Manuscrit version beta

Université de Bretagne Occidentale
U.F.R Sciences

THESE

Présentée pour obtenir le grade de

Docteur en Sciences

De l'Université de Bretagne Occidentale

Spécialité: Microbiologie

par

Mathieu GONNET

Génomique comparative des éléments génétiques de Thermococcales, un Ordre d'Archaea hyperthermophiles: Diversité et plasticité des génomes

Soutenance le 5 décembre 2008
devant le jury composé de:

FORTERRE Patrick	Rapporteur
SIMONET Pascal	Rapporteur
LE ROMANCER Marc	Examinateur
JEBBAR Mohamed	Examinateur
HUBLER Frédérique	Examinateur
DE ERAUSO Gaël	Invité
PRIEUR Daniel	Directeur de thèse

REMERCIEMENTS

Le travail présenté a été réalisé à l'Institut Universitaire Européen de la Mer avec le soutien financier du conseil régional de Bretagne. En préambule, je souhaite dédier ce travail en mémoire de Wolfram Zillig, un pionnier de l'étude des *Archaea*. Il a toutefois su transmettre sa passion et sa curiosité afin que des générations d'étudiants et de scientifiques prennent la suite de ses travaux fondateurs.

Je tiens tout d'abord à remercier Daniel Prieur, mon Directeur de Thèse, de m'avoir donné l'opportunité de faire une thèse sur un sujet passionnant, de m'avoir fait confiance tout au long de ces années, de m'avoir encouragé dans mes travaux et de m'avoir permis de les présenter à différents congrès internationaux. Je lui suis également très reconnaissant de m'avoir donné l'occasion de participer à la campagne océanographique Sweep Vents, invité par Ken Takaï du JAMSTEC. Je le remercie particulièrement de m'avoir offert l'extraordinaire expérience d'une plongée à 2000m dans le Pacifique Sud m'ayant permis d'observer les cheminées hydrothermales à travers le hublot du bathyscaphe Shinkaï6500 et de découvrir le biotope du grand Cthulhu. A ce titre je remercie les membres du LM2E avec qui j'ai partagé cette expérience : Jean-Louis pour sa bonne humeur, ses coups de gueule et son grand cœur, Joël pour ses discussions passionnantes et Valérie pour toutes les pauses que nous avons partagées. Je souhaite également avoir une pensée posthume envers Jacques Piccard, l'explorateur des abysses qui nous a quitté durant la rédaction de ce manuscrit.

Je voudrai également remercier Marc le Romancer, mon responsable scientifique, d'avoir canalisé mon énergie et de m'avoir impliqué dans des projets tels que l'ANR Modulome, m'ayant permis de travailler avec les informaticiens de l'équipe Symbiose de Jacques Nicolas de l'INRIA de Rennes dans le cadre de l'étude des CRISPR, mais également dans l'ACI DHV Genoarchaea qui m'a permis de collaborer avec P. Forterre et H. Van Tilbeurgh. Je remercie également Marc de s'être démené afin de m'obtenir des vacations d'enseignants au sein de l'université et de l'IUT m'ayant par la suite permis d'obtenir un poste d'Attaché Temporaire d'Enseignement et de Recherche.

Je remercie sincèrement Patrick Forterre et Pascal Simonet d'avoir accepté d'être les rapporteurs de ce manuscrit. J'avais eu la chance de rencontrer ces deux personnes charismatiques durant mon cycle universitaire prédoctoral et je suis très touché qu'ils aient accepté de juger ce travail. Patrick m'avais transmis le « virus » des *Archaea* lors de mes études à l'université Paris XI en 2002 et Pascal m'avait fait forte impression lors d'une conférence durant mon stage de DEA à

l'université Claude Bernard en 2003. Je remercie également Mohamed Jebbar et Frédérique Hubler d'avoir consenti à apporter leur examen critique sur ce manuscrit.

Finalement, je souhaite remercier Gaël de Erauso de m'avoir transmis sa passion des Thermococales et des éléments génétiques. Tu as réussi à me transmettre le « virus » et les compétences pratiques m'ayant permis de réaliser toutes ces expériences, de la culture de ces odorants micro-organismes en passant par l'extraction de leurs éléments génétiques et finalement utiliser une stratégie adaptée de séquençage. En dehors de nos longues discussions et de la multitude de publication dont tu abreuvais ma boite mail, je te suis infiniment reconnaissant de m'avoir inculqué les bases de la biologie moléculaire « old school », m'ayant permis de ne pas être un biologiste moléculaire « en kit » mais de devenir un vrai bricoleur d'ADN. A ce propos je pense qu'il était peut-être un peu extrémiste de séquencer l'intégralité du plasmide pAMT11 avec un antique séquenceur sur gel de type Licor… J'y ai gagné plus qu'un mentor, un véritable ami. Pour rester dans les compétences pratiques, je remercie également Adeline Toffin et Nadège Bienvenue de m'avoir appris certaines techniques de culture et d'avoir fait en sorte que le matériel nécessaire soit toujours à notre disposition, en stock suffisant et bien rangé dans le laboratoire.

J'adresse un grand merci aux différents stagiaires que j'ai eu l'honneur d'encadrer. Ludo le Golfeur, de l'IUT de Quimper, Bao le cuistot, Lucile la luciole, Elodie Barbie, Olivier le bucheron de Master 1 et Charlotte la pochette et Saïd Lebanon de Master 2. J'espère vous avoir transmis le virus des éléments génétiques et des échanges de gènes. De votre côté, vous m'avez tous fournis une aide précieuse, notamment ceux qui ont passés des semaines à faire des miniprep…

Je profite également de ces pages pour adresser mes plus vifs remerciements aux doctorants avec qui j'ai partagé le bureau… et parfois bien plus… Mélu, Hélène, Raja, Aurore, Pauline, Erwan, Nathalie, Anne, Zey, Maria ; ainsi que tous les membres du LM2E avec qui j'ai partagé la paillasse : Karine, Frédérique, Stéphane, Christian, Stéphane, Claire ; et les petites secrétaires qui ont fait en sorte que les bons de commandes transitent efficacement dans les rouages administratifs : Anne-So, Joëlle, Stéphanie et Christine. Je souhaite également remercier tous les autres membres du laboratoire qui n'auraient pas été cités : Didier, Ghislaine, Adeline, Patricia, Anne, Marianne, Jean-Paul, Audrey, Julien, Laurent, Cassandre.

Il me tient également à cœur d'adresser ma gratitude aux plateformes techniques prestataires de services qui m'ont été d'une grande aide au cours de ces travaux. Tout d'abord pour le séquençage : Morgan et Erwan de la plateforme de séquençage Ouest génopole, Richard Reinhardt du Max Planck Institute, mais également Catherine Chevalier, Isabelle et Remi

Houlgatte pour la création du prototype de puce à ADN de la plateforme transcriptomique de Nantes.

En dehors du laboratoire, j'ai trouvé au sein de l'IUEM de nombreuses personnes qui m'ont rendu le quotidien plus joyeux ; tout d'abord le LEMAR crew : Yannelle, Tata, Odile de Raie Monette, Czamczam, Kub, Carlito, Poulitos, Jona, Pierro, Pierru, Briva, Sorka, Fanny, Morgana, Fred, Mado, Manon, Remy, Hansy, Agnès, Manon et Maud… Au-delà de mon infinie reconnaissance pour vos corrections orthographiques apportées à ce manuscrit, vous avez toujours écoutés avec attention mes divagations scientifiques assez éloignées de vos préoccupations premières et m'avez permis de ne pas trouver trop longues les soirées passées à l'institut, vous m'avez remonté le moral et vous m'avez surtout beaucoup fait rigoler (vous et vos gâteaux au chocolat). Je souhaite également remercier tous les titulaires du LEMAR qui m'ont parfois considéré comme un membre de leur laboratoire en m'invitant à tous les pots et en particulier Christine, Marcel, Philippe, Fred, Martial, Christophe et Dario. Afin de finir le tour de l'institut, je remercie le personnel de l'accueil : Joëlle, Yannick et le veilleur qui est souvent venu prendre un café durant ronde nocturne…

Ces remerciements touchent à leur fin mais je voudrais remercier les amis non brestois qui m'ont soutenu dans cette aventure, car sans les Chorococos je ne suis rien : Teub, H, Gloubidio, Proutati, Vincouille, Anatole Bocal, Fredule, Geulafioul, Bubuzz, Aody, Laman, Jujubebed, Stephounet, Sihaya, Corksinet, Védé, Drac et Darkmorticious… Je remercie également mon lecteur mp3 qui a été mon plus fidèle compagnon de paillasse et de bureau, distillant du Krush, Cam, Shadow, Manuva, Flake, Gainsbourg, RJD2, Amon Tobin, Depth Affect, Funki Porcini, Bonobo, Beastie Boys, Burial, Cut Chemist, Fumuj, Herbaliser, Ez3kiel, High Tone, Svinkels…

Je voulais également remercier sincèrement toute ma famille pour le soutien sans faille qu'elle m'a toujours témoigné au cours de cette thèse et d'avoir accepté de ne pas me voir revenir souvent du bout du monde pour assister aux réunions familiales ; désolés cousinettes de ne pas être venu à vos mariage.

Finalement je remercie toutes les personnes avec qui je n'ai échangé ne serait-ce qu'un sourire au détour d'un couloir… ce simple geste a bien souvent été suffisant pour ensoleiller le ciel brestois…

Manuscrit version beta

Université de Bretagne Occidentale
U.F.R Sciences

THESE

Présentée pour obtenir le grade de

Docteur en Sciences

De l'Université de Bretagne Occidentale

Spécialité: Microbiologie

par

Mathieu GONNET

Génomique comparative des éléments génétiques de Thermococcales, un Ordre d'*Archaea* hyperthermophiles: Diversité et plasticité des génomes

Soutenance le 5 décembre 2008
devant le jury composé de:

FORTERRE Patrick	Rapporteur
SIMONET Pascal	Rapporteur
LE ROMANCER Marc	Examinateur
JEBBAR Mohamed	Examinateur
HUBLER Frédérique	Examinateur
DE ERAUSO Gaël	Invité
PRIEUR Daniel	Directeur de thèse

Liste des Tableaux

Liste des figures

SOMMAIRE

INTRODUCTION

I. Objectif général

Les conditions extrêmes dans lesquelles se développent les *Archaea* hyperthermophiles rendent ces organismes particulièrement intriguants.

D'un point de vu fondamental, la principale question soulevée concerne la stabilité de leurs macromolécules en conditions extrêmes de température, de pH, de pression, mais aussi de radiations ionisantes. De plus, la position phylogénétique de ces organismes soulève encore de nombreuses interrogations sur leur relation avec le progénote (premier ancêtre universel ou LUCA).

D'un point de vu industriel, les hyperthermophiles suscitent également un fort intérêt en raison d'un formidable potentiel génétique, source d'applications biotechnologiques à haute valeur ajoutée ou industrielle, telles que les biocatalyseurs thermostables.

Cependant, les programmes de séquençage complet de génomes ne permettent d'élucider qu'une infime partie des fonctions encodées par ces gènes. Actuellement, 69 génomes complets d'*Archaea*, dont celui de 4 espèces de Thermococcales sont disponibles. Parmi ces milliers de gènes, une part très importante ne possède pas d'homologue dans les banques de données. Ils représentent de nouvelles familles de gènes dont le rôle reste à déterminer afin d'apprécier l'étendue de la diversité et des mécanismes adaptatifs qui permettent le maintient de la vie en environnements extrêmes. Cette tâche est pour le moment encore à ses balbutiements en raison du manque d'outils génétiques appropriés pour ce groupe de micro-organismes.

La comparaison de ces génomes montre également l'importance évolutive des réarrangements chromosomiques et la présence de nombreux gènes transférés horizontalement, probablement hérités de la virosphère. Alors que les recherches portant sur les éléments génétiques des halophiles extrêmes, des méthanogènes et des thermoacidophiles des environnements terrestres chauds ont bien progressé au cours de la dernière décennie, très peu d'études ont été consacrées à ceux des Euryarchaea hyperthermophiles représentant pourtant la plus grande abondance et la plus large diversité chez les hyperthermophiles.

A l'heure actuelle, chez les Euryarchaea, seul un virus infectant la souche *Pyrococcus abyssi* GE 23 a été caractérisé. Ces particules virales en forme de citron renferment un génome de 18,1kb dont

le génome vient d'être séquencé. La plupart de ses ORFs ne possède pas d'homologue dans les bases de données, ils sont dits **orphelins**. Ce virus a la particularité d'établir une relation d'équilibre avec son hôte en se maintenant sous une forme typiquement plasmidique à faible nombre de copies. Cette particularité souligne l'étroite relation existant entre les plasmides et les virus, qui sont définis comme des éléments génétiques.

Les études sur les plasmides de Thermococcales ont, à ce jour, seulement permis la caractérisation de trois petits plasmides à réplication par cercle roulant cryptiques, stables et multicopies (20-30 copies par cellule). pGT5 (3,44 kb) isolé de la souche *Pyrococcus abyssi* GE5, pRT1 de la *Pyrococcus* sp. RT1 et pTN1 de *Thermococcus nautilii*. La découverte du petit plasmide cryptique pGT5 dans une souche de *Pyrococcus abyssi* GE5 a cependant permis d'initier un travail destiné à la mise au point de la première génération de vecteurs de clonage pour les Euryarchaeota hyperthermophiles. Le vecteur navette pYS2 a été construit en utilisant une partie d'un plasmide d'*E. coli*, une partie de pGT5 et le gène *pyrE* comme marqueur de sélection pour transformer des mutants auxotrophes pour l'uracyle (Watrin et al., 1999). Ces outils sont indispensables pour les études génétiques jusqu'alors impossibles chez les Thermococcales (Lucas et al., 2002). Ces outils génétiques ont depuis été améliorés par les groupes de Reeves et d'Imanaka, notamment par l'ajout d'un gène de résistance à l'antibiotique simvastatin.

Parmi les facteurs influençant la dynamique et l'évolution des populations microbiennes, les éléments génétiques (plasmides, transposons, virus) sont en effet des acteurs prépondérant car ils permettent la vectorisation de matériel génétique et donc une acquisition horizontal d'ADN. L'étude de ces entités fournit de bons modèles de comprenhension des processus fondamentaux comme la réplication, la propagation et l'évolution, mais aussi pour révéler l'existence de transferts d'informations génétiques par flux de gènes. Plasmides et virus ont été décrits dans les trois grands groupes d'*Archaea* : méthanogènes, halophiles extrêmes et hyperthermophiles, mais en nombre très limités, particulièrement pour le dernier groupe. Très peu de choses sont donc connues sur les fonctions encodées par ces éléments, notamment par un manque de données génomiques. Par définition, ces éléments sont capables de se répliquer de façon autonome et pour certains de se propager efficacement au sein d'une population microbienne donnée. Il est probable que ces éléments contribuent de façon significative aux transferts génétiques horizontaux. D'autre part, de récentes études ont révélé qu'à l'instar des génomes bactériens, ceux des archées contiennent des copies intégrées de virus ou de plasmides. Ils contribuent activement à la plasticité du génome en facilitant les réarrangements

chromosomiques, ainsi qu'aux transferts horizontaux par capture et propagation des gènes. L'ordre des Thermococcales représente un modèle d'étude chez les *Archaea.* En effet, il constitue une part prépondérante (en termes d'abondance et d'activité) et très diversifiée (plus de 40 espèces qui en font l'ordre le plus important) dans la composante hétérotrophe des écosystèmes microbiens hyperthermophiles. Le laboratoire dispose d'une collection originale de Thermococcales, isolées à partir d'échantillons prélevés sur diverses sources hydrothermales océaniques profondes des océans Atlantique, Pacifique et Indien. A ce jour, les petits plasmides cryptiques de *Pyrococcus* ne reflètent pas la diversité des éléments génétiques au sein de cet Ordre d'*Archaea*.

Ce travail de thèse consiste à élargir nos connaissances sur la diversité des éléments extrachromosomiques des Thermococcales, à appréhender leur diversité et à estimer leur contribution à la plasticité des génomes de leurs hôtes, notamment comme moyen d'adaptation face aux variations environnementales.

Le criblage de nombreux isolats de Thermococcales, d'origines géographiques variées, permettra la réalisation d'une banque de plasmides. Les plasmides seront caractérisés par la détermination de leur taille, l'établissement de leur profil de restriction et l'estimation de leur nombre de copies. Une classification préliminaire de ces plasmides sera proposée sur la base sur leurs origines géographiques, la comparaison de leurs profils de restriction et d'hybridations croisées ADN-ADN. Les génomes les plus intéressants (du point de vue de la biogéographie, de la taille et des homologies de séquence) seront séquencés.

L'analyse des séquences plasmidiques devrait apporter d'intéressantes informations sur la biologie de ces éléments, sur leur capacité à participer aux flux de gènes et à influencer l'évolution des génomes de leurs hôtes mais également leurs inter-relations existantes avec le monde viral.

Les séquences seront exploitées de façon aussi exhaustive que possible en mettant en œuvre des méthodes d'analyse bioinformatique afin de rechercher les structures impliquées dans les fonctions de maintenance, de réplication et de propagation. Les séquences plasmidiques seront ensuite comparées entres-elles et aux banques de données afin de caractériser ces éléments.

II. Les plasmides

1. Découverte et propriétés générales

La découverte des plasmides remonte aux années 1940. Cette découverte est liée à l'observation du transfert de matériel génétique entre des souches d'*Escherichia coli* (Lederberg *et al.,* 1946). Lederberg suggéra que ces gènes n'étaient pas portés par le chromosome, mais par une molécule d'ADN extrachromosomique qu'il décida d'appeler **plasmide**. Dans les années 70, l'étude et le « bricolage » de l'ADN des plasmides a également permis la création d'une technique révolutionnaire : le clonage moléculaire d'ADN, technique fondatrice de la biologie moderne (Cohen *et al.,* 1973). Aujourd'hui, les plasmides sont considérés comme d'importants acteurs dans le transfert de gènes intra- ou inter-espèces. Nombre d'études s'intéressent à leur contribution dans l'adaptation des micro-organismes aux changements d'environnementaux et dans la spéciation. Les plasmides ont donc un **impact évolutif clé**.

Les plasmides sont universels, ils sont présents dans les trois domaines du vivant. Bien que majoritairement rencontrés chez les procaryotes, certains eucaryotes unicellulaires possèdent ce type de réplicon additionnel, telle la levure *Saccharomyces cerevisiae* et son plasmide 2µ localisé dans le noyau de la cellule (Clark-Walker 1972; Livingston 1977). Ce plasmide a servi de base à la construction de plusieurs vecteurs de clonage (Parent *et al.,* 1985). Cependant, la diversité des plasmides décrits est inégale en fonction du domaine taxonomique. Les eubactéries étant étudiées depuis les années 40, leurs plasmides sont très bien documentés. Inversement, la découverte plus récente des *Archaea* au cours des années 80 et leurs relatives difficultés culturales en font le domaine le moins documenté. De plus, l'avancée des travaux relatifs aux plasmides est très inégale en fonction de chaque groupe taxonomique.

Définition

La fonction minimale d'un plasmide est sa capacité à contrôler sa réplication. Certains plasmides adoptent un style de vie parasitaire, représentant une charge pour la cellule et se comportent alors comme une forme d'ADN égoïste (Dawkins 1976), d'autres sont mutualistes et contribuent à l'adaptabilité de l'hôte à son environnement. Par leurs propriétés à transférer du matériel génétique, certains plasmides ont une influence sur la biologie de leurs hôtes en tant que vecteur, source de nouveaux gènes acquis de transfert horizontal.

Structure

Les plasmides sont des molécules d'ADN bicaténaires circulaires closes, généralement surenroulées négativement. Néanmoins, il existe quelques exceptions à cette définition générale d'un plasmide. Par exemple, des plasmides linéaires possédant des télomères existent chez certaines bactéries du genre *Borrelia* (Barbour *et al.*, 1987), *Streptomyces* (Kinashi *et al.*, 1991) ou *Escherichia* (Vostrov *et al.*, 1992).

La taille des plasmides peut considérablement varier, de 846pb pour le plasmide pRQ7 de la bactérie thermophile *Thermotoga maritima* (Harriott *et al.,* 1994) aux 690kb du méga-plasmide de *l'Archaea* halophile *Haloferax volcanii* DS2 (Ebert *et al.*, unpublished). Néanmoins, pour ces plasmides de grande taille, il est parfois difficile de les différencier des minichromosomes. La taille d'un plasmide limite également les possibilités de mécanismes de réplication de ces réplicons.

On peut distinguer 4 classes de tailles : les petits plasmides (1 à 10kb), les plasmides de taille moyenne (10-50kb), les gros plasmides (50-100kb) et les mégaplasmides. La taille d'un plasmide est souvent inversement proportionnelle au nombre de copies de cet élément dans la cellule hôte. Ils peuvent être présents en un ou deux exemplaires dans la cellule hôte (cas des mégaplasmides) et jusqu'à plusieurs dizaines de copies pour les petits plasmides. La taille d'un génome plasmidique va également, dans certaines mesures, conditionner son mode de réplication. Chaque mode de réplication possède des caractéristiques plus ou moins adaptées au plasmide : taille maximale du réplicon, vitesse de réplication et nombre maximal de copies générées.

Les plasmides sont doués de réplication autonome assurant un transfert vertical (de cellule mère à cellule fille) et pour certains un transfert horizontal (d'une cellule à une autre pouvant être des espèces différentes). Un réplicon extrachromosomique peut-être classé dans la catégorie des plasmides s'il ne code pas de gène conduisant à la formation de particules virales. Néanmoins, il est parfois ardu de classer certains réplicons. Les méga-plasmides sont considérés comme des « mini-chromosomes » additionnels au chromosome principal, en particulier chez les bactéries du genre *Rhizobium* (Hogrefe *et al.,* 1984; Margolin *et al.,* 1993) et chez les *Archaea* halophiles. De façon similaire, un génome viral ayant perdu sa capacité à produire des particules virales, reste capable de se maintenir dans le cytoplasme sous forme épisomale ou prophagique. A ce titre, tout au long de ce manuscrit de thèse sera utilisé **le terme générique de plasmide pour désigner les génomes extrachromosomiques de Thermococcales étudiés**. Aucune preuve ne peut être apportée que ces génomes extrachromosomiques aboutissent à la production de particules virales dans certaines conditions particulières physicochimiques de culture.

2. Origine des plasmides

La propriété des plasmides à transférer de l'ADN et leur abondance en grand nombre de copies dans la cellule a tout d'abord suscité l'intérêt pour le développement d'outils génétiques. En parallèle du côté « pratique », des études ont permis de formuler des hypothèses sur l'origine et l'évolution des plasmides et d'apporter des informations sur l'origine de l'ADN. L'analyse comparée des réplicons viraux et plasmidiques montre une certaine conservation, aussi bien pour les simples réplicons minimaux que pour les réplicons les plus évolués et complexes. L'éventuelle association de protéines fixées à l'ADN augmente l'efficacité de cette perpétuation moléculaire. La survie de la cellule assurant la survie du plasmide, la séquestration des gènes par un plasmide intervient quand l'environnement affecte négativement son hôte. Ce mécanisme de séquestration est fondamental, il est la pierre angulaire de la fabrication et de la propagation de macromolécules de taille croissante. Le niveau de sophistication des plasmides augmente avec l'addition de nouveaux gènes tels que ceux conférant à l'hôte la capacité d'occuper un environnement spécifique normalement inhospitalier pour la cellule.

Patrick Forterre a émis l'hypothèse que l'ADN aurait pu être inventé par les virus dans le monde à ARN (Forterre 2001). En effet, dans le conflit qui oppose le virus à la cellule, l'acquisition de l'ADN aurait présenté un avantage sélectif au virus, lui permettant de résister aux enzymes de dégradation des cellules (enzymes de restriction). Plusieurs arguments confortant cette hypothèse sont proposés : par exemple le fait que les enzymes nécessaires pour transformer le code génétique de l'ARN vers l'ADN n'existent que chez les virus ont été découverts (Warren 1980), tel le virus PBS1, dont le génome est un ADN modifié où l'uracile remplace la thymine (ADN-U),. Ces virus seraient des vestiges de la transition du monde ARN au monde ADN. Les premières cellules à ADN auraient en conséquence emprunté la machinerie virale afin de fabriquer de l'ADN, faisant ainsi des virus des entités très anciennes.

Tout comme l'« invention » de l'ADN, l'origine des plasmides est étroitement liée aux virus. Il est probable qu'un virus défectif puisse avoir été dompté par la cellule, détournant ainsi un réplicon égoïste à son avantage. L'acquisition d'un réplicon indépendant permet à la cellule de s'en servir comme d'un laboratoire d'expérimentation. Certains gènes du chromosome, transférés ou copiés sur plasmide, peuvent subir une évolution accélérée et tester différentes combinaisons génétiques aboutissant à une innovation nécessaire à l'évolution et à l'adaptation de l'hôte cellulaire.

3. Les fonctions codées par les plasmides

Les gènes portés par les plasmides peuvent se regrouper en **trois catégories**. La première catégorie répond à la définition des plasmides, elle comprend les gènes impliqués dans la **réplication**. La seconde est constituée de gènes fréquemment rencontrés conférant en général une plus grande stabilité du plasmide. Ils codent par exemple des systèmes de **maintenance**, de ségrégation, d'intégration ou de régulation du nombre de copies. La troisième catégorie correspond à des gènes **accessoires**.

3.1 *Fonction minimale : Réplication*

Par définition, la fonction minimale d'un plasmide est sa capacité à se répliquer. Certains plasmides ne possèdent qu'un seul gène, à l'image du plus petit plasmide connus à ce jour pRQ7 (846pb) de *Thermotoga* sp. RQ7 (Yu *et al.,* 1997). Ce patrimoine génétique est suffisant pour assurer la multiplication du plasmide en un grand nombre de copies au sein du cytoplasme et ne nécessite pas de mécanisme de ségrégation. La partition aléatoire des 300 copies du génome est suffisante pour assurer sa présence dans les cellules filles au cours des divisions cellulaires.

Le mode de réplication d'un plasmide est très important. Etant considéré comme la fonction minimale, c'est le principal critère de classification des plasmides. Les plasmides peuvent être classés en superfamilles en fonction du mode de réplication : les plasmides à réplication par cercle roulant, par déplacement de brin et selon le mode thêta. Le mode de réplication rencontré conditionnera également les mécanismes de ségrégation et de régulation du nombre de copies.

3.1.1 Réplication par cercle roulant

La réplication par cercle roulant (RCR) a été découverte il y a une vingtaine d'années, d'abord sur des plasmides de bactéries à Gram négatif, puis chez les bactéries à Gram positif et finalement chez les *Archaea*. La RCR permet la synthèse rapide d'un grand nombre de copies de molécules d'ADN circulaire. Ce mode de réplication se rapproche de celui observé chez certains bactériophages à ADN simple brin mais également de certains transposons de type II. La RCR nécessite trois éléments fonctionnels, une protéine Rep et deux origines de réplication, simple (*sso*) et double brin (*dso*). La principale caractéristique de la RCR est un découplage dans la réplication des deux brins d'ADN aboutissant à une réplication asynchrone et séquentielle. Le brin sens est répliqué dans un premier temps, vient ensuite réplication du brin antisens (Figure 1).

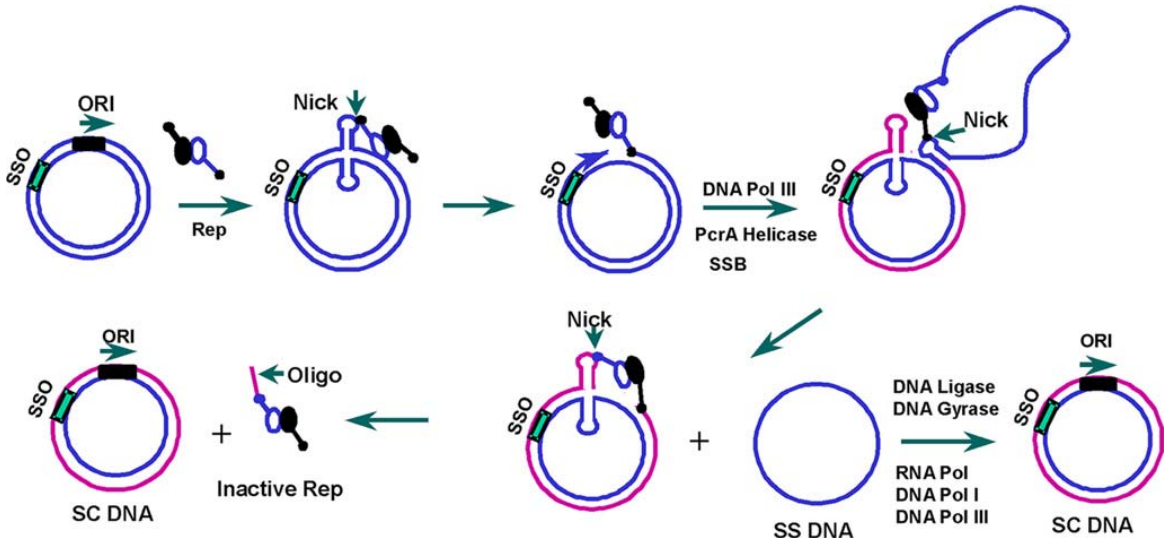


Figure 1 Réplication par cercle roulant. (D'après Khan, 2005)

Schématisation de la réplication par cercle roulant. L'ADN matrice est en bleu, le brin néosynthétisé est en violet. Les deux origines de réplication sont représentées par des rectangle, vert pour la *dso* et noir pour la *dso*.

Protéines Rep de type RCR

La réplication débute lorsque la protéine initiatrice Rep, codée par le plasmide, se fixe sur l'ADN au niveau du site appelé « origine de réplication double brin » (*dso*). Rep est une protéine à activité endonucléase-ligase site-spécifique, possédant des motifs conservés (Ilyina *et al.,* 1992). Elle coupe l'un des deux brins au niveau de la *dso*, libérant une extrémité 3' hydroxyle qui servira d'amorce à l'ADN polymérase III. La *dso* est une séquence d'environ 100pb contenant un site de fixation *bind* de la protéine Rep et un site de clivage simple brin appelé *nick* (Figure 2). Des séquences répétées inversées s'organisent en structures secondaires de type épingle à cheveux et cruciformes. Ces structures jouent un rôle important dans le recrutement de la protéine Rep (Figure 1).

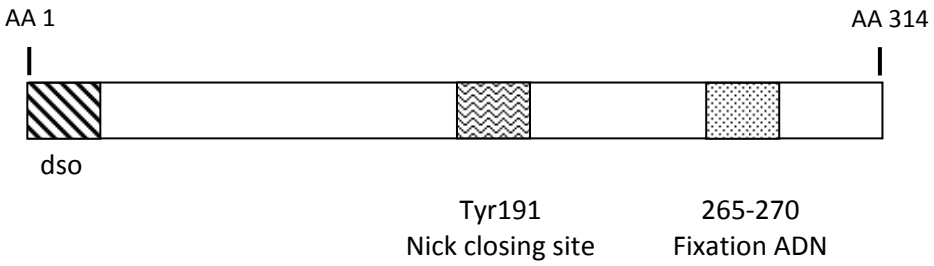


Figure 2 Organisation d'une protéine Rep.

Localisation des deux motifs fonctionnels de coupure et de fixation à l'ADN et de l'origine de réplication double brin (*dso*)

Il existe une étroite relation entre la séquence *dso* et la protéine Rep. L'analyse de leurs séquences révèle une coévolution entre ces deux éléments (Guglielmetti *et al.,* 2007). Agissant de concert, la protéine Rep et la séquence nucléotidique *dso* peuvent être considérés comme un module.

Après clivage, la protéine Rep reste fixée de façon covalente à l'extrémité 5' phosphate de l'ADN par une liaison phosphotyrosine (Figure 1). La réplication se déroule le long de l'ADN circulaire, déplaçant l'ADN coupé en molécule simple brin (brin leader). La synthèse continue d'ADN produit plusieurs copies du génome concaténées sous forme d'un intermédiaire simple brin. Ces concatémères linéaires de plasmides sont ensuite convertis en ADN double brin. La protéine Rep clive le brin leader simple brin au niveau d'une séquence nommée origine de réplication simple brin (*sso*) située dans une région non codante. Cependant, de nombreux plasmides ne possédant pas de *sso* ont été isolés ou créés artificiellement. Ces plasmides accumulent un grand nombre de copies simple brin circulaire et un faible nombre de copies double brin (Mendes *et al.,* 2000).

L'ADN polymérase III, de l'hôte, réplique ensuite l'ADN à partir de l'origine simple brin, réalisant un tour complet de la molécule afin de produire une molécule double brin. L'ADN polymérase I retire ensuite l'amorcegrâce à son activité 5'->3' exonucléase afin de le remplacer par de l'ADN. Au final, une ADN ligase se charge de ligaturer le fragment d'ADN néosynthétisé linéaire aboutissant à la formation d'une molécule double brin circulaire (Figure 1).

Les protéines de l'hôte jouent un rôle important dans la RCR, à l'exemple des ADN polymérases I, III, de l'ADN ligase et de l'ARN polymérase. D'autres protéines peuvent participer à la réplication par cercle roulant selon des mécanismes plus ou moins élucidés. Des hélicases, telles que PcrA et UvrD, possédant la bipolarité de débobinage de l'ADN 5'->3' et 3'->5' peuvent se fixer à la protéine Rep après coupure au niveau de la *dso*. Des protéines SSB (Single Strand Binding), se fixant à l'ADN simple brin interviennent également pour stabiliser les intermédiaires simples brins.

Les analyses phylogénétiques, basées sur les séquences des protéines Rep et des *dso,* ont permis de classer les réplicons de type RCR en 17 classes définis au sein de la base de données DPR Databse of Plasmid Replicons (www.essex.ac.uk/bs/staff/osborn/DPR_home.htm). Chaque classe de réplicons reflète une origine commune de protéine Rep. Bien qu'il existe des relations évolutives entre ces différentes classes, il est impossible de savoir si ce mécanisme de réplication est issu d'un unique ancêtre commun.

3.1.2 Réplication par mécanisme thêta

Le mécanisme thêta est plus classique. Il est également utilisé par les chromosomes circulaires bactériens et archéens. Il implique la séparation des brins parentaux par une **hélicase**, la synthèse d'une amorce ARN (ARNp) et l'initiation de la réplication par extension de cette amorce. La synthèse est continue sur le brin leader et discontinue sur le brin retardé. La synthèse d'ADN commence à partir d'une ou plusieurs origines de réplication et peut se faire uni- ou bidirectionnellement. Presque tous les plasmides affiliés à ce mode de réplication nécessitent une protéine Rep initiatrice de la réplication (à ne pas confondre avec les protéines Rep de réplication RCR) et l'ADN polymérase I durant les premières étapes de synthèse du brin leader.

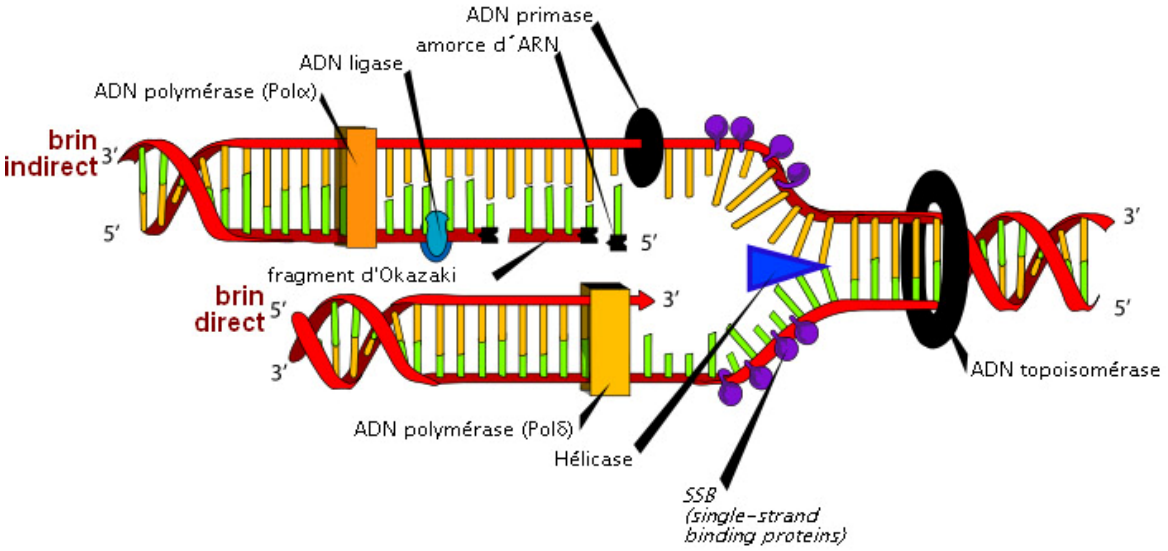


Figure 3 Machinerie de réplication théta

Représentation schématique d'une molécule d'ADN en cours de réplication et des différents partenaires protéiques intervenants dans la réplication de type théta chez les Eucaryotes. Figure adaptée de Maria Ruiz sous licence Creative Commons.

Origines de réplication de type théta

L'origine de réplication d'un plasmide se définit comme (i) la région minimale agissant en *cis* supportant la réplication autonome du plasmide, (ii) la région où les deux brins de l'ADN se séparent pour initier le processus de réplication ou (iii) le site de départ de la synthèse du brin leader. Ces origines de réplication contiennent les sites requis pour les interactions avec les protéines codées par le plasmide et/ou le chromosome de la cellule hôte, en particulier avec la

protéine initiatrice de la réplication. Elles sont généralement riches en AT et comportent des séquences répétées en tandem appelées itérons, espacées par un multiple de 11pb correspondant à la périodicité de l'hélice d'ADN. La présence d'abondantes méthylations joueraient un rôle post-réplicatif, notamment dans les mécanismes de correction des mésappariements (Liang *et al.,* 2002).

Les protéines Rep de type thêta

Elles se caractérisent par un domaine de fixation à l'ADN (HTH αHélice-Tour-αHélice) et un domaine d'interaction protéine-protéine (Leucine Zipper) permettant la dimérisation de Rep.

3.1.3 Réplication par déplacement de brin

Le plasmide à réplication par déplacement de brin le mieux documenté appartient à la classe d'incompatibilité IncQ. Son modèle est le plasmide RSF1010 isolé d'un *Pseudomonas* (Chakrabarty 1976). La réplication fait intervenir 3 protéines RepA, RepB, RepC codées par le plasmide, codant respectivement une hélicase 5'->3', une primase et une protéine d'initiation de la réplication (Scherzinger *et al.,* 1984).

Origines de réplication pour réplication par déplacement de brin

L'origine de réplication est définie comme la région minimale du plasmide RSF1010 nécessaire à l'initiation de la réplication par des protéines apportées en *trans* par un plasmide auxiliaire. Elle comporte une région de 174pb contenant une portion GC riche (28pb), une région AT riche (31pb) et trois itérons de 20pb servant de site de fixation à la protéine RepC. Deux palindromes sont également présents et forment une structure secondaire en épingle à cheveux reconnue par RepB, la primase codée par le plasmide.

Mécanisme de la réplication par déplacement de brin

La réplication nécessite également des protéines codées par l'hôte DnaA, DnaB, DnaC, DnaG, l'ADN polymérase III et des protéines SSB.

RepC se fixe au niveau des itérons. RepA se fixe à proximité dans la région AT riche et sépare les deux brins d'ADN (l'hélicase de l'hôte DnaN ne peut pas la remplacer). RepB intervient alors pour synthétiser des amorces ARN sur le brin antisens. La synthèse de chacun des brins se fait de manière continue, aboutissant au déplacement du brin complémentaire. L'ADN simple brin, par

l'intermédiaire des régions palindromiques, sert de matrice à la synthèse du brin complémentaire aboutissant à la formation d'ADN double brin.

3.2 *Fonctions communes*

En dehors d'un opéron réplicatif, de nombreux plasmides possèdent des gènes supplémentaires conférant des propriétés plus ou moins spécifiques aux plasmides. Ils codent en général pour des mécanismes de maintenance (partition), des protéines permettant l'intégration dans le chromosome de la cellule hôte ou bien des mécanismes impliqués dans le transfert intercellulaire du réplicon.

3.2.1 La maintenance plasmidique

La maintenance plasmidique regroupe l'ensemble des mécanismes permettant **d'assurer un héritage vertical des plasmides**. Elle est d'autant plus importante lorsque le plasmide ne confère pas d'avantage évolutif à son hôte en environnement sélectif. Les mécanismes de maintenance font intervenir la recombinaison site-spécifique reconstituant un réplicon fonctionnel après réplication (résolution des dimères de plasmide), des mécanismes de partition et d'addiction assurant la perpétuation du plasmide au cours des divisions cellumlaires successives.

La partition du plasmide est sa capacité à se distribuer équitablement entre les deux cellules filles après division cellulaire. La perpétuation du plasmide fait intervenir différents mécanismes en fonction du nombre de copies de l'élément. Les plasmides à haut nombre de copies se répartissent souvent aléatoirement lors de la division cellulaire. En effet, la probabilité qu'une cellule se retrouve sans copie du plasmide est de 2^{-n} où n est le nombre de copies avant la division. Ainsi, un plasmide présent en 10 copies risque d'être perdu toutes les 1024 divisions. Avec 20 copies, cette probabilité tombe à moins d'une chance sur un million (probabilité encore diminuée par l'encombrement volumique des plasmides dans la cellule). Bien que cette perte soit rare, elle confère dans certains cas un avantage évolutif pour la cellule qui n'aura plus à supporter la charge énergétique que représente la réplication d'un nombre trop important de copies du plasmide.

La stabilisation d'un plasmide à faible nombre de copies dans la population qui l'héberge nécessite d'autres mécanismes faisant intervenir des protéines spécialisées. Ces mécanismes sont appelés **actifs**. Ils appartiennent à différentes catégories et peuvent être cumulés sur un plasmide

pour une plus grande efficacité. D'une part, il existe des mécanismes actifs attribuant mécaniquement le plasmide à chacune des cellules filles. D'autre part, il existe des mécanismes passifs opérant une sélection négative des cellules n'ayant pas acquis de plasmides, c'est la mort cellulaire post-segregationelle (*PSK : Post Segregational Killing*)

3.2.1.a Partition par mécanismes actifs

Ces mécanismes actifs font intervenir des systèmes de partition. Un locus *par* code deux protéines (ParA et ParB) agissant en *trans* sur une région du plasmide nommée *parS* ressemblant au centromère des chromosomes eucaryotes (Li *et al.,* 2004).

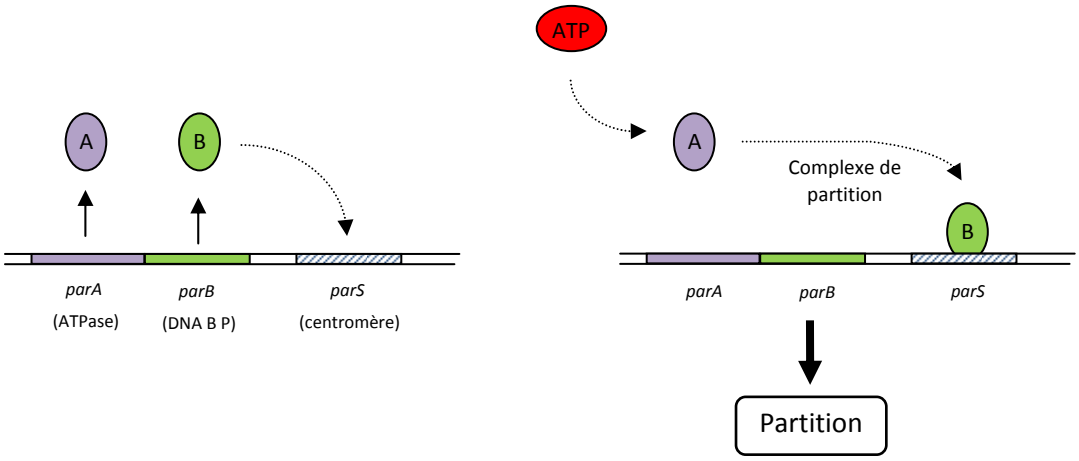


Figure 4 Organisation d'un opéron *par* et implication dans la partition

Le premier gène, *parA*, code une ATPase formant des structures filamenteuses, semblables au cytosquelette des eucaryotes, séparant et distribuant les plasmides aux cellules filles. Deux types de systèmes *par* existent. Les loci de type I codent des ATPases de type Walker alors que ceux de type II codent des ATPases de type actin-like. Le second gène code une protéine de fixation à l'ADN reconnaissant des répétitions directes ou inverse, localisées au niveau de la structure centromère-like. Il y a formation d'un complexe nucléoprotéique sur lequel s'accroche la structure filamenteuse ancrée à la membrane. La ségrégation intervient par migration dans des directions opposées de la cellule de chaque côté du septum.

- Système de partition I

Le système de partition I est le plus fréquent. Il code une protéine ATPase de la famille Walker A (également appelé P-loop). Ces ATPases WalkerA sont sous-divisées en deux groupes : type *Ia* et type *Ib*. Le type *Ia* comprend entre autre la protéine ParA du prophage P1 et SopA du plasmide F. Le type *Ib* inclut ParF du plasmide pTP228 de *Salmonela newport*, ParA des plasmides pTAR d'*Agrobacterium tumefaciens* et pB171 d'*Escherichia coli*. Quelques différences existent entre ces deux types de Walker, notamment la présence d'un domaine HTH à l'extrémité N-terminale impliquée dans la régulation transcriptionnelle de l'opéron *par*.

- Système de partition II

Le système de partition II le plus documenté est celui du plasmide R1. Il fait intervenir l'ATPase ParM (moteur) et la protéine de fixation à l'ADN ParR (répresseur). Le site *cis*-activateur servant de « centromère » est nommé *parC*.

3.2.1.b Partition par mécanismes "passifs" : l'addiction au plasmide

Ces mécanismes augmentent la maintenance plasmidique en tuant les cellules n'ayant pas acquis de plasmide. Ces systèmes sont rencontrés sous les appellations d'*addiction* au plasmide ou de *suicide* post-ségrégationel (PSK post segragational killing). Ils ont été identifiés sur de nombreux chromosomes et plasmides à faible nombre de copies, aussi bien chez les *Bacteria* que chez les *Archaea*. Le terme de *tuerie* est souvent abusif car les toxines possédent seulement un effet bactériostatique inhibant la croissance des cellules ayant perdu le plasmide.

Ces systèmes sont généralement codés sous forme d'un opéron codant deux protéines : un poison et son antidote. L'antidote est relativement instable par rapport au poison, il est rapidement dégradé par l'activité d'une protéase ATP-dépendante. L'antidote doit être maintenu à une concentration suffisante pour contrecarrer l'effet du poison, ceci requiert la présence du plasmide codant pour cette protéine. En effet, lors d'une division, une cellule ne reçevant pas de plasmides voit l'antidote encore présent dans le cytoplasme rapidement se dégrader sans qu'il ne puisse être synthétisé *de novo*. Le poison plus stable peut alors agir et tuer la cellule, d'où l'appellation de *post-segregational killing*. Le terme « module de dépendance » fait référence au fait que la cellule est « accro » (*addicted*) à sa « dose » d'antidote pour survivre.

L'antidote peut être soit un ARN antisens empêchant la traduction de l'ARN messager du poison comme dans le système *hok*/*sok* ou une protéine inhibant l'action du poison par formation d'un complexe avec celui-ci. Les systèmes protéine-protéine ont des caractéristiques communes : les protéines poison et antidote sont de petite taille (70 à 130AA) et le gène codant l'antidote précède celui du poison dans l'opéron. La transcription de cet opéron est auto-réprimée lors de la formation du complexe toxine-antitoxine. Les prédictions de structure secondaire suggèrent la présence d'un court brin ß à l'extrémité N-terminale des antidotes, le dimère d'antidote forme ainsi un feuillet ß impliqué dans la répression par liaison spécifique au promoteur de l'opéron.
Il existe une dizaine de systèmes de stabilisation plasmidiques. Certains systèmes sont spécifiques aux plasmides, alors que d'autres peuvent également être portés par les chromosomes. Plusieurs systèmes peuvent être présents sur un réplicon (plasmide ou chromosome) afin d'augmenter la stabilité de celui-ci. Les toxines portées par le chromosome sont également les premières enzymes exprimées lorsque l'on soumet la bactérie à un stress, en particulier lorsqu'il y a formation de lésions de l'ADN. Les cellules répondent à ce stress en utilisant l'effet bactériostatique des toxines pour « prendre leur temps » afin de réparer leur ADN et d'effectuer des recombinaisons (Williams *et al.,* 2007).

- *cdd*

Ce système est, à l'heure actuelle, le plus documenté. Il est localisé sur le plasmide F d'*E. coli (Tam et al., 1989)*. L'opéron *ccd* possède deux gènes, codant la protéine antidote CcdA (72AA) et le poison CcdB (101AA). Par isolement d'un mutant résistant à CcdB, il a été montré que la cible de ce poison est la gyrase, une topoisomérase de type II impliquée de la gestion du surenroulement de l'ADN (Van Melderen 2002). L'empoisonnement de la gyrase perturbe son fonctionnement et l'amène à provoquer des cassures double-brin aléatoirement dans l'ADN de manière ATP-dépendante. Ce type de lésion induit le système SOS et bloque le processus de division afin que la cellule ait le temps de réparer les dommages avant d'entamer une divison. La cellule continue à grandir sans division et apparaît alors filamenteuse.
L'autorépression de l'opéron *ccd* requiert la présence de CcdA et de CcdB. La forme impliquée dans la liaison à l'ADN est un tétramère $(CcdA)_2$-$(CcdB)_2$ alors que la forme libre est un hexamère $(CcdA)_2$-$(CcdB)_4$. Ceci suggère que la baisse relative de la quantité de CcdA diminue la répression et favorise ainsi sa synthèse. CcdA41 est une délétion ne comportant que les 41 acides aminés C-terminaux de CcdA perdant la capacité d'autorégulation tout en conservant ses propriétés d'antidote. Cette observation suggère que la partie N-terminale de CcdA est responsable de la liaison au promoteur de l'opéron *ccd*. Cette hypothèse est également étayée par l'observation

d'une mutation ponctuelle de l'arginine 4 en alanine dans CcdA conduisant à l'abolition l'activité de répression.

La protéine CcdA est dégradée par la protéase Lon, malgré une affinité pour Lon très inférieure à celle pour CddB. Deux modèles proposent que la dégradation des formes « accidentellement » libres de CcdA soit suffisante pour provoquer le *post-segregational killing* ou bien que qu'il existe un processus actif déstabilisant le complexe CcdA-CcdB.

L'étude des *ccd* portés par le chromosome de *Erwinia chrystanstemi* à montré que les homologues chromosomiques ne seraient pas principalement impliqués dans ce mécanisme d'addiction, mais interagiraient avec ceux d'un élément génétique mobile tentant de s'établir dans la cellule ; ils se comporteraient alors comme des modules d'anti-addiction et de protection contre l'invasion par un élément génétique (Saavedra De Bast *et al.,* 2008).

- *pem*

L'opéron *pem* du plasmide R100 d'*E. coli* code le poison PemK (110AA) et son l'antidote PemI (84AA) (Jensen *et al.,* 1995). Ce système est identique au système *parD* du plasmide R1, le poison étant alors appelé Kid et l'antidote Kis. Nous utiliserons la première appellation afin d'éviter toute confusion avec le système parDE détaillé plus bas.

Comme *ccd*, la stabilisation du plasmide résulte de la dégradation de l'antidote PemI par Lon et l'auto-répression complète nécessite à la fois le poison et l'antidote même si une faible répression est observée avec l'antidote seul.

La cible du poison PemK pourrait être DnaB, une hélicase impliquée dans l'initiation de la réplication de l'ADN. En effet, la surproduction de DnaB inhibe l'action toxique de PemK. PemK agirait comme un inhibiteur de la division cellulaire même si un effet de *post-segregational killing* a été observé chez certaines souches.

- *parDE*

RK2, également appelé RP4, est un gros plasmide de 60 kb possédant un large spectre d'hôtes au sein des bactéries à Gram négatif, telles que *E. coli*, *Agrobacterium tumefaciens* ou *Pseudomonas aeruginosa*. Selon l'espèce considérée, entre 4 et 8 copies du plasmide sont présentes dans le cytoplasme.

Un locus de 3,2 kb appelé *par* est responsable de la stabilisation de RK2 (Oberer *et al.,* 1999). Il contient cinq gènes organisés en deux opérons divergents. Le premier opéron, *parCBA* (2,3 kb), est responsable d'un système de résolution de multimères et le second, *parDE* (0,7 kb), est un système de *post-segregational killing* dont ParD est l'antidote (83AA) et ParE le poison (103AA). L'association de ces deux systèmes permet un très faible taux de cure plasmidique. La

contribution relative des deux opérons dans la stabilisation est fonction de la souche et de la température de croissance.

La cible du poison ParE n'a pas encore été déterminée, mais comme CcdB, ParE provoque la mort des cellules ayant perdu le plasmide, l'inhibition de la réplication, l'induction du système SOS et l'apparition de bactéries filamenteuses. Sachant que le système *parDE* est très répandu et qu'il est difficile de trouver des mutants résistants à ParE, on peut raisonnablement penser que ParE affecte une fonction essentielle. L'autorépression est accomplie par ParD ou par le complexe ParD-ParE sans que l'on ait pu déceler une quelconque action de ParE dans cette répression.

- *relBE*

Le système *relBE* a été découvert dans le génome d'*E. coli (Gotfredsen et al., 1998)*. L'opéron *relBE* code deux protéines : la toxine RelE, et son antidote RelB. La toxine possède une activité ribonucléase-ribosome dépendante. Elle se fixe au niveau du site A du ribosome et dégrade les ARNm en cours de synthèse. En présence de l'antitoxine, une autorépression transcriptionnelle intervient par fixation du complexe *relBE* sur le promoteur de l'opéron**.**

Ce système est le plus universel. Il est rencontré sur des plasmides d'autres espèces bactériennes à Gram négatif, de bactéries à Gram positif, et même d'*Archaea,* principalement des Euryarchaea. À ce jour, aucun homologue n'a été trouvé chez les eucaryotes même s'il a été montré que RelE d'*E. coli* est également actif dans *Saccharomyces cerevisiae* et que RelB en contrecarre son action (Kristoffersen *et al.,* 2000).

- *phd*/*doc*

Dans sa phase lysogène, le bactériophage P1 d'*E. coli* est présent sous forme d'un plasmide à faible nombre de copies. Il contient le module de PSK *phd*/*doc* où Phd (73 AA) est l'antidote instable et Doc (126AA), le poison stable (Jensen *et al.* 1995). Phd est dégradé par la protéase ClpPX. Phd est capable d'autoréguler seul l'opéron mais opère de façon plus efficace en présence de Doc.

La cible de Doc n'est pas encore connue mais elle serait impliquée dans une étape essentielle de la synthèse des protéines. Il a été établi que la présence d'un autre système d'addiction, *mazEF,* était requise pour la fonction *post-segregational killing* de *phd*/*doc*. Ces observations suggèrent que Doc provoque l'inhibition de la traduction de l'antidote MazE, menant ainsi à la mort. Ces deux mécanismes seraient donc couplés afin d'augmenter l'efficacité de partition.

- restriction-modification

Un autre type de système, proche des poison-antidote, est le système restriction-modification (RM). D'ordinaire, il est utilisé par les cellules pour se protéger de l'infection par de l'ADN étranger (un phage, par exemple). Il est caractérisé par la présence d'enzymes de restriction qui clivent la molécule d'ADN invasive au niveau d'un site spécifique. Le génome de la bactérie est protégé des enzymes de restriction grâce à une enzyme méthylant son ADN.

Ce système fonctionne également pour la stabilisation de plasmides. La différence fondamentale avec les systèmes « classiques » poison-antidote tient au fait que les deux enzymes ne forment pas un complexe inactivant l'enzyme de restriction. De plus, l'effet de *post-segregational killing* ne résulterait pas d'une dégradation plus rapide de la méthylase permettant la toxicité du poison (enzyme de restriction). Un modèle a été proposé dans lequel la dilution progressive des deux enzymes au cours des divisions successives mènerait à un point où la méthylation ne sont plus suffisantes à se protéger du peu d'enzymes de restriction encore présentes pouvant causer des dommages irréparables à l'ADN.

3.2.2 Intégration

Certains plasmides possèdent la capacité d'intégrer leur génome dans le chromosome de leur hôte. L'intégration est le résultat d'une recombinaison site-spécifique. L'intégration de plasmide, ou la formation de provirus, est effectuée par une recombinase de la famille des intégrases (Ints). L'intégration se déroule en plusieurs étapes catalysées uniquement par cette enzyme. Les mécanismes impliquant les recombinases et aboutissant à l'intégration de l'élément génétique mobile sont détaillés dans le chapitre consacré aux éléments génétiques intégrés des *Archaea* (Page 48)

4. Fonctions spécifiques

Les fonctions spécifiques ne sont pas indispensables à la survie du plasmide mais peuvent néanmoins apporter un avantage à l'hôte cellulaire en cas de changement d'environnement. De nombreuses souches ont été isolées du fait de la présence de plasmides conférant un avantage sélectif. Les cas les plus documentés pose d'énormes problèmes de santé publiques, ils concernent les plasmides codant des gènes de résistance aux antibiotiques (Schluter *et al.,* 2007). On peut également trouver des gènes impliqués dans la résistance aux métaux lourds (Unaldi *et al.,* 2003) ou l'exemple de certains *Pseudomonas* possédant l'opéron de gènes TOL impliqué dans la dégradation des composés hydrocarbures aromatiques (Greated *et al.,* 2002).

III. Les Thermococcales, un genre d'*Archaea* hyperthermophiles

1. Généralités sur les *Archaea*

La notion d'espèce biologique est essentiellement basée sur l'interfécondité entre individus. Ce fondement ne peut pas être utilisé pour la détermination des espèces de Procaryotes car leur reproduction est assurée par un mécanisme asexué de scissiparité. Durant la première moitié du XXème siècle, la taxonomie des Procaryotes fut d'abord basée sur des méthodes phénétiques regroupant les individus possédant des ressemblances (morphologiques, physiologiques, biochimiques, et écologiques). (Willy Henning, 1950)

Les organismes vivants sont constitués de macromolécules dont la composition est déterminée par le patrimoine génétique hérité de leurs ancêtres. C'est au niveau de ce patrimoine génétique et de son expression que s'opère la sélection naturelle. Un signal phylogénétique permet de retracer une histoire « généalogique » basée sur l'héritage vertical des gènes provenant de ses ancêtres. Sur ce postulat, Zuckerkandl et Pauling ont eu l'idée d'établir un arbre phylogénétique en se servant de la séquence d'une macromolécule comme d'un chronomètre moléculaire (Zuckerkandl *et al.,* 1965). Des mutations communes reflètent une relation de parenté entre organismes. L'avènement des techniques de biologie moléculaire a permis à Fitch et Margoliash d'appliquer ce concept à la protéine du Cytochrome C (Fitch *et al.,* 1967), mais ce marqueur moléculaire n'est pas universel et ne permet pas de répondre à la notion d'espèce chez les Procaryotes.

Woese et ses collaborateurs ont utilisé la molécule d'ARNr de la petite sous-unité du ribosome (16S pour les procaryotes et 18S pour les eucaryotes) pour mesurer les relations phylogénétiques entre micro-organismes, tel un chronomètre moléculaire (Woese *et al.,* 1977). Cette molécule a été choisie pour deux avantages principaux : une abondance importante dans la cellule et une taille suffisante pour réaliser des études statistiques. Cette séquence a l'avantage de présenter, d'une part, des régions très conservées permettant de différencier des espèces très éloignées et, d'autre part, des régions hypervariables permettant de discriminer des espèces voisines.

A l'origine de la méthode, l'ARNr 16S d'une espèce était digéré par l'enzyme RNAseT1 produisant un catalogue d'oligonucléotides de 6 à 12 paires de bases de séquence spécifique propre à chaque espèce. Par comparaison des catalogues de deux espèces, il était possible de retrouver la distance évolutive entre ces micro-organismes. Les premières études, basées sur l'analyse de séquences partielles, ont confirmé la dichotomie procaryotes/eucaryotes initialement basée sur la présence ou l'absence de noyau. L'application à plus grande échelle a mis en évidence un résultat

insoupçonné : il existe deux familles totalement différentes de procaryotes. L'étude de l'ARNr des bactéries anaérobies méthanogènes a montré qu'elles forment un groupe homogène représentant un taxon évolutif extérieur à celui des bactéries. Woese et ses collaborateurs ont baptisé ces micro-organismes archaeabactéries (Woese *et al.* 1977). La dichotomie Procaryotes/Eucaryotes a alors été scindée pour aboutir à un arbre du vivant constitué de trois règnes primaires : les archaeabactéries, les eubactéries et les eucaryotes. Le préfixe latin *Archaea* fait référence aux biotopes dans lesquels vivent ces organismes, supposés se rapprocher de l'atmosphère primitive terrestre (forte salinité ou pH, dépourvu d'O_2, riche en H_2 et $C0_2$ qui constituent les matières premières pour la production de méthane). Leur découverte dans les sources hydrothermales océaniques profondes a donné une nouvelle impulsion à l'hypothèse de l'origine chaude de la vie tout en faisant basculer son lieu d'apparition de la surface de la planète vers les profondeurs de la terre et des océans. Ces biotopes à grandes profondeurs auraient permis l'apparition de la vie et son développement à l'abri de l'intense bombardement de météorites dont la Terre a été le théâtre pendant les premiers 500 millions d'années de son existence. Ces gigantesques bombardements pouvaient en effet volatiliser la couche supérieure des océans, et stériliser la surface de la planète (Gogarten-Boekels *et al.,* 1995). De plus, l'apparition de la vie en profondeur règlerait le problème posé par le rayonnement UV, délétère pour les acides nucléiques, qui prévalait à cette époque en l'absence de couche d'ozone protectrice. Enfin, un grand nombre de ces micro-organismes atypiques présente un métabolisme basé sur l'utilisation du soufre et du fer, deux éléments que l'on retrouve dans le scénario de Wächtershaüser pour une origine autotrophe du vivant (Wachtershauser 1990).

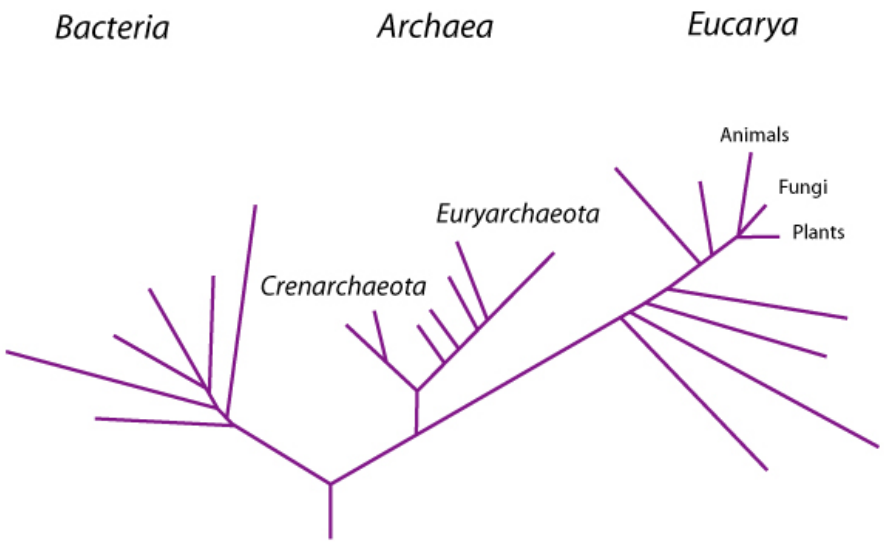


Figure 5 Arbre phylogénétique des trois domaines du vivant basé sur l'ADNr

La distance évolutive entre les deux règnes procaryotes est comparable à la distance avec les eucaryotes : le regroupement initial de ces organismes en un groupe (procaryote) devient impropre. Malgré l'absence de fossile, la comparaison de la longueur de la branche principale des *Archaea* à celle des deux autres règnes laisse penser que les *Archaea* seraient beaucoup plus anciennes (3,5 milliards d'années) et que leur vitesse d'évolution serait plus lente. Cependant, il s'agit d'hypothèses uniquement basées sur la comparaison d'ARNr. Certaines analyses phylogénétiques font apparaître les *Archaea* beaucoup plus tard, en même temps que les eucaryotes, aux alentours de 950 millions d'années.

La nomenclature proposée par Woese (Woese *et al.,* 1990) établit un nouvel arbre universel du vivant. Les trois domaines primaires sont nommés *Bacteria*, *Eukarya* et *Archaea*.

Les biotopes colonisés par les *Archaea* correspondent à l'ensemble des niches écologiques rencontrées sur le globe. Elles présentent de spectaculaires adaptations, notamment en environnements extrêmes. Elles sont aussi bien rencontrées en environnements psychrophiles, constituant notamment 34% du plancton marin arctique (DeLong *et al.,* 1994) qu'en environnements mésophiles ; les *Archaea* nitrifiantes jouent un rôle prépondérant dans le cycle de l'azote du sol (Leininger *et al.,* 2006). De remarquables adaptations leur permettent de résister à des températures extrêmes, le record actuel étant *Pyrolobus fumarii* supportant la température de 113°C (Cowen 2004), à des pH proches de 0 *Picrophilus torridus* (Schleper *et al.,* 1995), à des salinités très élevées *Halobacterium salinarium* (Dennis *et al.,* 1997), à d'importantes concentrations de métaux lourds *Ferroplasma acidophilum* (Golyshina *et al.,* 2005) ou encore à d'importantes doses de radiations ionisantes *Thermococcus gammatolerans* (Jolivet *et al.,* 2004).

Le domaine des *Archaea* est divisé en deux principaux phyla clairement établis, auxquels s'ajoutent d'autres phyla dont la pertinence reste à démontrer.

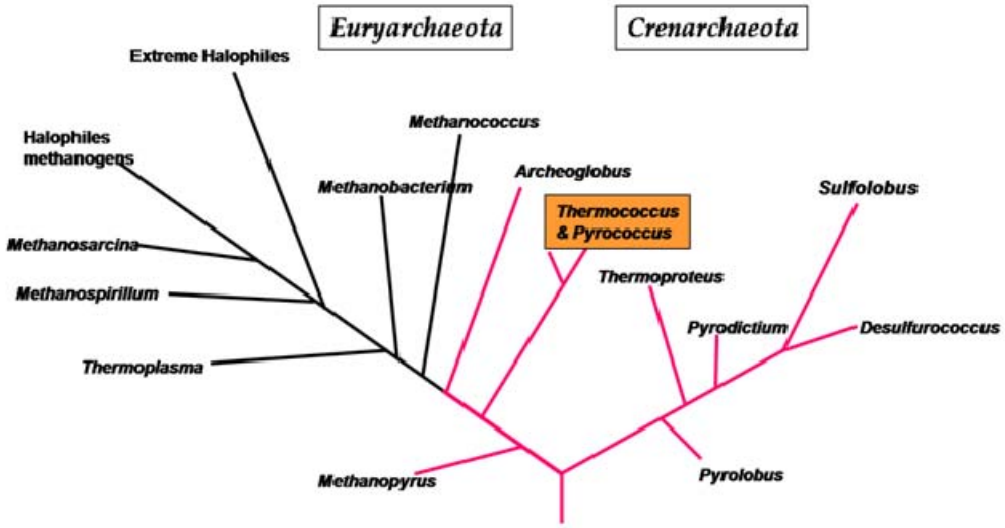


Figure 6 Arbre phylogénétique des Archaea basé sur la séquence de l'ADNr16S

Les ***Euryarchaeota***, du grec *euryos* 'diverses', comprennent 8 classes de micro-organismes couvrant à la fois un large spectre de niches écologiques et de métabolismes : des méthanogènes *Methanobacteria, Methanomicrobia, Methanococci* et *Methanopyri*, des halophiles extrêmes *Halobacteria*, ou des sulfato-réductrices hyperthermophiles *Archaeoglobi et Thermococci* et les *Thermoplasmatales.*

Les ***Crenarchaeota***, du grec *crenos* 'source', sont exclusivement composés d'espèces hyperthermophiles (Sulfolobales, Thermoproteales, Pyrodictiales, Desulfurococcales). La présence d'organismes mésophiles au sein de ce phylum a été suggérée par l'analyse de séquences environnementales d'ADNr16S. Néanmoins, l'accumulation des données pousse à croire que ces organismes appartiendraient à un phylum distinct nommé *Thaumarchaeota*.

Les ***Thaumarchaeota***, du grec *thaumas* 'merveille', sont une proposition de troisième phylum des *Archaea* (Brochier-Armanet *et al.,* 2008). Ces micro-organismes sont rarement rencontrés dans les études de diversité car ils ne répondent pas positivement aux amplifications réalisées avec les amorces universelles aux *Archaea* (Baker *et al.,* 2004). Ce phylum comporte de nombreux micro-organismes mésophiles non-cultivés, seulement détectés par analyses métagénomiques. Pour l'instant, ils sont considérés comme des Crenarchaea mésophiles. L'analyse du génome de *C.symbiosum* suggère en effet qu'un pan entier de la diversité des *Archaea* n'a pas encore été exploré (Hallam *et al.,* 2006). La compréhension de la biosphère et des cycles géochimiques nécessite d'accroitre nos connaissances sur ces organismes. Ils pourraient, par exemple, jouer un rôle prépondérant dans le cycle de l'azote (Prosser *et al.,* 2008).

Les ***Nanoarchaeota***, du grec *Nanos* 'petit', sont le dernier ajout au sein des *Archaea*. Un seul représentant, *Nanoarchaeum equitans,* a été décrit suite à son isolement à partir d'une source hydrothermale en Islande (Huber *et al.,* 2002). C'est un parasite de 0,4µm, vivant attaché au Crénote *Ignicoccus*. Cet organisme possède le plus petit génome connu. Son génome de 490 kb a été séquencé en 2003 (Waters *et al.,* 2003). La plupart des voies métaboliques essentielles, telles que la synthèse des nucléotides, lipides et cofacteurs n'existent pas. Des études phylogénétiques montrent qu'ils seraient issus des Euryarchaeota suite à une évolution rapide de membres des Thermococcales (Brochier *et al.,* 2005). La présence de caractères ancestraux, tels que des demi-ARNt et l'absence d'organisation génique sous forme d'opéron, pousse certains chercheurs à penser que cet organisme est un fossile vivant et que la racine de l'arbre universel du vivant serait fixée sur cette branche (Di Giulio 2007).

Les progrès technologiques réalisés durant les années 1990 ont permis d'utiliser la totalité de la séquence de l'ARNr pour la construction d'arbres phylogénétiques, rendant plus robuste la position et la longueur des branches au sein de l'arbre du vivant. Malheureusement, ces arbres ne sont pas toujours congruents lorsqu'ils sont réalisés avec d'autres gènes (Doolittle 1999). Les nombreux programmes de séquençage complet de génomes ont également remis en question le fondement de la notion d'espèce basée sur la molécule d'ARNr 16S (Konstantinidis *et al.,* 2005).

Cet historique de la définition d'une espèce procaryote montre une étroite **corrélation entre la révision de certains concepts et le développement des techniques d'analyses**. Alors que le séquençage d'un organisme est une photographie de son génome à un instant donné, l'essor actuel et en devenir des techniques de séquençage ouvre de nouvelles perspectives d'appréhension de l'évolution et la notion d'espèce. L'accès à une vision dynamique du génome permet de mesurer l'impact majeur des transferts horizontaux sur l'adaptation et la spéciation.

Bien que présentant des ressemblances morphologiques, *Archaea* et *Bacteria* possèdent de nombreux facteurs moléculaires témoignant une origine phylogénétique distincte. La membrane des *Archaea* est constituée de lipides à liaisons éther ou dibiphytanyl di glycérol tétraéther, formant une couche imperméable et stable particulièrement adaptée aux hautes températures. La faible perméabilité de cette membrane, notamment à la fuite d'ATP, et la présence de nombreuses voies de recyclage des métabolites secondaires, confèrent aux *Archaea* une physiologie adaptée aux environnements extrêmes ou changeants. Cette physiologie remarquable leur permet d'être très compétitives et de s'adapter au **stress chronique énergétique** dans les environnements oligotrophes (Valentine 2007). Elles possèdent également des propriétés

témoignant d'une origine commune avec les *Eukarya*, tels que les complexes de transcription et de réplication ou l'absence de peptidoglycane dans les enveloppes cellulaires.

2. Les Thermococcales

2.1 *Caractéristiques*

Les Thermococcales sont des coques hyperthermophiles vivant en environnements aquatiques. Leur croissance optimale est observée dans une gamme de température comprise entre 70 et 106°C. Elles possèdent souvent une touffe de flagelles polaires leur conférant une grande mobilité. Elles sont **anaérobies chimoorganotrophes** et nécessitent dans leur milieu de croissance la présence de protéines et de solutions riches en matière organique. Ces composés organiques sont oxydés et les électrons produits finissent sur l'accepteur final d'électron S_0 qui est réduit en H_2S. Néanmoins, ce métabolisme énergétique ne semble pas exclusif, comme le montre *T. onnurineus* qui est également carboxydotrophique (Lee *et al.,* 2008).

Les Thermococcales comprennent 3 genres : *Thermococcus*, *Pyrococcus* et *Palaeococcus*.

Thermococcus est le genre possédant le plus d'espèces décrites. Actuellement 29 souches provenant de différents environnements ont été décrites (Tableau 1). Elles sont principalement retrouvées au niveau des sites hydrothermaux le long des dorsales océaniques, à proximité du fluide émanant des fumeurs noirs. Elles peuvent également être isolées au niveau de sources marines côtières. Cet endémisme aux environnements océaniques est désormais levé. Trois souches ont été découvertes en eau douce. *T. waiotapuensis* (Gonzalez et al., 1999) et *T. zilligi (Ronimus et al., 1997)*, isolées du bassin hydrothermal terrestre Champagne Pool en Nouvelle-Zélande et une souche de *Pyrococcus* isolée d'une source d'eau douce en Algérie (Kecha *et al.,* 2007). La souche *T. sibiricus*, quant à elle, a été isolée d'un puits de pétrole en Sibérie (Miroshnichenko *et al.,* 2001).

Tableau 1 Souches de Thermococcales décrites

Type	Origine	Nom	Site	Nature de l'échantillon	Référence
Terrestre	Sibérie	*T. sibiricus*	Samotlor oil reservoir Nizhenevartovsk	Forrage -2350 m 84°C	Miroshnichenko et al. 2001
Eau douce	New Zeland	*T. zilligii*	Rotorua Kuirau Park	Terrestrial fresh water hot pool	Ronimus et al. 1999
	New Zeland	*T. waiotapuensis*	Waiotapu hot spring Lake Taupo area	Terrestrial fresh water hot spring 90°C	González et al. 2001
Côtière	New Zeland	*T. gorgonarius*	Whale Island	Sand from the shore 85°C	Miroshnichenko et al. 1998
	New Zeland	*T. pacificus*	Bay of Plenty	Shallow-water hot vent -40m 85°C	Miroshnichenko et al. 1998
	Japon	*T. kodakaraensis*	Kodakara Island Kagoshima	Solfatara on the shore 102°C	Atomi et al. 2005
	Grèce	*T. aegaeicus*	Palaeochori Bay Milos	Shallow water area, sédiments -4m 90-103°C	Arab et al. 2000
	Italie	*T. alcaliphilus*	Vulcano Porto Levante	Shallow marine hydrothermal system	Keller et al. 1997
	Italie	*T. litoralis*			Neuner et al. 2001
	Italie	*T. celer*	Vulcano Porto Levante	Solfataric marine water hole	Zillig 1983
	Italie	*T. woesei*	Vulcano Porto Levante	Shallow marine hydrothermal system	Zillig 1988
	Italie	*T. acidaminovorans*	Vulcano	Shallow marine hydrothermal system	Dirmeier et al. 2001
	Italie	*P. furiosus*	Vulcano	Shallow marine hydrothermal system	Fiala and Stetter 1986
Profond	Pacifique	*T. aggregans*	Basin Guaymas	Sédiments 2000m	Canganella et al. 1998
	Atlantique	*T. atlanticus*			Cambon-Bonavita et al. 2004
	Atlantique	*T. barophilus*	Mid Atlantique Ridge	SnakePit 3550m (23°22N, 44°56W)	Marteinsson et al.1999
	Pacifique	*T. barossii*	EPR Juan De Fuca		Duffaud et al. 1998
	Pacifique	*T. celericrescens*	Izu Bonin Arc	Suiyo Seamount (28°349N 140°389E) 1380m	Kuwabara et al. 2007
	Pacifique	*T. chitonophagus*	Basin Guaymas		Huber and Stetter 1996
	Pacifique	*T. coalescens*	Izu Bonin Arc	Suiyo Seamount (28°349N 140°389E) 1380m	Kuwabara et al. 2005
	Pacifique	*T. fumicolans*	North Fidji Basin	Hydothermal vent Chimney wall	Godfroy and Meunier 1996
	Pacifique	*T. gammatolerans*	Basin Guaymas	Hydothermal vent Chimney wall 2616m	Jolivet et al. 2003
	Pacifique	*T. guaymasensis*	Basin Guaymas	Sédiments 2000m	Canganella et al. 1998
	Pacifique	*T. hydrothermalis*	EPR 21°N	Chimney walls covered with Alvinnelids tubes	Godfroy et al. 1997
	Pacifique	*T. onnureneus*	PACMANUS field New guinea Australie	Hydrothermal fluid 1650m 3°44S, 151040E	Bae et al. 2006
	Pacifique	*T. nautilii*	EPR 13°N 104°W	Hydrothermal chimney 2330m	Soler et al. 2007
	Pacifique	*T. peptonophilus*			González et al. 1996
	Pacifique	*T. profundus*	Mid Okinawa Through	Hyfrothermal vent chimney 1400m	Kobayashi and Horikoshi 1995
	Pacifique	*T. siculi*	Mid Okinawa Through 27°32N 126°56E	Hydrothermal fluid 1394m	Grote et al. 2000
	Pacifique	*T. stetteri*	Marine solfararic fields, Northern Kurils	profond?	Miroshnichenko 1990
	Atlantique	*T. thioreducens*	Rainbow 36°N, 33°W	Black smoker chimney 2300m	Pikuta et al. 2007
	Pacifique	*P. horikoshii*	Okinawa Through 27°33N 125°56E	Hydrothermal fluid 1395m	González et al. 1999

La notion d'espèce est basée sur la divergence de séquence de l'ADNr 16S. Cette séquence permet d'affilier un genre à un micro-organisme (Figure 7). Cette technique de typage est très utile lors d'études de diversité microbienne. Néanmoins, en cas d'identité de séquences importante avec un autre ADNr16S, une hybridation ADN:ADN est nécessaire pour discriminer deux espèces.

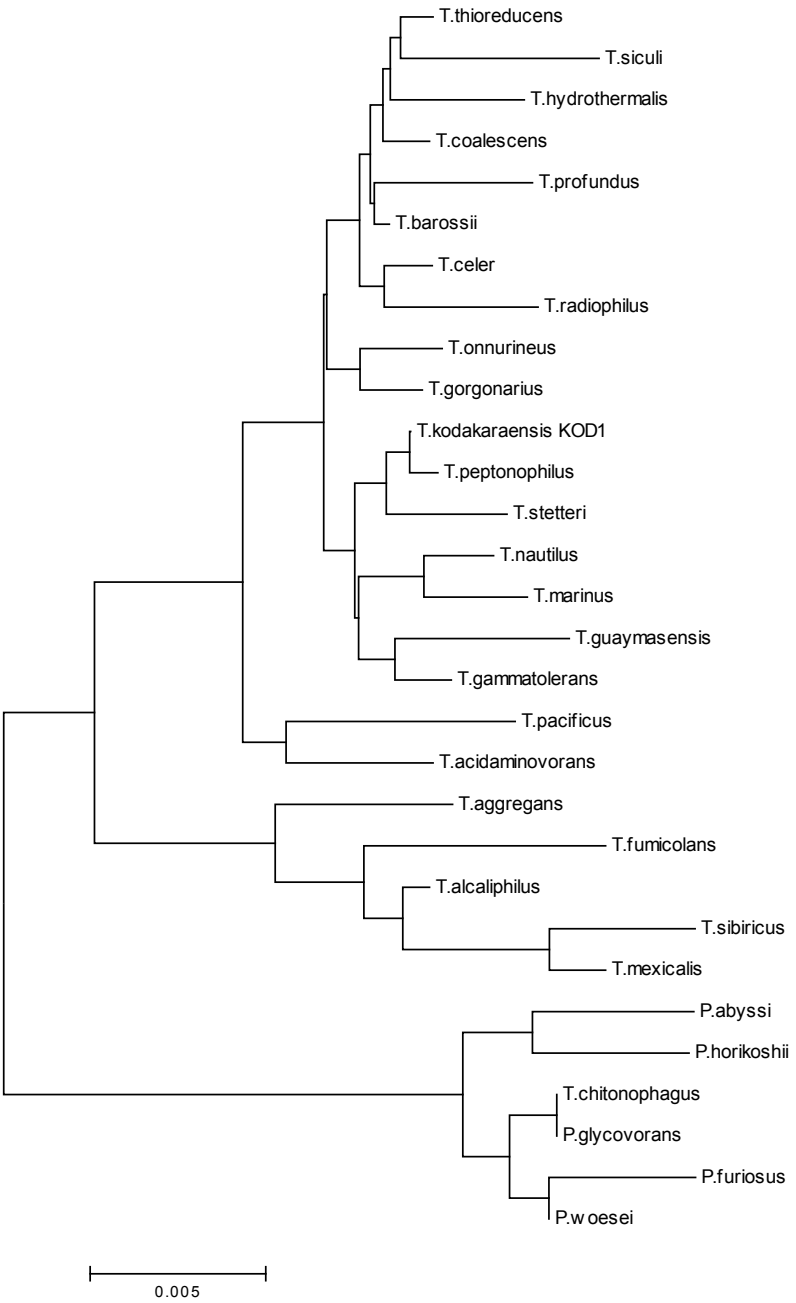


Figure 7 Arbre phylogénétique de l'ADNr16S des Thermococcales

Le critère d'interhybridation permet en théorie de discriminer deux espèces. Les révolutions technologiques, telles que l'accès au génome complet de certaines souches, montrent néanmoins que ce critère d'interhybridation possède certaines **limites**. Deux souches de *Pyrococcus* provenant du site de Vulcano en Italie ont été décrites au milieu des années 80: *Pyrococcus furiosus* (Fiala *et al.,* 1986) et *Pyrococcus Woesei* (Zillig, 1988). La disponibilité du génome de *Pyrococcus furiosus* a permis la construction d'une puce à ADN contenant tous ses ORFs (Hamilton-Brehm *et al.,* 2005). L'hybridation de l'ADN de *Pyrococcus woesei* sur cette puce montre que son génome est dépourvu de 105 ORFs organisés en deux clusters bordés par des séquences d'insertions (IS) ou des répétitions en tandem (LCTR). Selon les auteurs, *Pyrococcus woesei* aurait acquis cette centaine de gènes par transfert horizontal pour devenir *Pyrococcus furiosus*. Néanmoins, je pense qu'il faut également considérer la situation inverse, c'est-à-dire que la recombinaison entre deux LCTR aurait pu supprimer ces gènes. *Pyrococcus woesei* a depuis été renommée *Pyrococcus furiosus subsp. Woesei* (Kanoksilapatham et al., 2004). Ces exemples illustrent l'**importance de la plasticité des génomes et des transferts horizontaux de gènes dans la pangénomique**.

2.2 *Génomique comparée des Thermococcales*

Quatre génomes de Thermococcales sont disponibles dans les bases de données : *P. abyssi* (Cohen et al., 2003) *, P. horikoshii* (Kawarabayasi et al., 1998)*, P. furiosus* et *T. kodakaraensis KOD1* (Fukui et al., 2005). Ces génomes ont une taille voisine de 2Mb et comportent environ 2000 gènes.

Quatre autres génomes, d'espèces possédant des caractéristiques particulières, sont en cours de séquençage. *T.gammatolerans* (Jolivet *et al.* 2004) est un organisme à très haute résistance aux radiations ionisantes, *T. barophilus* (Marteinsson *et al.,* 1999) est le premier organisme isolé sous pression, *Thermococcus sp. AM4* (Sokolova *et al.,* 2004) est un organisme produisant son énergie par oxydation du monoxyde de carbone et libérant du H_2. Le dernier programme de séquençage, initié en aout 2007, concerne le génome de *T.onnurineus* (Cho et al., 2007). Bien que l'article concernant ce génome vienne d'être publié (Lee *et al.* 2008), sa séquence n'est toujours pas dans les bases de données. Il ne pourra donc pas être discuté dans cette étude.

L'obtention d'un plus grand nombre de génomes et leur comparaison pourrait permettre la détection d'opéron de gènes conférant les propriétés particulières (importance du retour à la microbiologie classique -> la post-génomique tend à revaloriser la microbiologie « classique », la

ce terme est d'ailleurs trompeur, il tend à cloisonner les secteurs d'acquisition d'informations). Il permettrait également d'estimer l'importance des réarrangements géniques et des transferts de gènes au sein des *Thermococcus* car la plasticité des génomes est liée à de fréquents réarrangements dynamiques, entre espèces phylogénétiquement éloignées mais aussi entre espèces proches (Bellgard *et al.,* 1999)

La synténie entre ces génomes se limite à quelques opérons. Les comparaisons de génomes ont également souligné l'extrême variabilité du contexte génomique. Les protéines universelles sont restreintes aux processus informationnels (réplication, transcription et traduction). Cette observation souligne la diversité des voies biochimiques utilisées et le fait que plusieurs solutions indépendantes ayant été « inventées » au cours de l'évolution pour accomplir ces processus fondamentaux. A plus court terme évolutif, plusieurs gènes montrent des phylogénies incongruentes avec celles de leurs hôtes basées sur l'ARNr16S. Ce mosaïsme témoigne l'importance des pertes et des acquisitions de gènes par transferts horizontaux dans la spéciation. L'importance des échanges d'ADN, entre procaryotes phylogénétiquement éloignés, pouvant appartenir à différents domaines, remet en question le scénario accepté d'émergence de la vie mais également la topologie de l'arbre universel du vivant

.

2.2.1 Les 3 génomes de Pyrococcus : *P. abyssi, P. furiosus, P. horikoshii*

Caractéristiques générales :

Les génomes des *Pyrococcus* sont assez uniformes. Leur taille est comprise entre 1,75 et 1,9Mb, un G+C% compris entre 40,8 et 44,7% et contiennent entre 1765 et 2208 ORFs (Tableau 2). Malgré des tailles homogènes, on observe une grande différence du nombre d'ORFs pouvant s'expliquer par la différence des méthodes d'annotations utilisées.

Tableau 2 Caractéristiques des génomes de *Pyrococcus,* Adapté de Lecompte et al., 2007

	P. abyssi	*P. horikoshii*	*P. furiosus*
Taille du chromosome (pb)	1 765 118	1 738 505	1 908 253
Régions codantes	91,10%	91,20%	92,50%
Contenu en G+C	44,70%	41,90%	40,80%
ARNs stables			
ARNt	46	46	46
ARNt avec introns	Trp, Met	Trp, Met	Trp, Met
ARNr	16S-23S	16S-23S	16S-23S
	5Sa, 5Sb	5Sa, 5Sb	5Sa, 5Sb
	7S	7S	7S
ORFs	1765	2061	2208
Familles de paralogues	621	606	845
Fonction putative	51%	20%	-
Gènes Informationnels	14%	-	-
Gènes Opérationnels	37%	-	-
Fonction inconnue	49%	80%	-
Eléments mobiles			
Intéines	14%	14	10
Séquences d'Insertion (IS)	1 vestige	1 vestige	24

8 intéines localisées au même site d'insertion dans les 3 *Pyrococcus* reflètent la forte conservation de ces éléments. *P. furiosus* diffère par un génome de plus grande taille et par la présence de séquences d'insertion (ISs) (Tableau 2). Le génome de *P. furiosus* contient également un plus grand nombre de paralogues. Ces différences sont en accord avec les phylogénies basées sur l'ARNr 16S indiquant que *P. abyssi* et *P. horikoshii* ont divergé après la spéciation de *P.furiosus*.

L'analyse par paire des 3 génomes montre que *P. abyssi* et *P. horikoshii* sont plus proches (1122kb en commun) que *P. abyssi* et *P. furiosus* (847) et que de *P. horikoshii* et *P. furiosus* (898kb). Le plus grand fragment conservé est observé entre *P. abyssi* et *P. horikoshii* (300kb). Un nombre important de réarrangements chromosomiques sont intervenus depuis la divergence de *P. furiosus* vis-à-vis de l'ancêtre commun à *P. abyssi* et *P. horikoshii*. Entre ces 2 proches génomes, il y a 17 inversions et transpositions.

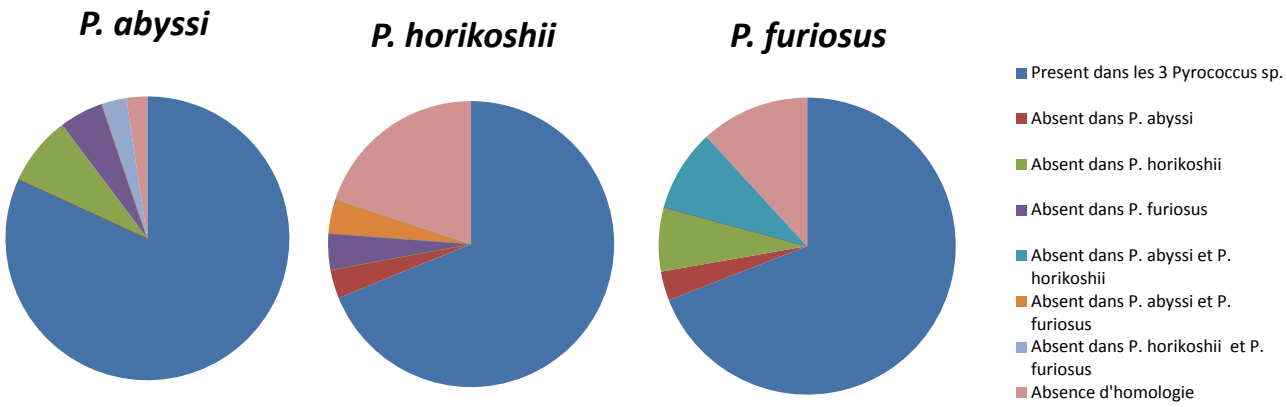


Figure 8 Répartition des ORFs homologues entre génomes de *Pyrococcus* Adapté de Lecompte et al., 2007

Ces génomes sont constitués de 4 grandes régions définies par des « motifs » conservés. Les régions les plus conservées sont la région I, contenant l'origine de réplication, et la région IV contenant l'opéron ribosomal. Dans la région II, la synténie est plus conservée entre *P. abyssi* et *P. horikoshii*. En intégrant le génome de *P. furiosus* dans la comparaison, le nombre d'évènements de réarrangements devient important. La région III contient le site de terminaison de la réplication ainsi que le point chaud de translocation et d'insertion-délétion également appelé *hospot* de recombinaison.

La principale différence entre *P. abyssi* et *P. horikoshii* est une inversion dans la région I, à proximité de l'origine de réplication. Parmi les 46 ARNt, 16 sont trouvés aux extrémités des régions réarrangées. Les ARNt sont une cible privilégiée pour la recombinaison et l'insertion-délétion de séquences. Cette hypothèse est confirmée par la présence de 2 régions propres à *P. horikoshii* (4kb à 21,6kb), bordées par des séquences répétées identiques aux extrémités 3' des $ARNt^{Val}$ et $ARNt^{Ala}$. Les séquences adjacentes (PHO1864 et PHO2000) correspondent à un gène codant une protéine similaire à l'intégrase du virus SSV1 infectant *Sulfolobus shibatae*. La conséquence de l'intégration est la duplication de l'extrémité 3' de l'ARNt, devenant un site préférentiel de recombinaison et pouvant aboutir à l'inversion de la région.

Pour trouver les gènes spécifiques de chacun des *Pyrococcus*, un seuil relativement bas de 20% d'identité de séquence protéique a été fixé. Ces gènes représentent des pertes ou des gains indépendants du genre *Pyrococcus*. Le nombre de gènes absents dans au moins un génome est assez important (278 chez *P. abyssi*, 232 chez *P. horikoshii* et 422 chez *P. furiosus*). Cette proportion est particulièrement significative chez *P. furiosus* pour lequel 192 gènes sont propres à cette espèce.

2.2.2 Thermococcus kodakaraensis

Alors que les *Thermococcus* possèdent un plus grand nombre de souches décrites, un unique génome de *Thermococcus* est actuellement disponible dans les bases de données. *Thermococcus kodakaraensis KOD1* est devenue la souche modèle pour l'étude des Thermococcales (Itoh 2003). De nombreuses enzymes d'intérêts industriels et biotechnologiques sont issues de cette souche, telles que des polymérases commercialisées pour leur haute fidélité et leur capacité d'édition (Nishioka *et al.,* 2001). Le développement des premiers **outils génétiques** fiables dédiés à cette souche permet d'envisager des expérimentations génétiques *in vivo* pour confirmer les découvertes *in vitro* et les prédictions réalisées *in silico* (Sato et al., 2005). La relative simplicité du protocole de transformation et la compétence naturelle de cette souche tendent à détrôner *P.abyssi* du statut de modèle des Euryarchées hyperthermophiles car elle semble plus compétente et plus difficilement transformable. Ce nouvel outil génétique permet d'inactiver un gène cible par double recombinaison. La sélection des cellules recombinantes est basée sur la résistance à la simvastatine portée par le vecteur. Cet antibiotique inhibe la HMG-CoA, une réductase impliquée dans la voie de biosynthèse des lipides. Ce système a permis de confirmer, expérimentalement qu'un opéron putatif de gènes était impliqué dans le transport des sucres (Matsumi *et al.,* 2007). Un second outil a été développé par l'équipe de John Reeve (Santangelo *et al.,* 2008) à partir du plasmide pTN1 de *Thermococcus nautilii* (Soler *et al.,* 2007). Ce vecteur navette entre *E.coli* et *T.kodakaraensis* possède le marqueur de sélection *trpE* (complément de l'autotrophie pour le tryptophane) ou le gène conférant la résistance à la mévinoline (antibiotique inhibiteur de la biosynthèse des lipides). Le dernier outil génétique développé pour *T.kodakaraensis* est un vecteur d'expression basé sur l'utilisation du gène TK1761 codant une β-galactosidase (Santangelo *et al.,* 2008). La première utilisation de ce vecteur a démontré que le découplage de la transcription et de la traduction a un effet négatif sur l'expression de gènes situés à proximité (Santangelo *et al.* 2008).

Cette souche a été isolée d'un solfatare côtier près de Kagoshima sur l'île de Kodakara au Japon (Morikawa *et al.,* 1994). Sa gamme de température de croissance est comprise, à pression atmosphérique, entre 60 et 100°C, avec un optimum de 85°C. Ce génome (Fukui *et al.* 2005) de 2,09 Mb possède un GC% de 52,0% plus élevé que celui des *Pyrococci* (40-45%) (Tableau 3).

Tableau 3 Caractéristiques des génomes de Thermococcales

	T. kodakaraensis	*P. abyssi*	*P. horikoshii*	*P. furiosus*
Taille du chromosome (pb)	2 088 737	1 765 118	1 738 505	1 908 253
Régions codantes	92,1%	91,1%	91,2%	92,5%
Contenu en G+C	52,0%	44,7%	41,9%	40,8%
CDS	2306	1765	2061	2208
ARNs stables				
ARNt	46	46	46	46
ARNt avec introns	Trp, Met	Trp, Met	Trp, Met	Trp, Met
ARNr	16S-23S	16S-23S	16S-23S	16S-23S
	5Sa, 5Sb	5Sa, 5Sb	5Sa, 5Sb	5Sa, 5Sb
	7S	7S	7S	7S
Eléments mobiles				
Intéines	16	14	14	10
Transposases	7	5	3	30
Virus-Related regions	4	0	2	0

Parmi les 2306 CDS, la moitié ne sont pas annotés. La génomique comparée montre que 1204 protéines sont conservées au sein des trois *Pyrococcus*. Ces protéines conservées incluent les processus informationnels et ceux des métabolismes primaires.

D'un autre côté, 689 protéines sont spécifiques de *T. kodakaraensis*. Quelques-unes sont très intriguantes et pourraient être responsables des traits spécifiques du genre *Thermococcus ;* notamment des protéines impliquées dans l'oxydation additionnelle du pyruvate, le métabolisme des nucléotides, la constitution des transporteurs d'ions métalliques, les systèmes améliorés de réponse au stress ou dans des systèmes de restriction-modification. Le génome de *Thermococcus kodakaraensis* a une taille plus importante que celui des *Pyrococcus.* La faible proportion de gènes dupliqués et des proportions d'espaces intergéniques comparables suggèrent une **acquisition de ces gènes supplémentaires par transfert horizontal** après la spéciation entre *Thermococcus* et *Pyrococcus*. Ce gain de gènes pourrait également résulter de l'action de grands transposons composites flanqués par des IS comme cela a été observé entre *Pyrococcus furiosus* et *Thermococcus litoralis* (Diruggiero et al., 2000).

Il n'existe pas de grandes régions contiguës du génome de *T. kodakaraensis* non retrouvées chez les *Pyrococci*. Cette observation suggère l'absence de transfert récent de grands fragments d'ADN à partir de lignées très éloignées. Alors que les contextes génomiques sont très conservés entre génomes de *Pyrococci*, leurs homologues sont très transloqués chez *T. kodakaraensis.* Ceci est

probablement la conséquence de multiples réarrangements, confirmant leur importance dans l'évolution. Il n'y a pas de longue synténie entre *T. kodakaraensis* et les *Pyrococci*. **Les réarrangements génomiques sont généralement plus rapides que l'évolution de la séquence des protéines**. La fréquence de translocation d'un gène unique a été corrélée au nombre d'IS et de séquences répétées dans le génome. Les nombreux réarrangements du génome de *T. kodakaraensis* vis-à-vis de ceux des *Pyrococci*, corrélés à la faible proportion d'IS et de répétitions reflètent simplement une distance évolutive ou suggèrent une accélération des réarrangements dans les environnements extrêmes.

La fantastique **plasticité** du génome de *Thermococcus kodakaraensis* peut s'expliquer par la multitude d'éléments mobiles rencontrés : intéines, transposases et régions virales intégrées.

Ce génome renferme 7 transposases proches de RofB, de la famille des IS605 (COG0675) qui pourraient participer à la duplication de séquences créant des points chauds de recombinaison.

Le point le plus intéressant, en relation avec les éléments génétiques mobiles, est la présence de **quatre éléments viraux intégrés** notés TKV1 à TKV4 dans le génome de *T.kodakaraensis*. Ces éléments ont été annotés en raison de leur ressemblance avec les formes intégrées de certains plasmides et virus de *Sulfolobus* conduisant à l'insertion de l'élément génétique mobile dans un ARNt (Peng *et al.,* 2000). Ce type de région est bordé par une intégrase partionnée. Des éléments intégrés « similaires » ont également été détectés dans les génomes des Euryarchaea *P. horikoshii* (Makino *et al.,* 1999) et *Methanocaldococcus jannaschii (She et al., 2001)*. L'intégrase portée par l'élément génétique mobile est une tyrosine recombinase dévolue à l'intégration. Or, en 2006 (Clore *et al.,* 2006), il a été montré que le virus SSV1 délété de son intégrase possède toujours la capacité à s'intégrer dans le génome de son hôte.

Dans TKV1 (23.592pb) TKV2 (27.204pb) TKV3 (27.886kb) et TKV4 (27.886kb), les gènes correspondant à l'extrémité N-terminale de l'intégrase chevauchent avec la partie 5' codant un $ARNt^{Val}$, $ARNt^{Glu}$ et $ARNt^{Arg}$. Ces régions chevauchantes (48pb) sont prédites pour être les sites d'attachement (*att*) lors de l'intégration. La partie codant l'extrémité C-terminale de l'intégrase possède également ces séquences (100% d'identité pour TKV1, 71% pour TKV2 et 29% pour TKV3). Ces séquences *att* sont très conservées dans îlots génomiques de *P. horikoshii* et *M. jannaschii* mais également chez les éléments intégratifs pRN et SSV1 de *Sulfolobus*. L'orientation du site *att* dans TKV4 est la même que celle de l'$ARNt^{Leu}$ alors qu'elle est différente pour TKV1-TKV3, suggérant une différence phylogénétique de TKV4. D'autre part, TKV4 contient de petits ORFs dont la plupart ne possèdent pas d'homologues dans les bases de données. TKV2 et TKV3

sont génétiquement proches, possédant notamment une région de 8,1kb dont la composition de bases est pratiquement identique.

Au sein de ces éléments viraux intégrés, une faible proportion de gènes possède des homologues dans les bases de données. Seulement quelques orthologues peuvent être trouvés. Ils sont généralement impliqués dans la réplication de l'ADN, le complexe MCM dans TKV1 (TK0096) et TKV4 (TK1361), le facteur de processivité (PCNA) dans TKV3 (TK0582). **Cette faible proportion d'homologues rencontrés dans les bases de données pourrait s'expliquer par une connaissance très restreinte des plasmides et des virus de Thermococcales.** Le faible nombre d'éléments génétiques mobiles caractérisés chez les Thermococcales limite les enseignements sur leur éventuelle mobilité et sur la dynamique de ces éléments viraux intégrés. De plus amples détails sur la diversité et les mécanismes d'intégration de ces îlots génomiques sont détaillés dans la section V de cette introduction (page 48).

IV. Les plasmides d'*Archaea*

Des plasmides ont été identifiés dans les trois grands groupes phénotypiques d'*Archaea* : halophiles extrêmes, méthanogènes et hyperthermophiles. Toutefois en raison de leur découverte relativement récente, les connaissances de ces plasmides sont encore **peu documentées** et loin d'être équivalents entre les différents groupes taxonomiques.

1. Les plasmides des halophiles extrêmes

Les halophiles extrêmes appartiennent au domaine des Euryarchaea. Leur génome est constitué de plusieurs réplicons circulaires double brin (Tableau 4).

Tableau 4 Génomes d'Halobacterium à réplicons multiples

Haloarcula marismortui	Baliga et al., 2004
chromosome I	3,1 Mb
chromosome II	290 kb
pNG100/200/300/400/500/600/700	33-410 kb
Halobacterium sp. NRC-1	
chromosome	2 Mb
pNRC100	191 kb
pNRC200	365 kb
Halobacterium sp. R1	Unpublished
chromosome	2 Mb
pHS1	148kb
pHS2	195 kb
pHS3	284 kb
pHS4	41 kb
Halobacterium sp. GRB	Ebert et al., 1984
chromosome	2 Mb
pGRB205	305 kb
pGR90	90 kb
pGRB37	37 kb
pGRB1	1,8 kb
Haloferax volcanii DS2	Séquençage en cours Ebert
chromosome	2,8 Mb
pHV4	690 kb
pHV3	442 kb
pHV2	86 kb
pHV1	6 kb
Halobacterium salinarium	Ebert et al., 1984
chromosome	2 Mb
pGRB205	305 kb
pGR90	90 kb
pGRB37	37 kb
pGRB1	1,8 kb

Le plus grand réplicon est considéré comme le chromosome. Les réplicons de taille intermédiaire sont classés en minichromosomes et mégaplasmides. Si un réplicon utilise un mécanisme de réplication similaire au chromosome et porte des gènes essentiels il est alorsconsidéré comme un minichromosome. Néanmoins, sans confirmation expérimentale, la présence sur un réplicon de protéines de réplication de type plasmidique n'est pas une preuve de son mode de réplication.

Une vingtaine de plasmides ont été décrits chez les halophiles, constituant un groupe varié. Certains de ces plasmides ont été séquencés (Tableau 5). Bien que la majorité de ces plasmides soient cryptiques, des fonctions spécifiques ont pu être attribuées à certains d'entre eux.

Tableau 5 plasmides d'haloarchées séquencés

Plasmide	Taille (bp)	GC %	Souche	Ref	Num accession	Réplication
pHGN1	1 765	62,7	*Haloarchaeal coccus LOC-1*	Hall 1989	NC_002124	RCR SFI
pNG100	33 303	54,2	*Haloarcula marismortui*	Baliga 2004	NC_006389	?
pNG200	33 452	55,6	*Haloarcula marismortui*	Baliga 2004	NC_006390	?
pNG300	39 521	60	*Haloarcula marismortui*	Baliga 2004	NC_006391	?
pNG400	50 060	57,4	*Haloarcula marismortui*	Baliga 2004	NC_006392	?
pNG500	132 678	54,5	*Haloarcula marismortui*	Baliga 2004	NC_006393	?
pNG600	155 300	=Pub	*Haloarcula marismortui*	Baliga 2004	NC_006394	Mini chromosome
pNG700	410 554	59,1	*Haloarcula marismortui*	Baliga 2004	NC_006395	?
pSCM201	3 463	59,9	*Haloarcula sp. AS7094*	Sun 2006	NC_006426	Théta uni
pGRB1	1 781	40,8	*Halobacterium halobium*	Hacket 1990	X52610	RCR SFI
pHSB1	1736	64	*Halobacterium sp. SP3*	Hacket 1989	X07128	RCR SFI
pHSB2	1 781		*Halobacterium sp. SP3*	Akhmanova 1993	X66324	RCR SFI
pHK2	4111		*Haloferax alicantei*	Holmes 1991	L29110	RCR SFI
pHH205	16 341	61,1	*Halobacterium salinarum J7*	Ye 2003	NC_003158	?
pPL47	46 867	47,7	*Haloquadratum walsbyi*	Bolhuis 2006	NC_008213	?
pZMX101	3 918	66	*Halorubrum saccharovorum*	Zhou 2007	NC_004531	Théta uni

Par exemple, les mégaplasmides pHH1 (150kb) et pNRC1 (200kb) possèdent une région de 9kb impliquée dans la synthèse de vacuoles de gaz qui permettent à leur hôte *Halobacterium halobium* de flotter à la surface de l'eau saumurée des marais salants. Ces *Archaea* étant photosynthétiques facultatives, la flottabilité conférée par le plasmide permet d'avoir accès plus facilement à la lumière et à l'oxygène (Pfeifer *et al.,* 1989). Ces gros plasmides sont, comme le génome de leur hôte, littéralement truffés de séquences d'insertions (IS) (Pfeifer *et al.,* 1989) et responsables de la forte variabilité génétique (par recombinaison et transposition) (Pfeifer *et al.,*

1988). Le plasmide pΦHL (10kb), quant à lui, confère à son hôte *H. salinarium* l'immunité au virus ΦH qui infecte normalement cette espèce (Gropp *et al.,* 1992).

Il existe également de nombreux petits plasmides à réplication par **cercle roulant**. Extrêmement abondant chez *Halobacterium salinarum,* HSB2 (1736pb) est le plus petit plasmide d'*Archaea* (Akhmanova *et al.*, 1993). Les réplicons à réplication par cercle roulant comportent également les plasmides d'halophiles : pHGN1 (Hall *et al.,* 1989), pGRB1 (Hackett *et al.,* 1990), pHSB1 (Hackett *et al.,* 1989). Ils possèdent un long ORF codant une protéine Rep partageant entre 30% à 88% d'identité de séquence malgré des origines géographiques très éloignées. Ces protéines possèdent les motifs typiques des phages et des plasmides de *Bacteria* se répliquant par cercle roulant (voir page 11). L'hypothèse de réplication par cercle roulant a été confirmée par la détection d'intermédiaire de réplication simple brin. La forte conservation de séquence a mis en évidence des phénomènes de microhétérogénéité de séquences. Cette **microhétérogénéité** correspond à des mutations ponctuelles entre isolats d'une même espèce. C'est le cas des plasmides pHSB1 (1736pb) et pHSB2 (1781pb) partageant 80% d'identité de séquence. Les protéines Rep correspondantes partagent quant à elles 88% d'identité. Ces mutations n'étant pas dispersées aléatoirement le long de la séquence, elles ne sont pas le résultat d'erreurs de réplication. Ces mutations sont organisées en clusters et leur principale conséquence est une différence d'un facteur 5 du nombre de copies entre ces deux plasmides (Akhmanova *et al.* 1993).

Certains plasmides halophiles ont servi de base à l'élaboration de vecteurs navettes : tels les plasmides pHV2 et pHK2 d'*Haloferax* et pHH1 d'*Halobium* (Cline et al., 1992). La plupart de ces vecteurs portent un gène de résistance à la novobiocine ou à la mévinoline qui constituent des marqueurs de sélections très efficaces (Holmes *et al.,* 1994). Ceci explique en partie pourquoi le développement d'outils génétiques a été plus précoce chez les halophiles qu'au sein des autres phyla d'*Archaea*.

2. Les plasmides des méthanogènes

Les plasmides mis en évidence chez les méthanogènes ont des tailles comprises entre 4,4 et 64kb. Le petit plasmide cryptique pME2001 (4,4kb) de *Methanobacterium thermoautotrophicum,* souche *Marburg,* est l'un des plasmides les mieux caractérisés chez les méthanogènes (Bokranz *et al.,* 1990). Il est présent au nombre de 15 à 30 copies par cellules et possède une grande stabilité. Un groupe de trois plasmides, pZF1, pZF2 et pFV1 retrouvés dans différentes souches de *M. thermoformicicum,*présente la particularité de coder un système de restriction-modification (R/M)

qui confère une protection à l'introduction d'ADN étranger, en particulier à celui des archéophages ΦF1 et ΦF3 décrits dans certaines souches de cette même espèce (Nolling *et al.,* 1992).

Le premier vecteur navette développé pour les méthanogènes utilise le plasmide cryptique pURB500 (8,3kb) de la souche *Methanococcus maripaludis* associé à un plasmide dérivé de pBR322 d'*E.coli* portant un gène de résistance bactérien à la puromycine (Tumbula *et al.,* 1997). Un autre vecteur navette, pWM207, capable de se répliquer dans sept espèces de *Methanosarcina* a été élaboré à l'aide du plasmide pC2A (5,5kb) isolé de *Methanosarcina acetivorans* (Metcalf *et al.,* 1997).

3. Les plasmides des hyperthermophiles

3.1 *Les plasmides des Crenarchaea hyperthermophiles*

Tous les plasmides décrits chez les Crenarchaea hyperthermophiles ont été isolés chez les Sulfolobales. Ce sont des micro-organismes thermoacidophiles (pH optimal 2 à 4), autotrophes ou hétérotrophes aérobies (ou anaérobies facultatives), susceptibles d'utiliser des composés soufrés comme accepteur terminal d'électron. Ces micro-organismes sont isolés d'environnements terrestres volcaniques comme les solfatares. En raison de plus simples conditions de cultures en aérobiose, ces micro-organismes sont plus documentés et les différents plasmides qu'ils hébergent, pour l'essentiel cryptiques, ont fait l'objet d'études approfondies (Lipps 2006).

Les premiers **plasmides conjugatifs** archéens, à ce jour les seuls connus, ont été découverts dans l'ordre des Sulfolobales. Le premier décrit, le plasmide pNOB8 (45kb), a été isolé d'une souche de *Sulfolobus japonicus* (Schleper *et al., 1995*). Le transfert horizontal s'effectue par contact direct entre les surfaces cellulaires à l'image de certaines bactéries gram-positives. Il semblerait que la conjugaison induise une forte réplication du plasmide (20-40 copies). Une fois le plasmide transféré dans les cellules receveuses, un système de contrôle codé par le plasmide réduirait progressivement le nombre de copies. Une dizaine d'autres plasmides conjugatifs ont depuis été identifiés chez les différentes souches de *Sulfolobus islandicus* (Prangishvili *et al.,* 1998) ; (Stedman *et al.,* 2000). Ils sont classés en 4 sous-familles : pNOB, pING, pSOG et le groupe des plasmides pARN et pHVE. La taille de ces plasmides s'échelonne entre 25 et 41kb. Ils sont sujets à de nombreuses variations génomiques du fait de la présence d'IS mais aussi de séquences

répétées directes et inversées impliquées dans des mécanismes de recombinaison, aboutissants à la formation de variants plus ou moins stables. Certains, ayant subits des délétions très importantes, ont perdu leur capacité à s'autotransférer mais peuvent néanmoins être **mobilisés** par des plasmides conjugatifs complets.

Seulement deux gènes impliqués dans la conjugaison ont pu être identifiés sur ces plasmides en se basant sur leurs similarités de séquences avec les protéines TraG et TrbE, codées par les plasmides conjugatifs bactériens (She *et al.,* 1998). Ces deux gènes font parti d'un ensemble de gènes très conservés chez tous les plasmides conjugatifs de *Sulfolobus*, regroupés en une région nommée *tra.* Elle s'étend sur environ 10kb, taille nettement plus restreinte que celle portée par les plasmides conjugatifs bactériens. De plus, chez pNOB8, des homologies avec des protéines appartenant à la famille des transposases, hélicases et systèmes de partition (ParA, ParB et ParC), ont pu être identifiées. pNOB8 est également à l'origine du vecteur d'expression pNOB8:lacS (Elferink *et al.,* 1996). Depuis ces travaux pionniers, d'autres plasmides conjugatifs ont été isolés, notamment pSOG1 et pSOG2 présents dans la souche *Sulfolobus* SOG2/4 (Erauso *et al.,* 2006).

Le groupe **des plasmides cryptiques** est nommé famille pRN. 5 petits plasmides, pDL10 (Kletzin *et al.,* 1999), pHEN7 (Kletzin *et al.* 1999), pRN1 et pRN2 (Keeling *et al.,* 1998), ainsi que pSSVx (Arnold *et al.,* 1999) constituent cette famille. Ils présentent une organisation génomique similaire et bipartite : une région commune très conservée et une partie variable. Tous ces plasmides ont été séquencés et sont présents en un nombre élevé de copies dans les cellules. La région conservée de la famille pRN comprend 3 ORFs (i) une protéine de régulation (Plr) qui pourrait constituer une nouvelle sous-classe de « Leucine Zipper » ; (ii) une protéine de régulation du nombre de copies CopG ; (iii) une protéine de réplication qui n'est pas conservée sur différents plasmides, pouvant être une protéine Rep ou une protéine multifonctionnelle qui possède les activités hélicase, primase et ADN polymérase (Lipps 2004; Lipps *et al.,* 2004). pSSVx est un plasmide de 5,7kb (Arnold *et al.* 1999), isolé de la souche *Sulfolobus islandicus* REY15/4. Cette souche contient également le fusellovirus SSV2. Le plasmide possède la capacité de se transférer à *Sulfolobus solfataricus* par co-transfection en présence de SSV2, mais dans des capsides plus petites. L'existence de répétitions en tandem identiques à celles rencontrées sur SSV2 serait impliquée dans ce mécanisme de transfert (Aucelli *et al.,* 2006; Wang *et al.,* 2007).

Un fait intéressant, mis en évidence pour certains plasmides précédemment cités (conjugatifs et cryptiques), est leur capacité d'intégration dans le chromosome de leur hôte au niveau d'un ARNt

selon un mécanisme site-spécifique médiée par une intégrase de type phagique codée par le plasmide (voir p.50).

3.2 *Les plasmides des Euryarchaea hyperthermophiles*

Chez les Euryarchaeota, un petit plasmide, pGS5 (2,8kb) a été découvert chez *Archaeoglobus fulgidus*, une archée hyperthermophile marine réduisant les sulfates. La particularité de ce plasmide est la **topologie** de son ADN qui est surenroulé négativement à l'inverse des autres plasmides d'*Archaea* hyperthermophiles (Lopez-Garcia *et al.,* 2000).

Des plasmides ont également été mis en évidence au niveau de l'ordre des *Thermoplasmatales* (hyperacidophiles terrestres capables de croissance à pH<1). Le plasmide pTA1 (15,2kb) a été décrit chez *Thermoplasma acidophilum* (Luo *et al.*, 1995), une thermoacidophile dont l'optimum de croissance est réalisé à une température de 56°C à pH 1,8. Au long des 15,7kb, 8 ORFs ont été annotés et seul l'ORF1 possède un homologue dans les bases de données. L'ORF1 possède des **similarités de séquences avec Cdc6**, une protéine impliquée dans l'initiation de la réplication chez les *Archaea* et les *Eukarya*. L'homologue trouvé dans pTA1 est très similaire à Tvo3, un des 3 homologues de Cdc6 du génome de *Thermoplasma volcanium*. Les analyses phylogénétiques suggèrent que pTA1 serait originaire du chromosome d'un *Thermoplasma* (Yamashiro *et al.*, 2006).

Plusieurs souches de *Picrophilus oshimae* se sont révélées porteuses d'un plasmide (Schleper *et al.* 1995) pKAW (8,3kb) et pKAV2 (8,8kb) (Schleper *et al.* 1995). Des hybridations en Southern ont montré que les deux plasmides partagent des régions homologues mais ils n'ont pas été séquencés.

Au sein de l'ordre des Thermococcales, plusieurs plasmides ont également été détectés et ont fait l'objet d'études au laboratoire LM2E depuis plusieurs années. Un criblage réalisé sur 52 souches du pacifique oriental a révélé que onze souches d'entre elles contenaient un plasmide (Benbouzid-Rollet *et al.,* 1997). Parmi celles-ci, trois souches proches de *Thermococcus stetteri* portent chacune deux plasmides, pSN559 (3kb) et pLN559 (18kb). Les autres souches contiennent chacune un plasmide dont la taille varie entre 3kb et plus de 20kb.

Aujourd'hui, **seul trois plasmides apparentés** ont été séquencés. Ils possèdent tous une taille voisine de 3,5kb et deux ORFs, dont l'un est une protéine Rep, initiatrice de la **réplication par cercle roulant** (RCR). Le plasmide pGT5, présent en grand nombre de copies (30 par cellule) dans la souche *Pyrococcus abyssi GE5* (Erauso *et al.,* 1993) a fait l'objet d'une étude détaillée. Les origines double et simple brin (*dso* et *sso*) nécessaires à ce mode de réplication ont également été identifiées dans la séquence. La fonction de ces différents éléments a été démontrée par des études *in vitro* (Erauso *et al.*, 1996). pGT5 est à la base de la construction de la première génération de vecteur de clonage dans le groupe des *Euryarchaeota* hyperthermophiles (Lucas *et al.,* 2002). Ce vecteur navette pYS2 a été obtenu en fusionnant une partie du plasmide pGT5 à un vecteur dérivé de pUC18 pour propagation dans *E.coli*. Il porte le gène pyrE de *S. acidolcaldarius* qui permet de complémenter efficacement des mutants auxotrophes à l'uracile ($pyrE^-$). Un second plasmide de *Thermococcales* a été isolé puis séquencé (Ward *et al.,* 2002). pRT1, de la souche *Pyrococcus sp. JT1*, présente 41% d'identité de séquence avec pGT5. Comme ce dernier, il se réplique par cercle roulant et possède la protéine p63 de la famille Rep.

Récemment un troisième plasmide, pTN1, isolé de *Thermococcus nautilii* a été décrit (Soler *et al.* 2007). La protéine Rep74 est homologue à Rep75 (environ 38% d'identité). Cette nouvelle séquence, alignée aux deux précédentes, met en lumière une erreur d'alignement faite par Ward lors de la description du plasmide pRT1 conduisant à une mauvaise définition du motif III de cette protéine. Cet alignement montre également que les protéines Rep74/75 constituent une nouvelle famille dont le domaine Rep ést fusionné à une extrémité N-terminale de fonction inconnue. Cet agencement rapproche d'avantage ces protéines des transposases de certaines séquences d'insertions que des plasmides à réplication par cercle roulant. Contrairement aux autres plasmides, le second ORFs de pTN1 codant la protéine p24 a été caractérisé. Cette protéine possède un segment hydrophobe, une région très chargée et un motif en doigt de zinc. Elle assure une très forte compaction de l'ADN simple et double brin par utilisation d'un segment. Etant le premier plasmide de *Thermococcus* caractérisé, il a servi de base à la construction d'un vecteur navette d'expression pour la souche modèle *Thermococcus kodakaraensis* (Santangelo *et al*. 2008).

Les éléments génétiques mobiles (plasmides et virus) possédant de nombreuses caractéristiques communes, il est important de noter qu'un **unique virus** de Thermococcales a été décrit. PAV1 est produit par la souche *Pyrococcus abyssi* GE23 (Geslin *et al.,* 2003). Ce virus, dont les particules virales sont en forme de citron, sont terminées par des fibres caudales et ressemblent morphologiquement au virus SSV1 de *Sulfolobus*. Malgré des essais d'infection sur un grand

nombre d'isolats de Thermococcales, son pouvoir infectieux n'a pas été démontré. PAV1 persisterait sous forme d'état porteur dans son hôte. Le séquençage de son génome (Geslin *et al.,* 2007) a montré qu'une soixantaine de copies du génome virale sont présente dans la cellule et l'absence de copies intégrées dans le chromosome. L'ADN circulaire de 18kb possède 25 ORFs dont 4 annotables. Les plus intéressants sont les ORFs 676 et 678, présentant des similarités avec le domaine lectine/glutamase concanavaline A impliqué dans la reconnaissance virus-hôte. Bien que morphologiquement proche de SSV1, aucune similarité de séquence n'a été mise en évidence. Ce virus appartiendrait à une nouvelle famille.

V. Les éléments génétiques intégrés dans les génomes d'Archaea

La recombinaison site-spécifique (RSS) est un processus évolutivement extrêmement bien conservé. Elle participe activement à la plasticité génomique. Elle est non seulement rencontrée au sein des trois domaines du vivant, mais également dans le monde des plasmides et des virus. A l'inverse des recombinaisons transpositionnelles et homologues, la RSS est conservative puisqu'elle n'engendre ni synthèse ni dégradation des fragments d'ADN impliqués.

1. Mécanisme d'intégration : recombinaison site-spécifique RSS

L'intégration fait intervenir une recombinase site-spécifiques portée par l'élément génétique mobile. Ces recombinases peuvent être de la même famille que celles impliquées dans divers autres mécanismes moléculaires, tels que la résolution des dimères de réplicons permettant une ségrégation correcte des cellules filles ou le contrôle de l'expression génique. La recombinaison est le principal moteur de la plasticité des génomes.

1.1 *Partenaires de la RSS*

On distingue deux familles de **recombinases** site-spécifique en fonction des résidus impliqués dans la catalyse (Tableau 6). Ces deux familles présentent quelques points communs en terme de catalyse, notamment une très grande spécificité des sites de recombinaison, l'indépendance vis-à-vis de cofacteurs énergétiques et la formation d'intermédiaires réactionnels covalents l'ADN.

Tableau 6 Les deux familles de recombinases

Famille des Tyrosine Recombinases >300 membres	Famille des Sérine Recombinases 30 membres
Très faible conservation de séquence primaire	Bonne conservation de séquence primaire
Tyrosine catalytique en C-terminal	Sérine catalytique en C-terminal
Complexe covalent intermédiaire 3'-P-Tyr	Complexe covalent intermédiaire 3'-P-Ser
Mécanisme séquentiel	Mécanisme concerté
Coupure décalée de 6 à 8pb	Coupure décalée de 6 à 8pb
Intermédiaire en jonction de Holliday	Rotation de 180°
Intégrases, Recombinases	Invertases, Résolvases, Intégrases
ex: λ-Int, Intl, Cre, Flp, Xer…	ex: Gin, Hin, γδ-Résolvases, ΦC31, R4…

Les **sites de recombinaison** sont quant à eux très divers et reflètent la variété des mécanismes concernés. Cependant, un certain nombre de points communs peuvent être définis puisque dans tous les cas le site minimal de recombinaison est un assemblage de séquences répétées inversées séparées par une séquence non palindromique de quelques paires de bases (Figure 9). Cette séquence intermédiaire appelée « spacer » sert à guider la réaction de recombinaison en lui fournissant sa directionalité. A une exception près (les intégrons), les deux sites de recombinaison ont un spacer strictement identique en séquence. Les sites de fixation qui l'encadrent ne sont pas nécessairement identiques et, en général, le consensus entre les quatre sites de fixation est assez flexible. Ce point est assez paradoxal compte tenu de la sélectivité très stricte des recombinases pour leur site de recombinaison.

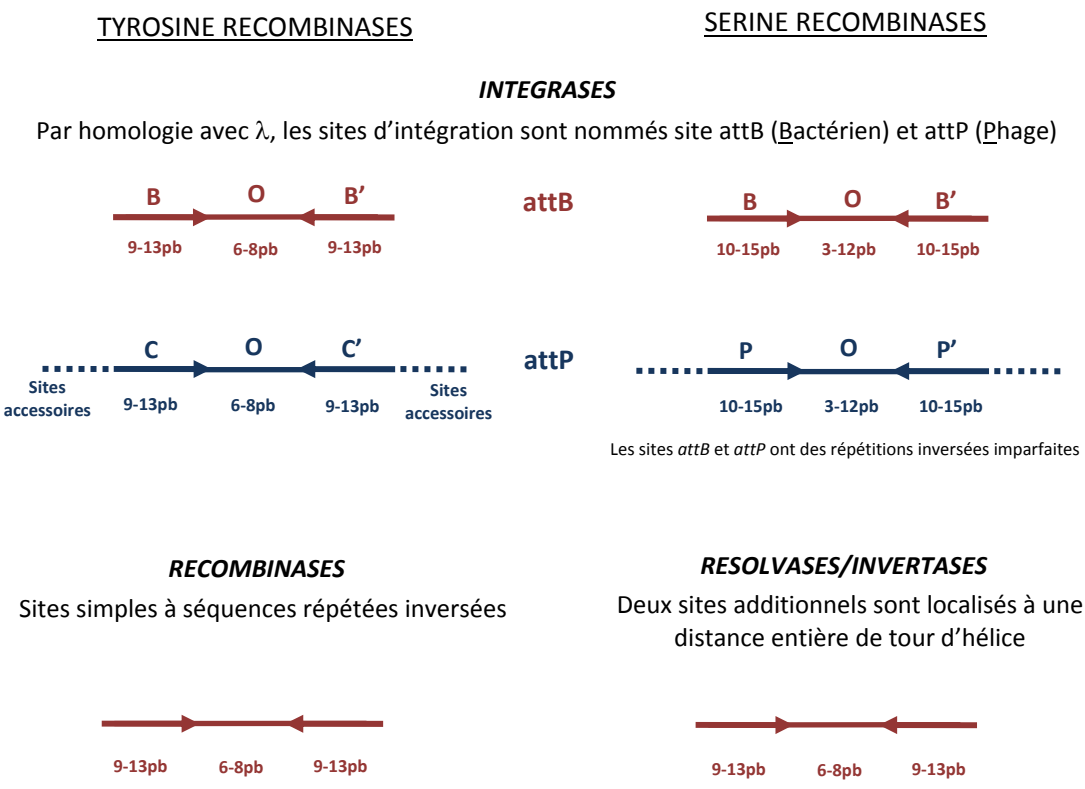


Figure 9 Sites de recombinaison des différentes familles de recombinases

Dans de nombreux cas, outre cette séquence minimale de recombinaison (appelée site CORE), le site de recombinaison comprend également des sites secondaires qui permettent la fixation de protéines accessoires indispensables à la formation du complexe synaptique. Ces sites secondaires sont dans certains cas des sites de fixation des recombinases qui, à cet endroit, n'exerceront pas une activité catalytique mais un rôle de structuration spatiale.

1.2 *Mode d'action des Recombinases*

Tyrosine Recombinases

La recombinaison requiert un assemblage quaternaire sous forme d'homoquadramère dans lequel chaque monomère fixe une des extrémités de l'ADN à recombiner. Le mode de coupure séquentiel implique l'activation de deux monomères et l'inactivation des deux autres. Le remodelage du complexe synaptique est indispensable pour que le second cycle de coupure-religature intervienne. Ce changement conformationnel est induit par la migration de la jonction de Holliday.

Sérine Recombinases

Comme pour les Tyrosine recombinases, quatre monomères sont requis pour couper les quatre brins d'ADN. Dans ce cas, les quatre évènements de coupure interviennent en même temps et le remodelage est lié à une rotation de 180° d'une moitié du complexe nucléoprotéique par rapport à l'autre. Cette rotation est le résultat de modifications dans les interactions protéine-protéine après formation du complexe covalent.

1.3 *Les intégrases des Archaea*

La première intégrase d'*Archaea* a été découverte dans le génome du **virus SSV1** de *Sulfolobus shibatae* (Palm *et al.*, 1991). Int-SSV1 est une tyrosine recombinase. Ce virus s'intègre sous forme de provirus par recombinaison site-spécifique au niveau d'un $ARNt^{Arg}$ (Schleper *et al.,* 1992; Muskhelishvili *et al.,* 1993). La particularité de l'intégration de SSV1 est la partition du gène de l'intégrase en deux parties, car le site *attP* est localisé dans la première moitié du gène correspondant à la région N-terminale de l'intégrase. Cette propriété contraste avec toutes les autres intégrases connues. Après intégration, le gène *int* code un petit fragment N-terminal IntN (70AA) et un plus grand fragment C-terminal IntC (270AA) bordant l'élément viral intégré. A l'heure actuelle, nous ne savons pas si IntN et IntC peuvent former une intégrase active. Néanmoins, les formes épisomales et intégrées étant toujours présentes simultanément dans la cellule, l'excision du provirus SSV1 pourrait être catalysée par la forme circulaire libre du virus SSV1, contenant une intégrase entière (Clore *et al.,* 2007).

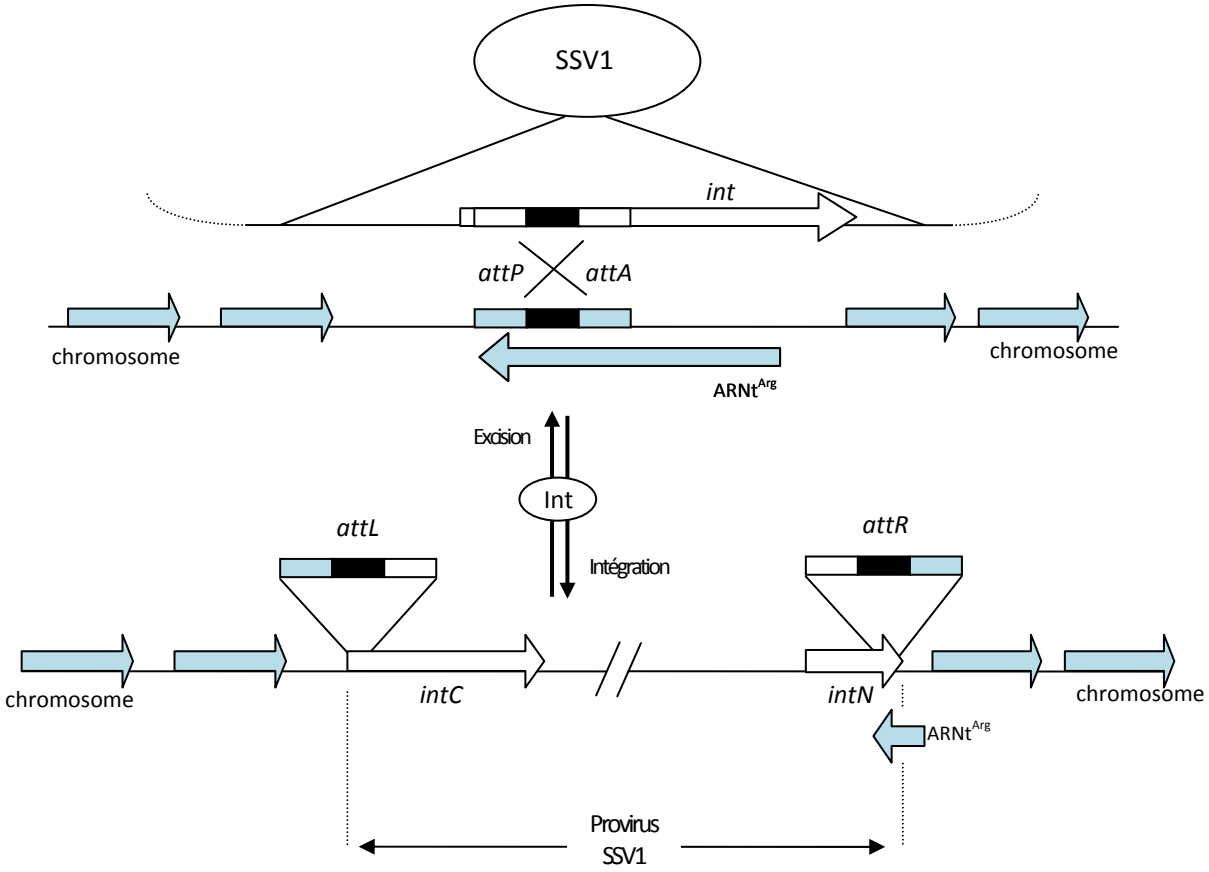


Figure 10 Intégration et de l'excision du virus SSV1

Représentation schématique de l'intégration du virus SSV1 sous forme de provirus dans le génome chromosomique de *Sulfolobus shibatae*.

Récemment, des intégrases produisant un gène *int* entier après intégration ont été découvertes dans les génomes chromosomiques de *Sulfolobus*. Ce type d'intégrase est également présent dans les génomes de certains plasmides conjugatifs de *Sulfolobus* (Greve et *al.,* 2004). De plus, il a été démontré que le plasmide pNOB8 peut s'intégrer de manière site-spécifique dans le chromosome de son hôte (She *et al.,* 2004).

Comme les recombinases bactériennes, leurs homologues archéennes sont très diverses en séquence et présentent seulement des identités de séquences au niveau du domaine catalytique. La conservation du site actif a été confirmée par les données de cristallographie. La comparaison de structures tridimensionnelles d'intégrases archéennes révèle l'importance de six résidus dans le site actif, organisés en motifs BoxI, BoxII et Kβ (Van Duyne 2001). Les alignements multiples de séquences révèlent une différence de consensus entre les intégrases bactériennes et archéennes de type SSV1 et de type pNOB8 (Tableau 7).

Tableau 7 Consensus des motifs d'intégrases

Intégrase		Consensus
Bactéries		R…K…H/KxxR…H/W…Y
Archaea	type SSV1	R…R…KxxR…R…Y
Archaea	type pNOB8	R…K…YxxR…R…Y

SSV1-Int est la seule intégrase ayant été caractérisée en détail. L'utilisation d'une enzyme recombinante de SSV1-Int a prouvé que l'intermédiaire 3'-phospotyrosine fixe les sites *attA* et *attP* par l'intermédiaire de la tyrosine conservée Y314 (Serre *et al.,* 2002). Ce travail montre que le mécanisme général d'intégration est similaire aux *Bacteria*.

2. Les différents types d'éléments intégrés (IEs).

Il existe deux groupes d'éléments intégrés chez les *Archaea,* définis par les deux familles d'intégrases (types SSV1 et pNOB8). Alors que les éléments de type pNOB8 s'intègrent selon un mécanisme similaire à celui des *Bacteria*, l'intégration des éléments de type SSV1 conduit à la partition du gène de l'intégrase en *intN* et *intC*. (Figure 11)

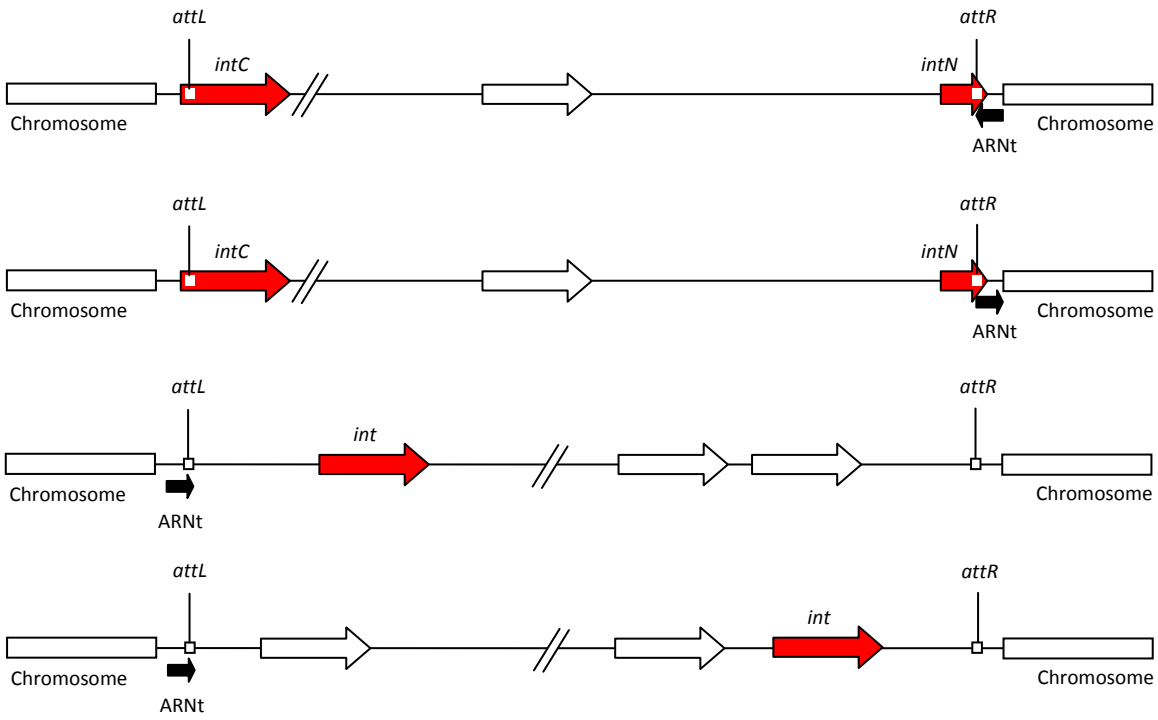


Figure 11 Profils d'intégration des éléments intégrés d'Archaea

2.1 *Les éléments de type SSV1*

La comparaison des génomes de *Sulfolobus solfataricus* et du plasmide PHEN7, a mis en évidence la présence d'une copie intégrée de ce plasmide sur le chromosome (Peng *et al.* 2000). Ce plasmide est bordé par des ORFs très similaires à *intN* et *intC* codés par le provirus SSV1. De la même manière, deux provirus putatifs ont été trouvés dans le génome de *Pyrococcus sp. OT3*, également bordés par *intN* et *intC* typiques de SSV1 (Makino *et al. 1999*). Sur cette base de nombreux *intN* et *intC* ont été identifiés dans d'autres génomes d'*Archaea* : *Aeropyrum pernix*, *Sulfolobus acidocaldarius* et *Sulfolobus tokodaii* (She *et al.* 2001), *Pyrococcus horikoshii* et *Thermococcus kodakaraensis* (Fukui *et al.* 2005) et chez les Methanococcales : *Methanococcus voltae A3, M. maripaludis* S2, *M. maripaludis* C6 et *M. maripaludis* C7, étrangement pas chez la souche voisine *M.maripaludis* C5. Cette dernière observation souligne l'**importance de la contribution des phages et de leurs formes intégrées dans la différentiation intra-espèce** (Keswani *et al.,* 1996). Contrairement aux Crenarchaea, ces éléments intégrés d'Euryarchaea ne possèdent pas encore d'équivalent sous forme épisomale… ils restent à découvrir…

Pour chaque *intC*, il y a un *intN* correspondant, car chaque paire de *intC-intN* partage la même répétition directe en tandem. De cette manière ils bordent les EIs. L'affiliation d'un EI au type SSV1 se fait sur la présence de *intN* et *intC* typique du provirus SSV1.

13 éléments de ce type ont été détectés (Tableau 8). Néanmoins, il existe certainement de nombreux autres éléments intégrés qui ne sont pas aussi facilement détectables. La plasticité des génomes pouvant faire disparaître ces bornes facilitant la détection des éléments intégrés lors des analyses automatiques des génomes.

2.2 *Les éléments de type pNOB8*

Les éléments intégrés codant l'intégrase de type pNOB8 ont d'abord été découverts dans le génome de *S. tokodaii (Kawarabayasi et al., 2001)*. Les critères permettant la détection de ce type d'élément sont les suivants : (i) pST2 contient un gène *int* typique de pNOB8 qui est resté intact après intégration ; (ii) il est bordé par une répétition directe dont une extrémité correspond à celle d'un $ARNt^{Met}$ et (iii) de nombreux ORFs sont très similaires à ceux du plasmide conjugatif pNOB8. Ce type d'IE est dit « type pNOB8 ». Peu de génomes ont été analysés à la recherche de ces éléments, seulement 8 éléments de type pNOB8 ont été découverts (Tableau 8)

Tableau 8 Eléments intégrés caractérisés dans les génomes d'Archaea

Organisme	Nom	Taille (pb)	Répétition directe (pb)	Gène chevauchant	Annotation	Référence
Type SSV1						
Euryachaeota						
Pyrococus sp. OT3	PT1	21,7	48	ARNt ^{Ala} <	Provirus putatif	
Pyrococus sp. OT3	PT2	4,1	46	ARNt ^{Val} <	Provirus putatif	Makino *et al.*, 1999
Thermococcus kodakaraensis	TKV1	23,6	48	ARNt ^{Val} <	MCM, régulateurs de transcription	Fukui *et al.*, 2005
Thermococcus kodakaraensis	TKV2	27,2	48	ARNt ^{Glu} <	ATPase, régulateurs de transcription	Fukui *et al.*, 2005
Thermococcus kodakaraensis	TKV3	27,9	48	ARNt ^{Arg} <	PCNA, enzymes	Fukui *et al.*, 2005
Thermococcus kodakaraensis	TKV4	18,8	42	ARNt ^{Leu} <	MCM, régulateurs de transcription	Fukui *et al.*, 2005
Crenarchaeota						
Aeropyrum pernix	AP1	17,8	65	ARNt ^{Leu} <	Inconnu	She *et al.*, 2001
Sulfolobus acidocaldarius	pSA1	8,7	44	ARNt ^{His} <	Plasmide	She *et al.*, 2004
Sulfolobus acidocaldarius	SA2	5,7	44	ARNt ^{Arg} <	Inconnu	She *et al.*, 2004
Sulfolobus acidocaldarius	pSA3	32,5	50	ARNt ^{Glu} <	plasmide conjugatif fégénéré, 2 enzymes	Chen *et al.*, 2005
Sulfolobus solfataricus	pXQ1	7,3	45	ARNt ^{Val} <	plasmide de type pRN	She *et al.*, 2001
Sulfolobus solfataricus	XQ2	67,7	45	ARNt ^{Arg} <	11 enzymes	She *et al.*, 2001
Sulfolobus tokodaii	pST1	6,7	48	ARNt ^{Ala} <	plasmide de type pRN	She *et al.*, 2002
Type pNOB8						
Euryachaeota						
Natromonas pharaonis	PL43	36,9	25	ARNt ^{Leu} <	3 copies de PL23 dégénérées	Falb *et al.*, 2005
Methanococcus jannaschii	MJ1	30,5	69	ARNt ^{Ser} <	MCM	Makino *et al.*, 1999
Crenarchaeota						
Sulfolobus tokodaii	pST2	44,8	39	ARNt ^{Met} <	ressemble à pNOB8	She *et al.*, 2002
Sulfolobus tokodaii	pST3	6,9	42	ARNt ^{Arg} <	plasmide de la famille pRN	She *et al.*, 2002
Sulfolobus tokodaii	ST4	66,1	29	ARNt ^{His} <	25 enzymes	She *et al.*, 2002
Sulfolobus tokodaii	ST5	nd	nd	ARNt ^{Gly} <	très dégénéré	She *et al.*, 2004
Sulfolobus acidocaldarius	SA4	7,5	43	ARNt ^{Val} <	ressemble à un plasmide	Chen *et al.*, 2005

3. Dynamique des IEs et Limitations de leur détection

La présence de nombreux gènes *int* dans les génomes d'*Archaea* montre **l'importance des phénomènes d'intégration**. Les méthodes de détection se limitent aux éléments « complets », probablement récemment intégrés. En effet, la présence de séquences répétées favorise les réarrangements du génome et donc la destruction de l'élément clairement bordé par *intN* et *intC*. Les séquences des génomes disponibles n'étant que des photos instantanées, les nouvelles technologies de séquençage permettront l'obtention de la séquence d'un génome plus facilement, plus rapidement et pour un faible coût, permettant ainsi d'avoir accès à la dynamique et la plasticité du génome (**dynamogénomique** !).

En plus des plasmides conjugatifs de *Sulfolobus*, de nombreux éléments génétiques mobiles, virus et plasmides, portent un gène *int*. Chez les Euryarchaea, trois virus portent une intégrase : le phage psiM2 de *Methanobacterium thermoautotrophicum* (Pfister *et al.,* 1998) et les halovirus HF1 et HF2 (Tang *et al.,* 2004). Deux petits plasmides cryptiques portant un gène *int* putatif ont également été isolés de méthanogènes : ce sont les plasmides pURB500 de *Methanococcus maripaludis* (Tumbula *et al.* 1997) et le plasmide pC2A de *Methanosarcina acetivorans* (Metcalf *et al.* 1997). Finalement, un grand nombre de plasmides, couvrant l'ensemble des tailles de réplicons des génomes des haloarchaea, possèdent au moins un gène codant une ADN recombinase. C'est le cas par exemple des plasmides pNRC100 et pNRC200 d'*Halobacterium sp.* (Ng *et al.,* 1998), cinq gros ou megaplasmides de *Haloarcula marismortui* (Baliga *et al.,* 2004) et pL23 de *Natromonas pharaonis* (Falb *et al.,* 2005).

Concernant les éléments « mobiles » de type SSV1, seuls les fusellovirus de *Sulfolobus* portent ce type de gène. Les gènes *int* de type SSV1 sont uniquement présents dans six génomes d'*Archaea*, dont trois sont des espèces de *Sulfolobus*. Ceci n'est certainement pas représentatif de la diversité mais résulterait de l'absence de recherche de ces éléments par criblage, comme cela a été réalisé sur des espèces de *Sulfolobus* (Zillig *et al.,* 1998). De nombreux éléments de ce type restent à isoler dans les autres phyla d'*Archaea*.

4. Etude de l'intégration

Récemment, de nombreux hôtes portant des éléments génétiques ont été séquencés ou sont en cours de séquençage, changeant les perspectives d'étude de l'intégration chez les *Archaea*. La première étape de caractérisation de l'élément intégré est l'identification du système d'intégration.

Le premier système d'intégration est celui de type SSV1 et de ses éléments associés, à l'exception de TKV4. La différence entre TKV4 et les autres IEs de type SSV1 concerne l'orientation de *intN* vis-à-vis de l'ARNt cible. Généralement *intN* et l'ARNt ont des orientations opposées. Dans TKV4, ils ont la même orientation car l'intégration s'est faite en amont du gène de l'ARNt.

Les autres mécanismes d'intégration ont été trouvés dans les IEs de type pNOB8, l'intégration se fait toujours en 3' du gène de l'ARNt. Dans un cas, l'ARNt cible et le gène *int* apparaîssent sous forme de tandem à la fin de *attL*. Pratiquement tous les IEs de type pNOB8 procèdent de cette manière. Le plasmide PL23, s'intégrant dans le génome de *Natromonas pharaonis*, est quelque peu différent. Son gène *int* est présent à l'extrémité de *attR*, tandis que l'ARNt cible est à l'extrémité de *attL*. Ceci est la conséquence d'un chevauchement en 3' de *attP* de PL23 (Falb *et al.* 2005)

Les gènes des ARNt sont des sites courant d'intégration chez les *Archaea* et les *Bacteria*. Des centaines de tyrosine recombinases sont connues pour s'intégrer en 3' de l'ARNt (Williams 2002), mais très rarement dans la partie 5'. Seulement 2 IEs, TKV4 (Fukui *et al.* 2005) et Mol38S de *Mesorhizobium loti* (Zhao *et al.*, 2002), présentent ce cas atypique. Ceci est d'autant plus surprenant que l'intégration dans l'extrémité 5' de l'ARNt le désolidarise de son promoteur en le plaçant à distance (Figure 11).

5. Avantage des IEs pour l'hôte ?

Il est possible de distinguer des IEs de grande ou de petite taille au sein des génomes d'*Archaea*. Les petits possèdent plusieurs ORFs codant des fonctions virales ou plasmidiques. Les plus grands sont généralement composés de gènes orphelins ou de fonctions non assignables, probablement car aucune forme libre de l'élément intégré n'a été isolée et séquencée. Néanmoins, quelques gènes possèdent des fonctions identifiables. Ce sont souvent des protéines impliquées dans la réplication ou dans la régulation de la transcription. L'abondance de ces IEs pose plusieurs questions : **que sont-ils, d'où viennent-ils et quel est leur rôle dans le domaine des *Archaea*** ?

Les IEs des chromosomes des trois *Sulfolobus* ont montré un héritage par transfert horizontal. Cette évidence tient à (i) une composition anormale en nucléotides, (ii) la présence d'ORFs affiliés aux plasmides et (iii) les ORFs possèdent un usage des codons atypique. Trois éléments, PXQ1, pST1 et pST3 semblent dérivés des plasmides de type pRN car ils codent au moins deux protéines putatives de la réplication des plasmides de la famille pRN. De plus, pST2 et pSA3 semblent dériver de plasmides conjugatifs de *Sulfolobus*. Bien que pSA3 possède moins d'ORFs conservés que pST2, les ORFs importants pour la conjugaison sont conservés (Chen *et al.,* 2005). pSA3 semble quant à lui dérivé d'un plasmide conjugatif plus divergent que ceux disponibles dans les bases de données. Malgré cette divergence, il a été montré que pS13 intervient dans la « conjugaison » du chromosome de *Sulfolobus acidocaldarius* (Chen *et al.* 2005). Il reste toutefois à déterminer l'importance de pST2 et pSA3 dans ce mécanisme.

Trois IEs de *Sulfolobus* ont transféré de l'information génétique conférant de nouvelles propriétés à leur hôte. Alors que ST4 code 25 enzymes, pSA3 ajoute deux nouveaux gènes dont les orthologues les plus proches sont rencontrés chez *M. jannaschii*, supposant un transfert horizontal entre Euryarchaea et Crenarchaea

XQ2, de son côté, code 11 enzymes, 12 IS complets ou partiels, 13 ORFs ayant une fonction hypothétique et 20 ORFs de fonction inconnue. Parmi les 11 enzymes, 8 sont redondantes chez *S. solfataricus*, mais présente en un seul exemplaire chez *S. tokodaii*. Les 3 autres sont uniques, leurs plus proches homologues proviennent des *Archaea* (34-44% d'indentité). Il reste à déterminer si ces enzymes confèrent un avantage à leur hôte.

Plusieurs autres IEs présents dans les génomes d'Euryarchaea codent des enzymes métaboliques, des régulateurs de transcription, des facteurs de réplication de maintenance du minichromosome (MCM) et des proliferating cell nuclear antigen (PCNA). Il a été suggéré (Forterre 1999) que les appareils de réplication des *Archaea* et des *Eukarya* sont originaires des virus et des plasmides. Ces IEs comportant les gènes MCM, PCNA ou des protéines de contrôle du cycle cellulaire (Cdc6) fournissent **une évidence de la capture de ces familles de gènes par le chromosome des *Archaea***.

6. Stabilité et évolution des IEs

L'étape cruciale d'un HGT est la maintenance stable de l'ADN étranger dans le génome de son hôte. Pour l'intégration de type SSV1, des IEs stables sont toujours produits lors de l'intégration,

car le gène *int* est partitionné en *intN* et *intC*. La forme épisomale de l'élément génétique ne semble alors plus capable de produire l'enzyme fonctionnelle.

La stabilité de l'intégration de l'élément pXQ1 dans *S. solfataricus* à été évaluée par amplification des sites *attA*, *attP*, *attL* et *attR*. Seule la forme intégrée pXQ1 a été retrouvée, et ceci dans toutes les cellules. Cela indique la stabilité de l'intégration de l'élément (She *et al.,* 2002; She *et al.* 2004).

Le mécanisme de capture d'IE de type pNOB8 a été étudié avec le plasmide conjugatif modèle pNOB8. Un mutant *Sulfolobus* sp. NOB8H2 portant seulement la forme intégrée de l'élément a été isolé. La caractérisation de ce mutant révèle la présence d'une délétion dans la région *int* devant intervenir après l'intégration. Il apparaît que la stabilisation de l'élément intégré demande la mutation d'au moins un des éléments nécessaires à l'excision.

Il est traditionnellement accepté que les IEs soient sujets à décomposition par dégénérescence de la séquence ou délétion sous la contrainte de pressions évolutives aboustissant à la production de reliques de ces éléments. Seuls les gènes conférant un avantage évolutif à l'hôte seront conservés. La dégénérescence des IEs démontre le processus séquentiel des HGT aboutissant à la capture d'un gène par un chromosome. Dans le génome de *S. tokodaii*, pST3 est un plasmide « selfish » de type pRN possédant un site *att* intact et une répétition directe de 44pb. Cette observation montre la récente capture de l'élément mobile dans le génome. Pour pST2, une répétition directe de 41pb, avec un mésappariement en son centre, ainsi qu'un gène *int* imparfait suggère que pST2 est un évènement intégratif plus ancien que pST3. pST4 résulterait d'un HGT encore plus ancien que pST2 et pST3 car il y a plus de changements dans les séquences répétées et dans le gène *int*. La répétition directe ne fait que 29pb et la séquence la plus importante, chevauchante au site d'attachement, n'existe même plus. Il y a également un autre vestige de *int* dans le génome de *S. tokodaii*, ST5 indiquant une intégration site-spécifique encore plus ancienne. Une des bornes ne peut être définie par manque de répétions directes et car le gène *int* est dégénéré en un ORF de 74AA difficilement caractérisable par les techniques d'annotation automatisée.

Se servant du génome de *S. tokodaii* comme référence, trois évènements interviennent dans la vie d'un IE. (i) de nombreux HGT s'intègrent de manière site spécifique, (ii) les IEs sont sujets à la délétions et/ou dégénération et (iii) seuls les gènes conférant un avantage évolutif seront conservés et améliorés par optimisation de la séquence. Ces données fournissent une base pour les études des mécanismes HGT conférés par les IEs.

VI. Les transferts horizontaux de gènes

Les échanges génétiques entre les procaryotes ont longtemps été considérés comme un phénomène marginal. Ils sont aujourd'hui de plus en plus perçus comme un facteur important évolutif. Effectivement, l'apport de la génomique soutient une remise en cause du concept de spéciation des bactéries ainsi que la phylogénie basée sur le traditionnel « arbre » en le nuançant par un nouveau modèle sous forme de réseau reflétant l'importance des transferts de gènes. Cette prise de conscience apparaît alors que les biologistes moléculaires révèlent l révèlent l'abondance des éléments génétiques médiant les transferts horizontaux de gènes (HGTs) et leurs implications dans les réarrangements de génomiques. D'autres études visent à mieux comprendre les mécanismes impliquant ces transferts ainsi que les barrières à leur établissement dans un génome, notamment par certaines protections utilisées par les microorganismes pour conserver l'intégrité de leur génome.

Ces informations suggèrent que les communautés microbiennes possèdent un réservoir dynamique de gènes collectifs où de nouvelles combinaisons génétiques obtenues par recombinaison agissent en tant que forces motrice de l'innovation génomique. Cette habilité compense l'inaptitude des espèces microbiennes à « innover » (=recombiner) par voie sexuelle.

L'étude des transferts horizontaux est également une préoccupation sanitaire. En effet, la dissémination de gènes de virulence ou de résistance aux antibiotiques au sein d'espèces pathogènes est une préoccupation de santé publique. Une autre menace encore peu documentée implique les plantes et animaux transgéniques, dont l'usage est encore limité par peur de transfert de gènes vers des organismes devenant potentiellement résistants ou dangereux. Le bénéfice des HGT provient de leur potentiel à étendre la diversité fonctionnelle des communautés microbiennes et à augmenter leurs performances dans des environnements changeants ou extrêmes.

Un transfert horizontal est accompli suite à l'établissement des gènes à l'intérieur d'un génome receveur nécessitant cinq étapes.

Un segment particulier d'ADN est préparé au sein de la cellule donneuse par différents procédés : excision et circularisation de transposon conjugatif, initiation du transfert d'un plasmide conjugatif par synthèse du complexe d'accouplement par formation d'un pilus ou d'une fusion de membrane, ou par empaquetage d'acide nucléique dans un virion. Ce mécanisme peut être beaucoup plus général, l'ADN donneur pouvant être de l'ADN extrachromosomique libéré après

lyse cellulaire. Un segment est alors transféré par conjugaison, nécessitant un contact entre le donneur et l'accepteur, ou par par transformation ou transduction sans contact direct entre cellule faisant intervenir des pores membranaires spécialisés dans l'acquisition de matériel génétique.

Le matériel génétique entre dans la cellule réceptrice mais des **mécanismes d'exclusion** peuvent empêcher le transfert.

L'ADN entrant peut être intégré dans le génome par recombinaison légitime site-spécifique ou par circularisation. Cette étape peut échouer car il existe certaines barrières aux transferts de gènes, telles que les systèmes de restriction-modification, l'incompatibilité avec un plasmide résidant ou par le système 'immunitaire' de type CRISPR (page 59).

L'étape finale est ensuite la réplication de l'élément au sein du génome receveur.

Ce processus fait appel à des mécanismes complexes et séquentiels. Les biologistes moléculaires se concentrent sur les mécanismes de transfert alors que les écologistes microbiens regardent plus largement le pool de gènes mobiles, parfois appelé **mobilome**. Concevoir les transferts horizontaux remet en cause les bases de génétique du XX^{ème} siècle développées pour expliquer la **théorie de Darwin** en contexte **génétique Mendélien** selon une doctrine centrale d'héritage vertical des gènes. La génomique comparative a permis de révéler l'existence de transferts horizontaux de gènes à des fréquences beaucoup plus importantes que supposées. Cette prépondérance des transferts horizontaux chez les Procaryotes et leur importance dans la spéciation permet de concevoir l'évolution sous forme d'un réseau et non d'un arbre purement linéaire. Elle permet également de conforter et de rapprocher certaines théories de l'évolution qui semblaient en désaccord lorsqu'elles étaient uniquement considérées au niveau d'organismes pluricellulaires.

Elle conforte tout d'abord la **théorie neutraliste** de l'évolution de Kimura, les mutations ponctuelles ne pouvant expliquer à elles seules la spéciation. Ces transferts horizontaux confortent également la théorie du gène égoïste de Dawkins. Plus un gène est « égoïste », dans le sens ou son activité est tributaire de partenaires protéiques, plus la viabilité de son transfert pourra être pérennisée. A ce titre, le gène codant l'ARNr16S est un bon marqueur moléculaire phylogénétique car les nombreuses interactions nécessaires à l'assemblage d'un ribosome limitent les probabilités de transfert au cours de l'évolution.

En apportant cette vision dynamique des génomes, cette théorie peut être réconciliée avec celle des Equilibres ponctués de Gould. Il implique une série de brutaux changements génétiques aboutissant à la création d'une nouvelle espèce issue d'une petite population isolée. L'apport de matériel génétique à partir de l'environnement est en accord avec ces différentes théories de l'évolution et permet aux procaryotes de répondre à un changement environnemental. Les mécanismes permettant ces transferts horizontaux sont d'ailleurs les premiers gènes exprimés lorsqu'un stress est appliqué chez certains organismes, tels que les *Streptococcus* (Solow *et al.,* 2001) ou chez *Pyrococcus furiosus* (Diruggiero *et al.* 2000),

Ces **transferts de gènes** sont probablement le pendant prépondérant à l'innovation génétique qui est assurée par la **reproduction sexuée** chez les eucaryotes (Ochman *et al.,* 2000). Ces deux processus permettant l'innovation génétique, font intervenir un **mécanisme commun : la recombinaison**.

1. Ce que disent les génomes à propos des HGT et leur impact sur les communautés microbiennes

La comparaison des génomes microbiens pour la recherche de HGT est une approche rétrospective. La détection de ces évènements est basée sur les traces laissées dans les séquences d'ADN, leur distribution, délétion ou insertion. La présence de phylogénies conflictuelles de certains gènes vis-à-vis de la phylogénie consensuelle basée sur l'ADNr16S est le cas le plus commun. Des analyses statistiques supplémentaires permettent de suggérer un héritage par un ancien HGT. La présence de gènes de composition atypique, profil d'usage de codon ou contenu en G+C, peut indiquer un transfert si récent que la machinerie cellulaire n'a pas été affectée par ces biais. Néanmoins, ces biais permettent seulement l'identification de transferts récents. Au fur et à mesure des divisions, des processus d'amélioration liés à la machinerie réplicative optimise la composition de la séquence afin de l'adapter à celle du génome hôte pour une meilleure adéquation avec la machinerie cellulaire.

Alors que ces études montrent des fréquences élevées d'HGT, leur implication dans la spéciation microbienne est encore vivement contestée. Tous les gènes ne sont pas équitablement assujettis aux HGT. Les ORFs non annotés semblent avoir une relation entre les phénotypes prédits et le degré d'héritage de HGTs. Les gènes accessoires, souvent trouvés dans le **mobilome**, sont plus souvent transférés. Ces gènes sont sujets à une forte pression de sélection et ne produisent pas de phylogénie pouvant être corrélée à celle de leur génome hôte.

Les gènes les plus transférés sont des gènes « opérationnels », ils codent des fonctions métaboliques, la synthèse des acides aminés ou des acides gras. Ces gènes sont plus souvent impliqués dans des HGT que les gènes « informationnels » codant des protéines impliquées dans la réplication, la transcription ou la traduction. Cette distinction entre HGT de gènes opérationnels et gènes fonctionnels a été développée après analyse de 312 groupes de gènes orthologues dans 4 génomes de *Bacteria* et deux d'*Archaea* (Jain *et al.,* 2003). Les fonctions informationnelles sont médiées par des structures complexes, comme les ribosomes, requérant un grand nombre d'interactions parmi la multitude de partenaires protéiques, alors que les gènes opérationnels nécessitent peu d'interaction pour être actifs. L'impact et la fréquence des HGT sur un gène sont donc bien souvent corrélées à la complexité des interactions de son produit.

Les transferts de gènes sont d'autant plus probables qu'ils interviennent entre organismes de familles proches avec des similarités de structures de génomes et des mécanismes contrôlant la recombinaison, la réplication et l'expression de gènes. Cette préférence peut également s'expliquer par des probabilités de rencontre plus importantes quand des microorganismes partagent un biotope commun. Néanmoins, cette barrière peut être ténue dans certaines souches mutatrices possèdant un mécanisme de réparation de l'ADN défectif (Aa *et al.,* 2006). Lake et son groupe ont utilisé des arbres phylogénétiques construits suivant la méthode de parcimonie pour 20000 gènes issus de quatre bactéries et deux *Archaea.* Des organismes partageant une même niche écologique sont plus enclins aux transferts de gènes que ceux ne vivant pas ensemble. La découverte de gènes d'*Archaea* dans le génome de la bactérie thermophile *Thermotoga maritima* (Mongodin *et al.,* 2005), et inversement la présence de gènes bactériens dans le génome de l'*Archaea* méthanogène mésophile *Methanosarcina mazei* confirme la barrière de niche écologique pour les HGT (Rest *et al.,* 2003).

Un autre élément confirmant l'importance des HGT dans le modelage des communautés microbiennes a été apporté par les projets de séquençage de métagénomes environnementaux. Rohwer a notamment trouvé de nombreux gènes d'éléments génétique d'*Archaea* et de bactéries dans les génomes viraux environnementaux (Casas *et al.,* 2007). La contribution des virus aux HGT et des recombinaisons est également suggérée par la présence de signatures spécifiques et des fréquences d'observation de sites de recombinaison dans le métagénome de communauté acidophiles (Rest *et al.* 2003) ou dans l'expérience pionnière de métagénomique de la mer des Sargasses.

2. Comment les gènes sont transférés au sein des communautés ?

La principale approche visant à mesurer les transferts de gènes au sein de communautés est rétrospective et ne prend pas en compte la dynamique des différents mécanismes de HGT. Une seconde approche détecte les HGT en cours, dans des communautés microbiennes intactes ou manipulées, comme le transfert de plasmide conjugatif entre bactéries donatrices et réceptrices au sein d'un microcosme. *Acitenobacter spp.* sert de modèle d'organisme récepteur pour les transferts de gène intra et interspécifique par transformation dans des biofilms ou des reconstitutions de communautés microbiennes (Ray *et al.,* 2005). D'autres études suggèrent que la compétence, phase durant laquelle la cellule peut acquérir de l'ADN extrachromosomique, est plus importante que précédemment admis. Ce mécanisme est d'autant plus important qu'il permet d'acquérir de plus grandes quantités d'ADN et d'origines beaucoup plus éloignées. La transduction est un autre procédé conduisant à l'innovation génétique dans les environnements aquatiques et terrestres surtout lorsqu'on considère que la biomasse virale est estimée dix fois supérieure à celle des procaryotes !

D'autres efforts ont été concentrés sur l'inventaire des mobilomes environnementaux. Par exemple, des plasmides conjugatifs ou mobilisables ont été capturés au sein de communautés stables en plaçant des souches « hameçon » réceptrices marquées (Haines *et al.,* 2006). Ce type d'approche est courant pour capter de nouveaux plasmides environnementaux. Des outils similaires ont été conçus pour capturer des transposons et des intégrons. Les premiers résultats montrent une immense diversité, probablement spécifique de l'écosystème. De cette façon, de nouvelles origines de réplication issues de plasmides de bactéries marines ont été détectées (Sobecky 2002). D'autres études, sur les intégrons issus d'ADN extrait du sol (Holmes *et al.,* 2003), suggèrent que ces éléments sont la clé centrale de l'évolution des communautés microbiennes.

La modulation de l'évolution microbienne par ces éléments est illustrée par l'élément Tn4371 de *Ralstonia* sp. A5 qui a récemment été reclassé comme un îlot génomique (Merlin *et al.,* 1999). Cet élément de 45kb porte les fonctions de transfert conjugatif, tels que celui de *mating pair formation* et de réplication de plasmides. Les motifs flanquants Tn4371 sont retrouvés dans les chromosomes de nombreuses protéobactéries et suggère une large distribution de ce type d'éléments. Les îlots génomiques peuvent comporter plusieurs degrés de mobilité horizontale. Leur composition mosaïste suggère donc un procédé d'assemblage gouverné par les HGT.

3. Approches et niches écologiques pour étudier les HGT en direct

Les transférants, organismes dont le génome est altéré par des transferts de gènes, sont détectés par sélection de souches donatrices et réceptrices. La communauté microbienne est conçue comme receveuse dans son ensemble. De nouvelles approches se fient à l'expression conditionnelle de phénotype. Cette méthode robuste est basée sur l'expression de protéines telles que les GFP (protéines fluorescentes vertes) (Heydorn *et al.,* 2000). Par exemple lorsqu'une une GFP est sous-exprimée par suite de la fixation d'un répresseur fort présent sur le chromosome de la souche donatrice, mais potentiellement absent dans la réceptrice. La lecture est fiable et rapide grâce à la présence de fluorescence.

Le couplage de la cytométrie en flux à l'approche GFP permet de recherche de transconjugants rares, le taux de conjugaison ainsi mesuré est 20 à 100 fois supérieur à celui observé avec le plasmide IncP1 par étalage sélectif sur boite (Sorensen *et al.,* 2003). De plus, cette expérience a démontré que des transconjugants du plasmide IncP1 sont également trouvés dans les bactéries à Gram positif alors que le transfert semblait limité aux bactéries à Gram négatif.

L'examen par microcopie confocale des bactéries de communautés naturelles sujettes aux HGT montre l'architecture d'un biofilm et les activités cellulaires influencant les populations de transconjugants. Malgré la fréquence des évènements de transfert, ils ne sont détectables qu'à partir du moment où les transférants deviennent majoritaires dans la population.

VII. L'immunité chez les Procaryotes : Le système CRISPR-CAS

Les phages sont la forme de vie la plus abondante sur Terre (Breitbart *et al.,* 2005). Le milieu océanique est celui pour lequel l'abondance virale a été la plus étudiée. Elle est estimée entre 5 et 10 particules virales par cellule bactérienne (Wommack *et al.,* 2000). Malgré cette submersion par les phages, les procaryotes parviennent à proliférer et évitent l'extinction en utilisant un large éventail de mécanismes de résistance innée, tels que l'interférence avec l'adsorption, la perturbation de l'injection d'ADN, la restriction/modification et l'infection abortive. (Sturino *et al.,* 2006).

Ce chapitre est consacré au premier mécanisme de résistance acquise chez les *Bacteria* et les *Archaea*. Les CRISPR, *Clustered Regularly Interspaced Palindromic Repeat,* ont récemment été découverts. Leur principale caractéristique est l'arrangement de courtes séquences répétées séparées par de courtes séquences non conservées, nommées *spacers*.

1. Historique de la découverte des CRISPRs

La première description des champs de CRISPRs a été effectuée en 1987 (Ishino *et al.,* 1987), grâce la découverte de 14 répétitions de 29 pb séparées par des séquences non répétées de 32-33pb à proximité d'un gène étudié par Ishino. Des structures similaires ont été observées les années suivantes chez *Mycobacterium tuberculosis* (Nakata *et al.,* 1989), *Haloferax volcanii* (Mojica *et al.,* 1995), *Methanocaldococcus jannaschii* (Bult *et al.,* 1996), *Thermotoga maritima* (Nelson *et al.,* 1999) et autres bactéries et Archaea. L'augmentation du nombre de génomes de procaryotes a permis la recherche à grande échelle des CRISPRs (Mojica *et al.,* 2000). Les analyses les plus récentes révèlent la présence de CRISPR dans approximativement 40% des génomes bactériens et 95% des génomes archéens (Grissa *et al.,* 2007; Kunin *et al.,* 2007).

En parallèle, quatre gènes systématiquement associés aux champs de répétitions ont été identifiés (Jansen *et al.,* 2002). Le nombre de ces gènes *cas*, *CRISPR Associated*, a ensuite été étendu entre 25 et 45 familles (Haft *et al.,* 2005). Les génomes ne possédant pas de CRISPRs ne possèdent pas de gènes *cas*.

Plusieurs hypothèses ont été émises sur la fonction des CRISPRs. En 1995, Mojica suggérait que ces répétitions étaient impliquées dans la partition (Mojica *et al.* 1995) car l'augmentation du nombre de répétitions chez *H. volcanii* altérait la ségrégation. Ce phénomène n'a pas été

reproduit lors d'expériences similaires menées chez *M. tuberculosis* (Jansen *et al.* 2002). La présence de plusieurs champs de CRISPRs dans certains génomes a suggéré qu'ils étaient des éléments génétiques mobiles (Jansen *et al.* 2002), alors que la présence de gènes *cas* impliqués dans la maintenance de l'ADN suggérait leur implication dans la réparation de l'ADN (Makarova *et al.,* 2002). En 2005, trois laboratoires rapportent que les séquences des spacers correspondent à des dérivés d'ADN phagiques ou plasmidiques et proposent un mécanisme immunitaire vis-à-vis les éléments extrachromosomiques (Bolotin *et al.,* 2005; Mojica *et al.,* 2005; Pourcel *et al.,* 2005). Une corrélation négative a également été rapportée entre la sensibilité d'une bactérie à l'infection phagique en fonction du nombre de spacers (Bolotin *et al.,* 2005). Cette hypothèse immunitaire a été confirmée expérimentalement en montrant que suite à une infection par un phage de nouveaux spacers peuvent être ajoutés dans le champ de CRISPR et conférent ainsi une résistance vis-à-vis de ce phage (Barrangou *et al.,* 2007).

2. Caractéristiques structurales des CRISPRs

Le champ de répétitions et les gènes *cas* composent le système CRISPR (Figure 12). Les séquences répétées (Godde *et al.,* 2006) et les gènes *cas* varient entre espèces (Makarova *et al.,* 2006). La taille des répétitions varie entre 24 et 47pb, celle des spacers entre 26 et 72pb. Le nombre de répétitions au sein d'un champ de CRISPR varie entre 2 et 249 pour *Verminephrobacter eiseniae* (Grissa *et al.,* 2007). Bien que la plupart des génomes contiennent un unique système CRISPR, certains en possèdent plusieurs, tel *M. jannaschii* qui en contient 18 (Bult *et al.* 1996). Finalement, le nombre de gènes *cas* est extrêmement variable, il est compris entre 4 et plus de 20. Malgré cette définition assez floue, la plupart des CRISPRs ont des caractéristiques communes.

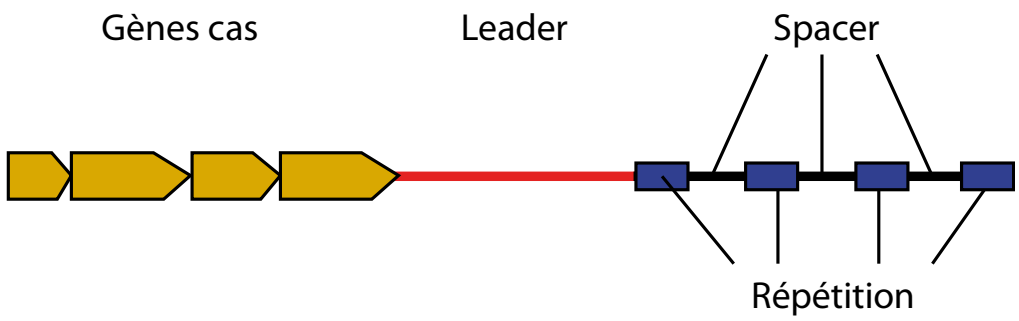


Figure 12 Représentation schématique d'un champs d'un système CRISPR

2.1 *Les répétitions*

Au sein d'un champ CRISPR, les répétitions sont pratiquement identiques. Malgré la divergence des séquences entre espèces, ces répétitions peuvent être groupées au sein de 12 classes en fonction de leurs compositions et des structures secondaires qu'elles génèrent (Jansen *et al.* 2002). Les classes majoritaires contiennent de courtes séquences palindromiques (5-7pb), caractère retranscrit dans l'acronyme CRISPR. Ces palindromes induisent une structure secondaire de tige-boucle, au niveau de l'ARN, supposée être impliquée dans la fonction immunitaire. En effet, des mutations compensatoires au sein des répétitions ont été observées, elles permettent le maintien de ce type de structure. La présence d'ARN issus de la transcription d'un champ de CRISPR a également été démontrée (Tang *et al.,* 2002; Tang *et al.,* 2005; Lillestol *et al.,* 2006). Chez *P. furiosus,* il semblerait que les spacers acquis le plus récemment sont également ceux qui sont le plus transcrits (Hale *et al.,* 2008). En dehors de la structure, de nombreuses répétitions possèdent une extrémité 3' conservée (GAAAS). La structure et la séquence conservée sont supposées servir de sites de fixation aux protéines Cas (Kunin *et al.,* 2007).

2.2 *Spacers*

Au sein d'un système CRISPR, les spacers sont généralement des séquences uniques. Cependant, quelques exceptions ont été observées et sont supposées être générées par duplication segmentale (Grissa *et al.,* 2007). Les recherches de similarités de séquences montrent que les spacers correspondent à des portions de génomes d'éléments génétiques mobiles. Mojica a étudié 4500 spacers provenant de 68 génomes, 2% sont strictement identiques à des séquences d'éléments génétiques. Ce faible pourcentage peut être expliqué par le manque de séquences de phages dans les bases de données, en accord avec les récentes estimations de l'importante diversité de ces éléments génétiques dans l'environnement (Edwards *et al.,* 2005; Casas *et al.* 2007; Delwart 2007). Néanmoins, chez *Streptococcus thermophilus*, pour lequel une douzaine de phages ont été séquencés, approximativement 40% des spacers ont des homologues dans les génomes phagiques (75%) ou plasmidiques (20%) (Bolotin *et al.,* 2005).

Les spacers semblent acquis de manière aléatoire à partir du génome phagique, aussi bien à partir du brin sens (codant) que du brin antisens (non codant). Deux études ont également rapporté l'existence d'un motif nucléotidique situé à 1 ou 2 pb en amont de la séquence capturée par le système CRISPR (Deveau *et al.,* 2007; Horvath *et al.,* 2008). Ce motif semble être important dans la reconnaissance et le clivage de l'ADN phagique, par le système CRISPR.

2.3 *La séquence leader*

Le leader est une séquence AT-riche de 550pb située à l'extrémité de 5' de la plupart des systèmes CRISPRs, immédiatement à proximité de la première répétition (Lillestol *et al.* 2006). De façon similaire aux répétions, elle ne contient pas de cadre ouvert de lecture et n'est généralement pas conservée entre espèces. Néanmoins, lorsque plusieurs loci CRISPR existent dans une souche, cette séquence est conservée. L'ajout d'une nouvelle unité « répétition-spacer » se fait toujours à proximité de la séquence leader, suggérant qu'elle agisse comme un site de reconnaissance pour orienter l'ajout de nouveaux spacers. Il a également été suggéré que le leader agisse comme un promoteur afin d'initier la transcription du champ de CRISPR (Tang *et al.* 2002; Tang *et al.* 2005).

2.4 *Les gènes cas*

Une partie des gènes *cas* a été détecté grâce à la présence d'homologues rencontrés spécifiquement à proximité des champs de CRISPRs (Haft *et al.* 2005). Les systèmes CRISPRs peuvent-être classés en 7 ou 8 sous-types en fonction de la présence des différents gènes *cas*. Chaque sous-type contient entre 2 et 6 gènes *cas* spécifiques. De plus, six gènes *cas* peuvent-être rencontrés dans différents sous-types, ils sont nommés *core cas gene* (*cas1-6*).

Le gène *cas1* sert de marqueur universel des systèmes CRISPR, à l'exception de *Pyrococcus abyssi* qui en est dépourvu. Seul le gène *cas2* possède une activité caractérisée. Il code une endoribunucléase intervenant dans la dégradation spécifique des ARNm phagiques (Beloglazova *et al.,* 2008). Il confère la protection vis-à-vis du phage selon un mécanisme analogue à l'enzyme Slicer de l'ARNi eucaryote.

D'autres gènes sont moins fortement associés aux CRISPRs, telles les protéines RAMP Repeat Associated Mysterious Protein, uniquement rencontrées dans les génomes contenant des CRISPRs mais pas nécessairement à proximité des CRISPRs. Certains domaines fonctionnels très généraux ont été identifiés sur les protéines Cas, endonucléase, exonucléase, hélicase, fixation aux acides nucléiques (Ebihara *et al.,* 2006), mais la fonction précise de ces protéines reste inconnue.

3. Les CRISPRs, un système de défense antiphage

En 2007, Barrangou et ses collaborateurs ont expérimentalement démontré que l'ajout d'un spacer d'origine phagique confère l'immunité vis-à-vis de celui-ci. Les auteurs ont ensuite infecté

S. thermophilus avec deux phages différents afin de récupérer des mutants résistants. Le séquençage de leurs loci CRISPR montre que tous les mutants ont indépendamment acquis entre un et quatre spacers à partir de l'ADN du phage et à proximité de la séquence leader. Dans tous les cas, ces spacers étaient dérivés de séquences phagiques. Si le spacer possède la séquence exacte du phage (100% d'identité), le mutant est résistant ; si un ou plusieurs changements de nucléotides interviennent, la bactérie est alors sensible au phage. Barrangou a ensuite ajouté à une souche sensible à un phage donné, un spacer afin de confirmer qu'il conférait la résistance. La délétion ultérieure de ce spacer permet de revenir au phénotype sensible. De manière intéressante, une faible proportion des phages maintient sa capacité à infecter les bactéries résistantes. Le séquençage du génome de ces phages montre qu'ils ont subi certaines mutations au niveau de séquence correspondant aux spacers de la bactérie. Certains phages ont des séquences identiques aux spacers conférant normalement l'immunité, mais ils présentent des mutations au niveau d'un motif AGAA situé en amont. Cette observation confirme l'importance de ce motif et son rôle putatif dans la reconnaissance par les protéines CAS.

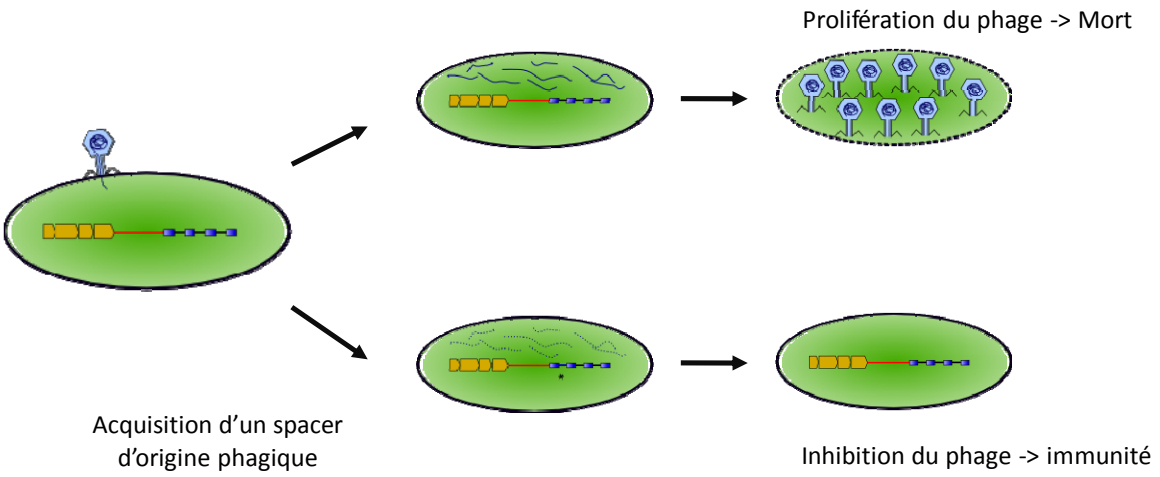


Figure 13 Effet de l'acquisition d'un spacer d'origine phagique dans un système CRISPR

Un mécanisme alternatif d'échappement vis-à-vis du système CRISPR a été proposé chez le phage SVR, infectant une souche de *Stygiolobus*, une archaea thermoacidophile appartenant à l'ordre des Sulfolobales (Vestergaard *et al.,* 2008). La redondance de séquence nécessaire à l'assemblage révèle que certains virus possèdent des polymorphismes ponctuels provoquant des mutations

silencieuses au niveau de la protéine codée, mais également des délétions de 12pb pouvant être générées en réponse au système de défense CRISPR de l'hôte.

4. Un modèle de l'activité des CRISPRs

Le mécanisme d'acquisition d'un spacer d'origine phagique et l'inhibition du phage est à l'heure actuelle inconnue. Néanmoins, un modèle approximatif peut-être énoncé. Le champ de CRISPR est transcrit en un unique ARN qui est ensuite clivé en plus petites unités (sRNA *small RNA*) de taille égale à celle d'une répétition et d'un spacer. Le clivage est préférentiellement réalisé au milieu du spacer et suppose que le sRNA est un spacer entouré de deux demi-répétitions. Ces répétitions étant palindromiques elles s'attacheraient afin de former une boucle.

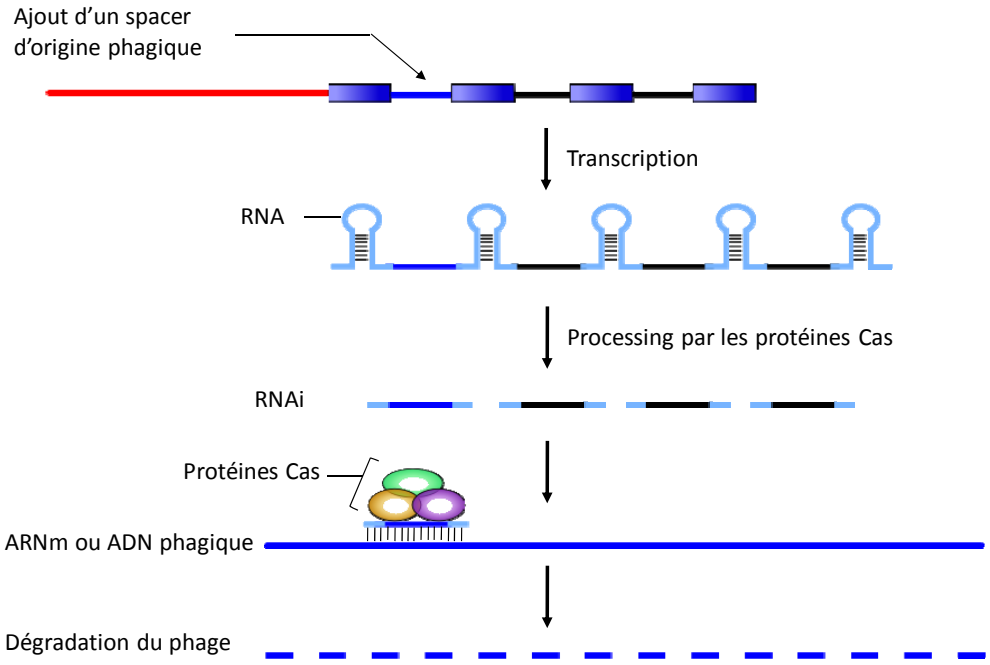


Figure 14 Mécanisme hypothétique de fonctionnement du système CRISPR

La maturation d'un long transcrit et la présence de protéines CAS impliquées dans la manipulation des ADN et des ARN font penser au système d'interférence à ARN des eucaryotes. Pour l'instant, seule l'enzyme intervenant dans la dégradation des ARNm phagiques à été caractérisée, il s'agit de la protéine Cas2 (Beloglazova *et al.* 2008).

5. Evolution des systèmes CRISPRs

Les CRISPRs évoluent très rapidement. Une étude métagénomique réalisée sur des sous-populations clonales de *Leptospirillum*, isolées d'un biofilm microbien acidophile, a montré que la collection de spacers était suffisante à conférer une individualité cellulaire (Tyson *et al.,* 2008). Les spacers les plus récents sont ajoutés en 5', à proximité de la séquence leader, les spacers les plus « âgés » peuvent-être conservés en 3', alors que les « nouveaux » spacers sont généralement uniques. L'inflation du nombre de répétitions est limitée par de fréquentes délétions de spacers. Ce mécanisme pourrait être passif, sous forme de recombinaison homologue entre séquences répétées ou bien actif en faisant intervenir certaines protéines Cas afin de contrôler la suppression de spacers.

Bien avant la découverte de la fonctionnalité des CRISPR, l'évolution rapide de ces séquences avait été utilisée en épidémiologie pour typer des souches proches : il s'agit du splogotypage (Kamerbeek *et al.,* 1997).

A un niveau évolutif plus élevé, les systèmes CRISPRs sont très diversifiés. Bien que les répétitions ne soient pas conservées entre organismes phylogénétiquement éloignés, quelques exceptions ont été observées. Par exemple, *E. coli* et *Mycobacterium tubercolusis* contiennent les mêmes répétitions malgré l'appartenance de ces organismes à deux phyla distincts. Cette observation peut être expliquée par un transfert horizontal d'un système CRISPR entre ces organismes, hypothèse confirmée lors de l'analyse phylogénétique des gènes *cas*. Le transfert horizontal pourrait être réalisé par des plasmides, en effet, un CRISPR est présent sur le plasmide pNOB8 de *Sulfolobus* (Greve *et al.* 2004).

La présence atypique d'un CRISPR dans un prophage de *Clostridium difficile* pourrait également lui être utile afin d'inhiber certains gènes de défense de la cellule et ainsi l'aider à détourner l'expression de l'hôte à son compte ou encore limiter l'invasion par des phages compétitifs (Sebaihia *et al.,* 2006).

MATERIELS ET METHODES

I. Souches et cultures

Les souches utilisées sont issues de sites hydrothermaux océaniques profonds couvrant l'ensemble des dorsales océaniques. Les échantillons ont été prélevés lors de six campagnes océanographiques effectuées entre 1989 et 2004 (Figure 15).

GE (1989) : Dorsale du bassin Nord Fidjien (15°N). Les échantillons ont été prélevés à proximité de cheminées hydrothermales par 1900m de profondeur.

AMISTAD (1999) : Dorsale du Pacifique Est (EPR, 13°N). Les échantillons ont été prélevés par 1600m de profondeur, au niveau de trois zones principales s'étendant sur 2km. Du nord au sud, ces sites sont Pulsar, La chainette, puis Genesis, Grandbonum, et plus au sud, le site Elsa.

IRIS (2001) : Dorsale médio-atlantique (MAR, 36°N). Les échantillons proviennent du site Rainbow (2300m) couvrant une surface de 400x200m à forte concentration de cheminées.

EXTREME (2001) : Dorsale du Pacifique Est (EPR 9°N), cheminées par 2500m de fond.

CIR (2001) : Triple jonction centrale de l'océan Indien (25°N/70°E), 2420m de profondeur.

SWEEP VENT (2004) Dorsale pacifique sud Valu Fa, bassin de Lau, au niveau des sites Mariner, Vai Lili et Hine Hina, aux profondeurs comprises entre 1700 et 1980m.

Les échantillons sont conditionnés et enrichis en microorganismes avec différents milieux de culture et différentes conditions physicochimiques, orientant les métabolismes des souches que l'on souhaite sélectionner. Ces « bouillons de culture » servent à isoler les souches, qui seront purifiées et éventuellement caractérisées. 295 souches issues de ces campagnes ont été purifiées, peu d'entre elles sont caractérisées (

Tableau 9). Ces souches purifiées, non caractérisées, supposent qu'il existe un certain degré de redondance dans la collection de travail.

Les cultures sont réalisées en fiole pénicilline sous atmosphère anoxique (N_2). Le milieu TRM, pour Thermococcales Rich Medium, est utilisé pour la croissance bactérienne. Ce milieu contient pour 1 litre d'eau : 23g de NaCl ; 5g de $MgCl_2.2H_2O$; 0,7g de KCl ; 0,5g de $(NH_4)_2SO_4$; 1ml d'une solution stérile de K_2HPO_4 à 5% (p/v) ; 1ml d'une solution stérile de $CaCl_2.2H_20$ à 2% (p/v) ; 1ml d'une solution stérile de $FeCl_3$ à 25mM ; 3,3g de tampon PIPES ; 1g d'extraits de levure ; 1g de tryptone

Tableau 9 Origines géographiques des isolats étudiés

Campagne	Abréviation	Année	Souches isolées
AMISTAD	AMT	1999	102
IRIS	IRI	2001	59
EXTREME	EXT	2001	16
CIR	CIR	2001	16
SWEEP VENT	SV	2004	33
STARMER	GE	1989	31

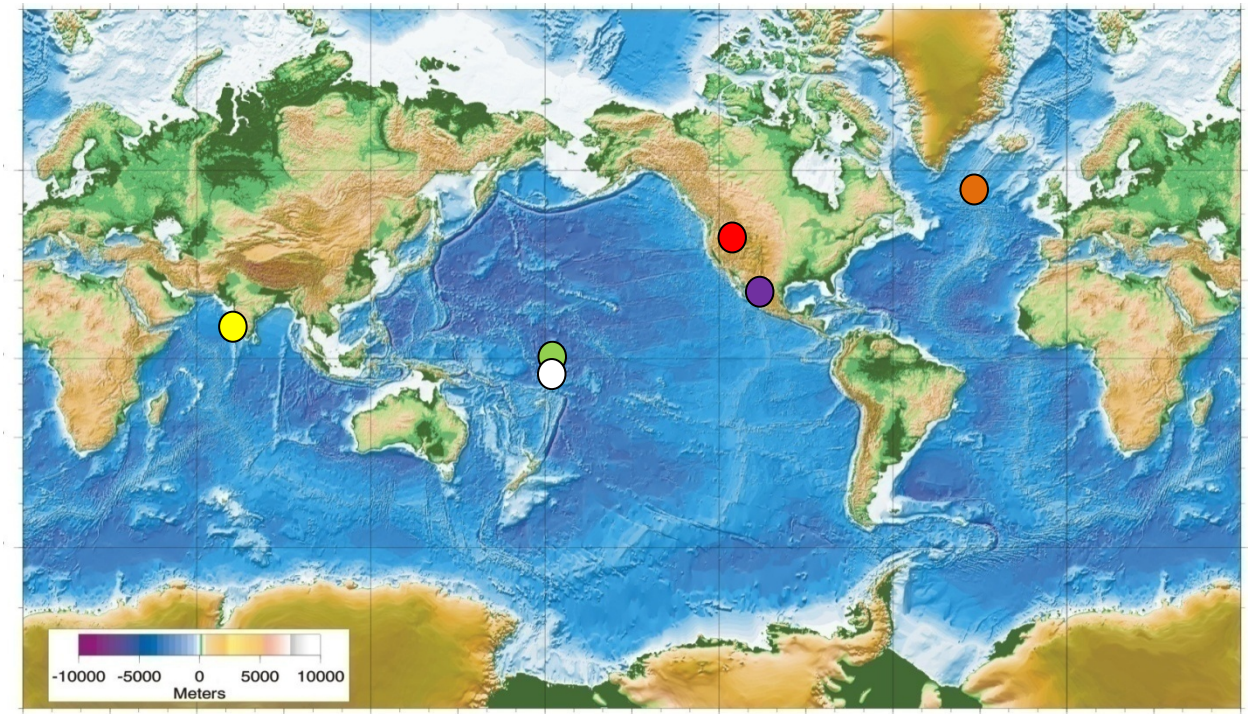


Figure 15 Origine géographique des isolats étudiés

et 1mg de résazurine. Le pH du milieu est ajusté à 6,8 avant d'être autoclavé 20 minutes à 120°C. 45ml de TRM sont répartis dans une fiole pénicilline additionnés d'un gramme de soufre élémentaire stérilisé par tyndallisation. Les fioles sont bouchées hermétiquement et le milieu est soumis à plusieurs cycles vide/gaz (N_2) pour éliminer l'oxygène. L'anaérobiose totale est réalisée par réduction chimique du milieu en ajoutant au 100ème une solution de Na_2S stérile à 3%. La résazurine, indicateur coloré du potentiel redox, permet de vérifier la réduction du milieu en devenant incolore en l'absence totale d'oxygène. Les fioles sont inoculées au 1/100ème avec une culture fraîche et incubées pendant 12h à 85°C, 90°C ou 95°C sous agitation modérée (150 rpm).

II. Extraction d'ADN de Thermococcales

1. Extraction d'ADN plasmidique

Le protocole est basé sur la technique de lyse alcaline (Birnboim *et al.,* 1979) modifiée pour la lyse des membranes d'*Archaea* et l'extraction de plasmides de Thermococcales. 45 ml de culture sont transférés dans des tubes Falcon de 50 ml placés dans la glace.

Ces tubes sont centrifugés à 7500rpm pendant 15 min à 4°C. Le surnageant est éliminé. Le culot cellulaire est resuspendu dans 600µl de tampon TNE (Tris HCl 0.1M, NaCl 0.1M, Na_2EDTA 50mM, PH 8). Les membranes sont lysées par 600µL de solution de lyse préparée le jour même (8mL d'eau milliQ autoclavée, 1mL de SDS 10%, 0.2mL de NaOH 10N). La lyse est stoppée par ajout de 700µL d'une solution de neutralisation. (60mL d'acétate de potassium 5M, 11.5mL d'acide acétique glacial, 28.5mL de NaOH 10N).

Cette solution est à nouveau centrifugée à 13000rpm pendant 15 min à 4°C faisant tomber les débris membranaires au fond du tube. Le surnageant contenant l'ADN est récupéré et additionné d'un volume de Phénol Chloroforme Alcool Isoamylique (25:25:1). Après une forte agitation visant à mélanger les phases, on centrifuge à 13000rpm pendant 15 min à 4°C. La phase phénol contenant les protéines complexées à l'ADN chromosomique est éliminée, la phase aqueuse supérieure contenant l'ADN plasmidique est transférée dans un nouveau tube. 0,7 volume d'isopropanol est ajouté pour précipiter l'ADN.

Une autre centrifugation à 13000rpm pendant 15 min à 4°C culotte l'ADN. L'isopropanol est aspiré et le culot d'ADN est lavé par ajout de 500µl d'éthanol 75%. Une dernière centrifugation à 13000rpm pendant 5 min à 4°C permet de culotter l'ADN. L'éthanol est aspiré et le culot est séché

au Speedvac. L'ADN plasmidique est finalement solubilisé dans 50µL de tampon TE (Tris-HCl 10mM, EDTA 1mM, PH8).

5µL d'ADN plasmidique, obtenus dans l'étape précédente, sont digérés par l'enzyme de restriction *HindIII*. Les produits de digestion sont ensuite séparés par électrophorèse sur gel d'agarose 0,8% TAE 1X coloré au BET sous courant de 80V. La qualité de l'extraction d'ADN plasmidique est révélée par transillumination sous UV. La taille des fragments est estimée grâce à la migration en parallèle du marqueur de taille Promega 1kb (12 bandes 10kb, 8kb, 6kb, 5kb, 4kb, 3kb, 2,5kb, 2kb, 1,5kb, 1kb, 500pb et 250pb. L'ADN est finalement quantifié au spectrophotomètre (Nanodrop ND-1000 Technologies™)

2. Extraction d'ADN total

2mL d'une culture en fin de phase exponentielle sont transférés dans un tube Eppendorf de 2mL. Les cellules sont culottées par 10min de centrifugation à 7000rpm à 4°C, et le surnageant est éliminé. Le culot est repris dans 800µL de TNE en vortexant à vitesse modérée pendant 30s, on ajoute 20µL de RNaseA (50µg/mL) puis 100µl de détergent SDS10% et 100µL de sarkosyl 10%. La lyse est immédiate, se traduisant par l'apparition de « neige » dans le tube. Les protéines sont dénaturées par l'action de 50µL de protéinase K (20mg/mL) incubé 1 h à 55°C.

1mL de PCI (25 :24 :1) est ajouté et les tubes sont agités. 15 min de centrifugation à 14000 rpm permettent de récupérer la phase aqueuse en prenant soin de ne pas aspirer la crêpe protéique blanche située à l'interphase. On ajoute 1mL de chloroforme et on centrifuge à 14000rpm, pendant 5min, à 4°C. La phase aqueuse supérieure est additionnée de 700µl d'isopropanol pour précipiter l'ADN. 5 min de centrifugation à 14000rpm permettent de culotter l'ADN et d'éliminer le solvant par aspiration. Le culot d'ADN est nettoyé par 500µL d'éthanol 75%. Le séchage de l'ADN sous vide (Speedvac™) permet d'éliminer l'éthanol résiduel. L'ADN est solubilisé dans 100 µL de TE durant une nuit à 4°C.

III. Criblage plasmidique de la collection de souches

La collection d'isolats de Thermococcales a été criblée à la recherche d'ADN plasmidique. Toutes les souches ont subi au minimum deux extractions d'ADN plasmidique. Les ADNp extraits ont été digérés par 2 unités des enzymes de restriction *HindIII* et *BamHI*. Cette restriction permet de savoir si la souche possède un (des) plasmide(s).

IV. Classification des plasmides

Une première estimation de la diversité et des relations entre plasmides extraits a été entreprise par hybridations croisées ADN-ADN en Southern blot.

1. Transfert d'ADN plasmidique sur membrane HybondN+

L'ADN des plasmides est digéré par 2u de l'enzyme de restriction HindIII et séparé sur gel d'agarose 0,8% TAE1X.

La première étape consiste à transférer l'ADN du gel d'agarose vers une membrane en nylon.

Dans une cuve contenant du tampon SSC10X, un support plastique est déposé. Il est recouvert d'un papier Whatman, servant de mèche, trempant dans le tampon SSC. Le gel d'électrophorèse est déposé sur ce support et recouvert d'une membrane HybondN+ prédécoupée à la taille du gel. Cet assemblage est recouvert de papier absorbant d'une épaisseur de 10cm empilé uniformément.

Cette couche de papier va absorber durant 12h le tampon contenu au fond de la cuve par capillarité. Le flux de liquide passant à travers le gel entraîne l'ADN vers la membrane sur lequel il sera fixé. Après transfert, la membrane est rincée avec du SSC5X pour éliminer les traces d'agarose. L'ADN est ensuite fixé covalemment à la membrane par une exposition de 2'30'' aux UV.

2. Marquage de la sonde

La sonde est de l'ADN plasmidique marqué par le kit à grande sensibilité ECL d'Amersham (Durrant *et al.,* 1990).

20µL d'ADN (500ng) sont chauffés durant 5 minutes à 95°C et dénaturés par plongeon immédiat dans la glace. Cet ADN dénaturé est marqué par ajout de 10µL de Nucléotide Mix 5X, 5µL d'un mélange d'héxamères aléatoires, 5 unités de Klenow et 14µL d'H_20 ultrapure. La réaction est incubée 1 heure à 37°C à l'abri de la lumière puis stoppée par ajout de 2%L d'une solution d'EDTA 0,5M.

3. Vérification du marquage

Le marquage de la sonde est vérifié par sept dépôts de 5µL de Nucléotide mix dilués respectivement au 1/5, 1/25, 1/50, 1/100, 1/250, 1/500, 1/1000 dans du TE sur une bandelette de membrane HybondN^{+}. Sur une seconde bandelette, 5µL de la sonde et 5µL de la sonde diluée au 1/5 2'30''. Ces ADN sont fixés à la membrane par exposition aux UV et lavées avec du SSC2X préchauffé à 60°C.

Les bandelettes sont ensuite transilluminées sous UV pour observer la fluorescence. La dilution au 1/5 sert de témoin négatif et doit émettre une faible fluorescence. Si la fluorescence est supérieure à la fluorescence émise par la dilution au 1/250 alors 30µL de sonde seront utilisés pour l'hybridation.

4. Hybridation d'ADN marqué ECL sur membrane ECL

La membrane est hydratée avec du SSC5X, déposée sur une toile de nylon et placé dans un tube à hybridation. 15mL de tampon d'hybridation préchauffé à 60°C sont ajoutés dans le tube à hybridation. La préhybridation dure 1h à 60°C. Pendant ce temps, 30µL de sonde sont chauffés 10 minutes à 100°C puis dénaturés par plongeon dans de la glace. La sonde dénaturée est ajoutée dans le tube à hybridation et incubée durant 14h à 60°C sous agitation rotative.

La membrane est rincée 15 minutes dans 200mL de SSC1X préchauffé à 60°C puis rincée selon les mêmes conditions dans du SCC 0,5X.

5. Révélation de la membrane

La lecture des signaux d'hybridation a été réalisée sur la plateforme OuestGénopole de Roscoff au moyen d'un lecteur Genex.

V. Obtention de la séquence d'un génome de plasmide

La longueur des runs de séquençage est limitée à environ 1kb. L'obtention d'un génome nécessite la fragmentation de son ADN en morceaux de plus petite taille. Chaque fragment est inséré dans un vecteur de clonage. L'utilisation d'amorces présentes sur le vecteur permet le séquençage de l'insert. Par chevauchement de séquence, l'intégralité du génome est reconstituée. Pour avoir un génome de qualité correcte pour l'annotation, une couverture globale de 10X est requise.

1. Nébulisation Précipitation

L'ADN est fragmenté de manière uniforme et aléatoire par des cassures mécaniques selon un processus nommé nébulisation (Hengen 1997) faisant appel à un appareil nommé nébuliseur. La solution d'ADN passe à travers un orifice très étroit sous la pression de gaz. La taille moyenne des fragments obtenus est fonction de la pression et de son temps d'application.

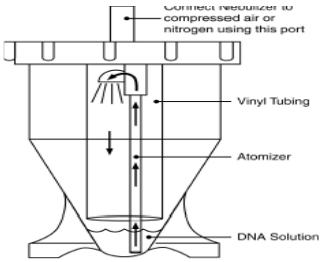


Figure 16 Schéma d'un nébuliseur

8µg d'ADNp, additionnés de 750µL de TE 20% glycérol sont introduits dans un nébuliseur. Le nébuliseur est placé dans la glace et connecté à une arrivée de gaz N_2 délivrant une pression de 0,8 bar. La pression de gaz fait passer l'ADN à travers une buse micrométrique brumisant le liquide. Dans ces conditions, 3 à 4 min sont nécessaires pour obtenir des fragments d'ADN de taille moyenne comprise entre 1 et 1.5 kb.

L'ADN nébulisé est transféré dans un microtube de 2mL et précipité par ajout de 75µL d'acétate de sodium 3M et 600µL d'isopropanol. La précipitation se déroule pendant 1h à 4°C. L'ADN est culotté par 1h de centrifugation à 14000rpm, à 4°C. L'isopropanol est aspiré avec une trompe à vide. Le culot d'ADN est nettoyé avec 500µL d'éthanol 75% et centrifugé 10 min à 14000rpm, à 4°C. Après évaporation de l'éthanol résiduel, l'ADN est resuspendu dans 32µL d'H_20 MQ.

2. Création de la banque d'ADNp nébulisé

L'ADN nébulisé est déposé sur gel d'agarose 0,8% TAE 1X BET, et confronté à un marqueur de poids moléculaire. L'application d'un champ électrique de 80V permet de séparer les fragments d'ADN en fonction de leur taille et de les visualiser par transillumination sous UV. L'agarose contenant les fragments compris entre 500 et 1000pb est découpé. Une seconde plage de fragments compris entre 1,5 et 4kb est également découpée. L'ADN est extrait de l'agarose par le kit QiaEXII de Qiagen selon le protocole détaillé par le fabricant.

Cette extraction est basée sur la solubilisation de l'agarose et l'absorption sélective des acides nucléiques par des billes de silice en présence d'une forte salinité. L'élution dépend du pH et de la concentration en sel. Un tampon, contenant un agent dénaturant rompt les liaisons hydrogènes des différents sucres de l'agarose polymérisé et solubilise le gel. L'ADN est élué par 22 µL d'eau milliQ autoclavée pH8.

3. Réparation

Le clonage d'ADN nécessitant des fragments double brin. La nébulisation de l'ADN ayant provoqué des cassures laissant des extrémités débordantes, il faut réparer l'ADN. Cette réparation consiste à synthétiser les extrémités des fragments d'ADN nébulisés pour les rendre franches, et à phosphoryler les extrémités 5' pour faciliter l'action ultérieure de la ligase. Le kit End-It^{TM} DNA End-Repair (Epicentre) permet de réaliser ces 2 réactions en une seule étape. Il contient un mélange de T4 DNA polymérase à l'activité 5'→3' polymérase et 3'→5' exonucléase permettant de remplir les extrémités 5' entrantes et de dégrader les extrémités 3' sortantes pour les convertir en bouts francs, et de T4 polynucléotide kinase pour la phosphorylation.

4. Clonage

Le clonage des fragments de la banque d'ADN shotgun est réalisé dans un vecteur de clonage permettant de cloner de tels fragments à bout-francs.

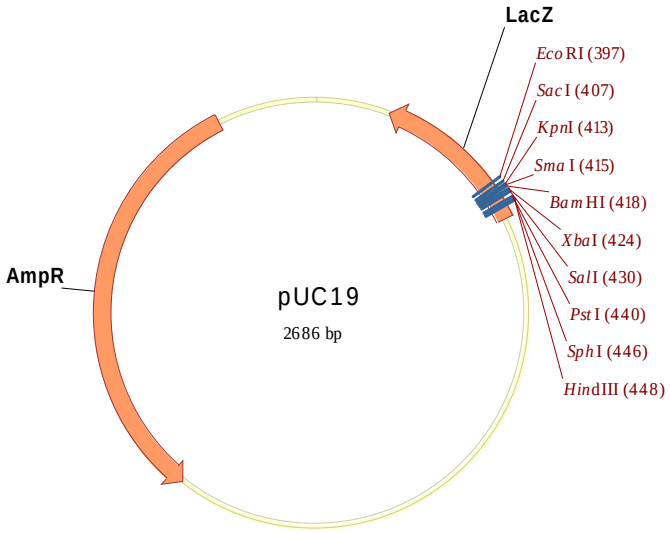


Figure 17 Carte de restriction du vecteur de pUC19

Le vecteur est préparé en transformant des cellules *E.coli* AmpS avec le plasmide pUC19 (Yanisch-Perron *et al.,* 1985). Le mélange est étalé sur milieu de culture solide contenant de l'ampicilline. Une colonie est mise en culture durant une nuit dans un tube contenant 4mL de milieu LB additionné d'ampicilline à concentration finale de 50µg/ml. Cette culture est inoculée dans 200mL de milieu LB Amp et incubée durant une nuit à 37°C. Le plasmide pUC19 est extrait par utilisation du kit Midi Prep de Qiagen suivant les recommandations du fabricant et resuspendu dans 100µL de TE 10mM pH8.

Le vecteur est ouvert par 4µL d'enzyme de restriction *SmaI*, 15µL de tampon de restriction 10X, 15µL de BSA 10X. La digestion est incubée pendant 4h à 25°C.

Dans le but de diminuer le nombre de faux positifs lors du clonage, il faut s'affranchir de l'ADN n'ayant pas été digéré. Pour cela, l'ADN digéré est déposé sur gel d'agarose 0,8% TAE 1X BET. L'ADN correspondant au plasmide linéarisé est excisé du gel avec un scalpel puis purifié avec le kit QiaEXII de Qiagen et resuspendu dans 22µL d'eau milliQ autoclavée.

Une autre cause de faux positifs de clonage est la recircularisation du plasmide sans insert lors de la ligature. Pour limiter ce phénomène, le vecteur linéarisé est déphosphorylé par ajout de 3 unités de SAP (Phosphatase alcaline de crevette) et 2,5µL de tampon de réaction 5X. La réaction est incubée 30minutes à 37°C et la phosphatase est inactivée 15 minutes à 65°C.

500ng d'ADN nébulisé sont ligaturés avec 50ng de vecteur pUC19 *SmaI*, 10u de ligase, et 0,5µL de Tampon de ligature 1X pendant 4h à 16°C dans un volume final de 5µL.

100µL de cellules compétentes *E.coli* XGold Kan^R Amp^S sont transformées avec les 5µL de produit de ligature par 45 secondes de choc thermique à 42°C. Les bactéries transformées sont sélectionnées par ensemencement sur une gélose nutritive sélective LB Amp IPTG Xgal.

5. Miniprep

Les colonies blanches contenant potentiellement un insert sont repiquées dans 4 mL de milieu LB en présence d'ampicilline et incubées durant 14h à 37°C. 2mL de cette culture sont culottés par 5 minutes de centrifugation à 3000rpm. Le surnageant est éliminé et le culot cellulaire est resuspendu par 250µL de solution I (50mM Glucose, 25mM TrisHcl pH8, 10mM EDTA pH8 et RNAseA à 50µg.mL^{-1}) Les cellules sont lysées par ajout de 250µL de solution II (0,2M NaOH, 1% Sodium Dodecyl Sulfate). La lyse est immédiatement stoppée par ajout de 350µL de solution III (3M Acétate de Potassium, 11,5% Acide Acétique glacial). Une centrifugation de 30 minutes à 14.000rpm permet de culotter les débris cellulaires, l'ADN chromosomique et les protéines. Le surnageant contenant l'ADN plasmidique est transféré dans un tube eppendorf 2mL. L'ADN est précipité par ajout de 500µL d'isopropanol et culotté par 15 minutes de centrifugation à 14.000rpm. L'isopropanol est aspiré à l'aide d'une trompe à vide et lavé par ajout de 500µL d'éthanol 75%. L'éthanol est aspiré à l'aide d'une trompe à vide et le culot d'ADN est séché au Speedvac. L'ADN est réhydraté dans 100µL de TrisHCl 10mM pH8. 5 µL de cette préparation sont digérés pendant 1h à 37°C par les enzymes de restriction *HindIII* et *EcoRI* bordant le site de clonage du vecteur pUC19. Une migration électrophorétique sur gel d'agarose 0,8% permet de confirmer la présence d'un insert avant séquençage.

6. Séquençage

Le séquençage a été effectué avec les amorces M13R et M13F présentes sur le vecteur de clonage pUC19. La réaction de séquence se fait avec le kit Big Dye Terminator v3. La séquence est produite par séquenceur capillaire ABI Prism 3100 de la plateforme technique Ouest Génopôle à la station biologique de Roscoff. La qualité des chromatogrammes est analysée ChromasPro. Le nettoyage des séquences consiste à éliminer la partie de la séquence correspondant au vecteur de clonage, il est réalisé avec Vecscreen (http://www.ncbi.nlm.nih.gov/VecScreen/VecScreen.html)

7. Assemblage

Les séquences sont assemblées pour former des contigs. Cet assemblage est réalisé avec le logiciel SeqMan 6.0. (Lasergene)

8. Lissage

Pour être valide, un génome doit avoir une couverture de séquençage moyenne de 10X et une couverture minimum de 6X. Chaque position nucléotidique doit être au minimum séquencée trois fois sur chaque brin. Le séquençage shotgun laisse invariablement des zones du génome dont la couverture de séquence n'est pas suffisante. L'amplification et le séquençage spécifique de ces zones sont nommés lissage, qui permet d'augmenter la couverture de séquençage pour obtenir une séquence de bonne qualité.

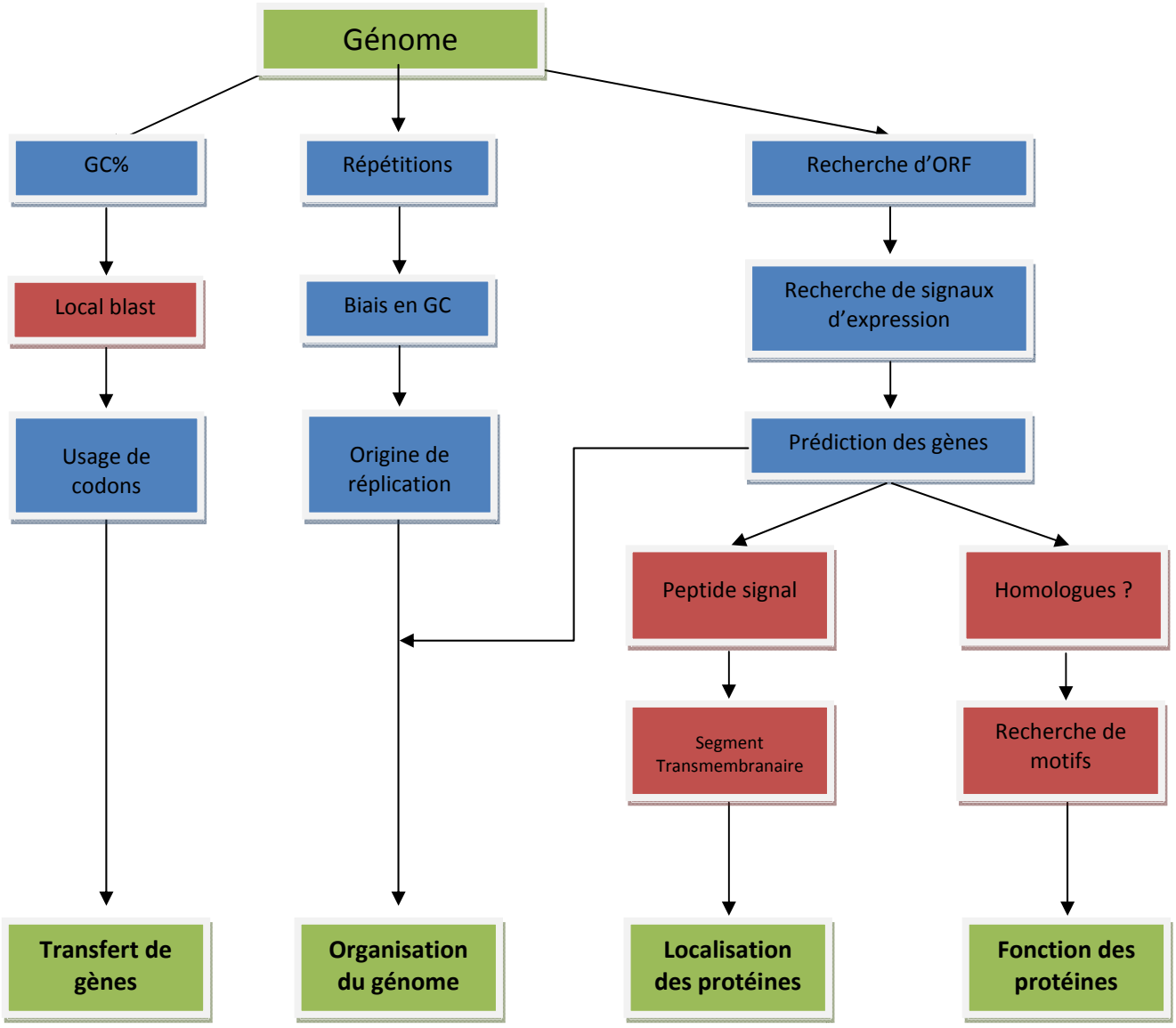
Des amorces flanquant les zones de faible couverture sont dessinées avec le logiciel FastPCR (http:/www.biocenter.helsinki.fi/bi/Programs/fastpcr.htm). L'amplification est réalisée avec 1,25 unité de polymérase à haute fidélité Pfu de *Pyrococcus furiosus*. 0,5µg d'ADN plasmidique servent de matrice, 1X Pfu DNA polymérase buffer, 200µM de chaque dNTP, 0,5µM des amorces sens et antisens et 21µL d'eau milliQ autoclavée. Le produit de PCR est ensuite directement séquencé en utilisant les amorces utilisées pour l'amplification.

9. Annotation

Les analyses *in silico* ont été réalisées sur une station de travail Sony VGN munie des systèmes d'exploitation Windows XP et Linux (distribution Ubuntu 7.04). Une panoplie de logiciels « on line » est installée sur la machine pour analyser les séquences nucléotidiques et protéiques. Deux principaux logiciels polyvalents ont servis à manipuler les séquences, à construire des banques de données locales, et à annoter les génomes : Vector NTI 10.3 (www.invitrogen.com) et Artemis

(www.sanger.ac.uk). Ces logiciels sont librement accessibles à la communauté scientifique sur simple demande. D'autre part ces analyses ont été automatisées par l'écriture de script perl utilisant la librairie bioperl permettant l'interaction avec les bases de données publiques et la plupart des logiciels utilisés en bioinformatique. Toutes les données ont été archivées dans une base de données MySQL interfacée en PHP.

Méthodologie générale



9.1 *Analyse de la séquence nucléotidique*

Les caractéristiques générales des séquences nucléotidiques comme le pourcentage en GC, la fréquence des bases sont données par les programmes généralistes tels que Vector NTI. La

recherche d'éventuelles séquences codant pour des ARN ribosomaux se fait avec le programme BlastN (voir plus loin) et celle des ARN de transfert se fait sur le site tRNA Scan (http://lowelab.ucsc.edu/tRNAscan-SE/). L'usage des codons a été calculé grâce aux programmes en ligne suivants : GCUA (http://bioinf.may.ie/gcua/) CodonW 1.3 (codonw.sourceforge.net) et SWAAP 1.02 (www.bacteriamuseum.org/SWAAP/). Les détails du mode opératoire peuvent être trouvés sur ces sites en consultant les manuels.

9.2 *Détermination des séquences répétées*

Les séquences répétées jouent un rôle important dans les mécanismes de réplication de l'ADN et dans la régulation de l'expression des gènes. Elles sont la cible de fixation de nombreux facteurs de transcription dimérique, chaque monomère reconnaissant une séquence. Les répétitions sont également à l'origine de la formation de structures secondaires d'ADN, impliquées dans la fixation de protéines et dans des mécanismes de régulation. Les origines de réplication sont généralement riches en séquences répétées.

Différents types de répétitions ont été recherchés :

Les **répétitions directes** (**DR** pour *direct repeats*) sont des séquences identiques (ou très similaires) présentent au minimum en deux copies sur le même brin de la même molécule d'ADN. Des répétitions adjacentes (sans espace entre les répétitions) sont appelées **répétitions en tandem (TR).**

Les **répétitions inversées** (**IR** pour *inverted repeats*) comportent au moins deux copies de la même séquence d'ADN répétées en sens inverse sur le même brin de la même molécule d'ADN. Des répétitions inversées adjacentes constituent un **palindrome (Pal)**. La présence de mésappariements et de gaps au sein de répétitions inversées forme des structures secondaires particulières nommées : *stem loop bulge*.

Divers outils de recherche utilisés proviennent du « package EMBOSS » et peuvent être installés sous environnement XP ou Linux. RepeatFinder est un outil en ligne permettant également la détermination des différents types de répétitions. Ces outils possèdent pour principale lacune de ne détecter seulement que les répétitions exactes et ne permettent pas la détection de structures secondaires telles que les boucles et épingles. L'analyse réalisée avec ces outils généralistes a été complétée par différents algorithmes d'analyse en ligne possédant chacun sa spécificité, dont la détection de mésappariements. REPFIND est spécialisé dans la recherche de répétitions directes

et de répétitions directes en tandem. PALINDROME permet la localisation de séquences palindromiques. EINVERTED localise les répétitions inverses situées à proximité dans le génome, sa tolérance des mésappariements et des gaps permet la détection de structures de type HBH.

Tableau 10 Programme utilisés pour la recherche de motifs protéiques et de répétitions

Programme	Nature de la répétition	Site d'Internet
Package Emboss	-Séquences répétées directes -Répétition en tandem -Séquences répétées inverses -Palindrome	http://emboss.sourceforge.net
REPFIND	-Séquences répétées directes -Répétition en tandem	http://zlab.bu.edu/repfind
Repeat Finder	-Séquences répétées directes -Répétition en tandem -Séquences répétées inverses -Palindrome	http://www.proweb.org/proweb/Tools/selfblast.html
EINVERTED	-Séquences répétés inverses proches -Tolère les mésappariements -Détection de structures HBH	http://bioweb.pasteur.fr/seqanal/
PALINDROME	-Détection de palindromes -Tolère les mésappariements	http://bioweb.pasteur.fr/docs/EMBOSS/palindrome.html

9.3 *Origine de réplication par la méthode du biais cumulatif en G+C et A+T*

La méthode du biais cumulatif en G+C et A+T (en anglais GC skew et AT skew) est basée sur l'hypothèse que l'asymétrie du mécanisme de réplication engendre une mutation différentielle entre les brins d'ADN direct et retardé. En effet, au cours de la réplication, le brin retardé est plus longtemps exposé sous forme simple brin, plus sujet à mutation que sous forme double brin. Combiné à la sélection naturelle, ce facteur peut être responsable d'une distribution biaisée en bases sur le génome. De fait, on observe généralement que le brin direct contient plus de résidus

guanine que de résidus cytosine. Le programme « cumulative skew » calcule le biais par la fonction Σ(G-C)/(G+C) (http://bioinformatics.upmc.edu/SKEW/index.html). Typiquement, une inversion de polarité au niveau de l'origine (valeur minimum) et de la terminaison (valeur maximum) est visible sur une représentation graphique du biais cumulatif en G+C. Cette inversion du biais en G+C, corrélée à la concentration de séquences répétées permet la prédiction de la localisation de l'origine de réplication.

9.4 *Prédiction des cadres ouverts de lecture*

L'identification des ORFs (cadres ouverts de lecture ou « Open Reading Frame ») est réalisée avec le logiciel VectorNTI 10.3 selon les paramètres suivants : un ORF démarre par un codon ATG, GTG ou TTG et se termine par à un codon TAA, TAG ou TGA. La longueur minimale de ce cadre ouvert de lecture est fixée arbitrairement à 100 paires de bases.

Quatre programmes de prédiction de gènes sont ensuite utilisés pour évaluer la robustesse de codage des ORFs. Le programme FGENESB (www.softberry.com/berry/), ainsi que deux algorithmes de recherche de chaînes de Markov cachées (HMMs) développées sous le nom de GeneMark, version standard (http://opal.biology.gatech.edu/GeneMark) et version Genemark.hmm (http://opal.biology.gatech.edu/GeneMark/hmmchoice.html). Ce dernier algorithme possède une très bonne acuité dans la localisation de l'extrémité 3' grâce à une analyse conjointe du positionnement du site de fixation du ribosome (RBS). Le dernier algorithme utilisé est nommé Glimmer 3.0 (http://www.cbcb.umd.edu/software/glimmer/), tout comme Genmark il est basé sur un modèle de Markov interpolé, particulièrement puissant pour la sélection d'ORFs chevauchant.

Tableau 11 Programmes utilisés pour la recherche de codons starts

Programme	Site d'Internet	Paramètres	Algorithme
Vector NTI	http://emboss.sourceforge.net	-Codon start -Taille min des ORFs	
FGENSB	http://www.softberry.com/berry/	-Codon start -Taille min des ORFs -Génome modèle -Prédiction des opérons	Modèle de Markov
GeneMark	http://opal.biology.gatech.edu/GeneMark/GeneMark/hmmchoice.html	-Codon start -Taille min des ORFs -Ordre du modèle RBS	Algorithme de Viberti Modèle de Markov
GeneMark.HMM	http://opal.biology.gatech.edu	-Codon start -Taille min des ORFs -Ordre du modèle RBS -Taille et pas de la fenêtre	Modèle de Markov Chaines cachées (HMM)
Glimmer 3.0	http://www.cbcb.umd.edu/software/glimmer/	-Codon start -Taille min des ORFs -Score seuil -Forme du génome	Modèle de Markov interpolé (IMM)

Pour être validé, un ORF doit avoir été prédit par VectorNTi et également par au moins un des 3 programmes supplémentaires. Bien souvent, plusieurs codons d'initiation de traduction sont possibles pour définir un ORF. Ces algorithmes ne sont pas très performants pour analyser les signaux de transcription et de traduction. Une analyse visuelle de la séquence est nécessaire pour déterminer le codon d'initiation de traduction le plus vraisemblable.

Les séquences situées en amont du codon «start» sont examinées avec Vecteur NTI. La localisation des sites de fixation du ribosome (Ribosome Binding Site RBS) également appelée séquence de Shine-Dalgarno, des promoteurs et des terminateurs se fait visuellement. La séquence RBS est par définition complémentaire de l'extrémité 3' de l'ARNr 16S de l'organisme considéré, elle est très conservée au sein d'un même groupe de procaryotes. C'est par complémentarité de bases que le ribosome se fixe sur l'ARNm en amont du site d'initiation de la

traduction. Dans le cadre de cette étude, la séquence complémentaire de l'extrémité 3' des Thermococcales est GGAGGTGA. Le second critère à prendre en compte pour localiser le bon codon d'initiation est sa distance par rapport au RBS, cette séquence doit être positionnée entre -13 et +4 par rapport au codon «start». Le programme RBSfinder (ftp://ftp.tigr.org/pub/software/RBSfinder/) est une méthode probabiliste qui permet d'améliorer la précision des systèmes d'identification de gènes par localisation des sites RBS. Le modèle probabiliste des RBS est généré par l'algorithme à partir d'une séquence consensus établie au préalable. Les motifs conservés sont recherchés en amont du codon d'initiation prédit (fenêtre de 4 à 20pb). Le calcul de similarité entre la séquence consensus et le motif trouvé permet de déterminer le RBS.

Une procédure itérative permet de déterminer tous les RBS, les codons d'initiation sont acceptés ou relocalisés selon le score du RBS et les préférences de codon (ATG>GTG>TTG).

La séquence consensus du promoteur est dérivée de l'analyse des génomes des *Pyrococci*. Selon le code IUPAC, cette séquence est YTTAWA (Y=C ou T et W=A ou T). Ce promoteur doit être situé dans une fenêtre de 60 nucléotides en amont du codon «start».

Les terminateurs potentiels chez les archées hyperthermophiles se caractérisent par 2-3 répétitions de 4-6 pyrimidines : TTTTT, espacées de quelques paires de bases situées dans une fenêtre de 50 nucléotides en aval du codon « stop ».

9.5 *Recherche d'homologues dans les bases de données*

La recherche de protéines homologues aux protéines déduites des ORF dans les bases de données s'effectue avec l'algorithme BlastP (www.ncbi.nlm.nih.gov/BLAST/) et la banque non redondante (nr) de Genbank. Le principal paramètre à régler est la matrice de substitution. En effet, le choix de la matrice de substitution peut fortement influencer les résultats de l'analyse. Les matrices PAM (Percent Accepted Mutation) sont basées sur des alignements globaux de protéines étroitement liées. La matrice PAM1 est calculée à partir de comparaisons de séquences possédant moins de 1% de divergence). Au contraire, les matrices BLOSUM (BLOck SUM) sont issues d'alignements locaux de domaines conservés. La matrice BLOSUM62 est celle utilisée par défaut. Elle est adaptée à la comparaison de protéines peu distantes, le premier BLAST réalisé est fait avec cette matrice. En fonction du résultat obtenu, un second BLAST avec une matrice optimisée permet d'affiner la recherche. Les matrices BLOSUM80 et 90, ainsi que les matrices PAM30 et 70, sont désignées pour la comparaison des séquences les moins divergentes. Au contraire, les matrices BLOSUM45 et PAM250 sont adaptées à la comparaison de séquences divergentes. Afin

de filtrer ou élargir la quantité de données récupérées, la taille du mot permettant l'accroche à une séquence (seed), ainsi que les pénalités d'ouverture et d'extension de gap ont été ajustées.

La recherche de similarité entre plasmides issus du séquençage est effectuée par BlastP en ligne de commande suite à la création d'une banque de données de séquences protéiques (FormatDB). Seules les similitudes ayant une E value < 10^{-3} sont retenues pour une recherche plus détaillée en utilisant l'algorithme itératif Psi-Blast.

Il est important de noter que les E value obtenues dépendent de la matrice de substitution et de la taille de la banque de données utilisée. Les comparaisons de E value ne sont pas significatives si l'on fait varier un de ces paramètres.

Le pourcentage d'identité en acides aminés entre deux protéines homologues est calculé par alignement global des deux molécules avec LALIGN (www.ch.embnet.org/software)

9.6 *Caractéristiques physicochimiques*

Les propriétés physico-chimiques générales : masse, potentiel isoélectrique, composition en acides aminés ont nécessité l'utilisation du logiciel VectorNTI ainsi que des serveurs PepStats (bioweb.pasteur.fr/seqanal/interfaces/pepstats.html) et ProtParam (www.expasy.ch/tools/).

9.7 *Recherche de motifs et de domaines*

Un motif est défini comme un ensemble de résidus aminoacides proches dans la séquence, conservés et importants pour la fonction d'une protéine donnée. Un domaine protéique est, quant à lui, une unité compacte qui forme une structure tridimensionnelle stable. Un domaine est caractérisé par des acides aminés conservés qui ne sont pas forcément situés à proximité dans la séquence. Généralement, une fois la protéine repliée, ces acides aminés conservés se retrouvent à proximité les uns des autres.

L'interrogation de banques de motifs et de domaines, généralistes et spécialisées, est nécessaire pour obtenir des informations sur la fonction d'une protéine. L'efficacité de la recherche dépend principalement de la syntaxe des motifs. En effet, les expressions régulières ne peuvent pas toujours rendre compte de la variabilité des motifs, en particulier chez les *Archaea* qui sont moins documentées que les autres règnes du vivant. Malgré les imperfections de définition des motifs, les banques PROSITE (motifs et profils de protéines), BLOCKS (motifs et alignements de séquences de protéines), PRINTS (motifs et signatures de protéines) sont des outils indispensables à l'analyse de protéines.

Une autre approche pour l'étude des protéines est la recherche de domaines via les banques SMART (famille de domaines), PFAM (alignement et modèle HMM des domaines de protéines) CDD (base de données des domaines conservés) et COG (cluster of orthologous groups). Ces banques sont issues d'alignements de séquences protéiques homologues.

La recherche de séquences informatives sur les protéines déduites a été réalisée avec un score d'alignement des séquences de 10 et une matrice de substitution BLOSUM62. Ces paramètres permettent d'extraire les motifs et les domaines qui varient des modèles consensus.

Tableau 12 Bases de données de motifs et domaines utilisés

Programme	Site d'Internet
ScanPROSITE	http://expasy.org/tools/scanprosite/
BLOCKS	http://blocks.fhcrc.org/blocks/blocks_search.html
PRINTS	http://www.bioinf.manchester.ac.uk/fingerPRINTScan/
SMART	http://smart.embl-heidelberg.de/
PFAM	http://www.sanger.ac.uk/Software/Pfam/search.shtml
CDD	http://www.ncbi.nlm.nih.gov.ilsprod.lib.neu.edu/Structure/cdd/cdd.shtml
COG	http://www-archbac.u-psud.fr/genomics/cog_guess.html

9.8 *Prédiction structurale*

L'identification des propriétés structurales des protéines, telles que le peptide signal, les segments transmembranaires ou les domaines coiled-coil apporte des informations sur la fonction de la protéine, même en absence d'homologues dans les bases de données.

9.8.1 Peptide signal

Un peptide signal est une courte séquence peptidique située à l'extrémité N-terminale d'une protéine. Il signale que cette protéine doit être secrétée dans le milieu extracellulaire ou s'insérer dans une membrane de la cellule. La recherche d'un éventuel peptide signal est faite grâce à SignalP (www.cbs.dtu.dk/services/SignalP/) qui est une méthode de réseaux de neurones développée à partir de peptides signaux caractérisés expérimentalement, ou encore avec le

programme iPSORT (http://psort.nibb.ac.jp/form.html) qui permet également de prédire l'adressage des protéines.

9.8.2 Segments transmembranaires et coiled-coil

Les méthodes de prédiction de segments transmembranaires, ou plus précisément d'hélices α membranaires dans les protéines, sont reliées aux profils d'hydrophobicité des chaînes de polypeptides. En complément de la localisation des segments transmembranaires, plusieurs programmes prédisent la topologie de la protéine comme Tmpred (www.ch.embnet.org/software/TMPRED_form.html) ou TMHMM (www.cbs.dtu.dk/services/)

9.8.3 Domaines coiled-coil et leucine zipper

Les domaines coiled-coil sont des superhélices α caractérisées, au niveau de la séquence, par une période de 7 résidus hydrophobes, habituellement des leucines. La recherche d'une telle périodicité est la base de l'algorithme de reconnaissance coiled-coil du programme COILS (www.ch.embnet.org/software/COILS_form.html). La prédiction et la visualisation de motifs leucine zippé peuvent également être effectuées avec le programme 2ZIP (www.bioinf.man.ac.uk/resources/)

9.9 *Etudes phylogénétiques*

La manipulation des séquences nucléiques ou protéiques est réalisée à l'aide de BioEdit. L'alignement des séquences est réalisé grâce aux programmes d'alignements sélectionnés dans le tableau Tableau 13.

Le choix du jeu de séquences à aligner, de la matrice de substitution utilisée ainsi que des pénalités d'ouverture et d'extension de gap sont des paramètres cruciaux pour l'analyse phylogénétique. La qualité de l'alignement conditionne l'arbre phylogénétique obtenu. L'alignement « automatique » obtenu avec les paramètres standards est affiné par un réalignement avec des paramètres optimisés. Finalement, cet alignement est corrigé manuellement *de visu*, avant tout pour retirer de l'analyse les régions où l'alignement est «ambigu».

Tableau 13 Programmes d'alignement de séquences

Programme	Site d'Internet	Référence
Muscle	http://www.drive5.com/muscle/	Edgar et al., 2004
Tcoffee	http://www.tcoffee.org/	Notredame et al., 2000
RDP9	https://rdp.cme.msu.edu/login/myrdp/	Cole et al., 2007
ClustalW	http://www.ebi.ac.uk/clustalw/	Thompson et al., 1994

Une fois que l'on dispose d'un alignement non ambigu (c'est-à-dire où toutes les positions homologues sont alignées, une analyse phylogénétique est réalisée selon différentes méthodes telles que Neigbhor Joining, Maximum likelihood, ou Maximum de Parcimonie reposant sur différents modèles évolutifs. La robustesse des arbres obtenus est évaluée par la technique de ré-échantillonnage aléatoire (bootstrap). Finalement, un cladogramme consensuel est dessiné.

Tableau 14 Programme de phylogénie

Programme	Site d'Internet	Référence
Phylip	http://evolution.genetics.washington.edu/phylip.html	Felsenstein et al., 1993
TreeCon	http://bioinformatics.psb.ugent.be/psb/Userman/treecon_userman.html	Van de Peer et al., 1997
Phylowin	http://pbil.univ-lyon1.fr/software/phylowin.html	Galtier et al., 1996
MEGA 4.0	http://www.megasoftware.net/	Tamura et al., 2007

10. Taxonomie de souches portant un plasmide

La taxonomie de la souche considérée est basée sur la séquence des gènes codant les ARNr 16S, 23S ainsi que de la région intergénique les séparant.

Ces gènes sont amplifiés à partir d'ADN total par réaction de polymérisation en chaîne produisant un fragment d'environ 1900pb. 1µL d'ADN est additionné de 2µL $MgCl_2$ 25mM, 5µL de Tampon10X, 1µL de dNTP 10mM, 1µL d'amorce 4F (5'-TCC GGT TGA TCC TGC CRG-3') 10µM, 1µL d'amorce 23R (5'- CTT TCG GTG GCC CCT ACT-3'), 1µL d'Uptitherm polymérase (Interchim^{TM}) et 40µL H_20 MQ. Le programme d'amplification comprend 30 cycles : 95°C 30sec, 53°C 30sec et 2min 72°C.

Le produit de PCR étant trop grand pour être séquencé avec le couple d'amorce utilisé pour l'amplification, une troisième amorce 1492R (5' CGG TTA CCT TGT TAC GAC TT-3') est utilisé durant le séquençage. L'assemblage (SeqMan, Lasergene™) des trois séquences permet de reconstituer l'intégralité des deux gènes ainsi que la région intergénique.

Un arbre phylogénétique est construit suivant le protocole détaillé précédemment.

11. Représentation graphique

La représentation graphique des cartographies de plasmides est réalisée par exportation des cartes annotées sous VectorNTI vers le logiciel de dessin vectoriel libre InkScape (http://www.inkscape.org)

12. Automatisation, Organisation et Pérénisation des donneés

L'automatisation des traitements bioinformatiques a été réalisée par l'écriture de script perl, en utilisant de nombreuses applications développées au sein de la librairie Bioperl. Les données servant de requetes sont contenues dans une base de données MySQL servant également d'entrée pour produire les analyses avec les scripts Perl. La communication entre la base de données et les scripts s'effectue en PHP par l'intermédiaire un simple navigateur web, tout comme l'administration et le remplissage de la base (Serveur Apache).

VI. Puce à ADN

La diversité de la collection de plasmides non séquencés a été explorée grâce à la création d'une puce à ADN comportant 400 des gènes des plasmides précédemment séquencés ainsi que les gènes des éléments viraux intégrés de *Thermococcus kodakaraensis* KOD1. Par hybridation compétitive entre l'ADN d'un plasmide inconnu marqué au Cy3 et de l'ADN contrôle correspondant aux gènes présents sur la puce, il est possible de connaître certains gènes présents sur notre plasmide inconnu.

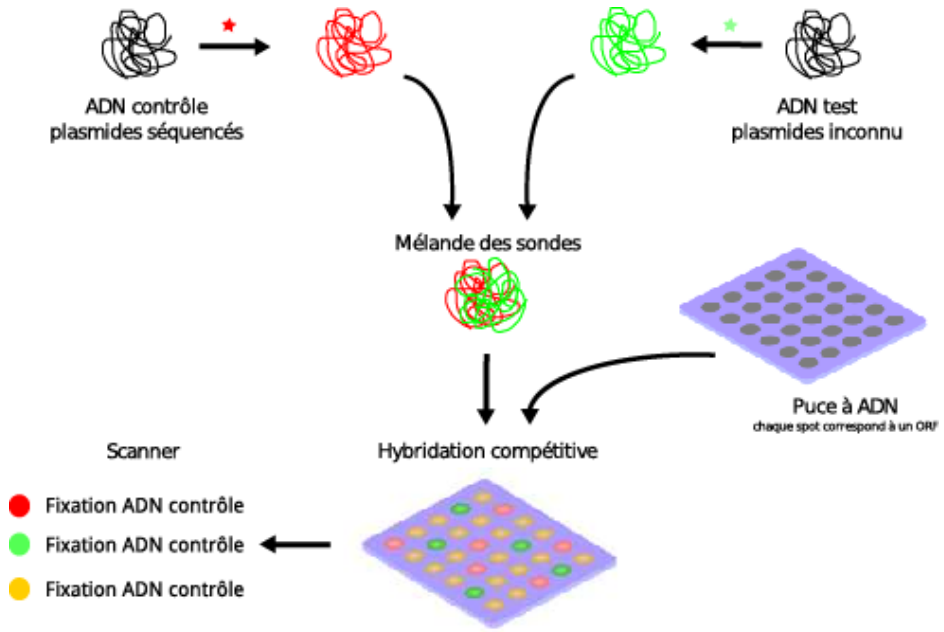


Figure 18 Principe de l'hybridation compétitive sur puce à ADN

Les étapes de fabrication, d'hybridation et de lecture de la puce sont réalisées sur la plateforme transcriptomique de la Ouest Génopôle de Nantes (IFR26).

1. Conception de la puce

1.1 *Amplification des gènes à déposer les puces*

L'ADN présent sur la puce correspond à un gène de plasmide amplifié par réaction de polymérisation en chaîne. L'amplification d'un gène se fait grâce à un couple d'amorces spécifiques dessinées avec le logiciel VectorNTI advance 10.3 (Invitrogen). La température d'hybridation de chaque couple d'amorces dépend à la fois de leur longueur et de leur

pourcentage en GC. Dans un but de facilité de manipulation, les amorces ont une longueur comprise entre 18 et 22 pb et hybrident à des températures comprises entre 53°C et 56°C.

Chaque gène est amplifié à partir de 500ng d'ADN plasmidique servant de matrice, 1 unité de Taq polymérase, 4µL de $MgCl_2$ 25mM, 1µL de dNTP 10mM 1µL de chacune des amorces à 10µM, 5µL de tampon 5X et 33µL d'H_20 MQ. 40 cycles d'amplification sont nécessaires : dénaturation 30sec à 95°C, hybridation des amorces 30sec à température variable et élongation 2min30sec. 3µL de produit d'amplification sont déposés sur gel d'agarose 0,8% TAE1X pour vérifier la spécificité de l'amplification.

1.2 *Purification des produits de PCR*

Les produits de PCR sont purifiés sur colonne de Résine Sephadex G50 présente au fond d'un puits d'une plaque Multiscreen (Millipore™). Cette colonne permet le dessalage des produits de PCR et l'élimination des nucléotides non incorporés, l'excès d'amorces et de polymérase. La limite d'exclusion de la résine G50 est d'environ 30 bases.

Les produits de PCR sont transférés dans cette plaque Multiscreen 96 puits avant d'être filtrés grâce à une pompe à vide (à une pression de 14Hg) jusqu'à ce que la résine soit sèche. Les produits de PCR restent attachés dans la colonne. L'élution de l'ADN est réalisée par ajout de 40µL de SSC 4,3X dans chaque puit de la plaque et 2 heures d'agitation orbitale à température ambiante. Après agitation, 37µL des produits de PCR sont transférés dans une plaque 96 puits Thermofast à jupe et 2µL sont prélevés pour un contrôle de pureté. Pour finir, 15µL de bétaïne 5M sont ajoutés aux 35µL de produit de PCR restants.

1.3 *Quantification des produits de PCR*

Les 2µL prélevés à l'étape précédente sont complétés de 6µL d'eau distillée et de 2µL de tampon de charge BBP5X. Le mélange est ensuite déposé sur un gel d'agarose 1,5% de 600mL.

1.4 *Dépôt de l'ADN sur la puce*

Les lames utilisées sont des lames GAPII (Corning™). Elles sont α-amino-propylsilanées et nécessitent 5 secondes d'hydratation à la vapeur d'un bain-marie à 60°C et un séchage de 5 secondes sur une plaque chauffante, préchauffée à 70°C et recouverte d'un papier aluminium. L'ADN est ensuite déposé sur la lame par un robot spotter Eurogrid d'Eurogentec (Corning™). Chaque produit PCR est déposé en quintuplicat sur la lame pour valider statistiquement les réactions d'hybridations.

1.5 *Fixation de l'ADN sur la puce*

Les lames α-amino-propylsilanées sont caractérisées par la présence de groupements -NH_3 permettant la fixation des groupements phosphates de l'ADN. Lors d'un passage au four « cross-link », les rayons UV (600×100µjoules) envoyés sur la lame transforment la liaison ionique qui existait entre les terminaisons –NH_3 de la lame et les groupements phosphates d'ADN en liaisons iono-covalentes fortes. L'ADN est alors fixé covalemmente à la lame.

Les lames sont ensuite plongées durant 15 minutes dans un bain agité contenant : 130 ml de 1-méthyl-2-pyrrolidone, 2,4g d'anhydre succinique et 10ml de tampon borate soumises à agitation pendant 15 minutes. Elles sont ensuite plongées dans un bain d'eau à 95°C pendant 2 minutes puis dans un bain d'éthanol à 95%. Une centrifugation de 3 minutes à 700g à température ambiante permet le séchage des lames.

2. Marquage et hybridation

L'ADN d'un plasmide inconnu est marqué avec le fluorophore Cy3 tandis qu'un contrôle (100% d'hybridation) est marqué au fluorophore Cy5. (Tab)

Tableau 15 Fluorophores utilisés pour les hybridations sur puce à ADN

cyanine	Nom	longueur d'onde d'émission
Cy3™	indodicarbocyanine 3-1-O-(2-cyanoethyl)-(N,N-diisopropyl)-phosphoramidite	563 - 570 nm
Cy5™	indodicarbocyanine 5-1-O-(2-cyanoethyl)-(N,N-diisopropyl)-phosphoramidite	662 - 670 nm

2.1 *Fragmentation de l'ADN plasmidique*

La réaction de marquage est sensible à la longueur des fragments d'ADN. L'ADN est donc fragmenté aléatoirement par nébulisation et précipité à l'isopropanol avant d'être marqué.

3. Etiquetage de l'ADN plasmidique

Avant de pouvoir incorporer le fluorophore, l'ADN doit être étiqueté avec des bases modifiées : les aha-dUTP. Cette modification permet la fixation du fluorophore. Dans chaque tube, on aliquote 20µl de solution 2X Random primer, 1µg d'ADN à marquer et on complète par de l'eau

ultra-pure traitée au DEPC jusqu'au 44 µl. L'ADN est incubé à 95°C pendant 5 min et directement plongé dans la glace pour être dénaturé.

5µl de 10X nucléotides Mix contenant des aha-dUTP et 1µL d'enzyme de Klenow sont ajoutés. Après homogénéisation du mélange, la réaction est incubée 2h à 37°C. La réaction est stoppée par ajout de 5µl de Stop Buffer. La réaction peut être utilisée immédiatement ou conservée à -20°C.

3.1 *Purification des réactions d'étiquetage*

Avant de réaliser le marquage fluorescent, il faut éliminer les amorces aléatoires ainsi que les aha-dUTP non incorporés.

Aux 55µL de réaction d'étiquetage, on ajoute 200µl de tampon de fixation. La solution est placée dans une colonne PureLink™ et centrifugée 1min à 10000g. 650 µl de tampon de lavage sont ajoutés avant de centrifuger à nouveau. Toute trace de liquide est éliminée par une centrifugation de 3min. La colonne est placée sur un nouveau tube eppendorf. L'élution de l'ADN est réalisée par ajout de 50µL d'eau traitée au DEPC. Une dernière centrifugation de 3 min à 10000g permet de récupérer l'ADN étiqueté et purifié au fond du tube.

3.2 *Couplage avec la fluorescence*

Le couplage avec la fluorescence consiste à incorporer le fluorophore au niveau des bases modifiées. A partir de cette étape, il faudra travailler à l'abri de la lumière directe car le fluorophore est photosensible. On ajoute à l'ADN purifié 10µL d'Acétate de sodium 3M, 2µl de glycogène et 300 µL d'éthanol absolu. Cette réaction est incubée 30 minutes à -20°C, et centrifugée à 14000rpm pendant 20minutes à 4°C. Le surnageant est éliminé, le culot est séché à l'air pendant dix minutes. Puis, il est solubilisé par un mélange contenant 5µl de tampon de couplage 2X, 3µL de d'eau traitée au DEPC, 2µL de DMSO, ainsi que 60µg de fluorophore. Le mélange réactionnel est incubé 1h à température ambiante**.**

3.3 *Purification des réactions de couplage à la fluorescence*

Le protocole de purification est le même que celui utilisé durant l'étape 3.1.

3.4 *Calcul de l'efficacité de marquage*

Avant de réaliser l'hybridation sur la puce, il faut s'assurer que le marquage a fonctionné. L'intensité de fluorescence étant proportionnelle à la quantité de fluorophore, l'efficacité de marquage est l'estimation du nombre de molécules de fluorophore incorporées à l'ADN. L'efficacité de marquage est normalisée en rapportant la fluorescence à 100pb. Les efficacités de

marquage sont également dépendantes du fluorophore utilisé, elles peuvent être calculées grâce aux formules empiriques suivantes :

Calcul de la quantité d'ADN : ADN(µg)=(A260-A320)×50(µg/ml)×volume en ml

Calcul de l'intensité de fluorescence :

$$\text{Cy3 (mole)} = \frac{(A555-A650)}{0{,}15 * vol\ (ml)} \qquad \text{Cy5 (mole)} = \frac{(A650-A750)}{0{,}24 * vol\ (ml)}$$

Base / ratio de fluorescence :

$$\text{Cy3 (mole)} = \frac{[(A260 - A320) - ((A555 - A650) * 0.04)]}{(A555 - A650) * 6600}$$

$$\text{Cy5 (mole)} = \frac{\left[(A260 - A320) - \left((A555 - A650) * 0\right)\right] * 239000}{(A650 - A750) * 6600}$$

555nm (pour Cy3™) et 650nm(pour Cy5™)

4. Hybridation des ADN plasmidiques sur la puce à ADN

4.1 *Identification des lames et fixation de l'ADN*

Dans le but de retrouver les lames durant l'hybridation, ces lames sont gravées du nom de l'ADN utilisé pour l'hybridation à l'aide d'une pointe de diamant.

4.2 *Blocage des sites*

Pour prévenir les hybridations aspécifiques, les lames sont plongées pendant 15 minutes dans un bain sous agitation contenant 130 ml de 1-méthyl-2-pyrrolidone (SIGMA), 2,4g d'anhydre succinique et 10ml de tampon borate 1M pH8.

4.3 *Dénaturation des lames*

L'hybridation nécessitant de l'ADN simple brin, les lames sont plongées dans un bain d'eau à 95°C pendant 2 minutes puis cinq fois dans un bain d'éthanol à 95%. Elles sont enfin centrifugées à 700g pendant 3 minutes à température ambiante pour les sécher au maximum.

4.4 *Préhybridation des lames*

Dans le but de limiter les hybridations aspécifiques, les lames sont préhybridées pendant une heure dans 150ml d'une solution SSC 3,5X, SDS 0,3%, BSA 1% préchauffée à 42°C. Elles sont

ensuite lavées par quelques allers-retours dans 5 bains d'eau distillée à température ambiante avant d'être centrifugées immédiatement à 700g pendant 3 minutes à température ambiante afin de les sécher.

4.5 *Hybridation*

Les cibles sont dénaturées 2 minutes à 98°C puis incubées 30 minutes à 37°C. Elles sont ensuite centrifugées 2 minutes à 13000 rpm. Les cibles sont déposées sur la zone à hybrider où ont été préalablement déposés les produits de PCR. Elles sont ensuite recouvertes d'une lamelle pour éviter l'évaporation. L'ensemble est placé dans une chambre à hybridation Telechem (ArrayIt^{TM}) qui est humidifiée par 20µL de SSC 3X. Une fois fermée hermétiquement, la chambre à hybridation est immergée pour hybridation dans un bain-marie à 42°C pendant une nuit à l'obscurité.

4.6 *Lavage*

Afin d'éliminer les sondes ne s'étant pas fixé de façon spécifique, les lames sont immergées dans un tampon SSC 2X, SDS 0,1%. C'est une étape importante de l'expérience. Le choix de la concentration du tampon de lavage permet de controler la stringence de l'hybridation. Les lames sont soumises à agitation pendant 2 minutes. Cette opération est réitérée avec un bain de SSC 1X puis deux bains de SSC 0,2X. Les puces sont ensuite centrifugées 3 minutes à 700g avant d'être scannées.

5. Lecture et interprétation des signaux

Lors de la lecture des puces, chaque spot est excité par un laser. En cas hybridation, il y émission de fluorescence mesurée à la longueur d'onde spécifique du fluorophore excité (Tableau 15). Les intensités de fluorescence mesurées pour chaque spot sont ensuite artificiellement colorées (Figure 19)

- Vert : seul l'ADN de l'échantillon testé s'est fixé. Le fluorophore correspondant est la Cy3
- Rouge : seul l'ADN contrôle du pool de référence s'est fixé. Le fluorophore correspondant est la Cy5.
- Jaune : les 2 ADN se sont fixés au même spot en quantité égale.

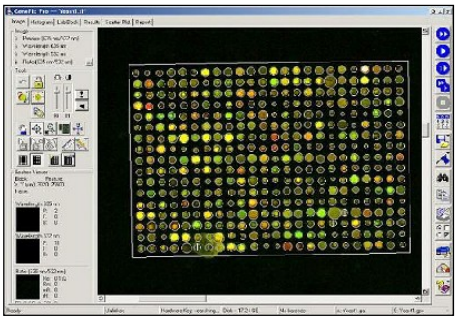


Figure 19 Capture d'écran de traitement des signaux d'hybridation d'une puce à ADN

Le scanner ScanArray ExpressHT (PerkinElmer Life Sciences™) est utilisé pour lire les lames. Il est muni de 2 lasers (excitation à 532nm et 635nm) qui permettent l'acquisition simultanée de la fluorescence émise par les 2 fluorophores.

Les images ainsi acquises sont enregistrées au format TIFF 16 bits en 65535 niveaux de gris. La gamme d'intensité détectée est donc comprise entre 0 (noir) et 65535 (blanc). Il est admis qu'à l'intérieur de cette gamme, l'intensité du signal augmente de manière linéaire avec le nombre de molécules de fluorophore sur le spot. Les images ainsi obtenues sont analysées avec le logiciel GenePix Pro 5.1 (Molecular Devices Corp.™) afin d'extraire les données numériques correspondantes à chaque spot. Les images sont colorées artificiellement (celle du canal Cy3 en vert et celle du canal Cy5 en rouge) et superposées pour visualisation.

Le logiciel permet de définir une grille sur l'image afin d'identifier chaque spot en lui assignant des coordonnées uniques, et de délimiter la surface du spot par rapport au reste de la lame. Le logiciel génère des données numériques correspondant pour chaque spot aux valeurs moyennes et médianes du signal, du bruit de fond local ainsi que d'autres paramètres (écart-type associé aux intensités des pixels d'un spot, le rapport signal/ bruit de fond). Pour notre analyse, les valeurs les plus importantes sont les valeurs brutes des intensités. A l'aide d'un tableur une colonne correspondant à la différence d'intensité de fluorescence entre la Cy3 (Echantillon) et le Cy5 (Contrôle) permet de discriminer les gènes portés par la puce présent sur le plasmide testé.

RESULTATS ET DISCUSSION

I. Abondance et Diversité

1. Collection de travail

257 isolats de Thermococcales, provenant de la souchothèque de Bretagne, ont servi de collection de travail.

Tableau 16 Souches utilisées comme collection de travail

Campagne	Abréviation	Origine	Année	Souches isolées
AMISTAD	AMT	Pacifique oriental	1999	102
IRIS	IRI	Atlantique	2001	59
EXTREME	EXT	Pacifique Est	2001	16
CIR	CIR	Indien	2001	16
SWEEP VENT	SV	Pacifique Sud	2004	33
STARMER	GE	Pacifique Sud	1989	31

Ces isolats ont été conservés en cryotubes dans des surgélateurs, pendant une durée comprise entre 3 mois et 5 ans.

18 des 257 isolats (7%) n'ont pas donné de croissance cellulaire après ensemencement : 7 (8%) de la campagne AMT, 3 (5%) de la campagne IRIS, et 8 (26%) de la campagne STARMER. Une cryoconservation trop longue et une microporosité à l'oxygène des cryotubes pourrait être néfaste à la pérennité de la souche conservée. Un répliquât de la collection de microorganismes a été réalisée depuis ces expériences en utilisant une autre technique de conservation sous forme de paillettes stockées dans de l'azote liquide.

2. Criblage de la collection à la recherche d'ADN plasmidique

Les 239 isolats de *Thermococcales* ont été criblés pour rechercher des plasmides, par extraction d'ADN extrachromosomique à l'aide d'une lyse alcaline. L'ADN obtenu a été digéré par des enzymes de restriction et visualisé par migration électrophorétique sur gel d'agarose. L'abondance relative de ces éléments est calculée, suivant l'origine géographique des souches, en

rapportant le nombre de souches présentant au moins un plasmide sur le nombre total de souches de la campagne (Tableau 17).

Tableau 17 Abondance plasmidique

Campagne	Nombre de couches criblées	Nombre de souches porteuses d'EG	Prévalence
AMISTAD	94	28	29,79%
IRIS	56	21	37,50%
EXTREME	16	4	25,00%
CIR	16	4	25,00%
SWEEP VENT	33	8	24,24%
STARMER	23	6	26,09%
Total	**239**	**71**	**29,83%**

L'observation de profils de restriction similaires indique l'existence de clones parmi les isolats. Ces clones sont identifiables car ils sont issus d'un même échantillon et portent des plasmides aux profils de restriction identiques. La présence de clones dans la collection de travail est une conséquence de la technique d'isolement. Après culture d'enrichissement, un nombre arbitraire d'isolats ont été isolés, purifiés et cryoconservés à partir d'une gélose nutritive.

La prévalence initialement calculée a donc été corrigée en extrapolant la quantité de clones portant un plasmide, vis-à-vis de l'ensemble des souches isolées sur une campagne (Tableau 18).

Tableau 18 Proportion de clones dans la collection

Campagne	Nombre de couches criblées	Nombre de souche porteuse d'EG	Nombre de clones	Proportion de clones
AMISTAD	94	28	4	13pb%
IRIS	56	21	8	38,1%
EXTREME	16	4	0	0%
CIR	16	4	0	0%
SWEEP VENT	33	8	1	12,5%
STARMER	23	6	0	0%
Total	**239**	**71**	**13**	**18,3%**

La proportion de clones parmi les souches non porteuses d'éléments génétiques a été extrapolée à l'ensemble de la collection de travail. La valeur exacte pourrait être déterminée en réalisant le

typage de chacune des souches, mais une telle entreprise serait trop coûteuse en temps et en argent. Cependant, une estimation peut être effectuée en se basant sur la proportion de clones parmi les souches porteuses de plasmides. Cette estimation, calculée pour chaque campagne est présentée dans le Tableau 19.

Tableau 19 Abondance Plasmidique

Campagne	Nombre de souches non clonales estimé	Nombre de souche porteuse d'EG	Abondance
AMISTAD	80	24	30,0%
IRIS	35	13	37,1%
EXTREME	16	4	25,0%
CIR	16	4	25,0%
SWEEP VENT	29	7	24,1%
STARMER	23	6	26,1%
Total	**199**	**58**	**29,1%**

Finalement, il existe des souches hébergeant plusieurs éléments génétiques. C'est ainsi que 9 isolats portent 2 plasmides et 1 isolat semble même porter 3 plasmides. (Tableau 20).

Tableau 20 Nombre de plasmides par isolat

Nombre de plasmides	Nombre de souches	Pourcentage
0	141	70,9%
1	48	24,1%
2	9	4,5%
3	1	0,5%
Total	**199**	**100%**

29,1%

Les profils de restriction ont également permis d'estimer la taille des plasmides. On peut définir trois classes de tailles : les petits plasmides <5kb, les plasmides de taille moyenne 5-20kb et les gros plasmides >20kb.

Bien que cette abondance paraisse importante, par rapport à ce qui a été observé, par exemple, chez les archées thermoacidophiles de l'ordre des Sulfolobales, elle est certainement sous-estimée. En effet, la technique d'extraction utilisée ne permet pas d'extraire les plasmides linéaires ainsi que les plasmides d'une taille supérieure à 75-100kb. Bien que de tels plasmides

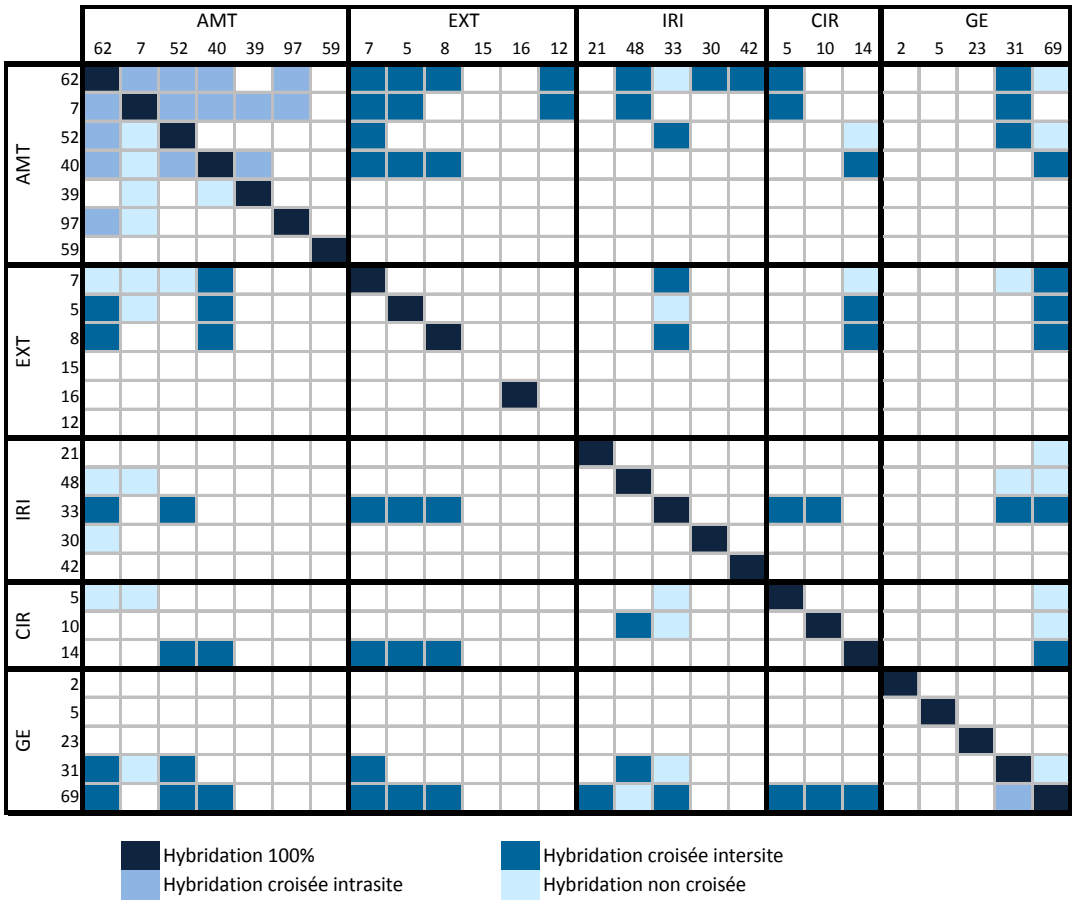
n'aient jamais été isolés chez les Crenarchaea hyperthermophiles, les Euryarchaea halophiles en possèdent.

L'appellation des plasmides décrits dans ce travail correspond au nom de la souche précédé du préfixe *p*. Par exemple, la souche numéro 9 de la campagne EXTREME possède un plasmide qui sera noté pEXT9. Dans les cas où plusieurs plasmides sont présents, une numérotation en suffixe utilise l'alphabet. La souche pEXT9 possède 2 plasmides, ce sont les plasmides pEXT9a et pEXT9b

3. Choix des plasmides à séquencer : classification en familles par hybridation

Afin d'explorer la diversité des éléments génétiques, le choix de plasmides candidats au séquençage a été dicté par des critères de dissemblance, ou au contraire, des critères de ressemblance (Tableau 21). Les plasmides présentant des similitudes sont regroupés en familles. L'obtention de plusieurs génomes issus d'une famille est nécessaire aux études de génomique comparée. Pour aboutir à cette classification, des expériences d'hybridations croisées ADN-ADN ont été réalisées. Un peu d'ADN de chaque plasmide est déposé sur une membrane en nylon. Cette membrane est ensuite hybridée avec une sonde générée à partir de l'ADN d'un plasmide à étudier (Figure 20). Cette méthode a l'avantage d'être rapide et permet de choisir avec parcimonie les candidats au séquençage afin d'accéder à la plus large diversité possible. Cette technique présente néanmoins deux inconvénients. Premièrement, des plasmides proches ne seront pas forcément regroupés par cette technique en cas de divergence de séquence trop importante. Deuxièmement, des regroupements artefactuels peuvent intervenir si un gène est très conservé entre deux plasmides de familles différentes (cas des régulateurs de transcription). Ce travail préliminaire a été réalisé, en partie, avant mon arrivée au laboratoire par Anne-Claire Mattenet (DEA Microbiologie Fondamentale et Appliquée 2002/2003). J'ai cependant vérifié la plupart des résultats obtenus avant de débuter le travail de séquençage.

Tableau 21 Résultats d'hybridation



Les ADN utilisés sont des ADNp correspondant à une souche : AMT : campagne AMISTAD ; EXT : campagne EXTREME ; IRI : campagne IRIS ; CIR : campagne Central Indian Ridge et GE : campagne STARMER. Ce tableau à double entrée utilise réciproquement l'ADNp comme sonde (lignes) et comme cible (colonne). L'observation d'un signal d'hybridation est indiquée par une couleur bleue.

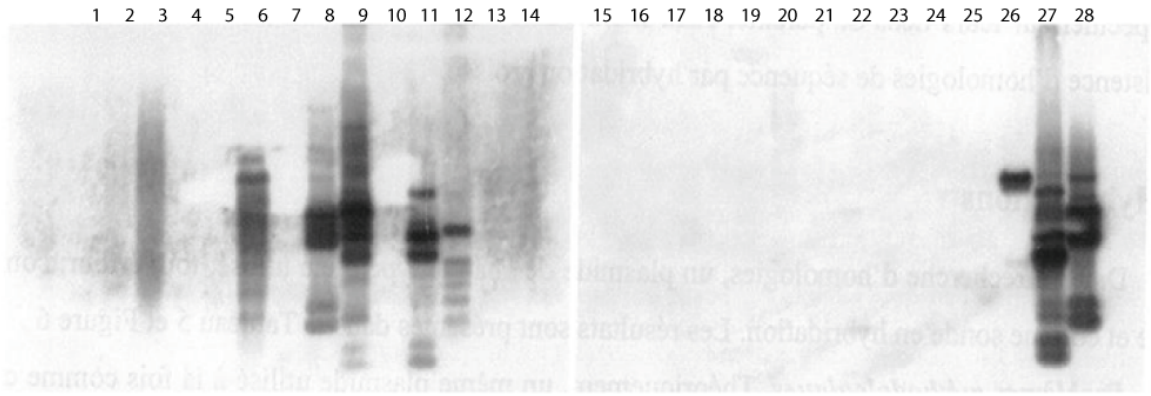


Figure 20 Hybridation avec la sonde pAMT40

1-pAMT97, 2-pAMT52, 3-pAMT62, 4-pAMT7, 5- pAMT39, 6- pAMT40 (témoin positif), 7- pAMT59, 8-pEXT7, 9-pEXT5, 10-pEXT16, 11-pEXT4, 12-pEXT8, 13-pEXT15, 14-pEXT12, 15-**Témoin négatif**, 16-pIRI48, 17-pIRI33, 18-pIRI30, 19-pCIR5, 20-pCIR10, 21-pCIR14, 22-pGE2, 23-pGT5, 24-pGE31, 25-pGE23, 26-p690, 27-pEXT5, 28-pEXT7

Problèmes méthodologiques

Théoriquement, un plasmide utilisé alternativement comme sonde et comme cible devrait toujours donner une réaction d'hybridation positive. Dans cette étude, plusieurs cas font exception. Ceci pourrait s'expliquer par une plus faible efficacité de marquage des petits fragments d'ADN (l'ADNp est digéré avant marquage) ou par la technique utilisée, adaptée pour l'hybridation d'importantes quantités d'ADN (ECL directe)

La qualité des hybridations dépend de la quantité et de la qualité de la cible et des cibles, mais aussi du nombre d'utilisation de la membrane. Après trois hybridations la membrane ne peut plus être utilisée. Il a donc été nécessaire de refaire très souvent des membranes. La reproductibilité inter-membranaires ne pouvant être contrôlée, il est impossible d'étudier les résultats d'un point de vue quantitatif ; cette étude se limitera donc à un aspect qualitatif.

Analyse des hybridations, prédiction d'homologies, choix des candidats au séquençage.

La première difficulté est l'analyse des signaux d'hybridation produits par les souches possédant plusieurs plasmides. Nous n'avons pas considéré les signaux d'hybridation obtenus en se servant de l'ADNp de ses souches comme sonde. Néanmoins, l'hybridation de certaines sondes sur l'ADN de ces plasmides permet de prédire l'appartenance de chacun des réplicon à une famille particulière.

Environ la moitié des plasmides n'ont pas produit de signal d'hybridation. Ils représentent des plasmides orphelins, supposés être les uniques représentants d'une famille au sein de notre collection. Ils peuvent également posséder une divergence de séquence suffisamment importante pour ne pas être détectés dans les conditions de stringence utilisées. Parmi ces plasmides « orphelins », nous avons choisi de séquencer les plasmides pAMT11, pGE2, pIRI42, et pEXT16.

Parmi les plasmides produisant un signal d'hybridation, la plupart hybrident entre eux, supposant l'existence d'une famille majoritaire de plasmides. Au sein de cette famille majoritaire, un représentant plasmidique de chaque océan a été sélectionné : pIRI33, pIRI48, pCIR10, pAMT7 et pEXT9 ont été séquencés.

II. Génomique comparative des plasmides de Thermococcales

1. Une famille ubiquiste de plasmides (pIRI33, pIRI48, pCIR10, pEXT9a & pAMT7)

La classification préliminaire, basée sur des hybridations Southern ADN/ADN a révélé la présence d'une famille majoritaire et ubiquiste de plasmides portés par des souches issues des différentes dorsales océaniques. Leur taille est homogène, comprise entre 8 et 13kb. Afin de caractériser cette famille dominante de plasmides, cinq souches porteuses de ce type d'éléments ont été sélectionnées. Le choix a pris en compte (i) une couverture géographique globale, augmentant potentiellement la diversité génétique en couvrant différentes zones géographiques, (ii) des plasmides provenant de souches isolées d'un même site, pIRI33 et pIRI48, (iii) la présence d'un de ces plasmides dans une souche contenant au moins un autre élément génétique (pEXT9a et pEXT9b).

1.1 *Caractérisation des souches porteuses*

Afin d'établir une éventuelle corrélation entre la présence d'un plasmide et son hôte cellulaire, les gènes codants les ARNr 16 et 23S ont été séquencés. L'analyse de ces séquences montre que toutes ces souches appartiennent au genre *Thermococcus*.

T. sp. EXT9 est issue d'un échantillon correspondant à la partie extérieure d'un fumeur noir colonisé par des vers tubicoles *Alvinella pompejana*. Il a été collecté sur la ride Pacifique Est au niveau du site pulsar. Cette souche possède deux plasmides, nommés pEXT9a et pEXT9b. Le plasmide pEXT9a appartient à la famille « ubiquiste », tandis que le plasmide pEXT9b appartient à une autre famille dont l'analyse sera présentée ultérieurement (page 180).

T. sp. AM7 a été isolé à partir d'un échantillon, correspondant à l'extrémité supérieure d'un fumeur noir, collecté sur la ride Pacifique est au niveau du site Pulsar (N12°'45'16'' W103°,59'20'') situé à 2500m de profondeur.

T. sp. IRI33 et *T.* sp. IRI48 ont été isolées à partir d'un échantillon correspondant à la partie externe d'un fumeur noir actif, prélevé au niveau du site Rainbow situé sur dorsale médio-atlantique (N36°14'06'' E33°56'30'').

T. sp CIR10 a été isolé à partir de fragments d'une cheminée collectée au niveau de la triple jonction de l'océan Indien.

La construction d'une phylogénie, basée sur les séquences d'ADNr 16S et 23S, ne montre pas de regroupement des souches porteuses de ce type de plasmide (Figure 21). Elle suggère une dissémination horizontale d'un plasmide ancestral à travers les océans. Cette absence de corrélation se confirme en analysant les génomes plasmidiques et en les confrontant à leurs origines géographiques. En particulier, les plasmides pIRI33 et pIRI48 apparaissent très différents alors qu'ils sont issus du même site.

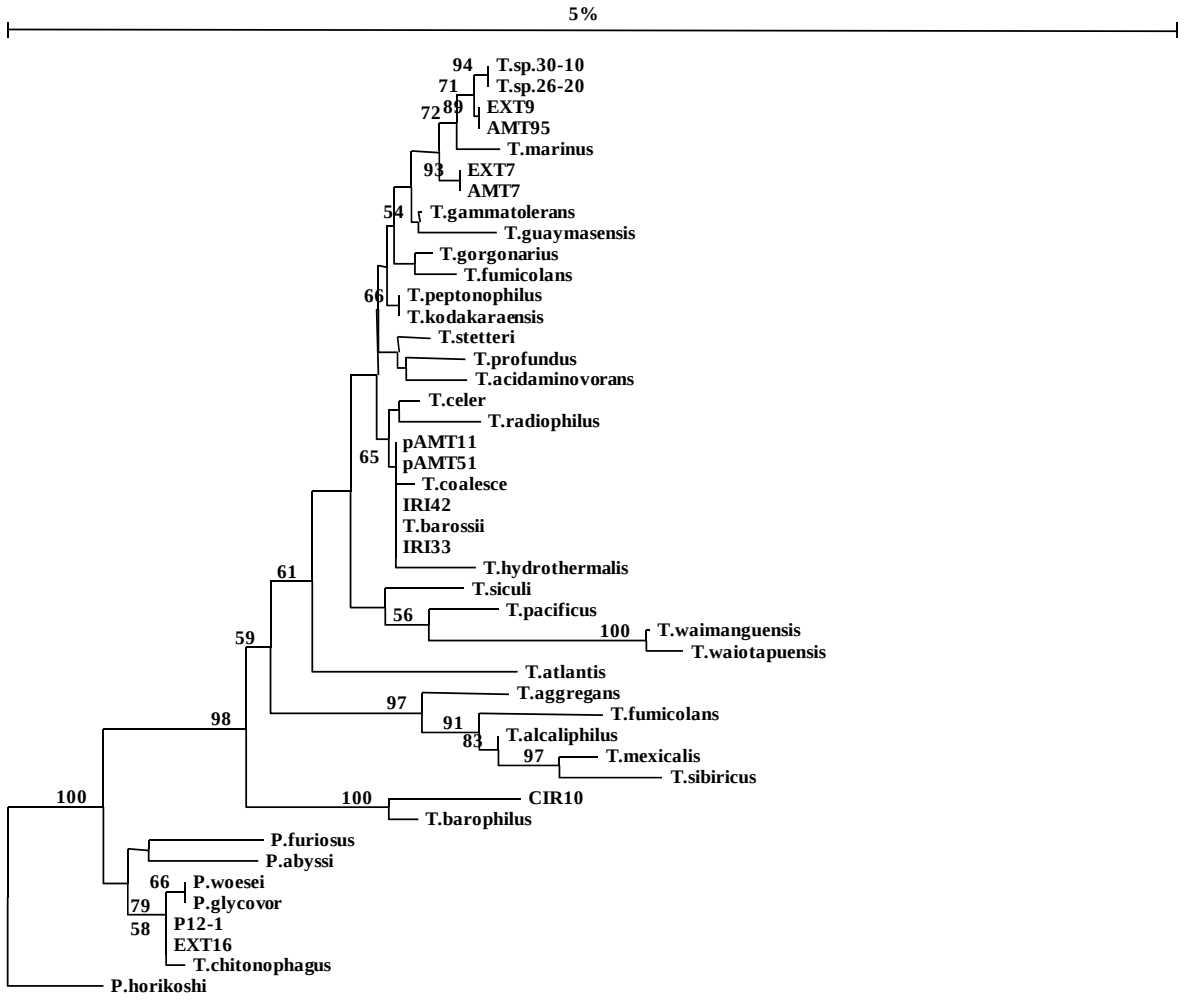


Figure 21 Phylogénie des Thermococcales basée sur le gène codant l'ADNr16S

1.2 *Propriétés générales : Organisation des génomes et caractéristiques nucléotidiques*

Les plasmides ont une taille comprise entre 8576 pb, pour pAMT7, et 13321 pb pour pCIR10. Le contenu en G+C est compris entre 44,8 et 50,0%. Il est plus faible que celui des chromosomes de *Thermococcus* et se rapproche de celui des *Pyrococcus* (40-44%). Parmi l'ensemble des ORFs prédits, entre 10 ORFs (pAMT7) et 17 ORFs (pEXT9a & pCIR10) sont des séquences potentiellement codantes, ce qui représente entre 80,2% (pAMT7) et 91,8% (pCIR10) du génome total (Tableau 22).

Tableau 22 Propriétés générales des plasmides pIRI33, pIRI48, pCIR10, pEXT9a et pAMT7.

Plasmide	Taille	Nbe ORFs	G+C%	Proportion codant	Densité génique	Origine
pIRI33	11040 pb	17	44,8	83,1%	1,53	Atlantique
pIRI48	12975 pb	14	50,0	91,0%	1,07	Atlantique
pCIR10	13321 pb	17	45,5	91,8%	1,27	Indien
pEXT9a	11040 pb	16	45,8	81,9%	1,51	Pacifique
pAMT7	8576 pb	10	45,6	80,2%	1,16	Pacifique

Cette forte proportion de séquences codantes est typique des éléments génétiques. Une corrélation peut être établie entre la taille des plasmides et leur proportion en séquences codantes. Elle permet de scinder cette famille en deux sous-groupes. Le premier, comportant pCIR10 et pIRI48, possède plus de 90% de séquences codantes, alors le second groupe, comportant pIRI33, pAMT7 et pEXT9a, n'en possède que 80%. Cette dichotomie, basée sur la proportion de séquences codantes, est également due à la densité génique.

Les ORFs sont, pour la plupart, colinéaires. Seuls 10 ORFs (13%) sont localisés sur le brin antisens, dont la moitié sont portés par le plasmide pEXT9a. Cette souche possédant un second plasmide (pEXT9b), des fréquents évènements de recombinaison entre les deux réplicons sont à envisager et seront discuté lors de l'analyse des différents ORFs (page 180).

Séquences répétées

Ces génomes sont riches en séquences répétées directes et inversées. Seules les séquences les plus répétées et les plus longues ont été prises en compte ; la taille de l'espacement entre les répétitions et sa régularité ont également été considérées afin d'interpréter une fonctionnalité biologique. Ces répétitions ne sont pas distribuées aléatoirement le long du génome, mais sont

généralement concentrées dans la plus grande région intergénique. L'illustration la plus remarquable est donnée par pCIR10, où la région 8500-9200 est particulièrement riche en séquences répétées inverses, notamment un motif de 10 pb répétés en tandem.

Diagramme de biais cumulatifs et non cumulatifs en G+C et A+T

L'analyse des biais cumulatifs en G+C et en A+T ne permet pas de visualiser les évènements de recombinaison/inversion les plus ostensibles ; elle ne permet pas non plus de prédire avec certitude une origine de réplication, laquelle se traduisant généralement par une inversion de polarité localisée par un tracé en V, où le minimum coïncide avec l'emplacement de l'*ori* et le maximum au point de terminaison de la réplication. On remarque cependant la présence d'une zone d'inflexion plus ou moins prononcée qui pourrait correspondre à l'origine de réplication.

On remarque également l'existence de deux types de profils ; celui des plasmides pIRI48 et pCIR10 possède un diagramme presque linéaire tandis que celui les plasmides pIRI33, pAMT7 et pEXT9a montre deux points d'inflexion plus prononcés. Ces inflexions coïncident également avec les régions précédemment notées comme riches en séquences répétées. Ainsi, les répétitions en tandem font penser aux itérons qui servent de sites de reconnaissance pour la protéine initiatrice de la réplication de certains plasmides.

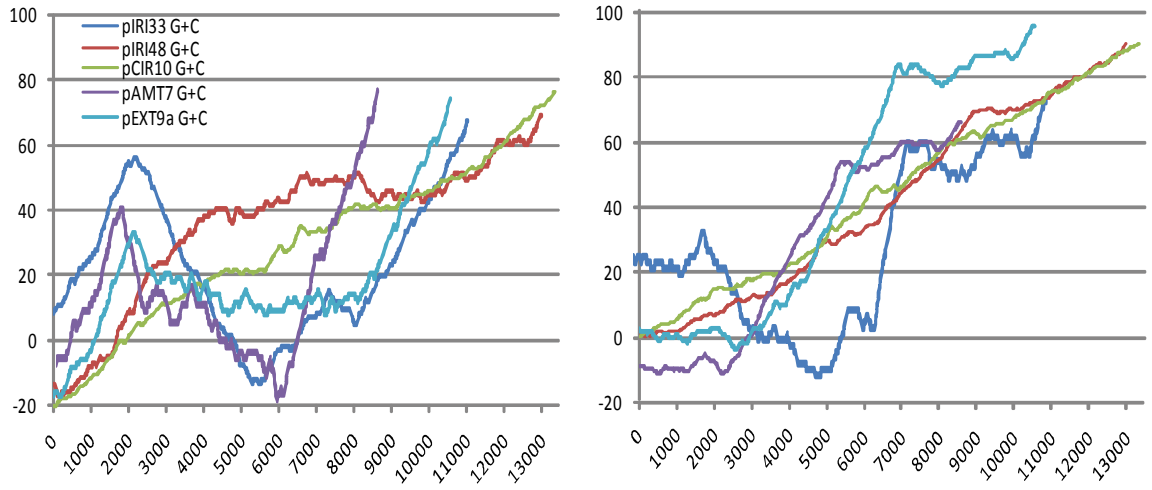


Figure 22 Biais en nucléotides des plasmides pIRI33, pIRI48, pCIR10, pAMT7 et pEXT9a

Représentation graphique du biais cumulatif, en abscisse la position sur le génome de chaque plasmide et en ordonnées les valeurs cumulées de biais. Le graphique de gauche représente le biais en G+C tandis que le graphique de droite représente le biais cumulatif en A+T.

1.3 *Contenu en gènes*

L'analyse comparée des génomes permet de distinguer trois régions dans ces génomes. La première est constituée de gènes strictement conservés et synténiques, la seconde est constituée de gènes spécifiques du sous-groupe de plasmides pIRI48 et pCIR10, et la troisième correspond à des gènes orphelins. L'analyse phylogénétique des gènes conservés confirme la distinction de deux sous-groupes au sein de cette famille. Cette distinction n'est pas le résultat d'un isolement géographique car deux plasmides issus du même site, pIRI33 et pIRI48, appartiennent à des sous-groupes différents.

Bien qu'un unique génome de virus de Thermomcoccales soit disponible dans les bases de données, certaines protéines sont homologues à celles rencontrées sur le virus PAV1 de *P. abyssi* GE23.

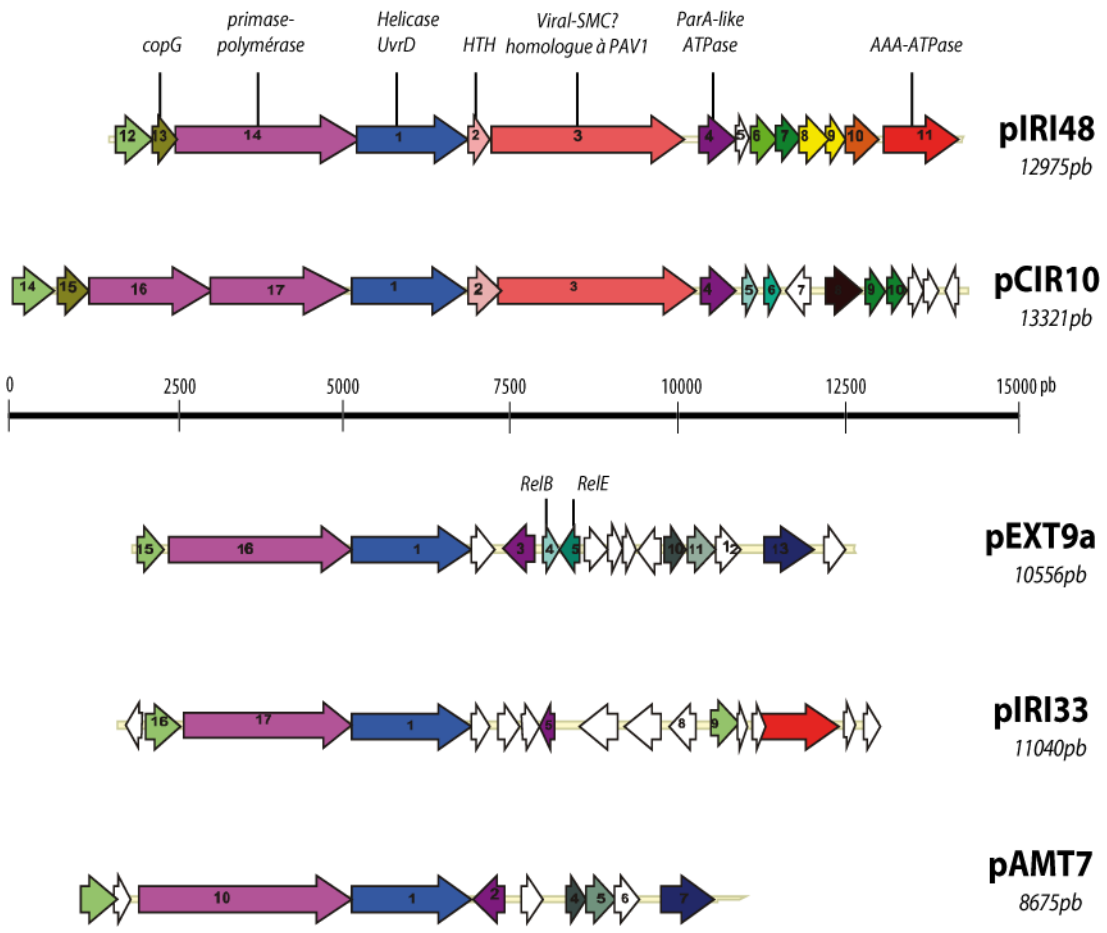


Figure 23 Génomique comparée des plasmides pIRI48, pCIR10, pEXT9a, pIRI33, pAMT7 et PAV1

Les ORFs sont représentés par des flèches. Chaque couleur représente des ORFs codant des protéines homologues.

1.3.1 Les gènes strictement conservés

Les gènes strictement conservés constituent le squelette de cette famille de plasmides. Quatre ORFs sont systématiquement rencontrés sous la forme d'un opéron synténique (

Tableau 23). Le premier gène (A) code un régulateur de transcription de la famille des CopG, le second (B) une hypothétique ADN polymérase-primase, le troisième (C) une hélicase de la famille Rep/UvrD, et le dernier (D) possède uniquement des similarités avec une protéine d'un îlot génomique de *Methanococcus voltae* A3. La principale fonction d'un réplicon extrachromosomique étant sa capacité à se perpétuer, il n'est pas surprenant de rencontrer des gènes impliqués dans la réplication.

Tableau 23 ORFs strictement conservés entre les plasmides pIRI48, pCIR10, pEXT9a, pIRI33 et pAMT7

	ORF	Taille (AA)	pI
A. CopG			
pIRI48	12	180	5,7
pCIR10	14	202	4,8
pEXT9a	15	132	5,6
pIRI33	16	175	4,8
pAMT7	?		
B. "Grand ORF"			
pIRI48	14	755	6,24
pCIR10	17	451	5,53
pEXT9a	16	675	5,8
pIRI33	17	671	5,9
pAMT7	10	673	5,7

	ORF	Taille (AA)	pI
C. Hélicase UvrD			
pIRI48	1	587	9,1
pCIR10	1	568	8,6
pEXT9a	1	591	8,4
pIRI33	1	591	8,8
pAMT7	1	591	8,8
D. DNA binding			
pIRI48	4	177	6,3
pCIR10	4	173	6,9
pEXT9a	3	164	8,7
pIRI33	5	85	9,4
pAMT7	2	155	9,1

Ces gènes sont strictement conservés sur les plasmides de la famille et sont organisés de façon synténique. Toutefois, deux organisations sont observées confirmant la distinction en sous-groupes : pCIR10/pIRI48 et pIRI33/pAMT7/pEXT9a (Figure 24). Ces gènes sont physiquement associés par la présence de chevauchements de séquences. L'analyse des transcrits a montré qu'ils étaient cotranscrits (Communication personnelle Gael Erauso). Cet opéron est également cerné par les deux inflexions du biais cumulatif en nucléotides (Figure 22).

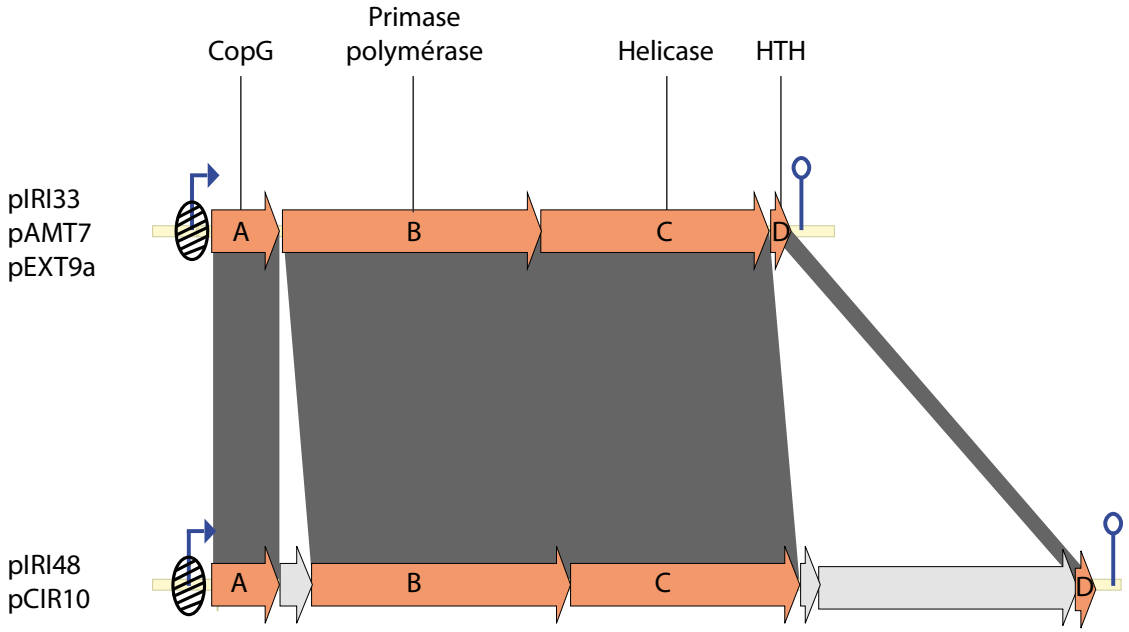


Figure 24 Organisation de l'opéron contenant les gènes conservés de pIRI48, pCIR10, pAMT7, pIRI33 et pAMT7

Les flèches oranges représentent les ORFs strictement conservés, les ORFs gris représentent les gènes spécifiques de la sous-famille pIRI48/pCIR10 ; L'ellipse hachurée représente une importante concentration en séquences répétées ; la flèche brisée bleue représente un promoteur et la sucette bleue un terminateur de transcription.

Le premier gène conservé (Tableau 23.A) est situé à l'extrémité 5' de l'opéron. Il code une protéine similaire à TK1815 (80AA), un répresseur de transcription de *T. kodakaraensis* KOD1. Comparativement, les protéines plasmidiques ont une taille plus importante (172±29AA) due à la présence de deux domaines. Le premier possède la signature RHH (Ruban-Hélice-Hélice) de la famille CopG (environ 50AA) ; il explique l'accroche avec la « petite » protéine du génome de *T. kodakaraensis* qui possède uniquement ce domaine. Ce gène, fréquent sur les éléments génétiques, est généralement situé en tête de l'opéron impliqué dans la réplication et assure la régulation du nombre de copies de l'élément génétique. Sans régulation, un trop grand nombre de copies du plasmide serait synthétisées, ce qui réduirait les ressources nécessaires aux autres processus cellulaires, la compétitivité de la cellule et donc la survie du plasmide. Ces protéines sont les plus petits régulateurs de transcription découverts (45AA). La première hélice du motif RHH permet la fixation à l'ADN, tandis que la seconde permet la dimérisation de la protéine. La chiralité du dimère induit la reconnaissance de répétitions inverses plus ou moins espacées, généralement localisées à proximité d'un promoteur, en amont du gène *copG*. La liaison sur ce type de site empêche la fixation de l'ARN polymérase et réprime ainsi l'opéron réplicatif. A l'exception de pAMT7, les protéines codées par les plasmides sont fusionnées à un second domaine très dégénéré. La bibliographie (Wales *et al.,* 2004) précise que les domaines CopG sont

souvent associés à un second domaine qui faciliterait le repliement RHH, et qui interviendrait également dans l'encombrement stérique contribuant à la répression la transcription. L'hypothèse d'un fonctionnement similaire à celui de CopG est renforcée par la présence, sur chacun des plasmides, d'un promoteur en aval de *copG* entouré de nombreuses séquences répétées pouvant servir de site de fixation à CopG (Figure 24).

Le second gène (Tableau 23.B) code les plus grosses protéines de ces plasmides (693±38AA). Ces protéines sont acides, pI=5,9±0,2. Cet ORF possède la caractéristique d'être scindé en deux ORFs chevauchant sur pCIR10. Ce chevauchement est constitué du tétranucléotide ATGA servant de codon stop et de codon start, respectivement pour le premier et le second ORF.

Alors que les protéines de pEXT9a, pIRI33 et pAMT7 sont très conservées (Id= 73%; similarité=82%), celles de pCIR10 et pIRI48 sont plus divergentes. Malgré des similarités très élevées entre protéines codées par ces plasmides, aucun homologue n'est détecté dans les bases de données. La détection de motif n'est guère plus fructueuse de prime abord. De nombreux motifs communs ou impliquant des motifs de faible complexité sont détectés et prédisent une activité ATPase. L'utilisation de l'algorithme SMART permet de détecter les signatures des ADN primases et les ADN polymérase B. Bien que des confirmations biochimiques soient nécessaires, il ne serait pas aberrant de trouver une ADN primase-polymérase exprimée en opéron avec une hélicase. Les trois fonctions indispensables à une réplication autonome pourraient donc être codées par cette famille de plasmide à l'intérieur du pool de gènes communs. De plus, l'analyse phylogénétique de ces protéines est congruente à celle de l'hélicase située en aval de gène. Cette observation confirme la relation évolutive entre l'hélicase et le gène codant cette ADN polymérase-primase putative permettant la distinction en deux sous-groupes de plasmides ne pouvant s'expliquer par un isolement géographique.

Le troisième gène (Tableau 23.C) code des protéines de tailles très homogènes (590±10AA) et de nature basique (pI=8,76±0,26). La recherche d'homologues et de domaines assignent cette protéine aux hélicases de la superfamille I, et plus particulièrement à la famille des hélicases Rep/UvrD. Les motifs caractérisant ces hélicases sont détectés sur la séquence peptidique (Figure 25). Il s'agit d'un motif de fixation à l'ADN (V), des motifs ATPases Walker A et Walker B (I et II), du motif hélicase (IV) et d'un motif permettant le couplage entre l'hydrolyse d'ATP et le débobinage de la double hélice d'ADN. L'analyse visuelle des motifs montre une plus grande divergende des motifs de pCIR10 et pIRI48.

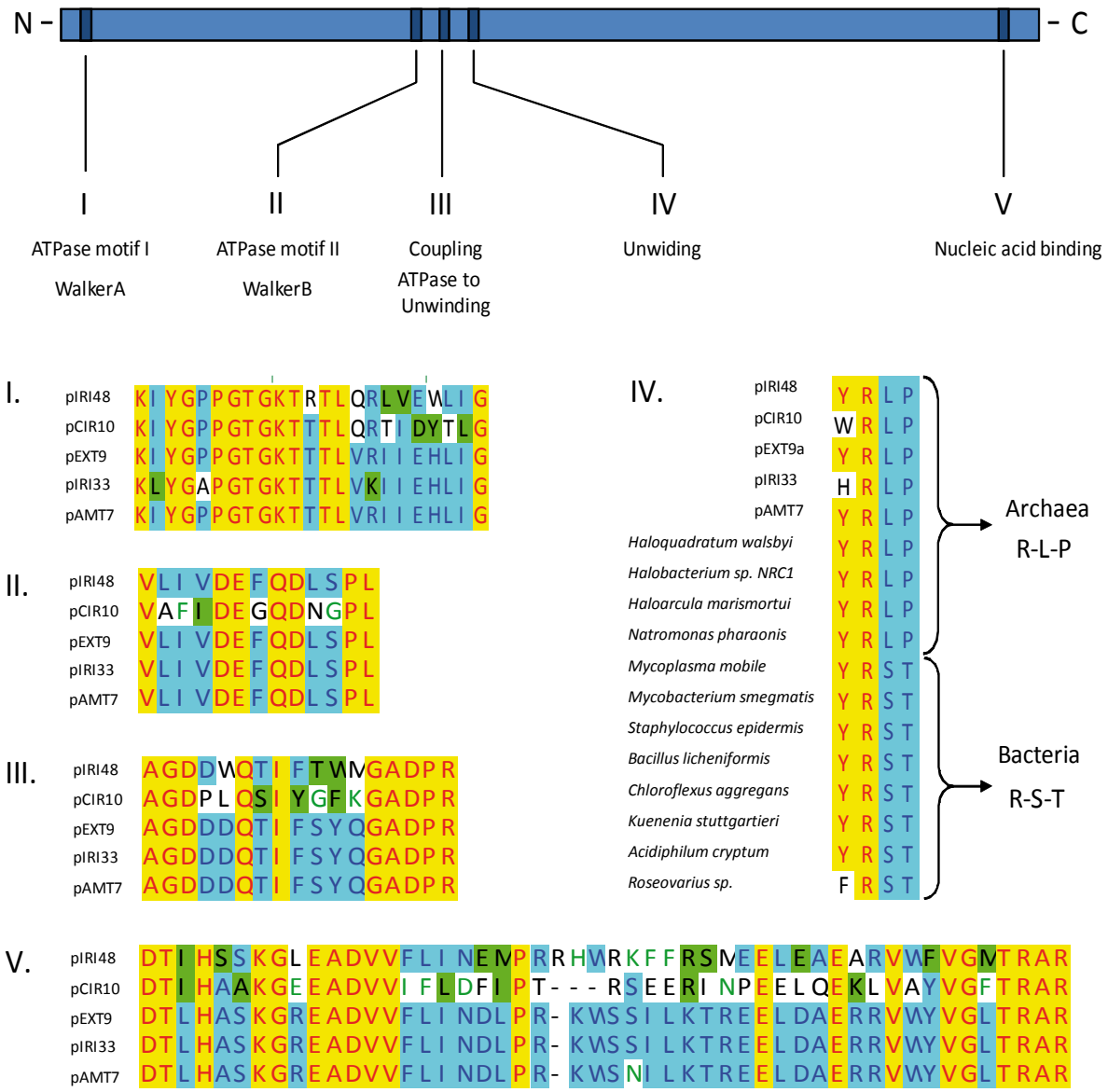


Figure 25 Analyse de la séquence protéique des hélicases UvrD codées par les plasmides pIRI48, pCIR10, pEXT9a, pIRI33 et pAMT7.

Représentation schématique des hélicases UvrD de la famille ubiquiste de plasmides avec la localisation des cinq motifs fonctionnels (chiffres romain). L'alignement de ces motifs est présentée en dessous du schéma de la protéine. La couleur des AA traduit les similarités (rouge sur fond jaune : identité stricte, bleue : identité > 80%, noir sur fond vert : AA similaire)

UvrD et Rep sont deux hélicases à ADN de bactéries à Gram négatif, partageant 40% d'identité au niveau des séquences aminoacides. Elles possèdent également des similarités avec les hélicases PcrA des bactéries à Gram positif. Ces hélicases permettent la séparation des deux brins d'ADN selon un mécanisme de déroulement catalysé par l'hydrolyse d'ADN. Les hélicases sont indispensables à la plupart des réactions de maintenance de l'ADN : réplication, réparation,

recombinaison et transcription. Cette famille d'hélicases est assez bien caractérisée *in vitro*, tant au point de vue biochimique que structural. Elles sont composées de quatre domaines structuraux et sont actives sous forme de dimères. Néanmoins, l'activité *in vivo* n'est que partiellement élucidée. Bien que la dimérisation implique majoritairement des homodimères Rep/Rep ou UvrD/UvrD, la présence d'hétérodimères Rep/UvrD multiplie les fonctionnalités de ces protéines (Wong *et al.,* 1993). Certaines activités sont correctement définies pour UvrD, PcrA et Rep : (i) Les UvrD participent à la réplication de nombreux éléments extrachromosomiques, tels que les bactériophages M13 et pX174 (Chao *et al.,* 1991); (ii) les UvrD assurent la réplication de certains plasmides à cercle roulant de bactéries à Gram négatif tandis que les PcrA interviennent, chez les bactéries à Gram positif, aussi bien dans la RCR de leurs plasmides que dans la réparation de l'ADN (Petit *et al.,* 1998). (iii) Suite à un stress oxydant (UV, radicaux oxygénés), UvrD est d'une part capable d'induire une stase cellulaire par induction du système SOS (Crowley *et al.,* 2001) puis d'interagir avec le complexe protéique UvrABC afin de permettre l'excision de l'ADN endommagé, aussi bien chez les bactéries (Ahn 2000) que les *Archaea* halophiles (Ahn 2000; McCready *et al.,* 2003). (iv) certaines versions « hybrides » entre UvrD et Rep interviennent aussi bien dans la réplication par cercle-roulant que dans des systèmes de réparation de mésappariement de bases, où elles favorisent l'ablation du segment d'ADN contenant le nucléotide erroné par le complexe MutSLH (Petit *et al.* 1998). La découverte de nouveaux homologues au sein des *Archaea* permet de récolter de précieuses informations sur les caractères partagés et donc sur le fonctionnement de ces protéines.

L'analyse phylogénétique montre une séparation des hélicases bactériennes et archéennes (Figure 26). Bien que formant un groupe distinct, la phylogénie des hélicases archéennes présente une dichotomie. Le premier clade comporte les hélicases d'halophiles, partageant de forts pourcentages d'identité, tandis que le second comporte celles des plasmides de Thermococcales. Ce dernier groupe ne semble pas monophylétique. La présence de branches relativement longues traduit une division en deux sous-groupes. Les hélicases de pEXT9a, pIRI33 et pAMT7 sont très conservées et partagent un pourcentage d'identité global d'environ 90%. L'ajout de l'hélicase de pCIR10 diminue cette valeur à 32,1%, et l'ajout de celle de pIRI48 fait chuter le pourcentage d'identité à 14,5%. Les plasmides isolés de souches provenant d'un même échantillon (pIRI33 et pIRI48) produisent l'alignement présentant le plus faible pourcentage d'identité. Ceci conforte l'existence de deux sous-groupes, ce qui avait été déjà supposé lors de l'analyse du biais cumulatif en nucléotides et de la synténie (Figure 22 et Figure 24).

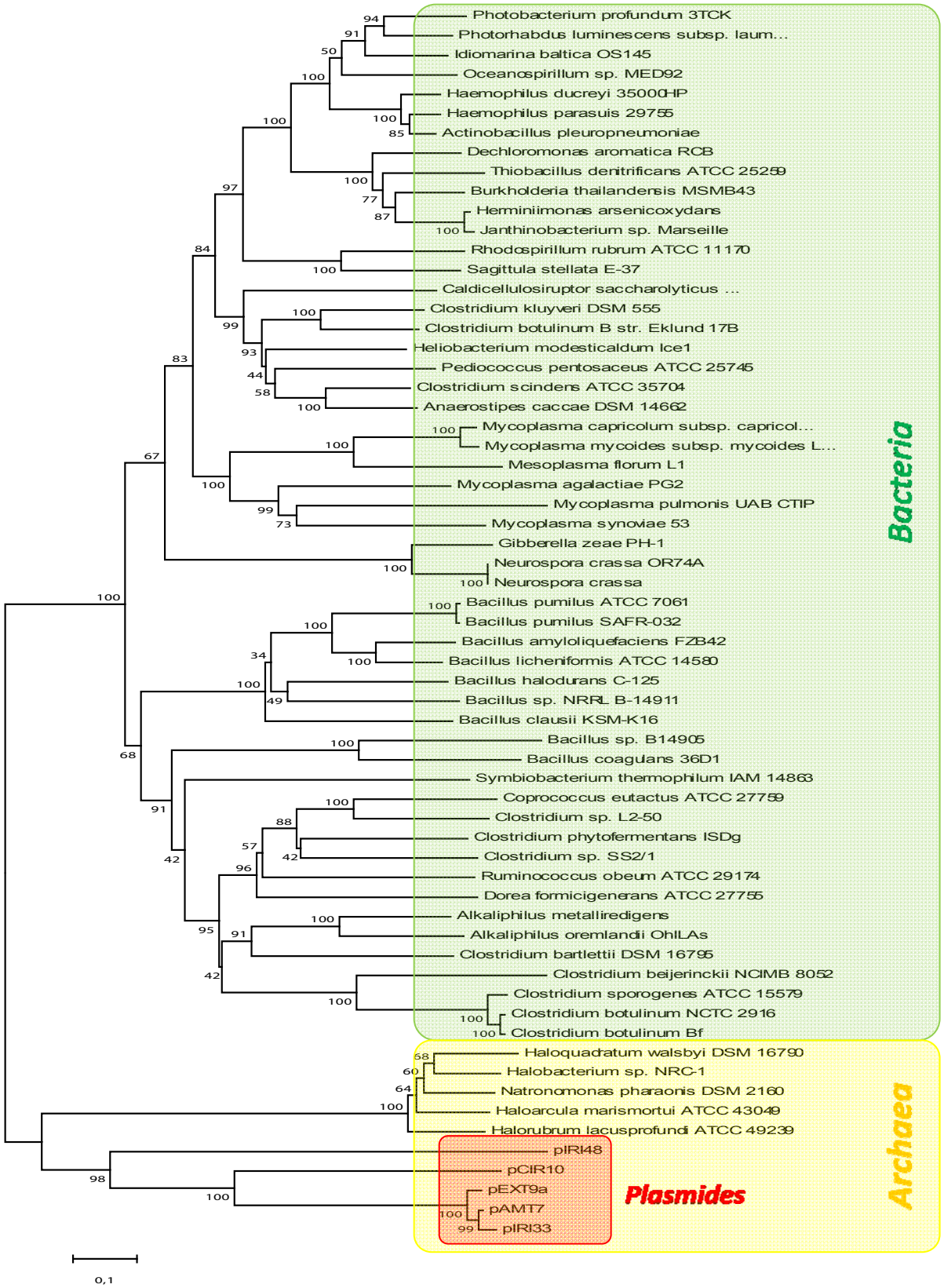


Figure 26 Phylogénie des hélicases UvrD

L'éloignement phylogénétique de la protéine codée par pIRI48 pourrait s'expliquer par l'absence d'une séquence spécifique portée par les autres plasmides. Cette séquence, appelée PIP-box, intervient dans le recrutement d'un partenaire protéique essentiel à la réplication : le facteur de processivité ou PCNA (*Proliferating Cell Nuclear Antigen*). Le recrutement du PCNA, codé par le chromosome permettrait la cohésion entre l'hélicase codée par le plasmide et l'ADN polymérase nécessaire à la réplication, augmentant ainsi la processivité.

Le dernier ORF conservé (Tableau 23.D) code des protéines de taille relativement homogène, 167±9AA. La prédiction de topologie de la protéine indique la présence de deux domaines. Le domaine N-terminal, absent de la protéine codée par pIRI33, est un leucine-zipper qui est généralement impliqué dans la dimérisation de nombreux facteurs de transcription. Le domaine C-terminal adopte un repliement de type HTH (Hélice-Tour-Hélice). L'exploration des bases de données révèle la présence de protéines possédant des similarités de séquences. Elles sont codées par des gènes présents sur le chromosome de différentes *Archaea* méthanogènes, l'espèce thermophile *Methanocaldococcus jannaschii (*MJ1503), et les espèces mésophiles : *Methanococcus maripaludis* C6 (MmarC6_116 et MmarC6_452), *M. maripaludis* C5 (MmarC5_1848 et MmarC5_1132) et, finalement, par le plasmide pURB500 (ORF3) également porté par cette dernière souche. Toutes ces protéines sont annotées comme étant des NTPases membranaires bien qu'aucune ne possède un peptide signal nécessaire à ladressage vers la membrane. Le second domaine, comportant les 60 AA de l'extrémité C-terminale n'a pas de fonction assignable. Néanmoins, l'alignement des séquences permet la définition d'un nouveau motif de fonction inconnue, proche des leucine-zipper (Figure 27).

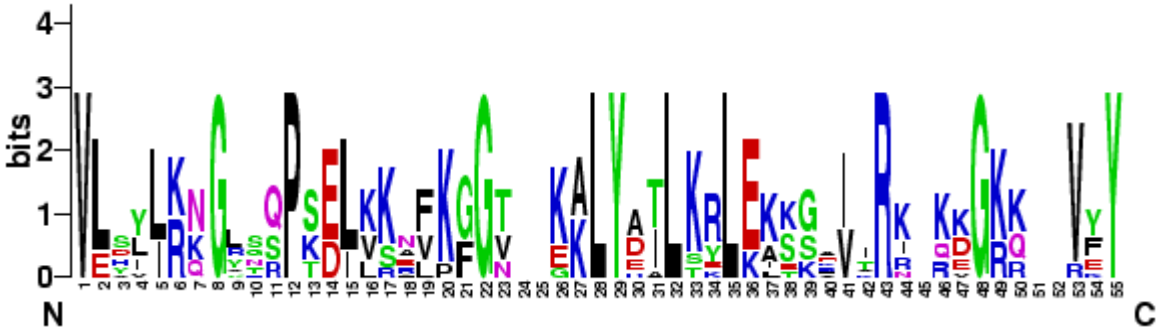


Figure 27 Représentation web logo

Alignement des 55 derniers AA des protéines codées par les ORFs de plasmides de Thermococcales ainsi que des homologues présents dans les génomes d'*Archaea* méthanogènes.

1.3.2 Les gènes spécifiques du sous groupe pCIR10 – pIRI48

L'organisation des génomes des plasmides et la phylogénie des gènes conservés ont montrés l'existence de deux sous-groupes au sein de cette famille de plasmides. Alors qu'aucun gène ne semble spécifique au sous-groupe pIRI33, pAMT7 et pEXT9a, quatre gènes sont spécifiques du sous-groupe pIRI48 et pCIR10. Trois sont localisés dans le même opéron que les gènes strictement conservés. Ces protéines présentent des similarités avec certains îlots génomiques ainsi qu'avec le virus PAV1 de *Pyrococcus abyssi* GE23, et sont principalement associées à la maintenance de l'ADN. Le quatrième gène de ce sous-groupe est situé en dehors de cet opéron.

Les protéines codées par l'ORF2 de pCIR10 (169AA) et par l'ORF2 de pIRI48 (110AA) sont parmi les protéines les plus basiques de ces plasmides (pI=9,8). L'alignement de séquence sur la partie commune de ces protéines produit un pourcentage d'identité de 89%. Un unique homologue est détecté, il s'agit d'une protéine orpheline portée par le plasmide cryptique pURB800 de *Methanocaldococcus jannaschii*. Le contexte génomique de cette protéine homologue, entourée par des ATPases de ségrégation, suggère son implication dans un mécanisme de partition ou de compaction de l'ADN.

Les ORFs 3 de pIRI48 et pCIR10, conservées au sein de ce sous-groupe, codent les plus grandes protéines de ces plasmides (980AA) et sont légèrement acides (pI=6). Les séquences de ces protéines partagent 76% d'identité. Un seul homologue est présent dans les bases de données, codé par l'ORF898 de PAV1, le virus infectant *P. abyssi* GE23. L'ajout de cette séquence orpheline dans l'alignement fait chuter l'identité de séquence à 31%. La prédiction de topologie de la protéine prédit un découpage en trois domaines séparés par des boucles localisées à la position des insertions/délétions sur l'alignement. L'extrémité C-terminale est basique et possède des répétitions de lysine rappelant les domaines « *coiled-coil* ». Ces extrémités riches en lysine sont bien décrites sur les protéines de condensation de l'ADN, histone eucaryote et histone-like des Archaea ; les lysines sont la cible de méthylation permettant de réguler l'affinité de la protéine pour l'ADN. La recherche de similarité itérative détecte diverses protéines SMC *Structural Maintenance Chromosome* (psi-blast evalue = 10e-85), la première d'entre-elles est codée par *Thermococcus kodakaraensis* (880AA). Les protéines SMC sont également appelées protéines motrices ; elles sont indispensables à la ségrégation et à la transmission des chromosomes lors de la division cellulaire. Typiquement, ce sont de grosses protéines qui ont une activité ATPasique (motifs Walker A et Walker B) et possèdent deux régions « *coiled-coiled »* repliées sur elles-mêmes. Les protéines SMC interviennent en coordination avec d'autres protéines dans une

variété d'opérations sur les chromosomes incluant la condensation, la cohésion entre les chromatides sœurs (eucaryotes), la recombinaison et la réparation de l'ADN. Bien qu'un domaine de type *coiled-coil* et un motif Walker B soit détecté (séquence ELDEFLQDLRE), seule une faible ressemblance avec le motif Walker A peut-être trouvée (séquence GLNVGKFT). Ces données renforcent l'hypothèse d'une implication dans la condensation de l'ADN, qui chez les virus, comme PAV1, est une étape indispensable à l'assemblage de particules virales.

Les ORFs 13 de pIRI48 et 15 de pCIR10 sont entourés par deux gènes strictement conservés au sein de la grande famille de plasmides, CopG et la primase-polymérase. Ces protéines sont de petite taille (165AA), partagent 32,5% d'identité mais ont des pI assez divergents (5,77 et 8,89). Aucune information n'est collectée lors de la recherche d'homologues dans les bases de données. Néanmoins, ces protéines ont en commun la présence d'un peptide signal et une forte proportion d'acides aminés polaires.

1.3.3 Les gènes accessoires

Un gène est dit accessoire lorsque qu'il est présent sur au moins deux plasmides appartenant aux deux sous-groupes précédemment définis.

Parmi ces gènes accessoires, il est possible de trouver des gènes agissant sous la forme de module. Les ORF5 et 6 de pCIR10 et les ORF4 et 5 de pEXT9a codent un système toxine-antitoxine RelBE, conférant l'addiction au plasmide (Page 18). Ce type de module est fréquemment rencontré sur les éléments génétiques ; il évite la perte du réplicon au sein de la population en inhibant la croissance des cellules ayant perdu cet élément. Lorsque ce type de module est présent sur un chromosome, sa fonction reste assez floue. Il peut, à la manière des éléments génétiques, augmenter la maintenance chromosomique afin d'éviter la perte de morceaux du chromosome. Il peut également agir selon un mécanisme de défense vis-à-vis des éléments génétiques, l'antitoxine chromosomique neutralisant la toxine codée par l'élément génétique. La dernière fonction découverte concerne la surexpression de ces gènes lors d'un stress ionisant chez *P. furiosus* (Williams *et al.* 2007). L'effet bactériostatique de la toxine permet à la cellule d'avoir le temps de réparer son ADN avant d'entreprendre une nouvelle division. Dans notre cas, il s'agit probablement d'un mécanisme augmentant la maintenance du plasmide.

Parmi les gènes accessoires, nous pouvons également noter la présence de quatre ORFs homologues adjacents sur le génome de pAMT7 (ORFs 4 à 7) et sur le génome de pEXT9a (ORFs 9, 10, 11 et 13) ; Ces ORFs codent des protéines aux propriétés physico-chimiques comparables (Tableau 24).

Tableau 24 Caractéristiques des ORFs conservés entre pEXT9a et pAMT7

pAMT7			pEXT9a			% identité
ORF	Taille	pI	ORF	Taille	pI	
4	107	8,5	10	107	9,7	76,6
5	141	4,5	11	141	4,6	84,4
6	119	9,3	12	122	9,7	78,7
7	261	9	13	246	6,2	29,2

Ces quatre gènes sont organisés en deux opérons (Figure 28). En effet, l'analyse des signaux de transcription prédit un promoteur à l'extrémité 5' et un terminateur à l'extrémité 3'. Le premier opéron comporte les trois premiers ORFs ; le dernier ORF, pour sa part, est transcrit de manière indépendante.

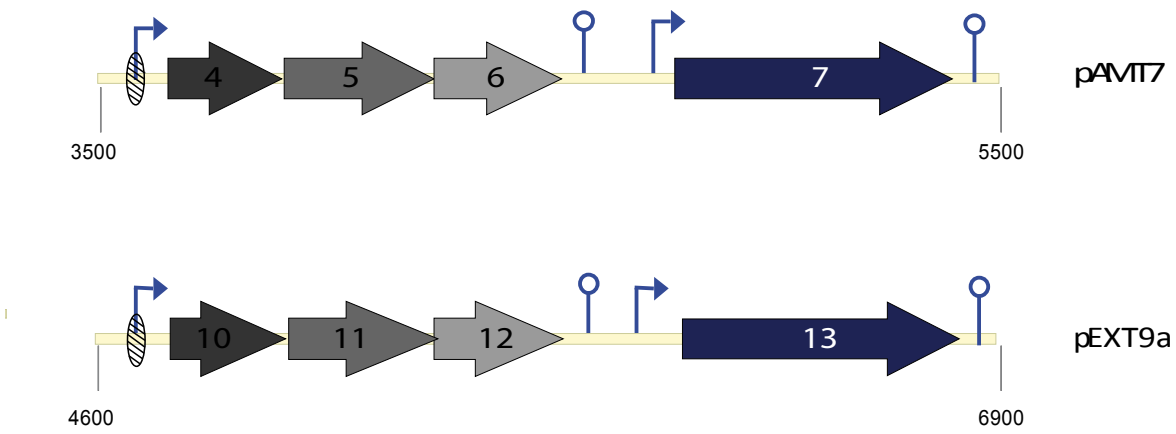


Figure 28 Organisation des ORFs conservés entre pEXT9a et pAMT7

Représentation graphique de la région 3500-5500 de pAMT7 et 4600-6900 de pEXT9a. Les flèches représentent les ORFs. Les promoteurs et terminateurs sont respectivement indiquée par des flèches brisées et des sucettes. L'ellipse hachurée représente une région contenant des répétitions inversées.

Seul le premier gène de l'opéron code une protéine à la fonction assignable ; il s'agit d'un régulateur de transcription de la famille RHH. Des séquences répétées inversées sont localisées,

de part et d'autre du promoteur. Elles pourraient servir de site de fixation à ce régulateur et, ainsi, réprimer la transcription de cet opéron en créant un encombrement stérique au niveau du promoteur. Le dernier gène, codé par l'ORF 7 de pAMT7 et l'ORF 13 de pEXT9a, possède seulement de faibles similarités de séquence avec les motifs A et B du domaine HMG2 (*High Mobility Group*). Ce type de domaine permet la fixation non spécifique à l'ADN et favorise la formation de structures secondaires dites distordues, jonctions à quatre brins et renflements d'ADN (*DNA bulges*). On les retrouve au niveau de facteurs de transcription mais aussi dans certains mécanismes de réparation de l'ADN au cours desquels ils se fixent sur les bases désappariées afin de faciliter leur excision.

1.3.4 Les gènes « uniques »

Cette catégorie regroupe les gènes rencontrés sur un plasmide. La plupart code des protéines orphelines. Cependant, certaines caractéristiques singulières nécessitent d'être commentées.

L'ORF13 de pIRI48 code une protéine de 368AA. Quatre homologues sont trouvés dans les bases de données. Ils sont codés par l'ORF375 de PAV1, par l'ORF1394 de *Methanococcus voltae* A3, par pNG4027 du plasmide pNG400 de *Haloarcula marismortui* et par l'ORF5 du virus AMDV3 (32% identité, 51% similarité) isolé d'un drainage de mine acide (Andersson *et al.,* 2008). Bien que ces protéines ne soient pas annotées, la présence des motifs Walker A et Walker B indiquent une fonction ATPase. Une analyse plus fine classe cette protéine dans la famille des AAA+ ATPases. Malgré la fréquence de ce type de domaine, il est surprenant de ne pas avoir détecté plus de protéines possédant des similarités dans les bases de données. La présence de trois hélices transmembranaires et d'un peptide signal prédit une localisation membranaire de cette protéine. Bien qu'aucune fonction ne puisse être prédite sur l'utilisation de l'énergie libérée lors de l'hydrolyse de l'ATP, cette ATPase membranaire possède une parenté avec celles des virus.

L'ORF7 de pIRI33 possède un domaine PIN, impliqué dans la fixation aux acides nucléiques. A l'inverse des autres motifs courants de fixation à l'ADN (RHH, HTH…), ce motif est souvent présent sur des protéines impliquées dans le métabolisme des acides nucléiques. Il est également rencontré sur les toxines des systèmes d'addiction. Dans ce cas, il possède une activité exonucléasique 5' -> 3' dégradant les acides nucléiques et conférant l'activité létale. Ce type de domaine assez fréquent permet la détection d'homologues chez les *Archaea* et les *Bacteria*. Certains génomes possèdent plusieurs homologues de cette protéine. Les homologues archéens

sont uniquement rencontrés chez les Euryarchaea hyperthermophiles ; aucun homologue n'est trouvé chez les Crenarchaea, ni chez les Euryarchaea mésophiles. De plus, la phylogénie de cette protéine n'est pas congruente à celle de l'ADNr16S (Figure 29).

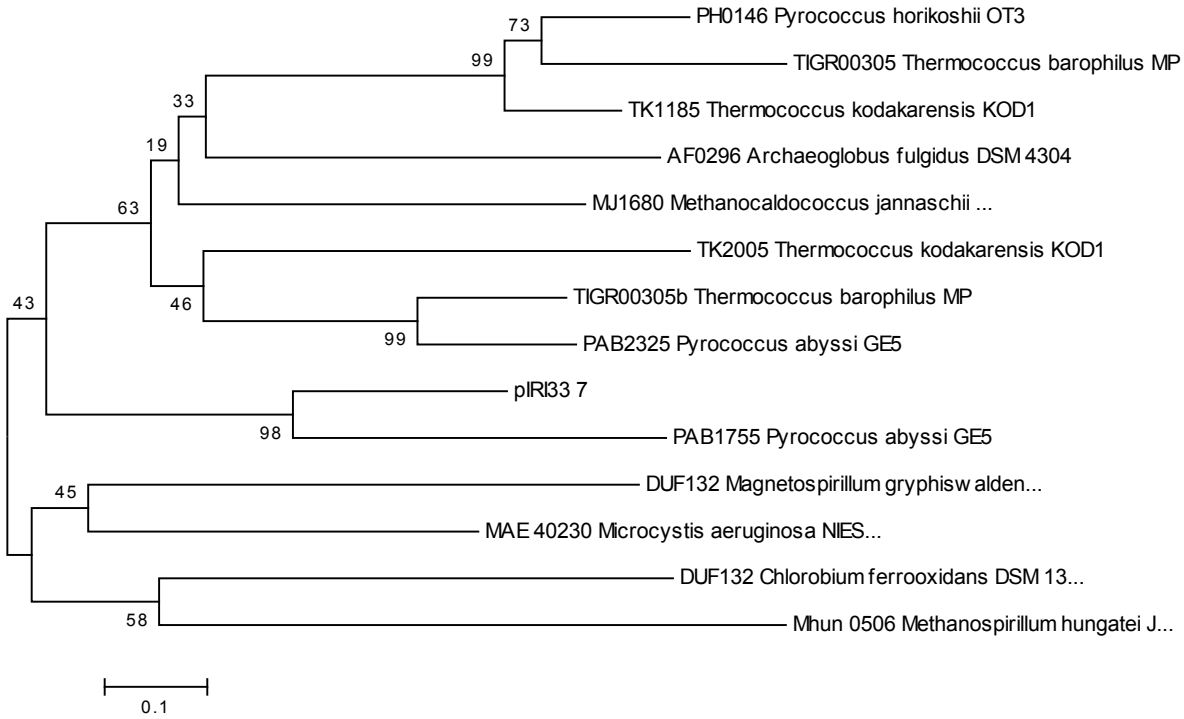


Figure 29 Phylogénie de l'ORF7 de pIRI33

La présence d'homologues intraspécifiques, non regroupés sur des branches communes, laisse présager une lointaine duplication avec perte de l'un des homologues dupliqués, ou bien un transfert horizontal. L'hypothèse de mobilité est confortée par la présence de cette protéine sur le plasmide pIRI33.

Parmi les gènes uniques de cette grande famille, beaucoup sont orphelins au sein des bases de données. Néanmoins, 8 gènes possèdent des homologues portés par d'autres plasmides de Thermococcales séquencés pendant ces travaux de thèse. Ce grand nombre d'accroches suppose l'existence de gènes affiliés aux éléments génétiques mobiles pouvant se déplacer d'un réplicon à un autre, comme c'est le cas par exemple pour les protéines de type CopG. Ils illustrent la fluidité de ces génomes et la fréquence des recombinaisons lorsque deux éléments cohabitent au sein d'une même souche.

Au sein de la sous-famille pIRI33, pEXT9a, pAMT7, le plasmide pIRI33 est celui qui est le moins conservé ; il ne possède pas l'opéron de gènes partagé entre pEXT9a et pAMT7 (Figure 28). Néanmoins, à la position occupée par ces gènes, on retrouve des gènes homologues présents dans le génome de pEXT9b, le second plasmide hébergé par la souche *T. sp* EXT9 (page.180). L'analyse comparative des génomes montre que les ORFs 8, 9, 10 et 12 de pIRI33 possèdent des fortes similarités de séquences avec les ORFs 1, 13 et 9 de pEXT9b. L'alignement de certaines protéines homologues produit des pourcentages d'identité très élevés (environ 90,0% !). L'hypothèse d'une cohabitation entre un ancêtre de pIRI33 et un plasmide apparenté à pEXT9b mérite d'être posée. En effet, des évènements de recombinaison auraient pu aboutir à la création d'un plasmide hybride perdant l'opéron conservé entre pEXT9a et pAMT7 (Figure 28), pour être remplacé par des gènes présents sur ce second plasmide.

1.3.5 Discussion sur cette famille de plasmides

Une famille majoritaire de plasmides avait été définie, cinq plasmides provenant de souches originaires des différents océans ont été séquencés (pIRI33, pIRI48, pCIR10, pAMT7, pEXT9a). Les expériences préliminaires d'hybridation ADN/ADN ont été vérifiées après séquençage de ces plasmides. L'analyse comparative des génomes révèle une structure génétique en trois compartiments. Le premier englobe les gènes strictement conservés. Le second permet la définition de deux sous-groupes au sein de cette famille, et finalement le troisième est constitué de gènes non conservés, souvent orphelins.

Les gènes conservés sont organisés en un opéron codant quatre protéines, potentiellement impliquées dans la réplication : un régulateur de transcription, une hélicase et une ADN primase-polymérase putative. Le régulateur de transcription contrôle l'expression de l'opéron et *in fine* la réplication et le nombre de copies de l'élément génétique. L'hélicase, de la superfamille I, est apparentée à la famille Rep/UvrD/PcrA. La phylogénie de ces hélicases forme un clade apparenté aux hélicases des Euryarchaea halophiles, distant des bactéries et des eucaryotes. Ce sont les premiers éléments génétiques d'*Archaea* à posséder des hélicases de cette famille. Les chromosomes de Thermococcales ne possédant pas d'hélicase de cette famille, il est légitime de se poser la question de leur origine. L'hypothèse de la présence de ce type d'hélicase dans un génome ancestral aux Euryarchaea semble peu probable car elle nécessiterait une perte dans l'ensemble des phyla et un maintien seulement dans les génomes des halophiles, ce type

d'hélicases n'étant présent que sur les chromosomes d'*Archaea* halophiles. Ces dernières ont probablement été transférées à partir d'un élément génétique vers ces chromosomes. Le génome des *Archaea* halophiles est souvent constitué de multiples réplicons, plusieurs chromosomes, mégaplasmides et plasmides. Ces génomes étant très plastiques, un évènement de recombinaison ancien a peut-être permis l'acquisition et la fixation de ce type d'hélicase à partir d'un élément génétique. La présence de ce type d'hélicase pourrait également expliquer la cohabitation stable de plusieurs chromosomes dans les génomes d'halophiles, la réplication des différents réplicons étant contrôlée par différentes familles de protéines. L'hypothèse d'une hélicase apportée par un élément génétique est renforcée par la présence de quatre gènes plasmidiques homologues avec des gènes portés par le virus PAV1. Parmi ces gènes, certains sont des parorphans car il n'existe pas d'homologues en dehors des éléments génétiques de Thermococcales. On peut éventuellement penser que ces éléments sont apparentés et dérivent d'un ancêtre commun. Dans ce cas, l'hélicase aurait été perdu dans le génome de PAV1 et remplacée par un autre gène codant une protéine assumant cette fonction.

La protéine codant l'ADN primase-polymérase possède de faibles similarités de séquences avec les protéines de même fonction. Néanmoins, les deux activités de cette protéine viennent d'être confirmées par l'équipe de P. Forterre, qui a également caractérisé deux plasmides appartenant à cette famille (communication personnelle Thèse N. Soler).

La génomique comparée, combinée à la phylogénie des gènes conservés, montre l'existence de deux sous-familles d'éléments génétiques. D'une part, les plasmides pCIR10 et pIRI48 possèdent des gènes spécifiques, codant majoritairement des protéines impliquées dans la fixation aux acides nucléiques, dont la plus grosse protéine des plasmides est homologue à un parorphan de PAV1 qui serait impliqué dans la compaction de l'ADN. D'autre part, les plasmides pEXT9a, pIRI33 et pAMT7 forme le second groupe. A l'inverse de la sous-famille précédente, ils ne possèdent pas de gènes qui leurs soient spécifiques. Néanmoins, les taux de similarités de séquence au sein de l'opéron réplicatif sont très importants malgré des origines géographiques très distantes.

Cette classification ne résulte pas d'origines géographiques différentes, ni de la phylogénie des souches hébergeant les plasmides. En effet, les hôtes des plasmides pIRI33 et pIR48 sont issus du même site géographique, alors que ces deux plasmides appartiennent chacun à des sous-familles différentes.

En dehors de gènes conservés entre plasmides, de nombreux parorphans ont été mis en évidence. Ils possèdent pour homologues des gènes codés par d'autres plasmides de Thermococcales séquencés au cours de ce travail. Cette observation illustre la plasticité des éléments génétiques et les potentialités de recombinaison lorsque deux éléments génétiques cohabitent dans une cellule. La comparaison des génomes de pEXT9a et pEXT9b (p.180), présents dans une même souche, renforce cet argument et suppose l'existence de transferts horizontaux entre ces deux réplicons, comme cela a été observé entre le virus SSV4 et le plasmide pXZ1 cohabitant dans une même souche de *Sulfolobus* (Peng 2008). Cet exemple illustre également la frontière ténue entre plasmide et virus.

A partir de ces observations, je souhaite émettre une hypothèse sur les deux sous-groupes : pCIR10 et pIRI48 seraient apparentés à des virus en raison de la présence de gènes conservés qui ne sont pas impliqués dans la réplication et qui sont homologues au virus PAV1 ; tandis que pIRI33, pAMT7 et pEXT9a ressemblent plus à des plasmides au sens strict du terme : seuls les gènes impliqués dans la réplication sont conservés. Bien que la distinction entre virus et plasmide soit assez précise chez les bactéries, la plasticité des génomes d'*Archaea* et les lacunes en matière de données génomiques et surtout expérimentales **ne permettent pas pour le moment d'établir une distinction entre ces deux classes d'éléments génétiques** chez les Thermococcales. Les connaissances des éléments génétiques des Crenarchaea thermoacidophiles sont plus nombreuses et révèlent une grande plasticité de ces éléments génétiques. Les deux meilleurs exemples sont pSSVx, un plasmide hybride recombiné avec un virus de type SSV (Arnold *et al.* 1999) ou bien la capacité d'encapsidation du plasmide pSSVi de la famille pRN ayant acquis des séquences d'attachement lui permettant de s'encapsider lorsqu'un virus de type SSV produit les protéines de capside (Wang *et al.,* 2007).

L'existence d'une sous-famille virale parmi ces « plasmides » nécessiterait une meilleure connaissance de la fonction et de l'implication des protéines homologues à celles du virus PAV1 dans la formation de particules virales.

2. pAMT11, un EG apparenté à un prophage de *T.kodakaraensis*

2.1 *Caractérisation du plasmide pAMT11 et de la souche porteuse*

La souche porteuse du plasmide pAMT11 a été isolée à partir d'un échantillon collecté sur la ride Pacifique Est (EPR East Pacifique Ridge N12°45'16'' W103°59'20'') lors de la campagne Amistad (1999). Cet échantillon est un morceau de paroi de cheminée hydrothermale, plus précisément d'un fumeur noir du site Pulsar (2500m de profondeur). Cette souche a une croissance optimale à 87,5°C et pH 6,5. Elle est affiliée au genre *Thermococcus* par séquençage du fragment d'ADN comportant les ADNr16S, ADNr23S et le court espace intergénique. Cette séquence de 1857pb (FJ182227) est proche de celle de *T. hydrothermalis*, également isolée du site EPR. La comparaison de ces séquences montre que seulement 8 mutations séparent *T.* sp AMT11 et *T. hydrothermalis*.

pAMT11 est un plasmide de 20 534 pb, possédant une composition en G+C de 55,3%. Le contenu en G+C n'est pas uniformément distribué au long de la séquence, il présente de nombreux points d'inflexions correspondants principalement aux séquences intergéniques riches en répétitions. Quelques répétitions remarquables seront analysées au cours de la discussion concernant l'origine de réplication et lors de la détection d'évènements potentiels de recombinaison (p.134).

30 ORFs ont été déterminés, de tailles comprises entre 147 et 1860pb (Tableau 25). Ces ORFs sont sur le même brin d'ADN qui, par convention, sera considéré comme le brin direct. Ils sont séparés par de courts espaces intergéniques représentant 8% du génome. Le plus large espace intergénique est situé entre les ORFs 20 et 21, c'est également la région ayant la plus faible composition en G+C et contenant les répétitions IR5 à IR8.

Les ORFs commencent principalement par ATG (76,7%), mais aussi par GTG (20%) ou TTG (3,3%) (Tableau 25). Les codons stop majoritairement utilisés sont TGA (90,1%), puis TAA (6,6%) et TAG (3,3%). Ces fréquences sont comparables à celles observées sur les chromosomes de Thermococcales. 24 ORFs possèdent un RBS, ce qui renforce la probabilité de codage d'un polypeptide.

Tableau 25 Tableau des ORFs de pAMT1

ORF	Position Début	Fin	Taille (nt)	Contenu G+C (%)	Frame	Codon Start	Stop	RBS
1	1	1737	1737	53,4	+1	GTG	TGA	GGAGG
2	1794	3140	1347	58	+3	ATG	TGA	GGAGGT
3	3270	4694	1371	59,1	+3	TTG	TGA	GGAGG
4	4694	4834	141	46,8	+2	ATG	TAG	GGAGGTG
5	4849	5103	255	52,6	+1	GTG	TGA	GGGGGTG
6	5097	5453	357	48,2	+3	ATG	TGA	
7	5611	5988	378	50,5	+1	ATG	TGA	GAGGTG
8	5994	6698	705	52,4	+3	ATG	TGA	GGAGGTG
9	6703	7389	687	56,3	+1	ATG	TGA	GGAGGG
10	7494	7832	339	59	+3	GTG	TGA	GAGGTG
11	7840	8733	894	56,9	+1	ATG	TGA	GGGGTG
12	8540	8941	402	54,2	+2	ATG	TGA	
13	8941	9375	435	61,1	+1	GTG	TGA	
14	9378	9689	312	57,6	+3	ATG	TGA	
15	9689	10954	1266	57,2	+2	ATG	TGA	GGAGG
16	10938	12185	1248	57,6	+3	ATG	TGA	GAGGTG
17	12188	12463	276	49,6	+2	ATG	TGA	GGAGGTG
18	12459	12776	318	48,1	+3	ATG	TGA	GGAGG
19	12783	13235	453	61,2	+3	ATG	TGA	GGAGG
20	13210	14535	1326	61,4	+1	ATG	TGA	
21	15659	15847	189	54,9	+2	GTG	TAA	GGAGGTG
22	15898	17757	1860	54,3	+1	GTG	TGA	GGAGGTG
23	17699	17926	228	58,8	+2	ATG	□□□	
24	17926	18231	306	57,8	+1	ATG	TGA	GGAGGTGA
25	18231	18434	204	59,9	+3	ATG	TGA	GGGGG
26	18540	18686	147	45,6	+3	ATG	TGA	GGTGG
27	18686	18928	243	59,6	+2	ATG	TGA	GGGGG
28	18928	19257	330	63,9	+1	ATG	TGA	GGAGG
29	19545	20192	648	54,3	+3	ATG	TGA	GGAGGTGA
30	20170	20487	318	40,6	+3	ATG	TAA	GGAGGT

19 ORFs sont chevauchants. Les séquences chevauchantes sont majoritairement constituées d'un simple dinucléotide TG, incluant le codon stop du premier ORFs (TGA) et le codon start du second (ATG). Les ORFs chevauchants ne sont pas distribués aléatoirement sur le génome, ils sont principalement localisés dans sa première moitié. Cette organisation diminue les probabilités de séparation des gènes lors de recombinaison et suggère une association fonctionnelle dirigée par une régulation et une expression coordonnée. La prédiction *in silico* des unités transcriptionnelles (TU) ajoute des informations supplémentaires sur l'association de ces gènes. Trois promoteurs, contenant les séquences BRE (*transcriptional factor B responsive element*) et TATA, ont été localisés. Ces promoteurs sont localisés à 50±2 pb du codon start ; une distance plus importante que celle observée sur les génomes de Thermococcales.

La moitié des ORFs ne possède qu'un unique homologue dans les bases de données, dans le génome de *Thermococcus kodakaraensis* KOD1. Ils sont qualifiés d'ORFans orthologues (Yin *et al.,* 2006). Ces homologues sont tous localisés dans la région TKV1, un îlot génomique de 23,6kb annoté *Virus like Integrated Element*. Cette région est bordée par une intégrase partitionnée, correspondant aux extrémités *intN* (TK0073) et *intC* (TK0103), selon un mécanisme d'intégration similaire aux virus de *Sulfolobus* (She et al. 2004).

Tableau 26 Tableau des protéines codées par pAMT11

Num	Taille (AA)	pI	SignalP	TMH	Fonction putative	Blast			
						Protéine	Espèce	eValue	Identité %
1	579	4,9	+	0		TK0074	*Thermococcus kodakaraensis*	10e-172	56
2	449	4,5	+	0		TK0075 TK0091	*Thermococcus kodakaraensis*	1.10e-102	49
3	457	5,2		0	Subtilisine, Serine protéase	TK0076	*Thermococcus kodakaraensis*	4.10e-179	64
4	50	4,5	+	0		TK0077	*Thermococcus kodakaraensis*	3.10e-9	70
5	85	3,7		0		TK0078	*Thermococcus kodakaraensis*	4.10e-24	62
6	119	5	+	3					
7	126	4,9	+	4		TK0079	*Thermococcus kodakaraensis*	5.10e-38	65
8	237	5,3	+	7		TK0080	*Thermococcus kodakaraensis*	4.10-42	40
9	229	8	+	0		TK0081	*Thermococcus kodakaraensis*	4.10e-67	61
10	113	4	+	4		TK0083	*Thermococcus kodakaraensis*	3.10e-16	47
11	306	8,9		2		TK0085	*Thermococcus kodakaraensis*	8.10e-91	56
12	134	11,8		0					
13	145	10,2		0					
14	104	9,4		0	Transcription regulator, ArsR	TK0086	*Thermococcus kodakaraensis*	1.10e-5	28
15	422	7,8		0		TK0087	*Thermococcus kodakaraensis*	9.10e-96	47
16	421	5,1	+	2		TK0088	*Thermococcus kodakaraensis*	8.10e-63	37
17	93	10,9	+	0					
18	109	9	+	3					
19	151	10,9		0		TK0089	*Thermococcus kodakaraensis*	3.10e-5	33
20	442	4,7		0		TK0090	*Thermococcus kodakaraensis*	1.10e-48	38
21	69	9,4		0	Transcription regulator, PhiH1	PhiH1	*Haloarcula marismortui*	0.01	35
22	620	9,5		0	Rep	p63	*Pyrococcus sp. JT1*	5.10e-54	29
23	92	8,1		0		TK0097 TK0411 TK0583	*Thermococcus kodakaraensis*	5.10e-6	36
24	102	5,4		0					
25	68	11,5		0		TK1368	*Thermococcus kodakaraensis*	1.10e-7	54
26	59	9,7		0					
27	81	4,2		0					
28	115	4,9		0	Transcription regulator, HTH	HTH	*Thermoanaerobacter tengcongensis*	0,31	29
29	216	10,1		0	Resolvase	PH1174	*Pyrococcus horikoshii*	1.10e-73	64
30	106	7		2					

Protéines codées par les ORFs de pAMT11, avec leurs tailles en AA, points isoélectriques, présence d'un peptide signal (SignalP), nombre d'hélices transmembranaires (TMH) et résultats de BLASTP.

70% (14,7/20,5kb) de la séquence d'AMT11 est partagée avec TKV1. Les ORFs 1 à 20 de pAMT11 sont homologues aux ORFs TK0074 à TK0090 de TKV1. Au sein de cette grande zone conservée, deux ORFs, TK0082 et TK0084, ne sont pas retrouvés dans pAMT11. Inversement, les ORFs 6, 12, 13, 17 et 18 de pAMT11 ne sont pas retrouvés dans la région TKV1.

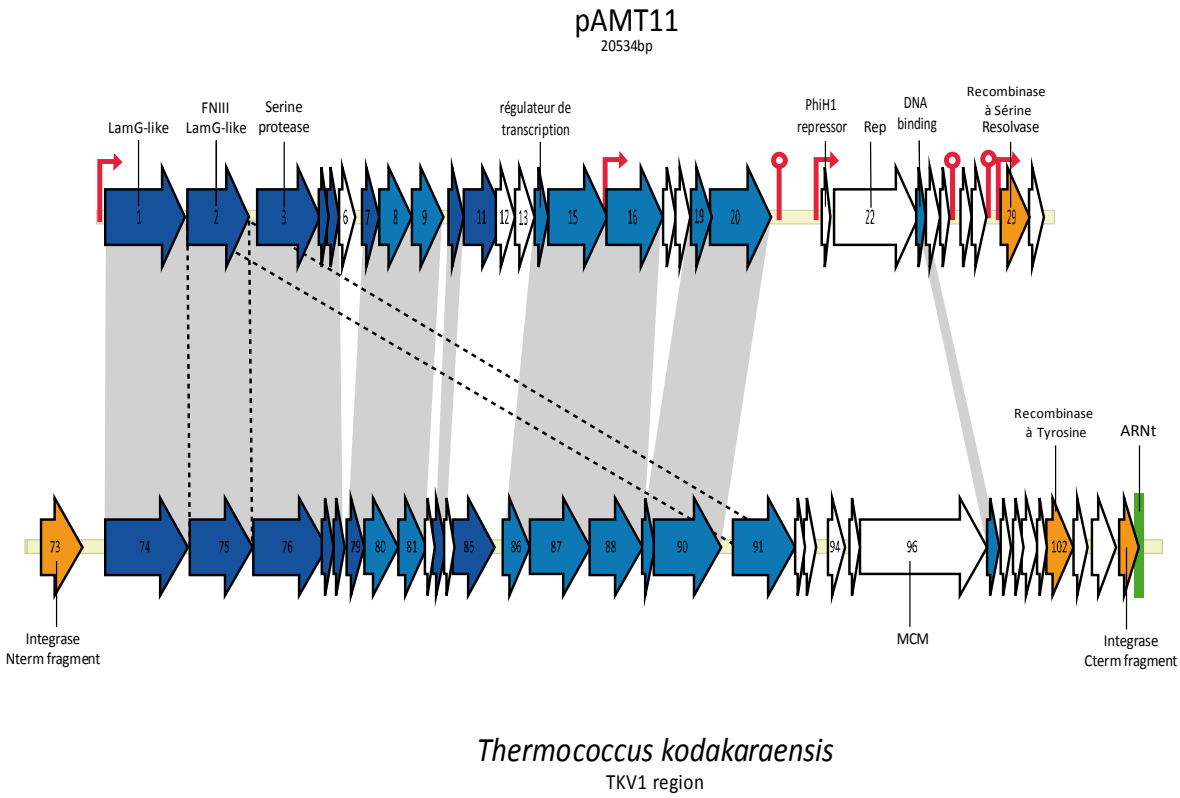


Figure 30 Comparaison du génome de pAMT11 et de la région TKV1 du génome de *Thermococcus kodakaraensis*

Les ORFs sont désignés par les flèches. La couleur bleue indique la présence d'homologues entre pAMT11 et TKV1. (bleu foncé : plus de 50% d'identité, bleu clair entre 25 et 50%, bleu très clair entre 15 et 25%). Les ORFs en orange codent des recombinases. Les symboles rouges désignent les signaux de transcription, les flèches brisées : promoteur, sucette : terminateur.

2.2 *ORFs conservés entre pAMT11 et TKV1*

L'ORF1 code la plus grosse protéine (579AA) conservée entre pAMT11 et TKV1. Son homologue dans TKV1 code la protéine TK0074 (22% d'identité, 29% de similarité). Des similarités sont également partagées avec la protéine Haur1999 de *Herpetosiphon auriantacus* (Ordre des *Chloroflexi*) et l'ORF678 de PAV1. Ces protéines sont toutes les deux annotés LamG. Ces laminines globulaires (LamG) sont de grosses glycoprotéines hétérotrimériques membranaires impliquées dans les adhésions cellulaires (Sasaki *et al.,* 2004). Dans PAV1, elles sont supposées être impliquées dans la reconnaissance virus-hôte (Geslin *et al.* 2007).

L'ORF2 code une protéine de 445AA. Deux homologues sont présents dans TKV1. Le premier est en position synténique avec pAMT11 (TK0075, 47,7% identité, 63,1% similarité) alors que le second est en bordure du bloc de gènes conservés entre pAMT11 et TKV1 (TK0091, 33,1% identité, 46,6% similarité). Ces deux orthologues synténiques partagent 35,5% d'identité. Les calculs de similarités montrent que les deux orthologues de TKV1 ne sont pas issus d'une duplication segmentale et que la protéine codée par le plasmide est plus proche de son homologue synténique que de celui situé à l'extrémité conservée. Aucun homologue n'est détecté dans les bases de données. La protéine se découpe en deux domaines. L'extrémité N-terminale ressemble aux immunoglobulines-like bactériennes (SM00634), et plus précisément à la classe des fibronectines de type III (SM00060), des protéines membranaires impliquées dans la fixation au collagène ou à l'actine des eucaryotes (Steward *et al.,* 2002), mais aussi chez les bactéries dont elles servent de cible de fixation et d'attachement (Beckmann *et al.,* 2002). Ce domaine est fréquemment rencontré sur les protéines de phages lytiques et tempérés. Il permet la fixation du virus à la surface cellulaire bactérienne par interaction avec les chaines glucosidiques (Fraser *et al.,* 2006). Le domaine C-terminal ressemble, quant à lui, aux Laminines globulaires (PF02210) et permettrait la formation d'hétéromultimères par interaction LamG-LamG avec la protéine codée par l'ORF1, mais également entre les deux orthologues codées par TKV1.

L'ORF3 code la protéine dont la fonction est la moins ambiguë. Il s'agit d'une endoprotéase à sérine (EC 3.4.21) contenant la triade catalytique Asp/Ser/His. L'arrangement des AA au niveau du site catalytique est DHS, classant cette protéase dans la famille des subtilisines. C'est la seconde famille la plus importante de protéases à sérine. Ces protéines sont trouvées chez les *Bacteria*, les *Archaea*, les eucaryotes et les virus (Siezen *et al.,* 1997). L'ORF3 est plus court que les homologues trouvés dans les bases de données. Les subtilisines sont des protéines mosaïques. Elles peuvent présenter de grandes extensions variables aux extrémités N- et C-terminales. Le domaine catalytique de ces protéases comporte un *core*, résidus présents dans les subtilisines, et des *inserts*, résidus supplémentaires. Plusieurs subtilisines ont été caractérisées chez les *Archaea*, permettant leur division en 2 sous-familles : les pyrolysines et les stetterlysines. Les pyrolysines comportent les subtilisines de *Pyrococcus furiosus* (Voorhorst *et al.,* 1996) et *P. horikoshii* (Kawarabayasi *et al.* 1998). La famille des stetterlysines comprend les subtilisines de *Thermococcus stetteri* (Voorhorst *et al.,* 1997), *T.kodakaraensis* (Kannan *et al.,* 2001) et d'*Aeropyrum pernix K1* (Catara *et al.,* 2003). La principale différence entre ces deux familles est la taille des protéines. Les pyrolysines (~1300AA) sont les plus grosses protéases à sérine

découvertes alors que les stetterlysines possèdent une taille similaire aux subtilisines des Eubactéries(~420AA). Cette différence de taille résulte de la présence de grands inserts en amont de l'histidine catalytique. L'ORF3 ne possède pas ces inserts caractéristiques des autres subtilisines des Thermococcales et peut être classée dans la famille des stetterlysines.

Les subtilisines sont synthétisées dans le cytoplasme sous forme de précurseur nommé préprosubtilisine. Sous cette forme, les préséquences et proséquences sont attachées en N-terminal de la protéine mature. La préséquence agit comme un peptide signal facilitant la sécrétion de la prosubtilisine à travers la membrane cytoplasmique. La proséquence agit comme une chaperonne intramoléculaire et permet le repliement correct de la protéine mature. La proséquence est ensuite clivée par autoprotéolyse afin de produire une subtilisine mature active. Ce processus de maturation des subtilisines des *Archaea* hyperthermophiles (Kannan *et al.* 2001) est similaire à celui des Eubactéries. La préséquence a été identifiée dans l'ORF3 en tant que signal d'excrétion par le programme SignalP 3.0.

Ces protéases, ATP-indépendantes, peuvent être impliquées dans de multiples voies cellulaires. La plupart des hyperthermophiles hétérotrophes peuvent croître sur des milieux dont le substrat protéique servira de source primaire en carbone et en énergie. Ces substrats doivent préalablement être transformés par des protéases extracellulaires ou associés à la membrane afin d'être importés. Les produits de l'hydrolyse extracellulaire sont transportés dans la cellule, probablement grâce à des transporteurs de type ABC. Les peptides ainsi acquis sont ensuite fermentés en acides libres, tels que l'acétate, isovalérate, butyrate ou phenylbutyrate servant à la production d'ATP. Une étude a mis en évidence que ces protéases ATP-indépendantes ne sont pas cantonnées à l'acquisition de substrats protéiques. Elles interviennent également dans la formation de la surface S (S-layer) en modifiant les facteurs d'adhésion, la quantité d'antigènes de surface et la morphologie de la cellule (Vickerman *et al.,* 2002), caractéristique de la réponse au choc thermique chez *Pyrococcus furiosus* (Shockley *et al.,* 2003). Néanmoins, il n'est pas impossible que ces endopeptidases puissent jouer un rôle dans la biologie du plasmide, par exemple dans des mécanismes de régulation. Récemment, Pa-Kae1 (*Pyrococcus abyssi* Kinase associated endopeptide 1), une protéine annotée comme endopeptidase dans *Pyrococcus abyssi*, s'est révélée ne pas avoir d'activité protéolytique mais elle peut en revanche se fixer à l'ADN et avoir une activité endonucléase de classe I (Hecker *et al.,* 2007). Bien que l'endopeptidase en question ne soit pas de la même famille que l'ORF3 de pAMT11, cet exemple illustre les limites de l'annotation et la nécessité de confirmations biochimie expérimentales. La dernière fonction, qui me semble la plus intéressante, est la présence d'une peptidase de procapside sur différents

phages, tels le phage lambda ou psiM2 de *Methanobacterium thermoautotrophicum*. Elle intervient, après attachement à la cellule hôte dans la pénétration et la lyse de la cellule. Cette fonction sera détaillée dans la discussion, permettant d'avoir une vue d'ensemble des fonctions codées par le génome de pAMT11.

L'ORF14, d'une taille de 105AA, possède pour homologue TK0086 de TKV1. Cet ORF est annoté en tant que régulateur de transcription de la famille ArsR (COG1414). L'identité globale de séquence entre les deux protéines n'est que de 14,6%, et s'explique par une différence de taille entre ces deux protéines. La protéine TK0086 est beaucoup plus grande et possède une extension N-terminale de 90AA absente de l'homologue plasmidique. Les régulateurs ArsR sont des protéines possédant un motif Hélice-Tour-Hélice en C-terminal et un domaine de fixation aux ions métalliques en N-terminal. L'alignement des deux protéines homologues montre que la région contenant le motif HTH est conservée et peut être identifiée lors d'une recherche dans les bases de motifs (PF01022). Les protéines à HTH possèdent souvent des extrémités peu conservées servant à stabiliser le domaine HTH (Religa *et al.,* 2007).

L'ORF23 est le seul gène conservé entre pAMT11 et TKV1, il est localisé en dehors du grand bloc synténique (Figure 30). Il code une protéine de 76AA possédant uniquement des homologues dans le génome de *T. kodakaraensis*. Ces trois paralogues, TK0097, TK0411 et TK0583, sont respectivement localisés dans les « éléments viraux intégrés » TKV1, TKV2 et dans un autre îlot génomique qui était passé jusqu'à présent inaperçu (Figure 31). Cette région, que nous avons appelée TKV5, est comprise entre les ORFs TK0579 et TK0599. Elle est bordée par une intégrase partitionnée chevauchante avec un ARNt, possède des biais dans la composition en dinucléotides en G+C et en usage des codons. De plus, elle code des protéines hypothétiques qui ne sont pas présentes dans les autres génomes de Thermococcales.

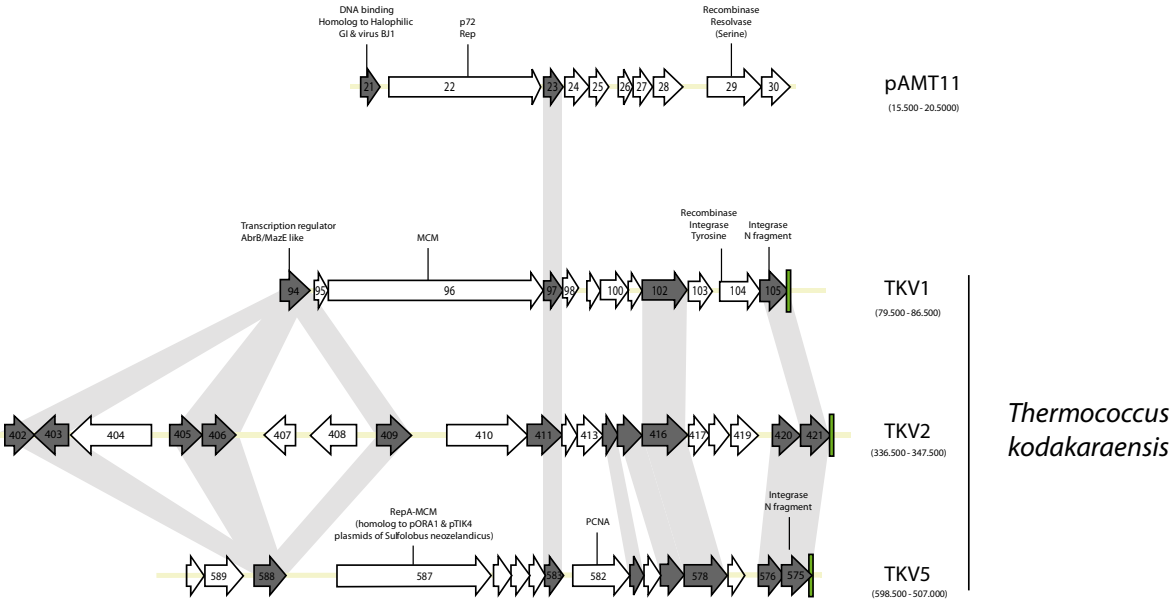


Figure 31 Génomique comparée des homologues de l'ORF23 de pAMT11

Représentation de l'extrémité 3' des îlots génomiques TKV1, TKV2, TKV5 et du plasmide pAMT11. Les flèches grises représentent les ORFs homologues. La fonction hypothétique de chaque protéine est mentionnée lorsqu'elle est renseignée.

On note également que le contexte génomique des différents paralogues est associé à des gènes impliqués dans la réplication. Sur pAMT11, l'ORF22 suit le gène *rep,* codant la protéine initiatrice de la réplication ; sur TKV1 elle suit l'hélicase MCM (TK0096) ; sur TKV5, elle est entourée par le PCNA (TK0587) et une protéine homologue à la RepA-MCM des plasmides pORA1 et pTIK4 de *Sulfolobus neozelandicus* (Figure 31). Ces observations suggèrent une relation avec de l'ADN mobile ou inversement une implication dans l'intégration/capture de gènes par le chromosome.

2.3 *ORFs non conservés pAMT11 et TKV1, implication réplication et recombinaison.*

L'ORF21 code un répresseur de la transcription. Cette protéine de 69AA possède un domaine de fixation à l'ADN de type WH (*Winged-helix*), de la famille GlpR (COG1369). Trois homologues sont détectés, uniquement dans les génomes d'Euryarchaea mésophiles : gp20 du virus intégratif BJ1 infectant *Halorubrum saccharovorum* et deux paralogues, rrnAC0584 et rrnAC2383, du génome de *Haloarcula marismortui* ATCC43049. Ils sont annotés comme étant des répresseurs de transcription et possèdent 35% d'identité avec la protéine de pAMT11. Le contexte génomique de ces paralogues est composé de nombreuses protéines hypothétiques codées par des gènes possédant des biais en dinucléotide et tétranucléotide, en usage des codons et en contenu en G+C. De plus, ces paralogues sont localisés à proximité d'une intégrase ce qui suggère une fois de plus leur appartenance à des îlots génomiques. Cette famille de protéines, associées aux éléments

génétiques mobiles d'Euryarchaea, est impliquée dans la fixation à l'ADN. Localisées en amont de gènes impliqués dans l'initiation de la réplication, ces protéines pourraient intervenir dans la régulation du nombre de copies du plasmide.

L'ORF22 code une protéine de 616AA, que nous avons appelée p72 en raison de sa masse théorique de 72kDa. Elle est homologue (27% identité, 40% similarité) à la protéine Rep p63 du plasmide à réplication par cercle roulant pRT1 de *Pyrococcus sp* JT1 (Ward *et al.* 2002). Cette protéine est une endonucléase site-spécifique, ligase et nucleotidyl-transférase impliquée dans l'initiation de la réplication des réplicons à cercle-roulant. Les protéines Rep ont une activité topoisomérase I permettant le clivage et la ligature d'ADN sur de nombreux éléments génétiques : plasmides, virus et transposons. Elles participent à l'initiation et la terminaison de la synthèse du brin direct médiée par cercle-roulant. La prédiction de topologie met en évidence quatre domaines séparés par trois boucles localisées à la position des insertions/délétions de l'alignement de séquence. Les quatre domaines sont strictement conservés entre ces protéines, suggérant que p72 est également une protéine Rep. De plus, les protéines Rep sont caractérisées par trois motifs détectés aussi bien sur p63 que p72 (voir page 11).

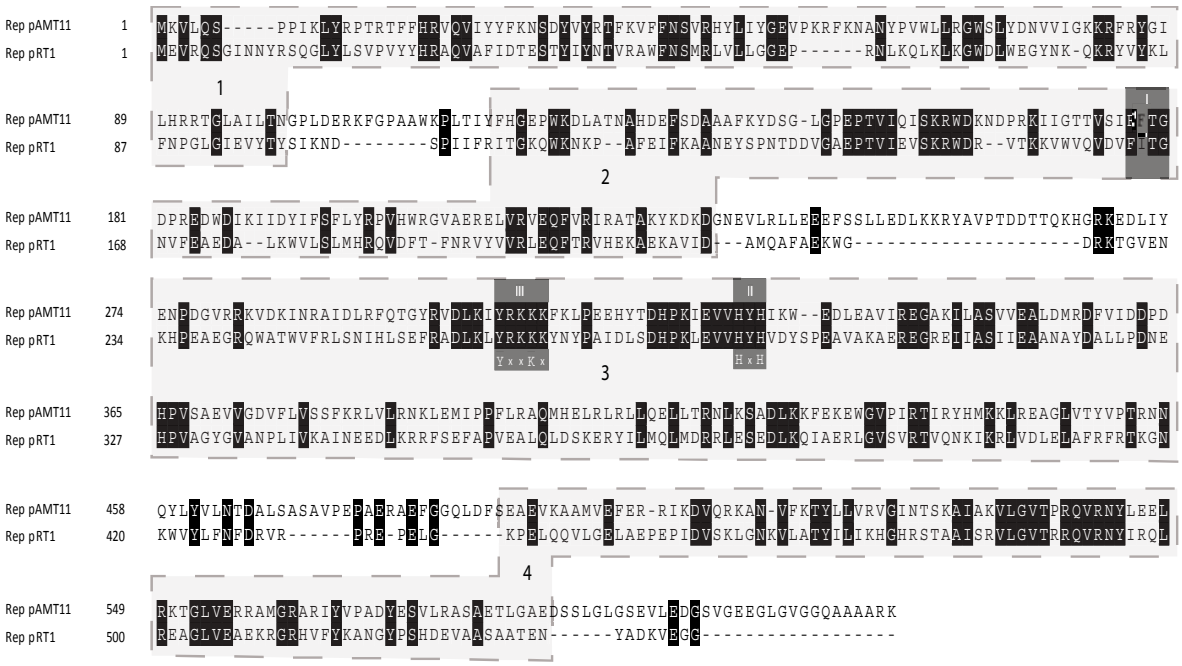


Figure 32 Alignement des protéines Rep de pAMT11 et pRT1

Alignement de la séquence peptidique des Rep de pAM11 et pRT1 : p72 et p63. Les AA conservés sont écrits en blanc sur fond noir. Les domaines, prédits par Scoobydomain, sont numérotés en chiffre arabe et sont délimités par une zone à fond gris et pointillés. Les motifs fonctionnels des protéines Rep et leurs séquences consensus sont délimités par un fond gris foncé numéroté en chiffres romains.

Le premier domaine (1) ne possède aucune caractéristique particulière. Le second domaine (2) possède le motif I, de fonction inconnue. Le troisième domaine (3) possède un motif HTH de la famille MerR (PF00376, PF01022), le motif II (HxH) et le motif III, YxKKK, identique à celui du plasmide pUB112 de la famille pE194/pLS1. L'agencement des motifs est habituellement I-II-III, alors que sur p63 et p72 il est de type I-III-II. C'est la principale différence observée avec les autres protéines Rep. Lors de la description de pTN1, N.Soler et ses collaborateurs avaient déjà suggéré un problème d'annotation des motifs de p63 qui appartiendrait en réalité à une nouvelle famille de protéines Rep (Soler *et al.* 2007). En effet, cette nouvelle protéine suggérait que p63 appartenait à une famille distincte de protéines Rep. L'alignement des quatre protéines Rep de Thermococcales confirme l'existence de deux familles distinctes et une prédiction erronée des motifs de p73 faute de posséder suffisamment de séquences homologues à l'époque de l'analyse de pRT1 (Ward *et al.* 2002).

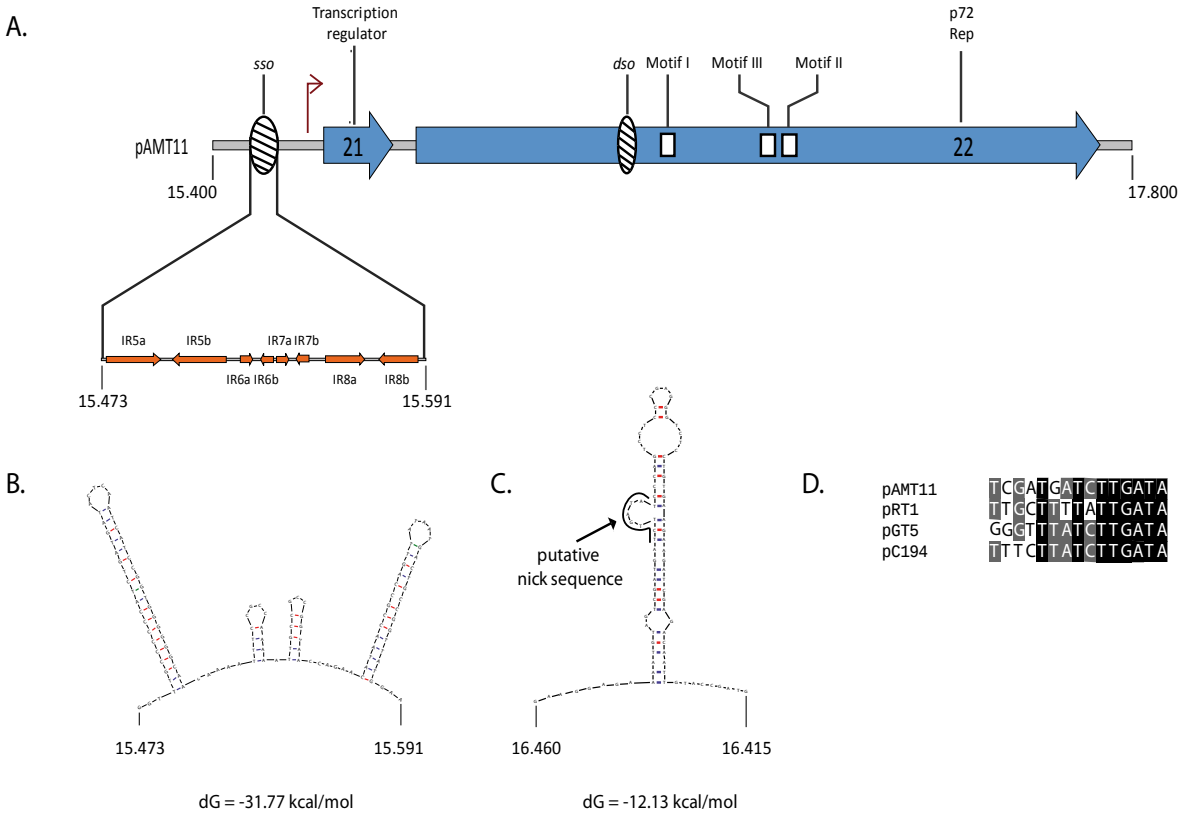


Figure 33 Organisation des motifs fonctionnels de la protéine Rep ; localisation des dso et sso putatives.

A. Représentation schématique de la région 15400 à 17800 de pAM11. Elle contient deux ORFS représentés par les flèches bleues. Les trois motifs fonctionnels des Rep protéines sont représentés par des carrés blancs. Les origines de réplication *sso* et *dso* sont localisées par l'ellipse hachurée. La *sso* est agrandie afin de montrer les répétitions inversées 5 à 8.

B. Prédiction de la structure secondaire de la *sso* par DNAfold

C. Prédiction de la structure secondaire de la *dso* par DNAfold

D. Alignement des séquences *nick* de tous les plasmides à RCR de Thermococcales

En dehors de la protéine Rep, la réplication par cercle roulant requiert la présence de deux origines de réplication DSO, *Double Strand Origin*, et SSO, *Single Strand Origin*. La DSO peut être assignée grâce à la présence du motif *nick* (TCTTGG/ATA), en amont de *rep*, dans la boucle d'une structure secondaire de l'ADN de type tige-boucle. La SSO ne possède pas de séquence consensus. Néanmoins, elle se caractérise par un repliement particulier de l'ADN. Cette structure spécifique est détectée en amont du gène *rep* entre les positions 15473 et 15591. Cette séquence est constituée de 4 répétitions inverses (IR5, IR6, IR7, IR8 Figure 33A) adoptant le repliement en épingle le plus stable du génome (Figure 33B).

L'ORF29 code une protéine de 216AA homologue d'une résolvase/invertase de *Pyrococcus horikoshii OT3* (PH1174). Les deux signatures caractérisant cette famille de sérine-recombinases sont localisées à l'extrémité N-terminale. La première contient la sérine impliquée dans la liaison covalente transitoire à l'ADN mais aussi dans l'interface permettant la dimérisation. La seconde signature est une région conservée, repliée sous forme d'un faisceau de trois hélices, incluant un motif de type HTH. Les résolvases permettent la recombinaison site-spécifique entre deux répétitions. Certaines, notamment sur les transposons, provoquent l'excision de la séquence située entre les deux répétitions, alors que d'autres provoquent un simple phénomène d'inversion. Sur les plasmides et virus, ces recombinases adoptent le nom de résolvases, caractérisant leur action dans la résolution des concatémères de réplication. La fonction de la recombinase de pAMT11 ne peut-être prédite par simple analyse *in silico*, l'implication dans des mécanismes plus généraux de recombinaison ne peut être exclue, notamment par la présence de nombreuses répétitions situées dans les courts espaces intergéniques. L'illustration la plus flagrante est la présence de trois répétitions directes de 18 nucléotides (DR1 : GGCCTTGGATCGGAGGTG) contenant un site de fixation du ribosome (RBS). Ces trois copies sont précisément localisées autours des ORFS 21 et 22 (Figure 34). Ces gènes sont intéressants car ils codent les protéines impliquées dans la réplication de pAMT11 et sont différents de ceux portés par TKV1, ce qui en fait la principale différence entre pAMT11 et l'élément intégré TKV1.

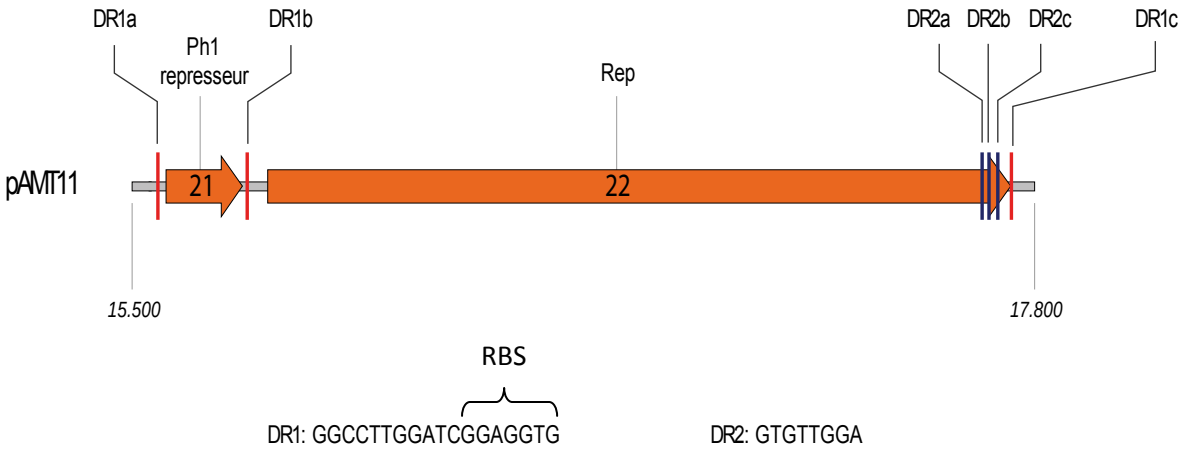


Figure 34 Représentation schématique des répétitions directes encadrant les ORFS 21 et 22 de pAMT11

Les ORFs sont représentés par les flèches. Les répétitions sont représentées par les barres verticales bleues pour DR1 et rouges pour DR2.

2.4 *Discussion des propriétés de pAMT11*

pAMT11 est un plasmide de *Thermococcus* affilié à pRT1, un plasmide à réplication par cercle roulant de *Pyrococcus* sp JT1. Cette parenté est uniquement basée sur la présence d'une protéine initiatrice de la réplication homologue à celle de ce petit plasmide cryptique. La découverte d'une protéine Rep dans pAMT11 a permis de prédire avec une plus grande acuité des motifs fonctionnels de ces protéines. Malgré la conservation des structures et des séquences constituant les origines de réplication *sso* et *dso*, l'agencement des motifs suggère l'existence d'une nouvelle famille de protéines Rep.

Les réplicons utilisant un mécanisme RCR sont généralement cryptiques et de petites tailles, à l'image des trois plasmides de Thermococcales disponibles actuellement dans les bases de données qui ne possédent que deux ORFs. L'intermédiaire de réplication simple brin est beaucoup moins stable que son équivalent simple brin, d'autant plus qu'il est thermolabile. C'est généralement pour cela qu'un second gène code une protéine de stabilisation de l'ADN, comme cela a été décrit sur le plasmide pTN1 de *T. nautilii* (Soler *et al.* 2007). En comparaison, pAMT11 est beaucoup plus grand. Il compte 30 ORFs codant entre autre de nombreuses protéines de fixation à l'ADN. Certaines pourraient être spécialisées dans la stabilisation des intermédiaires de réplication simple et double brin.

La plupart des gènes de pAMT11 sont homologues à ceux de l'îlot génomique TKV1, de *Thermococcus kodakaraensis*. Les deux souches portant ces éléments homologues ne sont pas originaires du même site : *T. sp* AMT11 provient d'une source hydrothermale profonde sur la dorsale Pacifique Est (Mexique), alors *que T.kodakaraensis* a été isolé d'une source côtière du Japon. N'étant à ce jour pas retrouvé sous forme libre ou intégrée au sein d'autres souches, cette observation suggère un héritage non-vertical et une mobilité à travers les océans.

Il est intéressant de constater qu'aucun gène conservé entre ces deux éléments n'est impliqué dans la réplication. En effet, pAMT11 possède une protéine Rep de type RCR alors que l'élément TKV1 possède une protéine initiatrice de la réplication de la famille des MCM, typique de la réplication des chromosomes, mais aussi rencontrée sur deux plasmides d'*Archaea* : pTAU4 de *Sulfolobus* (Greve *et al.,* 2005) et pEXT9b décrit pendant cette étude (page. 180). Les plasmides sont généralement considérés comme des réplicons autonomes constitués d'un set minimum de gènes impliqués dans la réplication et de différents autres gènes interchangeables constituant le *pool flexible*. La comparaison de pAMT11 et TKV1 va à l'encontre de ce concept, les gènes impliqués dans la réplication sont différents alors que les gènes du *pool flexible* sont conservés. L'association de ces gènes dans une fonction commune est d'autant plus probable que les protéines homologues codées par pAMT11 et TKV1 sont très conservées, malgré des sites géographiques très éloignés.

La plupart des gènes conservés codent des protéines orphelines membranaires. Néanmoins, deux protéines conservées font penser que ce module de gènes pourrait être impliqué dans la formation de particules virales. En effet, l'ORF1 code une protéine membranaire apparentée aux laminines globulaires (LamG), déjà rencontrées sur le virus PAV1 de *Pyrococcus abyssi* GE23 où elles sont supposées être un facteur d'attachement du virus à sa cellule hôte.

Au sein du même opéron, il existe une protéase à sérine fréquemment rencontrée dans la procaspide de nombreux bactériophages, tel le phage lambda, mais aussi d'archeophage, tel psiM2 de *Methanobacterium thermoautotrophicum*. Ces protéases interviennent dans différents processus phagiques comme la maturation de la procapside nécessaire à l'assemblage de la capside (Cheng *et al.,* 2004) ou lors du processus de pénétration cellulaire (Keller *et al.,* 1986).

Afin de poser des hypothèses sur les évènements ayant conduit à l'observation de la relation entre TKV1 et pAMT11, il faut tout d'abord définir quelques termes et considérer que ces deux éléments partagent un groupe de gènes homologues que nous nommerons « *Virus-like module* ». En effet, il code probablement pour une capacité de survie accrue en milieu extracellulaire (virus

ou vésicules). Il faut également considérer que TKV1 est issu d'un élément génétique intégratif ancestral pTKV1 possédant une MCM + « virus-like ». La viabilité d'un élément génétique de Thermococcales à MCM a de plus été confirmée lors de l'analyse du plasmide pEXT9b (page 180). TKV1 n'a peut-être plus la capacité de s'exciser, néanmoins, l'entrée dans le cytoplasme d'un autre élément possédant une intégrase pourrait l'exciser par l'action d'une intégrase agissant en trans. Cette action pourrait même être un prérequis à l'intégration du nouvel élément souhaitant libérer le site d'intégration afin d'y prendre la place. Cette hypothèse, déjà observée chez les bactéries, favorise les fréquences des rencontres entre réplicons et donc les évènements recombinaisons inter-éléments génétiques.

Plusieurs hypothèses peuvent être avancées (Figure 35). Elles sont exposées préférentiellement sous forme de schéma beaucoup plus explicite qu'un long discours. Premièrement, pTKV1 aurait pu s'intégrer ou recombiner de manière plus générale dans pRT1 afin de composer un plasmide hybride qui aurait ensuite recombiné pour exclure la partie « réplicative » de type MCM. La seconde hypothèse est basée sur le cheminement inverse entre ces réplicons. La troisième hypothèse ferait intervenir une recombinaison avec réplicon à MCM, apparenté à pEXT9b, permettant ainsi le transfert du module « virus-like ». La dernière hypothèse ferait intervenir un autre type de réplicon dans le genre de pEXT9b.

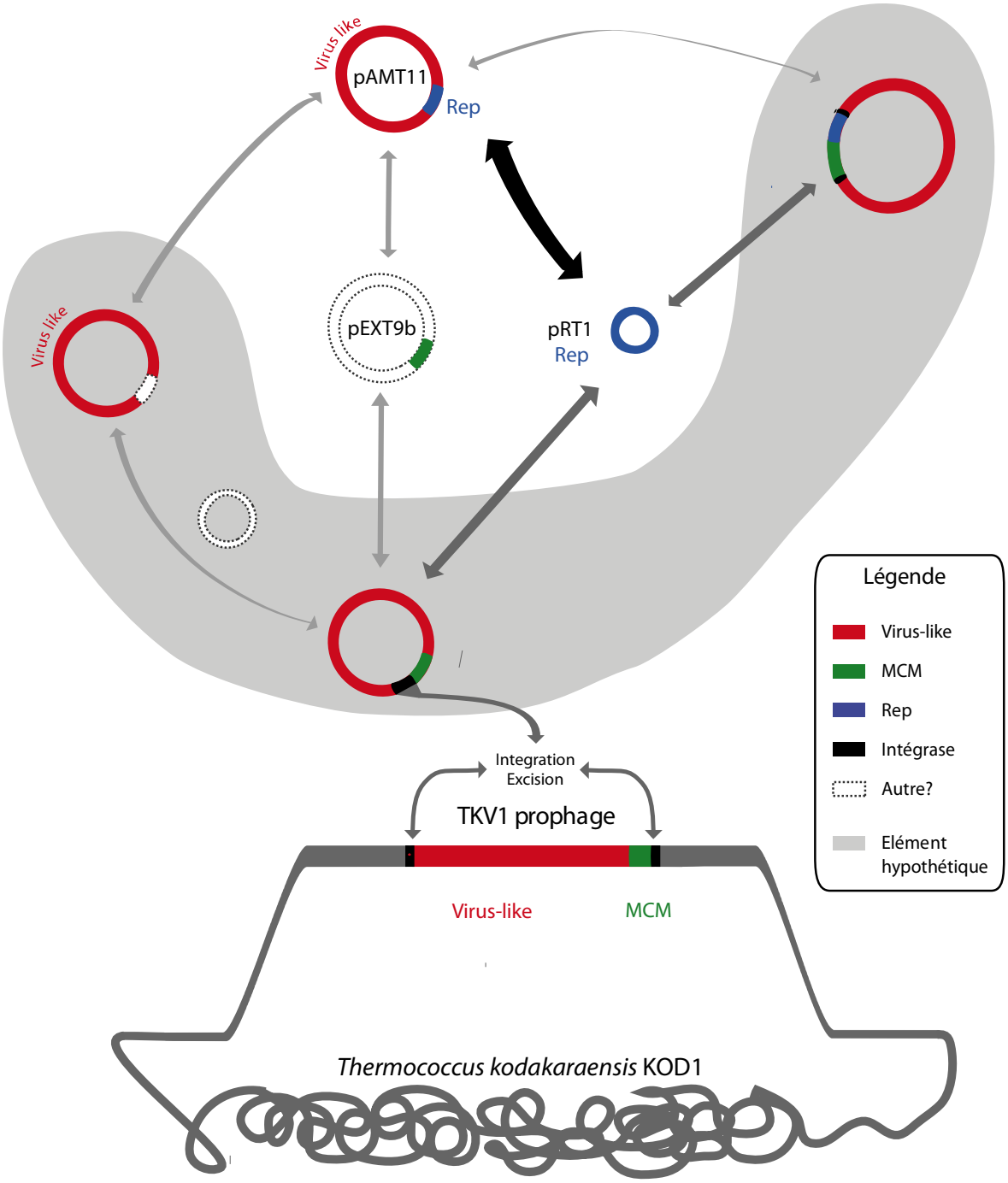


Figure 35 Hypothèses sur les relations entre les plasmides pAMT11 et pRT1 et l'élément intégré TKV1 de *Thermococcus kodakaraensis.*

pAMT11 illustre la notion de modularité des réplicons extrachromosomiques qui permettent l'acquisition et l'échange de groupes de gènes lors de la rencontre de plusieurs réplicons au sein d'un cytoplasme.

3. pGE2, un adénovirus intégratif ?

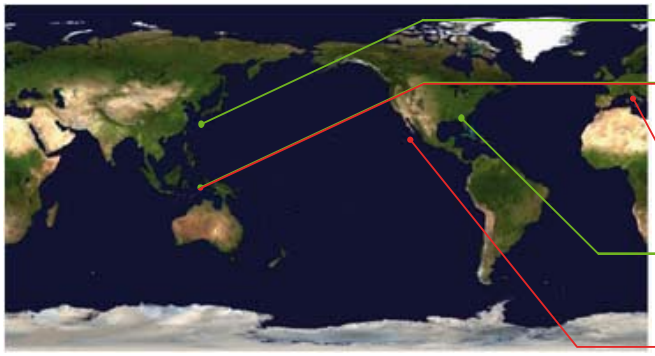
3.1 *Description générale de la souche P.abyssi GE2 et de son plasmide*

Le plasmide pGE2 est issu de la souche *Pyrococcus abyssi* GE2. Tout comme *Pyrococcus abyssi* GE5 (Erauso *et al.* 1993), cette souche a été isolée à partir d'un échantillon de cheminée hydrothermale collecté dans le bassin nord fidjien lors de la campagne du pacifique sud STARMER en 1989. Les séquences ADNr 16S, 23S et l'espace intergénique possèdent 100% d'identité avec *P.abyssi* GE5.

Ce plasmide de 23 976 pb possède une composition en G+C de 40,4%. 36 ORFs ont été déterminés, codant des protéines de tailles comprises entre 50 et 692 AA (Tableau 27). Ce plasmide est homologue à de nombreux îlots génomiques de Thermococcales mais également de Methanococcales mésophiles. pGE2 possède la capacité de s'intégrer dans le chromosome de la souche *P.abyssi* GE2. La présence de gènes affiliés aux adénovirus, notamment celui codant une ATPase de compaction de l'ADN viral, suggère que ce plasmide est capable de survivre en milieu extracellulaire dans « une coque protéique ».

3.2 *pGE2 est affilié à des îlots génomiques d'Euryarchaea*

A l'instar de pAMT11, pGE2 est apparenté à deux éléments viraux intégrés du génome de *T.kodakaraensis* : TKV2 et TKV3, mais aussi à un autre îlot génomique, PHV1, de *P.horikoshii*. L'homologie ne se cantonne pas aux Thermococcales, mais s'étend aux îlots génomiques présents chez certaines Methanococcales : *Methanococcus maripaludis* S2, *M.maripaludis* C6, *M.maripaludis* C7, *M.voltae A3* (Figure 36). Il est surprenant de trouver des éléments intégrés apparentés à pGE2 dans ces Methanococcales mésophiles. En effet, elles sont toutes issues d'échantillons prélevés dans des sédiments intertidaux de vasières côtières de la côte Ouest des USA : Floride, Géorgie et Caroline de Sud. Il aurait été moins surprenant de rencontrer ce genre d'éléments dans *Methanocaldococcus jannaschii,* par exemple, qui partage la niche écologique des Thermococcales, l'hyperthermophilie et les environnements profonds. Une insuffisance du nombre de génomes disponibles pourrait expliquer ce genre d'observation.



Génome	Homologue à pGE2	Thermophile
P. horikoshii	+	+
T.kodakaraensis	+	+
P.abyssi GE2	-	+
P.abyssi GE5	+	+
P. furiosus	+	+
M.aeolicus	-	-
M.maripaludis C5	+	-
M.maripaludis C6	+	-
M.maripaludis C7	+	-
M.maripaludis S2	+	-
M.voltae A3	+	-
M.jannaschii	-	+

Figure 36 Origine des Thermococcales et Methanococcales, présence ou non d'un élément apparenté à pGE2 et température optimale de croissance.

A l'image de pGE2, l'ensemble de ces îlots génomiques sont caractérisés par une intégrase, un bloc de gènes très conservés chez les Thermococcales et un peu moins conservés chez les méthanogènes (Figure 37). La comparaison des régions conservées suggère une relation avec les adénovirus, tandis que les régions non conservées possèdent des gènes impliqués dans la maintenanceet la réplication de l'ADN, rappelant ainsi les observations déjà effectuées lors de la comparaison de pAMT11 avec TKV1.

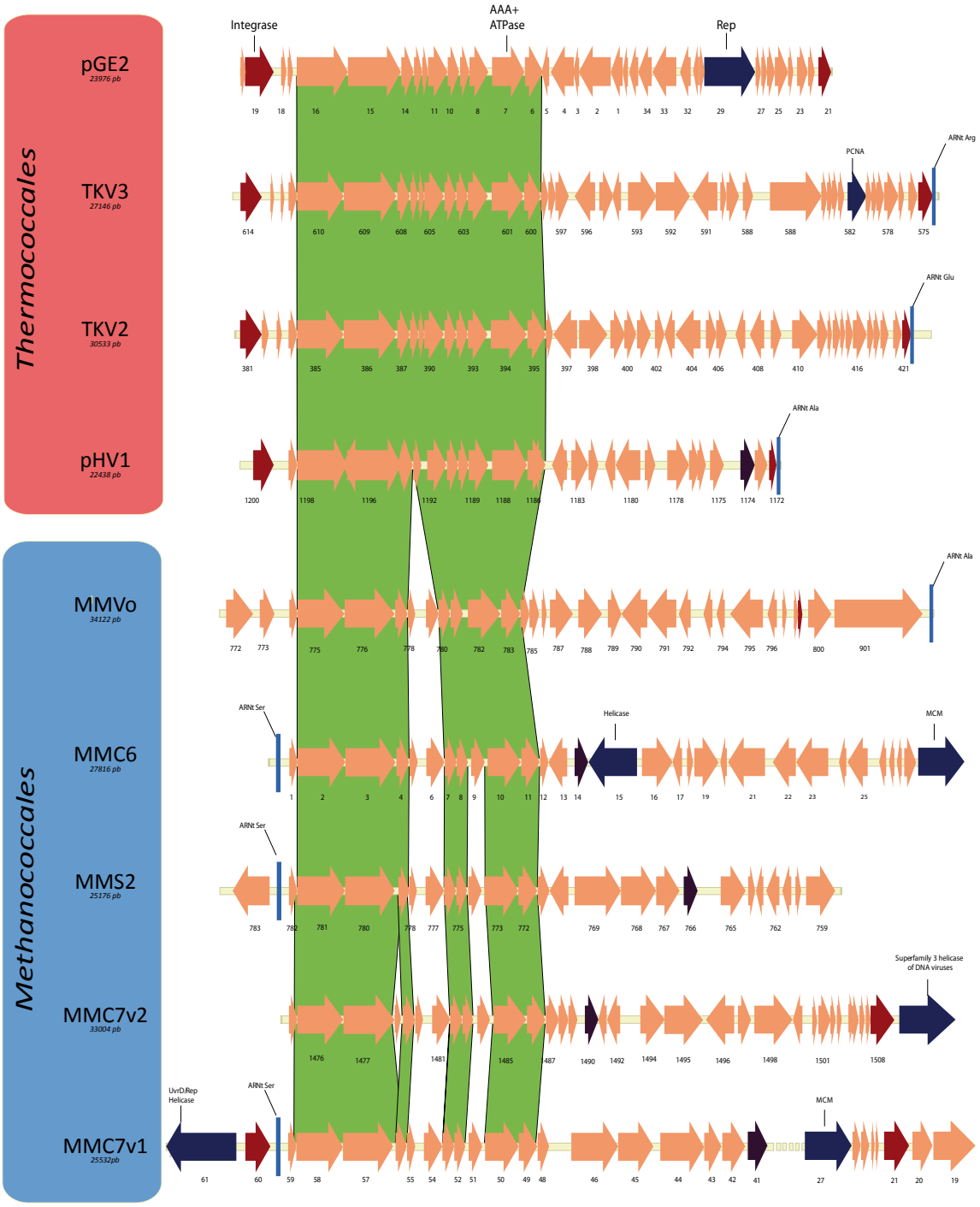


Figure 37 Génomique comparée de pGE2 et des éléments intégrés homologues

Représentation schématique du plasmide pGE2 de *P.abyssi* GE2 avec les éléments intégrés de *T.kodakaraensis* (TKV2 et TKV3), *P.horikoshii* (PHV1), *M.voltae* (MMVo), *M.maripaludis* C6 (MMC6), *M.maripaludis* S2 (MMS2) et de *M.maripaludis* C7 (MMC7v1 et MMC7v2). Les flèches représentent les ORFs. Les ORFs bleus interviennent dans la réplication. Les ORFs violet sont des recombinases alors que les recombinases site-spécifiques de type intégrase sont en pourpre. Les barres verticales bleues correspondent à des ARNt. Les ORFs homologues sont reliés par un fond vert.

Tableau 27 ORFs portés par le plasmide pGE2

Elements intégrés									ORF	Taille	igi	TMH	Fonction	BLAST				My Hits	Interpro	Pi	AA				Commentaire
MMC5	MMC6	Mvo	MMS2	MMC7V1	MMC7V2	TKV2	TKV3	Phor						ORF	Org	eval	Id				H	A	B	P	
x	x	x	x	x	x	x	x	x	1	163		0	Resolvase	Resolvase related protein	*P.abyssi*	2e-88	100	pfam_fs:Resolvase Resolvase, N terminal domain 15-131	PS00397 Recombinase PF00239 Resolvase Nterm domain	9,3	41,72	15,3	19	22,7	Le gene parA est une résolvase de plasmide d'E.coli
x	x	x	x	x	x	x	x	x	2	429			Transposase	Transposon ISC913	*P.abyssi*	0	100	pfam_ls:Transposase_35 Putative transposase DNA-binding domain 302-369 ressemble à un Zn finger	PF07282: Transposase_35	10,29	37,76	9,56	21,5	21,5	
x	x	x	x	x	x	x	x	x	3	50										9,52	40	16	22	10	
x	x	x	x	x	x	x	x	x	4	305										9,3	18,76	10,5	13,1	22,2	
x	x	x	x	x	x	x	x	x	5	95	+	2								8,03	13,68	4,21	5,26	25,8	
x	11	783	772	49	1486	395	600	1186	6	243				TK0600	*T.kodakaraensis*	4e-86	74	pfam_fs:HSF_DNA-bind HSF-type DNA-binding 167-188		7,11	38,68	16,1	16,1	17,7	
x	10	782	773	50	1485	394	601	1188	7	436			Helicase UvrD	TK0601 ATPase, AAA	*T.kodakaraensis*	0	74	prf:UVRD_HELICASE_ATP_BIND UvrD-like DNA helicase ATP-binding domain profile. 14-433		9,28	35,55	13,5	16,1	21,5	HerA helicase [Replication, recombination, andrepair] COG0433 pour PH1188 RuvB=""Holliday junction resolvasome, helicase subunit COG2255 pr TK601
x	9	781	x	x	x	393	602	1189	8	259	+	2		PH1189 et TK602	*P.horikoshii*	9e-67	54	pfam_fs:bZIP_1 bZIP transcription factor pfam_fs:Microtub_assoc Microtubule associated		9,79	31,27	10,8	17	27,4	Annoté smc chez *M.aeolicus*
x	x	x	x	x	x	392	603	1190	9	128	+	1		TK0392	*T.kodakaraensis*	1e-22	56		PS51257: ProK mB lipoprot lipid attachment site	5,48	26,56	10,9	10,2	30,5	TK0392 et TK603 sont plus courtes mais plus proches alors que PH1190 fait la meme taille mais plus divergent
x	8	780	775	52	1483	391	604	1191	10	157	+	1		TK0694 et TK0391	*T.kodakaraensis*	8e-49	57			6,35	33,12	12,1	11,5	26,8	
x	x	x	x	x	x	390	605	1192	11	267	+	2		TK605	*T.kodakaraensis*	6e-50	50			4,37	24,72	11,6	6,37	37,5	
x	x	x	x	x	x	389	606	1192.1n	12	77	+	2		TK606	*T.kodakaraensis*	2e-17	68			8,39	15,58	5,19	6,49	22,1	
x	x	x	x	x	x	388	607	1193	13	111	+	4		TK0607 et TK0388	*T.kodakaraensis*	4e-36	66			6,75	15,32	5,41	5,41	17,1	
x	47	777	776 779	53 56	1482 1479	387	608	1194	14	165	+	5		TK0387 et TK0698	*T.kodakaraensis*	5e-55	66			6,93	18,18	6,06	6,06	23,6	
1269	3541	1230	780	57	1477	386904	609	1196	15	716	+	4		PH1196	*P.horikoshii*	0	59			5,32	28,63	12	9,92	24	ORF15 et PH1196 possède un domaine Nterm supplémentaire (hérité de geobacillus?)
1270	3540	1229	781	58	1476	387903	610	1198	16	692	+	2		PH1198	*P.horikoshii*	0	62	pfam_fs:PEGA domain 316-350 pfam_fs: Phage tail fibre repeat 438-451		5,12	25,43	10,6	8,09	33	PH1198 et ORF16 possède un domaine Nterm supplémentaire (aucun résultat en Blast)
x	x	x	x	x	x	x	x	x	17	58										10,19	44,83	15,5	30,4	15,5	
x	x	x	x	x	x	x	x	x	18	64								pfam_ls:RHH_1 Ribbon-helix-helix protein, copG family 19-57		7,94	37,5	9,38	21,9	20,3	
x	x	x	x	x	x	381	x	1200 1172 1863	19	384			Integrase complete	PH1200 Cterm of integrase	*P.horikoshii*	3e-118	69			9,72	36,2	11,7	16,9	22,1	L'intégrase est complete, accroche les N et Cterm dans les génomes proches et la totalité de SSV1
x	x	x	x	x	x	372	x	x	20	74				TK0372	*T.kodakaraensis*	1e-3	42	pfam_fs:UPF0150 Uncharacterised protein family (UPF0150) 3-52		4,45	32,43	16,2	8,11	18,9	HRDC domain (Helicase and RNase D C-terminal) domain has a putative role in nucleic acid binding. pfam00570"" Homologue lointain mais dans région active
x	36110418	x	x	x	x	x	x	x	21	175			Integrase	TK0104 integrase/recomb	*T.kodakaraensis*	5e-84	86	pfam_ls: Phage integrase family 23-173	PF00589: Integrase, catalytic core, phage	9,8	39,43	10,3	18,9	28,6	
x	x	x	x	x	x	x	x	x	22	87										8,1	35,63	14,9	16,1	18,4	
x	x	x	x	x	x	x	x	x	23	140										9,77	38,57	11,4	19,3	17,1	
x	x	x	x	x	x	x	x	1175	24	72										4,61	36,11	19,4	12,5	18,1	
x	x	x	x	x	x	x	x	x	25	176										9,46	39,2	13,1	16,5	22,2	
x	x	x	x	x	x	x	x	x	26	106	+									10,32	33,02	12,3	17,9	14,2	
x	x	x	x	x	x	x	x	x	27	50										5,05	42	22	18	16	
x	x	x	x	x	x	x	x	x	28	74										10,07	36,49	10,8	20,3	14,9	
x	x	x	x	x	x	x	587	x	29	688			Rep	TK0587	*T.kodakaraensis*	1e-178	49	pfam_ls: ATPase family (AAA) 285-473 pfam_fs: Domain of unknown function (DUF771) 504-518		5,34	35,76	16,1	14,2	21,4	Egalement homologue à RepA de pTIK4 et pORA1 la primase polymerase!
x	x	x	x	x	x	x	x	x	30	56										9,4	42,86	17	21,4	21,4	
x	x	x	x	x	x	x	x	x	31	54										9,92	35,19	3,7	14,8	33,3	
x	x	x	x	x	x	x	x	x	32	149		2								9,78	14,09	2,68	6,71	27,5	
x	x	x	x	x	x	x	596	x	33	321	+	2							PF01105: emp24/gp25L/p24 family/GOLD	9,4	30,22	9,35	13,7	21,5	
x	x	x	x	x	x	x	x	x	34	182										8,61	26,37	9,89	11	24,7	
x	x	x	x	x	x	x	x	x	35	123										9,56	26,46	9,76	13,8	20,3	
x	x	x	x	x	x	x	x	x	36	52										9,04	34,62	11,5	15,4	26,9	

3.3 *pGE2 est un plasmide intégratif*

Chez pGE2, l'**ORF19**, code une **intégrase** de la famille des tyrosines recombinases. Elle est homologue aux intégrases ayant catalysé l'intégration des îlots génomiques homologues à pGE2 qui produisent une partition de l'intégrase en deux fragments, N-terminal et C-terminal, bordant l'élément intégré. La fonctionnalité de l'intégrase a été étudiée en se basant sur les données disponibles pour le modèle SSV1, et en considérant que le génome de la souche *P.abyssi* GE5 est proche de celui de *P.abyssi* GE2. La séquence *attP* de l'intégrase est complémentaire d'un ARNt situé entre les ORFs PAB1658 et PAB1659 de *P.abyssi*. Des expériences de PCR, directes et indirectes, ciblant les séquences *attL* et *attR* ont confirmé la **capacité intégrative du plasmide pGE2** au sein de l'ARNt Ala (codon CGG), possédant la séquence *attB*. (Communication personnelle G. Erauso)

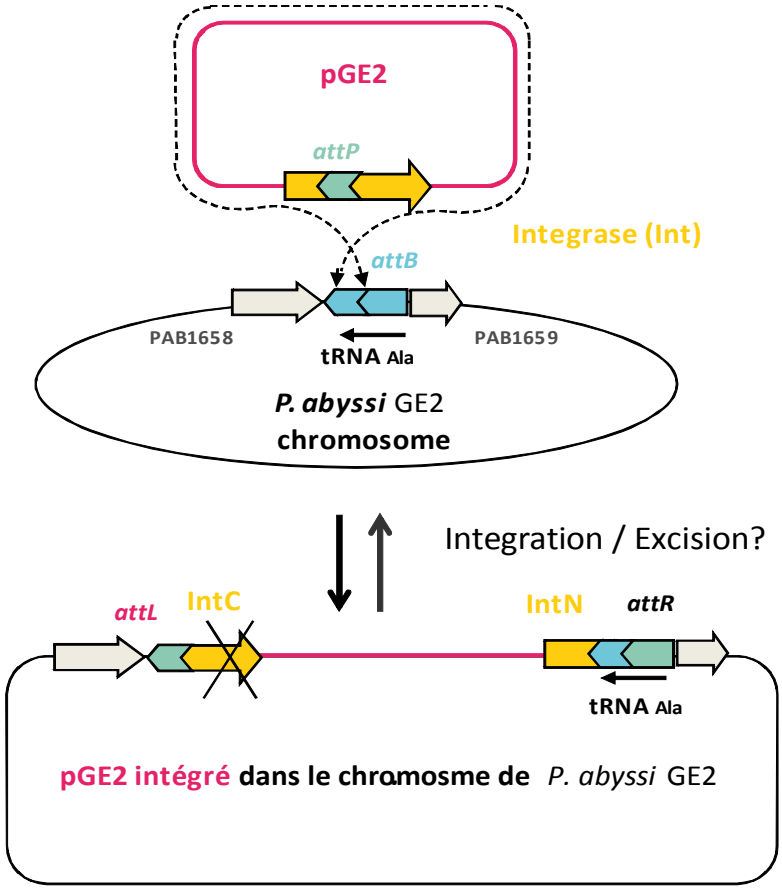


Figure 38 Intégration du plasmide pGE2 dans le chromosome de *P.abyssi* GE2

Schématisation de l'intégration du plasmide pGE2 dans le chromosome de son hôte *P.abyssi* GE2. Le plasmide, sous forme libre et intégrée, est représenté en rouge. Le chromosome de *P.abyssi* GE2 est en noir. Le gène codant l'intégrase est représenté par une flèche jaune, contenant une séquence d'attachement *attP* (vert). La cible de l'intégration est l'ARNt Ala, il est représenté en bleu.

3.4 *Les gènes conservés, relation avec les Adénovirus*

Les ORFs 6 à 16 de pGE2 sont conservés dans tous les éléments intégrés. A l'exception des ORFs 6 et 7, tous ces gènes codent des protéines membranaires (Tableau 27 et Figure 37).

L'**ORF7** code une protéine de 436AA qui possède un domaine ATPase AAA+ et dans lequel on trouve les motifs typiques Walker A et Walker B à l'extrémité N-terminale (Figure 39). La recherche d'homologues dans les bases de données en BlastP ne permet pas d'attribuer une fonction à cette ATPase. Une recherche itérative en utilisant PSI Blast permet d'isoler les ATPases impliquées dans l'empaquetage de l'ADN, un processus indispensable à la formation de particules virales chez les adénovirus. Bien que cette protéine et ses homologues aient des tailles plus importantes, l'alignement de séquence confirme l'existence du motif P9/A32 typique des ATPases d'empaquetage de la famille des adénovirus (Burroughs *et al.,* 2007).

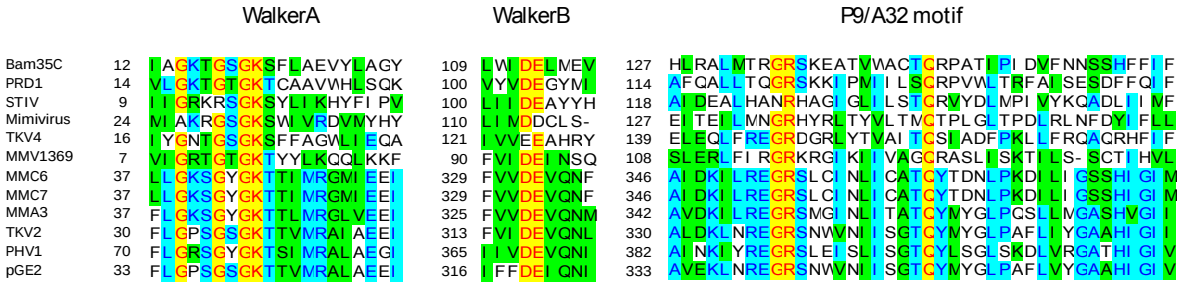


Figure 39 Alignement de l'ORF 7 de pGE2 avec les ATPases AAA+

Alignement des motifs ATPases (Walker A et Walker B) et du motif d'empaquetage P9/A32 de différents virus : Bam35c chez *Bacillus thuringiensis*, PRD1 chez *Escherichia coli*, STIV chez *Sulfolobus solfataricus*, Mimivirus ainsi que pGE2 et ses homologues codés par les éléments viraux intégrés.

L'**ORF8** code une protéine de 259 AA apparentée à une sous-unité de la préfoldine (PF02996) et possédant un ancrage membranaire. Les préfoldines sont des chaperonnes moléculaires qui capturent les intermédiaires protéiques et les transfèrent à d'autres chaperonnes assurant un repliement correct. Chez les *Archaea*, ce sont des hétérodimères $\alpha_2\beta_2$. Elles sont très étudiées pour leur rôle dans la formation du cytosquelette et leur implication dans la catalyse du repliement et de la polymérisation de l'actine lors de la formation des microtubules. La protéine codée par l'ORF8 est composé de deux domaines : un domaine d'association à l'actine (PF07989) et un domaine bZIP (PF0017à) servant à la dimérisation et à la fixation à l'ADN. Les homologues présents chez les Methanococcales ont des tailles sensiblement plus petites (150AA) en raison d'une partie N-terminale plus courte. Toutefois, la similarité de séquence est localisée dans la

partie C-terminale. Une telle protéine pourrait intervenir dans la ségrégation du génome extra chromosomique lors de la division cellulaire et/ou dans hypothétique assemblage de virion.

L'**ORF14** code une protéine de 165AA possédant 5 segments transmembranaires. Aucun motif ne permet de prédire une fonction. Cependant une surprenante observation suscite l'attention. Au sein des Methanococcales, cet ORF est « dupliqué » de part et d'autre d'un groupe de gènes qui n'est pas présent chez les Thermococcales (Figure 37). L'analyse phylogénétique suggère une évolution indépendante de chacun des deux paralogues (Figure 40). Ce phénomène de duplication-divergence ne semble pas être un évènement récent. La position phylogénétique de cette protéine dans les deux éléments intégrés de *M. maripaludis* C7 permet également de formuler une hypothèse sur l'origine des deux éléments de cette souche. Cette souche n'a pas été infectée deux fois par le même virus, mais par deux virus apparentés.

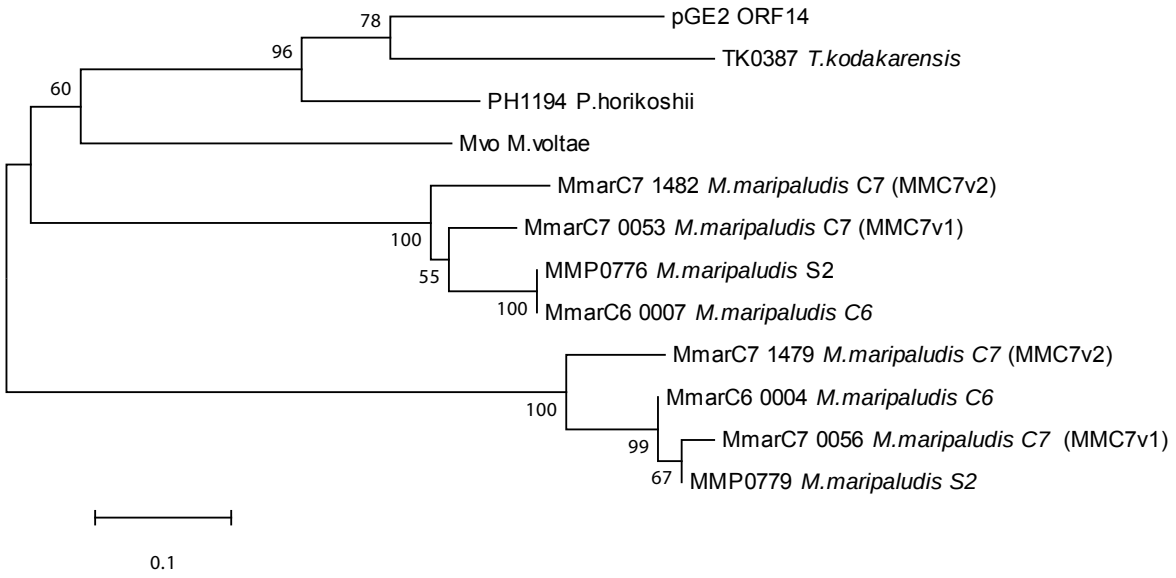


Figure 40 Arbre phylogénétique de l'ORF14 de pGE2

Les **ORF15** et **16** codent les deux plus grosses protéines du plasmide, respectivement 716AA et 692AA. Ces deux protéines membranaires ne sont pas uniquement rencontrées dans les différents îlots génomiques. En effet, on les retrouve dans le génome de *M.voltae* A3 (MVo 1238 et 1237), dans celui de *M.maripaludis* C6 (MmarC6 0541 et 0540), dans le génome de *M.maripaludis* C7 (MmarC7 0904 et 0903). Curieusement, le génome de *M.maripaludis* C5 ne possède pourtant pas ce type d'élément intégré (MmarC5 1269 et 1270). On remarque également que les homologues de ces deux ORFs sont systématiquement localisé l'un à côté de l'autre, comme dans les îlots génomiques.

La phylogénie de l'ORF16 est congruente avec celle de l'ORF15 (Figure 41), ce qui renforce l'idée selon laquelle ces deux ORFs sont étroitement associés.

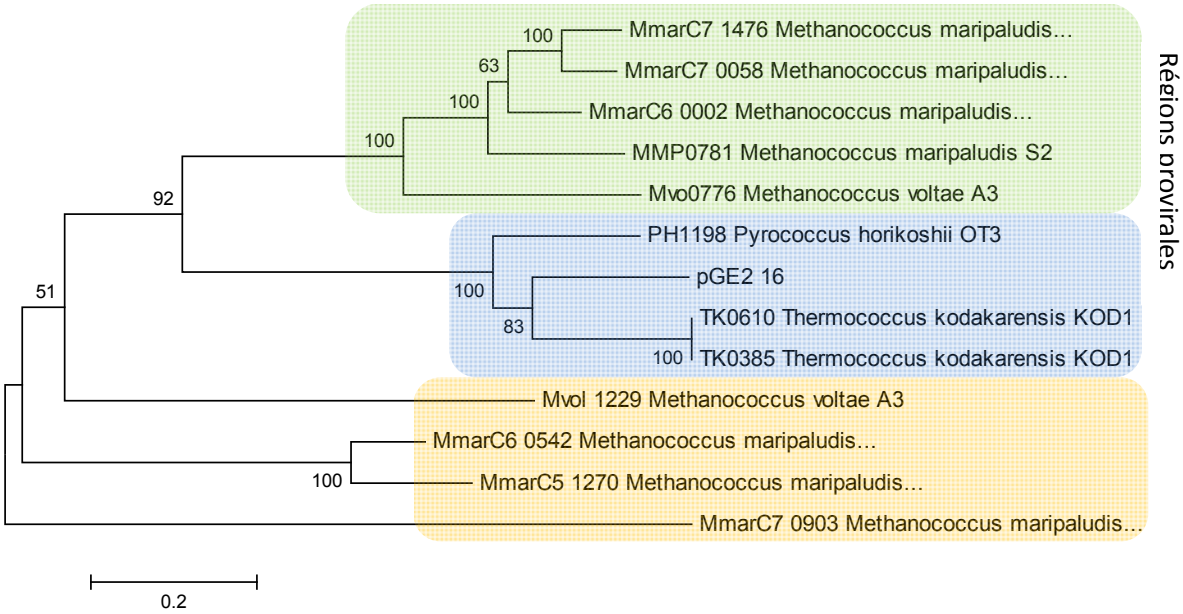


Figure 41 Phylogénie de l'ORF16 (congruente à celle de l'ORF15)

Cette analyse phylogénétique permet également de séparer ces protéines en deux groupes : d'un côté, les protéines codées par pGE2 et les îlots génomiques, et de l'autre côté les homologues chromosomiques situés en dehors des îlots génomiques. De plus, la distribution des homologues sur les éléments « mobiles » est séparée en fonction de l'affiliation aux Thermococcales ou aux Methanococcales. Cette observation suggère une évolution divergente entre les éléments mobiles et leurs homologues chromosomiques.

L'alignement de la protéine 15 avec ses homologues montre que seuls les 100 premiers AA de la région N-terminale sont conservés au sein des Thermococcales. Cette région est totalement différente chez les homologues de Methanococcales. Elle contient un motif antigène de surface (PFam 05453) essentiellement rencontré chez les *Bacteria*, chez quelques virus à ADN et chez *Sulfolobus* (Figure 42). Les prédictions de topologie suggèrent une localisation extracellulaire de ce domaine. Cette protéine pourrait servir de cible de reconnaissance par d'autres protéines codées par le plasmide pGE2.

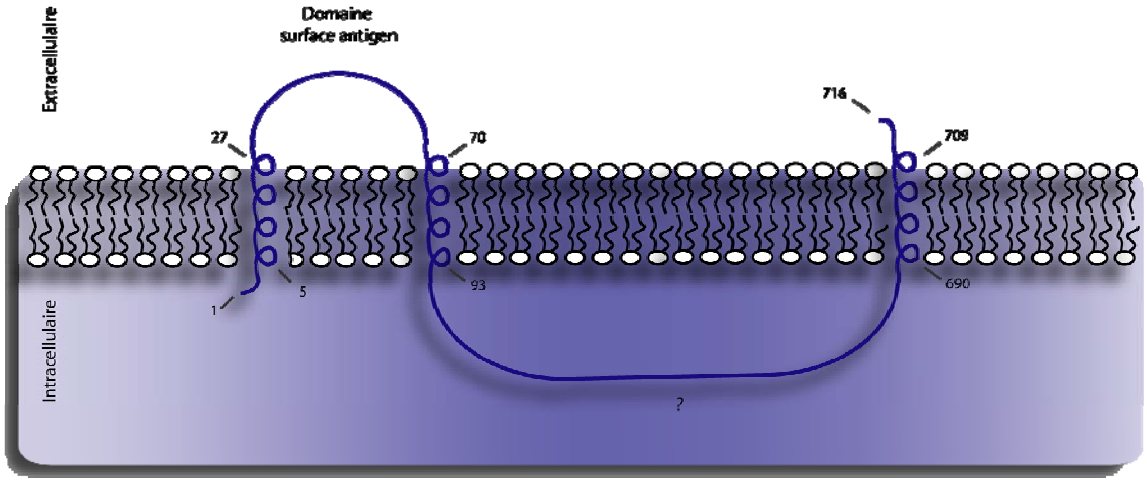


Figure 42 Schéma de la topologie de la protéine codée par l'ORF15 de pGE2

L'ORF16 possède deux domaines identifiables. Le domaine PEGA (PF08308) est situé dans la partie centrale de la protéine, entre les positions 310 et 350. Ce genre de domaine est rencontré chez les bactéries et certaines Euryarchaea. Il possède des similarités avec les protéines du S-layer (surface layé). La structure secondaire de ce domaine (Adindla *et al.,* 2004) adopte une conformation en brins beta favorisant les interactions entre protéines. De plus, un motif Phage_fiber (PF03335) des phages lambda et T4 est déterminé entre les positions 316 et 350.

Les ORF15 et 16 pourraient agir de concert en créant par interaction de type protéine-protéine. Les cibles pourraient être des protéines appartenant à la cellule hôte, ou dans l'hypothèse phagique intervenir dans la reconnaissance de la cellule cible.

3.5 *Les gènes non conservés – implication dans la réplication.*

L'ORF29 code une protéine de 788AA. Bien que cette protéine ne fasse pas partie de la région conservée des éléments intégrés, un homologue est rencontré dans le génome de *T.kodakaraensis*, TK0587, codant une protéine de fonction inconnue. La recherche d'homologues dans les bases de données permet également de trouver des similarités de séquence avec des protéines de tailles plus importantes : RepA des plasmides pORA1 (866AA) et pTIK4 (1053AA) de *Sulfolobus*. Ces protéines initiatrices de la réplication possèdent deux domaines. L'activité hélicase est portée par l'extrémité C-terminale alors que l'extrémité N-terminale possède la fonction ADN polymérase/primase. L'alignement de l'ORF29 de pGE2 avec ces protéines RepA explique la taille plus restreinte de notre protéine : le domaine Nterm possédant l'activité ADN

primase/polymérase n'est pas présent. La région alignée s'étend néanmoins sur l'ensemble de notre protéine.

Figure 43 Alignement de RepA de pGE2 avec les homologues de plasmides de *Sulfolobus*

3.6 *Discussion générale*

pGE2 est l'élément génétique de Thermococcales le plus intéressant, et ceci à plus d'un titre. Il possède un nombre de gènes orphelins moindre, par rapport aux autres éléments génétiques, permettant la détection de nombreux homologues et de suggérer des hypothèses sur son fonctionnement. Ses protéines, possédant des homologues, peuvent être annotées avec moins d'incertitude sur leur fonction.

La capacité d'interaction du réplicon extrachromosomique dans le chromosome tient en la présence d'une intégrase codée par le plasmide. Cette intégration de pGE2 dans le chromosome a été confirmée de manière expérimentale et procède selon un mécanisme similaire à celui décrit chez le virus SSV1 infectant *Sulfolobus shibatae*. D'autre part, nous avons remarqué une

diminution de la quantité de plasmide au fur et à mesure des repiquages de la souche en culture. Dans certains cas, le plasmide disparaissait totalement et nous obligeant à ressortir un cryotube de la souchothèque. Ce n'était pas une disparition totale mais un camouflage sous forme prophagique (vérifié par PCR). La culture « intensive » de la souche *P.abyssi* GE2 favorise l'intégration de pGE2 dans le chromosome. Il resterait ainsi protégé, se répliquant avec le chromosome de son hôte en attendant des conditions plus favorables pour s'exciser.

La capacité intégrative de ce type d'élément génétique a laissé de nombreuses traces de « virus » au sein des génomes. La génomique comparée montre qu'une grande région synténique de pGE2 est homologue à certains îlots génomiques de deux Thermococcales, *T. kodakaraensis* et *P. horikoshii* ; mais aussi dans les génomes de nombreuses Euryarchaea méthanogènes mésophiles.

Malgré des origines géographiques très distantes, la présence de ce groupe de gènes, possédant une organisation synténique conservée, est une information sur les gènes fonctionnant de concert afin de réaliser une action spécifique. La propriété des éléments génétiques étant leur propriété à contrôler leur réplication et se maintenir, il aurait été légitime que les gènes conservés soient impliqués dans cette fonction. Une comparaison globale des éléments montre que les gènes imputés à la maintenance du génome sont situés en dehors du large bloc de gènes conservés, et qu'ils font intervenir différents mécanismes de réplication. Cette observation peut aisément être mise en parallèle avec celles effectuées lors de la comparaison de pAMT11 avec TKV1 de *T. kodakaraensis*. Bien que peu de gènes conservés soient annotables, la présence d'une ATPase d'empaquetage de l'ADN spécifique des Adénovirus permet d'affilier ce groupe de gènes à la production de particules virales. L'existence d'adénovirus chez les Euryarchaea a récemment été proposée par la découverte d'une ATPase d'encapsidation homologue présente dans l'îlot génomique TKV4 de *T.kodakaraensis* (Krupovic *et al.,* 2008). Les homologues de cette ATPase d'encapsidation présent dans TKV2, TKV3 (et pGE2) n'avaient pas été mis en évidence lors de l'analyse à grande échelle effectuée par Krupovic à cause d'un filtre sur la taille des protéines. Cette découverte fait de pGE2 le premier élément génétique d'Euryarchaea affilié aux Adénovirus. A l'heure actuelle, aucune particule virale n'a été observée. Ceci pourrait s'expliquer par les conditions de cultures utilisées. En effet, le criblage de recherche de particules virales est effectué à partir de cultures en phase exponentielle de croissance. La production de particule virale nécessite peut-être une induction provoquée, par exemple, par un agent chimique ou tout simplement par une température optimale de production de particules virale différente de celle optimale pour la croissance de l'hôte. Les Thermococcales ayant un temps de génération très court, il ne serait pas improbable que la production de particules virales soit optimale lorsque la

souche se divise beaucoup moins vite, laissant le temps au virus de mettre en place l'usine virale (*viral factory*).

L'intégrase servant aussi bien à l'intégration qu'à l'excision, la même question peut se poser sur le « choix » pour l'élément génétique de rester préférentiellement/exclusivement sous forme libre ou intégrée. Il se pourrait qu'à la température optimale de croissance de la souche l'intégration soit favorisée, permettant au génome viral de se protéger en s'hébergeant dans le chromosome. A température plus faible, l'excision est peut-être favorisée afin de permettre la production de particules virales.

La présence d'éléments viraux intégrés affiliés à pGE2 chez les Thermococcales et les Methanococcales suppose l'existence d'un ancêtre viral commun ou une capacité toujours active de transferts horizontaux de gènes. Malgré des modes de vie différents, ces deux types d'organismes sont phylogénétiquement proches et possèdent des caractéristiques membranaires communes. L'importante quantité de protéines membranaires sur ces éléments suggère un mécanisme permettant à l'élément génétique de s'ancrer à la membrane puis de s'en affranchir sous la forme d'un virus enveloppé comme cela a été observé chez de nombreux virus d'*Archaea* tels que PSV, Pyrobaculum Spherical Virus (Haring *et al.,* 2004), ARV1, Acidianus RudiVirus1 (Vestergaard *et al.,* 2005), ou phiCh1 de *Natrialba magadii*. (Klein *et al.,* 2002)

En poursuivant ce raisonnement, je suis très intrigué par l'observation de vésicules de types virales, contenant de l'ADN « aléatoire » d'environ 4,5kb, aussi bien chez les Thermococcales (Soler *et al.,* 2008) que chez les Methanococcales (Bertani 1999; Eiserling *et al.,* 1999). Un mécanisme similaire existe chez les bactéries : le système GTA, *Gene Transfer Agent*, très étudié chez *Rhodobacter capsulatus* (Lang *et al.,* 2000). Il s'agit d'une relique de prophage ayant évolué de manière indépendante (dompté ?) dans le génome. Des senseurs du cycle cellulaire (kinases) activent le système GTA lorsque la cellule entre en phase stationnaire. La conséquence est la production de particules virales de petite taille encapsidant de l'ADN de manière aléatoire. Ces observations pourraient être reliées à la présence des îlots génomiques apparentés à pGE2. Ils pourraient contenir les gènes nécessaires à la production de vésicules, trop petites pour contenir l'ADN viral.

3.7 *Le transposon actif de pGE2*

pGE2 contient également un **transposon actif**. Pour obtenir le génome de pGE2, une première banque d'ADN a été réalisée et séquencée. L'assemblage de ces séquences ne permettait pas d'avoir une couverture du génome suffisante. Une seconde banque a été réalisée et séquencée. L'ajout des nouvelles séquences avec le précédent contig posait un problème au niveau d'une région précise. Un réassemblage complet des séquences montre une modification du génome du plasmide suite à l'insertion d'une séquence de 1865 pb. A l'exception d'un mésappariement, la recherche de similarité dans les bases de données par blastN produit un alignement parfait avec une région du génome *P.abyssi* GE5 codant les protéines PAB2076 et PAB2077. Cette séquence supplémentaire est un transposon non composite, codant les protéines TnpA et TnpB. Ces protéines sont respectivement une **transposase** et une **recombinase** à sérine catalysant la recombinaison spécifique entre deux copies de l'élément et résolvant les coïntégrats en achevant la transposition.

L'étude de ces protéines permet d'affilier cette séquence d'insertion (IS) aux transposons non DDE IS*607*, de la famille IS*200*/IS*605*/IS*607*. Ce groupe possède une grande hétérogénéité et une organisation complexe (Chandler *et al.,* 2002; Ton-Hoang *et al.,* 2005) (Figure 45). Cette famille d'IS est très atypique. Leurs extrémités ne sont pas bordées par des répétitions inverses (IRs) servant de site de fixation de la transposase ; elles sont asymétriques et possèdent des palindromes imparfaits générant des structures secondaires de type tige-boucle reconnues par la transposase. La transposition ne s'effectue pas vers un site aléatoire mais vers un tétra ou penta-nucléotidique spécifique (Kersulyte *et al.,* 1998) et ne provoque pas de duplication du site d'insertion. L'absence de nucléotide délété ou ajouté évite la perte de matériel génétique ainsi que la nécessité d'effectuer une réparation de l'ADN pour la cellule (Ton-Hoang *et al.* 2005). De nombreux éléments ne comportant qu'un des 2 gènes *tnpA* ou *tnpB* sont présents dans les bases de données. Deux transposases non-homologues, codées par *tnpA* permettent de discriminer le groupe IS*605* (*tnpA1*) et IS*607* (*tnpA2*). Celles du groupe IS*607* ressemblent aux recombinases S-site-spécifique (Grindley 2002) alors que celles du groupe IS*605* ressemblent aux protéines Rep, impliquées dans l'initiation de la réplication par cercle roulant (RCR) chez certains plasmides ou bactériophages simple brin (Curcio *et al.,* 2003) ou aux relaxases impliquées dans l'initiation de la conjugaison. Les structures secondaires de type-boucle reconnues par la transposase ressemblent aux SSO et DSO reconnues par les protéines de RCR. Les similitudes avec le mécanisme RCR viennent d'être étendues en élucidant ce mécanisme de transposition particulier (Barabas *et al.,* 2008). La transposition s'effectue par coupure d'un brin spécifique à proximité des structures tige-

boucle conduisant à la formation d'un intermédiaire simple brin circularisé, rappelant étrangement le mécanisme réplicatif des EG à RCR (Barabas *et al.,* 2008; Guynet *et al.,* 2008).

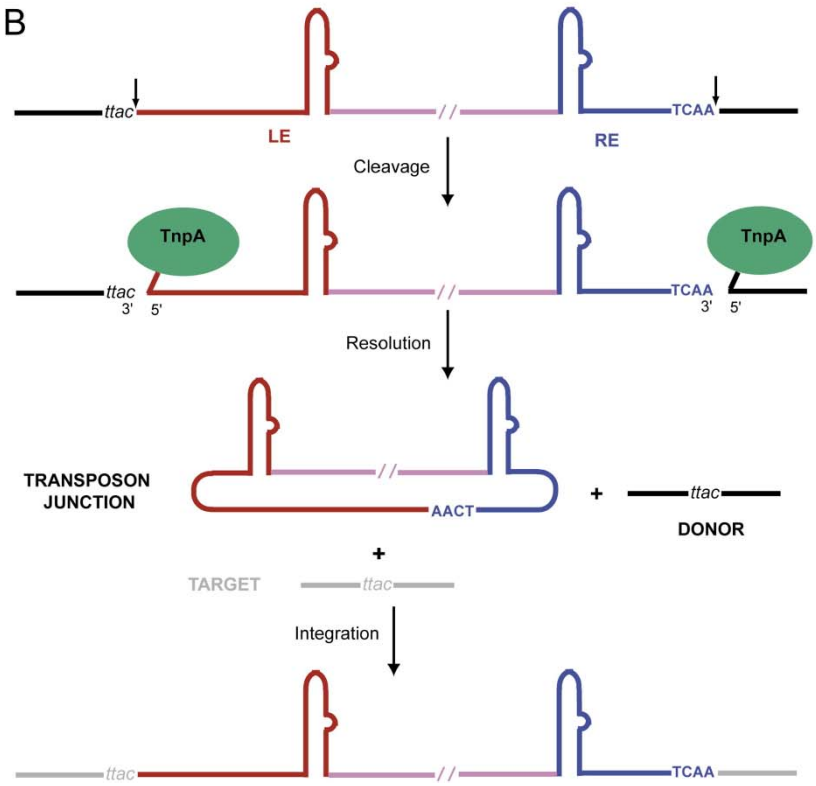


Figure 44 Mécanisme de transposition des IS200/IS605

De nombreuses séquences de cette famille sont rencontrées dans les génomes d'*Archaea* (Filee et al., 2007). Les génomes de *Sulfolobus* possèdent 6 éléments de ce type : IS*C1904* et IS*C1913* de *S.solfataricus* ; IS*C1926* de *S.* sp souche L0011 ; IS*Sto11*, IS*Sto12* et IS*Sto13* de *S.tokodaii* ; ainsi que IS1921 dans le génome d'*Acidianus ambivalens*. Quelques MITE (*miniature inverted-repeat transposable,)* éléments dérivés de IS*Sto12* et IS*Sto13,* sont également rencontrées dans le génome de *S.tokodaii*. Ces ISs ne sont pas uniquement présentes chez les Crenarchaea, elles sont également rencontrées de façon partielle dans les génomes de *M.jannaschii* (MG0012m/14) et *T.volcanium* IS*Tvo1*. Un seul élément complet est rencontré dans les génomes de Thermococcales : IS*Tko1* dans *T.kodakaraensis* et IS*Pfu4* dans *P.furiosus* bien qu'il existe des ISs partielles dans les génomes de *T.kodakaraensis* (TK1841/1842), *P.furiosus* (PF1985/1986) et *P.abyssi* GE5 (PAB2076/2077).

L'unique différence entre le transposon de *P.abyssi* GE5 (PAB2076/2077) et celui du plasmide pGE2 produit la mutation L371P dans le gène de la transposase, un acide aminé non-essentiel à la

fonction de transposition. L'histoire du cheminement « récent » du transposon de pGE2 peut facilement être reconstituée. La proximité des souches *P.abyssi* GE2 et GE5 impute l'origine du transposon au chromosome de *P.abyssi* GE2.

Les transposons peuvent être considérés comme le paroxysme de la théorie du gène égoïste de Dawkins, considérant le gène comme une unité fondamentale du vivant ; le génome étant une communauté de gènes vivant de manière plus ou moins symbiotique et possédant son lot de « cheater ». Ce transposon est constitué de deux gènes indispensables à la transposition. Néanmoins, retracer une histoire à long terme de cet élément est ardu de part sa nature composite et sa capacité à recombiner avec des éléments de même nature. Les considérations phylogénétiques séparées de chacun des deux gènes ne sont pas identiques. Alors que l'*orfB* est universellement présent dans les transposons de cette famille, l'*orfA* peut être une des deux protéines non homologue, *orfA1* pour IS*605* et *orfA2* pour IS*607*. La phylogénie de l'orfB (Figure 45A) montre un mélange de séquences bactériennes et archéennes et suggére une histoire évolutive complexe. La phylogénie de l'*orfA1* présente également un mélange de séquences entre les deux règnes. D'autre part, la phylogénie de l'*orfA2*, auquel appartient le transposon de pGE2, est monophylétique et favorise une hypothèse de transmission verticale de ces gènes. Prises de façons conjointes, ces observations indiquent l'existence de divers évènements de recombinaison entre les différentes copies des IS reflétant la nature composite et plastique de ces éléments. Le remplacement de l'*orfA* de IS607 par celui de IS605, et vice-versa, semble être un évènement très fréquent.

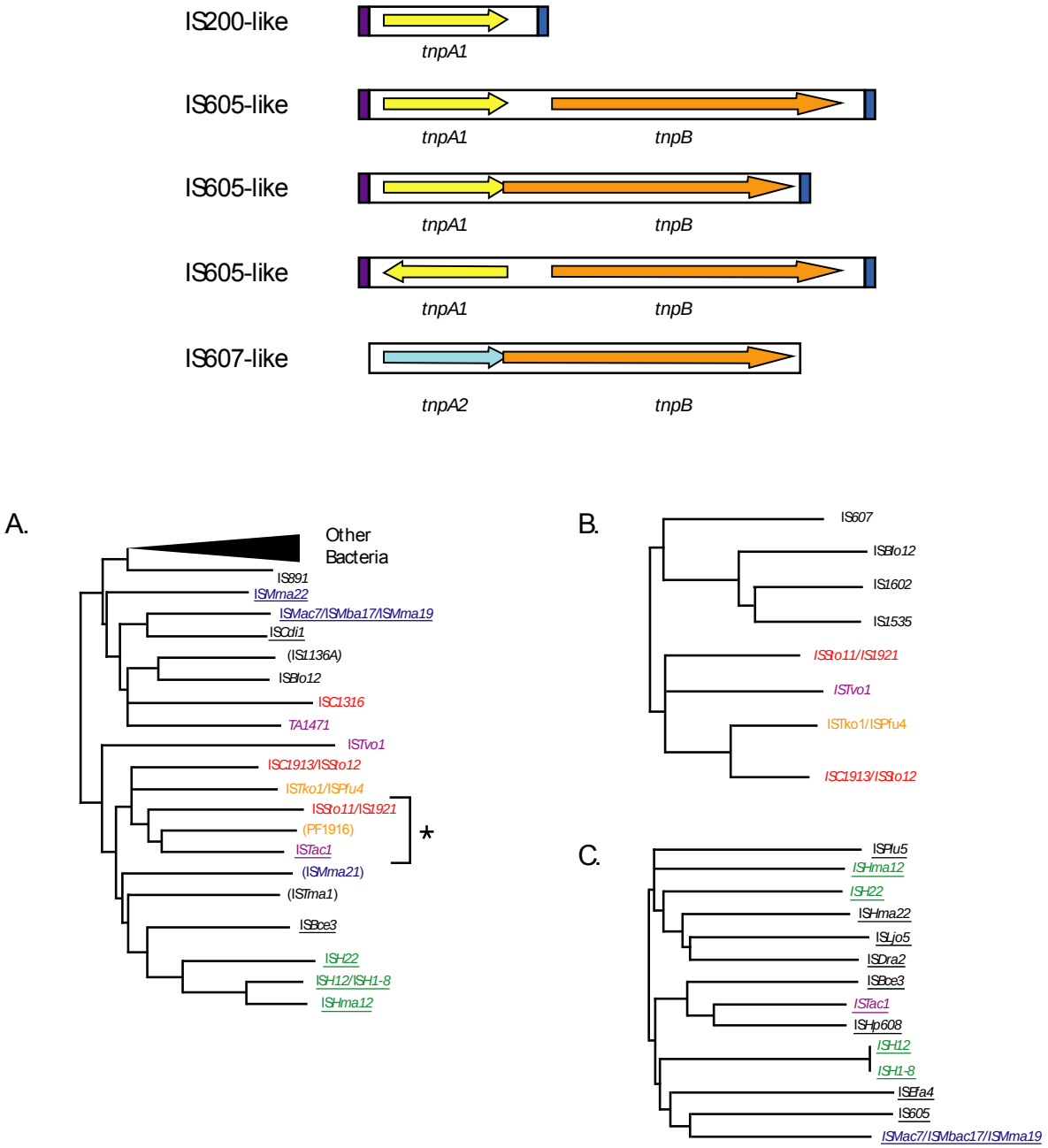


Figure 45 Famille des IS200/IS605/IS607

Le volet du haut montre l'organisation des différents membres de la famille. Jaune : tyrosine transposase TnpA1 ; Bleu : sérine TnpA2 ; Orange : TnpB de fonction inconnue. Les extrémités gauche et droite des transposons contenant TnpA1 sont respectivement en magenta et bleu. Ce ne sont pas des IRs mais elles ont le potentiel de former des structures secondaires. (A) Phylogénie de l'*orfB* de la famille IS200/IS605/IS607. (B) Phylogénie de l'*orfA1* famille IS607. (C) Phylogénie de l'*orfA2* famille IS605. Les éléments IS608 sont soulignés, les IS ne contenant que l'*orfB* sont entre parenthèses, l'astérisque indique une construction mosaïque. Les différentes Archaea ont été colorées : *Sulfolobales*, rouge ; *Thermoplasmatales*, magenta ; halophiles, vert ; méthanogènes : bleu ; autres : orange. Les *Bacteria* sont en noir.

Bien que l'absence d'IR aux extrémités d'un transposon composite soit courante, il faut noter l'existence de deux répétitions à l'extrémité 3' du gène codant la transposase (Figure 46). L'une

d'entre elles s'étend sur 30 pb et correspond à un palindrome imparfait : 4 mésappariements correspondant au pas d'un tour d'hélice sont localisés en son centre. La seconde répétition est à cheval sur l'extrémité 3' et la région intergénique, elle chevauche le codon stop de la transposase. Cet agencement de répétitions peut adopter une structure secondaire particulière servant de cible de reconnaissance de la transposase (Curcio *et al.* 2003).

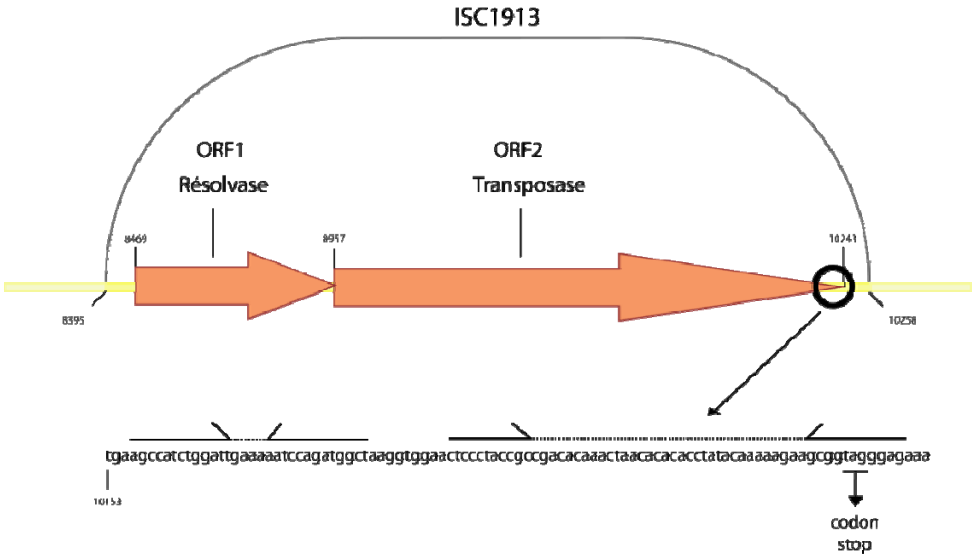


Figure 46 Transposon de pGE2

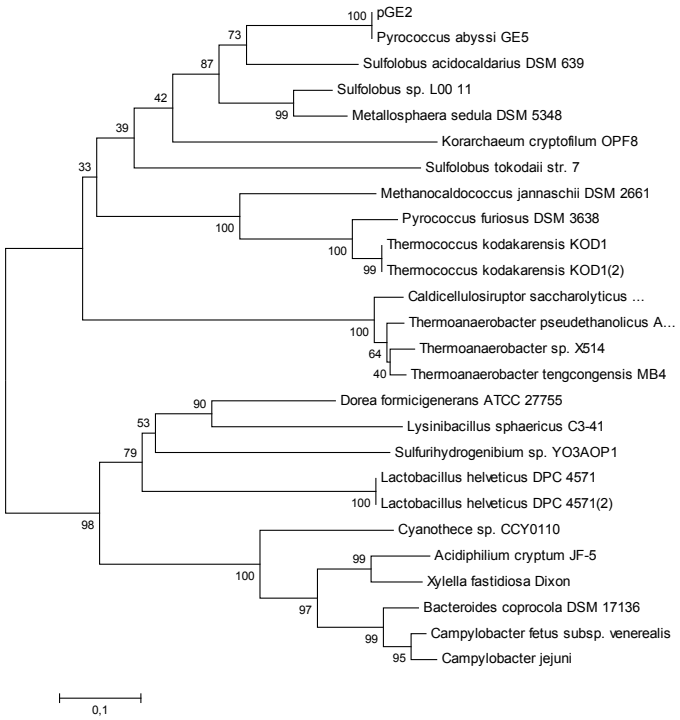


Figure 47 Arbre phylogénétique de la résolvase de pGE2.

4. pIRI42, un réplicon à grande plasticité ?

4.1 *Souche hébergeant pIRI42*

Ce plasmide est hébergé par la souche *Thermococcus* sp. IRI42, unique souche isolée de l'échantillon IR95-02 provenant du site Menez Gwen, situé au niveau de la ride médio-atlantique et de la triple jonction des Açores. Le ribotypage de cette souche la place à proximité des souches contenant le plasmide pIRI33, originaire du site voisin Rainbow, mais aussi de celle contenant pAMT11, isolée à partir d'un échantillon provenant de la ride pacifique est (Figure 21). Encore une fois, la détermination de l'espèce au sein des Thermococcales ne peut se cantonner à l'analyse du gène codant le 16S.

4.2 *Caractéristiques générales du plasmide pIRI42*

pIRI42 a une taille de 11924pb et une composition en G+C de 45,2%. 12 ORFs ont été déterminés, ils codent des protéines de tailles comprises entre 66 et 569AA. 6 sont portés par le brin direct et 6 par le brin complémentaire (Figure 48). Chose assez rare, presque la totalité des ORFs possèdent un RBS, seul l'ORF 2 en est dépourvu. L'organisation des ORFs ne respecte apparemment pas de règle de colinéarité, de nombreux gènes se retrouvent tête-bêche. A l'exception de celle codée par l'ORF5, aucune protéine ne semble excrétée, comme en témoigne l'absence de peptide signal et d'hélices transmembranaire. De plus, presque tous les ORFs annotés ont des homologues dans des génomes archéens génétiquement divers, aussi bien présents sur les chromosomes que les virus et les plasmides (Tableau 28).

Tableau 28 ORFs de pIRI42

ORF	Taille	P	TM	Fonction	BLAST ORF	BLAST Org	BLAST eval	BLAST Id	My Hits	Pi	AA H	AA C	AA A	AA B	AA P
1	569			Rep	putative Rep protein	*Pyrococcus sp* . JT1	8.10-26	26	pfam_ls:HTH_11 HTH domain 398-450	9,22	38,66	36,7	14,1	16,3	21,8
2	76									4,9	28,95	19,7	6,58	3,95	34,2
3	126				MJECL06 d'un plasmide	*Methanocaldococcus jannaschii*	5e-36	63	-	7,82	33,33	41,3	16,7	17,5	19,8
4	401	+	1				-		-	9,64	38,9	37,7	12,7	17,5	19,7
5	162				ATPase associated with various cellular activities	*Staphylothermus marinus* F1	8E-16	32	-	9,45	35,19	39,5	13,6	17,9	19,8
6	220			ADN glycosylase	pFV1_p10 Mismatch glycosylase du plasmide pFV10	*Methanothermobacter thermautotrophicus* pVF10	1e-53	45	pfam_ls: HhH-GPD superfamily base excision DNA repair protein 34-67	9,61	32,27	37,7	10,9	18,2	23,6
7	428				PAE3200	*Pyrobaculum aerophilum* str. IM2	0,4	24		7,22	35,05	34,6	14,5	14,5	22,4
8	336				PAE3200	*Pyrobaculum aerophilum* str. IM2	1	27		6,29	35,71	34	13,1	12,5	26,2
9	467			Cytosine Methylase	DNA (cytosine-5-)-methyltransferase	*Aeropyrum pernix* K1	6e-137	57	pfam_ls: C-5 cytosine-specific DNA me	9,3	36,4	34,9	13,9	16,1	18
10	66				Hypothetical protein APE_0871a.1	*Aeropyrum pernix* K1	1e-11	72	pfam_fs: Ribbon-helix-helix protein, copG family	9,7	32,84	31,3	11,9	17,9	25,4
11	156				leucine zipper and winged helix DNA-binding domain	*Pyrococcus abyssi* virus 1	3e-20	57	prf:MarR-type HTH XX-XXX Pfam:Cob_adeno_trans 13-110	9,2	34,96	46,6	20,6	25,2	14
12	150				MJECS06 et MJECL27 de plasmid	*Methanocaldococcus jannaschii*	0,084	26		9,5	29,33	40,7	14	20	23,3
13	79					-				9,75	49,37	15,2	0	12,7	25,3

Seuls les ORFs 2 et 4 sont orphelins. Leur localisation sur le génome ne permet pas d'affirmer qu'ils sont effectivement transcrits et traduits, en particulier l'ORF2 qui n'est pas inscrit dans une unité transcriptionnelle et ne possède pas de site de fixation du ribosome.

Une autre caractéristique de ce génome est la présence de nombreuses répétitions, directes et indirectes, parfaites et imparfaites (Figure 48). Ces répétitions sont majoritairement situées dans les espaces intergéniques à proximité du début ou de la fin de certains gènes, notamment au voisinage de l'ORF1 codant une protéine Rep initiatrice. La plus grande répétition, DRI, comprend trois unités répétées de 67pb situées dans les deux plus grands espaces intergéniques du génome, de part et d'autre de l'ORF3. D'autres répétitions sont situées à des places singulières, dans de courts espaces intergéniques à l'endroit où il y a inversion du biais cumulatif en G+C (DRIX), entre deux ORFs situées tête-bêche (DRX), ou à proximité immédiate de certains ORFs (DR II, III, VI, VIII et X Figure 48). L'annotation des gènes, discutée dans le paragraphe suivant, leur organisation non colinéaire et cette importante quantité de séquences répétées suggèrent une incroyable plasticité du génome de pIRI42.

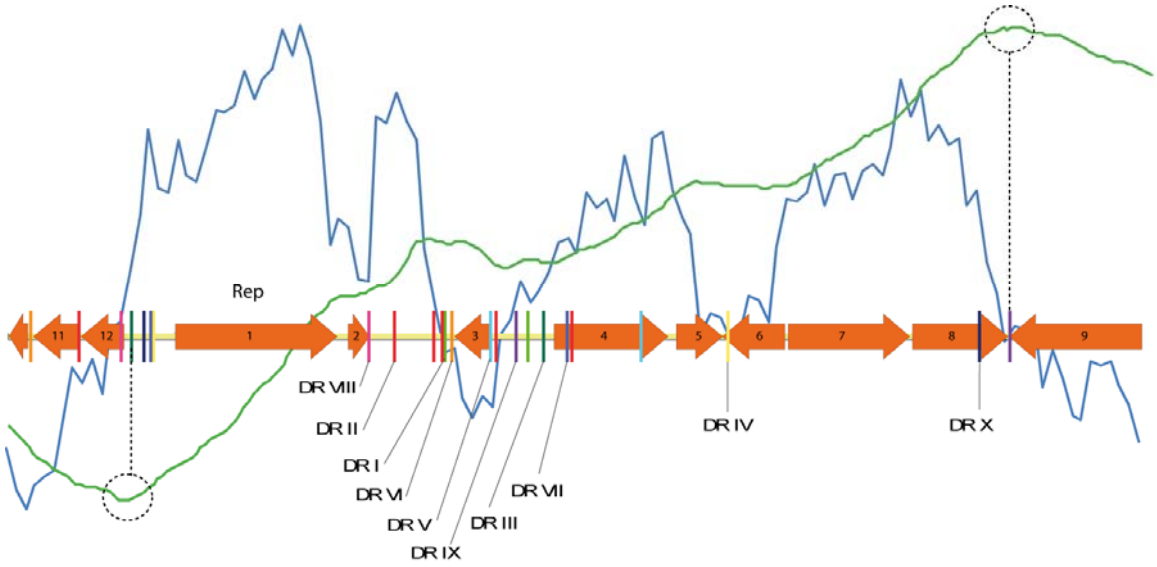


Figure 48 Biais cumulatif en GC et séquences répétées de pIRI42

La composition en GC et le biais cumulatif en GC sont représentés respectivement par la courbe bleue et la courbe verte ; les cercles pointillés correspondent aux points d'inflexion du biais cumulatif en GC. Les répétitions, annoté DRI au DRX, sont représentées par des barres verticales. Chaque couleur correspond à une répétition.

4.3 *Prédiction de la fonction des protéines codées par pIRI42*

L'**ORF1** code une protéine Rep, homologue à la protéine initiatrice de la réplication du plasmide pRT1 de la souche *Pyrococcus* sp. JT1 et à celle du plasmide pAMT11 (page 127). L'alignement de cette protéine avec ces deux homologues pose certains problèmes. Bien que le motif III déterminé lors de l'analyse de la Rep de pAMT1 soit confirmé, il est impossible de mettre en évidence le motif II (HxH) typique des protéines Rep. D'autres blocs de séquences identiques sont également mis en évidence, ils pourraient constituer un motif fonctionnel pour cette nouvelle famille de protéines Rep impliquées dans la réplication des « gros » plasmides, contrairement à Rep74/75 qui participent à la réplication des petits plasmides.

L'**ORF5** (162AA) code la seule protéine membranaire du plasmide. Les homologues les plus proches sont des ATPase AAA_5 de *Staphylothermus marinus* F1 et de *Pyrobaculum arsenaticum*. Ces protéines membranaires sont annotées ABC transporteurs (540AA), permettant un transport transmembranaire actif de nombreux composés par hydrolyse d'ATP. La protéine plasmidique s'aligne en totalité avec l'extrémité N-terminale du transporteur. La fonction ATPase présente en partie C-terminale ne peut être alignée, elle est donc absente de la protéine du plasmide.

L'**ORF6** code une protéine de 220AA ayant pour homologue *Mig.MthI*, l'ORF10 du plasmide pFV1 de *Methanothermobacter thermoautotrophicum* mais aussi la protéine APE_875 de *A.pernix* (Figure 50). Cette protéine est une ADN glycosylase, endonucléase III de la superfamille des « helix-hairpin-helix » (HhH). Ces nucléases impliquées dans la réparation de l'ADN sont présentes dans tous les domaines du vivant. Elles permettent la réparation de l'ADN présentant différents types de lésions, telles que les déaminations des résidus cytosines et 5-méthyle-cytosine fréquentes à haute température (Horst *et al.,* 1996), l'alkylation d'adénines ou des dimères de pyrimidines. La réparation se fait par un mécanisme nommé BER (Base Excision Repair). Ce mécanisme est très étudié car il est le premier à intervenir pour contrecarrer les effets potentiellement mutagènes et car il est similaire à celui des eucaryotes. Ce type de réparation est directement couplé à la réplication par interaction de l'ADN glycosylase avec le PCNA (Yang *et al.,* 2002). Le BER est (i) initié par l'ADN glycosylase coupant par hydrolyse la liaison glycosidique, suivi par (ii) une coupure du brin au niveau du site ADN venant de perdre sa base (iii) une synthèse d'ADN pour combler le trou (iv) et une ligature. Structuralement, ces glycosylases sont caractérisées par la présence d'un motif HhH et d'un cluster [4Fe-4S] ; la caractéristique commune est le basculement (flipping) de la base endommagée à partir de la double hélice vers

une cavité spécifique d'un type de nucléotide au niveau de l'enzyme. Ces caractéristiques sont partagées par les glycosylases des autres familles, telles que UDG de *Homo sapiens sapiens* (Kvaloy *et al.,* 2001) ou MUG de *E.coli et Deinococcus radiodurans* (Mokkapati *et al.,* 2001; Moe *et al.,* 2006).

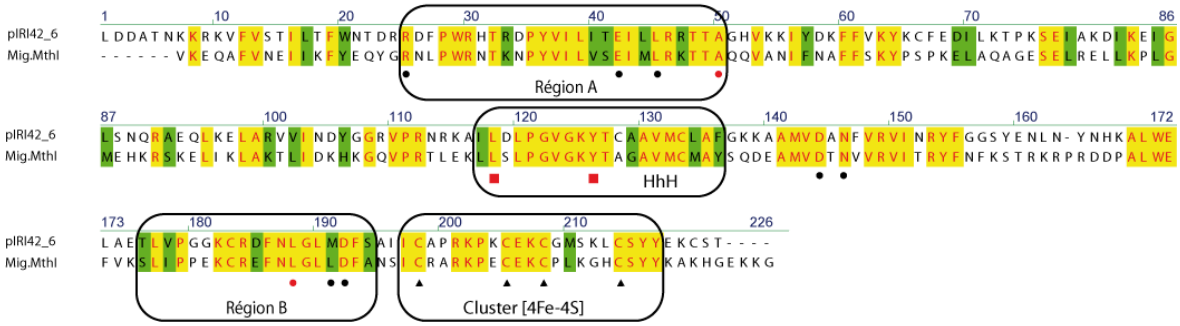


Figure 49 Alignement de l'ADN glycosylase de pIRI42 et de pFV1

Alignement de la séquence de l'ORF6 de pIRI46 avec Mig.Mth1, ORF10 du plasmide pFV1 de *Methanothermobacter thermoautotrophicus*. Les résidus identiques sont en rouge sur fond jaune et les résidus similaires sont en noir sur fond vert. Les AA impliqués dans la liaison à l'ADN sont indiqués par des disques, rouges lorsque cet AA a une importance dans la spécificité de substrat. Les carrés rouges indiquent les AA impliqués dans l'attaque nucléophile. Les triangles noirs localisent les cystéines formant le cluster [4Fe-4S].

La préférence des erreurs à corriger a été étudiée par mutagénèse dirigée sur la protéine de *Methanothermobacter* dont la version sauvage à une préférence pour la correction des thymines lors de mésappariement T/G (Yang *et al.,* 2000). Les mutations A50V ou L187Q permettent la conversion de préférence pour la reconnaissance et la correction des adénines lors de mésappariements A/G, comme c'est le cas pour la glycosylase MutY de *E.coli* (Liu *et al.,* 2008). Une analyse détaillée de l'alignement de séquence permet de formuler l'hypothèse que la protéine codée par l'ORF6 de pIRI42 code une thymine glycosylase. Ce type de correction serait biologiquement cohérent avec la nécessité de corriger les fréquentes apparitions de mésappariements G/T à haute température par déamination de 5-méthyle-cytosines (m5C) (Horst *et al.* 1996). Cette hypothèse est également étaillée par la présence d'une cytosine méthylase codée par l'ORF9.

L'**ORF9** code une protéine de 467AA, dont le plus proche homologue est une méthyltransférase (MTase) codée par APE_872 de *A. pernix* K1 ; elles partagent 53,6% d'identité. Cette protéine est une m5c-MTase, elle méthyle le carbone 5 de certaines cytosines et conduit à la formation de 5-méthyl-cytosine (m5C). Cette protéine est rencontrée dans les 3 domaines du vivant et possède

de multiples fonctionnalités. Chez les procaryotes, elle est impliquée dans les systèmes de restriction-modification. Chez les eucaryotes, elle sert principalement à la régulation de l'expression des gènes en méthylant les îlots CpG (Wang *et al.,* 2004) et à la réparation de l'ADN par BER (Visnes *et al.,* 2008). Dans ce cas précis, elle est chargée de faciliter la sortie de la base erronée de la double hélice (base flipping) afin de la cliver et de la réparer (Dodson *et al.,* 2002).

Deux hypothèses sur le fonctionnement couplé des ORF6 et 9 peuvent être énoncées : un système de restriction/modification ou bien un système de réparation de l'ADN. Malheureusement, ces hypothèses ne peuvent être tranchées sans expérimentation. La présence de nombreux gènes homologues rencontrés sur les chromosomes de Crenarchaeota ainsi que la présence de nombreuses séquences répétées et réarrangements dénotent un mécanisme actif de capture pouvant être affilié au système de RM. Néanmoins l'hypothèse d'un système de réparation de l'ADN ne peut être exclue. Un tel système, composé d'une endonucléase T/G et d'une cytosine méthylase, a récemment été décrit sur le plasmide p4C de *Thermus* sp. 4C (Ruan *et al.,* 2007). Le curage plasmidique de cette souche conduit à la diminution de l'optimum de température de 10°C. Les auteurs ont conclus à l'implication du plasmide dans la thermotolérance.

De nombreuses protéines ayant montré leur implication dans divers processus très différents, on ne peut exclure une implication dans les deux mécanismes. En plus de garantir la pérennité du génome, une correction plus efficace lors de la réplication pourrait assurer le maintien des séquences nécessaires à l'activité de RM.

L'analyse des contextes génomiques, ne se référant pas seulement à la protéine produisant le meilleur score lors du Blast, montre que plusieurs gènes du plasmide sont situés à proximité dans le génome d'A.pernix. En effet, les ORF 6, 7, 8, 9 et 10 du plasmide sont homologues à des gènes contigus du génome de *A.pernix*, entre APE_0871 et APE_0875 (Figure 50). A proximité de ces gènes se trouvent une ATPase possédant des motifs viraux et également une hélicase. Une analyse globale du génome de *A.pernix* suggère un héritage horizontal de cette région car elle possède un biais dans l'usage des dinucléotides ainsi qu'une composition anormale en G+C (IslandPath, PredictBias).

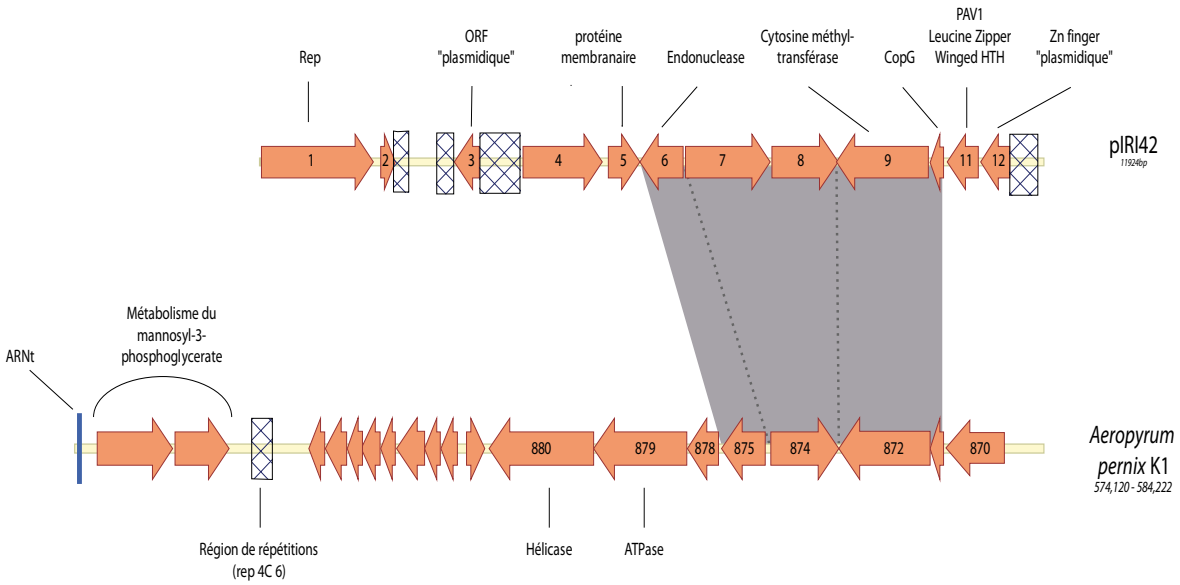


Figure 50 Génomique comparée de pIRI42 et de *Aeropyrum pernix* K1

Les **ORFs 7 et 8** sont des paralogues issus d'un évènement de duplication segmentale probablement lié à un évènement de recombinaison. Les protéines codées possèdent pour seul homologue la protéine hypothétique PAE3200 de *Pyrobaculum aerophilum*. La fiche genbank indique une implication dans la réplication et la réparation de l'ADN. Cette information résulte d'une recherche dans la banque de motif PFAM avec une accroche pour le motif RuvB_N. RuvB est une hélicase bactérienne du résolvasome RuvABC catalysant la résolution des jonctions de Holliday apparaissant lors de la recombinaison. Une recherche itérative en psi-blast montre également des accroches avec le gène APE_874 de *Aeropyrum pernix* K1. Cette information d'importance sera intégrée et discutée ultérieurement en apportant une vision du contexte génomique de ce gène chez *A.pernix* et chez pIRI42.

4.4 *Discussion sur le plasmide pIRI42*

Contrairement aux autres éléments génétiques, pratiquement tous les ORFs de ce plasmide possèdent des homologues présents dans des génomes d'*Archaea*. Seuls deux gènes sont orphelins. Ces deux gènes pourraient ne pas réellement coder de protéines dans le sens où ils ne sont pas inscrits dans une unité transcriptionnelle et sont les seuls à ne pas posséder de site de fixation du ribosome. De plus ces gènes sont truffés de séquences répétées ayant peut-être servies à générer cette « structure génique » par recombinaison. Cette hypothèse est confortée par l'observation d'une grande plasticité de ce génome.

A titre de trace de cette plasticité, on peut noter la présence d'une duplication segmentale pouvant être générée soit par une erreur de réplication, soit par une recombinaison entre différentes copies du plasmide dans la cellule. Le terme erreur est en effet à pondérer. D'un point de vue anthropocentrique, ce genre d'évènement est souvent synonyme de dérangement génétique alors que dans le cas de ce genre de réplicon ils participent activement à leur évolution. Même si un gène est indispensable au fonctionnement du réplicon, la duplication d'un gène permet à l'une des deux copies d'accumuler des mutations. En cas de transfert horizontal de ce génome dans un nouvel hôte, le paralogue ayant accumulé les mutations pourrait être plus adapté au nouvel environnement, avoir une composition (usage dans codon) ou produire une protéine à meilleure efficacité.

La plasticité est également suspectée si l'on s'intéresse à l'agencement des gènes sur le réplicon. Au contraire des autres plasmides possédant des ORFs généralement colinéaires ou organisés en deux blocs anticolinéraires. L'orientation des gènes de pIRI42 ne respectent pas les précédentes règles observées. De plus, il existe de nombreuses répétions localisées dans les espaces intergéniques dont certaines correspondent aux positions d'inversion du biais cumulatif en G+C. La fluidité de ce génome et la présence de nombreuses répétitions pourraient résulter d'un mécanisme actif favorisant les recombinaisons. En effet, il existe un couple de gènes, cytosine méthyl-transférase ADN glycosylase, également retrouvé sur les plasmides pFV1 et pFZ1 de *Methanobacterium thermoautotrophicum*. Il a été démontré qu'il code pour un système de restriction-modification. Un hypothétique système de ce genre pourrait reconnaître et cliver l'ADN au niveau des répétitions situées dans les espaces intergéniques et ensuite recoller les morceaux d'un autre ordre. De cette façon, lorsque qu'un autre réplicon (ou d'un ADN extrachromosomique importé par la cellule) rentre dans la cellule, il pourrait être incorporé grâce à ce système.

D'un point de vue cellulaire, il est plus intéressant que de l'ADN étranger transite d'abord par un réplicon indépendant du chromosome. Ceci faciliterait les processus d'adaptation, afin d'optimiser la séquence nucléotidique de l'ADN étranger avant de l'incorporer au chromosome.

5. pEXT16, un réplicon à deux origines de réplication ?

5.1 *Description générale du plasmide pEXT16 et de la souche Pyrococcus sp. EXT16*

pEXT16 est un plasmide porté par la souche *Pyrococcus* sp. EXT16, isolée à partir d'un échantillon collecté au niveau de la crête Pacifique ouest (*East Pacific Ridge*) durant la campagne EXTREME en octobre 2001. La nature de cet échantillon est une paroi organique d'un *Alvinella pompejana,* un annélide polychète rencontré uniquement à proximité des cheminées hydrothermales de l'océan Pacifique. Egalement appelé vers de Pompéi, il présente une thermotolérance exceptionnelle pour un eucaryote (plus de 80°C chez l'adulte) et peut être qualifié d'extrêmophile.

Ce plasmide est le plus gros séquencé durant cette étude. Les 34196 bp qui le constituent présentent un contenu en G+C de 45,8%. La distribution de cette composition en G+C n'est pas uniformément répartie. Les régions les plus pauvres en G+C correspondent aux espaces intergéniques (Figure 51). Ce génome contient également de nombreuses répétitions, localisées préférentiellement dans les espaces intergéniques. 30 ORFs sont prédits pour coder des protéines de tailles comprises entre 51 et 1142AA.

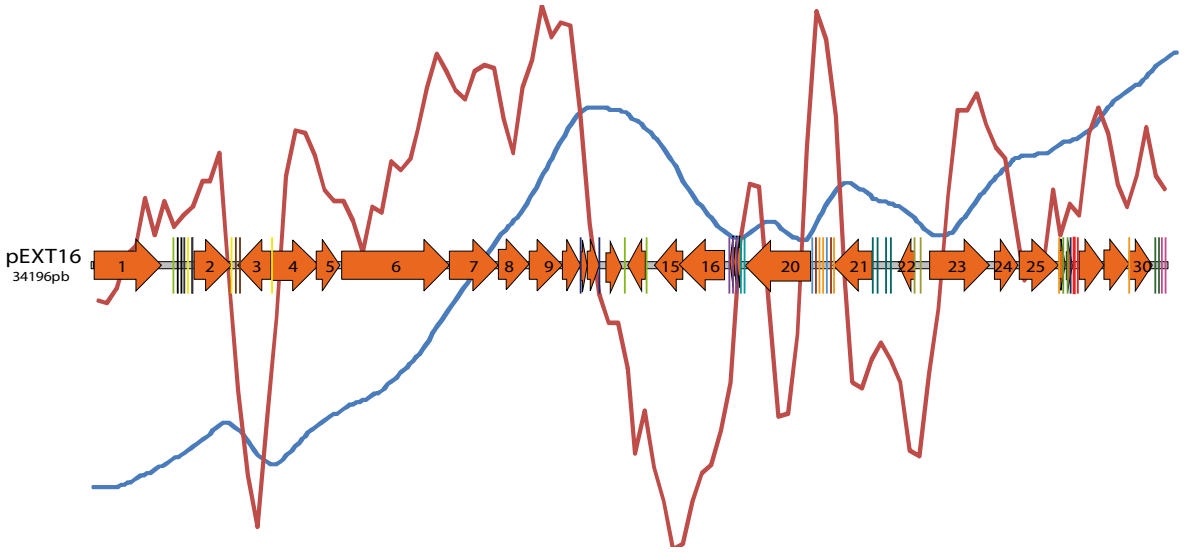


Figure 51 Carte du génome de pEXT9b

Les ORFs sont représentés par des flèches. Les répétitions sont représentées par des barres verticales colorées, chaque couleur correspondant à une répétition précise. La courbe rouge représente le biais en G+C et la courbe bleue le biais cumulatif en G+C.

Contrairement aux autres plasmides qui possèdent un grand nombre de gènes orphelins, pEXT16 possèdent de nombreux homologues dans les bases de données (Tableau 28). Nombre de ces gènes sont impliqués dans des processus liés à l'ADN bien que les fonctions précises ne puissent

être déterminées sur la base de la séquence protéique. Nous pouvons retrouver certains gènes présents au sein des plasmides de la grande famille, des protéines composites probablement issues de la fusion de deux ORFs, de nombreuses ATPases et un hypothétique système de restriction contribuant à la dynamique de ce génome. Seules deux protéines ont une fonction claire, une intégrase et une protéine dupliquée typique de l'initiation de la réplication des chromosomes d'Archaea et des eucaryotes.

Tableau 29 Protéines codées par le plasmide pEXT16

ORF	Taille	pI	PS	TMH	Fonction putative	Blast			
						Protéine	Espèce	Evalue	Id
1	718	5,8			Helicase DEAD	Mhun_2466 DEAD/DEAH box helicase-li	*Methanoculleus marisnigri JR1*	1e-44	29
2	401	6,81	0		Orc1/cdc6	TERMP_2054 orc1/cdc6	*Thermococcus barophilus*	2e-62	39
3	352	8,99							
4	491	9,51			Cytosine Methyltransférase	DNA-cytosine methyltransferase	*Microcoleus chthonoplastes PCC 7420*	9e-84	44
5	250	9,22			DNA glycosylase	putative DNA glycosylase	*uncultured methanogenic archaeon RC-I*	2e-34	36
6	1142	5,65	+		ATPase AAA+ Segregation	Hbut0359	*Hyperthermus butylicus*	2e-2	23
7	526	5,3							
8	328	6,11			DEAD -> not helicase	helicase domain-containing protein	*Sphingomonas wittichii RW1*	3e-24	35
9	351	9,7	+		DUF883	Mycobacterium vanbaalenii PYR-1	*helicase domain-containing protein*	2e-27	37
10	187	9,62							
11	125	10,1	+						
12	67	6,72		2		ribonuclease PH	*Anabaena variabilis ATCC 29413*	2	37
13	170	9,68	+	3					
14	205	9,82			Integrase	TK0104 integrase/recombinase	*Thermococcus kodakarensis KOD1*		
15	300	5,52							
16	482	8,38							
17	66	10,3							
18	64	10,7			DNA binding - RHH				
19	84	7,99	+	1					
20	694	4,54							
21	425	7,78			Orc1/Cdc6	TERMP_2054 orc1/cdc6	*Thermococcus barophilus*	9e-118	54
22	160	8,65			DNA binding	TERMP_2062	*Thermococcus barophilus*	1e-58	77
23	638	7,88			HerA Helicase + Phospholipase D	aq_1852	*Aquifex aeolicus VF5*	5e-79	32
24	259	4,84							
25	414	9,41							
26	62	8,73	+						
27	51	9,49							
28	257	9,49			DNA binding - HTH	TERMP_2050	*Thermococcus barophilus MP*	8e-132	87
29	291	7,86			DUF1814	TERMP_2048	*Thermococcus barophilus MP*	2e-139	89
30	238	9,67			Resolvase	TERMP_2052	*Thermococcus barophilus MP*	7e-53	87

ORFs du plasmide pEXT16 et caractéristiques des protéines traduites : taille, point isoélectrique, présence d'un peptide signal, hélices transmembranaires, fonction hypothétique. Présentation du meilleur résultat obtenu lors d'un BlastP.

5.2 *Contenu en gène en plasmide pEXT16*

L'**ORF1** code une protéine de 718AA. La recherche d'homologues dans les bases de données, combinée à la recherche de motifs, prédit la fonction d'hélicase ATP-dépendante de la famille DEAD. La fréquence de ce type de domaine conduit à des annotations automatiques très hétérogènes, allant de la simple fonction ATPase, ne donnant pas d'information sur l'utilisation de l'énergie libérée par hydrolyse de l'ADN, jusqu'à des protéines annotées ski2-hélicases. Cette piste a été creusée grâce à la récente caractérisation, biochimique et structurale, de la protéine

PH1280, une hélicase Ski2 codée par *P. horikoshii* OT3 (Zhang *et al.,* 2008). L'alignement de séquence confirme la présence des six motifs de ce type particulier d'hélicases (de la Cruz *et al.,* 1999). Ces motifs sont impliqués dans la fixation de l'ATP, le débobinage des duplex ARN et/ou la rupture des associations ARN-protéine et protéine-protéine. Alors que ces motifs sont très conservés, la spécificité de substrats reconnus serait liée aux domaines flanquant le domaine hélicase ou bien par l'interaction avec d'autres facteurs protéiques. L'hélicase **Ski2p** s'associe avec les protéines Ski3p et Ski8p pour former le complexe Ski, l'exosome de *S. cerevisiae*. Sa fonction est la dégradation 3' -> 5' et le recyclage des ARNm chez les eucaryotes. Il intervient dans tous les processus de dégradation des ARNm présentant des aberrations, ARNm non sens (NMD : *Non-sens Mediated mRNA Decay*) (Maquat *et al.,* 2001), présentant des codons stop incongrus (PTCs : *Premature Translation termination Decay*) ou ne possédant pas de codon stop (NSD : *Non Stop Decay*). Il a récemment été montré que le complexe Ski intervient également dans la dégradation *no-go* des ARN (Doma *et al.,* 2006; Doma *et al.,* 2006) et lors du clivage par RISC (Orban *et al.,* 2005) ou bien dans des mécanismes de défense antiviraux.

La protéine codée par le plasmide pEXT16 est probablement une hélicase à ARN, néanmoins la spécificité des ARN ciblés et les partenaires protéiques recrutés ne peut être prédite par analyse *in silico*. Cette protéine pourrait fonctionner seule, en régulant par exemple l'affinité des protéines de fixation aux ARN, nombreuses chez les organismes hyperthermophiles (Bini *et al.,* 2002) ou en ciblant certains type d'ARN (Grosjean *et al.,* 2008). Ces protéines sont très nombreuses chez les hyperthermophiles afin de contrecarrer la thermolabilité des ARN à haute température. La faible durée de vie des ARN libres dans la cellule sert de base à un mécanisme rapide de régulation par recyclage et de reprogrammation de l'expression des gènes, notamment lors de changements des conditions environnementales (Andersson *et al.,* 2006).

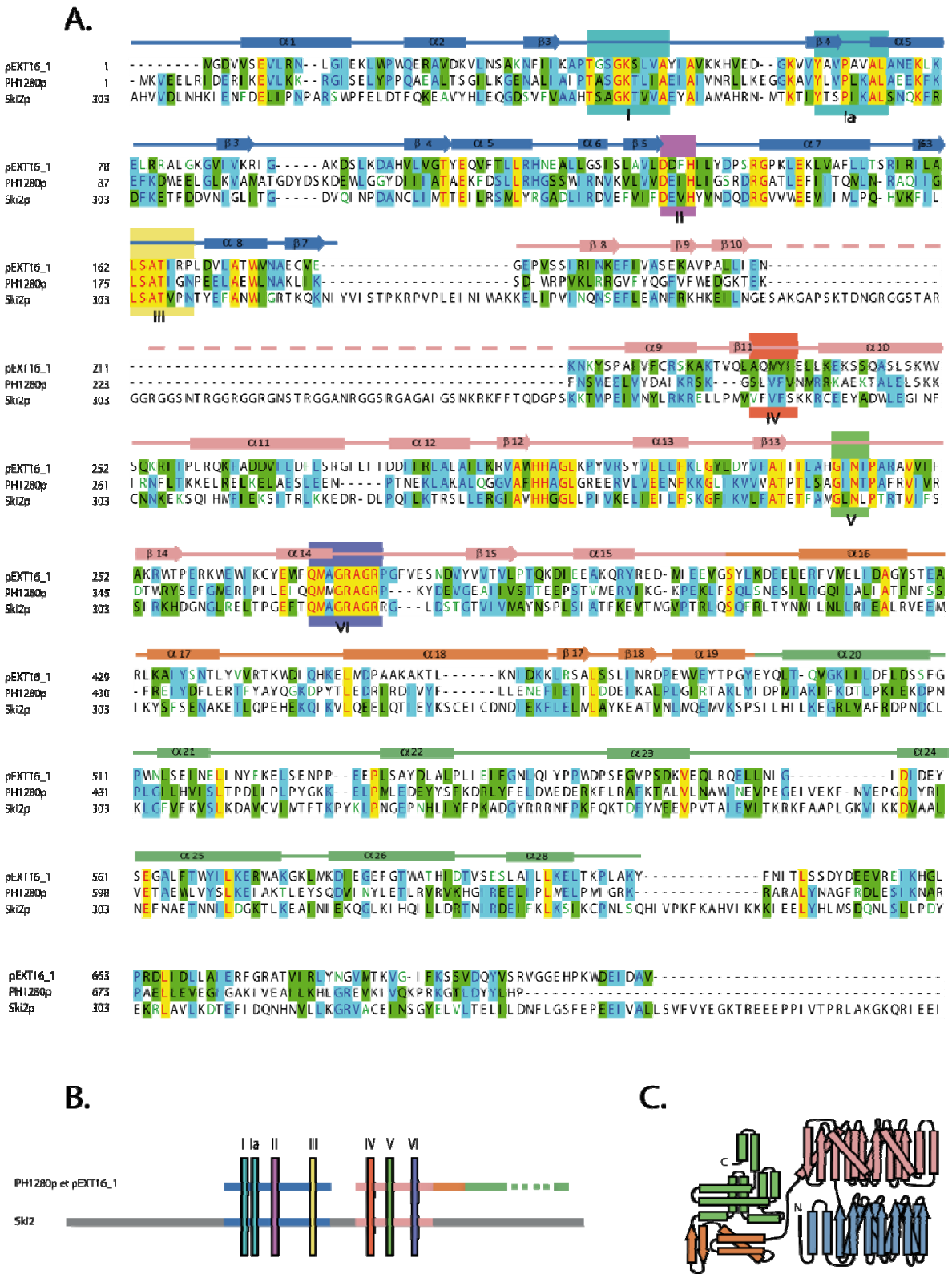


Figure 52 Alignement des hélicases Ski2 de pEXT16, Pyrococcus horikoshii et S. cerevisiae

A. Alignement de séquences des protéines codées par l'ORF1 de pEXT16 avec les hélicases Ski2 caractérisées biochimiquement et structuralement : PH1280p *Pyrococcus horikoshii* ; Ski2P *Saccharomyces cerevisiae*. Les motifs fonctionnels sont numérotés en chiffre romain et indiqués par des rectangles. La représentation de la structure secondaire déterminée pour PH1280P est représentée au dessus de

l'alignement (rectangle : feuillet beta, flèche : hélice alpha). La structure tertiaire, comportant quatre domaines est indiqué par les couleurs (bleue, rose, orange, vert).

B. Représentation schématique des hélicases Ski2p, localisation des motifs fonctionnels et des quatre domaines.

C. Représentation de la conformation de hélicase Ski2p de *P.horikoshii* (d'après (Zhang *et al.* 2008))

Les ORFs **2** et **21** codent deux protéines homologues partageant 36% d'identité de séquence (51% similarité). Une recherche d'homologues dans les bases de données prédit la fonction **Orc1/Cdc6** (COG1474, origin recognition complex 1 / cell division control). Ces protéines sont considérées comme le facteur clé initiant la réplication du chromosome. Ce sont des ATPases de la famille des AAA+. Chez les eucaryotes, la réplication est initiée au niveau de différentes origines sous le contrôle d'un complexe de reconnaissance (ORC : *Origin Recognition Complex*). Les protéines Orc1/Cdc6 des *Archaea* possèdent des similarités de séquence avec les protéines Cdc6, mais aussi avec l'extrémité C-terminale des protéines Orc1 des eucaryotes. A l'instar de leurs homologues eucaryotes, après fixation à l'origine de réplication ces protéines Orc1/Cdc6 recrutent l'hélicase MCM (Mini-Chromosome-Maintenance) afin d'initier la réplication. Ces protéines possèdent une faible affinité pour l'ADNdb quelconque mais une affinité très élevée pour l'origine de réplication. L'implication dans la reconnaissance de l'origine de réplication a aussi bien été démontrée chez *P. abyssi* (Matsunaga *et al.,* 2001) que chez *S. solfataricus* (Robinson *et al.,* 2004). Elle implique un motif HTH situé dans un domaine « winged helix ». Cette fixation nécessite également l'hydrolyse d'ATP par l'intermédiaire du motif ATPase de la famille AAA+. Toutes ces caractéristiques sont retrouvées sur les protéines codées par les ORF2 et 21 et confirme leurs appartenances à la famille Orc1/Cdc6.

Une analyse phylogénétique révèle néanmoins une position singulière (Figure 53), ne regroupant pas ces protéines avec leurs homologues chromosomiques de Thermococcales, mais avec ceux portés par les deux plasmides, pFV1 et pFZ1, hébergés par deux isolats de *Methanothermobacter thermoautotrophicus*. Il semble donc exister une famille divergente d'Orc1/Cdc6 spécifique des plasmides d'Euryarchaea. Un dernier plasmide, pTA1, hébergé par *Thermoplasma acidophilum*, possède également une protéine Orc1/Cdc6. Néanmoins cette protéine homologue ne peut être correctement alignée et produit donc une longue branche lors de l'analyse phylogénétique conduisant à des artéfacts d'attraction. Une exploration plus poussée des réplicons extrachromosomiques chez les Crenarchaea montre qu'aucun élément génétique de ce groupe taxonomique ne code pour ce type de protéines impliquées dans la réplication des chromosomes. En considérant cette phylogénie de manière globale, nous pouvons également observer qu'elle n'est pas congruente avec celle utilisée pour le typage, l'ADNr16S. Bien que la plupart des souches

phylogénétiquement proches soient regroupées, nous n'observons pas deux blocs monophylétiques séparant les Crenarchaea et les Euryarchaea. Cette observation traduirait l'existence de transferts horizontaux ancestraux qui seraient intervenus peu de temps après l'établissement de ces deux phyla.

Le chromosome des *Archaea* peut présenter plusieurs origines de réplication (Robinson *et al.* 2004). Chez *Sulfolobus*, une des trois origines de réplication a été acquise par capture d'un élément génétique (Robinson *et al.,* 2007). Cette capture a permis l'établissement d'une nouvelle origine de réplication et de la protéine initiatrice correspondante, apparentée à la protéine eucaryote Cdt1. Si au sein d'un chromosome, une origine de réplication phylogénétiquement distante est possible, l'hypothèse de transferts horizontaux entre systèmes homologues est tout à fait probable.

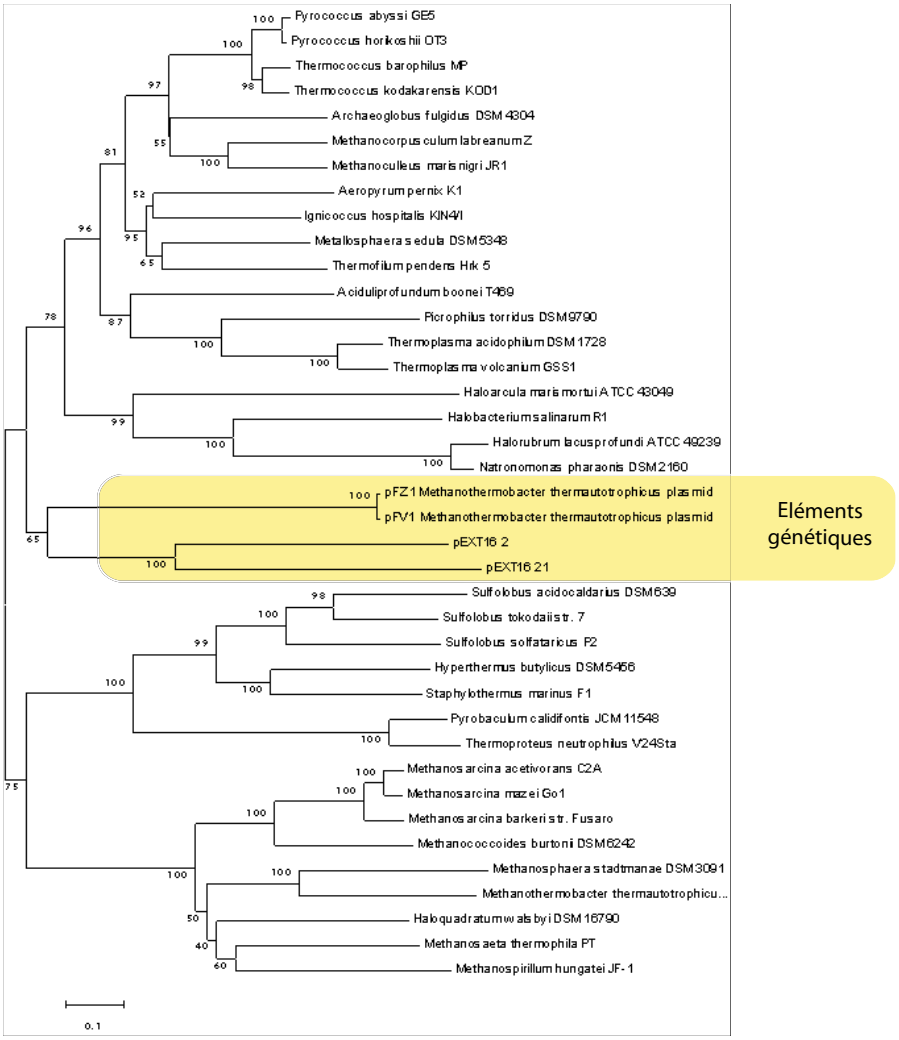


Figure 53 Analyse phylogénétique de la protéine codée par l'ORF2 de pEXT16, de fonction hypothétique Orc1/Cdc6

Ces protéines Orc1/Cdc6 se fixant à l'origine de réplication, nous avons tenté de déterminer l'existence de telles origines sur le plasmide pEXT16. Chez *P.abyssi* (Matsunaga *et al.,* 2007), elles se caractérisent comme des régions riches en A/T flanquées d'une séquence de 13pb souvent répétée. Cette séquence, annotée mini-ORB (*Origin Recoginition Box :* YTNCANNNGAAM*),* est généralement considérée comme la version minimale de l'origine de réplication des Archaea. Plusieurs séquences mini-ORB peuvent être localisées dans les espaces intergéniques situés en amont des deux gènes *orc1/cdc6.* Elles répondent au consensus archéen mais sont différentes de celles déterminées sur le chromosome des Thermococcales. Ces séquences sont localisées dans les deux plus grands espaces intergéniques du plasmide et contiennent de nombreuses répétitions, directes et inverses de grande taille, riches en A/T. Les plus remarquables s'étendent respectivement sur 55pb (IR7) et 26pb (IR8). Malgré la conformité de ces séquences avec le consensus archéen, ces mini-ORB divergent de celles des Thermococcales (Matsunaga *et al.* 2007). La divergence des séquences mini-ORB est certainement corrélée à celle des protéines Orc1/Cdc6 de pEXT16. La caractérisation biochimique des sites de fixation de ces protéines permettrait de mieux comprendre les préférences de sites de fixation et de savoir s'il existe des interférences avec les paralogues codés par le chromosome de *P.* sp. EXT16. Ceci permettrait de confirmer l'indépendance de la réplication de pEXT16 vis-à-vis du chromosome de la cellule hôte. Néanmoins, ce plasmide ne possédant pas d'hélicase, l'hypothèse la plus vraisemblable voudrait que les Orc1/Cdc6 de pEXT16 se fixent spécifiquement au plasmide afin d'initier sa réplication tout en permettant le recrutement de l'hélicase chromosomique MCM et du complexe de réplication associé, lui assurant une réplication indépendante du chromosome.

Les **ORFs 4 et 5** codent respectivement une C5-cytosine-méthyltransférase et une ADN glycosylase/endonucléase III. Ce couple de gènes a déjà été rencontré dans le plasmide pIRI42 (ORF8 et 9). Ils pourraient-être impliqués dans la réparation de l'ADN suivant un mécanisme de type BER et/ou dans un système de restriction-modification. Cette seconde hypothèse est soutenue par la présence du même couple de gènes sur les plasmides pFV1 et pFZ1 de *M. thermoautotrophicus* (possédant également le gène codant une protéine Orc1/Cdc6). En effet, un couple de gènes homologues a été caractérisé sur ces plasmides, il code un système de restriction-modification nommé *Mth*FI reconnaissant le site CTAG (Nolling *et al.* 1992). Ce système est homologue à celui de la bactérie mésophile *Neisseria gonorrhoeae* et suppose un transfert horizontal entre des *Archaea* et *Bacteria*. La présence sur pEXT16 de nombreuses

répétitions, séquences dupliquées, et gènes issus de chromosomes reflète la plasticité de ce gène qui pourrait résulter de l'action d'un système de restriction performant.

L'**ORF6** code la plus grosse protéine du plasmide (1149AA). Un seul homologue est détecté dans les bases de données ; il s'agit de la protéine Hbut_0359 de *Hyperthermus butyilicus*. Ces deux protéines sont composées de trois domaines strictement conservés. Le domaine N-terminal possède une fonction ATPase spécifique des *Archaea* (PF01637). Cette fonction très commune ne renseigne pas sur la finalité du mécanisme nécessitant la production d'énergie par hydrolyse d'ATP. Bien que le domaine central ne possède pas de motif détectable, le domaine C-terminal possède un motif DUF2408 (PF10303) de fonction inconnue et uniquement rencontré chez les champignons et un motif Mis12 (PF05859) intervenant dans la formation de leur kinétochore. Le kinétochore est un assemblage supramoléculaire de protéines au niveau des régions centromériques des chromosomes mitotiques des eucaryotes. Il sert de plateforme d'ancrage aux microtubules sur le chromosome permettant ainsi le placement des chromosomes sur le plan équatorial et leur partage en deux lots identiques au cours de l'anaphase et de la métaphase par association avec les microtubules polaires. Un faisceau supplémentaire de convergence avec notre protéine est la présence de l'activité ATPasique du kinétochore, permettant la dépolymérisation active au niveau des microtubules entraînant leur raccourcissement et donc la ségrégation. L'ORF6 coderait une protéine potentiellement impliquée dans la ségrégation du plasmide.

L'**ORF8** code une protéine de 328AA. Une recherche de similarité dans les bases de données détecte des hélicases uniquement affiliées aux bactéries et de tailles beaucoup plus importantes. La similarité est cantonnée à la portion N-terminale des hélicases. L'analyse détaillée des motifs confirme la fonction ATPase de cette protéine mais ne confirme pas l'activité hélicase malgré la présence du motif DExD. Qui plus est, la recherche de similarité et de motifs montre une affiliation aux protéines de la famille SNF2 qui utilisent l'énergie produite par l'hydrolyse de l'ATP, pour désassembler les histones liés à l'ADN chez les *Archaea* et les eucaryotes (Flaus *et al.,* 2006), favorisant ainsi l'accessibilité des facteurs de transcription. En effet, ces protéines possèdent un domaine proche de celui des hélicases DExD sans toutefois posséder une quelconque activité hélicase. Cet ORF8 code donc une protéine à activité ATPase dont la fonction reste à élucider. Cet exemple montre la limite de détection d'homologues par la recherche de similarités lorsque l'on sonde avec des séquences très divergentes. Les résultats doivent être analysés avec attention et ne pas se cantonner aux dix premières protéines détectées par BLAST. L'annotation des génomes

étant souvent automatique, il faut prendre garde à la qualité de l'annotation des séquences présente dans les bases de données.

L'**ORF14** code une protéine de 187AA. La recherche d'homologues dans les bases de données, ainsi que la prédiction de motifs et domaines, permet d'affilier cette protéine à la famille des **intégrases** phagiques (PF00589), une classe spécifique de **tyrosines recombinases** permettant l'intégration de réplicons extrachromosomiques dans le chromosome. Les homologues détectés sont portés par les virus intégratifs de *Sulfolobales*, STIV infectant *Sulfolobus* (Maaty *et al.,* 2006) ou ATV infectant *Acidianus* (Prangishvili *et al.,* 2006). Ce sont également les intégrases ayant permis l'intégration des îlots génomiques dans le génome de chromosomes d'Euryarchaea thermophiles et mésophiles. Bien que l'affiliation à cette classe de protéine soit sans ambiguïté, l'alignement de séquences présente une modification remarquable au niveau du motif catalytique. Alors que ce motif est très conservé (R…HxxR…Y), la protéine codée par le plasmide pEXT16 est de séquence R…TxxR…Y. Les intégrases sont également caractérisées par une séquence nucléotidique, nommée *att,* identique à une portion d'un ARNt et permettant la recombinaison site-spécifique conduisant à l'intégration de l'élément génétique dans le chromosome. La recherche de séquence *att,* caractérisée chez les Crenarchaea et chez les bactéries, a été infructueuse et ne permet pas de déterminer une séquence d'intégration potentielle. Nous sommes en mesure de nous demander si ce gène code une intégrase fonctionnelle ou bien si la protéine en question appartient à une nouvelle famille d'intégrase possédant un site catalytique légèrement différent et produisant une recombinaison site-spécifique ciblant une séquence *att* différente.

L'**ORF22** code une protéine de 160AA possédant seulement deux homologues dans les bases de données. Le premier est TERMP2062 de *T.barophilus* et génère un alignement de séquence produisant une identité de séquence de 77% (90% similarité). Le second, plus divergent, est APE_2505 de *Aeropyrum pernix* K1 (26% identité, 50% similarité). En dehors des bases de données publiques, cet ORF est également présent sur l'ensemble des plasmides de la grande famille.

L'**ORF23** code une protéine de 692 AA. La recherche d'homologue dans les bases de données permet seulement la détection de deux protéines possédant des similarités sur l'ensemble de la séquence. Il s'agit des protéines hypothétiques HG1295_15821 de *Hydrogenivirga* sp. 128-5-R1-1 et aq_1852 de *Aquifex aeolicus*. L'alignement de ces trois protéines produit un alignement avec un pourcentage d'identité global de 24% et 80,4% d'identité. Néanmoins, de nombreuses autres

protéines possèdent des similarités de séquences cantonnées à la portion couvrant les AA 200 à 692. Elles sont annotées protéines hypothétiques, ATPases, hélicases HerA ou protéines contenant le motif DUF87. La prédiction de domaine et l'analyse des motifs révèlent une probable fusion entre deux domaines de fonctions distinctes. Le domaine N-terminal possède une activité **Phosholipase D** (COG3886). La phospholipase D catalyse l'hydrolyse de la liaison phosphodiester des glycérophospholipides. Des activités phospholipase D ont été découvertes chez des virus, des procaryotes et des eucaryotes. Elles sont généralement régulées par des protéines kinases et des protéines liant le GTP appartenant aux familles des protéines Rho et des protéines d'ADP-ribosylation. Des approches génétiques montrent que la phospholipase D intervient dans divers processus cellulaires incluant la signalisation par récepteur, la régulation du transport membranaire intracellulaire et la réorganisation du cytosquelette d'actine. L'activation coordonnée de différentes voies biochimiques dépendant de la phospholipase D explique vraisemblablement les rôles pléiotropes de ces enzymes dans plusieurs facettes de la régulation cellulaire.

Le domaine C-terminal de cette protéine possède un motif ATPase de la superfamille ABC. L'hydrolyse d'ATP produit l'énergie nécessaire à la fonction générale du domaine **hélicase HerA** (COG0433). Ces protéines sont les premières hélicases bipolaires caractérisées (5'->3' et 3'->5'). Elles sont ubiquistes chez les Archaea hyperthermophiles et codées en opéron avec les gènes mre11-rad50 et nudA (Constantinesco *et al.,* 2004). Chez *Sulfolobus*, ces protéines interagissent pour effectuer une recombinaison nécessaire à la réparation de l'ADN ayant subi des lésions double brin comme c'est le cas suite à l'exposition aux UV (Quaiser *et al.,* 2008). Il faut toutefois faire attention à la bibliographie, car cette protéine à tout d'abord été caractérisée chez *P.abyssi* sous le nom de Mla (Mre11-linked ATPase) (Manzan *et al.,* 2004) sans s'être intéressé au potentiel d'une fonction hélicase. L'absence de gènes homologues à ceux décrits dans l'opéron précédent et la présence du sous-motif FtsK-SpoIIIE (PF01580) oriente plus cette hélicase dans une fonction de ségrégation de l'ADN ou d'encapsidation virale comme cela est suggéré dans la bibliographie (Iyer *et al.,* 2004; Burroughs *et al.* 2007).

Malgré la convergence entre la fonction de ces domaines, pouvant d'une part être impliqués dans transport membranaire et l'assemblage du cytosquelette et d'autre part dans la ségrégation/encapsidation, cette association d'une **hélicase HerA** avec une **phosholipase D** n'a encore jamais été décrite.

Les ORF **28** et **29** (PF08843) ne possèdent pas de motif permettant de leur inférer une fonction. Néanmoins, ils présentent de nombreux homologues dans les bases de données, majoritairement dans les génomes d'Archaea, mais aussi dans certains génomes de bactéries thermophiles. La particularité de ces homologues est qu'ils sont tous situés à coté les uns des autres. Après alignement de séquences, les similarités sont très importantes, 88% d'identité avec les TERMP2048 et TERMP2050 de *T. barophilus* (96% de similarité), 43% avec PAB2321 et PAB2322 de *P.abyssi* (63% similarité), 42% avec Tlet1204 et Tlet1205 de *Thermotoga lettingae* (62% similarité). L'analyse de l'**ORF30** de pEXT16 pourrait expliquer cette association des homologues des ORF28 et 29. Il code une protéine de 230AA possédant un domaine **PinR** (COG1961), une classe particulière de recombinase site-spécifique de la famille des **sérines recombinases** (PF0039). L'homologue le plus conservé dans les bases de données est également présent dans le génome de *T. barophilus*, bien qu'il soit présent sous forme partionnée. L'alignement des deux séquences produit une identité de 87% (94% de similarité). Tout comme pour les ORF28 et 29, l'homologue de l'ORF 30 est adjacent à ceux de *T.barophilus*. De plus, ce trio de gènes est flanqué par les deux plus grandes séquences palindromiques du génome de pEXT16 (Pal2, Pal5), et les ORF28 et 29 sont bornés par une répétition inverse (IR9a et IR9b). Cette association suppose une implication dans une fonction commune et une prédisposition à la mobilité. Cette organisation n'est pas sans rappeler celle observée au sein des séquences d'insertions et pourrait expliquer la mobilité de cet opéron.

5.3 *Relation avec un hypothétique îlot génomique de T. barophilus*

Cette description des ORFs montre que de nombreux gènes sont apparentés au génome de *T. barophilus*. En représentant de manière schématique la conservation des gènes entre pEXT16 et *T. barophilus* (Figure 54) nous pouvons mettre en évidence la présence d'un îlot génomique potentiel au sein de *T.barophilus,* possédant de nombreuses similarités avec pEXT16 mais qui ne sont pas retrouvées au sein des autres génomes de Thermococcales.

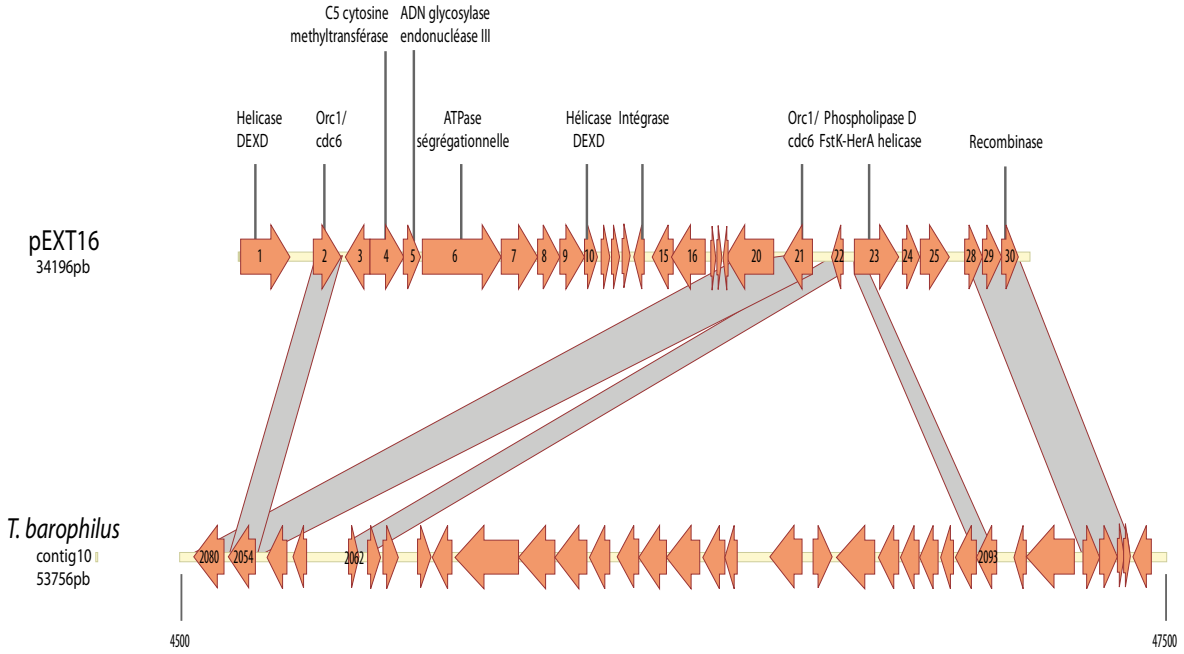


Figure 54 Comparaison de pEXT16 avec le contig 10 de *T. barophilus*

Certains pourcentages d'identité atteignent des taux record de l'ordre de 85%, comme cela a uniquement été observé pour les gènes codant le transposon de pGE2 (page 154). Cette observation pourrait également expliquer la fusion potentielle entre l'hélicase HerA et une phosholipase D de l'ORF23. En effet, il existe également une phospholipase D dans cette portion du génome de *T. barophilus*. La présence d'un groupe de gènes apparentés à un transposon, d'une intégrase et d'un système de restriction-modification sur le plasmide, favorise l'hypothèse de fréquentes interactions avec le chromosome des Thermococcales.

5.4 *Discussion générale autours de pEXT16*

En dehors d'être le plus gros plasmide étudié, ainsi que celui possédant le moins de gènes orphelins, pEXT16 est l'élément génétique le plus atypique de Thermococcales. L'analyse de la séquence nucléotidique supposait une extraordinaire plasticité de ce génome par une distribution anarchique du biais en nucléotide et la présence de nombreuses séquences répétées, directes, inversées ou palindromiques localisées dans les courts espaces intergéniques. La prédiction des séquences codantes révèle la présence de nombreux gènes conférant des capacités à échanger des gènes et à intergair avec d'autres réplicons : la présence d'un système de restriction-modification (ORF4 et 5) et de deux recombinases. La première est une intégrase (ORF14),

permettant l'intégration dans un autre réplicon et la seconde une résolvase (ORF30) affiliée à un transposon par l'analyse de son contexte génomique. Cette plasticité est également illustrée par la présence d'une protéine chimérique constituée de la fusion d'une hélicase bipolaire HerA avec une phospholipase D (ORF23).

D'un point de vue réplicatif, cet élément génétique possède deux protéines initiatrices de la réplication Orc1/Cdc6 (ORF2 et 21). Elles interviennent habituellement dans la réplication des chromosomes d'*Archaea* selon un système homologue à celui des eucaryotes permettant de recruter l'hélicase MCM et la machinerie réplicative associée. Une réplication apparentée à celle du chromosome peut donc être avancée, d'autant plus que certains motifs typiques des origines de réplications chromosomiques ont été mis en évidence en amont des deux gènes *orc1/cdc6* de pEXT16. Néanmoins, certaines différences au niveau des protéines Orc1/Cdc6 et des origines de réplication supposent une réplication indépendante de l'élément génétique vis-à-vis du chromosome. pEXT16 n'est pas le premier plasmide d'*Archaea* à posséder ce type d'origine de réplication. En effet, trois autres plasmides codant des protéines Orc1/Cdc6 ont déjà été caractérisés. pFV1 et pFZ1 sont présent dans *Methanobacterium thermoautotrophicum* dont les homologues Orc1/Cdc6 sont apparentés à ceux présents sur le plasmide pEXT16. Une autre similitude est observée entre ces plasmides : ils codent pour un système de restriction-modification homologue à celui de pEXT16. Ce système pourrait expliquer le transfert de blocs de gènes, notamment le système *orc1/cdc6* et l'origine de réplication associée.

pTA1, hébergé par *Thermoplasma acidophilum*, est le troisième plasmide possédant ce type de protéine. La séquence peptidique de sa protéine Orc1/Cdc6 est suffisamment divergente des autres plasmides pour qu'il ne soit pas considéré comme issus d'un proche ancêtre commun des autres plasmides.

L'observation de deux origines de réplication putatives apparentées à celle des chromosomes dans un plasmide pose de multiples questions : sont-elles indépendantes l'une de l'autre, chaque Orc1/Cdc6 est-il réellement spécifique de l'origine de réplication qui le précède ? Dans quelle mesure les protéines homologues codées par le chromosome sont-elles capables de se fixer sur celles du plasmide ? Ces observations confortent également certaines hypothèses émises sur la présence de multiples origines de réplication dans les chromosomes d'*Archaea*. En effet, chez *Sulfolobus*, il a été démontré que la troisième origine de réplication (*oriC3*) est différente des deux autres (*oriC1* et *oriC2*). Cette origine de réplication ne présente pas de gène *orc1/cdc6* à proximité, mais une protéine WhiP apparentée à Cdt1, un autre facteur d'initiation de la réplication eucaryote. Cette troisième origine de réplication aurait été acquise par un mécanisme

de capture d'élément génétique mobile (Robinson *et al.* 2007). pEXT16 possède les caractéristiques typiques des réplicons chromosomiques et une intégrase. Son intégration dans le chromosome permettrait-elle l'acquisition d'une origine de réplication supplémentaire pour le chromosome ? Si l'on appréhende la question sous l'angle opposé, pEXT16 pourrait être une origine de réplication chromosomique devenue indépendante ?

Le maintient d'un réplicon apparenté à celui du chromosome ainsi que le faible nombre de copies de pEXT16 dans la cellule (estimé par la faible abondance d'ADN extrachromosomique) suppose un mécanisme actif de partition permettant la pérennité du plasmide. Une ATPase présentant des similitudes avec les protéines centromériques du kinétochore eucaryote pourrait être candidate à cette fonction (ORF6).

Finalement, deux autres protéines retiennent tout particulièrement notre attention quant à leur fonctionnement et leur implication dans la « physiologie » du plasmide et/ou de l'hôte. La première est une hélicase à ARN de la famille ski2. Elle est la première étape conduisant à la dégradation des ARN par l'exosome. Son implication dans des mécanismes antiviraux a été démontrée chez de nombreuses espèces par dégradation d'ARN viraux. Je propose donc une hypothèse, que j'admets quelque peu farfelue, sur l'utilité d'un tel gène porté par un plasmide/virus. Je supposerais une implication dans une mesure contre défensive vis-à-vis du système CRISPR. Ce système a récemment été découvert, il est présent chez toutes les *Archaea* et confère une immunité de la cellule suite à l'établissement d'un élément extrachromosomique dans le cytoplasme par un système analogue à l'ARN interférent des eucaryotes. Dans la course à l'armement entre le chromosome et les virus, dans quelle mesure un système de contre-défense pourrait être codé par un génome viral afin de se soustraire à ce mécanisme « immunitaire » ?

6. pEXT9b, un plasmide à hélicase MCM

6.1 *Description générale du plasmide pEXT9b*

pEXT9b est un plasmide porté par la souche *Thermococcus* pEXT9. Ce plasmide coexiste avec le plasmide pEXT9a qui appartient à la grande famille (page 107). La présence de deux réplicons au sein d'une souche suppose qu'ils appartiennent à des classes d'incompatibilités différentes, sinon la compétition entre les réplicons aboutirait à l'exclusion de l'un d'entre eux.

Ce plasmide a une taille de 11.237pb, pour une composition en GC de 43%. 14 ORFs codent des protéines de tailles comprises entre 52 et 951AA. Ces ORFs sont disposés suivant deux blocs d'orientations inverses. Les ORFs 1 à 4 sont sur le brin direct tandis que les ORFs 5 à 14 sont sur le brin complémentaire. Parmi ces 14 ORFs, la moitié sont orphelins. La plupart de ces orphelins possèdent de grandes quantités de séquences répétées qui pourraient résulter d'évènements de recombinaison produisant ainsi des ORFs codant des protéines non fonctionnelles.

4 promoteurs sont mis en évidence, 2 sur le brin direct et 2 sur le brin complémentaire.

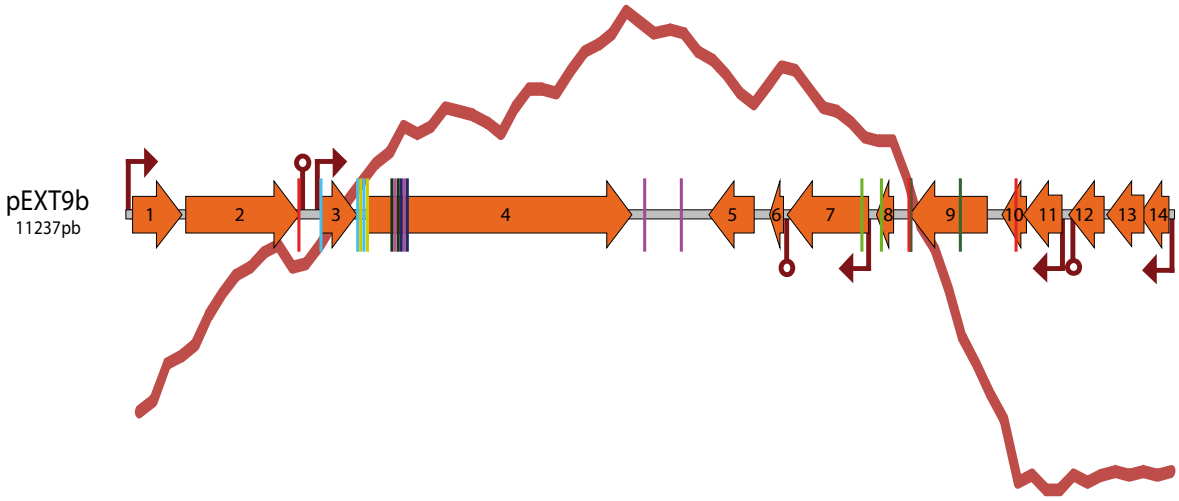


Figure 55 Carte de plasmide pEXT9b

Les ORFs sont représentés par des flèches, numérotées 1 à 14. Les promoteurs et terminateurs de transcription sont respectivement localisés par des flèches brisées et des sucettes. Le biais cumulatif en G+C est indiqué par la courbe rouge. Les répétitions sont représentées par des barres verticales colorées, chaque couleur représentant une séquence donnée (pour une plus grande lisibilité, l'espacement entre les répétitions ne correspond pas exactement à l'échelle de taille)

6.2 *Contenu en gènes du plasmide pEXT9b*

Les fonctions putatives des protéines codées par le plasmide pEXT9b sont consignées dans le Tableau 30.

Tableau 30 Protéines codées par le plasmide pEXT9b

	Taille	pI	SignalP	TMH	Fonction hypothétique	BlastP				Motif
						Protéine	Espèce	evalue	%Id	
1	177	9,2								
2	407	9,6	+	12	Perméase	TK0253	*Thermococcus kodakaraensis*	4e-140	60	MFS_1
						TK0833	*Thermococcus kodakaraensis*	3e-30	29	
3	128	8,8			Régulateur de transcription	TK2270	*Thermococcus kodakaraensis*	2e-3	40	RHH_1
						PAE1484	*Pyrobaculum calidifontis*	2e-3	40	
4	951	6,5			Helicase MCM		*Thermococcus kodakaraensis*	5e-97	35	MCM
5	164	6,8			Régulateur de transcription	ORF3	*Sulfolobus sp NOB8H2*	0,2	44	HTH
6	52	5,7						0,3	22	
7	293	9,7								
8	61	9,9								
9	278	8,9			Régulateur de transcription		*Granulibacter bethesdensis*	4e-4	23	Bzip
10	89	9,6								
11	138	9,2	+	2						
12	125	4,7			Toxine	PAB1755	*Pyrococcus abyssi GE5*	9e-45	76	DUF132
13	133	5,41			Antitoxine?	PAB1754	*Pyrococcus abyssi GE5*	2e-51	84	RHH_1
14	105	9,67	+	1						

L'**ORF2** code une protéine de 407AA. Elle partage 60,9% d'identité avec TK2270 de *T.kodakaraensis*. Cette protéine est un transporteur de la famille ubiquiste des MFS (Major Facilitator Superfamily) également appelé uniporter-symporter-antiporter. Le transport peut se faire vers l'extérieur ou vers l'intérieur de la cellule suivant des gradients chimiques parallèles ou antiparallèles. Les molécules transportées sont généralement de petits solutés, mono et disaccharides, phosphate, nitrate, sialate, ou l'efflux de drogues simples. La nature de l'élément transporté permet l'affiliation à l'une des 18 familles de perméase MFS. La protéine fonctionne seule, aucun partenaire n'est nécessaire pour le transport. A l'image des protéines de cette famille, il est possible de localiser les motifs fonctionnels (PF07690) et 12 segments transmembranaires permettant d'enchâsser la protéine dans la membrane. La caractérisation fonctionnelle de certains homologues permet de prédire l'implication de cette protéine dans l'acquisition de sucres par l'intermédiaire d'un gradient protoionique. Cette fonctionnalité, apportée par le plasmide, pourrait augmenter le fitness de l'hôte cellulaire en favorisant sa capacité d'acquisition de nutriments dans les environnements oligotrophes auxquels sont soumis les Thermococcales.

L'**ORF4** code la plus grosse protéine rencontrée dans cette étude, d'une taille de 951AA. La protéine est une hélicase 3' -> 5' de la famille MCM, MiniChromosome Maintenance. Chez les

eucaryotes le complexe MCM est essentiel aux phases d'initiation et d'élongation lors de la réplication de l'ADN. Il est formé d'un heterohexamultimère des sous-unités : MCM2 à MCM7. Ces sous-unités possèdent des similarités de séquences et sont des paralogues présumés dériver d'un ancêtre commun (Iyer *et al.,* 2004). Les *Archaea* ont un mécanisme de réplication similaire aux eucaryotes (Kelman *et al.,* 2003) dont l'initiation de la réplication implique un unique orthologue MCM formant un homomultimère de stœchiométrie variable, hexamère (Pape *et al.,* 2003), heptamère (Yu *et al.,* 2002), filamenteux (Chen *et al.,* 2005) ou la forme majoritaire en solution de double hexamère (Gomez-Llorente *et al.,* 2005). Les études structurales et hydrodynamiques montrent des assemblages similaires entre les deux domaines du vivant; le complexe forme un anneau autour de l'ADN afin d'en séparer les deux brins (Fletcher *et al.,* 2003). Jouant un rôle clé dans la réplication de l'ADN, le complexe MCM est une cible de choix pour la lutte anti-tumorale. La composition simplifiée du complexe MCM des *Archaea* en fait un modèle d'étude particulièrement intéressant. Les études réalisées ont principalement visé à caractériser les MCM de *Sulfolobus solfataricus* et *Methanobacter thermoautrophicum*. La comparaison de ces MCM montre une certaine modularité de leur organisation en fonction des différents domaines rencontrés. Néanmoins, un certains nombre de traits communs peuvent être définis. La principale caractéristique est un découpage en quatre domaines possédant chacun plusieurs motifs. A l'heure actuelle, seul le plasmide pTAU4 de *Sulfolobus* possède une protéine de type MCM (Greve *et al.* 2005).

L'**ORF3** code une protéine de 128AA. Cette protéine est composée de deux domaines. Le domaine N-terminal possède un motif Ribbon-Helix-Helix de type CopG. Les protéines possédant ce type de domaine interagissent avec les acides nucléiques et sont fréquemment rencontrés sur les facteurs de transcription. La sous-famille CopG est également connue sous le nom de RepA, responsable de la régulation du nombre de copies des plasmides. Elle se fixe sur le promoteur *repAB* et contrôle la synthèse de RepB, protéine initiatrice de la réplication du plasmide. Alors que de nombreux régulateurs de la transcription de cette famille fixent les acides nucléiques au moyen du domaine RHH, les protéines de la sous-famille CopG utilisent ce motif pour leur dimérisation, structure essentielle à la fonction de régulation du nombre de copies du réplicon. Une cherche d'homologues dans les bases de données permet d'accrocher des protéines possédant le domaine CopG mais aucune fonction ne peut être attribuée au domaine C-terminal. Le gène *copG* possède en amont un promoteur. Ce promoteur est bordé par une séquence répétée inverse de 10 pb servant probablement de site de fixation à CopG. Une fois fixée, cette protéine empêcherait la

fixation de l'ARN polymérase et la transcription de cet opéron impliqué dans la régulation du nombre de copies du réplicon.

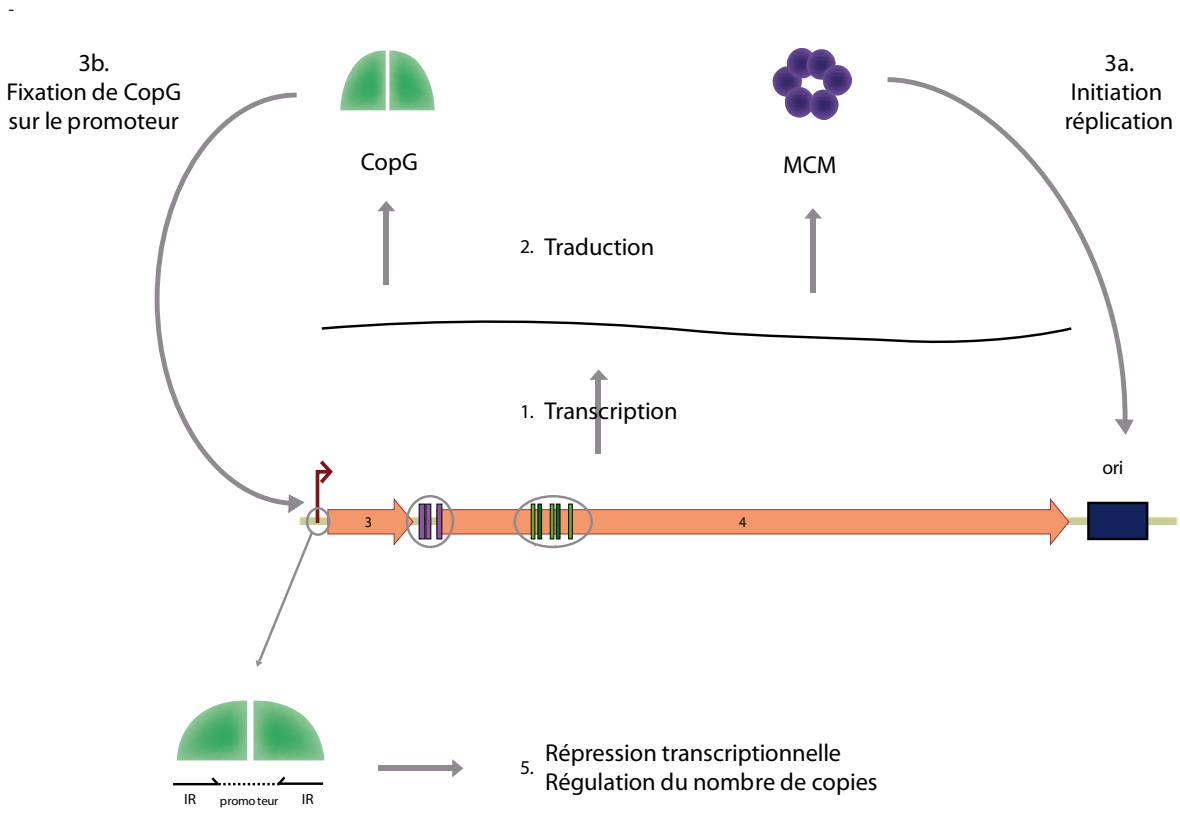


Figure 56 Modèle supposé de la réplication et de la régulation du nombre de copies de pEXT9b

Le contexte génomique de l'homologue TK2270, codé par le chromosome de *T.kodakaraensis*, possède une organisation similaire : une séquence répétée inversée est située au niveau d'un promoteur, suivi également d'un opéron de gènes impliqués dans la maintenance du génome et plus précisément dans la partition.

L'**ORF5** code une protéine de 164AA, composée de deux domaines. Le domaine C-terminal se replie suivant le modèle Winged Helix (WH), de la famille des motifs de fixation à l'ADN de type Helix-Turn-Helix. La recherche d'homologue, en sondant les banques de données avec ce domaine, montre que les plus proches homologues sont de petites protéines s'alignant en totalité. Ces homologues sont codés par les plasmides pNOB8 de *Sulfolobus* sp. NOB8H2 et pUR500 de *Methanococcus maripaludis* C5. Le domaine N-terminal accroche des protéines SMC codées par diverses *Bacteria* et *Archaea*. Ce sont de grosses protéines, multidomaines,

intervenant dans la ségrégation du chromosome. En détaillant l'alignement, il est surprenant de remarquer que les AA conservés correspondent à une zone non annotée située entre deux domaines. La conservation d'une telle séquence entre plusieurs domaines du vivant ainsi que son association à d'autres domaines fonctionnels amène à se poser des questions sur sa fonction. Une recherche de motif confirme la ressemblance avec les protéines impliquées dans la partition du réplicon. Elle révèle une faible similarité avec les préfoldines, chaperonnes impliquées dans le repliement des protéines et leur polymérisation, telle la mise en place du cytosquelette par polymérisation de tubuline et d'actine. Néanmoins, aucun motif ATPase fournissant l'énergie nécessaire à la polymérisation ne peut être mis en évidence.

Les **ORF12** et **13** peuvent être discutés conjointement. Les deux protéines codées sont de petites tailles, respectivement 125 et 133 AA, et les plus acides du génome. Les homologues rencontrés dans les bases de données sont également associés et contigus sur les chromosomes pour lesquels ils sont rencontrés ; à l'image des protéines PAB1755 et PAB 1754 de *P. abyssi* GE5, possédant respectivement 76 et 84% d'identité. Cette particularité suppose que ces deux protéines codées participent à une action commune. L'**ORF12** possède un motif de fonction inconnue : DUF132. Ce motif est assez largement distribué, il est rencontré chez les Eucaryotes mais principalement sur les chromosomes de *Bacteria* et chez les *Euryarchaeota*. Une seule Crenarchaea, *Aeropyrum pernix K1*, possède ce motif. Cette protéine se caractérise également par la présence d'un domaine Pin (COG1569), acronyme de PilT amino terminus. C'est un domaine compact ~100 AA replié sous forme de sandwich $\alpha/\beta/\alpha$. La poche ainsi formée est le site actif, constitué de quatre AA conservés impliqués dans la chélation d'un cation bivalent Mg^{2+} ou Mn^{2+} (Jeyakanthan *et al.,* 2005). Les *Archaea* possèdent beaucoup de protéines ayant un unique domaine PIN et d'autres possédants des fusions avec des domaines ATPase ou C4 zing-finger (Makarova *et al.,* 1999). Les domaines PIN sont moins communs chez les bactéries, à l'exception d'un grand nombre chez les Mycobacteria, indépendante de l'expansion des Archaea (Arcus *et al.,* 2005) et chez les eucaryotes, malgré leur implication dans les mécanismes d'ARN interférant et des ARN messagers non sens, *non sens mediated ARNm* (Clissold *et al.,* 2000). La fonction du domaine PIN est longtemps restée incertaine même s'il semblait avoir un effet dans la signalisation, tel le répresseur de transcription StbB du plasmide pNR1 (Tabuchi *et al.,* 1992). L'activité principale de ce domaine semble néanmoins se confirmer, notamment par son activité 3'-5' exonucléase utilisée différemment selon les organismes. Chez les eucaryotes on peut citer les télomérases humaines (Takeshita *et al.,* 2007), alors que chez les procaryotes son implication dans la dégradation d'ARNm cellulaires a permis d'assigner ce domaine PIN à un nouveau type

d'opéron toxine-antitoxine VapBC (Arcus *et al.* 2005; Bunker *et al.,* 2008). L'ORF13 possède quant à lui un motif RHH de fixation à l'ADN. Ce motif très commun est également trouvé sur les antitoxines et confirmerait l'hypothèse émise. Le complexe toxine-antitoxine formé possède alors la capacité de se fixer sur le promoteur de l'opéron et inhibe l'expression du système.

Aucun des homologues, également disposés sous la forme de couple sur les chromosomes, n'est annoté comme appartenant à des systèmes toxine-antitoxine. Ces nouvelles données issues de la génomique comparée appuient pourtant cette hypothèse sur la fonction de ce couple de gènes.

L'**ORF7** possède des ressemblances avec une protéine d'initiation de la réplication acquise par transfert horizontal dans le chromosome d'*A.pernix* à partir d'un élément génétique (Piekarowicz *et al.,* 2007).

6.3 *Comparaison avec les autres plasmides de Thermococcales*

Au delà des homologues rencontrés dans les bases de données publiques, le plasmide pEXT9b possède des homologues chez les autres plasmides étudiés durant ces travaux, notamment avec les plasmides de la famille ubiquiste. Ces résultats sont consignés dans le Tableau 31.

Tableau 31 Homologues de pEXT9b rencontrés sur des plasmides de Thermococcales

pEXT9b	Homologues					
1	pIRI33	9	pCIR10	8		
3	pEXT9a	13				
5	pIRI42	3	pEXT9a	3	pAMT7	2
12	pIRI33	7				
13	pIRI33	8				

Il faut tout d'abord noter que pEXT9b est issu d'une souche possédant un autre plasmide appartenant à la famille ubiquiste, pEXT9a (page 107). Certains gènes affiliés à cette famille de plasmides ne sont pas présents sur le plasmide pEXT9a mais ils sont présents sur le plasmide pEXT9b. Nous pouvons supposer que la cohabitation de deux réplicons au sein d'une souche produit des recombinaisons aboutissant à un certain mosaïsme.

Le premier exemple concerne les ORF12 et 13, précédemment décrits. L'association de ces deux gènes avait déjà été observée en observant la répartition de leurs homologues au sein des bases de données. On peut remarquer que le plasmide pIRI33 possède également ces deux gènes. Ces homologues sont toujours strictement associés et aucun homologue d'un de ces deux gènes n'a été observé seul. Cette observation traduit la nécessité de garder associer les différents partenaires protéiques aboutissant à une fonction donnée. En effet, un gène seul, sans son partenaire, n'est pas très utile voir néfaste si l'on considère que ce couple de gènes coderait un système d'addiction au plasmide.

Le second exemple est encore plus remarquable. Il concerne l'ORF1 de pEXT9b, homologue aux ORF9 de pIRI33 et ORF8 de pCIR10 (54% identité). Ce sont des protéines ne possédant pas d'autres homologues au sein des bases de données. Avant le séquençage du plasmide pEXT9b, cette protéine semblait associée à la famille ubiquiste de plasmides. Nous pouvons nous demander dans quelle mesure la protéine codée par le plasmide pEXT9b ne proviendrait pas d'une recombinaison avec le plasmide pEXT9a ? Cette supposition est confortée par la présence de nombreuses séquences répétées de grandes tailles favorisant les évènements de recombinaison intra et intermoléculaires (Oliveira *et al.,* 2008).

6.4 *Discussion autours de pEXT9b*

Le plasmide pEXT9b est pour l'instant l'unique représentant de plasmide à hélicase MCM chez les Euryarchaea. Néanmoins, il existe chez les Crenarchaea thermoacidophiles une famille de trois petits plasmides cryptiques possédant une protéine de type RepA-MCM. Les plasmides pTIK4 et pOR1 possèdent une protéine RepA alors que pTAU4 possède une protéine MCM (Greve *et al.* 2005). La souche possédant le plasmide pTIK4 est notamment capable de surpasser toutes les autres en croissance sur boite même si l'influence du plasmide dans ce mécanisme n'a pas été confirmée (Zillig *et al.* 1998). Bien que fonctionnant différemment, les protéines RepA et MCM sont toutes les deux des P-loop ATPases de la famille AAA+, affiliées aux hélicases de la superfamille III.

Chez les eucaryotes, les complexes MCM sont des hétéromultimères composés de différentes sous-unités MCM. Chez les *Archaea*, il existe en général une seule protéine MCM active sous forme d'un homomultimère. La structure quaternaire du complexe MCM, lorsqu'un élément code une protéine homologue à celle du chromosome serait à déterminer. Il existe certainement des complexes homomultimériques et d'autres hétéromultimériques. La présence d'hétéromultimères pourrait permettre la réplication du plasmide, les sous-unités codées par le

plasmide étant capable d'interagir avec l'ADN du réplicon extrachromosomique pendant que les sous-unités codées par le chromosome garderaient leurs capacités d'interaction avec les différents partenaires protéiques nécessaires à la réplication.

Cette hélicase intervenant dans la réplication est placée dans un opéron contenant un régulateur de transcription faisant office de régulateur du nombre de copies de l'élément. Actif sous forme de dimère, Il serait capable de fixer un promoteur situé en amont et entouré par une répétition inversée.

Ce plasmide possède également un transporteur membranaire passif. Bien que le soluté transporté ne puisse être prédit par analyse *in silico*, il pourrait favoriser l'acquisition de nutriments ou d'oligo éléments. Ceci constituerait un avantage sélectif pour la souche hébergeant ce plasmide dans la compétition pour les ressources en milieu oligotrophe.

Aucun système de partition actif n'a été observé. Néanmoins un hypothétique système d'addiction peut-être annoté. Un couple de gène adjacent, dont l'un se fixe à l'ADN et l'autre possède une activité exonucléase, possède de nombreux homologues également adjacents au sein des bases de données.

Deux hypothèses peuvent-être avancées sur l'origine de ce plasmide pEXT9b. La première considérerait ce plasmide comme une relique d'un élément ancestral ayant transféré son hélicases MCM au progénote. La seconde serait que le plasmide pEXT9 appartenait à l'origine à la famille ubiquiste et l'arrivée d'un second plasmide de la même classe d'incompatibilité aurait normalement éliminé l'un des deux plasmides. Un potentiel évènement de recombinaison serait intervenu afin de créer une sorte d'hybride possédant un autre type d'hélicase, héritée du chromosome. Cette hypothèse de recombinaison peut être soutenue lorsque l'on compare d'une part le génome de pEXT9a à celui de pEXT9b et lorsque l'on regarde la famille de plasmide pORA1, pTIK4 et TAU4 de *Sulfolobus*. En effet, ces trois plasmides appartiennent à la même famille mais possèdent des hélicases différentes ; d'un côté RepA (pTIK4 et pORA1) et de l'autre MCM pour pTAU4 (Greve *et al.* 2005). Cette hypothèse de recombinaison est favorisée dans par la cohabitation dans le cytoplasme avec le plasmide pEXT9a de la famille ubiquiste et plusieurs observations supposent qu'il y a eu des transferts horizontaux entre ces deux réplicons.

III. Prototype de Puce à ADN

1. Présentation générale

Afin de poursuivre l'exploration de la diversité et de la répartition géographique des génomes d'éléments génétiques, nous avons entrepris d'utiliser les premières séquences d'éléments génétiques caractérisés pour sonder la collection d'éléments génétiques non séquencés. Pour ce faire, une puce à ADN comportant 251 gènes issus de génomes plasmidiques et du virus PAV1 a été conçue. Ayant observé de nombreuses relations entre les éléments génétiques séquencés et des éléments de type « viraux intégrés », les gènes de ces iqlots génomiques ont également été inclus sur cette puce à ADN. Nous espérons ainsi détecter des éléments génétiques apparentés à des îlots génomiques qui n'ont pas encore été rencontrés sous forme libre. La question inverse peut également se poser : existe-t-il des génomes de Thermococcales possédant des éléments génétiques intégrés ? Pour cela, des hybridations utilisant l'ADN du chromosome de certaines souches de Thermococcales de référence comme ont également été réalisées.

Cette approche est à rapprocher de la classification des plasmides réalisée lors de l'étude de l'abondance. Les hybridations ADN/ADN nous avaient permis de regrouper les plasmides en familles afin choisir des candidats au séquençage. Ayant désormais la composition en gènes de chaque plasmide, la puce à ADN devrait permettre d'avoir une idée plus fine de la présence de certains gènes sur ces plasmides inconnus.

Au sein de la puce à ADN, chaque gène est déposé sous forme de produit de PCR.L'hybridation fait intervenir deux sondes. La première sert de témoin positif, l'intensité de son signal correspondant à 100% d'hybridation tandis que la seconde est générée à partir d'ADN à étudier, plasmide ou chromosomique. L'hybridation compétitive permet la comparaison de l'intensité de signal des deux sondes. Une intensité supérieure au bruit de fond permet l'identification d'un gène présent sur un ADN que l'on souhaite étudier (Figure 57).

Cette puce à ADN est encore à l'état de prototype. Les conditions optimales de stringence permettant de diminuer le nombre de faux négatifs, sans toutefois avoir trop de faux positifs, restent à affiner. En effet, seulement deux analyses ont été effectuées. La première s'est révélée trop stringente alors que la seconde ne fut pas assez stringente.

2. Analyse préliminaire

La première analyse englobe 28 plasmides inconnus issus de notre collection (

Tableau 32). La première étape fut de s'assurer de l'efficacité de marquage, gage d'une sonde de bonne qualité. Elle est mesurée en calculant le nombre de flurophores incorporés par fragment de 100pb (

Tableau 32).

Tableau 32 Efficacité de marquage des sondes générées à partir de plasmides à étudier

Echantillon	Nbe de flurophores par 100pb	Echantillon	Nbe de flurophores par 100pb
pAMT3	25,66	pEXT5	10,12
pAMT 5	20,72	pEXT7	7,48
pAMT 8	7,49	pEXT15	9,89
pAMT 15	5,52	pGE31	5,75
pAMT 19	3,64	pIRI17	5,97
pAMT 21	5,69	pIRI 22	8,9
pAMT 51	3,54	pIRI 29	4,65
pAMT 67	4,17	pSV9	3,35
pAMT 70	5,93	pSV10	7,04
pAMT 76	8,39	pSV11	6,32
pAMT 85	9,55	pSV13	-189,38
pAMT 94	10,42	pSV17	8,58
pAMT 95	4,85	pSV18	7,49
pCIR5	9,45	pSV19	8,45

L'efficacité de marquage est comprise entre 3,35 et 10,12. Le protocole utilisé nécessite une efficacité supérieure à 2 pour réaliser une hybridation de bonne qualité. On peut toutefois noter une certaine hétérogénéité d'efficacité de marquage. Elle s'explique par des différences de rendement de récupération d'ADN marqué durant les différentes phases de lavage et de précipitation de l'ADN. Seul le plasmide pSV13 n'a pas été correctement marqué et, pour cette raison, il a été écarté de cette étude. Les hybridations ont été réalisées en dupliquat. Ces duplicatas produisent des intensités de signal comparables, confirmant l'homogénéité du signal interlames et la reproductibilité des expériences. Chaque ORF est présent en quintuplicats dispersés au sein de la lame. Chacun a produit des signaux d'intensités équivalentes et permet de s'affranchir de variation intralame ou d'éventuels « effets de bord ».

9 plasmides (33%) produisent des signaux d'hybridation, tandis que 18 (66%) ne produisent pas de signal positif. L'ensemble des gènes détectés sont consignés dans le Tableau 33.

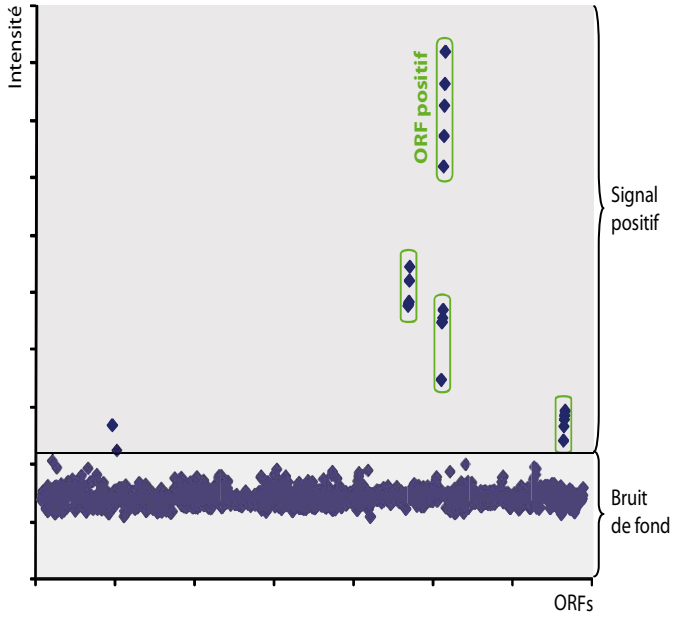


Figure 57 Interprétation des signaux d'hybridation sur une puce.

Sonde utilisée : pCIR10. Représentation de l'intensité du signal d'hybridation en fonction des différents 'ORFs présents sur la puce. Un ORFs est considéré positif, donc présent sur le plasmide inconnu testé, si, au moins 4 des quintuplicats, produisent une intensité supérieure au bruit de fond.

Tableau 33 Hybridations obtenues sur la puce à ADN

Plasmide testé	Hybridation	
	Plasmide Cible	Numéro d'ORF
pAMT5	pEXT9	4 et 5
pAMT 67	pAMT11	1à33 (tout le plasmide)
pAMT 70	pCIR10	3, 4, 5, 10, 13
	pIRI48	1à18 (tout le plasmide)
	PAV1	62, 109
	pEXT16	3
pAMT 76	pCIR10	3, 4, 13
	pIR48	1à18 (tout le plasmide)
pAMT 94	pIRI33	5
pEXT7	TKV4	1364
pIRI29	pGE2	27, 28
	pIRI42	1à13 (tout le plasmide)
pSV10	pEXT9	13
pSV19	pIRI33	1

Quatre plasmides (pAMT67, pAMT70, pAMT76 et pIRI29) possèdent tous les gènes de certains plasmides. Cette observation soulève à nouveau la question de la présence de clones au sein de collection d'isolats de Thermococcales (p.101). La détection de clones potentiels nécessite au préalable de tracer l'échantillon dont sont originaires les souches supposées clonales. **pAMT67** hybride avec tous les ORFs de pAMT11. Ces deux plasmides sont issus de souches provenant d'un même enrichissement. Néanmoins, la comparaison de leurs profils de restriction prouve que ces deux plasmides ne sont pas des clones. Cette différence de profil de restriction pourrait également être le résultat d'un ou de plusieurs gènes supplémentaires, comme le suggère l'hybridation avec l'ORF191 de PAV1. Cette hybridation n'est pas artéfactuelle et ne résulte pas d'hybridations croisées aspécifiques. En effet, cet ORF supplémentaire ne possède pas de

similarité de séquence avec pAMT11 malgré le fait qu'il code un motif très commun de type RHH-CopG, fréquemment rencontré sur les éléments génétiques.

Les plasmides **pAMT70** et **pAMT76** hybrident avec tous les ORFs de pIRI48. Les souches dont ils sont issus proviennent d'une autre campagne océanographqiue que celle ayant permis d'isoler le plasmide pIRI48 ; ce ne sont donc pas des clones de la souche IRI48. De plus, les probabilités que ces plasmides soient issus de clones est minime car ils présentent des profils de restriction différents. Ces plasmides appartiennent très certainement à la grande famille de plasmides. pAMT76 hybride en plus avec les ORFs 5 et 10 de pCIR10, l'ORF3 de pEXT16 et les ORFs 109 et 62 du virus PAV1. Ces gènes n'étant pas similaires à ceux de pIRI48, ils témoigneraient de la présence de gènes supplémentaires sur pAMT76. En dehors des gènes de la grande famille, l'hybridation avec l'ORF4 de pEXT16 suggère la présence d'une cytosine methyltransférase liée à un système de restriction-modification conférant une prédisposition au transfert de gènes. Finalement, la détection d'ORFs hybridant avec le génome de PAV1 pourrait témoigner de potentiels évènements de recombinaison entre virus et plasmides. Cette hypothèse est étayée par le fait que, au sein de la grande famille de plasmides, le sous-groupe pIRI48-pCIR10 présente de nombreux gènes homologues à PAV1.

Le dernier plasmide, **pIRI29** hybride avec l'intégralité de pIRI42, isolé d'une souche provenant du même échantillon. L'analyse du profil de restriction montre que la souche IRI29 possède en fait deux réplicons. Le premier, d'environ 12kb, possède un profil de restriction identique à pIRI42, il explique les signaux d'hybridation obtenus. Le second réplicon a une taille d'environ 2kb, il porte certainement les gènes ayant produit les signaux d'hybridation supplémentaires, c'est-à-dire avec les ORFs 27 et 28 de pGE2.

Cinq plasmides testés (pEXT7, pSV10, pSV19, pAMT5 et pEXT7) produisent seulement des signaux d'hybridation avec quelques gènes. **pEXT7** présente un signal positif avec TK1364, un ORF de la région virale intégrée TKV4, seul îlot génomique de *Thermococcus kodakaraensis* à ne pas avoir encore trouvé un équivalent sous forme libre. Néanmoins, cette région prophagique a été caractérisée comme étant affiliée aux Adénovirus (Jellyroll fold), par la présence d'une ATPase AAA+ d'encapsidation. TK1364 code une protéine possédant de fortes similarités de séquences avec les protéines P1 et P3 issues de la maturation d'une polyprotéine du virus de la mosaïque de la canne à sucre. La présence de régions de faible complexité pourrait-être expliquée cette annotation.

pSV10 hybride avec l'ORF 13 de pEXT9a, une protéine putative orpheline de 122AA. La présence de ce gène ou sein d'un nouveau génome plasmidique provenant d'un site très éloigné pourrait témoigner un transfert de gène entre la dorsale pacifique est (pEXT9a) et la dorsale pacifique sud (pSV10).

pAMT95 est un petit plasmide. La carte de restriction estime sa taille à environ 4kb, il possède mathématiquement peu de gènes. pAMT95 hybride seulement avec l'ORF5 de pIRI33, un gène conservé au sein de la grande famille de plasmide codant un régulateur de transcription à motif HTH. Il n'est pas surprenant de retrouver ce gène, en effet, il est très fréquent sur les éléments génétiques.

pAMT5 hybride avec les ORFs 4 et 5 et pEXT9a. Ces ORFs codent un système toxine-antitoxine de type RelBE. Ce module génique agit comme un module d'addiction, mobile et indépendant du type de réplicon. Il est intéressant de les retrouver associés sur un autre plasmide. Il aurait été difficilement concevable de rencontrer la toxine sans son antitoxine. (à part si elle avait été codée par le chromosome !)

pSV19 hybride avec l'hélicase de pIRI33, principal marqueur de la grande famille de plasmides. Malheureusement, aucune autre accroche avec des gènes de la famille de plasmides n'a été détectée. Nous ne pouvons exclure l'existence d'un plasmide possédant seulement l'hélicase. Néanmoins, ce n'est pas l'unique cas pour lequel nous nous serions attendus à trouver des hybridations croisées avec l'ensemble des gènes homologues. Ceci est valable pour tous les signaux d'hybridation détectés avec des gènes de la grande famille de plasmides. En effet, un plasmide hybridant avec un des ORF de la grande famille (pIRI33, pIRI48, pCIR10, pEXT9a, pAMT7) aurait dû hybrider avec les autres homologues. Quelques hybridations croisées sont cependant détectées avec les ORFs des plasmides pIRI48 et pCIR10. Ces plasmides constituant un sous-groupe au sein de la grande famille, nous pouvons penser que la stringence de notre hybridation était trop forte pour détecter des homologues possédant une séquence un peu trop divergente. Cette forte stringence limite l'intensité des signaux hybridations mais elle permet de diminuer le bruit de fond et le nombre de faux positifs. Ces résultats préliminaires peuvent donc être interprétés assez sereinement.

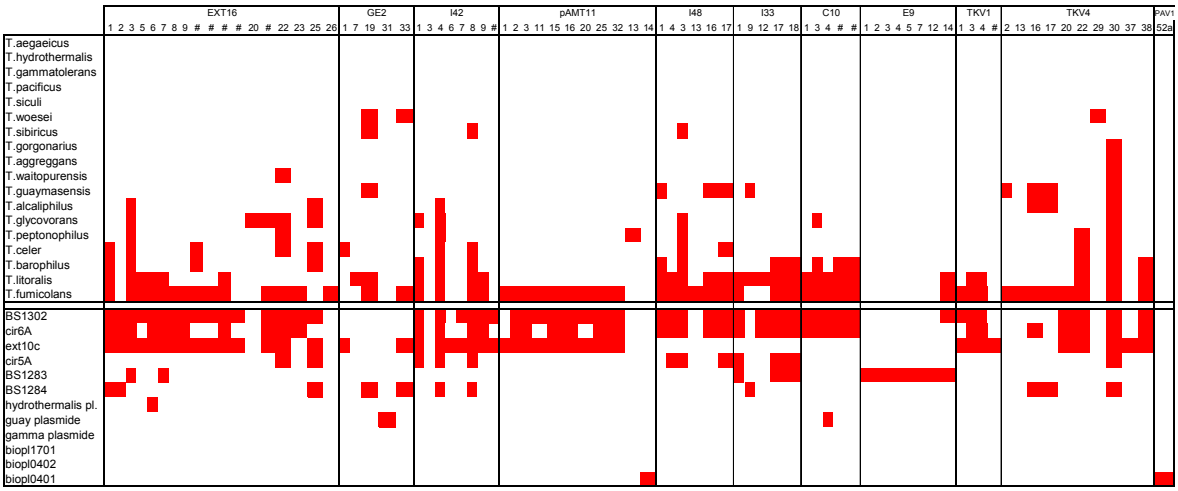
3. Analyse à stringence plus faible

Après avoir été confronté à une stringente trop forte, nous avons réalisé de nouvelles hybridations, à stringence plus faible par diminution de la salinité des tampons de lavage. Cette seconde expérience d'hybridation a étudié les plasmides pCIR5, pCIR6a, pEXT10, pBS1302, pBS1283, pBS1284, pBIOPL1701, pBIOPL0402, pBIOPL0401 ainsi que les plasmides des souches caractérisées *T. hydrothermalis, T. guaymasensis* et *T. gammatolerans*.

Le séquençage des différents plasmides de Thermococcales ayant montré de nombreuses relations avec des éléments de type virus intégrés, nous avons également décidé d'utiliser notre outil pour sonder l'ADN chromosomique de certaines souches de références : *T. aegaeicus. T. hydrothermalis. T. gammatolerans. T. pacificus. T. siculi, T. woesei. T. sibiricus, T. gorgonarius, T. aggregans, T. waiatopuensis, T. guaymasensis. T. alcaliphilus, T. glycovorans, T. peptonophilus, T. celer, T. barophilus, T. litoralis et T.fumicolans.*

Cette seconde expérience est moins concluante que la première. La diminution de la stringence a provoqué une augmentation du bruit de fond. La présence de nombreux faux positifs potentiels ne permet pas une analyse sérieuse des résultats. Nénamoins, ces hybridations positives sont consignées dans le Tableau 30.

Tableau 34 Signaux d'hybridation obtenus lors de la seconde exploration de la diversité des plasmides de Thermococcales par puce à ADN.



4. Conclusions et perspectives de l'outil puce à ADN

L'outil puce à ADN, développé pour élargir nos connaissances de la diversité des plasmides de Thermococcales et choisir de nouveaux candidats au séquençage est un semi-échec. Bien que l'idée soit intéressante, la technologie mise en place ne semble pas adaptée. Attirés par les sirènes de la modernité et les technologies proposées par les plateformes technologiques, nous avons été orientés sur la technique de microarray.

Une puce à ADN de type macroarray, utilisant des membranes de nitrocellulose comme dans notre étape de caractérisation de préliminaire, aurait été plus productive. Tout d'abord, de plus nombreux tests auraient été possibles afin de déterminer les paramètres physicochimiques optimaux. Une puce « maison » nous aurait également affranchit du travail avec une plateforme technologique, riche en contrainte administratives, logistiques et financières.

Les conditions d'hybridation de notre puce à ADN étant très stringentes il y a de fortes chances pour que le plasmide pAMT76 possède des gènes homologues au virus PAV1. Cette observation souligne la relation entre les plasmides et les virus. Il serait très intéressant de séquencer le plasmide pAMT76 pour répondre à ces questions.

IV. World of CRISPR, une boite à outils pour détecter et classer les CRISPRs

La présence de nombreuses séquences répétées sur les génomes chromosomiques et des éléments génétiques pose des questions sur leur signification biologique et leur implication dans différents mécanismes cellulaires. Afin de bénéficier de compétences en matière de linguistique et d'algorithmique, nous avons débuté une collaboration en 2005 avec l'équipe INRIA Symbiose de J. Nicolas à Rennes, dans le cadre de l'ANR Modulome. Le but de cette collaboration est le développement d'un programme d'analyse des répétitions afin d'obtenir une représentation graphique simplifiant l'interprétation pour le biologiste moléculaire.

Les CRISPRs étant une structuration particulière des répétitions, nous avons optimisé cet outil afin de détecter ce type d'élément impliqué dans l'immunité vis-à-vis des phages. (Barrangou *et al.* 2007). La nature très dynamique des CRISPRs et la diversité des gènes associés (*cas*) posent des problèmes de détection, en particulier pour les CRISPRs en début de formation ou pour ceux en train de disparaître. La détection précise des CRISPRs et de nouveaux gènes associés (*cas*) reste à optimiser afin d'établir de nouvelles hypothèses et comprendre le fonctionnement de ce système immunitaire procaryote. Dans cette optique, une formalisation de la définition de CRISPR a été nécessaire afin de permettre leur détection. La méthode choisie vise à réduire au maximum le nombre de paramètres utilisés afin de ne pas avoir d'*a priori* sur la détection de CRISPRs (taille des répétitions, variations entre répétitions, nombre minimum de répétitions dans un champ de CRISPRs…). Au cours des différentes étapes de ce projet, nous avons développé cet outil en utilisant les génomes d'*Archaea* hyperthermophiles comme modèle.

1. Détection et visualisation des CRISPRs

D'un point de vue informatique, les génomes sont des textes possédant des caractéristiques particulières. La première étape de structuration de l'information est la création d'un arbre des suffixes (Kurtz *et al.,* 1999). Par cette méthode, la séquence est décomposée en lexique contenant les différents mots présents dans cette séquence, le nombre de leurs occurrences et les coordonnées associées. Il présente également l'avantage d'avoir un temps de construction linéaire, proportionnel à la taille des mots, et non pas proportionnel à la taille de la séquence. Cette méthode est donc beaucoup moins « gourmande » en ressources de calcul que les

alignements dont le temps de calcul est proportionnel à la taille de la séquence et exponentielle du nombre de séquences à aligner.

Les répétitions de mots de grande taille ne posent pas trop de problèmes car leur probabilité d'occurrence est faible. Au contraire, les mots utilisés pour la détection des CRISPRs sont de petites tailles et possèdent des fréquences d'occurrence assez élevées, comparables à celles de mots aléatoires de taille équivalente. De plus, les CRISPRs présentent une structuration particulière des répétitions nécessitant d'être discriminées vis-à-vis des répétitions contiguës (tandem) et des répétitions distribuées aléatoirement au sein du génome. Ce problème a pu être résolu par la détection des répétitions maximales exactes dans une portée limitée, traduisant la distance entre les répétitions, sous l'appellation de *local maximum repeats.* Des variations entre les répétions ont également été tolérées afin d'être en adéquation avec les observations biologiques. En effet, bien que la plupart des répétitions soient identiques, certaines mutations ont été observées parmi les unités répétées. A titre d'exemple, CRISPRFinder (Grissa *et al.,* 2007) tolère un mésappariement dans les répétitions maximales lors de la recherche d'unités répétées. Au lieu d'introduire un seuil, il a été choisi de considérer que des unités répétées dégénérées étaient composées de deux sous-unités plus petites chevauchantes avec une unité plus grande.

Cet outil de détection a été couplé à un programme de représentation graphique basé sur le script AGE/Pyramide. Ce premier outil, disponible sous le nom de Pygram (Durand *et al.,* 2006), permet la détection des répétitions, la représentation graphique et l'extraction des séquences détectées.

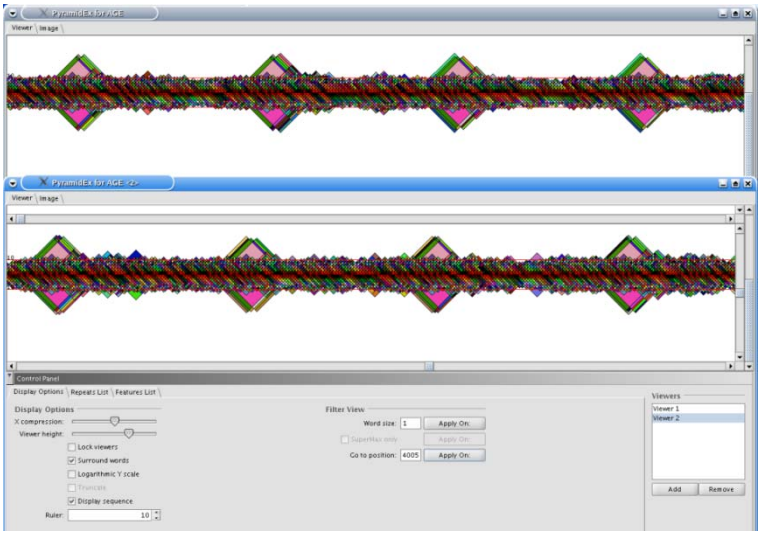


Figure 58 Capture d'écran du logiciel Pygram sur le génome de *P.abyssi*

Les pyramides de couleurs identiques représentent des répétitions.

Afin de répondre à différentes questions biologiques et formuler des hypothèses sur le fonctionnement des CRISPRs, cet outil a été amélioré afin d'intégrer les dernières données expérimentales, concernant notamment la recherche de gènes associés ou la comparaison de séquences présentes dans les CRISPRs.

2. Analyse du voisinage des CRISPRs, recherche de gènes *cas*

Il est intéressant de définir les gènes *cas* associés à un système CRISPR afin de mieux le caractériser et de formuler de nouvelles hypothèses sur le fonctionnement des protéines codées par ce système. Cette recherche s'appuie sur les gènes détectés par Haft (Haft *et al.* 2005), regroupés en différentes familles composées de 45 sous-types de gènes *cas*. Ces familles ont été établies par une approche de recherche de chaines de Markov cachées (HMM). L'alignement des protéines conservées a ainsi permis la définition d'un profil HMM pour chaque famille.

La robustesse des profils HMM définis par Haft a été vérifiée. Pour cela, les 84 séquences de protéines CAS de Thermococcales annotées ont été rapatriées. Ces séquences ont été utilisées pour procéder à des recherches de similarité (BlastP) contre tous les génomes de Thermococcales afin de vérifier si d'autres protéines CAS, ne répondant par forcément au consensus du profil HMM, pouvaient être détectées (Figure 59).

323 protéines sont ainsi détectées dans les génomes de Thermococcales. Afin de ne garder que les vrais positifs, deux filtres ont été utilisés. Le premier est basé sur un critère de taille, il permet d'éliminer une partie des protéines pouvant générer des faux positifs, notamment par la présence de motifs fréquents (Walker, RHH…). Le second traduit la présence de ces gènes à proximité d'un champ de CRISPR.

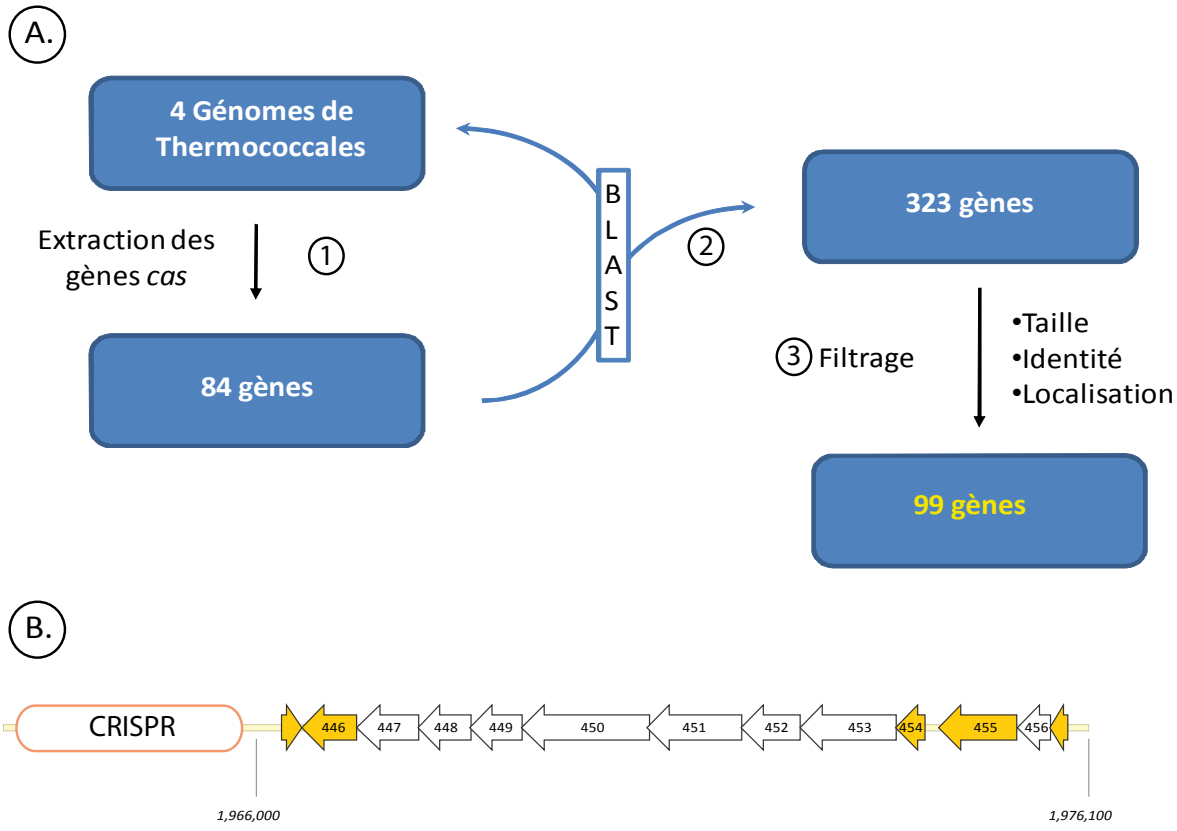


Figure 59 Stratégie utilisée pour la détection de nouveaux gènes *cas*

A. Schéma de la stratégie utilisée pour détecter de nouveaux gènes *cas* chez les Thermococcales.

B. Exemple de nouveaux gènes cas détectés au sein de *T. kodakaraensis* (jaune). Les autres gènes (blanc) ne sont pas des gènes *cas*, ou bien, ils sont déjà annotés *cas*.

L'utilisation des filtres réduit le nombre de protéine détectées à 99, soit 15 protéines de plus que celles annotées dans les génomes de Thermococcales par la recherche de profils HMM. L'appartenance de ces 15 nouvelles protéines à la famille des CAS a été validée par alignement avec les protéines CAS ayant servies de requête. Un exemple est donné pour *T.kodakaraensis* (Figure 59 B), où cinq nouveaux gènes *cas* ont pu être identifiés.

Cette analyse a également servi à calibrer notre outil de détection des gènes *cas*. La recherche de gènes *cas* n'a pas été effectuée sur l'intégralité du génome, mais seulement sur les 35kb situés à proximité d'un CRISPR détecté. Cette limitation de la zone de recherche permet l'utilisation de valeurs seuils plus larges lors de la recherche par HMM, augmentant ainsi le nombre gènes trouvés tout en diminuant le nombre de faux positifs. L'autre résultat intéressant est la détection par cette méthode d'un gène *cas1* chez *P.abyssi*, alors que les autres outils prédisaient que c'était le seul génome à ne pas posséder ce gène *cas1*.

3. CRISPI, une base de données sur les systèmes CRISPRs

L'algorithme développé, pour la recherche des CRISPRs et des gènes *cas* associés, a été employé sur l'ensemble des génomes de procaryotes disponibles dans la Genbank en date du 29 octobre 2008 (release 167.0). Le processus a été automatisé afin que la base de donnée soit mise à jour dès qu'une nouvelle version de la Genbank est disponible. Ce processus d'automatisation nous a semblé important afin de garantir l'actualisation de la base de données. En effet, de nombreuses bases de données génomiques sont créées et ne sont pas forcément mises à jour. A ce titre, je souhaiterais prendre l'exemple de la DPR *Database of Plasmid Replicon*, une très bonne initiative de classification des plasmides en fonction du réplicon utilisé. Malheureusement, cette base de données n'est plus maintenue depuis le mois de septembre 2000 lui faisant perdre toute utilité.

Notre analyse a porté sur 53 génomes d'*Archaea* et 707 de *Bacteria* (Tableau 35)

.

Tableau 35 Recherche de CRISPRs dans les génomes de procaryotes

	Archaea	*Bacteria*
Génomes Analysés	53	707
CRISPRs détectés	291	1941
Génomes sans CRISPRs	4 (5,5%)	131 (18,5%)
Unités détectées	5955	22788
Spacers détectés	5664	20847
Gènes cas putatifs	948	ND

Environ 95% des génomes d'Archaea possède un système CRISPR. Cette valeur est comparable à celles énoncées dans la littérature. L'analyse effectuée sur les génomes de *Bacteria* détecte des systèmes CRISPRs dans 81,5% des génomes. Cette proportion est environ deux fois plus élevée que celles déterminées par les autres outils de recherche de CRISPRs. Ces pourcentages et nombres de systèmes CRISPRs sont toutefois à analyser avec précaution. L'abondance plus élevée détectée avec notre outil peut s'expliquer par notre méthode visant à utiliser le moins de paramètres possibles afin de ne pas avoir d'*a priori* sur ce type de structure. Il y a donc un certain nombre de faux positifs mais également des CRISPRs en cours de formation ou en sénescence. La détection de ce type de structures « dégénérées » est primordiale pour comprendre la dynamique des systèmes CRISPRs. De nombreux gènes *cas* supplémentaires ont été détectés. Une analyse

complémentaire sera nécessaire afin de définir de nouveaux profils HMM et éventuellement de nouveaux sous-type *cas*, notamment lorsque l'on considère que ceux actuellement définis ne permettent pas la détection du gène *cas1* de *P.abyssi*.

4. World of CRISPR, une interface web disposant de nombreux outils pour le biologiste moléculaire

Etant conscient qu'il existe une séparation entre les personnes produisant les séquences et celles développant les outils d'analyse, il est apparu très utile d'interfacer l'outil de recherche de CRISPRs afin que toute personne, ne possédant pas de connaissances en informatique puisse utiliser cet outil sans avoir à taper la moindre ligne de commande. Nous autres, biologistes moléculaires y sommes souvent réfractaires.

Le site web est disponible à l'adresse : http://crispi.genouest.org/

Cette interface permet tout d'abord la visualisation des CRISPRs et des gènes *cas* pour un génome d'intérêt présent dans la genbank. Un exemple est donné avec le CRISPR3 de *P.furiosus* sous forme de captures d'écrans dans la Figure 60. Elle permet également d'effectuer la recherche de CRISPR et de gènes *cas* pour un nouveau génome. Pour cela une fenêtre permet l'importation d'une séquence (Maximum 10Mb) et l'analyse est effectuée en environ 10 minutes (pour 2Mb). Les résultats sont finalement présentés sous la même forme que les génomes précalculés (Figure 60).

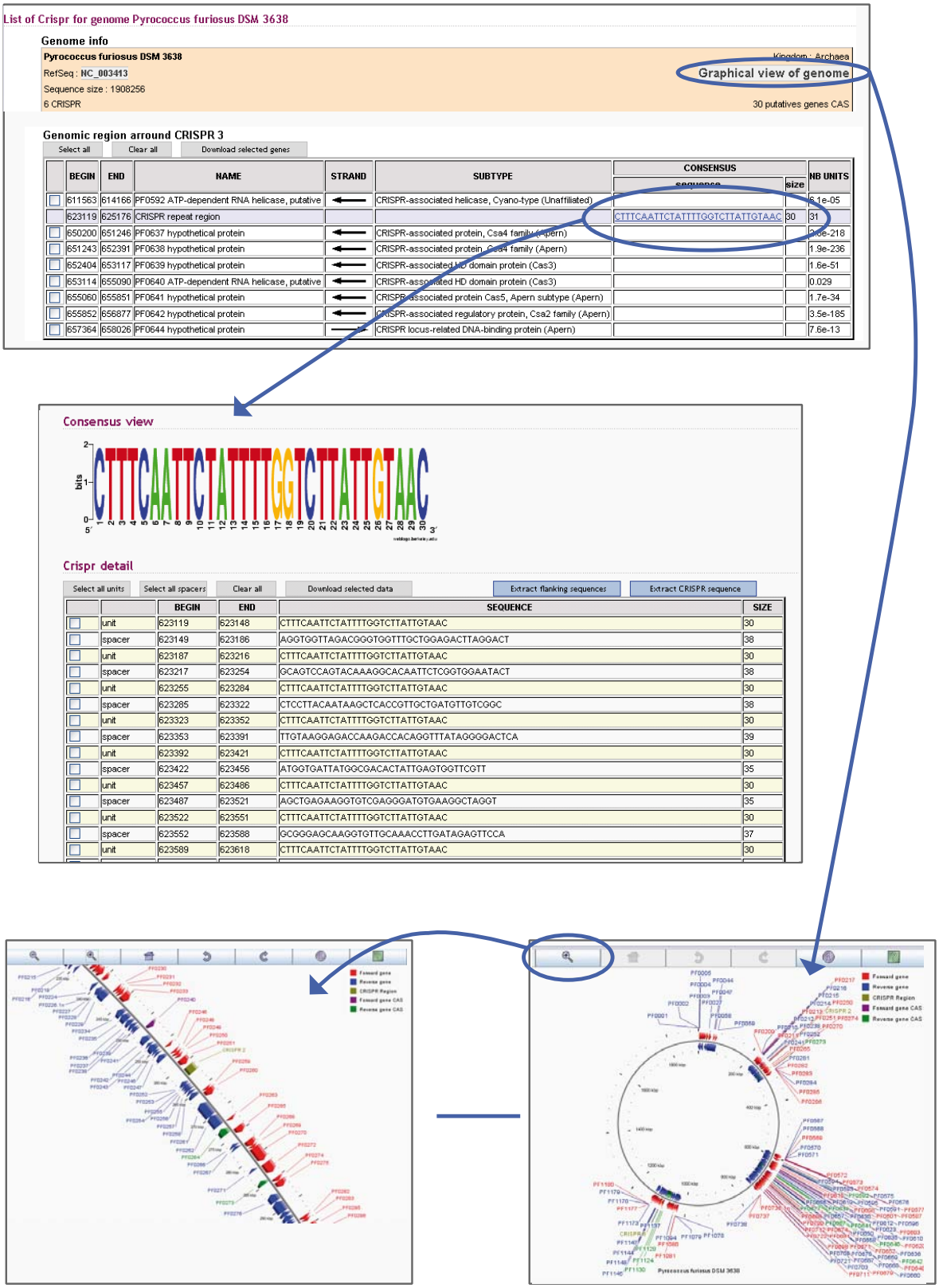


Figure 60 Capture d'écran de l'interface web permettant la consultation d'un système CRISPR

Notez la présence de multiples boutons permettant de rapatrier les différentes séquences : spacer, répétitions, gènes *ca.* De nombreux liens hypertextes sont fournis, notamment vers la Genbank pour consulter les génomes et protéines.

Afin de répondre aux attentes des personnes étudiant les éléments génétiques, qui souhaitent par exemple prédire la résistance d'une souche vis-à-vis d'un virus ou trouver des traces d'infections passées dans un génome, un outil permettant la comparaison d'une séquence (virale) vis-à-vis de la banque de spacers a été développé. Il est basé sur l'algorithme BLASTn avec des paramètres adaptés pour la recherche de similarités sur des séquences de petite taille. Pour une plus grande souplesse d'utilisation, ces paramètres peuvent également être ajustés par l'utilisateur.

Nous avons appliqué cet outil aux Thermococcales afin de savoir si les éléments génétiques séquencés pendant cette étude avaient des séquences similaires au sein de la base de données contenant les séquences répétées et les spacers de CRISPRs. Cette information pourrait refléter la trace d'une infection par cet élément génétique et la résistance de la souche. Un cas unique a été observé. Il y a fort à penser qu'en possédant plus de génomes de Thermococcales, et en particulier de souches isolées du même endroit que les éléments génétiques, nous aurions détecté de plus nombreuses similarités, comme cela a été observé lors de la comparaison du génome du virus SRV avec les spacers de 11 génomes de Sulfolobales (Vestergaard *et al.* 2008).

La similarité que nous avons détecté concerne le plasmide pGE2, issus de *P. abyssi* GE2. Une séquence de 27pb de l'ORF15 possède 100% d'identité de séquence avec un spacer du CRISPR3 de *P. abyssi* GE5. De plus, sur pGE2, il est possible de détecter la séquence particulière GAAAC, en amont de ce spacer, qui servirait de site de reconnaissance pour l'acquisition et/ou l'inactivation par le système CRISPR. Ce résultat est très intéressant car *P.abyssi* GE2 et *P.abyssi* GE5 sont issus du même échantillon. *P.abyssi* GE5 ne possède pas cet élément génétique intégratif, aussi bien sous forme libre que sous forme intégrée. La présence d'un spacer identique à pGE2 supposerait que *P.abyssi* GE5 a été infecté par un « virus » similaire à pGE2 mais que la souche a réussi à contrecarrer cette infection par l'action d'un système CRISPR.

5. Discussion sur les CRISPRs et sur l'outil World Of CRISPRs

Afin d'étendre les fonctionnalités de l'outil World Of CRISPR et de répondre à certaines questions biologiques, quatre chantiers sont actuellement en cours.

Le premier concerne l'amélioration de la méthode de détection de similarités entre un spacer et une séquence virale. En effet, le virus peut contrecarrer l'inhibition par le CRISPR grâce des mutations synonymes (Vestergaard *et al.* 2008). La recherche de similarité pourrait être

accomplie en utilisant la séquence des protéines d'un virus et les comparer au spacer traduit dans les 6 cadres de lecture.

Le second est la mise à disposition d'un outil facilitant le splogotypage, une technique de typage infraspécifique reposant sur l'évolution très rapide des séquences spacer au sein d'un champ de CRISPR. En effet, plusieurs études d'épidémiologie sont basées sur la caractérisation d'isolats bactériens par séquençage de loci CRISPRs et permettent la discrimination de souches très proches (Mokrousov *et al.,* 2007). Considérant que notre outil permet la recherche de CRISPR et l'extraction des spacers, un utilisateur souhaitant typer plusieurs souches pourrait fournir plusieurs séquences de loci CRISPRs et obtenir immédiatement une représentation graphique des spacers partagés ou différents entre souches. La filiation entre souches sera représentée sous forme d'un dendrogramme.

La troisième évolution vise à caractériser la séquence leader lorsque plusieurs systèmes CRISPRs sont présents dans un génome. Aucune étude n'a pour l'instant été réalisée sur la séquence leader. Néanmoins, elle semble importante pour l'insertion de nouveaux spacers au sein d'un champ de CRISPR. Elle pourrait contenir des sites de fixation de protéines CAS indispensables à l'insertion de nouvelles séquences et également jouer le rôle de promoteur à la transcription du locus. La recherche exhaustive de ces séquences permettrait la définition de motifs fonctionnels, de tester des hypothèses d'interaction avec des protéines CAS et ainsi accroître nos connaissances sur le fonctionnement du système CRISPR.

Le quatrième chantier concerne un site putatif de fixation des protéines CAS. En effet, chez *Streptococcus*, le motif GAAAS situé en amont d'une séquence phagique identique à un spacer est indispensable à l'inactivation du phage. Malheureusement aucune analyse à grande échelle n'a été menée sur le contexte génomique lorsqu'il existe des similarités de séquence entre un spacer et un élément génétique. Cette évolution de notre outil permettrait la prédiction de nouveaux motifs et d'établir une corrélation avec la présence de certains sous-types de protéines CAS.

CONCLUSION ET PERSPECTIVES

Forte des connaissances acquises depuis 60 ans, initiées par la découverte du support de l'hérédité et de la structure de l'ADN, la microbiologie des *Bacteria* est entrée dans l'ère post-génomique grâce à l'accumulation de données physiologiques, biochimiques et génomiques. Ceci a permis d'émettre de nombreuses hypothèses issues de l'analyse des données génomiques qui ont été par la suite testées et confirmées grâce au développement d'une multitude d'outils moléculaires. Ces informations apportent une nouvelle couche de complexité dans la compréhension de l'évolution des procaryotes. L'avènement de nouvelles technologies de séquençage à haut débit permet aujourd'hui d'appréhender le génome sous un angle dynamique mettant en lumière l'importance des flux de gènes dans le modelage des chromosomes bactériens. L'arbre du vivant, basé sur le gène de l'ADNr16S, doit être considéré comme une trame générale de l'évolution à laquelle s'ajoute l'impact de nombreux transferts horizontaux. Cet apport latéral de matériel génétique est la principale source d'innovation, il constitue une clé permettant de comprendre l'adaptation et la spéciation des micro-organismes. Les principaux vecteurs de ces transferts sont les éléments génétiques. Les phages, et en particulier leur forme intégrée prophagique, jouent un rôle prépondérant dans l'adaptation des micro-organismes au sein de différentes niches écologiques en augmentant leur fitness. Leur abondance, sous forme intégrée dans les génomes, peut s'accompagner d'éventuels avantages apportés à leur hôte. Ils sont d'importants acteurs de variabilité inter-espèce et certainement le principal moteur de la spéciation. A titre d'exemple, ce phénomène est bien illustré dans les génomes modèles de *Listeria* ou des *Streptococcus agalactiae*. Lorsqu'une bactérie lysogène est soumise à des changements de conditions environnementales, les gènes prophagiques montrent en premier des changements de leur expression. Ces résultats suggèrent que les prophages ne sont pas seulement des cargos génétiques passifs du chromosome bactérien mais qu'ils contribuent vraiment à la physiologie de leur hôte. (Peut-être pour eux-mêmes se protéger, mais c'est une autre question évolutive!). Grâce aux importantes quantités de données génomiques accumulées, la classification des éléments génétiques au sein des « compartiments » plasmides ou virus s'est affinée afin de devenir assez claire. Néanmoins, la nature de certains éléments peut être plus complexe, à l'image des prophages P1 et N15 ou des prophages de *Borrelia* présentant d'intrigantes relations avec les plasmides. Finalement, plasmides et virus forment un *continuum* par échange de groupes de gènes lorsqu'ils se rencontrent au sein d'une cellule.

L'autre domaine procaryotique, les *Archaea*, semble évoluer suivant les mêmes préceptes. Néanmoins, leur découverte plus récente et la présence de mécanismes de maintenance de l'information génétique *fondamentalement* différents de ceux des *Bacteria*, font de ce domaine une *terra quasi incognita* concernant les mécanismes permettant le transfert et la capture de gènes. La diversité de leurs éléments génétiques est à ce jour très hétérogène en fonction des groupes taxonomiques. Cette connaissance est pourtant un prérequis indispensable à la compréhension de l'évolution et de l'adaptation de ces micro-organismes. Ceci nécessite le développement de nouveaux outils génétiques qui ne pourront être créés qu'à partir d'éléments génétiques propres aux *Archaea*. Comme les *Archaea* partagent avec les eucaryotes certains des processus fondamentaux de maintenance et de réplication de l'ADN, elles sont également considérées comme des modèles indispensables à la compréhension des organismes eucaryotes.

Les Thermococcales constituent un modèle d'étude archéen séduisant. Ce sont des Euryarchaea colonisant des biotopes extrêmes caractérisés par de fortes températures, pressions, concentrations en métaux lourds et/ou radioactivités. Elles sont ubiquistes au niveau des sources hydrothermales océaniques profondes, mais également rencontrées au niveau de certaines sources côtières et terrestres. De récentes études montrent qu'elles sont également majoritaires dans les couches les plus profondes de la subsurface. Bien qu'essentiellement chimioorganotrophes, certaines souches, à l'image de *T.onnurineus*, sont également capables d'utiliser le monoxyde de carbone comme source de carbone et d'énergie. Actuellement, cinq génomes sont disponibles. Leur comparaison révèle une plasticité conduisant à un dynamisme qui semble principalement lié à la présence d'éléments génétiques intégrés. L'impact de ces éléments est notamment illustré chez *P. furiosus*, dont le génome apparaît activement remanié par l'activité de transposons. De même, il apparaît que chez *T.kodakaraensis* la majorité des gènes spécifiques de cette espèce sont en réalité localisés dans des îlots génomiques.

Au contraire des connaissances acquises chez les Sulfolobales, pour lesquelles une large variété de plasmides et de virus ont été décrits, seuls quatre génomes d'éléments génétiques de Thermococcales sont disponibles : trois plasmides cryptiques à réplication par cercle roulant et le virus PAV1.

Le sujet de cette étude visait à étendre nos connaissances sur l'abondance, la diversité et la biogéographie des éléments génétiques de Thermococcales. Pour cela, une collection d'environ 300 isolats de Thermococcales a été criblée à la recherche d'éléments génétiques, montrant qu'environ 1/3 des isolats possédaient au moins un réplicon extrachromosomique. Une analyse

préliminaire a permis d'établir une classification de ces éléments. Elle montre qu'une majorité de ces éléments forme une vaste famille ubiquiste. D'autres éléments génétiques représentent au contraire d'uniques représentants d'hypothétiques familles. Afin d'accéder à la diversité de ces éléments génétiques à travers les océans, cinq génomes de la famille ubiquiste et six génomes « orphelins » ont été séquencés et analysés.

Le premier enseignement de ce travail montre qu'il n'est pas encore possible d'aborder la notion de biogéographie qui avait été posée lors de l'initiation de cette thèse. Ce volet ne pourra pas être étudié tant que certaines bases indispensables ne seront pas validées, telles que la biogéographie des hôtes qui ne peut être résolue sur la base du seul marqueur phylogénétique ADNr16S. Il possède en effet un faible pouvoir résolutif intra-genre dû au biais introduit par la forte composition en G+C indispensable à la cohésion structurale de l'ARNr16S à haute température ainsi qu'à la présence de mécanismes de réparation de l'ADN favorisant les mutations T->C. La phylogénie des espèces thermophiles présente des branches plus courtes et de nombreux évènements de réversion de mutations masquent en partie le signal phylogénétique. De plus, l'étude phylogénétique de nombreux gènes n'est pas forcément congruente avec celle basée sur le marqueur universel. La seconde raison est le que le mode de dissémination de ces organismes reste à ce jour inconnu. En effet, la découverte récente de Thermococcales dans les échantillons de subsurface profonde suggère l'existence de niches évoluant indépendamment, peut-être pendant des millions d'années et qui peuvent-être libérées lors d'un bouleversement géologique.

L'analyse des génomes des éléments génétiques étudiés révèle l'existence d'une grande diversité des réplicons extrachromosomiques de Thermococcales. Ils sont globalement différents de ceux observés chez les Crenarchaea thermophiles. L'absence de similarité entre les éléments génétiques de ces deux phyla suggère une évolution indépendante des réplicons extrachromosomiques suite à la divergence entre les *Euryarchaea* et *Crenarchaea*. Cette évolution divergente pourrait résulter d'une barrière créée par la nécessité pour certains de ces réplicons de recruter des partenaires protéiques de l'hôte cellulaire afin de se répliquer. Néanmoins, nous avons détecté la présence de similarités avec certains génomes d'Euryarchaea méthanogènes mésophiles. La notion de « qui se ressemble s'assemble » semblerait plus se rapporter aux propriétés phylogénétiques qu'à la présence de caractères communs de modes de vie en milieu thermophile.

Le second constat est la présence de nombreux gènes orphelins. Malgré l'accumulation des données génomiques, la quantité de gènes orphelins ne diminue pas. Au sein des génomes chromosomiques, la fonction et l'origine de ces gènes orphelins reste un énigmatique problème de la biologie moléculaire moderne. En effet, ces orphans peuvent représenter une fraction importante des gènes spécifiques entre espèces proches. L'importance de ces gènes dans la spéciation et dans l'adaptation est primordiale.

L'origine de ces orphans est très discutée, néanmoins, leur abondance plus importante sur les génomes de virus et plasmides suggére que les éléments génétiques sont des usines à innovation permettant de tester de nouvelles combinaisons génétiques. Plusieurs hypothèses ont été émises sur l'origine de ces orphans, elles classent les orphans en trois classes concordantes avec les observations effectuées sur les éléments génétiques séquencés. La première hypothèse serait une sur-annotation des génomes procaryotes, c'est-à-dire qu'une partie d'entre eux ne seraient pas de vrais gènes mais des artefacts ou des pseudogènes résultant de recombinaisons ou de mutations. Ce phénomène explique que le nombre de gènes orphelins continue à augmenter malgré l'accumulation des données génomiques. Lors de l'annotation de nos éléments génétiques, nous avons recherché des signaux de traduction et de transcription. La plupart des ORFs qui ne possèdent pas ces signaux sont des gènes orphelins. De plus, ces gènes sont généralement localisés dans des portions précises du génome, de petite taille et possèdent d'importantes quantités de séquences répétées. Ces observations, en accord avec les données bibliographiques précédemment citées, traduiraient la présence de faux gènes parmi ces orphelins. Leur existence pourrait s'expliquer par la grande plasticité des génomes d'éléments génétiques et de fréquents réarrangements conduisant à la création de nombreux gènes hybrident, certains fonctionnels et sources d'innovations génétiques, mais surtout à une énorme quantité de gènes « déchets » que l'on peut qualifier de pseudogènes ou d'artéfact d'annotation.

La seconde hypothèse sur l'origine des gènes orphelins serait une accélération de leur vitesse évolutive ne permettant plus de retrouver des homologues dans les bases de données. La dernière catégorie de gènes orphelins est composée de gènes n'ayant pas encore d'homologues séquencés. En considérant cette dernière catégorie vis-à-vis des plasmides étudiés, on remarque que les plasmides de la famille ubiquiste possèdent moins de gènes orphelins que les autres. Cette observation s'explique par la présence de gènes homologues entre plasmides mais qui ne possèdent pas d'autres homologues dans les bases de données. Il serait donc très intéressant d'isoler de nouveaux plasmides apparentés à ceux qui sont d'uniques représentants d'hypothétiques familles. L'obtention et le séquençage de nouveaux génomes homologues

permettraient de diminuer ce nombre de gènes orphelins et également de définir les gènes conservés au sein de ces familles, supposés indispensables à une activité commune (module). De plus l'alignement de ces protéines homologues doit permettre la prédiction de nouveaux motifs fonctionnels. La détermination biochimique de leur fonction est très importante, car ils pourraient constituer un réservoir de nouvelles familles de protéines. Ils pourraient être élaborés sur les éléments génétiques, possédant une grande plasticité et par la suite véhiculés par transfert horizontal sur le chromosome. Afin de comprendre le fonctionnement de ces éléments génétiques et accéder à un réservoir potentiel de nouvelles familles de protéines, les séquences issues de ces plasmides ont servi à alimenter un programme d'expression de protéines *in vitro*, de caractérisation fonctionnelle et de cristallographie (ANR Genoarchaea en partenariat avec P. Forterre et H. Van Tilbeurgh)

Le critère fondamental de classification des réplicons repose sur la combinaison des gènes participant à la réplication. Cependant, au cours de ce travail, nous avons mis en évidence la présence d'éléments génétiques possédant des opérons réplicatifs différents alors qu'un large cluster de gènes, de fonction inconnue, était en revanche conservé. Cette observation illustre la présence de différents modules géniques possédant chacun une fonction particulière, et la **nature mosaïque** des éléments génétiques. Cette vision des génomes est en accord avec la théorie du gène égoïste de Dawkins. Cette théorie est souvent poussée à la caricature en ne prenant en compte que son illustration par les transposons. Néanmoins, appliquer cette théorie aux réplicons extrachromosomiques revient à ne plus considérer chaque gène comme un individu, mais à considérer ces modules géniques comme des communautés, inscrites dans une méta-communauté qu'est le génome. C'est sur cette base que l'appartenance d'un élément génétique à la classe des plasmides ou des virus peut-être définie. Une famille de plasmides se caractérise par un module réplicatif commun et l'absence de module permettant la production de particules virales, alors que l'appartenance au monde viral serait basée sur la composition de ce module permettant la production des particules virales.

Cet aspect peut être illustré par la description dans ce travail d'une famille ubiquiste de « plasmides ». L'analyse comparative des génomes de cette famille révèle une structure réplicative commune, impliquant un nouveau type d'opéron réplicatif. Il se caractérise par la présence d'une hélicase de la famille UvrD et d'une ADN primase-polymérase appartenant à une nouvelle famille, démontrant le potentiel de réservoir génique des éléments génétiques. La présence des trois fonctions indispensables à la réplication confère certainement une plus grande

capacité d'établissement de ce genre de réplicon suite à un transfert horizontal. En effet, cela minimise la nécessité de recruter des partenaires protéiques codés par son hôte. En dehors de la présence d'une excitante nouvelle famille d'ADN primase-polymérase, ce sont les premiers éléments génétiques à posséder ce type d'hélicase qui, habituellement, est présente chez presque toutes les *Bacteria*, alors que les seuls homologues archéens sont seulement portés par des réplicons chromosomiques d'Euryarchaea halophiles qui sont les micro-organismes portant le plus grand nombre de réplicons au sein des *Archaea*. La distinction entre chromosome, minichromosome, megaplasmide et plasmide n'est pas encore clairement définie. Ce manque de connaissance est en partie lié aux difficultés d'étude de la réplication parmi tous ces réplicons possédant une forte plasticité liée à la fréquence des recombinaisons et des transferts de gènes entre réplicons. Néanmoins, la cohabitation de plusieurs réplicons au sein d'une cellule soulève des questions sur la mise en place de différents mécanismes de maintenance assurant la stabilité de chacun de ces réplicons. L'étude de l'hélicase homologue portée par les plasmides de Thermococcales pourrait certainement aider à mieux comprendre leur fonctionnement chez les halophiles.

D'un point de vue évolutif, la découverte d'éléments génétiques à hélicase UvrD et ADN primase-polymérase pose de nombreuses questions sur leur(s) origine(s). Cette hélicase étant affiliée à celle des *Bacteria*, leur existence au sein des halophiles pourrait résulter d'un transfert horizontal à partir des *Bacteria*. Inversement, il est possible que ce type d'hélicase ait été présent par le passé sur le chromosome, et qu'il ait été perdu après un éventuel transfert vers les plasmides. Néanmoins, une hypothèse alternative doit être considérée après une analyse plus poussée de cette famille de plasmides. En effet, cette famille peut être scindée en deux sous-groupes distincts ne résultant pas d'un isolement géographique. L'un des sous-groupes possède en plus des gènes homologues au virus PAV1 intercalés au sein de l'opéron réplicatif. On peut éventuellement penser que ces éléments sont apparentés et dérivent d'un ancêtre commun viral. Dans ce cas, cela signifierait que le virus PAV1 aurait perdu l'hélicase qui aurait été remplacée par un autre gène codant une protéine de fonction analogue ou capable de recruter l'hélicase de l'hôte cellulaire.

Assigner ces réplicons au sein de la classe des plasmides ou des virus n'est donc pas évident. Admettons qu'après avoir caractérisé le génome d'un élément génétique, nous n'observions pas de particules virales, ceci ne signifierait pas forcément que ce n'est pas un virus. En effet, les conditions expérimentales utilisées pour effectuer la recherche de ces particules virales sont en général celles utilisées pour la croissance optimale de la souche hôte. Or, dans ce cas le virus

profite de la multiplication de son hôte pour également se multiplier, n'attendrait-il pas que celle-ci soit stressé pour tenter de sortir de la cellule ? Quand est-il lorsqu'une mutation intervient sur le génome et empêche par exemple la formation de particules virales ? Le génome peut continuer à se maintenir dans la cellule sous la forme épisomale, mais doit-on pour autant parler de plasmide ? Cet exemple illustre la frontière ténue entre plasmide et virus. Elle suggère également une hypothèse sur l'origine des plasmides. Ils dériveraient de virus dégénérés et/ou domptés par la cellule. A titre d'exemple prenons PAV1, seul virus de Thermococcales décrit à ce jour, qui a longtemps été considéré comme un plasmide de la souche *P. abyssi* GE23.

De plus, nous avons également observé que deux éléments génétiques, pAMT11 et pGE2, étaient apparentés à des îlots génomiques. Les homologies entre ces éléments et les îlots génomiques inclus une large portion du génome constituant un module qui n'est pas impliqué dans la réplication. Après analyse de ces gènes conservés, un faisceau d'indices laisse penser que pAMT11 et pGE2 sont certainement des virus. Il est indispensable de rechercher des particules phagiques afin de confirmer l'implication de ce module de gènes dans la production et l'assemblage de protéines virales.

Les plasmides étant par définition des réplicons, il n'est pas surprenant de trouver d'autres familles de gènes impliqués dans la réplication. Parmi les plasmides étudiés, deux possèdent des gènes normalement codés par le chromosome : des protéines initiatrices de la réplication de type Orc1/Cdc6 chez pEXT16 et une hélicase MCM chez pEXT9b. En plus de ces protéines, pEXT16 possède une origine de réplication similaire à celle des chromosomes de Thermococcales. La notion de module ne doit pas uniquement considérer la présence de certains gènes, mais également la présence de séquences non codantes participant intégralement dans la fonction à laquelle est dévolue le module fonctionnel. Outre le fait qu'ils sont l'illustration d'une grande diversité, ces plasmides posent aussi des questions fondamentales sur la cohabitation de ces éléments génétiques avec leurs homologues chromosomiques. Ils seraient également de très bons modèles d'étude de la réplication des Euryarchaea, en particulier pEXT16 qui possède à la fois des protéines initiatrices de la réplication et une origine de réplication typique du chromosome. Cette découverte pose également des questions sur la présence et l'origine de multiples origines de réplication chez les *Archaea*. Ces plasmides sont-ils issus d'une origine de réplication chromosomique devenue indépendante ? Ce type de réplicon contribue à la présence de plusieurs origines de réplication chez les *Archaea* ? Et dans cette mesure, suivant les hypothèses émises par P. Forterre, sont-ils une relique du système ancestral qui a été transféré au

chromosome du progénote ? Cette question mérite d'autant plus d'être posée suite à la découverte qu'une des origines de réplication de *Sulfolobus* a été capturée d'un élément extrachromosomique.

Au-delà de ces nombreuses questions, certains axes prioritaires peuvent d'ores et déjà être définis afin d'améliorer notre connaissance des éléments génétiques *d'Archaea*.

Afin de mieux caractériser la diversité des éléments génétiques, il faudrait tout d'abord détecter et séquencer des éléments génétiques apparentés aux « plasmides » qui n'ont pas encore de « petits frères ». Ces nouvelles données permettraient la définition d'un squelette génique minimal, supposé indispensable à la maintenance de ces familles de réplicons extrachromosomiques. Ces gènes « indispensables » pourraient ensuite être utilisés comme marqueurs afin de classer rapidement un nouvel élément génétique inconnu. L'obtention de nouveaux gènes homologues permettrait également la définition de motifs conservés et ainsi orienter les expériences de caractérisations fonctionnelles. La définition de gènes essentiels à un type de réplicon pourrait également permettre la création et l'optimisation de nouveaux outils génétiques assez limités pour le moment chez les Thermococcales.

Un autre point essentiel est l'étude de la production de vésicules contenant de l'ADN, observées chez les Thermococcales et les Methanococcales. La présence d'un îlot génomique homologue entre ces organismes et affilié à l'élément génétique intégratif apparenté aux Adénovirus pourrait-être la réponse à ces observations. Ce mécanisme ressemble en effet fortement au système GTA, *Gene Transfer Agent*, décrit chez *Rhodobacter capsulatus*, où la présence dans le chromosome de gènes hérités d'un phage permet la production de particule « virus-like » transférant de l'ADN aléatoire. Je souhaiterai également faire part de mes interrogations sur l'observation de cellules de Thermococcales ayant la capacité de fusionner lorsqu'on les soumet à un stress endommageant l'ADN. Au-delà de savoir ce qu'il advient des chromosomes une fois la fusion de cellules effectuée, un mécanisme d'origine phagique pourrait coder des protéines membranaires permettant cette fusion. Des recombinaisons homologues entre les deux chromosomes permettraient de régénérer le patrimoine génétique original, à l'image du processus décrit chez les *Sulfolobus* lors d'un stress oxydatif, provoquant une agrégation des cellules et l'expression de gènes de conjugaison, autorisant la mise en commun du patrimoine génétique afin de reconstituer un chromosome viable.

Un autre point d'intérêt illustrant la plasticité des génomes des éléments génétiques mobiles et les transferts horizontaux concerne le gène codant l'intégrase de pGE2. Il est le plus proche homologue du gène permettant au virophage Sputnik de s'intégrer dans le génome de Mamavirus. Il a clairement été établit que ce gène avait été transféré par voie horizontale à partir des *Archaea* vers les virus eucaryotes. La découverte d'un élément génétique intégratif chez les Euryarchaea élargit les potentialités d'origine de ce gène.

En conclusion, le XXIème siècle s'annonce comme étant l'ère de la révolution des technologies de l'information. La plus grande d'entre elle pourrait-être l'acquisition des données génétiques permettant d'appréhender le vivant sous un nouvel angle. La connaissance bénéficie à la connaissance (du latin *scientia* = connaissance), nous permettant d'aller toujours plus loin dans les moyens et les outils développés afin de nous permettre d'apporter des réponses aux questions liées à l'évolution et l'origine du vivant. Bien que les données génomiques connaissent une inflation exponentielle, le nombre de nouvelles questions suit également un développement du même ordre. L'élargissement du fossé entre ces deux courbes est également exponentiel. Cet écart entre les questions et réponses représente la courbe exponentielle de notre ignorance. A croire que la science est une méthode qui développe plus notre ignorance que notre connaissance. Quoiqu'il en soit nous n'avons pas atteint notre maximum d'ignorance…

BIBLIOGRAPHIE

Aa, E., J. P. Townsend, R. I. Adams, K. M. Nielsen and J. W. Taylor (2006). "Population structure and gene evolution in Saccharomyces cerevisiae." FEMS Yeast Res **6**(5): 702-15.

Adindla, S., K. Guruprasad and L. Guruprasad (2004). "Three-dimensional models and structure analysis of corynemycolyltransferases in Corynebacterium glutamicum and Corynebacterium efficiens." Int J Biol Macromol **34**(3): 181-9.

Ahn, B. (2000). "A physical interaction of UvrD with nucleotide excision repair protein UvrB." Mol Cells **10**(5): 592-7.

Akhmanova, A. S., V. K. Kagramanova and A. S. Mankin (1993). "Heterogeneity of small plasmids from halophilic archaea." J Bacteriol **175**(4): 1081-6.

Andersson, A. F. and J. F. Banfield (2008). "Virus population dynamics and acquired virus resistance in natural microbial communities." Science **320**(5879): 1047-50.

Andersson, A. F., M. Lundgren, S. Eriksson, M. Rosenlund, R. Bernander and P. Nilsson (2006). "Global analysis of mRNA stability in the archaeon Sulfolobus." Genome Biol **7**(10): R99.

Arcus, V. L., P. B. Rainey and S. J. Turner (2005). "The PIN-domain toxin-antitoxin array in mycobacteria." Trends Microbiol **13**(8): 360-5.

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." Mol Microbiol **34**(2): 217-26.

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 Res **34**(17): e114.

Baker, G. C. and D. A. Cowan (2004). "16 S rDNA primers and the unbiased assessment of thermophile diversity." Biochem Soc Trans **32**(Pt 2): 218-21.

Baliga, N. S., R. Bonneau, M. T. Facciotti, M. Pan, G. Glusman, E. W. Deutsch, P. Shannon, Y. Chiu, R. S. Weng, R. R. Gan, P. Hung, S. V. Date, E. Marcotte, L. Hood and W. V. Ng (2004). "Genome sequence of Haloarcula marismortui: a halophilic archaeon from the Dead Sea." Genome Res **14**(11): 2221-34.

Barabas, O., D. R. Ronning, C. Guynet, A. B. Hickman, B. Ton-Hoang, M. Chandler and F. Dyda (2008). "Mechanism of IS200/IS605 family DNA transposases: activation and transposon-directed target site selection." Cell **132**(2): 208-20.

Barbour, A. G. and C. F. Garon (1987). "Linear plasmids of the bacterium Borrelia burgdorferi have covalently closed ends." Science **237**(4813): 409-11.

Barrangou, R., C. Fremaux, H. Deveau, M. Richards, P. Boyaval, S. Moineau, D. A. Romero and P. Horvath (2007). "CRISPR provides acquired resistance against viruses in prokaryotes." Science **315**(5819): 1709-12.

Beckmann, C., J. D. Waggoner, T. O. Harris, G. S. Tamura and C. E. Rubens (2002). "Identification of novel adhesins from Group B streptococci by use of phage display reveals that C5a peptidase mediates fibronectin binding." Infect Immun **70**(6): 2869-76.

Bellgard, M. I., T. Itoh, H. Watanabe, T. Imanishi and T. Gojobori (1999). "Dynamic evolution of genomes and the concept of genome space." Ann N Y Acad Sci **870**: 293-300.

Beloglazova, N., G. Brown, M. D. Zimmerman, M. Proudfoot, K. S. Makarova, M. Kudritska, S. Kochinyan, S. Wang, M. Chruszcz, W. Minor, E. V. Koonin, A. M. Edwards, A. Savchenko and A. F. Yakunin (2008). "A novel family of sequence-specific endoribonucleases associated with the Clustered Regularly Interspaced Short Palindromic Repeats." J Biol Chem.

Benbouzid-Rollet, N., P. Lopez-Garcia, L. Watrin, G. Erauso, D. Prieur and P. Forterre (1997). "Isolation of new plasmids from hyperthermophilic Archaea of the order Thermococcales." Res Microbiol **148**(9): 767-75.

Bertani, G. (1999). "Transduction-like gene transfer in the methanogen Methanococcus voltae." J Bacteriol **181**(10): 2992-3002.

Bini, E., V. Dikshit, K. Dirksen, M. Drozda and P. Blum (2002). "Stability of mRNA in the hyperthermophilic archaeon Sulfolobus solfataricus." Rna **8**(9): 1129-36.

Birnboim, H. C. and J. Doly (1979). "A rapid alkaline extraction procedure for screening recombinant plasmid DNA." Nucleic Acids Res **7**(6): 1513-23.

Bokranz, M., A. Klein and L. Meile (1990). "Complete nucleotide sequence of plasmid pME2001 of Methanobacterium thermoautotrophicum (Marburg)." Nucleic Acids Res **18**(2): 363.

Bolotin, A., B. Quinquis, A. Sorokin and S. D. Ehrlich (2005). "Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin." Microbiology **151**(Pt 8): 2551-61.

Breitbart, M. and F. Rohwer (2005). "Here a virus, there a virus, everywhere the same virus?" Trends Microbiol **13**(6): 278-84.

Brochier-Armanet, C., B. Boussau, S. Gribaldo and P. Forterre (2008). "Mesophilic Crenarchaeota: proposal for a third archaeal phylum, the Thaumarchaeota." Nat Rev Microbiol **6**(3): 245-52.

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 Biol **6**(5): R42.

Bult, C. J., O. White, G. J. Olsen, L. Zhou, R. D. Fleischmann, G. G. Sutton, J. A. Blake, L. M. FitzGerald, R. A. Clayton, J. D. Gocayne, A. R. Kerlavage, B. A. Dougherty, J. F. Tomb, M. D. Adams, C. I. Reich, R. Overbeek, E. F. Kirkness, K. G. Weinstock, J. M. Merrick, A. Glodek, J. L. Scott, N. S. Geoghagen and J. C. Venter (1996). "Complete genome sequence of the methanogenic archaeon, Methanococcus jannaschii." Science **273**(5278): 1058-73.

Bunker, R. D., J. L. McKenzie, E. N. Baker and V. L. Arcus (2008). "Crystal structure of PAE0151 from Pyrobaculum aerophilum, a PIN-domain (VapC) protein from a toxin-antitoxin operon." Proteins **72**(1): 510-8.

Burroughs, A. M., L. M. Iyer and L. Aravind (2007). "Comparative Genomics and Evolutionary Trajectories of Viral ATP Dependent DNA-Packaging Systems." Genome Dyn **3**: 48-65.

Casas, V. and F. Rohwer (2007). "Phage metagenomics." Methods Enzymol **421**: 259-68.

Catara, G., G. Ruggiero, F. La Cara, F. A. Digilio, A. Capasso and M. Rossi (2003). "A novel extracellular subtilisin-like protease from the hyperthermophile Aeropyrum pernix K1: biochemical properties, cloning, and expression." Extremophiles **7**(5): 391-9.

Cavalier-Smith, T. (2002). "Origins of the machinery of recombination and sex." Heredity **88**(2): 125-41.

Chakrabarty, A. M. (1976). "Plasmids in Pseudomonas." Annu Rev Genet **10**: 7-30.

Chandler, M. and J. Mahillon (2002). Insertion sequences revisited. Mobile DNA, vol 2. A. Press. Washington, DC., ASM Press**:** 305–366.

Chao, K. L. and T. M. Lohman (1991). "DNA-induced dimerization of the Escherichia coli Rep helicase." J Mol Biol **221**(4): 1165-81.

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." J Bacteriol **187**(14): 4992-9.

Chen, Y. J., X. Yu, R. Kasiviswanathan, J. H. Shin, Z. Kelman and E. H. Egelman (2005). "Structural polymorphism of Methanothermobacter thermautotrophicus MCM." J Mol Biol **346**(2): 389-94.

Cheng, H., N. Shen, J. Pei and N. V. Grishin (2004). "Double-stranded DNA bacteriophage prohead protease is homologous to herpesvirus protease." Protein Sci. **13**(8): 2260–2269.

Cho, Y., H. S. Lee, Y. J. Kim, S. G. Kang, S. J. Kim and J. H. Lee (2007). "Characterization of a dUTPase from the Hyperthermophilic Archaeon Thermococcus onnurineus NA1 and Its Application in Polymerase Chain Reaction Amplification." Mar Biotechnol (NY) **9**(4): 450-8.

Clark-Walker, G. D. (1972). "Isolation of circular DNA from a mitochondrial fraction from yeast." Proc Natl Acad Sci U S A **69**(2): 388-92.

Cline, S. W. and W. F. Doolittle (1992). "Transformation of members of the genus Haloarcula with shuttle vectors based on Halobacterium halobium and Haloferax volcanii plasmid replicons." J Bacteriol **174**(3): 1076-80.

Clissold, P. M. and C. P. Ponting (2000). "PIN domains in nonsense-mediated mRNA decay and RNAi." Curr Biol **10**(24): R888-90.

Clore, A. J. and K. M. Stedman (2006). "The SSV1 viral integrase is not essential." Virology.

Clore, A. J. and K. M. Stedman (2007). "The SSV1 viral integrase is not essential." Virology **361**(1): 103-11.

Cohen, G. N., V. Barbe, D. Flament, M. Galperin, R. Heilig, O. Lecompte, O. Poch, D. Prieur, J. Querellou, R. Ripp, J. C. Thierry, J. Van der Oost, J. Weissenbach, Y. Zivanovic and P. Forterre (2003). "An integrated analysis of the genome of the hyperthermophilic archaeon Pyrococcus abyssi." Mol Microbiol **47**(6): 1495-512.

Cohen, S. N., A. C. Chang, H. W. Boyer and R. B. Helling (1973). "Construction of biologically functional bacterial plasmids in vitro." Proc Natl Acad Sci U S A **70**(11): 3240-4.

Constantinesco, F., P. Forterre, E. V. Koonin, L. Aravind and C. Elie (2004). "A bipolar DNA helicase gene, herA, clusters with rad50, mre11 and nurA genes in thermophilic archaea." Nucleic Acids Res **32**(4): 1439-47.

Cowen, D. A. (2004). "The upper temperature of life--where do we draw the line?" Trends Microbiol **12**(2): 58-60.

Crowley, D. J. and P. C. Hanawalt (2001). "The SOS-dependent upregulation of uvrD is not required for efficient nucleotide excision repair of ultraviolet light induced DNA photoproducts in Escherichia coli." Mutat Res **485**(4): 319-29.

Curcio, M. J. and K. M. Derbyshire (2003). "The outs and ins of transposition: from mu to kangaroo." Nat Rev Mol Cell Biol **4**(11): 865-77.

Dawkins, R. (1976). "Selfish Genes ans Selfish Memes." Oxford University Press.

de la Cruz, J., D. Kressler and P. Linder (1999). "Unwinding RNA in Saccharomyces cerevisiae: DEAD-box proteins and related families." Trends Biochem Sci **24**(5): 192-8.

DeLong, E. F., K. Y. Wu, B. B. Prezelin and R. V. Jovine (1994). "High abundance of Archaea in Antarctic marine picoplankton." Nature **371**(6499): 695-7.

Delwart, E. L. (2007). "Viral metagenomics." Rev Med Virol **17**(2): 115-31.

Dennis, P. P. and L. C. Shimmin (1997). "Evolutionary divergence and salinity-mediated selection in halophilic archaea." Microbiol Mol Biol Rev **61**(1): 90-104.

Deveau, H., R. Barrangou, J. E. Garneau, J. Labonte, C. Fremaux, P. Boyaval, D. A. Romero, P. Horvath and S. Moineau (2007). "Phage response to CRISPR-encoded resistance in Streptococcus thermophilus." J Bacteriol.

Di Giulio, M. (2007). "The tree of life might be rooted in the branch leading to Nanoarchaeota." Gene.

Diruggiero, J., D. Dunn, D. L. Maeder, R. Holley-Shanks, J. Chatard, R. Horlacher, F. T. Robb, W. Boos and R. B. Weiss (2000). "Evidence of recent lateral gene transfer among hyperthermophilic archaea." Mol Microbiol **38**(4): 684-93.

Dodson, M. L. and R. S. Lloyd (2002). "Mechanistic comparisons among base excision repair glycosylases." Free Radic Biol Med **32**(8): 678-82.

Doma, M. K. and R. Parker (2006). "Endonucleolytic cleavage of eukaryotic mRNAs with stalls in translation elongation." Nature **440**(7083): 561-4.

Doma, M. K. and R. Parker (2006). "Revenge of the NRD: preferential degradation of nonfunctional eukaryotic rRNA." Dev Cell **11**(6): 757-8.

Doolittle, W. F. (1999). "Phylogenetic classification and the universal tree." Science **284**(5423): 2124-9.

Durand, P., F. Mahe, A. S. Valin and J. Nicolas (2006). "Browsing repeats in genomes: Pygram and an application to non-coding region analysis." BMC Bioinformatics **7**: 477.

Durrant, I., L. C. Benge, C. Sturrock, A. T. Devenish, R. Howe, S. Roe, M. Moore, G. Scozzafava, L. M. Proudfoot, T. C. Richardson and et al. (1990). "The application of enhanced chemiluminescence to membrane-based nucleic acid detection." Biotechniques **8**(5): 564-70.

Ebihara, A., M. Yao, R. Masui, I. Tanaka, S. Yokoyama and S. Kuramitsu (2006). "Crystal structure of hypothetical protein TTHB192 from Thermus thermophilus HB8 reveals a new protein family with an RNA recognition motif-like domain." Protein Sci **15**(6): 1494-9.

Edwards, R. A. and F. Rohwer (2005). "Viral metagenomics." Nat Rev Microbiol **3**(6): 504-10.

Eiserling, F., A. Pushkin, M. Gingery and G. Bertani (1999). "Bacteriophage-like particles associated with the gene transfer agent of methanococcus voltae PS." J Gen Virol **80 (Pt 12)**: 3305-8.

Elferink, M. G., C. Schleper and W. Zillig (1996). "Transformation of the extremely thermoacidophilic archaeon Sulfolobus solfataricus via a self-spreading vector." FEMS Microbiol Lett **137**(1): 31-5.

Erauso, G., S. Marsin, N. Benbouzid-Rollet, M. F. Baucher, T. Barbeyron, Y. Zivanovic, D. Prieur and P. Forterre (1996). "Sequence of plasmid pGT5 from the archaeon Pyrococcus abyssi: evidence for rolling-circle replication in a hyperthermophile." J Bacteriol **178**(11): 3232-7.

Erauso, G., A. Reysenbach, A. Godfroy, J. Meunier, B. Crump, F. Partensky, J. A. Baross, V. T. Marteinsson, G. Barbier, N. R. Pace and D. Prieur (1993). "Pyrococcus abyssi sp. nov., a new hyperthermophilic archaeon isolated from a deep-sea hydrothermal vent." Arch Microbiol **160**: 338-349.

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 **152**(Pt 7): 1951-68.

Falb, M., F. Pfeiffer, P. Palm, K. Rodewald, V. Hickmann, J. Tittor and D. Oesterhelt (2005). "Living with two extremes: conclusions from the genome sequence of Natronomonas pharaonis." Genome Res **15**(10): 1336-43.

Fiala, G. and K. O. Stetter (1986). "Pyrococcus furiosus sp. nov. represents a novel genus of marine heterotrophic archaebacteria growing optimally at 100°C. ." Arch. Microbiol. **145**: 56-61.

Filee, J., P. Siguier and M. Chandler (2007). "Insertion sequence diversity in archaea." Microbiol Mol Biol Rev **71**(1): 121-57.

Fitch, W. M. and E. Margoliash (1967). "Construction of phylogenetic trees." Science **155**(760): 279-84.

Flaus, A., D. M. Martin, G. J. Barton and T. Owen-Hughes (2006). "Identification of multiple distinct Snf2 subfamilies with conserved structural motifs." Nucleic Acids Res **34**(10): 2887-905.

Fletcher, R. J., B. E. Bishop, R. P. Leon, R. A. Sclafani, C. M. Ogata and X. S. Chen (2003). "The structure and function of MCM from archaeal M. Thermoautotrophicum." Nat Struct Biol **10**(3): 160-7.

Forterre, P. (1999). "Displacement of cellular proteins by functional analogues from plasmids or viruses could explain puzzling phylogenies of many DNA informational proteins." Mol Microbiol **33**(3): 457-65.

Forterre, P. (2001). "Genomics and early cellular evolution. The origin of the DNA world." C R Acad Sci III **324**(12): 1067-76.

Fraser, J. S., Z. Yu, K. L. Maxwell and A. R. Davidson (2006). "Ig-like domains on bacteriophages: a tale of promiscuity and deceit." J Mol Biol **359**(2): 496-507.

Fukui, T., H. Atomi, T. Kanai, R. Matsumi, S. Fujiwara and T. Imanaka (2005). "Complete genome sequence of the hyperthermophilic archaeon Thermococcus kodakaraensis KOD1 and comparison with Pyrococcus genomes." Genome Res **15**(3): 352-63.

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): 4510-9.

Geslin, C., M. Le Romancer, G. Erauso, M. Gaillard, G. Perrot and D. Prieur (2003). "PAV1, the first virus-like particle isolated from a hyperthermophilic euryarchaeote, "Pyrococcus abyssi"." J Bacteriol **185**(13): 3888-94.

Godde, J. S. and A. Bickerton (2006). "The Repetitive DNA Elements Called CRISPRs and Their Associated Genes: Evidence of Horizontal Transfer Among Prokaryotes." J Mol Evol.

Gogarten-Boekels, M., E. Hilario and J. P. Gogarten (1995). "The effects of heavy meteorite bombardment on the early evolution--the emergence of the three domains of life." Orig Life Evol Biosph **25**(1-3): 251-64.

Golyshina, O. V. and K. N. Timmis (2005). "Ferroplasma and relatives, recently discovered cell wall-lacking archaea making a living in extremely acid, heavy metal-rich environments." Environ Microbiol **7**(9): 1277-88.

Gomez-Llorente, Y., R. J. Fletcher, X. S. Chen, J. M. Carazo and C. San Martin (2005). "Polymorphism and double hexamer structure in the archaeal minichromosome maintenance (MCM) helicase from Methanobacterium thermoautotrophicum." J Biol Chem **280**(49): 40909-15.

Gonzalez, J. M., D. Sheckells, M. Viebahn, D. Krupatkina, K. M. Borges and F. T. Robb (1999). "Thermococcus waiotapuensis sp. nov., an extremely thermophilic archaeon isolated from a freshwater hot spring." Arch Microbiol **172**(2): 95-101.

Gotfredsen, M. and K. Gerdes (1998). "The Escherichia coli relBE genes belong to a new toxin-antitoxin gene family." Mol Microbiol **29**(4): 1065-76.

Greated, A., L. Lambertsen, P. A. Williams and C. M. Thomas (2002). "Complete sequence of the IncP-9 TOL plasmid pWW0 from Pseudomonas putida." Environ Microbiol **4**(12): 856-71.

Greve, B., S. Jensen, K. Brugger, W. Zillig and R. A. Garrett (2004). "Genomic comparison of archaeal conjugative plasmids from Sulfolobus." Archaea **1**(4): 231-9.

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 **1**(5): 319-25.

Grindley, N. D. F. (2002). The movement of Tn3-like elements: transposition and cointegrate resolution. Mobile DNA, vol 2. A. Press. Washington, DC., ASM Press**:** 230-279.

Grissa, I., G. Vergnaud and C. Pourcel (2007). "The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats." BMC Bioinformatics **8**: 172.

Grissa, I., G. Vergnaud and C. Pourcel (2007). "CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats." Nucleic Acids Res **35**(Web Server issue): W52-7.

Gropp, F., B. Grampp, P. Stolt, P. Palm and W. Zillig (1992). "The immunity-conferring plasmid p phi HL from the Halobacterium salinarium phage phi H: nucleotide sequence and transcription." Virology **190**(1): 45-54.

Grosjean, H., C. Marck, C. Gaspin, W. A. Decatur and V. de Crecy-Lagard (2008). "RNomics and Modomics in the halophilic archaea Haloferax volcanii: identification of RNA modification genes." BMC Genomics **9**(1): 470.

Guglielmetti, S., D. Mora and C. Parini (2007). "Small rolling circle plasmids in Bacillus subtilis and related species: organization, distribution, and their possible role in host physiology." Plasmid **57**(3): 245-64.

Guynet, C., A. B. Hickman, O. Barabas, F. Dyda, M. Chandler and B. Ton-Hoang (2008). "In vitro reconstitution of a single-stranded transposition mechanism of IS608." Mol Cell **29**(3): 302-12.

Hackett, N. R. and S. DasSarma (1989). "Characterization of the small endogenous plasmid of Halobacterium strain SB3 and its use in transformation of H. halobium." Can J Microbiol **35**(1): 86-91.

Hackett, N. R., M. P. Krebs, S. DasSarma, W. Goebel, U. L. RajBhandary and H. G. Khorana (1990). "Nucleotide sequence of a high copy number plasmid from Halobacterium strain GRB." Nucleic Acids Res **18**(11): 3408.

Haft, D. H., J. Selengut, E. F. Mongodin and K. E. Nelson (2005). "A Guild of 45 CRISPR-Associated (Cas) Protein Families and Multiple CRISPR/Cas Subtypes Exist in Prokaryotic Genomes." PLoS Comput Biol **1**(6): e60.

Haines, A. S., P. Akhtar, E. R. Stephens, K. Jones, C. M. Thomas, C. D. Perkins, J. R. Williams, M. J. Day and J. C. Fry (2006). "Plasmids from freshwater environments capable of IncQ retrotransfer are diverse and include pQKH54, a new IncP-1 subgroup archetype." Microbiology **152**(Pt 9): 2689-701.

Hale, C., K. Kleppe, R. M. Terns and M. P. Terns (2008). "Prokaryotic silencing (psi)RNAs in Pyrococcus furiosus." Rna.

Hall, M. J. and N. R. Hackett (1989). "DNA sequence of a small plasmid from Halobacterium strain GN101." Nucleic Acids Res **17**(24): 10501.

Hallam, S. J., K. T. Konstantinidis, N. Putnam, C. Schleper, Y. Watanabe, J. Sugahara, C. Preston, J. de la Torre, P. M. Richardson and E. F. DeLong (2006). "Genomic analysis of the uncultivated marine crenarchaeote Cenarchaeum symbiosum." Proc Natl Acad Sci U S A **103**(48): 18296-301.

Hamilton-Brehm, S. D., G. J. Schut and M. W. Adams (2005). "Metabolic and evolutionary relationships among Pyrococcus Species: genetic exchange within a hydrothermal vent environment." J Bacteriol **187**(21): 7492-9.

Haring, M., X. Peng, K. Brugger, R. Rachel, K. O. Stetter, R. A. Garrett and D. Prangishvili (2004). "Morphology and genome organization of the virus PSV of the hyperthermophilic archaeal genera Pyrobaculum and Thermoproteus: a novel virus family, the Globuloviridae." Virology **323**(2): 233-42.

Harriott, O. T., R. Huber, K. O. Stetter, P. W. Betts and K. M. Noll (1994). "A cryptic miniplasmid from the hyperthermophilic bacterium Thermotoga sp. strain RQ7." J Bacteriol **176**(9): 2759-62.

Hecker, A., N. Leulliot, D. Gadelle, M. Graille, A. Justome, P. Dorlet, C. Brochier, S. Quevillon-Cheruel, E. L. Cam, H. V. Tilbeurgh and P. Forterre (2007). "An archaeal orthologue of the universal protein Kae1 is an iron metalloprotein which exhibits atypical DNA-binding properties and apurinic-endonuclease activity in vitro." Nucleic Acids Res.

Hengen, P. N. (1997). "Shearing DNA for genomic library construction." Trends Biochem Sci **22**(7): 273-4.

Heydorn, A., B. K. Ersboll, M. Hentzer, M. R. Parsek, M. Givskov and S. Molin (2000). "Experimental reproducibility in flow-chamber biofilms." Microbiology **146 (Pt 10)**: 2409-15.

Hogrefe, C. and B. Friedrich (1984). "Isolation and characterization of megaplasmid DNA from lithoautotrophic bacteria." Plasmid **12**(3): 161-9.

Holmes, A. J., M. R. Gillings, B. S. Nield, B. C. Mabbutt, K. M. Nevalainen and H. W. Stokes (2003). "The gene cassette metagenome is a basic resource for bacterial genome evolution." Environ Microbiol **5**(5): 383-94.

Holmes, M., F. Pfeifer and M. Dyall-Smith (1994). "Improved shuttle vectors for Haloferax volcanii including a dual-resistance plasmid." Gene **146**(1): 117-21.

Horst, J. P. and H. J. Fritz (1996). "Counteracting the mutagenic effect of hydrolytic deamination of DNA 5-methylcytosine residues at high temperature: DNA mismatch N-glycosylase Mig.Mth of the thermophilic archaeon Methanobacterium thermoautotrophicum THF." Embo J **15**(19): 5459-69.

Horvath, P., D. A. Romero, A. C. Coute-Monvoisin, M. Richards, H. Deveau, S. Moineau, P. Boyaval, C. Fremaux and R. Barrangou (2008). "Diversity, activity, and evolution of CRISPR loci in Streptococcus thermophilus." J Bacteriol **190**(4): 1401-12.

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): 63-7.

Ilyina, T. V. and E. V. Koonin (1992). "Conserved sequence motifs in the initiator proteins for rolling circle DNA replication encoded by diverse replicons from eubacteria, eucaryotes and archaebacteria." Nucleic Acids Res **20**(13): 3279-85.

Ishino, Y., H. Shinagawa, K. Makino, M. Amemura and A. Nakata (1987). "Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in Escherichia coli, and identification of the gene product." J Bacteriol **169**(12): 5429-33.

Itoh, T. (2003). "Taxonomy of nonmethanogenic hyperthermophilic and related thermophilic archaea." J Biosci Bioeng **96**(3): 203-12.

Iyer, L. M., D. D. Leipe, E. V. Koonin and L. Aravind (2004). "Evolutionary history and higher order classification of AAA+ ATPases." J Struct Biol **146**(1-2): 11-31.

Iyer, L. M., K. S. Makarova, E. V. Koonin and L. Aravind (2004). "Comparative genomics of the FtsK-HerA superfamily of pumping ATPases: implications for the origins of chromosome segregation, cell division and viral capsid packaging." Nucleic Acids Res **32**(17): 5260-79.

Jain, R., M. C. Rivera, J. E. Moore and J. A. Lake (2003). "Horizontal gene transfer accelerates genome innovation and evolution." Mol Biol Evol **20**(10): 1598-602.

Jansen, R., J. D. Embden, W. Gaastra and L. M. Schouls (2002). "Identification of genes that are associated with DNA repeats in prokaryotes." Mol Microbiol **43**(6): 1565-75.

Jensen, R. B. and K. Gerdes (1995). "Programmed cell death in bacteria: proteic plasmid stabilization systems." Mol Microbiol **17**(2): 205-10.

Jeyakanthan, J., E. Inagaki, C. Kuroishi and T. H. Tahirov (2005). "Structure of PIN-domain protein PH0500 from Pyrococcus horikoshii." Acta Crystallogr Sect F Struct Biol Cryst Commun **61**(Pt 5): 463-8.

Jolivet, E., E. Corre, S. L'Haridon, P. Forterre and D. Prieur (2004). "Thermococcus marinus sp. nov. and Thermococcus radiotolerans sp. nov., two hyperthermophilic archaea from deep-sea hydrothermal vents that resist ionizing radiation." Extremophiles **8**(3): 219-27.

Kamerbeek, J., L. Schouls, A. Kolk, M. van Agterveld, D. van Soolingen, S. Kuijper, A. Bunschoten, H. Molhuizen, R. Shaw, M. Goyal and J. van Embden (1997). "Simultaneous detection and strain differentiation of Mycobacterium tuberculosis for diagnosis and epidemiology." J Clin Microbiol **35**(4): 907-14.

Kannan, Y., Y. Koga, Y. Inoue, M. Haruki, M. Takagi, T. Imanaka, M. Morikawa and S. Kanaya (2001). "Active subtilisin-like protease from a hyperthermophilic archaeon in a form with a putative prosequence." Appl Environ Microbiol **67**(6): 2445-52.

Kanoksilapatham, W., J. M. Gonzalez, D. L. Maeder, J. DiRuggiero and F. T. Robb (2004). "A proposal to rename the hyperthermophile Pyrococcus woesei as Pyrococcus furiosus subsp. woesei." Archaea **1**(4): 277-83.

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 strain7." DNA Res **8**(4): 123-40.

Kawarabayasi, Y., M. Sawada, H. Horikawa, Y. Haikawa, Y. Hino, S. Yamamoto, M. Sekine, S. Baba, H. Kosugi, A. Hosoyama, Y. Nagai, M. Sakai, K. Ogura, R. Otsuka, H. Nakazawa, M. Takamiya, Y. Ohfuku, T. Funahashi, T. Tanaka, Y. Kudoh, J. Yamazaki, N. Kushida, A. Oguchi, K. Aoki and H. Kikuchi (1998). "Complete sequence and gene organization of the genome of a hyper-thermophilic archaebacterium, Pyrococcus horikoshii OT3 (supplement)." DNA Res **5**(2): 147-55.

Kecha, M., S. Benallaoua, J. P. Touzel, R. Bonaly and F. Duchiron (2007). "Biochemical and phylogenetic characterization of a novel terrestrial hyperthermophilic archaeon pertaining to the genus Pyrococcus from

an Algerian hydrothermal hot spring." Extremophiles **11**(1): 65-73.

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 **2**(4): 391-3.

Keller, B. and T. A. Bickle (1986). "The nucleotide sequence of gene 21 of bacteriophage T4 coding for the prohead protease." Gene **49**(2): 245-51.

Kelman, L. M. and Z. Kelman (2003). "Archaea: an archetype for replication initiation studies?" Mol Microbiol **48**(3): 605-15.

Kersulyte, D., N. S. Akopyants, S. W. Clifton, B. A. Roe and D. E. Berg (1998). "Novel sequence organization and insertion specificity of IS605 and IS606: chimaeric transposable elements of Helicobacter pylori." Gene **223**(1-2): 175-86.

Keswani, J., S. Orkand, U. Premachandran, L. Mandelco, M. J. Franklin and W. B. Whitman (1996). "Phylogeny and taxonomy of mesophilic Methanococcus spp. and comparison of rRNA, DNA hybridization, and phenotypic methods." Int J Syst Bacteriol **46**(3): 727-35.

Kinashi, H., M. Shimaji-Murayama and T. Hanafusa (1991). "Nucleotide sequence analysis of the unusually long terminal inverted repeats of a giant linear plasmid, SCP1." Plasmid **26**(2): 123-30.

Klein, R., U. Baranyi, N. Rossler, B. Greineder, H. Scholz and A. Witte (2002). "Natrialba magadii virus phiCh1: first complete nucleotide sequence and functional organization of a virus infecting a haloalkaliphilic archaeon." Mol Microbiol **45**(3): 851-63.

Kletzin, A., A. Lieke, T. Urich, R. L. Charlebois and C. W. Sensen (1999). "Molecular analysis of pDL10 from Acidianus ambivalens reveals a family of related plasmids from extremely thermophilic and acidophilic archaea." Genetics **152**(4): 1307-14.

Konstantinidis, K. T. and J. M. Tiedje (2005). "Towards a genome-based taxonomy for prokaryotes." J Bacteriol **187**(18): 6258-64.

Kristoffersen, P., G. B. Jensen, K. Gerdes and J. Piskur (2000). "Bacterial toxin-antitoxin gene system as containment control in yeast cells." Appl Environ Microbiol **66**(12): 5524-6.

Krupovic, M. and D. H. Bamford (2008). "Archaeal proviruses TKV4 and MVV extend the PRD1-adenovirus lineage to the phylum Euryarchaeota." Virology.

Kunin, V., R. Sorek and P. Hugenholtz (2007). "Evolutionary conservation of sequence and secondary structures in CRISPR repeats." Genome Biol **8**(4): R61.

Kurtz, S. and C. Schleiermacher (1999). "REPuter: fast computation of maximal repeats in complete genomes." Bioinformatics **15**(5): 426-7.

Kvaloy, K., H. Nilsen, K. S. Steinsbekk, A. Nedal, B. Monterotti, M. Akbari and H. E. Krokan (2001). "Sequence variation in the human uracil-DNA glycosylase (UNG) gene." Mutat Res **461**(4): 325-38.

Lang, A. S. and J. T. Beatty (2000). "Genetic analysis of a bacterial genetic exchange element: the gene transfer agent of Rhodobacter capsulatus." Proc Natl Acad Sci U S A **97**(2): 859-64.

Lederberg, J. and E. L. Tatum (1946). "Gene recombination in Eschericia coli." Nature **58**: 558.

Lee, H. S., S. G. Kang, S. S. Bae, J. K. Lim, Y. Cho, Y. J. Kim, J. H. Jeon, S. S. Cha, K. K. Kwon, H. T. Kim, C. J. Park, H. W. Lee, S. I. Kim, J. Chun, R. R. Colwell, S. J. Kim and J. H. Lee (2008). "The complete genome sequence of Thermococcus onnurineus NA1 reveals a mixed heterotrophic and carboxydotrophic metabolism." J Bacteriol **190**(22): 7491-9.

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): 806-9.

Li, Y., A. Dabrazhynetskaya, B. Youngren and S. Austin (2004). "The role of Par proteins in the active segregation of the P1 plasmid." Mol Microbiol **53**(1): 93-102.

Liang, G., F. Chan, Y. Tomigahara, Y. Tsai, F. Gonzales, E. Li, W. Laird and P. Jones (2002). "Cooperativity between DNA Methyltransferases in the Maintenance

Methylation of Repetitive Elements." Mol Cell Biol **22**(22): 480-491.

Lillestol, R. K., P. Redder, R. A. Garrett and K. Brugger (2006). "A putative viral defence mechanism in archaeal cells." Archaea **2**(1): 59-72.

Lipps, G. (2004). "The replication protein of the Sulfolobus islandicus plasmid pRN1." Biochem Soc Trans **32**(Pt 2): 240-4.

Lipps, G. (2006). "Plasmids and viruses of the thermoacidophilic crenarchaeote Sulfolobus." Extremophiles **10**(1): 17-28.

Lipps, G., A. O. Weinzierl, G. von Scheven, C. Buchen and P. Cramer (2004). "Structure of a bifunctional DNA primase-polymerase." Nat Struct Mol Biol **11**(2): 157-62.

Liu, P., J. A. Theruvathu, A. Darwanto, V. V. Lao, T. Pascal, W. Goddard, 3rd and L. C. Sowers (2008). "Mechanisms of base selection by the Escherichia coli mispaired uracil glycosylase." J Biol Chem **283**(14): 8829-36.

Livingston, D. M. (1977). "Inheritance of the 2 micrometer m DNA plasmid from Saccharomyces." Genetics **86**(1): 73-84.

Lopez-Garcia, P., P. Forterre, J. van der Oost and G. Erauso (2000). "Plasmid pGS5 from the hyperthermophilic archaeon Archaeoglobus profundus is negatively supercoiled." J Bacteriol **182**(17): 4998-5000.

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." Appl Environ Microbiol **68**(11): 5528-36.

Luo, Y., T. Leisinger and A. Wasserfallen (1995). "The plasmids found in isolates of the thermoacidophilic archaebacterium Thermoplasma acidophilum." FEMS Microbiol Lett **128**: 157-161.

Maaty, W. S., A. C. Ortmann, M. Dlakic, K. Schulstad, J. K. Hilmer, L. Liepold, B. Weidenheft, R. Khayat, T. Douglas, M. J. Young and B. Bothner (2006). "Characterization of the archaeal thermophile Sulfolobus turreted icosahedral virus validates an evolutionary link among double-stranded DNA viruses from all domains of life." J Virol **80**(15): 7625-35.

Makarova, K. S., L. Aravind, M. Y. Galperin, N. V. Grishin, R. L. Tatusov, Y. I. Wolf and E. V. Koonin (1999). "Comparative genomics of the Archaea (Euryarchaeota): evolution of conserved protein families, the stable core, and the variable shell." Genome Res **9**(7): 608-28.

Makarova, K. S., L. Aravind, N. V. Grishin, I. B. Rogozin and E. V. Koonin (2002). "A DNA repair system specific for thermophilic Archaea and bacteria predicted by genomic context analysis." Nucleic Acids Res **30**(2): 482-96.

Makarova, K. S., N. V. Grishin, S. A. Shabalina, Y. I. Wolf and E. V. Koonin (2006). "A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action." Biol Direct **1**: 7.

Makino, S., N. Amano, H. Koike and M. Suzuki (1999). "Prophages inserted in archaebacterial genomes." Proc. Japan Acad Ser. B **75**: 166-171.

Manzan, A., G. Pfeiffer, M. L. Hefferin, C. E. Lang, J. P. Carney and K. P. Hopfner (2004). "MlaA, a hexameric ATPase linked to the Mre11 complex in archaeal genomes." EMBO Rep **5**(1): 54-9.

Maquat, L. E. and X. Li (2001). "Mammalian heat shock p70 and histone H4 transcripts, which derive from naturally intronless genes, are immune to nonsense-mediated decay." Rna **7**(3): 445-56.

Margolin, W. and S. R. Long (1993). "Isolation and characterization of a DNA replication origin from the 1,700-kilobase-pair symbiotic megaplasmid pSym-b of Rhizobium meliloti." J Bacteriol **175**(20): 6553-61.

Marteinsson, V. T., J. L. Birrien, A. L. Reysenbach, M. Vernet, D. Marie, A. Gambacorta, P. Messner, U. B. Sleytr and D. Prieur (1999). "Thermococcus barophilus sp. nov., a new barophilic and hyperthermophilic archaeon isolated under high hydrostatic pressure from a deep-sea hydrothermal vent." Int J Syst Bacteriol **49 Pt 2**: 351-9.

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." J Bacteriol **189**(7): 2683-91.

Matsunaga, F., P. Forterre, Y. Ishino and H. Myllykallio (2001). "In vivo interactions of archaeal Cdc6/Orc1 and minichromosome maintenance proteins with the replication origin." Proc Natl Acad Sci U S A **98**(20): 11152-7.

Matsunaga, F., A. Glatigny, M. H. Mucchielli-Giorgi, N. Agier, H. Delacroix, L. Marisa, P. Durosay, Y. Ishino, L. Aggerbeck and P. Forterre (2007). "Genomewide and biochemical analyses of DNA-binding activity of Cdc6/Orc1 and Mcm proteins in Pyrococcus sp." Nucleic Acids Res.

McCready, S. and L. Marcello (2003). "Repair of UV damage in Halobacterium salinarum." Biochem Soc Trans **31**(Pt 3): 694-8.

Mendes, M. V., J. F. Aparicio and J. F. Martin (2000). "Complete nucleotide sequence and characterization of pSNA1 from pimaricin-producing Streptomyces natalensis that replicates by a rolling circle mechanism." Plasmid **43**(2): 159-65.

Merlin, C., D. Springael and A. Toussaint (1999). "Tn4371: A modular structure encoding a phage-like integrase, a Pseudomonas-like catabolic pathway, and RP4/Ti-like transfer functions." Plasmid **41**(1): 40-54.

Metcalf, W. W., J. K. Zhang, E. Apolinario, K. R. Sowers and R. S. Wolfe (1997). "A genetic system for Archaea of the genus Methanosarcina: liposome-mediated transformation and construction of shuttle vectors." Proc Natl Acad Sci U S A **94**(6): 2626-31.

Miroshnichenko, M. L., H. Hippe, E. Stackebrandt, N. A. Kostrikina, N. A. Chernyh, C. Jeanthon, T. N. Nazina, S. S. Belyaev and E. A. Bonch-Osmolovskaya (2001). "Isolation and characterization of Thermococcus sibiricus sp. nov. from a Western Siberia high-temperature oil reservoir." Extremophiles **5**(2): 85-91.

Moe, E., I. Leiros, A. O. Smalas and S. McSweeney (2006). "The crystal structure of mismatch-specific uracil-DNA glycosylase (MUG) from Deinococcus radiodurans reveals a novel catalytic residue and broad substrate specificity." J Biol Chem **281**(1): 569-77.

Mojica, F. J., C. Diez-Villasenor, J. Garcia-Martinez and E. Soria (2005). "Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements." J Mol Evol **60**(2): 174-82.

Mojica, F. J., C. Diez-Villasenor, E. Soria and G. Juez (2000). "Biological significance of a family of regularly spaced repeats in the genomes of Archaea, Bacteria and mitochondria." Mol Microbiol **36**(1): 244-6.

Mojica, F. J., C. Ferrer, G. Juez and F. Rodriguez-Valera (1995). "Long stretches of short tandem repeats are present in the largest replicons of the Archaea Haloferax mediterranei and Haloferax volcanii and could be involved in replicon partitioning." Mol Microbiol **17**(1): 85-93.

Mokkapati, S. K., A. R. Fernandez de Henestrosa and A. S. Bhagwat (2001). "Escherichia coli DNA glycosylase Mug: a growth-regulated enzyme required for mutation avoidance in stationary-phase cells." Mol Microbiol **41**(5): 1101-11.

Mokrousov, I., E. Limeschenko, A. Vyazovaya and O. Narvskaya (2007). "Corynebacterium diphtheriae spoligotyping based on combined use of two CRISPR loci." Biotechnol J **2**(7): 901-6.

Mongodin, E. F., I. R. Hance, R. T. Deboy, S. R. Gill, S. Daugherty, R. Huber, C. M. Fraser, K. Stetter and K. E. Nelson (2005). "Gene transfer and genome plasticity in Thermotoga maritima, a model hyperthermophilic species." J Bacteriol **187**(14): 4935-44.

Morikawa, M., Y. Izawa, N. Rashid, T. Hoaki and T. Imanaka (1994). "Purification and characterization of a thermostable thiol protease from a newly isolated hyperthermophilic Pyrococcus sp." Appl Environ Microbiol **60**(12): 4559-66.

Muskhelishvili, G., P. Palm and W. Zillig (1993). "SSV1-encoded site-specific recombination system in Sulfolobus shibatae." Mol Gen Genet **237**(3): 334-42.

Nakata, A., M. Amemura and K. Makino (1989). "Unusual nucleotide arrangement with repeated sequences in the Escherichia coli

K-12 chromosome." J Bacteriol **171**(6): 3553-6.

Nelson, K. E., R. A. Clayton, S. R. Gill, M. L. Gwinn, R. J. Dodson, D. H. Haft, E. K. Hickey, J. D. Peterson, W. C. Nelson, K. A. Ketchum, L. McDonald, T. R. Utterback, J. A. Malek, K. D. Linher, M. M. Garrett, A. M. Stewart, M. D. Cotton, M. S. Pratt, C. A. Phillips, D. Richardson, J. Heidelberg, G. G. Sutton, R. D. Fleischmann, J. A. Eisen, O. White, S. L. Salzberg, H. O. Smith, J. C. Venter and C. M. Fraser (1999). "Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of Thermotoga maritima." Nature **399**(6734): 323-9.

Ng, W. V., S. A. Ciufo, T. M. Smith, R. E. Bumgarner, D. Baskin, J. Faust, B. Hall, C. Loretz, J. Seto, J. Slagel, L. Hood and S. DasSarma (1998). "Snapshot of a large dynamic replicon in a halophilic archaeon: megaplasmid or minichromosome?" Genome Res **8**(11): 1131-41.

Nishioka, M., H. Mizuguchi, S. Fujiwara, S. Komatsubara, M. Kitabayashi, H. Uemura, M. Takagi and T. Imanaka (2001). "Long and accurate PCR with a mixture of KOD DNA polymerase and its exonuclease deficient mutant enzyme." J Biotechnol **88**(2): 141-9.

Nolling, J. and W. M. de Vos (1992). "Identification of the CTAG-recognizing restriction-modification systems MthZI and MthFI from Methanobacterium thermoformicicum and characterization of the plasmid-encoded mthZIM gene." Nucleic Acids Res **20**(19): 5047-52.

Oberer, M., H. Lindner, O. Glatter, C. Kratky and W. Keller (1999). "Thermodynamic properties and DNA binding of the ParD protein from the broad host-range plasmid RK2/RP4 killing system." Biol Chem **380**(12): 1413-20.

Ochman, H., J. G. Lawrence and E. A. Groisman (2000). "Lateral gene transfer and the nature of bacterial innovation." Nature **405**(6784): 299-304.

Oliveira, P. H., F. Lemos, G. A. Monteiro and D. M. Prazeres (2008). "Recombination frequency in plasmid DNA containing direct repeats--predictive correlation with repeat and intervening sequence length." Plasmid **60**(2): 159-65.

Orban, T. I. and E. Izaurralde (2005). "Decay of mRNAs targeted by RISC requires XRN1, the Ski complex, and the exosome." Rna **11**(4): 459-69.

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 archaebacterium Sulfolobus shibatae." Virology **185**(1): 242-50.

Pape, T., H. Meka, S. Chen, G. Vicentini, M. van Heel and S. Onesti (2003). "Hexameric ring structure of the full-length archaeal MCM protein complex." EMBO Rep **4**(11): 1079-83.

Parent, S. A., C. M. Fenimore and K. A. Bostian (1985). "Vector systems for the expression, analysis and cloning of DNA sequences in S. cerevisiae." Yeast **1**(2): 83-138.

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**(Pt 2): 383-91.

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." J Mol Biol **303**(4): 449-54.

Petit, M. A., E. Dervyn, M. Rose, K. D. Entian, S. McGovern, S. D. Ehrlich and C. Bruand (1998). "PcrA is an essential DNA helicase of Bacillus subtilis fulfilling functions both in repair and rolling-circle replication." Mol Microbiol **29**(1): 261-73.

Pfeifer, F. and U. Blaseio (1989). "Insertion elements and deletion formation in a halophilic archaebacterium." J Bacteriol **171**(9): 5135-40.

Pfeifer, F., U. Blaseio and P. Ghahraman (1988). "Dynamic plasmid populations in Halobacterium halobium." J Bacteriol **170**(8): 3718-24.

Pfeifer, F., U. Blaseio and M. Horne (1989). "Genome structure of Halobacterium halobium: plasmid dynamics in gas vacuole

deficient mutants." Can J Microbiol **35**(1): 96-100.

Pfister, P., A. Wasserfallen, R. Stettler and T. Leisinger (1998). "Molecular analysis of Methanobacterium phage psiM2." Mol Microbiol **30**(2): 233-44.

Piekarowicz, A., A. Klyz, M. Majchrzak, M. Adamczyk-Poplawska, T. K. Maugel and D. C. Stein (2007). "Characterization of the dsDNA prophage sequences in the genome of Neisseria gonorrhoeae and visualization of productive bacteriophage." BMC Microbiol **7**: 66.

Pourcel, C., G. Salvignol and G. Vergnaud (2005). "CRISPR elements in Yersinia pestis acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies." Microbiology **151**(Pt 3): 653-63.

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., 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." J Mol Biol **359**(5): 1203-16.

Prosser, J. I. and G. W. Nicol (2008). "Relative contributions of archaea and bacteria to aerobic ammonia oxidation in the environment." Environ Microbiol **10**(11): 2931-41.

Quaiser, A., F. Constantinesco, M. F. White, P. Forterre and C. Elie (2008). "The Mre11 protein interacts with both Rad50 and the HerA bipolar helicase and is recruited to DNA following gamma irradiation in the archaeon Sulfolobus acidocaldarius." BMC Mol Biol **9**: 25.

Ray, J. L. and K. M. Nielsen (2005). "Experimental methods for assaying natural transformation and inferring horizontal gene transfer." Methods Enzymol **395**: 491-520.

Religa, T. L., C. M. Johnson, D. M. Vu, S. H. Brewer, R. B. Dyer and A. R. Fersht (2007). "The helix-turn-helix motif as an ultrafast independently folding domain: the pathway of folding of Engrailed homeodomain." Proc Natl Acad Sci U S A **104**(22): 9272-7.

Rest, J. S. and D. P. Mindell (2003). "Retroids in archaea: phylogeny and lateral origins." Mol Biol Evol **20**(7): 1134-42.

Robinson, N. P. and S. D. Bell (2007). "Extrachromosomal element capture and the evolution of multiple replication origins in archaeal chromosomes." Proc Natl Acad Sci U S A **104**(14): 5806-11.

Robinson, N. P., I. Dionne, M. Lundgren, V. L. Marsh, R. Bernander and S. D. Bell (2004). "Identification of two origins of replication in the single chromosome of the archaeon Sulfolobus solfataricus." Cell **116**(1): 25-38.

Ronimus, R. S., A. Reysenbach, D. R. Musgrave and H. W. Morgan (1997). "The phylogenetic position of the Thermococcus isolate AN1 based on 16S rRNA gene sequence analysis: a proposal that AN1 represents a new species, Thermococcus zilligii sp. nov." Arch Microbiol **168**(3): 245-8.

Ruan, L. and X. Xu (2007). "Sequence analysis and characterizations of two novel plasmids isolated from Thermus sp. 4C." Plasmid **58**(1): 84-7.

Saavedra De Bast, M., N. Mine and L. Van Melderen (2008). "Chromosomal toxin-antitoxin systems may act as antiaddiction modules." J Bacteriol **190**(13): 4603-9.

Santangelo, T. J., L. Cubonova, R. Matsumi, H. Atomi, T. Imanaka and J. N. Reeve (2008). "Polarity in archaeal operon transcription in Thermococcus kodakaraensis." J Bacteriol **190**(6): 2244-8.

Santangelo, T. J., L. Cubonova and J. N. Reeve (2008). "Shuttle vector expression in Thermococcus kodakaraensis: contributions of cis elements to protein synthesis in a hyperthermophilic archaeon." Appl Environ Microbiol **74**(10): 3099-104.

Sasaki, T., R. Fassler and E. Hohenester (2004). "Laminin: the crux of basement membrane assembly." J Cell Biol **164**(7): 959-63.

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." Appl Environ Microbiol **71**(7): 3889-99.

Scherzinger, E., M. M. Bagdasarian, P. Scholz, R. Lurz, B. Ruckert and M. Bagdasarian (1984). "Replication of the broad host range plasmid RSF1010: requirement for three plasmid-encoded proteins." Proc Natl Acad Sci U S A **81**(3): 654-8.

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." J Bacteriol **177**(15): 4417-26.

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." Proc Natl Acad Sci U S A **89**(16): 7645-9.

Schleper, C., G. Puehler, I. Holz, A. Gambacorta, D. Janekovic, U. Santarius, H. P. Klenk and W. Zillig (1995). "Picrophilus gen. nov., fam. nov.: a novel aerobic, heterotrophic, thermoacidophilic genus and family comprising archaea capable of growth around pH 0." J Bacteriol **177**(24): 7050-9.

Schluter, A., R. Szczepanowski, A. Puhler and E. M. Top (2007). "Genomics of IncP-1 antibiotic resistance plasmids isolated from wastewater treatment plants provides evidence for a widely accessible drug resistance gene pool." FEMS Microbiol Rev **31**(4): 449-77.

Sebaihia, M., B. W. Wren, P. Mullany, N. F. Fairweather, N. Minton, R. Stabler, N. R. Thomson, A. P. Roberts, A. M. Cerdeno-Tarraga, H. Wang, M. T. Holden, A. Wright, C. Churcher, M. A. Quail, S. Baker, N. Bason, K. Brooks, T. Chillingworth, A. Cronin, P. Davis, L. Dowd, A. Fraser, T. Feltwell, Z. Hance, S. Holroyd, K. Jagels, S. Moule, K. Mungall, C. Price, E. Rabbinowitsch, S. Sharp, M. Simmonds, K. Stevens, L. Unwin, S. Whithead, B. Dupuy, G. Dougan, B. Barrell and J. Parkhill (2006). "The multidrug-resistant human pathogen Clostridium difficile has a highly mobile, mosaic genome." Nat Genet **38**(7): 779-86.

Serre, M. C., C. Letzelter, J. R. Garel and M. Duguet (2002). "Cleavage properties of an archaeal site-specific recombinase, the SSV1 integrase." J Biol Chem **277**(19): 16758-67.

She, Q., K. Brugger and L. Chen (2002). "Archaeal integrative genetic elements and their impact on genome evolution." Res Microbiol **153**(6): 325-32.

She, Q., X. Peng, W. Zillig and R. A. Garrett (2001). "Gene capture in archaeal chromosomes." Nature **409**(6819): 478.

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 **2**(4): 417-25.

She, Q., B. Shen and L. Chen (2004). "Archaeal integrases and mechanisms of gene capture." Biochem Soc Trans **32**(Pt 2): 222-6.

Shockley, K. R., D. E. Ward, S. R. Chhabra, S. B. Conners, C. I. Montero and R. M. Kelly (2003). "Heat shock response by the hyperthermophilic archaeon Pyrococcus furiosus." Appl Environ Microbiol **69**(4): 2365-71.

Siezen, R. J. and J. A. Leunissen (1997). "Subtilases: the superfamily of subtilisin-like serine proteases." Protein Sci **6**(3): 501-23.

Sobecky, P. A. (2002). "Approaches to investigating the ecology of plasmids in marine bacterial communities." Plasmid **48**(3): 213-21.

Sokolova, T. G., C. Jeanthon, N. A. Kostrikina, N. A. Chernyh, A. V. Lebedinsky, E. Stackebrandt and E. A. Bonch-Osmolovskaya (2004). "The first evidence of anaerobic CO oxidation coupled with H2 production by a hyperthermophilic archaeon isolated from a deep-sea hydrothermal vent." Extremophiles **8**(4): 317-23.

Soler, N., A. Justome, S. Quevillon-Cheruel, F. Lorieux, E. Le Cam, E. Marguet and P. Forterre (2007). "The rolling-circle plasmid pTN1 from the hyperthermophilic archaeon Thermococcus nautilus." Mol Microbiol **66**(2): 357-70.

Soler, N., E. Marguet, J. M. Verbavatz and P. Forterre (2008). "Virus-like vesicles and extracellular DNA produced by hyperthermophilic archaea of the order Thermococcales." Res Microbiol **159**(5): 390-9.

Solow, B. T. and G. A. Somkuti (2001). "Molecular properties of Streptococcus thermophilus plasmid pER35 encoding a restriction modification system." Curr Microbiol **42**(2): 122-8.

Sorensen, S. J., A. H. Sorensen, L. H. Hansen, G. Oregaard and D. Veal (2003). "Direct detection and quantification of horizontal gene transfer by using flow cytometry and gfp as a reporter gene." Curr Microbiol **47**(2): 129-33.

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." J Bacteriol **182**(24): 7014-20.

Steward, A., S. Adhya and J. Clarke (2002). "Sequence conservation in Ig-like domains: the role of highly conserved proline residues in the fibronectin type III superfamily." J Mol Biol **318**(4): 935-40.

Sturino, J. M. and T. R. Klaenhammer (2006). "Engineered bacteriophage-defence systems in bioprocessing." Nat Rev Microbiol **4**(5): 395-404.

Tabuchi, A., Y. N. Min, D. D. Womble and R. H. Rownd (1992). "Autoregulation of the stability operon of IncFII plasmid NR1." J Bacteriol **174**(23): 7629-34.

Takeshita, D., S. Zenno, W. C. Lee, K. Saigo and M. Tanokura (2007). "Crystal structure of the PIN domain of human telomerase-associated protein EST1A." Proteins **68**(4): 980-9.

Tam, J. E. and B. C. Kline (1989). "The F plasmid ccd autorepressor is a complex of CcdA and CcdB proteins." Mol Gen Genet **219**(1-2): 26-32.

Tang, S. L., S. Nuttall and M. Dyall-Smith (2004). "Haloviruses HF1 and HF2: evidence for a recent and large recombination event." J Bacteriol **186**(9): 2810-7.

Tang, T. H., J. P. Bachellerie, T. Rozhdestvensky, M. L. Bortolin, H. Huber, M. Drungowski, T. Elge, J. Brosius and A. Huttenhofer (2002). "Identification of 86 candidates for small non-messenger RNAs from the archaeon Archaeoglobus fulgidus." Proc Natl Acad Sci U S A **99**(11): 7536-41.

Tang, T. H., N. Polacek, M. Zywicki, H. Huber, K. Brugger, R. Garrett, J. P. Bachellerie and A. Huttenhofer (2005). "Identification of novel non-coding RNAs as potential antisense regulators in the archaeon Sulfolobus solfataricus." Mol Microbiol **55**(2): 469-81.

Ton-Hoang, B., C. Guynet, D. R. Ronning, B. Cointin-Marty, F. Dyda and M. Chandler (2005). "Transposition of ISHp608, member of an unusual family of bacterial insertion sequences." Embo J **24**(18): 3325-38.

Tumbula, D. L., T. L. Bowen and W. B. Whitman (1997). "Characterization of pURB500 from the archaeon Methanococcus maripaludis and construction of a shuttle vector." J Bacteriol **179**(9): 2976-86.

Tyson, G. W. and J. F. Banfield (2008). "Rapidly evolving CRISPRs implicated in acquired resistance of microorganisms to viruses." Environ Microbiol **10**(1): 200-7.

Unaldi, M. N., H. Korkmaz, B. Arikan and G. Coral (2003). "Plasmid-encoded heavy metal resistance in Pseudomonas sp." Bull Environ Contam Toxicol **71**(6): 1145-50.

Valentine, D. L. (2007). "Adaptations to energy stress dictate the ecology and evolution of the Archaea." Nat Rev Microbiol **5**(4): 316-23.

Van Duyne, G. D. (2001). "A structural view of cre-loxp site-specific recombination." Annu Rev Biophys Biomol Struct **30**: 87-104.

Van Melderen, L. (2002). "Molecular interactions of the CcdB poison with its bacterial target, the DNA gyrase." Int J Med Microbiol **291**(6-7): 537-44.

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." J Bacteriol **190**(20): 6837-45.

Vickerman, M. M., N. M. Mather, P. E. Minick and C. A. Edwards (2002). "Initial characterization of the Streptococcus gordonii htpX gene." Oral Microbiol Immunol **17**(1): 22-31.

Visnes, T., M. Akbari, L. Hagen, G. Slupphaug and H. E. Krokan (2008). "The rate of base excision repair of uracil is controlled by the initiating glycosylase." DNA Repair (Amst) **7**(11): 1869-81.

Voorhorst, W. G., R. I. Eggen, A. C. Geerling, C. Platteeuw, R. J. Siezen and W. M. Vos (1996). "Isolation and characterization of the hyperthermostable serine protease, pyrolysin, and its gene from the hyperthermophilic archaeon Pyrococcus furiosus." J Biol Chem **271**(34): 20426-31.

Voorhorst, W. G., A. Warner, W. M. de Vos and R. J. Siezen (1997). "Homology modelling of two subtilisin-like proteases from the hyperthermophilic archaea Pyrococcus furiosus and Thermococcus stetteri." Protein Eng **10**(8): 905-14.

Vostrov, A. A., A. Malinin, V. N. Rybchin and A. N. Sravchevskii (1992). "[Construction of linear plasmid vectors for cloning in Escherichia coli cells]." Genetika **28**(7): 186-8.

Wachtershauser, G. (1990). "Evolution of the first metabolic cycles." Proc Natl Acad Sci U S A **87**(1): 200-4.

Wales, T. E., J. S. Richardson and M. C. Fitzgerald (2004). "Facile chemical synthesis and equilibrium unfolding properties of CopG." Protein Sci **13**(7): 1918-26.

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.

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): 124-33.

Wang, Y., E. P. Rocha, F. C. Leung and A. Danchin (2004). "Cytosine methylation is not the major factor inducing CpG dinucleotide deficiency in bacterial genomes." J Mol Evol **58**(6): 692-700.

Ward, D. E., I. M. Revet, R. Nandakumar, J. H. Tuttle, W. M. de Vos, J. van der Oost and J. DiRuggiero (2002). "Characterization of plasmid pRT1 from Pyrococcus sp. strain JT1." J Bacteriol **184**(9): 2561-6.

Warren, R. A. (1980). "Modified bases in bacteriophage DNAs." Annu Rev Microbiol **34**: 137-58.

Waters, E., M. J. Hohn, I. Ahel, D. E. Graham, M. D. Adams, M. Barnstead, K. Y. Beeson, L. Bibbs, R. Bolanos, M. Keller, K. Kretz, X. Lin, E. Mathur, J. Ni, M. Podar, T. Richardson, G. G. Sutton, M. Simon, D. Soll, K. O. Stetter, J. M. Short and M. Noordewier (2003). "The genome of Nanoarchaeum equitans: insights into early archaeal evolution and derived parasitism." Proc Natl Acad Sci U S A **100**(22): 12984-8.

Williams, E., T. M. Lowe, J. Savas and J. Diruggiero (2007). "Microarray analysis of the hyperthermophilic archaeon Pyrococcus furiosus exposed to gamma irradiation." Extremophiles **11**(1): 19-29.

Williams, K. P. (2002). "Integration sites for genetic elements in prokaryotic tRNA and tmRNA genes: sublocation preference of integrase subfamilies." Nucleic Acids Res **30**(4): 866-75.

Woese, C. R. and G. E. Fox (1977). "Phylogenetic structure of the prokaryotic domain: the primary kingdoms." Proc Natl Acad Sci U S A **74**(11): 5088-90.

Woese, C. R., O. Kandler and M. L. Wheelis (1990). "Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya." Proc Natl Acad Sci U S A **87**(12): 4576-9.

Wommack, K. E. and R. R. Colwell (2000). "Virioplankton: viruses in aquatic ecosystems." Microbiol Mol Biol Rev **64**(1): 69-114.

Wong, I., M. Amaratunga and T. M. Lohman (1993). "Heterodimer formation between Escherichia coli Rep and UvrD proteins." J Biol Chem **268**(27): 20386-91.

Yamashiro, K., S. I. Yokobori, T. Oshima and A. Yamagishi (2006). "Structural analysis of

the plasmid pTA1 isolated from the thermoacidophilic archaeon Thermoplasma acidophilum." Extremophiles.

Yang, H., J. H. Chiang, S. Fitz-Gibbon, M. Lebel, A. A. Sartori, J. Jiricny, M. M. Slupska and J. H. Miller (2002). "Direct interaction between uracil-DNA glycosylase and a proliferating cell nuclear antigen homolog in the crenarchaeon Pyrobaculum aerophilum." J Biol Chem **277**(25): 22271-8.

Yang, H., S. Fitz-Gibbon, E. M. Marcotte, J. H. Tai, E. C. Hyman and J. H. Miller (2000). "Characterization of a thermostable DNA glycosylase specific for U/G and T/G mismatches from the hyperthermophilic archaeon Pyrobaculum aerophilum." J Bacteriol **182**(5): 1272-9.

Yanisch-Perron, C., J. Vieira and J. Messing (1985). "Improved M13 phage cloning vectors and host strains: nucleotide sequences of the M13mp18 and pUC19 vectors." Gene **33**(1): 103-19.

Yin, Y. and D. Fischer (2006). "On the origin of microbial ORFans: quantifying the strength of the evidence for viral lateral transfer." BMC Evol Biol **6**: 63.

Yu, J. S. and K. M. Noll (1997). "Plasmid pRQ7 from the hyperthermophilic bacterium Thermotoga species strain RQ7 replicates by the rolling-circle mechanism." J Bacteriol **179**(22): 7161-4.

Yu, X., M. S. VanLoock, A. Poplawski, Z. Kelman, T. Xiang, B. K. Tye and E. H. Egelman (2002). "The Methanobacterium thermoautotrophicum MCM protein can form heptameric rings." EMBO Rep **3**(8): 792-7.

Zhang, X., T. Nakashima, Y. Kakuta, M. Yao, I. Tanaka and M. Kimura (2008). "Crystal structure of an archaeal Ski2p-like protein from Pyrococcus horikoshii OT3." Protein Sci **17**(1): 136-45.

Zhao, S. and K. P. Williams (2002). "Integrative genetic element that reverses the usual target gene orientation." J Bacteriol **184**(3): 859-60.

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 **2**(3): 131-40.

Zuckerkandl, E. and L. Pauling (1965). "Molecules as documents of evolutionary history." J Theor Biol **8**(2): 357-66.