11057.
- Naoe, C., M. Tsunoda and K. T. Nakamura (2004). "Structural study of the archaeal homologous recombinant protein complexed with DNA fragments." Nucleic acids symposium series (2004)(48): 115-116.
- Newbury, S. F. (2006). "Control of mRNA stability in eukaryotes." Biochemical Society Transactions **34**(Pt 1): 30-34.
- Ng, W. V., S. P. Kennedy, G. G. Mahairas, B. Berquist, M. Pan, H. D. Shukla, S. R. Lasky, N. S. Baliga, V. Thorsson, J. Sbrogna, S. Swartzell, D. Weir, J. Hall, T. A. Dahl, R. Welti, Y. A. Goo, B. Leithauser, K. Keller, R. Cruz, M. J. Danson, D. W. Hough, D. G. Maddocks, P. E. Jablonski, M. P. Krebs, C. M. Angevine, H. Dale, T. A. Isenbarger, R. F. Peck, M. Pohlschroder, J. L. Spudich, K. W. Jung, M. Alam, T. Freitas, S. Hou, C. J. Daniels, P. P. Dennis, A. D. Omer, H. Ehardt, T. M. Lowe, P. Liang, M. Riley, L. Hood and S. DasSarma (2000). "Genome sequence of *Halobacterium* species NRC-1." Proceedings of the National Academy of Sciences of the United States of America **97**(22): 12176-12181.
- Nguyen-Duc, T., E. Peeters, S. Muyldermans, D. Charlier and G. Hassanzadeh-Ghassabeh (2013). "Nanobody(R)-based chromatin immunoprecipitation/micro-array analysis for genome-wide identification of transcription factor DNA binding sites." Nucleic acids research **41**(5): e59.

- Nguyen-Duc, T., L. van Oeffelen, N. Song, G. Hassanzadeh-Ghassabeh, S. Muyldermans, D. Charlier and E. Peeters (2013). "The genome-wide binding profile of the *Sulfolobus solfataricus* transcription factor Ss-LrpB shows binding events beyond direct transcription regulation." *BMC Genomics* 14(1): 828.
- Nitta, K. R., E. Jacox, R. Vincentelli, A. Jolma, D. Sobral, A. Mistral, A. Ohtsuka, A. Kubo, C. Cambillau, Y. Satou, J. Taipale and P. Lemaire (2012). "An atlas of Ascidian transcription factor DNA-binding sites reveals extensive conservation of DNA-binding specificities among chordates." In submission.
- Noll, K. M. and M. Vargas (1997). "Recent advances in genetic analyses of hyperthermophilic archaea and bacteria." *Archives of Microbiology* 168(2): 73-80.
- Nunoura, T., Y. Takaki, J. Kakuta, S. Nishi, J. Sugahara, H. Kazama, G. J. Chee, M. Hattori, A. Kanai, H. Atomi, K. Takai and H. Takami (2011). "Insights into the evolution of Archaea and eukaryotic protein modifier systems revealed by the genome of a novel archaeal group." *Nucleic acids research* 39(8): 3204-3223.
- Ochs, S. M., S. Thumann, R. Richau, M. T. Weirauch, T. M. Lowe, M. Thomm and W. Hausner (2012). "Activation of archaeal transcription mediated by recruitment of transcription factor B." *The Journal of biological chemistry* 287(22): 18863-18871.
- Oke, M., M. Kerou, H. Liu, X. Peng, R. A. Garrett, D. Prangishvili, J. H. Naismith and M. F. White (2011). "A dimeric Rep protein initiates replication of a linear archaeal virus genome: implications for the Rep mechanism and viral replication." *Journal of Virology* 85(2): 925-931.
- Okutan, E., L. Deng, S. Mirlashari, K. Uldahl, M. Halim, C. Liu, R. A. Garrett, Q. She and X. Peng (2013). "Novel insights into gene regulation of the rudivirus SIRV2 infecting *Sulfolobus* cells." *RNA biology* 10(5): 875-885.
- Omichinski, J. G., P. V. Pedone, G. Felsenfeld, A. M. Gronenborn and G. M. Clore (1997). "The solution structure of a specific GAGA factor-DNA complex reveals a modular binding mode." *Nature Structural Biology* 4(2): 122-132.
- Orell, A., E. Peeters, V. Vassen, S. Jachlewski, S. Schalles, B. Siebers and S. V. Albers (2013). "Lrs14 transcriptional regulators influence biofilm formation and cell motility of Crenarchaea." *The ISME journal* 7(10): 1886-1898.
- Ouhammouch, M. (2004). "Transcriptional regulation in Archaea." *Current Opinion in Genetics and Development* 14(2): 133-138.
- Ouhammouch, M., R. E. Dewhurst, W. Hausner, M. Thomm and E. P. Geiduschek (2003). "Activation of archaeal transcription by recruitment of the TATA-binding protein." *Proceedings of the National Academy of Sciences of the United States of America* 100(9): 5097-5102.
- Ouhammouch, M. and E. P. Geiduschek (2001). "A thermostable platform for transcriptional regulation: the DNA-binding properties of two Lrp homologs from the hyperthermophilic archaeon *Methanococcus jannaschii*." *The EMBO journal* 20(1-2): 146-156.
- Ouhammouch, M. and E. P. Geiduschek (2005). "An expanding family of archaeal transcriptional activators." *Proceedings of the National Academy of Sciences of the United States of America* 102(43): 15423-15428.
- Ouhammouch, M., W. Hausner and E. P. Geiduschek (2009). "TBP domain symmetry in basal and activated archaeal transcription." *Molecular Microbiology* 71(1): 123-131.
- Ouhammouch, M., G. E. Langham, W. Hausner, A. J. Simpson, N. M. El-Sayed and E. P. Geiduschek (2005). "Promoter architecture and response to a positive regulator of archaeal transcription." *Molecular Microbiology* 56(3): 625-637.

- Ouhammouch, M., F. Werner, R. O. Weinzierl and E. P. Geiduschek (2004). "A fully recombinant system for activator-dependent archaeal transcription." The Journal of biological chemistry **279**(50): 51719-51721.
- Pace, N. R., J. Sapp and N. Goldenfeld (2012). "Phylogeny and beyond: Scientific, historical, and conceptual significance of the first tree of life." Proceedings of the National Academy of Sciences of the United States of America **109**(4): 1011-1018.
- Palangat, M., C. T. Hittinger and R. Landick (2004). "Downstream DNA selectively affects a paused conformation of human RNA polymerase II." Journal of Molecular Biology **341**(2): 429-442.
- Palm, P., C. Schleper, B. Grampp, S. Yeats, P. McWilliam, W. D. Reiter and W. Zillig (1991). "Complete nucleotide sequence of the virus SSV1 of the archaeobacterium *Sulfolobus shibatae*." Virology **185**(1): 242-250.
- Pastre, D., L. Hamon, F. Landousy, I. Sorel, M. O. David, A. Zozime, E. Le Cam and O. Pietrement (2006). "Anionic polyelectrolyte adsorption on mica mediated by multivalent cations: a solution to DNA imaging by atomic force microscopy under high ionic strengths." Langmuir : the ACS journal of surfaces and colloids **22**(15): 6651-6660.
- Patikoglou, G. A., J. L. Kim, L. Sun, S. H. Yang, T. Kodadek and S. K. Burley (1999). "TATA element recognition by the TATA box-binding protein has been conserved throughout evolution." Genes & Development **13**(24): 3217-3230.
- Peck, R. F., S. DasSarma and M. P. Krebs (2000). "Homologous gene knockout in the archaeon *Halobacterium salinarum* with *ura3* as a counterselectable marker." Molecular Microbiology **35**(3): 667-676.
- Peeters, E., S. V. Albers, A. Vassart, A. J. Driessen and D. Charlier (2009). "Ss-LrpB, a transcriptional regulator from *Sulfolobus solfataricus*, regulates a gene cluster with a pyruvate ferredoxin oxidoreductase-encoding operon and permease genes." Molecular Microbiology **71**(4): 972-988.
- Peeters, E. and D. Charlier (2010). "The Lrp family of transcription regulators in archaea." Archaea (Vancouver, B.C.) **2010**: 750457.
- Peeters, E., B. T. Hoa, I. Zegers, D. Charlier and D. Maes (2005). "Overexpression, purification, crystallization and preliminary X-ray diffraction analysis of the C-terminal domain of Ss-LrpB, a transcription regulator from *Sulfolobus solfataricus*." Acta crystallographica. Section F, Structural biology and crystallization communications **61**(Pt 11): 985-988.
- Peeters, E., N. Peixeiro and G. Sezonov (2013). "Cis-regulatory logic in archaeal transcription." Biochemical Society Transactions **41**(1): 326-331.
- Peeters, E., T. L. Thia-Toong, D. Gigot, D. Maes and D. Charlier (2004). "Ss-LrpB, a novel Lrp-like regulator of *Sulfolobus solfataricus* P2, binds cooperatively to three conserved targets in its own control region." Molecular Microbiology **54**(2): 321-336.
- Peeters, E., L. van Oeffelen, M. Nadal, P. Forterre and D. Charlier (2013). "A thermodynamic model of the cooperative interaction between the archaeal transcription factor Ss-LrpB and its tripartite operator DNA." Gene **524**(2): 330-340.
- Peeters, E., C. Wartel, D. Maes and D. Charlier (2007). "Analysis of the DNA-binding sequence specificity of the archaeal transcriptional regulator Ss-LrpB from *Sulfolobus solfataricus* by systematic mutagenesis and high resolution contact probing." Nucleic acids research **35**(2): 623-633.

- Peeters, E., R. Willaert, D. Maes and D. Charlier (2006). "Ss-LrpB from *Sulfolobus solfataricus* condenses about 100 base pairs of its own operator DNA into globular nucleoprotein complexes." *The Journal of biological chemistry* **281**(17): 11721-11728.
- Peixeiro, N., J. Keller, B. Collinet, N. Leulliot, V. Campanacci, D. Cortez, C. Cambillau, K. R. Nitta, R. Vincentelli, P. Forterre, D. Prangishvili, G. Sezonov and H. van Tilbeurgh (2013). "Structure and function of AvtR, a novel transcriptional regulator from a hyperthermophilic archaeal lipothrixvirus." *Journal of Virology* **87**(1): 124-136.
- Peng, N., X. Ao, Y. X. Liang and Q. She (2011). "Archaeal promoter architecture and mechanism of gene activation." *Biochemical Society Transactions* **39**(1): 99-103.
- Peng, N., L. Deng, Y. Mei, D. Jiang, Y. Hu, M. Awayez, Y. Liang and Q. She (2012). "A synthetic arabinose-inducible promoter confers high levels of recombinant protein expression in hyperthermophilic archaeon *Sulfolobus islandicus*." *Applied and Environmental Microbiology* **78**(16): 5630-5637.
- Peng, N., Q. Xia, Z. Chen, Y. X. Liang and Q. She (2009). "An upstream activation element exerting differential transcriptional activation on an archaeal promoter." *Molecular Microbiology* **74**(4): 928-939.
- Peng, X., H. Blum, Q. She, S. Mallok, K. Brugger, R. A. Garrett, W. Zillig and D. Prangishvili (2001). "Sequences and replication of genomes of the archaeal rudiviruses SIRV1 and SIRV2: relationships to the archaeal lipothrixvirus SIFV and some eukaryal viruses." *Virology* **291**(2): 226-234.
- Peng, X., I. Holz, W. Zillig, R. A. Garrett and Q. She (2000). "Evolution of the family of pRN plasmids and their integrase-mediated insertion into the chromosome of the crenarchaeon *Sulfolobus solfataricus*." *Journal of Molecular Biology* **303**(4): 449-454.
- Peng, X., A. Kessler, H. Phan, R. A. Garrett and D. Prangishvili (2004). "Multiple variants of the archaeal DNA rudivirus SIRV1 in a single host and a novel mechanism of genomic variation." *Molecular Microbiology* **54**(2): 366-375.
- Perera, I. C. and A. Grove (2010). "Molecular mechanisms of ligand-mediated attenuation of DNA binding by MarR family transcriptional regulators." *Journal of molecular cell biology* **2**(5): 243-254.
- Perez-Rueda, E., J. Collado-Vides and L. Segovia (2004). "Phylogenetic distribution of DNA-binding transcription factors in bacteria and archaea." *Computational Biology and Chemistry* **28**(5-6): 341-350.
- Perez-Rueda, E. and S. C. Janga (2010). "Identification and genomic analysis of transcription factors in archaeal genomes exemplifies their functional architecture and evolutionary origin." *Molecular Biology and Evolution* **27**(6): 1449-1459.
- Perugini, G., A. Valenti, A. D'Amaro, M. Rossi and M. Ciaramella (2009). "Reverse gyrase and genome stability in hyperthermophilic organisms." *Biochemical Society Transactions* **37**(Pt 1): 69-73.
- Pfeiffer, F., S. C. Schuster, A. Broicher, M. Falb, P. Palm, K. Rodewald, A. Ruepp, J. Soppa, J. Tittor and D. Oesterhelt (2008). "Evolution in the laboratory: the genome of *Halobacterium salinarum* strain R1 compared to that of strain NRC-1." *Genomics* **91**(4): 335-346.
- Phillips, G. and V. de Crecy-Lagard (2011). "Biosynthesis and function of tRNA modifications in Archaea." *Current Opinion in Microbiology* **14**(3): 335-341.
- Pietila, M. K., E. Roine, L. Paulin, N. Kalkkinen and D. H. Bamford (2009). "An ssDNA virus infecting archaea: a new lineage of viruses with a membrane envelope." *Molecular Microbiology* **72**(2): 307-319.

- Pina, M., A. Bize, P. Forterre and D. Prangishvili (2011). "The archeoviruses." FEMS Microbiol Rev **35**(6): 1035-1054.
- Popescu, A., A. Karpay, D. A. Israel, R. M. Peek, Jr. and A. M. Krezel (2005). "Helicobacter pylori protein HP0222 belongs to Arc/MetJ family of transcriptional regulators." Proteins **59**(2): 303-311.
- Porter, K. and M. L. Dyll-Smith (2008). "Transfection of haloarchaea by the DNAs of spindle and round haloviruses and the use of transposon mutagenesis to identify non-essential regions." Molecular Microbiology **70**(5): 1236-1245.
- Prangishvili, D. (2013). "The wonderful world of archaeal viruses." Annual Review of Microbiology **67**: 565-585.
- Prangishvili, D., S. V. Albers, I. Holz, H. P. Arnold, K. Stedman, T. Klein, H. Singh, J. Hiort, A. Schweier, J. K. Kristjansson and W. Zillig (1998). "Conjugation in archaea: frequent occurrence of conjugative plasmids in Sulfolobus." Plasmid **40**(3): 190-202.
- Prangishvili, D., H. P. Arnold, D. Gotz, U. Ziese, I. Holz, J. K. Kristjansson and W. Zillig (1999). "A novel virus family, the Rudiviridae: Structure, virus-host interactions and genome variability of the sulfolobus viruses SIRV1 and SIRV2." Genetics **152**(4): 1387-1396.
- Prangishvili, D., P. Forterre and R. A. Garrett (2006). "Viruses of the Archaea: a unifying view." Nature reviews. Microbiology **4**(11): 837-848.
- Prangishvili, D. and R. A. Garrett (2004). "Exceptionally diverse morphotypes and genomes of crenarchaeal hyperthermophilic viruses." Biochemical Society Transactions **32**(Pt 2): 204-208.
- Prangishvili, D. and R. A. Garrett (2005). "Viruses of hyperthermophilic Crenarchaea." Trends in Microbiology **13**(11): 535-542.
- Prangishvili, D., R. A. Garrett and E. V. Koonin (2006). "Evolutionary genomics of archaeal viruses: unique viral genomes in the third domain of life." Virus Research **117**(1): 52-67.
- Prangishvili, D., H. P. Klenk, G. Jakobs, A. Schmiechen, C. Hanselmann, I. Holz and W. Zillig (1998). "Biochemical and phylogenetic characterization of the dUTPase from the archaeal virus SIRV." The Journal of biological chemistry **273**(11): 6024-6029.
- Prangishvili, D., E. V. Koonin and M. Krupovic (2013). "Genomics and biology of Rudiviruses, a model for the study of virus-host interactions in Archaea." Biochemical Society Transactions **41**(1): 443-450.
- Prangishvili, D. and M. Krupovic (2012). "A new proposed taxon for double-stranded DNA viruses, the order "Ligamenvirales". " Archives of Virology **157**(4): 791-795.
- Prangishvili, D. and T. E. Quax (2011). "Exceptional virion release mechanism: one more surprise from archaeal viruses." Current Opinion in Microbiology **14**(3): 315-320.
- Prangishvili, D., K. Stedman and W. Zillig (2001). "Viruses of the extremely thermophilic archaeon Sulfolobus." Trends in Microbiology **9**(1): 39-43.
- Prangishvili, D., G. Vestergaard, M. Haring, R. Aramayo, T. Basta, R. Rachel and R. A. Garrett (2006). "Structural and genomic properties of the hyperthermophilic archaeal virus ATV with an extracellular stage of the reproductive cycle." Journal of Molecular Biology **359**(5): 1203-1216.
- Prangishvili, D., W. Zillig, A. Gierl, L. Biesert and I. Holz (1982). "DNA-dependent RNA polymerase of thermoacidophilic archaeobacteria." European journal of biochemistry / FEBS **122**(3): 471-477.

- Prangishvili, D. A., R. P. Vashakidze, M. G. Chelidze and I. Gabriadze (1985). "A restriction endonuclease Sual from the thermoacidophilic archaeobacterium *Sulfolobus acidocaldarius*." FEBS Letters **192**(1): 57-60.
- Prangishvili, D., W. Zillig, A. Gierl, L. Biesert and I. Holz (1982). "DNA-dependent RNA polymerase of thermoacidophilic archaeobacteria." European journal of biochemistry / FEBS **122**(3): 471-477.
- Pritchett, M. A., S. P. Wilkinson, E. P. Geiduschek and M. Ouhammouch (2009). "Hybrid Ptr2-like activators of archaeal transcription." Molecular Microbiology **74**(3): 582-593.
- Pritchett, M. A., J. K. Zhang and W. W. Metcalf (2004). "Development of a markerless genetic exchange method for *Methanosarcina acetivorans* C2A and its use in construction of new genetic tools for methanogenic archaea." Applied and Environmental Microbiology **70**(3): 1425-1433.
- Probst, A. J., A. K. Auerbach and C. Moissl-Eichinger (2013). "Archaea on human skin." PloS one **8**(6): e65388.
- Prosser, J. I. and G. W. Nicol (2012). "Archaeal and bacterial ammonia-oxidisers in soil: the quest for niche specialisation and differentiation." Trends in Microbiology **20**(11): 523-531.
- Pryor, E., E. Waligora, B. Xu, S. Delloso-Nolan, D. Wozniak and T. Hollis (2012). "The Transcription Factor AmrZ Utilizes Multiple DNA Binding Modes to Recognize Activator and Repressor Sequences of *Pseudomonas aeruginosa* Virulence Genes." PLoS Pathog **8**(4): e1002648.
- Quax, T. E., M. Krupovic, S. Lucas, P. Forterre and D. Prangishvili (2010). "The *Sulfolobus* rod-shaped virus 2 encodes a prominent structural component of the unique virion release system in Archaea." Virology **404**(1): 1-4.
- Quax, T. E., S. Lucas, J. Reimann, G. Pehau-Arnaudet, M. C. Prevost, P. Forterre, S. V. Albers and D. Prangishvili (2011). "Simple and elegant design of a virion egress structure in Archaea." Proceedings of the National Academy of Sciences of the United States of America **108**(8): 3354-3359.
- Quax, T. E., M. Voet, O. Sismeiro, M. A. Dillies, B. Jagla, J. Y. Coppee, G. Sezonov, P. Forterre, J. van der Oost, R. Lavigne and D. Prangishvili (2013). "Massive activation of archaeal defense genes during viral infection." Journal of Virology **87**(15): 8419-8428.
- Quemin, E. R., S. Lucas, B. Daum, T. E. Quax, W. Kuhlbrandt, P. Forterre, S. V. Albers, D. Prangishvili and M. Krupovic (2013). "First insights into the entry process of hyperthermophilic archaeal viruses." Journal of Virology.
- Qureshi, S. A. (2007). "Protein-DNA interactions at the *Sulfolobus* spindle-shaped virus-1 (SSV1) T5 and T6 gene promoters." Canadian Journal of Microbiology **53**(9): 1076-1083.
- Qureshi, S. A., S. D. Bell and S. P. Jackson (1997). "Factor requirements for transcription in the Archaeon *Sulfolobus shibatae*." The EMBO journal **16**(10): 2927-2936.
- Rachel, R., M. Bettstetter, B. P. Hedlund, M. Haring, A. Kessler, K. O. Stetter and D. Prangishvili (2002). "Remarkable morphological diversity of viruses and virus-like particles in hot terrestrial environments." Archives of Virology **147**(12): 2419-2429.
- Rachel, R., I. Wyschkony, S. Riehl and H. Huber (2002). "The ultrastructure of *Ignicoccus*: evidence for a novel outer membrane and for intracellular vesicle budding in an archaeon." Archaea (Vancouver, B.C.) **1**(1): 9-18.
- Rafferty, J. B., W. S. Somers, I. Saint-Girons and S. E. Phillips (1989). "Three-dimensional crystal structures of *Escherichia coli* met repressor with and without corepressor." Nature **341**(6244): 705-710.

- Raumann, B. E., M. A. Rould, C. O. Pabo and R. T. Sauer (1994). "DNA recognition by beta-sheets in the Arc repressor-operator crystal structure." *Nature* **367**(6465): 754-757.
- Ream, T. S., J. R. Haag, A. T. Wierzbicki, C. D. Nicora, A. D. Norbeck, J. K. Zhu, G. Hagen, T. J. Guilfoyle, L. Pasa-Tolic and C. S. Pikaard (2009). "Subunit compositions of the RNA-silencing enzymes Pol IV and Pol V reveal their origins as specialized forms of RNA polymerase II." *Molecular Cell* **33**(2): 192-203.
- Redder, P. and R. A. Garrett (2006). "Mutations and rearrangements in the genome of *Sulfolobus solfataricus* P2." *Journal of Bacteriology* **188**(12): 4198-4206.
- Reeve, J. N., K. A. Bailey, W. T. Li, F. Marc, K. Sandman and D. J. Soares (2004). "Archaeal histones: structures, stability and DNA binding." *Biochemical Society Transactions* **32**(Pt 2): 227-230.
- Reiter, W. D. and P. Palm (1990). "Identification and characterization of a defective SSV1 genome integrated into a tRNA gene in the archaeobacterium *Sulfolobus* sp. B12." *Molecular & general genetics : MGG* **221**(1): 65-71.
- Reiter, W. D., P. Palm and S. Yeats (1989). "Transfer RNA genes frequently serve as integration sites for prokaryotic genetic elements." *Nucleic acids research* **17**(5): 1907-1914.
- Reiter, W. D., P. Palm, S. Yeats and W. Zillig (1987). "Gene expression in archaeobacteria: physical mapping of constitutive and UV-inducible transcripts from the *Sulfolobus* virus-like particle SSV1." *Molecular & general genetics : MGG* **209**(2): 270-275.
- Reiter, W. D., P. Palm and W. Zillig (1988). "Analysis of transcription in the archaeobacterium *Sulfolobus* indicates that archaeobacterial promoters are homologous to eukaryotic pol II promoters." *Nucleic acids research* **16**(1): 1-19.
- Ren, Y., Q. She and L. Huang (2013). "Transcriptomic analysis of the SSV2 infection of *Sulfolobus solfataricus* with and without the integrative plasmid pSSVi." *Virology* **441**(2): 126-134.
- Reno, M. L., N. L. Held, C. J. Fields, P. V. Burke and R. J. Whitaker (2009). "Biogeography of the *Sulfolobus islandicus* pan-genome." *Proceedings of the National Academy of Sciences of the United States of America* **106**(21): 8605-8610.
- Revyakin, A., C. Liu, R. H. Ebricht and T. R. Strick (2006). "Abortive initiation and productive initiation by RNA polymerase involve DNA scrunching." *Science (New York, N.Y.)* **314**(5802): 1139-1143.
- Rose, R. W., T. Bruser, J. C. Kissinger and M. Pohlschroder (2002). "Adaptation of protein secretion to extremely high-salt conditions by extensive use of the twin-arginine translocation pathway." *Molecular Microbiology* **45**(4): 943-950.
- Roulet, E., S. Busso, A. A. Camargo, A. J. Simpson, N. Mermoud and P. Bucher (2002). "High-throughput SELEX SAGE method for quantitative modeling of transcription-factor binding sites." *Nat Biotechnol* **20**(8): 831-835.
- Rowlands, T., P. Baumann and S. P. Jackson (1994). "The TATA-binding protein: a general transcription factor in eukaryotes and archaeobacteria." *Science (New York, N.Y.)* **264**(5163): 1326-1329.
- Saint-Girons, I., N. Duchange, G. N. Cohen and M. M. Zakin (1984). "Structure and autoregulation of the metJ regulatory gene in *Escherichia coli*." *J Biol Chem* **259**(22): 14282-14285.
- Sakofsky, C. J., L. A. Runck and D. W. Grogan (2011). "Sulfolobus mutants, generated via PCR products, which lack putative enzymes of UV photoproduct repair." *Archaea (Vancouver, B.C.)* **2011**: 864015.

- Samson, R. Y., T. Obita, S. M. Freund, R. L. Williams and S. D. Bell (2008). "A role for the ESCRT system in cell division in archaea." Science (New York, N.Y.) **322**(5908): 1710-1713.
- Samson, R. Y., Y. Xu, C. Gadelha, T. A. Stone, J. N. Faqiri, D. Li, N. Qin, F. Pu, Y. X. Liang, Q. She and S. D. Bell (2013). "Specificity and function of archaeal DNA replication initiator proteins." Cell reports **3**(2): 485-496.
- Sandler, S. J., L. H. Satin, H. S. Samra and A. J. Clark (1996). "recA-like genes from three archaean species with putative protein products similar to Rad51 and Dmc1 proteins of the yeast *Saccharomyces cerevisiae*." Nucleic acids research **24**(11): 2125-2132.
- Sandman, K., J. A. Krzycki, B. Dobrinski, R. Lurz and J. N. Reeve (1990). "HMf, a DNA-binding protein isolated from the hyperthermophilic archaeon *Methanothermobacter fervidus*, is most closely related to histones." Proceedings of the National Academy of Sciences of the United States of America **87**(15): 5788-5791.
- Sandman, K. and J. N. Reeve (2005). "Archaeal chromatin proteins: different structures but common function?" Current Opinion in Microbiology **8**(6): 656-661.
- Santangelo, T. J., L. Cubonova, K. M. Skinner and J. N. Reeve (2009). "Archaeal intrinsic transcription termination in vivo." Journal of Bacteriology **191**(22): 7102-7108.
- Santangelo, T. J. and J. N. Reeve (2006). "Archaeal RNA polymerase is sensitive to intrinsic termination directed by transcribed and remote sequences." Journal of Molecular Biology **355**(2): 196-210.
- Sato, T., T. Fukui, H. Atomi and T. Imanaka (2005). "Improved and versatile transformation system allowing multiple genetic manipulations of the hyperthermophilic archaeon *Thermococcus kodakaraensis*." Applied and Environmental Microbiology **71**(7): 3889-3899.
- Sauer, R. T., R. R. Yocum, R. F. Doolittle, M. Lewis and C. O. Pabo (1982). "Homology among DNA-binding proteins suggests use of a conserved super-secondary structure." Nature **298**(5873): 447-451.
- Schelert, J., V. Dixit, V. Hoang, J. Simbahan, M. Drozda and P. Blum (2004). "Occurrence and characterization of mercury resistance in the hyperthermophilic archaeon *Sulfolobus solfataricus* by use of gene disruption." Journal of Bacteriology **186**(2): 427-437.
- Schelert, J., M. Drozda, V. Dixit, A. Dillman and P. Blum (2006). "Regulation of mercury resistance in the crenarchaeote *Sulfolobus solfataricus*." Journal of Bacteriology **188**(20): 7141-7150.
- Schlenker, C., A. Goel, B. P. Tripet, S. Menon, T. Willi, M. Dlakic, M. J. Young, C. M. Lawrence and V. Copie (2012). "Structural studies of E73 from a hyperthermophilic archaeal virus identify the "RH3" domain, an elaborated ribbon-helix-helix motif involved in DNA recognition." Biochemistry **51**(13): 2899-2910.
- Schlenker, C., S. Menon, C. M. Lawrence and V. Copie (2009). "(1)H, (13)C, (15)N backbone and side chain NMR resonance assignments for E73 from *Sulfolobus* spindle-shaped virus ragged hills, a hyperthermophilic crenarchaeal virus from Yellowstone National Park." Biomolecular NMR assignments **3**(2): 219-222.
- Schleper, C. (2010). "Ammonia oxidation: different niches for bacteria and archaea?" The ISME journal **4**(9): 1092-1094.
- Schleper, C., I. Holz, D. Janekovic, J. Murphy and W. Zillig (1995). "A multicopy plasmid of the extremely thermophilic archaeon *Sulfolobus* effects its transfer to recipients by mating." Journal of Bacteriology **177**(15): 4417-4426.
- Schleper, C., K. Kubo and W. Zillig (1992). "The particle SSV1 from the extremely thermophilic archaeon *Sulfolobus* is a virus: demonstration of infectivity and of transfection with viral

- DNA." Proceedings of the National Academy of Sciences of the United States of America **89**(16): 7645-7649.
- Schleper, C. and G. W. Nicol (2010). "Ammonia-oxidising archaea--physiology, ecology and evolution." Advances in Microbial Physiology **57**: 1-41.
- Schleper, C., R. Roder, T. Singer and W. Zillig (1994). "An insertion element of the extremely thermophilic archaeon *Sulfolobus solfataricus* transposes into the endogenous beta-galactosidase gene." Molecular & general genetics : MGG **243**(1): 91-96.
- Schmid, A. K., D. J. Reiss, M. Pan, T. Koide and N. S. Baliga (2009). "A single transcription factor regulates evolutionarily diverse but functionally linked metabolic pathways in response to nutrient availability." Molecular Systems Biology **5**: 282.
- Schreiter, E. R. and C. L. Drennan (2007). "Ribbon-helix-helix transcription factors: variations on a theme." Nat Rev Microbiol **5**(9): 710-720.
- Schreiter, E. R., S. C. Wang, D. B. Zamble and C. L. Drennan (2006). "NikR-operator complex structure and the mechanism of repressor activation by metal ions." Proc Natl Acad Sci U S A **103**(37): 13676-13681.
- Schut, G. J., S. L. Bridger and M. W. Adams (2007). "Insights into the metabolism of elemental sulfur by the hyperthermophilic archaeon *Pyrococcus furiosus*: characterization of a coenzyme A-dependent NAD(P)H sulfur oxidoreductase." Journal of Bacteriology **189**(12): 4431-4441.
- Schwaiger, R., C. Schwarz, K. Furtwangler, V. Tarasov, A. Wende and D. Oesterhelt (2010). "Transcriptional control by two leucine-responsive regulatory proteins in *Halobacterium salinarum* R1." BMC Molecular Biology **11**: 40.
- Selth, L. A., S. Sigurdsson and J. Q. Svejstrup (2010). "Transcript Elongation by RNA Polymerase II." Annual Review of Biochemistry **79**: 271-293.
- Sencilo, A., L. Paulin, S. Kellner, M. Helm and E. Roine (2012). "Related haloarchaeal pleomorphic viruses contain different genome types." Nucleic acids research **40**(12): 5523-5534.
- Servin-Garciduenas, L. E., X. Peng, R. A. Garrett and E. Martinez-Romero (2013). "Genome sequence of a novel archaeal rudivirus recovered from a mexican hot spring." Genome announcements **1**(1).
- Sevostyanova, A., V. Svetlov, D. G. Vassilyev and I. Artsimovitch (2008). "The elongation factor RfaH and the initiation factor sigma bind to the same site on the transcription elongation complex." Proceedings of the National Academy of Sciences of the United States of America **105**(3): 865-870.
- Sharon, E., Y. Kalma, A. Sharp, T. Raveh-Sadka, M. Levo, D. Zeevi, L. Keren, Z. Yakhini, A. Weinberger and E. Segal (2012). "Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters." Nature Biotechnology **30**(6): 521-530.
- She, Q., H. Phan, R. A. Garrett, S. V. Albers, K. M. Stedman and W. Zillig (1998). "Genetic profile of pNOB8 from *Sulfolobus*: the first conjugative plasmid from an archaeon." Extremophiles : life under extreme conditions **2**(4): 417-425.
- She, Q., B. Shen and L. Chen (2004). "Archaeal integrases and mechanisms of gene capture." Biochemical Society Transactions **32**(Pt 2): 222-226.
- She, Q., R. K. Singh, F. Confalonieri, Y. Zivanovic, G. Allard, M. J. Awayez, C. C. Chan-Weiher, I. G. Clausen, B. A. Curtis, A. De Moors, G. Erauso, C. Fletcher, P. M. Gordon, I. Heikamp-de Jong, A. C. Jeffries, C. J. Kozera, N. Medina, X. Peng, H. P. Thi-Ngoc, P. Redder, M. E. Schenk, C. Theriault, N. Tolstrup, R. L. Charlebois, W. F. Doolittle, M. Duguet, T. Gaasterland, R. A. Garrett, M. A. Ragan, C. W. Sensen and J. Van der Oost (2001). "The complete genome of the

- crenarchaeon *Sulfolobus solfataricus* P2." Proceedings of the National Academy of Sciences of the United States of America **98**(14): 7835-7840.
- She, Q., C. Zhang, L. Deng, N. Peng, Z. Chen and Y. X. Liang (2009). "Genetic analyses in the hyperthermophilic archaeon *Sulfolobus islandicus*." Biochemical Society Transactions **37**(Pt 1): 92-96.
- Shockley, K. R., D. E. Ward, S. R. Chhabra, S. B. Connors, C. I. Montero and R. M. Kelly (2003). "Heat shock response by the hyperthermophilic archaeon *Pyrococcus furiosus*." Applied and Environmental Microbiology **69**(4): 2365-2371.
- Snyder, J. C., S. K. Brumfield, K. M. Kerchner, T. E. Quax, D. Prangishvili and M. J. Young (2013). "Insights into a viral lytic pathway from an archaeal virus-host system." Journal of Virology **87**(4): 2186-2192.
- Soares, D., I. Dahlke, W. T. Li, K. Sandman, C. Hethke, M. Thomm and J. N. Reeve (1998). "Archaeal histone stability, DNA binding, and transcription inhibition above 90 degrees C." Extremophiles : life under extreme conditions **2**(2): 75-81.
- Sollner, S., S. Berkner and G. Lipps (2006). "Characterisation of the novel restriction endonuclease *Suil* from *Sulfolobus islandicus*." Extremophiles : life under extreme conditions **10**(6): 629-634.
- Somers, W. S. and S. E. Phillips (1992). "Crystal structure of the met repressor-operator complex at 2.8 Å resolution reveals DNA recognition by beta-strands." Nature **359**(6394): 387-393.
- Song, N., T. Nguyen Duc, L. van Oeffelen, S. Muyldermans, E. Peeters and D. Charlier (2013). "Expanded target and cofactor repertoire for the transcriptional activator *LysM* from *Sulfolobus*." Nucleic acids research **41**(5): 2932-2949.
- Soppa, J. (2010). "Protein acetylation in archaea, bacteria, and eukaryotes." Archaea (Vancouver, B.C.) **2010**.
- Spitalny, P. and M. Thomm (2008). "A polymerase III-like reinitiation mechanism is operating in regulation of histone expression in archaea." Molecular Microbiology **67**(5): 958-970.
- Stedman, K. M., C. Schleper, E. Rumpf and W. Zillig (1999). "Genetic requirements for the function of the archaeal virus *SSV1* in *Sulfolobus solfataricus*: construction and testing of viral shuttle vectors." Genetics **152**(4): 1397-1405.
- Stedman, K. M., Q. She, H. Phan, H. P. Arnold, I. Holz, R. A. Garrett and W. Zillig (2003). "Relationships between fuselloviruses infecting the extremely thermophilic archaeon *Sulfolobus*: *SSV1* and *SSV2*." Research in Microbiology **154**(4): 295-302.
- Stedman, K. M., Q. She, H. Phan, I. Holz, H. Singh, D. Prangishvili, R. Garrett and W. Zillig (2000). "pING family of conjugative plasmids from the extremely thermophilic archaeon *Sulfolobus islandicus*: insights into recombination and conjugation in Crenarchaeota." Journal of Bacteriology **182**(24): 7014-7020.
- Steitz, T. A. (2008). "A structural understanding of the dynamic ribosome machine." Nature reviews. Molecular cell biology **9**(3): 242-253.
- Stetter, K. O. (1996). "Hyperthermophiles in the history of life." Ciba Foundation Symposium **202**: 1-10; discussion 11-18.
- Stormo, G. D. and Y. Zhao (2010). "Determining the specificity of protein-DNA interactions." Nature reviews. Genetics **11**(11): 751-760.
- Suzuki, T., T. Iwasaki, T. Uzawa, K. Hara, N. Nemoto, T. Kon, T. Ueki, A. Yamagishi and T. Oshima (2002). "*Sulfolobus tokodaii* sp. nov. (f. *Sulfolobus* sp. strain 7), a new member of the genus

- Sulfolobus isolated from Beppu Hot Springs, Japan." Extremophiles : life under extreme conditions **6**(1): 39-44.
- Szabo, Z., M. Sani, M. Groeneveld, B. Zolghadr, J. Schelert, S. V. Albers, P. Blum, E. J. Boekema and A. J. Driessen (2007). "Flagellar motility and structure in the hyperthermoacidophilic archaeon *Sulfolobus solfataricus*." Journal of Bacteriology **189**(11): 4305-4309.
- Tabatabaei, M., R. A. Rahim, N. Abdullah, A.-D. G. Wright, Y. Shirai, K. Sakai, A. Sulaiman and M. A. Hassan (2010). "Importance of the methanogenic archaea populations in anaerobic wastewater treatments." Process Biochemistry **45**(8): 1214-1225.
- Tenorio-Salgado, S., A. Huerta-Saquero and E. Perez-Rueda (2011). "New insights on gene regulation in archaea." Computational Biology and Chemistry **35**(6): 341-346.
- Terwilliger, T. C. (2000). "Maximum-likelihood density modification." Acta Crystallogr D Biol Crystallogr **56** (Pt 8): 965-972.
- Teske, A., K. U. Hinrichs, V. Edgcomb, A. de Vera Gomez, D. Kysela, S. P. Sylva, M. L. Sogin and H. W. Jannasch (2002). "Microbial diversity of hydrothermal sediments in the Guaymas Basin: evidence for anaerobic methanotrophic communities." Applied and Environmental Microbiology **68**(4): 1994-2007.
- Teufel, K., A. Bleiholder, T. Griesbach and F. Pfeifer (2008). "Variations in the multiple *tbp* genes in different *Halobacterium salinarum* strains and their expression during growth." Archives of Microbiology **190**(3): 309-318.
- Teufel, K. and F. Pfeifer (2010). "Interaction of transcription activator GvpE with TATA-box-binding proteins of *Halobacterium salinarum*." Archives of Microbiology **192**(2): 143-149.
- Thomm, M. and G. Wich (1988). "An archaeobacterial promoter element for stable RNA genes with homology to the TATA box of higher eukaryotes." Nucleic acids research **16**(1): 151-163.
- Torsvik, T. and I. D. Dundas (1974). "Bacteriophage of *Halobacterium salinarum*." Nature **248**(450): 680-681.
- Toussaint, A., D. Prangishvili and I. J. Molineux (2012). "Introduction to special issue on viruses and microbes." Virology **434**(2): 137.
- Tso, J. Y., D. J. Van Den Berg and L. J. Korn (1986). "Structure of the gene for *Xenopus* transcription factor TFIIIA." Nucleic acids research **14**(5): 2187-2200.
- Uzan, M. and E. S. Miller (2010). "Post-transcriptional control by bacteriophage T4: mRNA decay and inhibition of translation initiation." Virology journal **7**: 360.
- Valentine, D. L. (2007). "Adaptations to energy stress dictate the ecology and evolution of the Archaea." Nature reviews. Microbiology **5**(4): 316-323.
- van de Vossenberg, J. L., T. Ubbink-Kok, M. G. Elferink, A. J. Driessen and W. N. Konings (1995). "Ion permeability of the cytoplasmic membrane limits the maximum growth temperature of bacteria and archaea." Molecular Microbiology **18**(5): 925-932.
- Vassart, A., M. Van Wolferen, A. Orell, Y. Hong, E. Peeters, S. V. Albers and D. Charlier (2013). "Sa-Lrp from *Sulfolobus acidocaldarius* is a versatile, glutamine-responsive, and architectural transcriptional regulator." MicrobiologyOpen **2**(1): 75-93.
- Vassilyev, D. G. (2009). "Elongation by RNA polymerase: a race through roadblocks." Current Opinion in Structural Biology **19**(6): 691-700.
- Vestergaard, G., R. Aramayo, T. Basta, M. Haring, X. Peng, K. Brugger, L. Chen, R. Rachel, N. Boisset, R. A. Garrett and D. Prangishvili (2008). "Structure of the acidianus filamentous virus 3 and comparative genomics of related archaeal lipothrixviruses." J Virol **82**(1): 371-381.

- Vestergaard, G., M. Haring, X. Peng, R. Rachel, R. A. Garrett and D. Prangishvili (2005). "A novel rudivirus, ARV1, of the hyperthermophilic archaeal genus Acidianus." *Virology* **336**(1): 83-92.
- Vestergaard, G., S. A. Shah, A. Bize, W. Reitberger, M. Reuter, H. Phan, A. Briegel, R. Rachel, R. A. Garrett and D. Prangishvili (2008). "Stygiolobus rod-shaped virus and the interplay of crenarchaeal rudiviruses with the CRISPR antiviral system." *Journal of Bacteriology* **190**(20): 6837-6845.
- Vierke, G., A. Engelmann, C. Hebbeln and M. Thomm (2003). "A novel archaeal transcriptional regulator of heat shock response." *The Journal of biological chemistry* **278**(1): 18-26.
- Waage, I., G. Schmid, S. Thumann, M. Thomm and W. Hausner (2010). "Shuttle vector-based transformation system for *Pyrococcus furiosus*." *Applied and Environmental Microbiology* **76**(10): 3308-3313.
- Wagner, M., S. Berkner, M. Ajon, A. J. Driessen, G. Lipps and S. V. Albers (2009). "Expanding and understanding the genetic toolbox of the hyperthermophilic genus *Sulfolobus*." *Biochemical Society Transactions* **37**(Pt 1): 97-101.
- Wagner, M., M. van Wolferen, A. Wagner, K. Lassak, B. H. Meyer, J. Reimann and S. V. Albers (2012). "Versatile Genetic Tool Box for the Crenarchaeote *Sulfolobus acidocaldarius*." *Frontiers in microbiology* **3**: 214.
- Wagner, M., A. Wagner, X. Ma, J. C. Kort, A. Ghosh, B. Rauch, B. Siebers and S. V. Albers (2013). "Investigation of the malE promoter and MalR, a positive regulator of the maltose regulon, for the vector system in *Sulfolobus acidocaldarius*." *Applied and Environmental Microbiology*.
- Wang, J. C. (2002). "Cellular roles of DNA topoisomerases: a molecular perspective." *Nature reviews. Molecular cell biology* **3**(6): 430-440.
- Wang, Y., Z. Duan, H. Zhu, X. Guo, Z. Wang, J. Zhou, Q. She and L. Huang (2007). "A novel *Sulfolobus* non-conjugative extrachromosomal genetic element capable of integration into the host genome and spreading in the presence of a fusellovirus." *Virology* **363**(1): 124-133.
- Wardleworth, B. N., R. J. Russell, S. D. Bell, G. L. Taylor and M. F. White (2002). "Structure of Alba: an archaeal chromatin protein modulated by acetylation." *The EMBO journal* **21**(17): 4654-4662.
- Weihofen, W. A., A. Cicek, F. Pratto, J. C. Alonso and W. Saenger (2006). "Structures of omega repressors bound to direct and inverted DNA repeats explain modulation of transcription." *Nucleic acids research* **34**(5): 1450-1458.
- Weininger, U., M. Zeeb, P. Neumann, C. Löw, M. Stubbs, G. Lipps and J. Balbach (2009). "Structure-based stability analysis of an extremely stable dimeric DNA binding protein from *Sulfolobus islandicus*." *Biochemistry* **48**(42): 10030-10037.
- Weininger, U., M. Zeeb, P. Neumann, C. Low, M. T. Stubbs, G. Lipps and J. Balbach (2009). "Structure-based stability analysis of an extremely stable dimeric DNA binding protein from *Sulfolobus islandicus*." *Biochemistry* **48**(42): 10030-10037.
- Welander, P. V. and W. W. Metcalf (2008). "Mutagenesis of the C1 oxidation pathway in *Methanosarcina barkeri*: new insights into the Mtr/Mer bypass pathway." *Journal of Bacteriology* **190**(6): 1928-1936.
- Werner, F. (2007). "Structure and function of archaeal RNA polymerases." *Molecular Microbiology* **65**(6): 1395-1404.
- Werner, F. (2008). "Structural evolution of multisubunit RNA polymerases." *Trends in Microbiology* **16**(6): 247-250.

- Werner, F. (2012). "A nexus for gene expression-molecular mechanisms of Spt5 and NusG in the three domains of life." Journal of Molecular Biology **417**(1-2): 13-27.
- Werner, F. and D. Grohmann (2011). "Evolution of multisubunit RNA polymerases in the three domains of life." Nature reviews. Microbiology **9**(2): 85-98.
- Werner, F. and R. O. Weinzierl (2002). "A recombinant RNA polymerase II-like enzyme capable of promoter-specific transcription." Molecular Cell **10**(3): 635-646.
- Werner, F. and R. O. Weinzierl (2005). "Direct modulation of RNA polymerase core functions by basal transcription factors." Molecular and Cellular Biology **25**(18): 8344-8355.
- Werner, M., P. Thuriaux and J. Soutourina (2009). "Structure-function analysis of RNA polymerases I and III." Current Opinion in Structural Biology **19**(6): 740-745.
- Werner, T. (2000). "Identification and functional modelling of DNA sequence elements of transcription." Briefings in bioinformatics **1**(4): 372-380.
- Whitaker, R. J., D. W. Grogan and J. W. Taylor (2005). "Recombination shapes the natural population structure of the hyperthermophilic archaeon *Sulfolobus islandicus*." Molecular Biology and Evolution **22**(12): 2354-2361.
- Whittington, J. E., R. F. Delgadillo, T. J. Attebury, L. K. Parkhurst, M. A. Daugherty and L. J. Parkhurst (2008). "TATA-binding protein recognition and bending of a consensus promoter are protein species dependent." Biochemistry **47**(27): 7264-7273.
- Wiedenheft, B., K. Stedman, F. Roberto, D. Willits, A. K. Gleske, L. Zoeller, J. Snyder, T. Douglas and M. Young (2004). "Comparative genomic analysis of hyperthermophilic archaeal Fuselloviridae viruses." Journal of Virology **78**(4): 1954-1961.
- Wilkinson, S. P., M. Ouhammouch and E. P. Geiduschek (2010). "Transcriptional activation in the context of repression mediated by archaeal histones." Proceedings of the National Academy of Sciences of the United States of America **107**(15): 6777-6781.
- Wirth, J. F., J. C. Snyder, R. A. Hochstein, A. C. Ortmann, D. A. Willits, T. Douglas and M. J. Young (2011). "Development of a genetic system for the archaeal virus *Sulfolobus turreted icosahedral virus* (STIV)." Virology **415**(1): 6-11.
- Wisedchaisri, G., D. M. Dranow, T. J. Lie, J. B. Bonanno, Y. Patskovsky, S. A. Ozyurt, J. M. Sauder, S. C. Almo, S. R. Wasserman, S. K. Burley, J. A. Leigh and T. Gonen (2010). "Structural underpinnings of nitrogen regulation by the prototypical nitrogen-responsive transcriptional factor NrpR." Structure (London, England : 1993) **18**(11): 1512-1521.
- Woese, C. R., O. Kandler and M. L. Wheelis (1990). "Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya." Proceedings of the National Academy of Sciences of the United States of America **87**(12): 4576-4579.
- Wojtas, M. N. and N. G. Abrescia (2013). "Archaeal transcription: making up for lost time." Biochemical Society Transactions **41**(1): 356-361.
- Wojtas, M. N., M. Moggi, O. Millet, S. D. Bell and N. G. Abrescia (2012). "Structural and functional analyses of the interaction of archaeal RNA polymerase with DNA." Nucleic acids research **40**(19): 9941-9952.
- Wong, J., J. Lu, R. Edwards, L. Frost and M. Glover (2011). "Structural basis of cooperative DNA recognition by the plasmid conjugation factor, TraM." Nucleic acids research **39**(15): 6775-6788.

- Worthington, P., V. Hoang, F. Perez-Pomares and P. Blum (2003). "Targeted disruption of the alpha-amylase gene in the hyperthermophilic archaeon *Sulfolobus solfataricus*." Journal of Bacteriology **185**(2): 482-488.
- Wu, J., S. Wang, J. Bai, L. Shi, D. Li, Z. Xu, Y. Niu, J. Lu and Q. Bao (2008). "ArchaeaTF: an integrated database of putative transcription factors in Archaea." Genomics **91**(1): 102-107.
- Xie, Y. and J. N. Reeve (2004). "Transcription by *Methanothermobacter thermautotrophicus* RNA polymerase in vitro releases archaeal transcription factor B but not TATA-box binding protein from the template DNA." Journal of Bacteriology **186**(18): 6306-6310.
- Xue, H., R. Guo, Y. Wen, D. Liu and L. Huang (2000). "An abundant DNA binding protein from the hyperthermophilic archaeon *Sulfolobus shibatae* affects DNA supercoiling in a temperature-dependent fashion." Journal of Bacteriology **182**(14): 3929-3933.
- Yan, M. and J. D. Gralla (1997). "Multiple ATP-dependent steps in RNA polymerase II promoter melting and initiation." The EMBO journal **16**(24): 7457-7467.
- Yang, H., G. L. Lipscomb, A. M. Keese, G. J. Schut, M. Thomm, M. W. Adams, B. C. Wang and R. A. Scott (2010). "SurR regulates hydrogen production in *Pyrococcus furiosus* by a sulfur-dependent redox switch." Molecular Microbiology **77**(5): 1111-1122.
- Yokoyama, K., S. A. Ishijima, L. Clowney, H. Koike, H. Aramaki, C. Tanaka, K. Makino and M. Suzuki (2006). "Feast/famine regulatory proteins (FFRPs): *Escherichia coli* Lrp, AsnC and related archaeal transcription factors." FEMS Microbiology Reviews **30**(1): 89-108.
- Yokoyama, K., S. A. Ishijima, H. Koike, C. Kurihara, A. Shimowasa, M. Kabasawa, T. Kawashima and M. Suzuki (2007). "Feast/famine regulation by transcription factor FL11 for the survival of the hyperthermophilic archaeon *Pyrococcus* OT3." Structure (London, England : 1993) **15**(12): 1542-1554.
- Yokoyama, K., H. Nogami, M. Kabasawa, S. Ebihara, A. Shimowasa, K. Hashimoto, T. Kawashima, S. A. Ishijima and M. Suzuki (2009). "The DNA-recognition mode shared by archaeal feast/famine-regulatory proteins revealed by the DNA-binding specificities of TvFL3, FL10, FL11 and Ss-LrpB." Nucleic acids research **37**(13): 4407-4419.
- You, X. Y., C. Liu, S. Y. Wang, C. Y. Jiang, S. A. Shah, D. Prangishvili, Q. She, S. J. Liu and R. A. Garrett (2011). "Genomic analysis of *Acidianus hospitalis* W1 a host for studying crenarchaeal virus and plasmid life cycles." Extremophiles : life under extreme conditions **15**(4): 487-497.
- Zhang, C., L. Guo, L. Deng, Y. Wu, Y. Liang, L. Huang and Q. She (2010). "Revealing the essentiality of multiple archaeal pcna genes using a mutant propagation assay based on an improved knockout method." Microbiology (Reading, England) **156**(Pt 11): 3386-3397.
- Zhang, C., D. J. Krause and R. J. Whitaker (2013). "*Sulfolobus islandicus*: a model system for evolutionary genomics." Biochemical Society Transactions **41**(1): 458-462.
- Zhang, C., B. Tian, S. Li, X. Ao, K. Dalgaard, S. Gokce, Y. Liang and Q. She (2013). "Genetic manipulation in *Sulfolobus islandicus* and functional analysis of DNA repair genes." Biochemical Society Transactions **41**(1): 405-410.
- Zhang, C. and R. J. Whitaker (2012). "A broadly applicable gene knockout system for the thermoacidophilic archaeon *Sulfolobus islandicus* based on simvastatin selection." Microbiology (Reading, England) **158**(Pt 6): 1513-1522.
- Zheng, T., Q. Huang, C. Zhang, J. Ni, Q. She and Y. Shen (2012). "Development of a simvastatin selection marker for a hyperthermophilic acidophile, *Sulfolobus islandicus*." Applied and Environmental Microbiology **78**(2): 568-574.

- Zillig, W., H. P. Arnold, I. Holz, D. Prangishvili, A. Schweier, K. Stedman, Q. She, H. Phan, R. Garrett and J. K. Kristjansson (1998). "Genetic elements in the extremely thermophilic archaeon *Sulfolobus*." Extremophiles : life under extreme conditions **2**(3): 131-140.
- Zillig, W., D. Prangishvilli, C. Schleper, M. Elferink, I. Holz, S. Albers, D. Janekovic and D. Gotz (1996). "Viruses, plasmids and other genetic elements of thermophilic and hyperthermophilic Archaea." FEMS Microbiology Reviews **18**(2-3): 225-236.
- Zillig, W., K. O. Stetter and M. Tobien (1978). "DNA-dependent RNA polymerase from *Halobacterium halobium*." European journal of biochemistry / FEBS **91**(1): 193-199.
- Zolghadr, B., A. Klingl, A. Koerdt, A. J. Driessen, R. Rachel and S. V. Albers (2010). "Appendage-mediated surface adherence of *Sulfolobus solfataricus*." Journal of Bacteriology **192**(1): 104-110.
- Zolghadr, B., S. Weber, Z. Szabo, A. J. Driessen and S. V. Albers (2007). "Identification of a system required for the functional surface localization of sugar binding proteins with class III signal peptides in *Sulfolobus solfataricus*." Molecular Microbiology **64**(3): 795-806.
- Zuckerandl, E. and L. Pauling (1965). "Molecules as documents of evolutionary history." Journal of Theoretical Biology **8**(2): 357-366.

Annexes

Annexe 1. C. Jaubert *et al.*, Open Biology - Supplementary figures

Nodes families	Present in <i>S. islandicus</i> pan-genome	Present in <i>S. islandicus</i> LAL14/1
<i>ISH3</i> transposases, ISC1200	142	3
<i>IS110</i> transposases, ISC1229	113	2
CoA pathway genes	68	5
ATPase subunit of the ABC transporters	63	5
<i>IS1</i> transposases, ISC796	53	9
<i>IS607</i> transposases, ISC1921	48	0
<i>IS6</i> transposases, ISSis1 and ISSto2	48	2
Putative molybdopterins	40	4

Table S1. Functions associated to the most represented *nodes* in the *S. islandicus* pan-genome and their representatives in *S. islandicus* LAL14/1.

Locus_tag	Predicted function
SiL_0127	Hypothetical Protein
SiL_0155	Hypothetical Protein
SiL_0303	Hypothetical Protein
SiL_0329	Hypothetical Protein
SiL_0384	Hypothetical Protein
SiL_0409	Hypothetical Protein
SiL_0481	Putative Acyl-coenzyme A synthetases/AMP-(fatty) acid ligases
SiL_0498	Hypothetical Protein
SiL_0587	Putative methyltransferase small
SiL_0600	CRISPR-associated protein, Cmr3
SiL_0601	CRISPR-associated protein, Cmr2
SiL_0603	CRISPR-associated protein, Cmr5
SiL_0604	CRISPR-associated protein, Cmr1
SiL_0614	Hypothetical Protein
SiL_0630	Related to a CRISPR-associated protein
SiL_0657	Hypothetical Protein
SiL_0686	Hypothetical Protein
SiL_0706	Hypothetical Protein
SiL_0762	Hypothetical Protein
SiL_0794	Hypothetical Protein
SiL_0799	Hypothetical Protein
SiL_0810	Hypothetical Protein
SiL_0818	Putative Glycosyltransferase
SiL_0823	Hypothetical Protein
SiL_0837	Hypothetical Protein
SiL_0838	Hypothetical Protein; distantly related to ATV_gp71
SiL_0839	Membrane protein involved in the export of O-antigen and teichoic acid
SiL_0855	Hypothetical Protein
SiL_0857	Hypothetical Protein

SiL_0860	Hypothetical Protein
SiL_0874	Hypothetical Protein
SiL_0876	Hypothetical Protein
SiL_0890	Hypothetical Protein
SiL_0909	Hypothetical Protein
SiL_1006	Hypothetical Protein
SiL_1086	Hypothetical Protein
SiL_1316	Hypothetical Protein
SiL_1318	Hypothetical Protein
SiL_1322	Hypothetical Protein
SiL_1325	Hypothetical Protein
SiL_1330	Hypothetical Protein
SiL_1477	Hypothetical Protein
SiL_1655	Hypothetical Protein
SiL_1730	Hypothetical Protein
SiL_1761	Hypothetical Protein
SiL_1951	Hypothetical Protein
SiL_1973	Hypothetical Protein
SiL_2104	Hypothetical Protein
SiL_2115	Hypothetical Protein
SiL_2365	Hypothetical Protein
SiL_2366	Hypothetical Protein

A

Locus_tag	Predicted function
SiL_0408	Hypothetical protein
SiL_0507	Hypothetical protein
SiL_1048	Hypothetical protein
SiL_1321	Conjugative plasmid protein
SiL_1473	Fe-S oxidoreductase
SiL_1752	Hypothetical protein
SiL_1913	Hypothetical protein

B

Locus_tag	Predicted function
SiL_0405	Putative MarR family transcriptional regulator
SiL_0602	CRISPR-associated protein Cmr6
SiL_0605	CRISPR-associated protein Cmr4
SiL_1009	Hypothetical protein
SiL_1319	Predicted secreted endonuclease
SiL_1320	Conjugative plasmid protein
SiL_2368	Viral protein (SSV-like)

C

Table S3. List of the genes of *S. islandicus* LAL14/1 specific to (A) *S. islandicus* LAL14/1 (B) Sulfolobus; (C) *Crenarchaeota*.

Word length	Total number of distinct words	Most underrepresented words	Most overrepresented words
2	16	ag/ct; cc/gg	ac/gt; <u>cg</u>
3	64	acg/cgt; cca/tcc; gga/tgg	aca/tgt; cga/tcg; <u>gac/gtc</u>
4	256	cagc/gctg; catc/gatg; cctc/gagg; gaaa/ttcc	attc/gaat; cagg/cctg; gaac/gttc; gagc/gctc;
5	1024	aatcc/ggatt; aattc; catta; taagg; tgatg	aatca; agtaa; attca; catcc; gaaac; ggatg; gtttc; tgaat; tgatt; ttact
6	4096	aaatta; ctacta ; cttaac; cttagc; gcagga; gctaag; gttaag; taaaaa; taattt; tagtag; tcctgc; ttttta	aaaaaa/tttttt; aagtag; <u>ggtacc</u> ; <u>gttaac</u> ; gttatc; <u>taatta</u>
7	16384	atTTTT; caacttc; cattttc; gaaatag; gaatttg; gatactg; gattacg	aaaaaaa/ttttttt; catatta; catttca; taaatag; tagattg; taggatg; taggtg
8	65536	actataga/tctatagt ; attgaaag/ctttcaat; ctaatcta ; ctttgtgt; taatctac ; tagaattg ; tctactat ; ttcaattc	aaatctac; aatagaat; <u>actatagt</u> ; cttgtacc; gattttac; tagattaa/ttaatcta; <u>tctataga</u> ; ttgtgtt

Table S4. Most under- and overrepresented words in the genome of *S. islandicus* LAL14/1.

Some of the most exceptional words are shown and were selected according to their R'MES rank. *Palindromic words* are indicated as underlined words in italic characters. Words of length 6 to 8 related to CRISPR spacers are indicated in bold characters. The words are grouped with their reverse complement (word/ reverse complement) whenever relevant.

Dot-plot LAL14/1 vs HVE10/4		
Position		Description
LAL14/1	HVE 10/4	
344.209 bp - 381.042 bp (36,8kb)	-	Absent in HVE 10/4. LAL14/1 – contains CRISPRs family I
-	392.309 bp - 452.067 bp (59,8kb)	Absent in LAL 14/1
381.042 bp - 871.850 bp (490,8kb)	452.067 bp - 1.014.813 bp (562,7kb)	Inversion. Variable region. Presence of CRISPRs of family III in LAL14/1.
766.321 bp - 833.855 bp (67,5kb)	-	Absent in HVE 10/4. Contains 73ORFs including methyltransferase (20 ORFs) and glycosylase (11 ORFs) genes
871.850 bp - 344.209 bp (1.957kb)	1.014.813 bp - 392.309 bp (2.032kb)	Conserved in both strains.
1.230.660 bp - 1.265.311 bp (34,7kb)	1.374.116 bp - 1.421.105 bp (47kb)	Inversion. Contains 3 ARNt genes and 43 ORF of unknown function

A

Dot-plot LAL14/1 vs REY15A		
Position		Description
LAL 14/1	REY15A	
317.849 bp - 353.843 bp (36kb)	721.233 bp - 768.241 bp (47,1kb)	Inversion. Variable region. LAL14/1 -presence of CRISPRs family I
353.843 bp -374.796 bp (21kb)	-	Absent in REY15A.
374.796 bp - 736.187 bp	333.277 bp - 721.233 bp	Conserved in both strains. Variable region

(361,3kb)	(388kb)	
736.187 bp -842.637 bp (106,5kb)	768.241 bp -874.957 bp (106,7kb)	Inversion. Variable region.
762.101 bp -806.736 bp (44,6kb)	-	Absent in REY15A.
842.637 bp -875.499 bp (32,8kb)	-	Absent in REY15A. Contains genes <i>vapBC</i> and some genes coning for the ABC transporters.
875.499 bp - 317.849 bp (1.907,5kb)	874.957 bp - 333.277 bp (1.981,3kb)	Conserved in both strains. Well conserved in all <i>S. islandicus</i> .
2.153.028 bp - 2.264.012bp (111kb)	2.189.662bp - 2.313.104bp (123.4kb)	Inversion

B

Table S5. Comparison of genome structures by dot-plot analysis. (A) *S. islandicus* LAL14/1 vs HVE 10/4. (B) *S. islandicus* LAL14/1 vs REY15A.

<i>S. islandicus</i> strain	Position in the genome from..to..	Variable region size in kb
LAL 14/1	282-890	608
HVE 10/4	294-1019	725
REY 15A	304-891	587
LD8.5	294-1013	719
LS2.15	310-1078	768
M14.25	302-1001	699
M16.27	296-1090	794
M16.4	310-980	670
YG57.14	300-973	673
YN15.51	2407-1605	802

Table S6. Positions and sizes of variable regions in different strains of *S. islandicus*

Position	tRNA type		Intron start	End	Intron
	Anti-codon				
994191..994264	Ala	CGC	-	-	No
1484693..1484766	Ala	GGC	-	-	No
1594052..1594125	Ala	TGC	-	-	No
1570342..1570416	Arg	CCG	-	-	No
1526901..1526988	Arg	CCT	1526940	1526952	Yes
1353142..1353216	Arg	GCG	-	-	No
1161877..1161951	Arg	TCG	-	-	No
1568617..1568706	Arg	TCT	1568656	1568670	Yes
1759525..1759611	Asn	GTT	1759574	1759561	Yes
1633028..1633100	Asp	GTC	-	-	No
1104887..1104959	Gln	CTG	-	-	No
1432100..1432172	Gln	TTG	-	-	No
1178938..1179030	Glu	CTC	-	-	No
1376626..1376715	Glu	TTC	-	-	No
1247049..1247124	Gly	CCC	-	-	No
1431771..1431846	Gly	GCC	-	-	No
1479146..1479221	Gly	TCC	-	-	No
1638129..1638202	His	GTG	-	-	No
996201..996286	Ile	GAT	996239	996250	Yes
1057784..1057882	Leu	CAA	1057843	1057829	Yes
1186783..1186867	Leu	CAG	-	-	-
111325..111409	Leu	GAG	-	-	No
1188894..1188993	Leu	TAA	1188953	1188939	Yes
1162261..1162345	Leu	TAG	-	-	No
1678998..1679093	Lys	CTT	1679036	1679057	Yes
1697966..1698062	Lys	TTT	1698004	1698026	Yes
1365026..1365117	Met	CAT	1365065	1365081	Yes
1632832..1632934	Met	CAT	1632871	1632895	Yes
1185332..1185405	Met	CAT	-	-	No
1230694..1230767	Phe	GAA	-	-	No

1486173..1486248	Pro	CGG	-	-	No
1011465..1011561	Pro	GGG	1011505	1011525	Yes
1760144..1760219	Pro	TGG	-	-	No
1184871..1184978	Ser	CGA	1184908	1184931	Yes
1759153..1759237	Ser	GCT	-	-	No
1631494..1631577	Ser	GGA	-	-	No
1476812..1476896	Ser	TGA	-	-	No
1610286..1610373	Thr	CGT	1610325	1610337	Yes
2207684..2207758	Thr	GGT	-	-	No
1105124..1105213	Thr	TGT	1105163	1105177	Yes
1736117..1736255	Trp	CCA	1736155	1736219	Yes
1183357..1183445	Tyr	GTA	1183397	1183409	Yes
1698229..1698303	Val	CAC	-	-	No
1589773..1589847	Val	GAC	-	-	No
1179203..1179277	Val	TAC	-	-	No

Table S7. tRNA genes of *S. islandicus* LAL14/1

Strain	Atypical coding genes % by genome	Fraction of atypical coding genes in large variable region	% of atypical genes situated in large variable region
LAL14/1	21,4%	49,0%	55,2%
REY15A	19,9%	46,4%	55,7%
HVE10/4	18,9%	50,9%	63,2%
LD85	22,6%	47,4%	48,4%
YG5714	22,3%	51,1%	51,1%
YN1551	21,5%	50,7%	62,5%
M1425	12,4%	35,7%	59,9%
LS215	17,0%	42,3%	59,9%
M1627	14,4%	40,4%	68,8%
M164	16,0%	40,9%	61,2%

Table S8. CAG frequency and distribution in 10 *S. islandicus* genomes

Replication origin	OriC-1	OriC-2	OriC-3
Gene linked	<i>cdc6-1</i>	<i>cdc6-3</i>	<i>whiP</i>
strains			
<i>S. solfataricus</i> P2	220kb	2Mb	740kb
<i>S. solfataricus</i> 98/2	1,15Mb	800pb	1,65Mb
<i>S. tokodaii</i>	322kb	2,16Mb	1,25Mb
<i>S. acidocaldarius</i>	578kb	100pb	1,19Mb
<i>S. islandicus</i> LAL 14/1	1,59Mb	800pb	1,15Mb
<i>S. islandicus</i> HVE 10/4	1,75Mb	800pb	1,30Mb
<i>S. islandicus</i> REY15A	1,6Mb	800pb	1,16Mb
<i>S. islandicus</i> M16.4	1,73Mb	800pb	1,26Mb
<i>S. islandicus</i> M16.27	1,83Mb	800pb	1,36Mb
<i>S. islandicus</i> M14.25	1,75Mb	800pb	1,280Mb
<i>S. islandicus</i> YG57.14	1,76Mb	800pb	1,25Mb
<i>S. islandicus</i> YN15.51	863kb	800pb	1,31Mb
<i>S. islandicus</i> LD8.5	1,82Mb	800pb	1,29Mb
<i>S. islandicus</i> LS2.15	1,81Mb	800pb	1,34Mb

Table S9. Positions of the three origins of replication *oriC* in different *Sulfolobus* genomes.

<i>S. islandicus</i> strains	Number of <i>vapBC</i> loci	Number of <i>HEPN-NT</i> * loci
LAL 14/1	15	5
HVE 10/4	18	5
REY 15A	17	6
LD8.5	21	6 (3)
LS2.15	24	6 (1)
M14.25	21	6 (1)
M16.4	21	6 (1)
M16.27	21	6 (1)
YG57.14	20	6
YN15.51	19	6 (1)

Table S10. *vapBC* and *HEPN-NT* loci in *S. islandicus*. The digits in brackets indicate the number of loci carrying at least one disrupted ORF; these copies of *HEPN-NT* are considered to be non-functional

		REY 15A	LS85	M1425	LS215	YG5714	YN1551	M1627	M164	HVE10/4	LAL14/1	Genomic region	
Family I	HEPN	SiRe_094 7	LD85_1191	M1425_1074	LS215_1171	YG5714_106 4	YN1551_179 8	M1627_1138	M164_1063	SiH_1034	SiL_0948	Conserved region	
	NT	SiRe_094 8	LD85_1190	M1425_1073	LS215_1170	YG5714_106 3	YN1551_179 9	M1627_1137	M164_1062	SiH_1035	SiL_0947		
Family II	HEPN	SiRe_093 8	LD85_1187	M1425_1070	LS215_1167	YG5714_106 0	YN1551_180 2	M1627_1134	M164_1059	SiH_1027	SiL_0944		
	NT	SiRe_093 9	LD85_1188	M1425_1071	LS215_1168	YG5714_106 1	YN1551_180 1	M1627_1135	M164_1060	SiH_1028	SiL_0945		
	HEPN	SiRe_091 7				YG5714_203 1	YN1551_022 1						
	NT	SiRe_091 8				YG5714_203 0	YN1551_022 2						
	HEPN	SiRe_045 8	LD85_0337	M1425_0330	LS215_0355	YG5714_033 3	YN1551_272 5	M1627_0330	M164_0352	SiH_0331	SiL_0488		Variable region
	NT	SiRe_045 9	LD85_0336	M1425_0329	LS215_0354	YG5714_033 2	YN1551_252 6	M1627_0329	M164_0351	SiH_0330	SiL_0487		
	HEPN	SiRe_092 4	LD85_0452*	M1425_0753 *	LS215_0865	YG5714_092 0	YN1551_195 9	M1627_0758*	M164_0795*	SiH_1009	SiL_1923		
	NT	SiRe_092 3	LD85_0451	M1425_0752	LS215_0864	YG5714_091 9	YN1551_196 0	M1627_0757	M164_0794	SiH_1008	SiL_1922		
	HEPN	SiRe_090 2	LD85_1061*	M1425_0454	LS215_0536	YG5714_104 4	YN1551_182 0	M1627_1098	M164_1041	SiH_0561	SiL_0783		
	NT	SiRe_090 3	LD85_1060	M1425_0455	LS215_0535	YG5714_104 3	YN1551_181 9	M1627_1099	M164_1040	SiH_0562	SiL_0782		
	HEPN	SiRe_073 1	LD85_0839*	M1425_0456	LS215_0829 *	YG5714_096 5	YN1551_089 2*	M1627_0468	M164_0490				
	NT	SiRe_073 0	LD85_0838	M1425_0458	LS215_0828	YG5714_096 6	YN1551_089 3	M1627_0469	M164_0491				
	HEPN		LD85_0841				YN1551_206 2	M1627_0851					
	NT		LD85_0842				YN1551_206 3	M1627_0852					

Table S11. Diversity of the *HEPN-NT* gene pairs in 10 *S. islandicus* genomes

* truncated version of corresponding genes

<i>Sulfolobus</i> species	<i>ups</i> operons				
SSP2	SSO0117	SSO0118	SSO0119	SSO0120	SSO0121
SS98/2	Ssol98_02032	Ssol98_02037	Ssol98_02042	Ssol_1095	Ssol98_02052
LAL 14.1	SiL_1862	SiL_1861	SiL_1860	SiL_1859	SiL_1858
M.14.25	M1425_2008	M1425_2007	M1425_2006	M1425_2005	M1425_2004
M.16.4	M164_2015	M164_2014	M164_2013	M164_2012	M164_2011
M.16.27	M1627_2086	M1627_2085	M1627_2084	M1627_2083	M1627_2082
HVE 10/4	SiH_1954	SiH_1953	SiH_1952	SiH_1951	SiH_1950
L.S.2.15	LS215_2173	LS215_2172	LS215_2171	LS215_2170	LS215_2169
L.D.8.5	LD85_2273	LD85_2272	LD85_2271	LD85_2270	LD85_2269
REY15A	SiRe_1882	SiRe_1881	SiRe_1880	SiRe_1879	SiRe_1878
Y.G.57.14	YG5714_2134	YG5714_2133	YG5714_2132	YG5714_2131	YG5714_2130
Y.N.15.51	YN1551_0785	YN1551_0786	YN1551_0787	YN1551_0788	YN1551_0789
<i>S.tokodaii</i>	ST1400	ST1399	ST1398	ST1397	ST1396
<i>S.acidocaldarius</i>	not annotated	Saci_1496	Saci_1495	Saci_1494	Saci_1493
<i>M.sedulla</i>	Msed_2107	Msed_1193	Msed_2105	Msed_2104	Msed_2103

Table S12. Positions the operons *ups* coding for the pili of type IV in *Sulfolobales*

		<i>S. islandicus</i> strains					
		HVE 10/4*		REY 15A*		LAL 14/1	
Family	Element	Intact orf	Total	Intact orf	Total	Intact orf	Total (variable)
<i>IS1</i>	ISC796	1	4	0	5	1 (0)	9 (2)
<i>IS5</i>	ISC1058	0	1	0	2	0	3 (0)
<i>IS5</i>	ISC1234	5	7	6	8	0	0
<i>IS6</i>	ISSis1	1	2	1	2	0	0
<i>IS6</i>	ISSto2	4	5	2	4	0	2 (1)
<i>IS110</i>	ISC1190	0	1	1	3	0	3 (0)
<i>IS110</i>	ISC1229	9	10	2	4	1 (0)	2 (1)
<i>IS110</i>	ISSto4	0	0	0	0	1 (1)	1 (1)
<i>IS200/605</i>	ISC1733	2	2	8	8	4 (2)	4 (2)
<i>IS256</i>	ISC1332	1	1	1	1	0	1 (1)
<i>IS605</i>	OrfB	17	18	18	19	10	12 (10)
<i>IS607</i>	ISC1921	0	0	1	1	0	0
<i>IS630</i>	ISC1048	0	12	0	10	0	8 (4)
<i>IS630</i>	ISC1078	0	1	0	1	0	0
<i>IS630</i>	ISC1395	0	0	1	2	0	2 (2)
<i>ISH3</i>	ISC1200	3	7	11	22	0	3 (2)
<i>ISCNY</i>	ISC1205	2	4	0	3	0	3 (3)
<i>ISL3</i>	ISC1043	0	1	0	1	0	0
Total		45	76	52	96	7	53

Table S13. Comparison of the *IS* content of the genomes of three *S. islandicus* strains isolated from Iceland.

* data from Guo et al., *J Bact* 2011;193(7):1672-1680.

	*HVE 10/4	*REY15A	LAL 14/1	LD8.5	LS2.15	M14.25	M16.27	M16.4	YG57.14	YN15.51
SM3A	2(1)	2(1)	11(7)	3(2)	3(2)	3(2)	4(4)	4(3)	3(2)	2(1)
SMN1	9(8)	7(7)	9(9)	0	0	2(2)	1	5(4)	0	0

Table S14 MITEs distribution in *S. islandicus*. In brackets the number of MITEs situated in the variable region.

* - as published by Guo et al., *J Bact* 2011;193(7):1672-1680.

Spacers	Putative targets	Targeted ORF	Length of spacer vs. target	Number of mismatches
1_11	SIRV1 VIII	ORF98	45/45	0
1_36	SIRV2	ORF119c	39/39	2
	SIRV1 XX	ORF119	39/39	3
	SIRV1 VIII	ORF117	39/39	3
1_66	SIRV1 XX	ORF112	41/41	4
	SIRV1 VIII	ORF112	41/41	4
1_77	SSV2	ORF72	45/45	1
1_83	SIRV2	ORF436	38/44	3
	SIRV1 XX	ORF440	38/44	4
	SIRV1 VIII	ORF440	38/44	4
1_85	STIV2	STIV2_B265	37/40	3
1_86	SSV7	SSV7_D336	40/41	2
	SSV2	ORF328	40/41	2
2_3	SIRV2	-	41/46	5
2_6	pING1	ORF1042	40/42	2
	pSOG2	ORF1094	42/42	3
2_13	STIV2	STIV2_F98	39/39	1
2_38	SSV2	ORF328	40/40	1
2_40	pXZ1	ORF82	38/40	2
2_58	SIRV1	ORF1070	40/40	1
	SIRV1 XX	ORF1070	40/40	5
	SIRV2	ORF1070	40/40	5
2_64	STIV2	STIV2_C141 STIV2_E65	34/39	1
2_67	pLD8501	LD85_3160	41/42	2
	L.D.8.5	LD85_1584	41/42	2
	pAH1	ORF78	41/42	2
	M.16.27	M1627_1565	42/42	3
	98/2	Ssol_1586	42/42	4
	P2	-	42/42	4
2_83	M.14.25	M1425_1876	41/41	4
	M.16.4	M164_1349	41/41	5
	Y.N.15.51	YN1551_0201	41/41	5
	Y.G.57.14	YG5714_2694	41/41	5
2_88	SIRV1 XX	-	35/39	1
	SIRV1	-	35/39	1
2_92	pARN4	-	40/42	3
2_93	pARN3	ORF609	41/41	0
2_98	SIRV1 XX	ORF1070	41/41	1
	SIRV2	ORF1070	41/41	2
2_112	SIRV1 VIII	ORF356	41/42	2
	SIRV1 XX	ORF356	41/42	2
	SIRV2	ORF356	41/42	3
2_113	M.16.45	CRISPR region	40/41	2

	M .16.43	CRISPR region	40/41	2
3_8	SIRV1 XX	ORF335	34/34	2
	SIRV1 VIII	ORF335	34/34	2
4_1	SIRV1 XX	ORF105	33/33	1
	SIRV1	ORF90	33/33	1
4_5	SIRV1 XX	ORF74	36/36	2
	SIRV VIII	ORF84a	36/36	2
	LS2.15	CRISPR	34/36	2
4_6	SIRV2	ORF309/158a	36/36	2
	SIRV1 XX	ORF417/158b	36/36	3
	SIRV VIII	ORF417/158b	36/36	3
4_17	SIRV1 XX	ORF356	36/36	1
	SIRV1 VIII	ORF356	36/36	1
	SIRV2	ORF356	35/36	2
5_1	SIRV1 XX	ORF440	37/37	0
	SIRV1 VIII	ORF440	37/37	0
	SIRV-2	ORF436	37/37	3
5_2	SIRV2	ORF56b/119c	38/38	2
5_3	SIRV2	ORF156	32/35	1
	SIRV1 XX	ORF156	33/35	2
5_6	SIRV2	ORF436	36/39	1
	SIRV1 XX	ORF440	35/39	1
	SIRV1 VIII	ORF440	35/39	1
5_7	SIRV1 XX	ORF105	33/33	1
	SIRV1 VIII	ORF105	33/33	1
	SIRV2	ORF102	33/33	2

Table S15. Spacers/targets correspondence for the CRISPRs of *S. islandicus* LAL14/1. For the spacers shorter than targets only the spacers corresponding to 85% or more of the putative target length retained. Spacers with 100% of correspondence to the targets are in bold.

Supplementary Figures

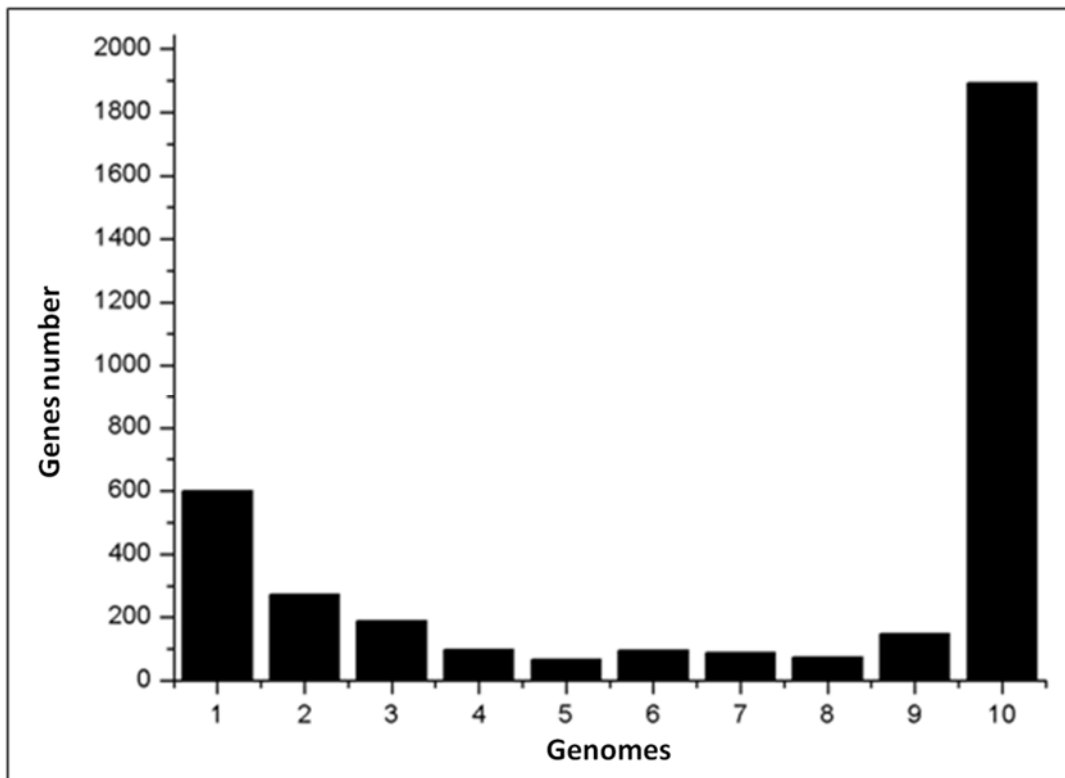


Figure S1. *S. islandicus* pan genome analysis. The histogram indicates the number of genes shared by indicated number of *S. islandicus* genomes. Plot 1 – unique genes present in only one from 10 genomes (*singletons*); plot 10 – number of ubiquitous genes shared by all *S. islandicus* (*core genome*).

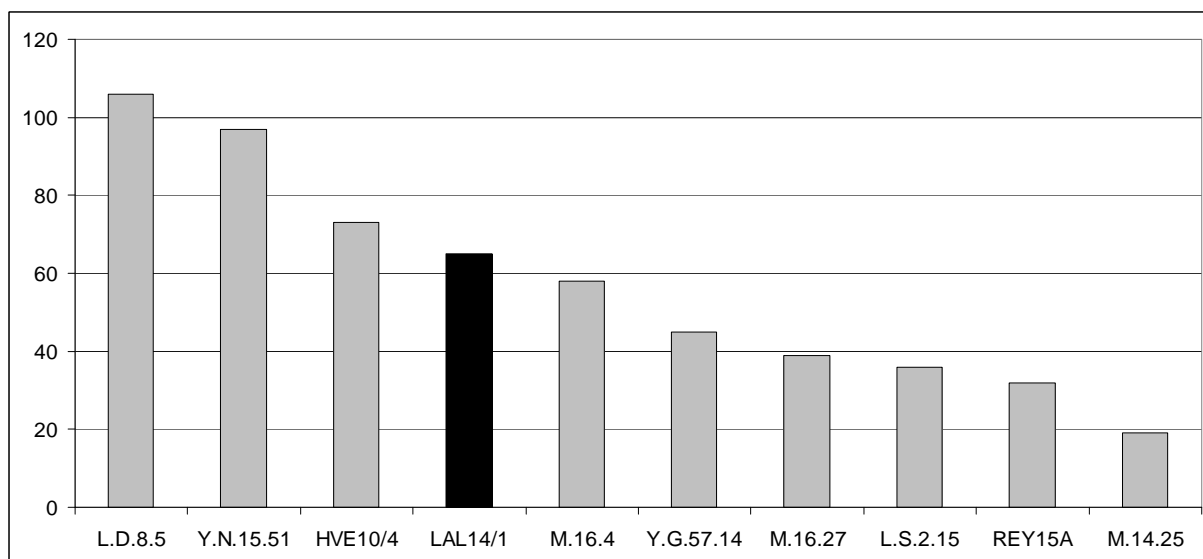


Figure S2. Singleton distribution in ten *S. islandicus* strains.

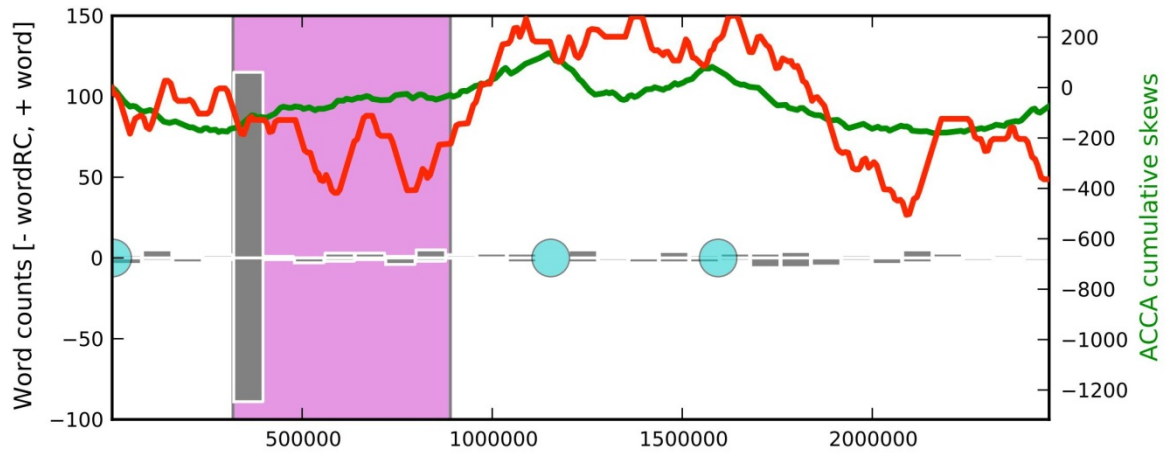
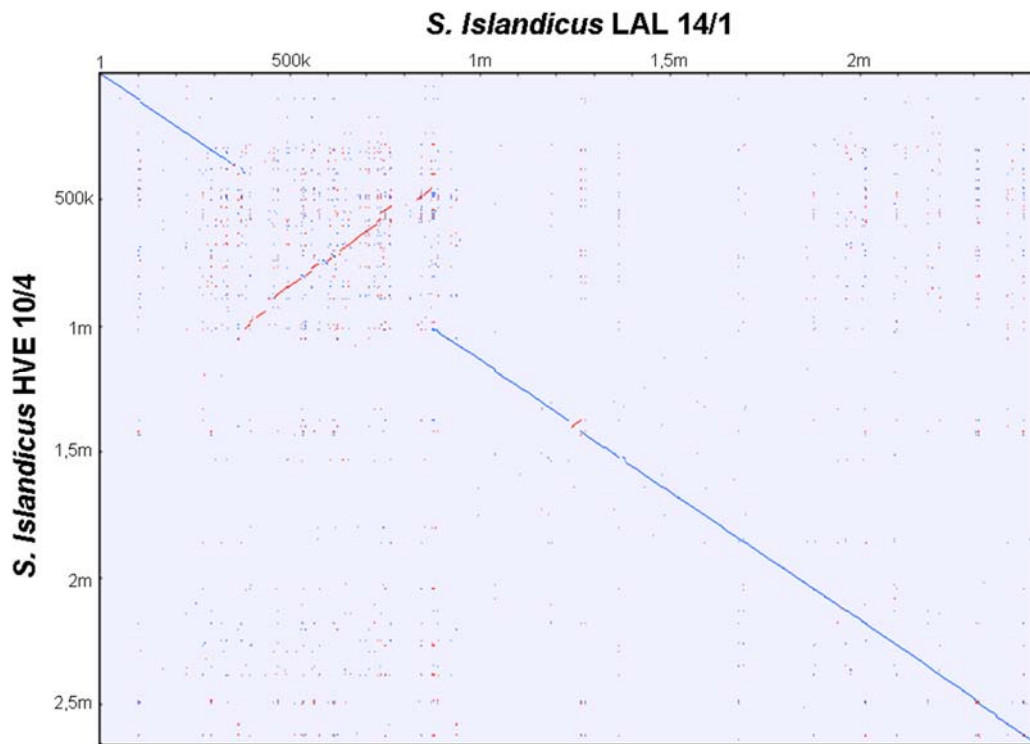
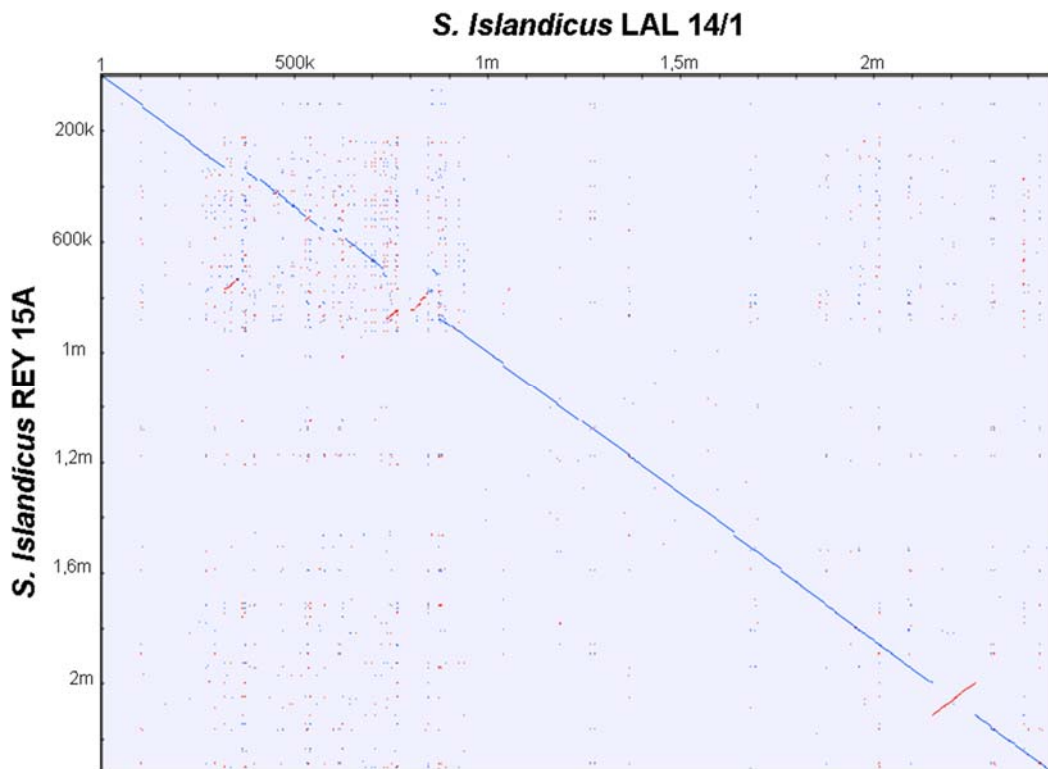


Figure S3 . Representation of the word ACTATAGA in the LAL14/1 genome. The word counts are indicating along the genome sequence by gray bars. Red curve represents the skew of the word ACTATAGA; green curve shows the skew of ACCA indicating the positions of the replication origins, marked as circles. The hyper-variable region is represented by a pink rectangle.



A.



B.

Figure S4. Dot-plot comparison of (A) *S. islandicus* LAL14/1 vs HVE10/4 and (B) LAL14/1 vs REY15A.
Commented in Table S4.

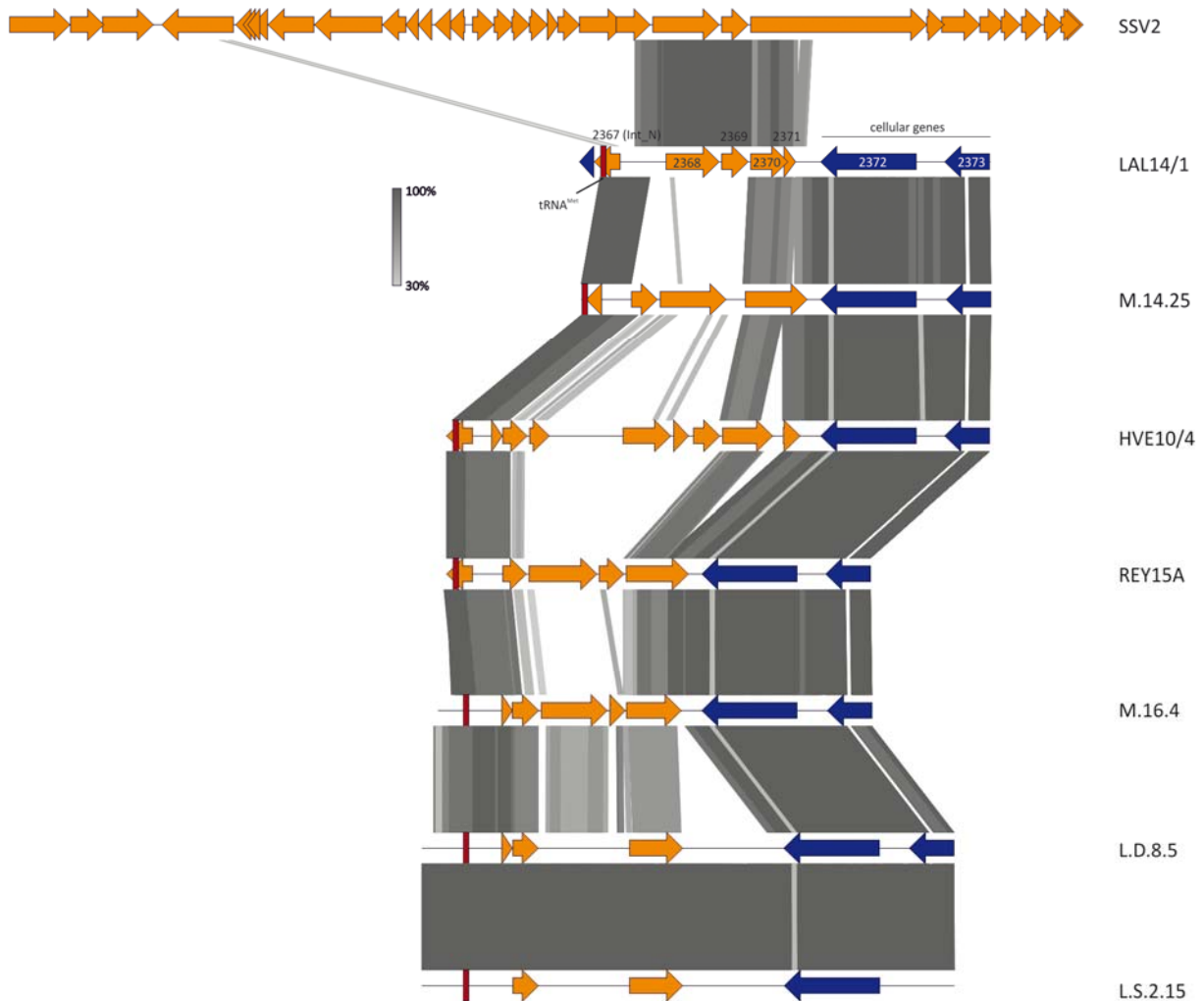


Figure S5. Remnants of fuselloviral genomes in different *S. islandicus* strains. Equivalent genomic loci of selected *S. islandicus* strains encompassing remnants of SSV-like proviruses are aligned; linearized genome map of fusellovirus SSV2 is also included in the alignment. Pairwise tblastx hits between the corresponding loci are illustrated by different shades of grey (the identity scale is included in the figure). Genome accession numbers of the depicted *S. islandicus* strains are provided below the corresponding strain names. Homologues of fuselloviral genes, blue arrows; non-viral genes, red arrows; tRNA Thr [GGT] genes, green rectangles.

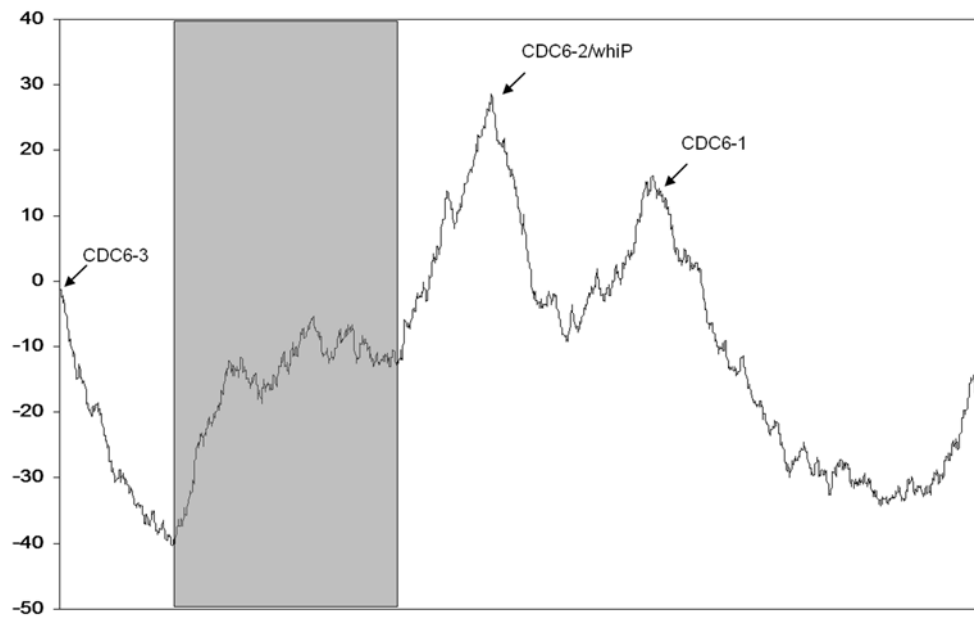
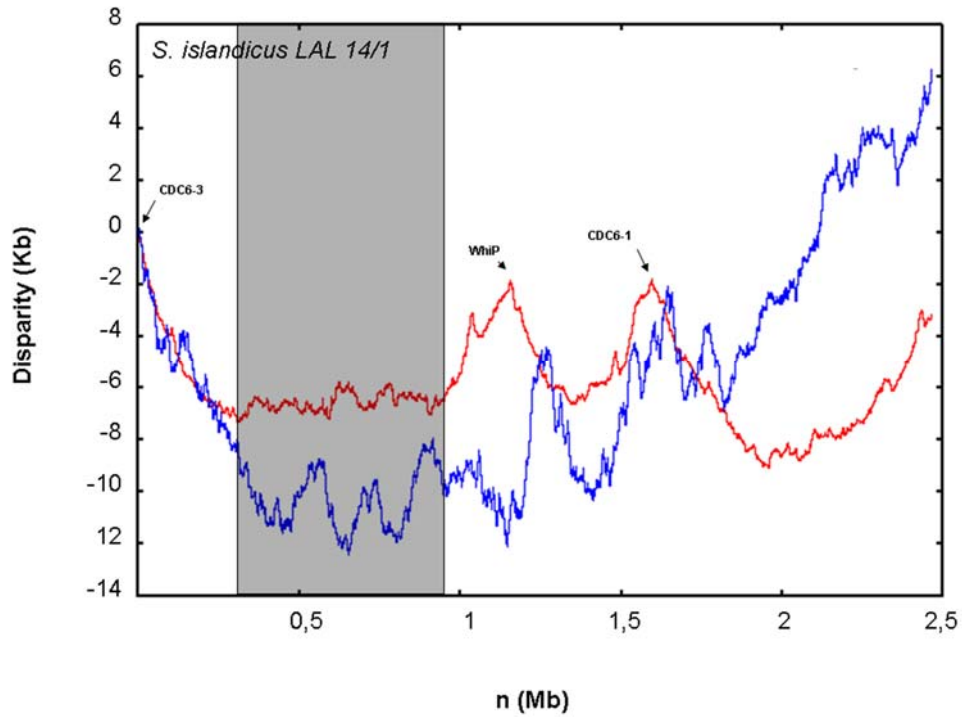


Figure S6. Identification of origins of replication in *S. islandicus* LAL14/1 genome by (A) approach Z-curve; (B) ACCA plot. Arrows indicate the positions of some conserved genes linked to three *oriC*. The position of long variable region is indicated by shadow rectangle.

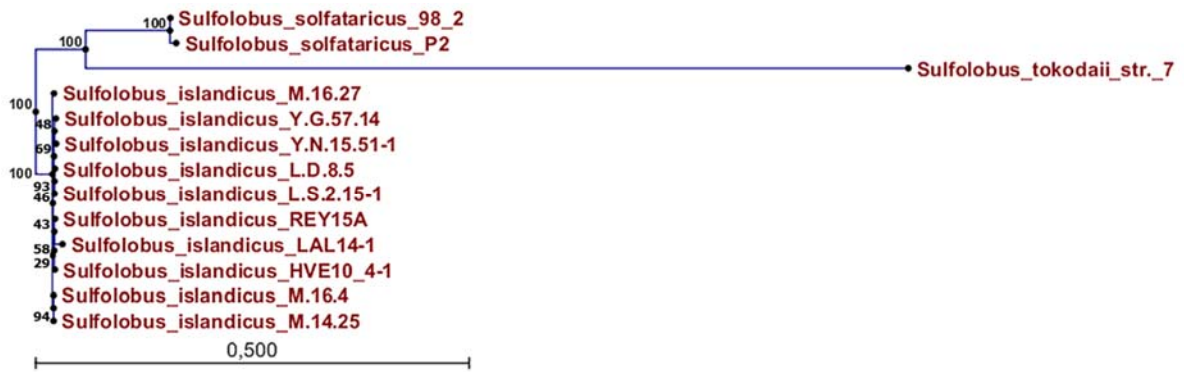


Figure S7. Phylogenetic tree of the proteins of Ups family in *Sulfolobales*

Annexe 2. Guillière *et al.*, PlosOne - Supplementary figures

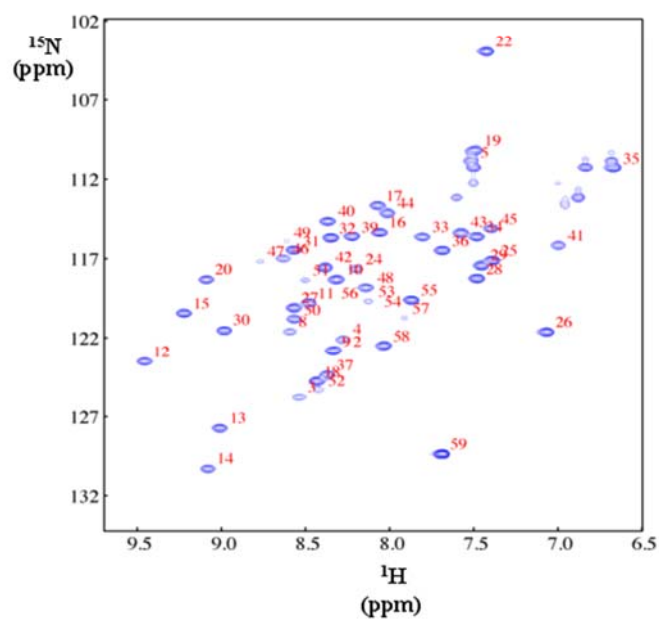


Figure S1: Assigned ^1H - ^{15}N HSQC spectrum of AFV1p06 obtained at 25 °C in buffer E. Assignments are indicated in red. Assignments of side-chain amide groups are not shown.

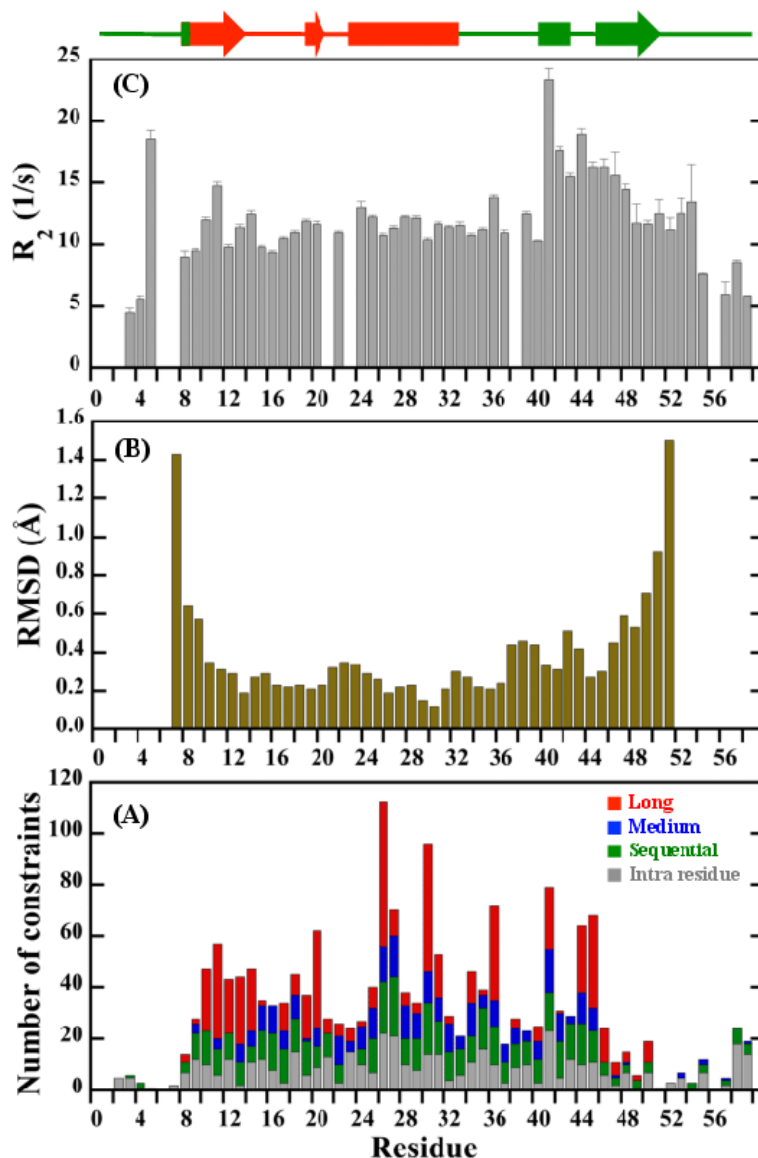


Figure S2: Number of nOe-derived distance constraints per residue used for structure calculations **(A)**; mean pairwise root mean square deviation (RMSD) of backbone atoms (N,C,O,CA) of the structure ensemble **(B)**; amide ^{15}N transverse relaxation rate (R_2) **(C)**. The intra-residue, sequential, medium and long range distance constraints are colour coded in **(A)**. In **(B)**, the RMSD of residues 1-6 and 52-59, which are disordered, are not shown. The secondary structure is displayed on the top of the figure, with helices represented as rectangles and strands as arrows. The ZNF region is displayed in red and the rest of the protein in green.



Figure 53: Phylogenetic tree of the distribution of the AFV1p06-like ZNF fold in *Eukarya* and *Archaea*. The tree was built from a psi-blast search of the NCBI database using the alignment of the AFV1p06-like ZNF motif (presented on Fig. 4). Although the separation on two sister groups *Crenarchaea* and *Eukarya/Euryarchaea* was strongly confirmed by the boot strap value, it cannot be excluded that the ZNF proteins found in *Eukarya* and *Euryarchaea* form a single phylogenetic clade.