Etude de la réplication de l'ADN chez les Archaea

Jonathan Berthon

UNIVERSITÉ PARIS-SUD 11

ARCHAEA
Laboratoire de Biologie Moléculaire du Gène chez les Extremophiles

cnrs

U.F.R. SCIENTIFIQUE D'ORSAY

THESE
présentée pour obtenir le grade de
DOCTEUR EN SCIENCES DE L'UNIVERSITE PARIS-SUD 11

N° d'ordre : 9256

Discipline : BIOLOGIE

par

Jonathan BERTHON

Etude de la réplication de l'ADN chez les Archaea

Thèse dirigée par le Professeur Patrick FORTERRE

Soutenue le Jeudi 27 Novembre 2008

Composition du jury :

M. Nicolas GLANSDORFF	Rapporteur
M. Joël QUERELLOU	Rapporteur
M. Daniel GAUTHERET	Président
M. Laurent JANNIERE	Examinateur
M. Patrick FORTERRE	Directeur de thèse

Prologue

Ma première rencontre anonyme avec le professeur Patrick Forterre remonte à la lecture d'un article sur l'origine du génome publié en Novembre 2000 dans un numéro du magazine La Recherche consacré aux origines de la vie ; Patrick y développait déjà son fétichisme des virus. Ma seconde rencontre quelque peu insolite avec le professeur Patrick Forterre eut lieu dans un amphithéâtre. Ce jour là, Patrick, que je n'avais jusque-là jamais rencontré, arriva à son cours en retard. Il ne prit pas le temps de se présenter, bredouilla quelques excuses pour justifier son retard, nous révélant que, le nez plongé dans un article passionnant, il n'avait pas levé les yeux à temps et avait manqué son arrêt de RER. Il commença ensuite son cours, toujours sans avoir fait les présentations, et parla de l'origine du génome et, entre autres choses, de l'étymologie du mot ribosome. Réalisant soudain que le professeur anonyme que j'avais devant moi parlait avec les mots du professeur Patrick Forterre, je m'indignais que celui-ci parla, sans même citer ses sources, des idées d'un autre, mais compris tout aussi vite qu'il s'agissait certainement du professeur Patrick Forterre lui-même. Lorsque le cours prit fin et qu'il se présenta enfin, j'eus la confirmation que j'attendais, et m'en tenais à cette première impression : curieux personnage !

Je recroisais ensuite Patrick dans le module Evolution de mon cursus de maîtrise, en apprit davantage sur les Archaea, sur Carl Woese et l'origine du vivant. Je dus ensuite faire face à Patrick pour l'examen oral de ce module pendant lequel je flattai bassement son ego en lui disant que j'appréciai sa théorie de la thermoréduction (que je trouve réellement séduisante). Nous parlâmes aussi du roi Carl Woese dont j'avais pris la peine, en ma qualité d'étudiant zélé, de lire quelques articles pour préparer l'examen. Je dis avec franchise que

j'avais trouvé la prose du bonhomme un brin élitiste, presque ésotérique et son ton péremptoire, voire cynique. C'était sans savoir que Carl avait fait sa traversée du désert, s'efforçant d'imposer son concept des trois domaines cellulaires du vivant à une communauté scientifique qui n'en voulait pas. Je dus malgré tout faire bonne impression lors de cet oral car Patrick s'en alla chercher quelques tirés à part de ses articles qu'il me glissa dans la main avant notre séparation, et sur ces entrefaites j'abandonnai ma première impression pour ma deuxième impression : sympa le gars !

La suite est un concours de circonstances : je viens frapper à la porte de Patrick car je suis dans la panade, sans DEA pour m'accepter. Il se démène, me trouve deux stages pour que je reste au contact de la science. Rebelote l'année suivante, à l'issue du DEA, mais Patrick a des projets pour moi, une thèse, et veut m'envoyer au Japon pendant un an, voire deux ans.

Quelle idée, je n'ai jamais dit aimer les sumotoris !

Le reste n'est qu'anecdotes : la confrontation entre deux cultures, du riz, une charmante entremetteuse, un restaurant de grillades, encore du riz, un gâteau à la patate douce, un anorak fuchsia, une pommade à lèvres goût cerise, toujours du riz, un mariage, et bientôt qui sait des petits sumos ?

Remerciements

Je n'aurais pas de mots assez forts et assez justes pour remercier Patrick pour tout ce qu'il a fait pour moi ces dernières années. Je le remercie solennellement pour m'avoir donné l'opportunité de réaliser une thèse entre France et Japon. Je remercie également Evelyne pour sa formidable générosité.

Je souhaite ensuite remercier le professeur Yoshizumi Ishino pour m'avoir accueilli dans son laboratoire au Japon. Je remercie aussi l'ensemble des membres de son équipe pour leur accueil chaleureux, leur bonne humeur, leur gentillesse empruntée.

Je tiens à remercier plus particulièrement Kenzo Fukunaga pour m'avoir épaulé lorsque je subissais le contrecoup du choc culturel ; merci également à Shinichi Kiyonari pour son aide avec le Biacore.

Je me dois de remercier Gilles Henckès, Robert Aufrère et leurs étudiants pour m'avoir assisté dans le clonage des gènes de réplication et de traduction.

Je remercie aussi Sophie Quevillon-Cheruel pour son aide technique et ses conseils. Je remercie au même titre Arnaud Hecker pour m'avoir coaché lors de ma première année de thèse au laboratoire BMGE.

Je tiens aussi à remercier Ryosuke Fujikane pour avoir répondu avec une inaltérable gentillesse et une inébranlable persévérance à mes stupides interrogations.

Je remercie vivement Céline Brochier-Armanet pour sa relecture détaillée et critique des épreuves du Chapitre IV, ses conseils d'écriture, ses recommandations et ses corrections.

Je remercie Monsieur Nicolas Glansdorff et Monsieur Joël Querellou pour avoir gentiment accepté d'être mes rapporteurs. Merci, également à Monsieur Laurent Jannière et à Monsieur Daniel Gautheret pour avoir bien voulu faire partie de mon jury de thèse.

Je remercie également mes soutiens financiers lors de mon équipée japonaise : le Collège Doctoral Franco-Japonais et la Japan Society for Promotion of Science.

Je remercie aussi tous les anonymes de ma classe de salsa pour avoir égayé de manière hebdomadaire ma dernière année de thèse.

Je remercie ma famille et mes proches pour leur soutien indéfectible et leurs nombreux encouragements.

Un immense merci à Tsukiko et Toshihiko Kaneco pour avoir accueilli un *gaijin* dans leur famille et m'avoir considéré comme un fils malgré mon pesant mutisme japonais. Merci à Kayoko, ma grande sœur du bout du monde, pour sa gentillesse.

Enfin et surtout, je remercie infiniment ma femme Yumiko pour son soutien moral. Je tiens aussi à lui exprimer mon admiration pour son abnégation et pour le courage dont elle fait preuve chaque jour en cette terre d'exil pour vivre à mes côtés.

Summary

Study of DNA replication in Archaea

Cellular organisms belong to one of the three domains of life: Archaea, Bacteria, and Eucarya. Archaea are unicellular organisms with a bacterial phenotype, yet they exhibit many eucaryotic features at the molecular level. In particular, archaeal DNA replication machinery is a homologous and simplified version of that in eucaryotes. In this work, I have studied archaeal DNA replication with both *in vitro* and *in silico* approaches.

First, I have tried to purify the archaeal replication initiator protein Cdc6/Orc1 in its native form, in the hope to set up the first *in vitro* archaeal DNA replication system. Unfortunately, this approach was not successful, because of the aggregation properties and instability of the protein.

Secondly, I have carried out a comparative analysis of the genome context of DNA replication genes in archaeal genomes. This analysis has led to the identification of a widely conserved association between DNA replication and ribosome-related genes. This genomic organization suggests the existence of a mechanism coupling DNA replication and translation. Remarkably, experimental evidence supporting this idea is emerging piecemeal from biochemical studies in bacteria and eucaryotes. I have then set up some experimental tools that will allow testing some of these *in silico* predictions in the near future.

Finally, I have examined the taxonomic distribution of DNA replication genes in archaeal genomes in order to infer the probable composition of the DNA replication machinery of the last archaeal common ancestor. Overall, the phyletic patterns of archaeal DNA replication genes suggest that the DNA replication apparatus of this ancestor was likely more complex than that of contemporary archaeal cells.

Keywords: Archaea | DNA replication | Cdc6/Orc1 | Genomic context | Comparative analysis | Evolution | Functional coupling | Ribosome

Résumé

Etude de la réplication chez les Archaea

Les organismes cellulaires appartiennent à l'un des trois domaines du vivant : Archaea, Bacteria, Eucarya. Les Archaea sont des organismes unicellulaires avec un phénotype bactérien mais qui possèdent de nombreux caractères moléculaires eucaryotes. En particulier, la machinerie de réplication archéenne est une version homologue et simplifiée de celle des eucaryotes. Au cours de cette thèse, j'ai étudié la réplication de l'ADN chez les Archaea en combinant des approches *in vitro* et *in silico*.

Premièrement, j'ai essayé de purifier la protéine initiatrice de la réplication Cdc6/Orc1, sous une forme native, dans l'espoir de mettre au point le premier système de réplication de l'ADN *in vitro* chez les Archaea. Malheureusement, cette approche a été infructueuse en raison de l'instabilité et des propriétés d'agrégation de la protéine.

Deuxièmement, j'ai réalisé une analyse comparative du contexte génomique des gènes de réplication dans les génomes d'Archaea. Cette analyse nous a permis d'identifier une association très conservée entre des gènes de la réplication et des gènes liés au ribosome. Cette organisation suggère l'existence d'un mécanisme de couplage entre la réplication de l'ADN et la traduction. De manière remarquable, des données expérimentales obtenues chez des modèles bactériens et eucaryotes appuient cette idée. J'ai ensuite mis au point des outils expérimentaux qui permettront d'éprouver la pertinence biologique de certaines des prédictions effectuées.

Finalement, j'ai examiné la distribution taxonomique des gènes de la réplication dans les génomes d'Archaea afin de prédire la composition probable de la machinerie de réplication de l'ADN chez le dernier ancêtre commun des Archaea. Dans leur ensemble, les profils phylétiques des gènes de la réplication suggèrent que la machinerie ancestrale était plus complexe que celle des organismes archéens contemporains.

Mots clefs : Archaea | Réplication de l'ADN | Cdc6/Orc1 | Contexte génomique | Analyse comparative | Evolution | Couplage fonctionnel | Ribosome

Table des matières

Liste des tableaux et figures

Science is impelled by two main factors, technological advance and a guiding vision (overview). A properly balanced relationship between the two is key to the successful development of a science: without the proper technological advances the road ahead is blocked. Without a guiding vision there is no road ahead; the science becomes an engineering discipline, concerned with temporal practical problems. […] A society that permits biology to become an engineering discipline, that allows that science to slip into the role of changing the living world without trying to understand it, is a danger to itself. […] Biology today is little more than an engineering discipline. […] biology is here to understand the world, not primarily to change it. Biology's primary job is to teach us.

Woese, C.R. (2004). A new biology for a new century. Microbiol Mol Biol Rev 68, 173-186.

Note liminaire

Les termes ‘procaryote’ et ‘eucaryote’ font partie intégrante du vocabulaire du microbiologiste contemporain au même titre que bactérie, microbe, microorganismes — parfois archaebactérie, voire archée. Dans les manuels scolaires ou les ouvrages de référence, ces deux termes sont généralement introduits au travers de la dichotomie procaryote/eucaryote qui sert à distinguer les cellules possédant un noyau (les eucaryotes) de celles qui n’en possèdent pas (les bactéries et les archées). Or, l’acception courante et moderne des termes ‘procaryote’ et ‘eucaryote’ correspond à un travestissement de sens par rapport à la définition entendue par les auteurs qui en ont réintroduit l’usage (Sapp, 2005; Woese, 2004). D’autre part, l’origine historique de la distinction entre procaryotes et eucaryotes, attribuée au protistologiste français Edouard Chatton, est confuse (pour des revues, voir (Katscher, 2004; Sapp, 2005)).

La réintroduction des termes ‘procaryote’ et ‘eucaryote’ a eu lieu dans les années 1960 sous l’impulsion des microbiologistes Stanier et van Niel (Stanier and Van Niel, 1962). Or, le recours à cette nomenclature s’inscrivait dans un contexte scientifique particulier (Sapp, 2006). Selon ces deux auteurs, les tentatives d’élaboration d’une classification phylo-génétique bactérienne cohérente se heurtaient à des difficultés insurmontables qui rendaient vain l’espoir d’aboutir à une classification taxonomique en ce qui concerne les organismes bactériens (Stanier and Van Niel, 1962). Néanmoins, ils considéraient qu’il était possible de distinguer les bactéries des eucaryotes en se basant sur des critères d’organisation interne de la cellule et définirent dans cet objectif le concept de bactérie (Stanier and Van Niel, 1962). Par conséquent, en aucune manière la distinction faite par Stanier et van Niel entre

‘procaryote’ et ‘eucaryote’ dans le cadre de leur réflexion ne se voulait une distinction d’ordre taxonomique ou évolutive (voir, (Sapp, 2006)). Cette distinction était purement organisationnelle et visait à distinguer les organismes eucaryotes, qui possèdent une organisation interne complexe (organelles, compartimentation membranaire), des organismes non-eucaryotes, à l’architecture interne plus simple (voir, (Sapp, 2005)).

Or, dans son usage moderne, le terme ‘procaryote’ est employé au même rang que le terme ‘eucaryote’, c’est-à-dire avec un sens taxonomique (Woese, 1994). Le terme ‘eucaryote’ définit un groupe monophylétique et a donc un sens au regard de l’évolution des organismes ; le terme ‘procaryote’ ne désigne pas un groupe monophylétique mais doit se comprendre comme la part non-eucaryote des organismes cellulaires (Woese, 2004). A ce titre, la distinction entre organismes ‘procaryote’ et ‘eucaryote’ apparaît superfétatoire, voire fallacieuse d’un point de vue évolutif, au même titre que les grades ‘reptiles’, ‘poissons’ et autres ‘invertébrés’. En attribuant une valeur taxonomique au mot ‘procaryote’, ses utilisateurs cherchent souvent à minimiser les différences fondamentales existant entre les types cellulaires bactérien et archéen parce qu’ils sont convaincus que la distinction évolutive majeure se situe entre les eucaryotes d’une part et les bactéries et les archées d’autre part (Woese, 1994, 2004). Par ailleurs, la littérature scientifique fourmille d’exemples où les auteurs emploient improprement le terme ‘procaryote’ en lieu et place de ‘bactérie’, ce qui témoigne du fait que l’acception courante du mot ‘procaryote’ est vague, donc que son utilisation est souvent inappropriée.

Certains auteurs prônent un changement lexical pour éviter que ne se perpétue l’utilisation de cette division évolutive arbitraire (Goldenfeld and Woese, 2007; Pace, 2006). D’autres auteurs défendent au contraire son utilisation soit parce qu’ils ne reconnaissent pas l’existence des trois domaines (Mayr, 1998), soit parce qu’ils considèrent les archées comme de vulgaires bactéries (Martin, 2005), soit parce qu’ils considèrent que la différence

d'organisation cellulaire entre les organismes eucaryotes d'une part et les organismes archéen et bactérien d'autre part justifie le maintien de la dichotomie procaryote/eucaryote (Martin and Koonin, 2006; Walsh and Doolittle, 2005). Une alternative pragmatique consistant à substituer les termes 'procaryote' et 'eucaryote' par des termes neutres pour effacer toute connotation évolutive a été proposée (Forterre, 1992), mais cette nomenclature n'a pas été adoptée par la communauté scientifique.

En tant que fervent *archaeologue*, j'ai choisi de m'inscrire dans la révolution lexicale prônée par Carl Woese et Norman Pace et de ne pas employer le terme 'procaryote' dans la suite de ce mémoire. Aussi, j'utiliserai les termes Archaea, Bacteria et Eucarya pour désigner les différents domaines cellulaires ; les noms communs archée, bactérie et eucaryote pour désigner les différents types d'organismes cellulaires ; les adjectifs archéen, bactérien et eucaryotique pour qualifier les caractères propres aux domaines Archaea, Bacteria et Eucarya, respectivement.

INTRODUCTION

Classification phylogénétique du vivant

Présentation générale des Archaea

La réplication de l'ADN

Analyse comparative des génomes

Présentation de la thématique de recherche

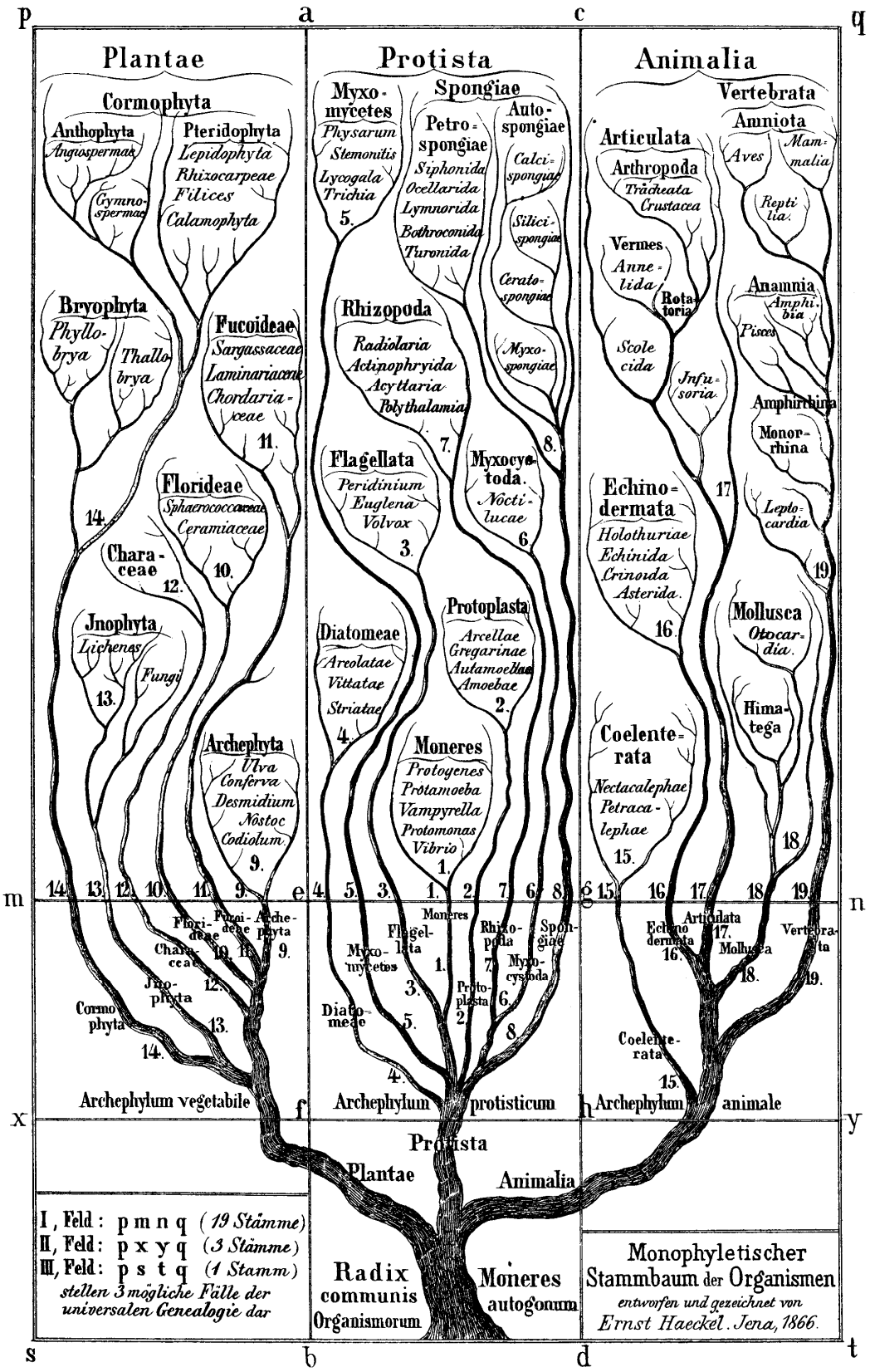

Figure 1 : Le premier arbre généalogique du vivant, publié par Haeckel. Haeckel distingue trois royaumes : les plantes (Plantae), les protistes (Protista) et les animaux (Animalia)
D'après (Haeckel, 1866).

Classification phylogénétique du vivant

Les classifications naturelles

Depuis l'Antiquité, les naturalistes s'efforcent de classer les êtres vivants en regroupant en des ensembles cohérents des organismes partageant des ressemblances significatives et des caractères distinctifs. Pendant de nombreux siècles, les observateurs se basèrent sur le canevas élaboré par Aristote — qui distinguait deux règnes (les animaux et les végétaux) — pour dresser leurs classifications mais celles-ci souffraient de nombreuses lacunes. Au XVIII^{e} siècle, le naturaliste suédois Carl von Linné réforma la systématique en formulant les bases d'une classification naturelle des êtres vivants, encore en vigueur aujourd'hui.

Au cours des XVIII^{e} et XIX^{e} siècles, les progrès constants de la biologie stimulèrent l'évolution conjointe des idées scientifiques et des courants philosophiques et contribuèrent à l'amélioration graduelle de la compréhension de l'unité du vivant et à la reconnaissance des bases fondamentales de l'évolution. La théorie de l'évolution formulée en 1859 par Charles Darwin dans son livre *De l'origine des espèces* marque un tournant majeur dans l'histoire de la biologie car elle cristallise l'ensemble des idées progressistes contemporaines et prend le contrepied des doctrines téléologiques défendues par l'Eglise. Dès l'énoncé par Darwin de sa théorie de la descendance avec modification, le naturaliste allemand Ernst Haeckel s'empara de l'idée et proposa le premier arbre généalogique du vivant dans lequel il distinguait trois royaumes : les plantes (Plantae), les protistes (Protista) et les animaux (Animalia) (Haeckel, 1866) **(Figure 1)**. La classification du monde vivant fut ensuite régulièrement révisée au cours du XXème siècle et la description d'Haeckel fut supplantée par la classification à

quatre royaumes de Copeland (Copeland, 1938) puis celle à cinq royaumes de Whittaker (Whittaker, 1969).

Pendant de nombreux siècles, les naturalistes se servirent de critères morphologiques pour classer les organismes vivants mais ceux-ci se révélèrent vite d'un intérêt restreint en ce qui concerne les microorganismes ; ils montrèrent aussi certaines limites pour classer les animaux. Pour certains microbiologistes, l'élaboration d'une taxonomie pour les bactéries apparaissait même insurmontable, si bien qu'ils abandonnèrent l'idée de jamais y parvenir (Sapp, 2005; Stanier and Van Niel, 1962). La popularisation des concepts de systématique phylogénétique de Willi Hennig et la formulation des principes de la phylogénie moléculaire par Zuckerkandl et Pauling permirent finalement de dépasser cette apparente difficulté (Zuckerkandl and Pauling, 1965).

Emergence du concept d'Archaea

La représentation du monde vivant changea radicalement dans les années 1970, lorsque Carl Woese entreprit de caractériser la structure de l'arbre phylogénétique du vivant en s'appuyant sur l'analyse de la séquence de l'ARN ribosomique — une molécule présente dans tous les systèmes autoréplicatifs, facilement isolable, qui évolue à un rythme qui s'accorde bien avec l'étude des relations de parenté entre des espèces très éloignées. La technique utilisée par Carl Woese consiste à cultiver des organismes dans un milieu marqué au ^{32}P, d'isoler l'ARN ribosomique de la petite sous-unité puis d'analyser le profil de digestion par la RNase T1 par autoradiographie. L'image obtenue est un nuage de tâches dont chacune correspond à un oligonucléotide de séquence donnée. L'ensemble des tâches, le catalogue oligonucléotidique, définit l'organisme à partir duquel il a été préparé. La différence entre deux catalogues A et B est représentée par la valeur d'un coefficient de similitude S_{AB} ; ce coefficient donne une idée

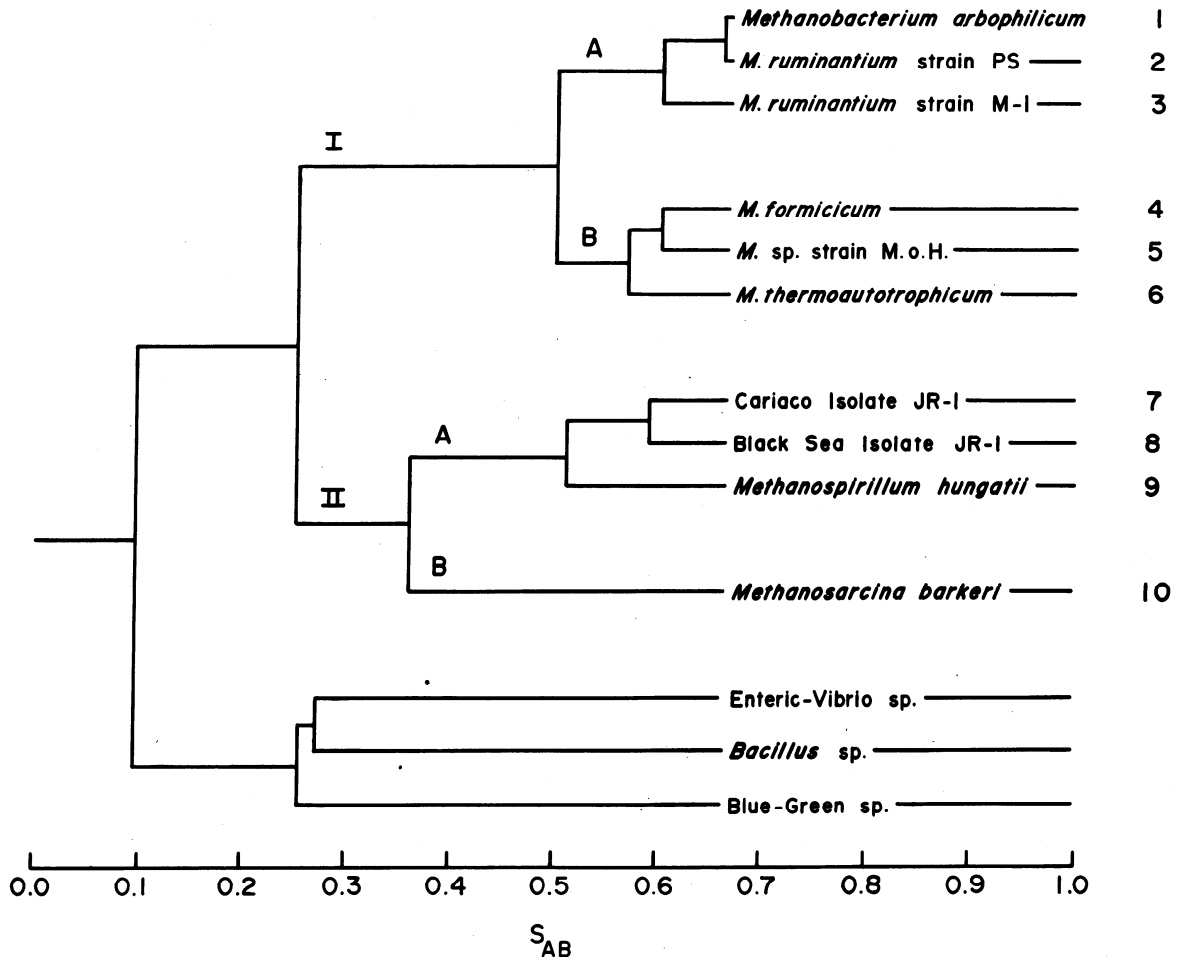

Figure 2 : Dendrogramme des relations entre les méthanogènes et les bactéries typiques. La distance entre les différentes espèces est proportionnelle au degré de ressemblance entre leur catalogue oligonucléotidique respectif, symbolisé par la valeur du coefficient de similitude S_{AB} (échelle en bas de la figure). Les dix espèces méthanogènes (chiffres situés à droite) sont clairement distincts des bactéries représentées ici par *Vibrio*, *Bacillus* et une cyanobactérie. Les méthanogènes comprennent deux divisions (I et II). Chacune des divisions comprend elle-même deux sous-groupes (A et B).
D'après (Fox et al., 1977).

de la distance entre les organismes considérés. En suivant ce principe, Carl Woese et ses collaborateurs montrèrent que les organismes couramment désignés 'bactéries méthanogènes' sont en réalité clairement distincts des bactéries typiques et forment probablement un groupe phylogénétique à part entière (Fox et al., 1977) **(Figure 2)**. Ensuite, ils compilèrent les coefficients de similitude mesurés entre des représentants des eucaryotes, des bactéries et des méthanogènes et montrèrent que ceux-ci se répartissent en trois groupes distincts. Aussi, Carl Woese et George Fox remirent en cause la vision dichotomique du vivant et proposèrent de répartir les organismes cellulaires non pas en deux mais en trois grands groupes d'ordre

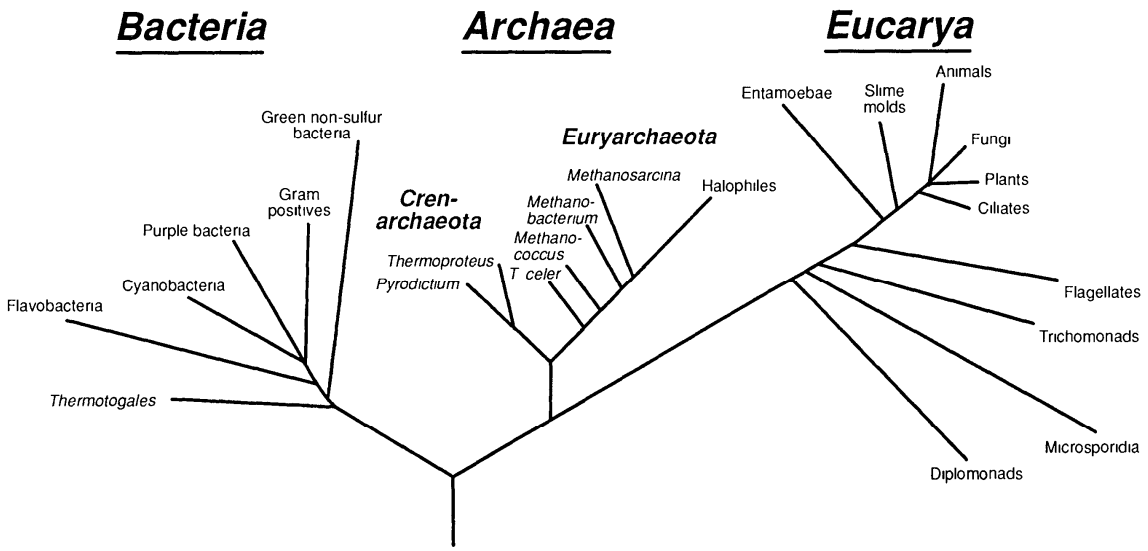

Figure 3 : Arbre phylogénétique universel raciné montrant les relations entre les trois domaines cellulaires du vivant. Cet arbre est basé sur la comparaison des séquences des ARN ribosomiques. Les organismes cellulaires se répartissent en trois domaines appelés *Bacteria*, *Archaea* et *Eucarya*. Les *Archaea* sont divisées en deux phylums nommés *Crenarchaeota* et *Euryarchaeota*.
D'après (Wheelis et al., 1992).

équivalent. Ces trois groupes — originellement 'urkingdoms' (Woese and Fox, 1977), puis renommés domaines (Woese et al., 1990) — furent baptisés eubacteria, archaebacteria et urkaryotes (Woese and Fox, 1977). Ce concept fut mal reçu par la communauté des microbiologistes qui considérait les archaebactéries comme des bactéries un peu spéciales — de vieilles bactéries comme leur nom l'indiquait. Pourtant, les éléments corroborant la singularité de ce nouveau groupe phylogénétique s'accumulaient petit à petit (nature des lipides membranaires, coenzymes spécifiques, composition de l'ARN polymerase). Dans le but d'asseoir la légitimité de la vision tripartite du monde cellulaire vivant, Carl Woese et ses collaborateurs proposèrent de réformer la classification systématique en introduisant le terme de domaine pour désigner le taxon de plus haut rang (Woese et al., 1990). En outre, afin d'éviter que ne se perpétue la tendance des microbiologistes à confondre les bactéries et les archaebactéries, ils suggèrent d'abandonner les termes archaebactéries, eubactéries et urkaryotes, et proposèrent d'appeler les trois domaines du vivant *Archaea, Bacteria* et *Eucarya* (Woese et al., 1990) **(Figure 3)**. Ce vent de réforme fut une fois de plus mal accueilli

et certains tentèrent de contrer l'initiative de Carl Woese en cherchant à souligner la nature bactérienne des Archaea pour tenter de maintenir la classification en place, qui distinguait les organismes eucaryotes des organismes non-eucaryotes (Mayr, 1998; Wheelis et al., 1992; Woese, 1998).

Les nombreux travaux réalisés sur les Archaea, les nombreuses analyses comparatives des génomes effectuées, les phylogénies moléculaires construites à partir de marqueurs choisis, etc., l'ensemble des données accumulées depuis soutient la pertinence de la division du monde cellulaire vivant en trois domaines d'ordre équivalent (Forterre et al., 2002; Gribaldo and Brochier-Armanet, 2006). Pourtant, le scepticisme, voire l'ignorance de certains microbiologistes à l'égard du concept d'Archaea persiste encore aujourd'hui (Woese, 2007).

L'arbre du vivant : une chimère ?

Alors que certains auteurs voient dans les résultats des analyses comparatives des génomes, des arguments en faveur de l'existence de trois domaines cellulaires distincts, d'autres y voient au contraire matière à contester la pertinence même de chercher à construire des arbres phylogénétiques (pour des revues, voir (McInerney et al., 2008; Wolf et al., 2002)). En effet, si ces analyses ont bien montré que chacun des trois domaines cellulaires présente des caractères génétiques distinctifs, elles ont également mis en lumière l'existence d'un important flux de gènes entre génomes parfois très éloignés. Or, selon certains auteurs ces transferts horizontaux de gènes sont si nombreux qu'ils compromettent le concept d'espèce (Doolittle and Papke, 2006), rendent futile l'idée de chercher à construire un arbre phylogénétique du vivant (Doolittle, 1999), voire appellent à repenser les bases de notre système de classification naturelle (Bapteste and Boucher, 2008). D'autres auteurs estiment toutefois que l'impact réel des transferts horizontaux de gènes est surestimé et que la

descendance verticale reste le principal moteur de l'évolution des organismes cellulaires (Glansdorff, 2000; Kurland et al., 2003). L'influence de l'acquisition de gènes en provenance d'autres espèces est difficile à estimer mais des études suggèrent que certains transferts de gènes ont bien joué un rôle dans l'évolution des organismes récepteurs. La colonisation des niches hyperthermophiles par les bactéries après l'acquisition de la reverse gyrase en provenance des archées en est l'exemple le plus marquant (Forterre et al., 2000).

Néanmoins, il semble qu'une certaine classe de gènes soit majoritairement réfractaire aux transferts entre organismes parce que les protéines qu'ils codent sont au cœur d'un réseau d'interactions fonctionnelles complexes et intimes (Jain et al., 1999; Woese, 2004). Par conséquent, ces gènes devraient permettre de retracer l'histoire généalogique des organismes qui les portent. Récemment, des auteurs ont rapporté la mise au point d'une procédure d'automatisation de la construction de l'arbre du vivant basée sur l'ensemble des protéines universellement conservées dans les génomes (Ciccarelli et al., 2006). Mais, le principe de cette approche a été tourné en dérision dans un article au titre provocateur : "The Tree of one percent" (Dagan and Martin, 2006) — le nombre de protéines universelles représente environ un pour cent du protéome d'une bactérie (voir également (McInerney et al., 2008)).

Aujourd'hui encore, les questions autour de l'arbre du vivant, de la position de la racine, de la nature du dernier ancêtre cellulaire commun font l'objet d'intenses débats. D'un côté, certains auteurs mettent l'accent sur le rôle fondamental joué par les transferts horizontaux de gènes et contestent le concept même d'arbre du vivant pour la représentation de l'évolution des organismes (Doolittle and Bapteste, 2007), lui préférant l'image d'un réseau (Doolittle, 2000). Cette famille d'auteurs s'intéresse principalement à la dynamique de l'évolution des génomes dans leur ensemble. A ce titre, ils insistent sur l'influence des transferts horizontaux car ceux-ci brouillent le signal phylogénétique. La conséquence de ce signal parasite est que l'arbre des gènes et l'arbre des espèces qui les portent ne sont pas

congruents. Aussi, ces scientifiques cherchent à mettre au point des modèles mathématiques et des programmes informatiques permettant de mieux décrire l'évolution de l'ensemble des gènes contenus dans les génomes (Dagan et al., 2008). Une autre famille d'auteurs s'intéresse principalement à l'héritage universel des organismes cellulaires vivants. Ces auteurs estiment que, en dépit de l'influence des transferts de gènes entre organismes, il est envisageable d'identifier un cœur de gènes qui a été hérité uniquement de manière verticale depuis la période du dernier ancêtre cellulaire universel. Aussi, il demeure possible de s'appuyer sur les gènes qui fondent l'identité de la cellule qui les porte pour en retracer l'histoire évolutive et prédire la nature éventuelle du dernier ancêtre cellulaire commun.

Par ailleurs, le recours à des méthodes globales, connues sous le nom de méthodes phylogénomiques, semble ouvrir des voies prometteuses vers la résolution, au moins partielle, de la structure de l'arbre phylogénétique du vivant (Delsuc et al., 2005). Au-delà du 'simple' exercice de reconstruction de l'arbre du vivant, le défi majeur de la biologie évolutive reste de comprendre comment évoluent les organismes et leurs génomes, qu'il s'agisse d'une évolution verticale ou horizontale. En outre, les analyses comparatives de génomes pourraient permettre de remonter artificiellement le cours de l'évolution et tenter de prédire le contenu probable du génome du dernier ancêtre cellulaire commun.

Présentation générale des Archaea

Caractères biologiques propres aux Archaea

Très tôt après la mise en évidence de l'existence de trois domaines cellulaires phylogénétiquement distincts, des données expérimentales corroborant cette idée furent rassemblées par Carl Woese et ses collaborateurs. Les données accumulées auparavant sur les méthanogènes, les premiers organismes reconnus comme Archaea, avaient déjà révélé la présence de coenzymes uniques mais aussi l'absence d'un attribut typiquement bactérien, le peptidoglycane (voir (Woese and Fox, 1977)). Par la suite, les caractères atypiques d'autres 'bactéries', présentés jusqu'alors comme des adaptations, furent reconsidérés comme de possibles particularités biologiques des Archaea (voir (Woese, 2007)). Le premier des caractères reconnu comme archéen est celui qui fit réaliser à Carl Woese l'importance de sa découverte : la signature autoradiographique de l'ARN ribosomique (ARNr) (voir (Woese, 2007)). D'autres phénotypes furent rapidement reconnus comme propres aux Archaea, telle l'absence de peptidoglycane dans la paroi cellulaire ou l'existence d'une version modifiée de la séquence 'canonique' T- ψ-C-G au niveau du bras commun de l'ARN de transfert (voir (Woese et al., 1978)). Depuis les travaux pionniers de Woese, une revue bibliographique des études menées chez les Archaea permet de faire ressortir trois caractères qui distinguent les archées des bactéries et des eucaryotes :

- la stéréochimie du glycérol phosphate sur lequel sont bâtis les phospholipides membranaires (voir (Boucher, 2007; Pereto et al., 2004)) ;
- le caractère non pathogène des archées — plus précisément, l'absence de cas circonstancié faisant état d'un lien direct entre le développement d'une maladie et la présence d'une

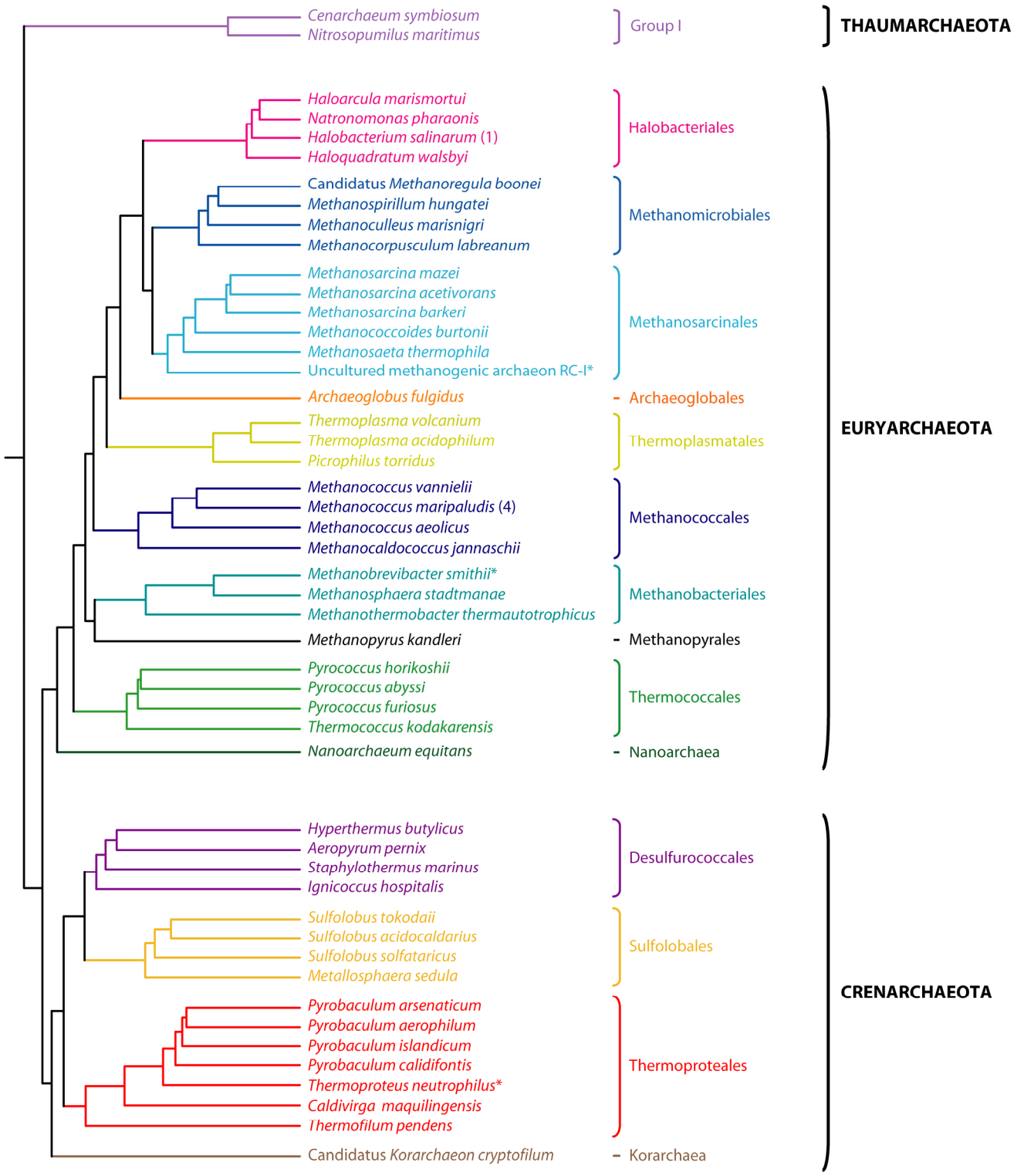

Figure 4 : Phylogénie consensuelle des Archaea. Cette phylogénie des Archaea correspond à une compilation d'analyses récentes basées sur différents marqueurs moléculaires (à l'exception des organismes signalés par un astérisque). Ces analyses soutiennent l'existence de trois phylums principaux au sein du domaine Archaea : les Thaumarchaeota, les Crenarchaeota et les Euryarchaeota (Brochier-Armanet et al., 2008). La position du méthanogène non cultivé isolé à partir d'une rizière 'Uncultured methanogenic archaeon RC-I (Methanosarcinales) est basé sur l'analyse de l'ARNr 16S (Erkel et al., 2006). Les positions de *Thermoproteus neutrophilus* (Thermoproteales) et *Methanobrevibacter smithii* (Methanobacteriales) ont été inférées respectivement à partir de la classification taxonomique et du contenu génomique (Samuel et al., 2007) ; elles sont donc provisoires.

archée (pour une vue d'ensemble, voir (Cavicchioli et al., 2003; Eckburg et al., 2003; Lepp et al., 2004; Reeve, 1999; Vianna et al., 2006)) —, qui pourrait reposer sur des critères biochimiques (Martin, 2004) ou, plus vraisemblablement, écologiques (Valentine, 2007) ;

- la méthanogenèse (Ferry and Kastead, 2007) et l'oxydation anaérobique du méthane (Boetius et al., 2000; DeLong, 2000; Hallam et al., 2004; Hinrichs et al., 1999; Orphan et al., 2001; Raghoebarsing et al., 2006; Thauer and Shima, 2006).

Phylogénie des Archaea

Historiquement, les Archaea étaient réparties en deux groupes majeurs comprenant d'une part les thermophiles extrêmes, d'autre part les méthanogènes et les organismes apparentés (Woese, 1987). Par la suite, ces deux phylums furent baptisés Crenarchaeota et Euryarchaeota (Woese et al., 1990). Les Crenarchaeota — également décrites sous le terme Eocytes (Lake, 1988) — correspondent à un groupe d'organismes thermophiles sulfate dépendant (Woese et al., 1990). Le nom Crenarchaeota vient du grec *crenos*, qui signifie source, pour indiquer le fait que ce groupe d'organismes ressemblerait au phénotype de l'ancêtre commun des Archaea, supposé hyperthermophile (Forterre et al., 2002; Woese et al., 1990). Les Crenarchaeota cultivables comprennent trois ordres : les Desulfurococcales, les Sulfolobales et les Thermoproteales **(Figure 4)**. Les Euryarchaeota correspondent à un groupe d'organismes phénotypiquement hétérogène comprenant des méthanogènes, des halophiles, des thermophiles et une espèce réduisant le soufre (Forterre et al., 2002; Woese et al., 1990). Les Euryarchaeota cultivables comprennent neuf ordres : les Thermococcales, les Methanopyrales, les Methanobacteriales, les Methanococcales, les Thermoproteales, les Archaeoglobales, les Methanosarcinales, les Methanomicrobiales et les Halobacteriales

(Figure 4). L'existence d'un troisième phylum, nommé Korarchaeota, a été proposée sur la base de l'analyse de séquences d'ARNr 16S, amplifiées par PCR à partir d'un échantillon prélevé dans une source chaude du parc Yellowstone (Barns et al., 1996). La séquence génomique d'un représentant de ce groupe taxonomique suggère qu'il s'agit d'une lignée ancienne, présentant un mélange de caractères propres aux Crenarchaeota et aux Euryarchaeota, qui branche vraisemblablement à la base des Crenarchaeota ((Elkins et al., 2008) ; Céline Brochier, communication personnelle) **(Figure 4)**. L'archée nanoscopique *Nanoarchaeum equitans*, qui vit en symbiose avec une archée hyperthermophile du genre *Ignicoccus*, a été proposée comme le membre fondateur d'un quatrième phylum baptisé Nanoarchaeota sur la base de l'analyse de l'ARNr 16S (Huber et al., 2002). Des analyses phylogénétiques plus récentes, basées sur l'utilisation de marqueurs moléculaires alternatifs, suggèrent que *N. equitans* représente plutôt une lignée euryarchéenne évoluant rapidement, probablement affiliée aux Thermococcales (Brochier et al., 2005) ; cette affiliation est supportée par d'autres analyses génomiques comparatives (Dutilh et al., 2008; Makarova and Koonin, 2005). Le récent séquençage du génome de l'archée mésophile *Cenarchaeum symbiosum* a suscité des questions quant à l'affiliation du groupe d'organismes qu'il représente (groupe I selon la désignation adoptée par De Long (DeLong, 1992)) vis-à-vis des organismes cultivables des phylums Crenarchaeota et Euryarchaeota. Historiquement, ce groupe d'organismes marins avait été associé au phylum des Crenarchaeota sur la base de l'amplification de la séquence des ARNr 16S (Brochier-Armanet et al., 2008; DeLong, 1992; Fuhrman et al., 1992). Or, des doutes ont depuis été émis sur la fiabilité des reconstructions basées sur l'ARNr 16S, notamment lorsqu'il s'agit de résoudre la position d'un embranchement basal. A ce titre, l'utilisation des protéines ribosomales semble une alternative intéressante pour clarifier la position de groupes atypiques (Brochier et al., 2004; Brochier et al., 2005). En s'appuyant sur l'analyse de ces marqueurs moléculaires, il a été

montré que *C. symbiosum* et les organismes mésophiles qui lui sont apparentés représentent probablement un nouveau phylum : les Thaumarchaeota (Brochier-Armanet et al., 2008). Une analyse comparative des génomes d'Archaea avait d'ailleurs déjà souligné la nature singulière du génome de *C. symbiosum* (Makarova et al., 2007). Par conséquent, les données les plus récentes appuient l'idée que les Archaea se divisent en trois phylums principaux : les Crenarchaeota, les Euryarchaeota et les Thaumarchaeota (Brochier-Armanet et al., 2008) **(Figure 4)**.

Diversité, habitats et écologie des Archaea

La majorité des organismes cellulaires peuplant la planète Terre sont des microorganismes unicellulaires : archées, bactéries ou eucaryotes (protistes). Et, même si l'impact des activités humaines sur l'environnement est notable, les cycles géochimiques sur la planète Terre reposent exclusivement sur l'activité biologique des microorganismes (pour une revue, voir (Falkowski et al., 2008)). Or, le rôle joué par les archées dans les cycles géochimiques terrestres a longtemps été sous-estimé car les premiers organismes cultivables ont été isolés à partir d'environnements inhospitaliers. Dès lors, l'épithète extrêmophile fut utilisée pour qualifier l'ensemble de ces organismes. Pourtant, les premières applications des techniques d'amplification par PCR à la détection de marqueurs moléculaires dans des échantillons environnementaux ont attestées de la présence d'archées dans des milieux plus communs (Dawson et al., 2006; DeLong, 1992; DeLong et al., 1994; Pace, 1997). Depuis, les études métagénomiques ont confirmé le caractère cosmopolite et la diversité des niches écologiques de ces organismes (Chaban et al., 2006; Dawson et al., 2006; Schleper et al., 2005). Dans leur ensemble, les études récentes en écologie microbienne suggèrent que les Archaea jouent un rôle fondamental dans les cycles biogéochimiques terrestres (Karner et al., 2001; Reysenbach

et al., 2006), notamment au niveau du cycle de l'azote (Cavicchioli et al., 2007; de la Torre et al., 2008; Hallam et al., 2006; Hatzenpichler et al., 2008; Konneke et al., 2005; Leininger et al., 2006).

Si les archées n'occupent pas exclusivement des biotopes extrêmes, il est néanmoins possible que tous les écosystèmes dans lesquels ces micro-organismes prospèrent aient une caractéristique commune. En effet, il a été proposé que les archées soient particulièrement adaptées aux environnements qui imposent un stress énergétique à la cellule, soit parce que les propriétés physicochimiques de l'écosystème (température, salinité, acidité, etc.) sont extrêmes, et donc éprouvantes, soit parce que les ressources énergétiques du milieu sont limitées (Valentine, 2007). Les archées prospéreraient dans de tels milieux parce qu'elles préserveraient mieux leur potentiel énergétique, grâce à une membrane douée d'une imperméabilité renforcée, et parce qu'elles posséderaient des voies métaboliques singulières, peu coûteuses en énergie (Valentine, 2007).

Nature du dernier ancêtre commun des Archaea

Outre le fait de révéler les relations de parenté entre les êtres vivants, les arbres phylogénétiques permettent de s'intéresser à l'évolution des organismes et d'inférer la nature probable de leur dernier ancêtre commun. Selon la nomenclature proposée par Woese et ses collaborateurs pour le domaine Archaea, le phénotype hyperthermophile représenté chez les Crenarchaeota correspondrait au phénotype ancestral (Woese et al., 1990). Par la suite, l'analyse de la séquence de l'ARNr 16S de l'archée hyperthermophile méthanogène *Methanopyrus kandleri*, qui fait apparaître cet organisme comme un représentant ancien qui se situe à la base de l'arbre des Archaea, amena Woese et ses collaborateurs à proposer que le dernier ancêtre commun des Archaea était un organisme hyperthermophile méthanogène

(Burggraf et al., 1991). Il a depuis été démontré que la position de *M. kandleri* à la base de l'arbre est vraisemblablement artefactuelle (Slesarev et al., 2002), probablement parce que cette lignée évolue rapidement (Brochier et al., 2004). Des analyses phylogénétiques des gènes impliqués dans la méthanogenèse suggèrent néanmoins que ce métabolisme est ancien (Bapteste et al., 2005), une hypothèse qui semble confortée par les analyses isotopiques de sédiments géologiques (Ueno et al., 2006). De la même façon, il a ensuite été proposé que l'archée hyperthermophile *N. equitans* soit un représentant actuel d'une forme microbienne primitive, un 'fossile vivant' (Di Giulio, 2006; Huber et al., 2002; Waters et al., 2003). Le profil du dernier ancêtre commun des Archaea correspondrait, selon cette hypothèse, à un organisme hyperthermophile possédant un génome compact avec un grand nombre de gènes scindés (par rapport à la situation observée dans les génomes des organismes actuels). D'autres analyses suggèrent au contraire que cet organisme est un Euryarchaeota, probablement affilié aux Thermococcales, qui évolue vite en raison de son style de vie parasitaire (Brochier et al., 2005) ; cette hypothèse est appuyée par des analyses génomiques comparatives (Dutilh et al., 2008; Makarova and Koonin, 2005). De manière remarquable, la notion rémanente est que le dernier ancêtre commun des Archaea était un organisme hyper-thermophile, une idée qui s'accorde bien avec l'hypothèse populaire d'une origine chaude de la vie (pour une revue récente, voir (Martin et al., 2008)). De façon intéressante, les analyses basées sur les protéines ribosomales suggèrent que les organismes mésophiles du phylum Thaumarchaeota, brancheraient à la base de l'arbre des Archaea **(Figure 4)**. Par conséquent, le dernier ancêtre commun des Archaea pourrait bien avoir été un organisme mésophile et non un hyperthermophile.

Par ailleurs, une étude génomique récente suggère que cet ancêtre était probablement aussi complexe que les organismes archéens modernes (Makarova et al., 2007). De manière remarquable, une étude comparative de la distribution des gènes codant les protéines

ribosomales dans les trois domaines cellulaires du vivant indique que des gènes ribosomaux ont été perdus au cours de l'évolution des Archaea. Aussi, cela suggère que l'ancêtre commun des Archaea pourrait même avoir été davantage complexe que les archées contemporaines en ce qui concerne certains systèmes cellulaires (voir le Chapitre IV pour une discussion sur l'évolution de la machinerie de réplication de l'ADN chez les Archaea).

La réplication de l'ADN

Vue d'ensemble historique et scientifique des travaux sur la réplication de l'ADN

L'acide désoxyribonucléique (ADN) a été identifié dans la deuxième moitié du XIX^{e} siècle par Friedrich Miescher. Ce dernier isola, à partir de cellules de pus récoltées sur des plaies infectées de patients, une molécule riche en phosphore qu'il appela 'nucléine', car extraite du noyau. Au début du XX^{e} siècle, la nature nucléique du chromosome fut établie et la nucléine rebaptisée acide désoxyribonucléique. Néanmoins, les biologistes accordaient toujours peu d'importance à cette molécule et lui préféraient les protéines comme candidat au support de l'hérédité. Il fallut attendre plus de soixante-dix ans à partir de la découverte de Miescher pour qu'émergent les premières indications scientifiques sur le rôle de l'ADN dans l'hérédité.

L'introduction des techniques physiques en biologie (radiomarquage, centrifugation) va jouer un rôle décisif dans la détermination de la nature chimique du gène. Les premières indications viennent des travaux d'Avery, MacLeod et McCarty qui cherchent à identifier la nature du principe transformant de la maladie du pneumocoque, mais elles ne convainquent pas la communauté scientifique qui accorde toujours du crédit à la théorie d'un support protéique de l'hérédité (Avery et al., 1944). Il fallut attendre presque une décennie supplémentaire pour que l'ADN soit reconnu comme le support de l'hérédité après que Hershey et Chase aient montré la fonction respective des protéines et des acides nucléiques dans la croissance d'un bactériophage en procédant à un marquage différencié de chacune des deux espèces moléculaires (Hershey and Chase, 1952). Un an plus tard, James Watson et Francis Crick proposaient le modèle de la structure en double-hélice de l'ADN puis

énonçaient les implications génétiques de ce modèle structural (Watson and Crick, 1953a, b). Dans le second article, les auteurs s'interrogeaient notamment sur le mode de réplication de l'ADN — qu'ils envisageaient avec préscience comme procédant selon un mécanisme semi-réplicatif —, sur l'intervention ou non d'une enzyme dans le processus de réplication, sur le rôle joué par les protéines, sur les contraintes topologiques inhérentes à l'enroulement en double hélice (Watson and Crick, 1953a).

L'utilisation de précurseurs nucléotidiques radiomarqués à l'étude du mécanisme de la réplication permit d'en déterminer les caractéristiques principales : la réplication de l'ADN est un processus semi-conservatif (Meselson and Stahl, 1958) ; la réplication du chromosome débute au niveau d'un point précis appelé origine puis se propage selon un mouvement bidirectionnel, les fourches de réplication s'éloignant l'une de l'autre dans des directions opposées (Cairns, 1963) ; la réplication de l'ADN est semi-discontinue, l'un des deux brins étant synthétisé par la ligature de courts segments polynucléotidiques appelés fragments d'Okazaki (Okazaki et al., 1968).

Par la suite, de nombreuses expériences biochimiques et génétiques permirent d'identifier progressivement l'ensemble des protéines impliquées dans le processus de la réplication de l'ADN chez les bactéries et les eucaryotes. Des systèmes de réplication de l'ADN *in vitro* furent également mis au point à partir de protéines purifiées afin de délimiter avec davantage de précision le spectre d'action des différents facteurs et d'explorer leurs interactions fonctionnelles. Par ailleurs, le recours à des modèles viraux a été extrêmement profitable à la détermination du mécanisme de base de la réplication, même si ces systèmes ne permettaient pas de récapituler l'ensemble des évènements qui se déroule dans un contexte chromosomique cellulaire. Enfin, les approches génétiques ont été décisives pour ordonner la séquence des évènements qui régissent l'initiation de la réplication de l'ADN chez les eucaryotes. Par conséquent, l'étude de la réplication de l'ADN chez les bactéries et les

eucaryotes et la caractérisation des différents acteurs moléculaires ont essentiellement reposées sur des approches biochimiques et génétiques.

A l'inverse, le séquençage précoce du génome de l'archée *M. jannaschii* et l'avènement des méthodes d'analyse comparative des génomes ont influencé les premières années d'étude sur la réplication de l'ADN chez les Archaea (Bult et al., 1996; Olsen and Woese, 1996). En particulier, la génomique comparative a vite permis de souligner la ressemblance entre la machinerie de réplication des archées et celle des eucaryotes. Aussi, la majorité des protéines impliquées dans la réplication de l'ADN chez les Archaea ont été identifiées par homologie avec les protéines eucaryotes (pour une revue, voir (Edgell and Doolittle, 1997)). Les seules exceptions sont l'ADN polymérase D (PolD) (Cann et al., 1998; Imamura et al., 1995; Uemori et al., 1997) et l'ADN topoisomérase VI (Topo VI) (Bergerat et al., 1997; Bergerat et al., 1994), identifiées par des approches biochimiques, et la protéine initiatrice WhiP, dont le gène a été identifié car il occupe une position conservée à proximité de plusieurs origines de réplication de Crenarchaeota (Robinson and Bell, 2007). De la même façon, l'analyse du biais nucléotidique par des approches *in silico* a permis de prédire la position des origines de réplication dans les génomes d'Archaea avant que la première origine de réplication ne soit identifiée expérimentalement chez *Pyrococcus abyssi* (Lopez et al., 1999; Matsunaga et al., 2001; Myllykallio et al., 2000). Par la suite, les méthodes biochimiques et génétiques prirent le pas sur les approches de génomique comparative et une véritable communauté scientifique s'organisa autour de l'étude de la réplication chez les Archaea, en particulier parce que les Archaea recèlent des clefs importantes concernant l'évolution de la machinerie de réplication des eucaryotes. En outre, l'apparente simplicité du modèle archéen par rapport au modèle eucaryote contribua à ce que les archées s'imposent comme des modèles alternatifs à l'étude de la réplication chez les eucaryotes (voir, (MacNeill, 2001; Tye, 2000; Vas and Leatherwood, 2000)). L'identification de trois origines de réplication chez *Sulfolobus*

solfataricus et *Sulfolobus acidocaldarius* a renforcé l'intérêt du modèle archéen pour l'étude de la réplication chez les eucaryotes car ces organismes modèles permettent d'aborder les questions relatives aux modalités de la réplication de l'ADN en présence de plusieurs origines (Lundgren et al., 2004; Robinson et al., 2004).

Mécanisme de la réplication de l'ADN : principes généraux

Les machineries de réplication diffèrent d'un domaine cellulaire à l'autre, même si un certain nombre de protéines présentent des liens évolutifs (voir ci-après). En revanche, le processus de la réplication de l'ADN dans les trois domaines cellulaires du vivant repose sur les mêmes principes fondamentaux, principes qui découlent de la structure de la double-hélice d'ADN et des propriétés des enzymes qui en assurent la duplication. Bien que la réplication de l'ADN soit un processus continu son déroulement est rythmé par trois étapes majeures : l'initiation, l'élongation et la terminaison.

Initiation

L'initiation consiste en une déstabilisation localisée de la double-hélice au niveau d'une région particulière appelée origine afin de permettre l'assemblage de la machinerie de réplication. Cette modification transitoire de la structure de l'ADN est déclenchée par la fixation d'une ou de plusieurs protéines initiatrices. Ce remodelage de l'ADN ouvre la voie au chargement de l'hélicase, une enzyme dont l'activité a pour effet de séparer les deux brins de la double-hélice d'ADN pour les rendre accessibles. L'exposition temporaire de l'ADN sous forme simple brin, dont l'intégrité est préservée par la mobilisation d'une protéine spécialisée, permet aux différentes protéines impliquées dans la synthèse de l'ADN, ou réplisome, de prendre place sur la matrice et d'y bâtir deux fourches de réplication **(Figure 5)**.

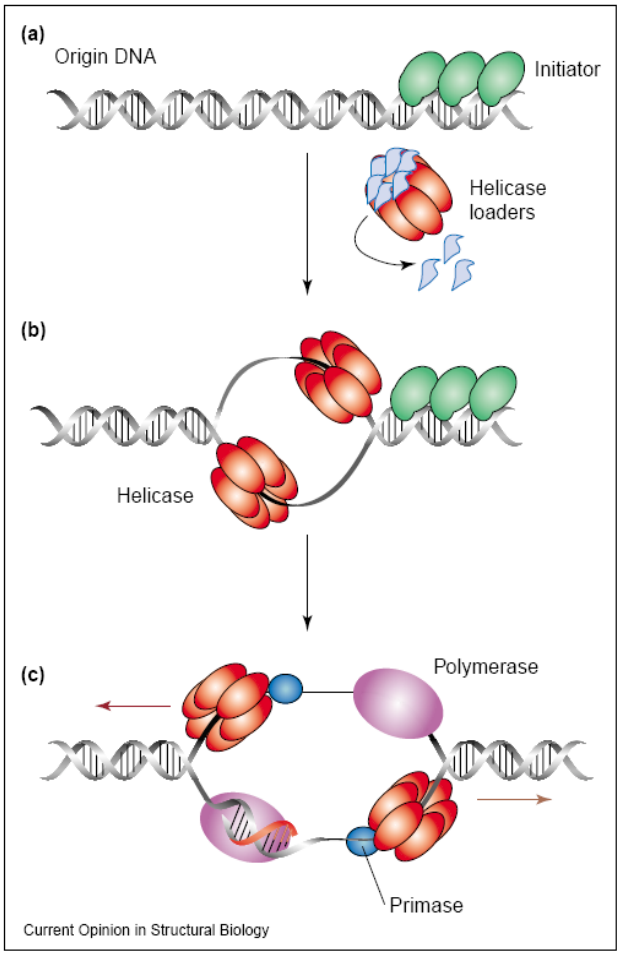

Figure 5: Vue générale de l'initiation de la réplication. (a) La ou les protéines initiatrices se fixent à l'origine de réplication puis remodèlent l'ADN et recrutent d'autres facteurs, dont l'hélicase. (b) L'hélicase ouvre davantage l'ADN et expose l'ADN simple-brin, dont l'intégrité est protégée par la mobilisation d'une protéine spécifique. (c) D'autres facteurs de réplication (ADN primase, ADN polymérase,) se fixent alors au niveau de l'ADN pour y bâtir deux fourches de réplication. Une fois que le facteur de processivité et son facteur de chargement sont présents, la machinerie de réplication quitte la phase d'initiation pour rentrer dans la phase d'élongation. Le schéma présenté correspond à un modèle bactérien mais les étapes décrites sont également valables dans un contexte archéen ou eucaryote ; seuls la nature et le nombre de facteurs varient selon les modèles.
D'après (Cunningham and Berger, 2005).

Elongation

L'élongation correspond à la synthèse de fragments nucléotidiques par la machinerie de réplication selon un mouvement coordonné, depuis l'élaboration des fourches de réplication jusqu'à leur démantèlement. Les deux fourches de réplication vont progresser dans des directions opposées à partir du point d'initiation de la réplication. Chacune des deux fourches synthétise simultanément deux hémi-molécules en se servant des deux matrices parentales pour produire les brins complémentaires. Le déplacement de la fourche de réplication au travers de la double-hélice d'ADN est rendu possible par l'action combinée de deux enzymes. D'une part, l'hélicase sépare progressivement les deux brins de l'ADN en aval de la machinerie de réplication afin que les deux matrices soient continuellement accessibles. D'autre part, les ADN topoisomérases relâchent constamment les modifications topologiques

de l'ADN induites par le déplacement du réplisome afin de préserver les conditions favorables à l'ouverture de la molécule d'ADN par l'hélicase. Les protéines qui assemblent les désoxyribonucléotides, les ADN polymérases, présentent un certain nombre de particularités qui influencent la dynamique de la fourche de réplication. Premièrement, ces enzymes sont incapables d'initier une réaction de polymérisation *de novo* et sont entièrement dépendantes de la présence d'une amorce d'ARN ou d'ADN. Ce court fragment d'ARN est synthétisé par une ARN polymérase spécialisée appelée ADN primase. Deuxièmement, les ADN polymérases ont besoin d'interagir avec une protéine auxiliaire pour rester au contact de l'ADN, faute de quoi elles se comportent de manière distributive, c'est-à-dire qu'elles alternent entre des phases d'activité et d'inactivité. Ce facteur auxiliaire, appelé facteur de processivité, se présente sous la forme d'un anneau capable d'encercler la molécule d'ADN et de maintenir l'ADN polymérase au contact de sa matrice. Cet anneau est ouvert puis refermé autour d'une molécule d'ADN grâce à une protéine de chargement dédiée. Troisièmement, les ADN polymérases allongent exclusivement les chaînes polynucléotidiques depuis l'extrémité 5' vers l'extrémité 3' car le métabolisme des nucléotides produit des désoxynucléotides 5' triphosphates. Compte tenu de l'arrangement antiparallèle des deux chaînes au sein de la molécule d'ADN, l'un des brins est synthétisé de manière continue et l'autre de manière discontinue **(Figure 6)**. Concrètement, l'ADN polymérase qui synthétise le brin continu glisse le long de sa matrice au fur et à mesure de la dénaturation de la double-hélice par l'hélicase réplicative. Au contraire, le brin discontinu est synthétisé par l'intermédiaire de courts segments, appelés fragments d'Okazaki, initiés à intervalles réguliers par l'ADN primase, qui dépose sur ce brin des amorces ARN au fur et à mesure de l'ouverture de la molécule d'ADN par l'hélicase réplicative. Finalement, les amorces ARN sont éliminées afin d'obtenir une molécule homogène avant qu'une enzyme spécialisée, appelée ADN ligase, raboute les fragments d'Okazaki les uns aux autres pour obtenir une molécule d'ADN continue. Dans

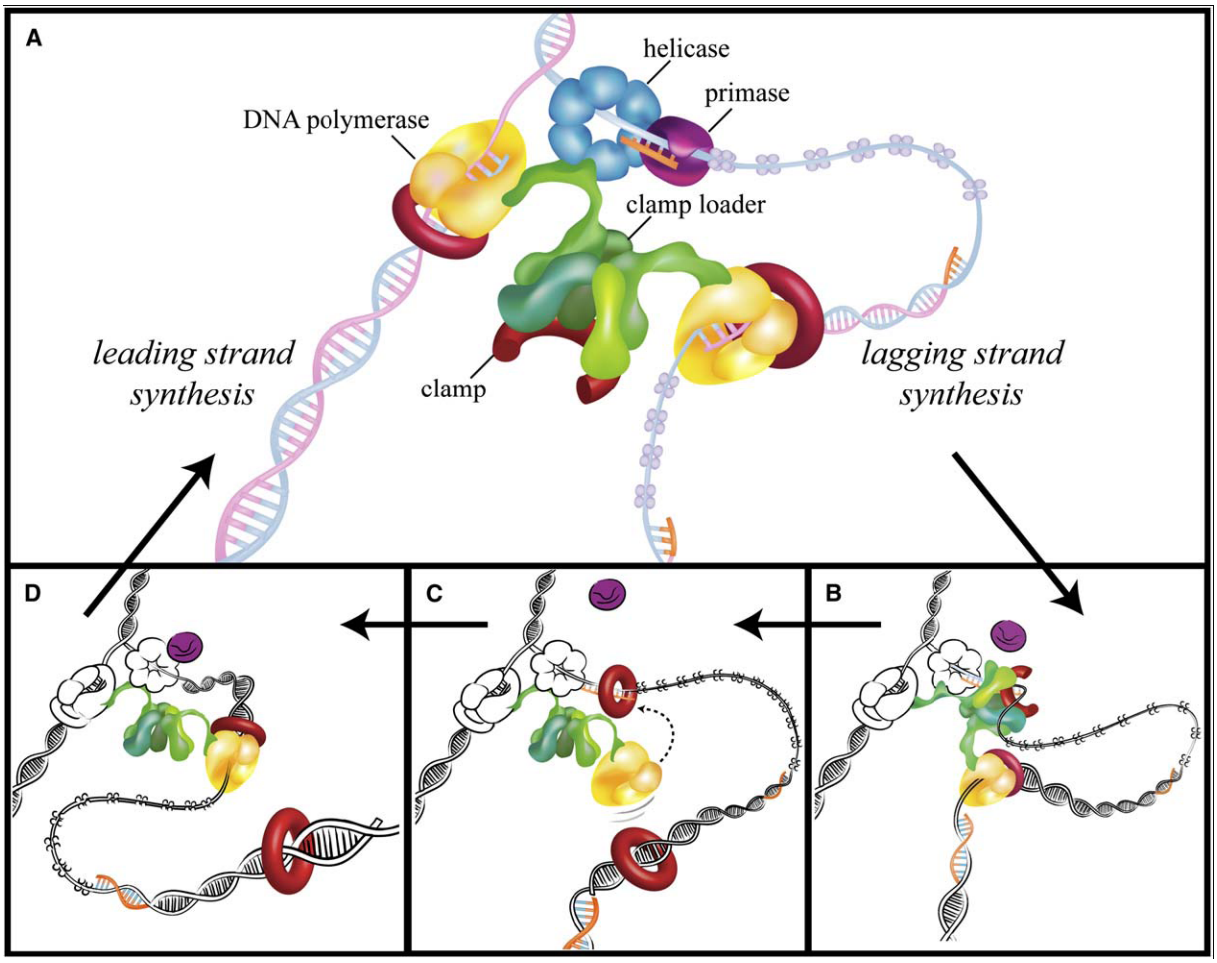

Figure 6 : Le cycle de synthèse du brin discontinu. (A) Tandis que l'ADN polymérase du brin discontinu synthétise un fragment d'Okazaki, un nouveau facteur de processivité est ouvert et la primase initie la synthèse du fragment suivant en déposant une amorce ARN au pied de la fourche de réplication. (B) Après la synthèse de l'amorce ARN, le facteur de processivité est chargé au niveau de la jonction entre l'amorce et la matrice ADN simple brin, par ailleurs recouverte par une protéine spécialisée. (C) L'achèvement de la synthèse du fragment d'Okazaki provoque le recyclage de l'ADN polymérase du brin discontinu au niveau du facteur de processivité situé sur la nouvelle amorce. Le facteur de processivité est abandonné à proximité de l'intervalle entre deux fragments d'Okazaki. (D) L'ADN polymérase du brin discontinu synthétise un nouveau fragment d'Okazaki. L'ouverture de la double-hélice et la synthèse du brin continu se poursuit tout au long du cycle. Le schéma illustre le modèle établi à partir des études menées chez *Escherichia coli*.
D'après (Langston and O'Donnell, 2006).

des conditions favorables, la fourche de réplication continue sa progression jusqu'à rencontrer une autre fourche, un évènement qui provoque leur démantèlement.

Terminaison

La terminaison correspond au démantèlement des fourches de réplication et à la résolution des liens topologiques unissant les deux molécules d'ADN néosynthétisées avant leur séparation en deux molécules filles. La résolution des liens topologiques requiert l'intervention d'ADN topoisomérases. Le mécanisme de la terminaison est assez bien connu chez les bactéries, mal connu chez les eucaryotes et totalement inconnu chez les archées.

Mécanisme de la réplication de l'ADN chez les bactéries

Initiation de la réplication : mécanismes et régulation [1]

Le chromosome bactérien contient une seule origine de réplication appelée *oriC*. Chez la bactérie modèle *Escherichia coli*, l'origine de réplication correspond à une région d'environ 250 paires de bases au niveau de laquelle se trouvent des sites de fixation spécifiques pour la protéine initiatrice de la réplication, DnaA. La protéine DnaA est une ATPase de la famille AAA^+ (ATPase that are associated with various cellular activities). La protéine DnaA se fixe à l'origine au niveau de sites à haute affinité (DnaA boxes) tout au long du cycle cellulaire. L'initiation de la réplication est déclenchée par la fixation, en présence d'ATP, de molécules DnaA supplémentaires au niveau de sites de faible affinité (ATP-DnaA boxes) distribués au niveau de l'origine. La forme DnaA liée à l'ATP (DnaA-ATP) forme, selon un mode coopératif, une structure homo-oligomérique capable d'étirer la molécule d'ADN en un filament. Ce remodelage provoque l'ouverture de la molécule au niveau de régions riches en

[1] D'après Mott, M.L., and Berger, J.M. (2007). DNA replication initiation: mechanisms and regulation in bacteria. Nature reviews *5*, 343-354.

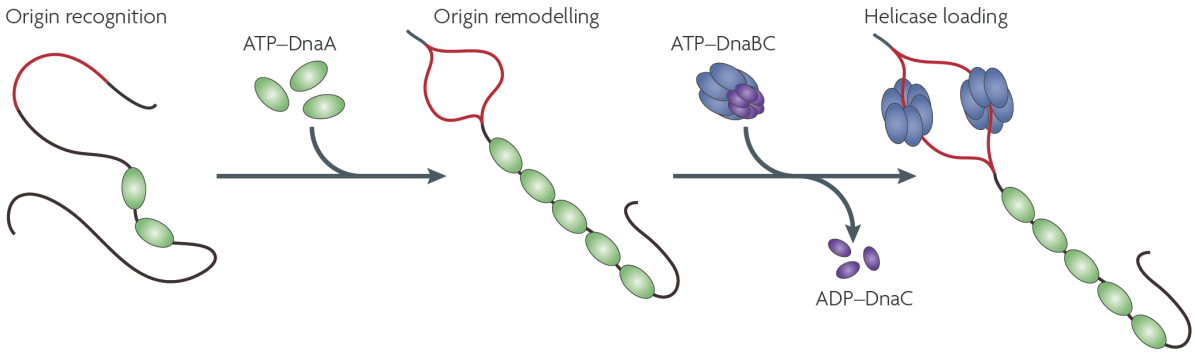

Figure 7: Remodelage de l'origine par la protéine DnaA et chargement de l'hélicase chez la bactérie modèle *Escherichia coli*. La protéine DnaA, représentée en vert, est associée à plusieurs sites de fixation à haute affinité tout au long du cycle cellulaire. Au moment de l'initiation, des molécules supplémentaires de DnaA liées à l'ATP se fixent à l'origine, au niveau de laquelle elles s'assemblent pour former un complexe nucléoprotéique. Cela a pour effet de faciliter l'ouverture d'une région adjacente riche en AT (en rouge). Une fois cette région dénaturée, le chargement de l'hélicase DnaB (en bleu) intervient sous le contrôle de la protéine DnaC (en violet).
D'après (Mott and Berger, 2007).

AT appelées DUE (DNA unwinding elements). Ensuite, la protéine DnaA recrute l'hélicase DnaB avec l'assistance de la protéine DnaC, une ATPase apparentée à la molécule DnaA (les protéines DnaA et DnaC sont paralogues). A l'image du déclenchement de l'initiation, l'étape de chargement de l'hélicase DnaB par DnaC est régulée par la présence d'ATP **(Figure 7)**. Les stratégies de régulation de l'initiation de la réplication comprennent : i) la séquestration de l'origine ; ii) la titration de DnaA ; iii) l'inhibition de la transcription du gène *dnaA* ; iv) l'inactivation de la protéine DnaA. Le processus d'inactivation, qui consiste à promouvoir la formation de la forme DnaA-ADP, dépend de l'interaction entre le facteur de processivité de l'ADN polymérase, DnaN ou β-clamp, et la protéine Hda. Outre le fait de prévenir les évènements de re-réplication, ce processus pourrait marquer la transition entre les phases d'initiation et d'élongation de la réplication.

La protéine initiatrice de la réplication DnaA et l'hélicase réplicative DnaB (appelée DnaC chez *Bacillus subtilus*) sont universellement conservées chez les bactéries. En revanche, le mode de régulation du déclenchement de l'initiation chez la bactérie *Caulobacter crescentus* diffère notablement par rapport à *E. coli*, dans la mesure où la fixation de la protéine DnaA au niveau de l'origine est contrôlée par la protéine régulatrice CtrA. De façon

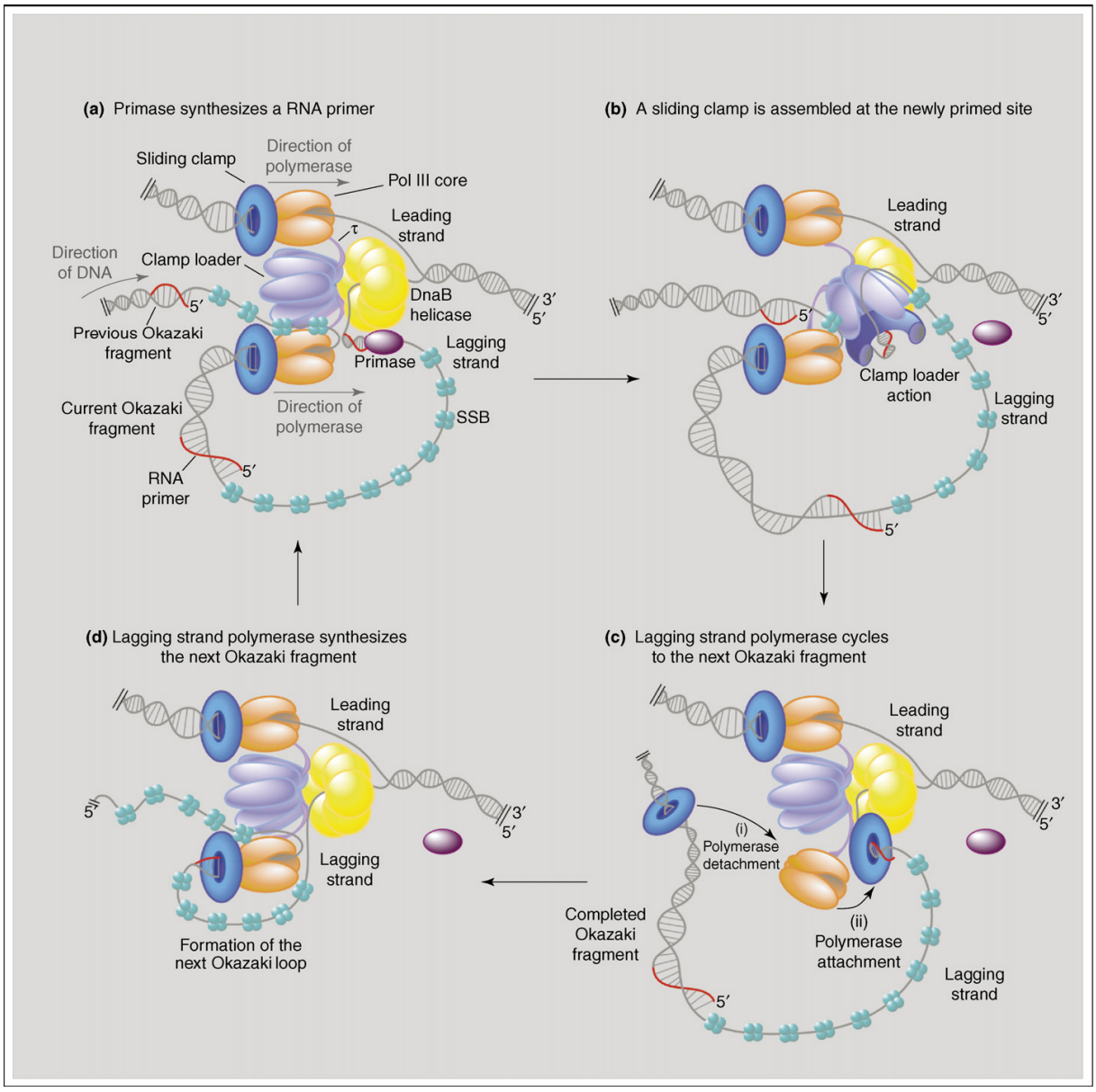

Figure 8 : Dynamique de la fourche de réplication chez *Escherichia coli*. La synthèse du brin continue s'effectue d'un seul tenant et ne requiert qu'un ou quelques évènements d'amorçage. A l'inverse, le brin discontinu est synthétisé sous la forme de multiples segments d'ADN appelés fragments d'Okazaki initiés au fur et à mesure de l'ouverture du duplex par l'hélicase réplicative (en jaune). (a) Un fragment d'Okazaki est initié par la synthèse d'une courte amorce ARN (en rouge) par l'ADN primase. Une fois l'amorce achevée, l'ADN primase se dissocie de la matrice ADN. (b) Le complexe γ (en violet) exploite l'énergie de l'ATP pour ouvrir et assembler le facteur de processivité β (en bleu) au niveau de l'amorce ARN. (c) L'ADN polymérase qui procède à la synthèse du brin discontinu se dissocie de l'anneau β après avoir achevé un fragment d'Okazaki et s'ancre à un nouvel anneau beta assemblé au niveau de l'amorce d'un fragment naissant. L'ancien anneau β demeure en place. (d) L'ADN polymérase du brin discontinue débute la synthèse d'un nouveau fragment d'Okazaki. D'après (Pomerantz and O'Donnell, 2007)

intéressante, la protéine impliquée dans le chargement de l'hélicase réplicative chez *E. coli* n'est pas non plus conservée chez les autres bactéries modèles. Chez *B. subtilis* le chargement de l'hélicase fait intervenir les protéines DnaB, DnaD et DnaI (la protéine DnaB de *B. subtilis* ne correspond pas à la protéine DnaB trouvée chez *E. coli*). En outre, le mode d'action des facteurs assistant le chargement de l'hélicase chez *B. subtilis* est différent de celui adopté par le couple DnaA/DnaC pour charger DnaB chez *E. coli* (Davey and O'Donnell, 2003). Chez *C. crescentus*, aucun facteur de chargement de l'hélicase n'a, à ce jour, été identifié. Une hypothèse séduisante est que le chargement de l'hélicase repose uniquement sur DnaA chez *C. crescentus* sachant que la présence de la protéine initiatrice au niveau de l'origine est finement régulée par CtrA. Aussi, il semble que les protéines impliquées dans l'initiation de la réplication diffèrent selon les organismes bactériens considérés.

Dynamique de la fourche de réplication [2]

Une fois que la protéine DnaC a chargé l'hélicase DnaB sur l'ADN, celle-ci devient une hélicase réplicative mature. Durant la phase d'élongation, DnaB dénature la molécule d'ADN et libère progressivement la matrice simple brin nécessaire à la synthèse des brins continu et discontinu **(Figure 8)**. Cette synthèse requiert l'intervention de deux ADN polymérases, une pour chaque brin, réunies au sein d'un complexe qui se caractérise par une asymétrie fonctionnelle (Glover and McHenry, 2001). En effet, le brin continu est synthétisé a priori sans interruption à partir de l'amorce initiale produite par l'ADN primase **(Tableau 1)**. En revanche, la synthèse du brin discontinu nécessite que l'ADN primase initie la synthèse de fragments de manière régulière, au fur et à mesure de l'ouverture du duplex ADN. Une fois la synthèse de l'amorce ARN achevée, l'ADN primase est déplacée de la matrice par la sous-unité χ du complexe γ, qui charge le facteur de processivité β au niveau de l'extrémité 3'OH

[2] D'après Pomerantz, R.T., and O'Donnell, M. (2007). Replisome mechanics: insights into a twin DNA polymerase machine. Trends in microbiology *15*, 156-164.

Tableau 1 : Protéines de la réplication de l'ADN chez les bactéries

Initiation	
Protéine initiatrice	DnaA
Hélicase	DnaB [1]
Chargeur de l'hélicase	DnaC [2]
Elongation	
Hélicase	DnaB (5′ vers 3′)
ADN primase	DnaG
Protéine fixant l'ADN simple brin	SSB
Topoisomérases	Topo IV ADN gyrase
Polymérases réplicatives	Pol III (Famille C) PolC [3] (Famille C)
Facteur de processivité	Sous-unité β
Protéine chargeant le facteur de processivité	Complexe γ
Maturation des fragments d'Okazaki	
Polymérase	Pol I
Nucléase ARN	Pol I (activité exonucléase 5′-3′) RNase HII ?
Nucléase ADN	Pol I (activité exonucléase 5′-3′)
ADN ligase	NAD-ligase

[1] Nomenclature *E. coli*
[2] La nature du chargeur de l'hélicase est divergente selon les phylums (voir texte)
[3] La PolC est présente dans certains phylums bactériens

de l'amorce ARN (Yuzhakov et al., 1999). Puis, l'ADN polymérase III (Pol III) s'ancre à l'ADN via le facteur β **(Tableau 1)**. Le complexe protéique formé par la réunion entre la Pol III, le facteur β et le complexe γ constitue l'holoenzyme Pol III, la réplicase d'*E. coli*, une enzyme capable de synthétiser des fragments d'ADN de manière processive. Lorsque la Pol III officiant sur le brin discontinu entre en contact avec l'amorce ARN du fragment d'Okazaki précédent, le complexe γ transfère l'ADN polymérase à un autre anneau β assemblé au niveau d'une nouvelle amorce ARN ; l'ancien anneau β demeure à l'extrémité du fragment d'Okazaki qui vient d'être achevé (Pomerantz and O'Donnell, 2007). La fréquence d'initiation des fragments d'Okazaki sur le brin discontinu est régulée via l'interaction fonctionnelle distributive entre les protéines DnaB et DnaG (Corn and Berger, 2006). Les amorces ARN de chacun des fragments d'Okazaki sont ensuite remplacées par de l'ADN avant que l'ADN

ligase n'intervienne pour rabouter les différents segments (Kornberg and Baker, 1992). Les anneaux β abandonnés au cours du cycle de synthèse du brin discontinu marquent la position des amorces d'ARN et servent probablement de points d'ancrage aux protéines impliquées dans le processus de maturation. L'élimination des amorces ARN et leur remplacement par de l'ADN est assurée par l'ADN Pol I selon une activité de déplacement de coupure, c'est-à-dire que l'ADN Pol I dégrade l'ARN grâce à son activité exonucléase 5' vers 3' tout en allongeant l'extrémité 3'OH du fragment d'Okazaki adjacent (Kornberg and Baker, 1992). La RNase HII pourrait également avoir un rôle à jouer dans le processus de maturation des fragments d'Okazaki (Fukushima et al., 2007).

De manière intéressante, la bactérie *Bacillus subtilis* et les organismes qui lui sont apparentés utilisent deux ADN polymérases de la famille C pour répliquer leur génome (Dervyn et al., 2001), selon un dispositif qui s'apparente à celui rencontré chez les eucaryotes (voir ci-après). Chez *B. subtilis*, la synthèse des brins continu et discontinu est probablement répartie entre les PolC et Pol III respectivement (Dervyn et al., 2001). Par ailleurs, certains génomes bactériens, dont celui de *B. subtilis*, codent une protéine présentant un domaine exonucléase 5'-3'. Cette protéine, nommée YpcP, pourrait assister ou se substituer à la Pol I au cours de la maturation des fragments d'Okazaki (Fukushima et al., 2007).

Mécanisme de la réplication de l'ADN chez les eucaryotes [3]

L'étude de la réplication du virus eucaryote SV40 a été extrêmement utile pour établir le mécanisme de base de la réplication de l'ADN chez les eucaryotes, en particulier le processus de recrutement des ADN polymérases et l'initiation de la synthèse de l'ADN (Waga and Stillman, 1998). Il apparut néanmoins assez rapidement que ce système ne reflétait pas

l'ensemble des évènements se déroulant au niveau de la fourche de réplication recopiant le chromosome cellulaire eucaryote. L'identification et la caractérisation partielle de la plupart des autres facteurs de réplication nécessaires à l'assemblage du complexe de pré-réplication sont le fruit d'expériences menées chez la levure *Saccharomyces cerevisiae* **(Tableau 2)**. Le modèle présenté par la suite se base sur les données recueillies chez la levure ; certaines différences entre modèles seront signalées. En outre, certains facteurs de réplication identifiés chez la levure n'ont pas d'orthologues apparents chez les organismes eucaryotes supérieurs. Pour une représentation illustrée du modèle d'initiation de la réplication chez la levure voir http://www.dnareplication.net/ (Cotterill and Kearsey, 2008).

Définition de l'origine [4]

Chez les eucaryotes, l'origine est reconnue par le complexe ORC (Origin Recognition Complex), une structure formée par l'assemblage de six sous-unités (Orc1-6), dont trois (Orc1, Orc4 et Orc5) sont membres de la famille des ATPases AAA^+ (Bell and Stillman, 1992). Le chargement du complexe ORC sur l'ADN dépend de la fixation d'ATP par la sous-unité Orc1 (Bell and Stillman, 1992). La spécificité de séquence du site de fixation du complexe ORC n'a pu être déterminée que chez les organismes eucaryotes unicellulaires (e.g., les levures), en particulier chez la levure de boulanger *Saccharomyces cerevisiae*, mais ces éléments ne sont pas conservés de manière stricte. En fait, la caractéristique la plus fondamentale de ces origines semble être la richesse en nucléotides AT. Chez les métazoaires, les origines de réplication n'arborent aucune caractéristique évidente en dehors d'une richesse en nucléotides AT et l'utilisation ou non d'une origine de réplication semble davantage

[3] D'après Kelly, T.J., and Stillman, B. (2006). Duplication of DNA in Eukaryotic Cells. In DNA replication and human disease, M.L. DePamphilis, ed. (Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory Press), pp. 1-29.
[4] D'après Aladjem, M.I., Falaschi, A., and Kowalski, D. Ibid. Eukaryotic DNA Replication Origins. pp. 31-61.

Tableau 2 : Protéines de la réplication de l'ADN chez les eucaryotes

Initiation	
Protéine initiatrice	ORC1-6 Cdc6
Hélicase	MCM2-7
Chargeur de l'hélicase	Cdt1
Assemblage de la fourche	GINS (Sld5, Psf1, Psf2, Psf3) Mcm10 Cdc45 Sld2 Sld3 Dpb11
Kinases	Cdc7-Dbf4 CDK-Cyclin
Elongation	
Synthèse ADN	
Hélicase	MCM2-7 (3' vers 5')
ADN primase	Complexe Pol α-primase PriSL
Protéine fixant l'ADN simple brin	RPA
Topoisomérases	Topo IB Topo II
Polymérases réplicatives	Pol δ (Famille B) Pol ε (Famille B)
Facteur de processivité	PCNA
Protéine chargeant le facteur de processivité	RFC
Maturation des fragments d'Okazaki	
Polymérase	Pol δ
Nucléase ARN	RNase HII ?
Endonucléase ADN	FEN-1 Dna2
Hélicase	Dna2 Pif1
ADN ligase	ATP-ligase

dépendre du contexte chromosomique. En particulier, une étude récente suggère que, chez les mammifères, le choix du site d'initiation est déterminé par l'organisation structurale de la chromatine, laquelle est influencée par la vitesse de la fourche au cours de l'évènement de réplication du cycle cellulaire précédent (Courbet et al., 2008).

Assemblage du complexe de pré-réplication (chargement de l'hélicase réplicative) [5]

Après s'être fixé à l'ADN, le complexe ORC recrute la protéine Cdc6 puis la protéine Cdt1, lesquelles assistent le chargement d'une forme inactive de l'hélicase réplicative, le complexe MCM2-7. Cette séquence d'évènements, appelée assemblage du complexe de pré-réplication, est finement régulée par la fixation et l'hydrolyse de l'ATP (Randell et al., 2006). Dans un premier temps, le complexe ORC-ATP recrute à l'origine la protéine Cdc6 liée à une molécule d'ATP, une étape indépendante de l'hydrolyse d'ATP (Randell et al., 2006). Dans un second temps, le complexe ORC-ATP et la protéine Cdc6-ATP recrutent, avec l'aide de la protéine Cdt1, le complexe MCM2-7 au niveau de l'origine (Randell et al., 2006). A ce stade, le complexe MCM2-7 n'est pas associé de façon stable à l'ADN ; il n'est donc pas, à proprement parler, chargé sur l'ADN. Dans un troisième temps, la protéine Cdc6 hydrolyse sa molécule d'ATP sous l'influence conjointe du complexe ORC-ATP et de l'ADN (Randell et al., 2006). Cet évènement stimule le chargement du complexe MCM2-7 et provoque la dissociation des protéines Cdt1 et Cdc6 (Randell et al., 2006). Finalement, la fixation du complexe MCM2-7 stimule l'hydrolyse de l'ATP par le complexe ORC.

Chez les métazoaires, d'autres protéines participent à l'assemblage du complexe pré-réplicatif. En particulier, la protéine geminin restreint la période d'activité de la protéine Cdt1 (Wohlschlegel et al., 2000) et procure un niveau de contrôle supplémentaire (voir plus bas). D'autre part, il a été proposé que la protéine MCM8 humaine assiste la fixation de la protéine Cdc6 à la chromatine (Volkening and Hoffmann, 2005). Néanmoins, des résultats obtenus chez le xénope suggèrent au contraire que MCM8 participe à la phase d'élongation en tant qu'hélicase (Maiorano et al., 2005). En outre, MCM8 participerait au recrutement d'une sous-unité RPA et de l'ADN polymérase α, c'est-à-dire à l'établissement de la fourche de réplication (Maiorano et al., 2005). Enfin, des résultats récents suggèrent que la protéine

[5] D'après Sivaprasad, U., Dutta, A., and Bell, S.P. Ibid. Assembly of Pre-replication Complexes. pp. 63-88.

MCM9 assiste Cdt1 dans le processus de chargement de MCM2-7 et contrebalance l'effet inhibiteur de la protéine geminin (Lutzmann and Mechali, 2008).

Régulation de l'assemblage et de l'activation du complexe de pré-réplication [6]

La transition entre la phase d'assemblage et la phase d'activation du complexe de pré-réplication se trouve sous le contrôle de deux kinases (DDK, Dbf4 dependent kinase ; CDK, cyclin dependent kinase), dont le niveau d'activité fluctue au cours du cycle cellulaire. Au moment de la phase G_1, l'activité kinase est faible ce qui est propice à l'assemblage du complexe de pré-réplication. La concentration des sous-unités régulatrices des kinases (Dbf4 et cycline) augmentent progressivement jusqu'au début de la phase S. Au-delà d'un certain seuil d'activité kinase, l'assemblage du complexe de pré-réplication est défavorisé mais l'activation des complexes existants est favorisée. Ce découplage temporel entre la phase d'assemblage du complexe de pré-initiation et la phase d'activation du complexe de pré-initiation constitue la barrière principale aux évènements de re-réplication.

L'assemblage du complexe de pré-initiation est défavorisé en présence d'une forte activité kinase car CDK cible l'ensemble des protéines initiatrices de la réplication. Les conséquences de cette phosphorylation diffèrent selon les organismes eucaryotes considérés (pour des revues, voir (Kearsey and Cotterill, 2003; Machida et al., 2005)). La régulation de l'activité de la protéine Cdt1 semble privilégiée, en particulier chez les métazoaires chez lesquels l'activité de Cdt1 est contrôlée d'une part par un inhibiteur (la protéine geminin) (Wohlschlegel et al., 2000), d'autre part par un facteur de stimulation (la protéine MCM9) (Lutzmann and Mechali, 2008).

[6] D'après Ibid.

Activation du complexe de pré-réplication et assemblage du complexe de pré-initiation [7]

L'activation de l'hélicase requiert l'activité de protéines kinases DDK et CDK et l'association interdépendante d'au moins une douzaine de facteurs de réplication supplémentaires en une structure appelée complexe de pré-initiation. L'activité des protéines kinases DDK et CDK augmentent progressivement au cours de la phase G_1 jusqu'à atteindre un seuil suffisant pour déclencher la phosphorylation d'un certain nombre de facteurs de réplication. Ces modifications ont deux effets complémentaires : d'une part elles inactivent les protéines initiatrices de la réplication — prévenant l'assemblage de nouveaux complexes de pré-réplication —, d'autre part elles activent les complexes de pré-réplication préalablement assemblés. En particulier, les résultats obtenus chez la levure indiquent que les phosphorylations catalysées par DDK (Jiang et al., 1999; Sheu and Stillman, 2006; Yabuuchi et al., 2006) et CDK (Tanaka et al., 2007; Yabuuchi et al., 2006; Zegerman and Diffley, 2007) initient un processus aboutissant à l'assemblage du complexe de pré-initiation qui comprend dans son ensemble : Mcm10, Cdc45, Sld3, GINS (Sld5, Psf1, Psf2 et Psf3), Sld2, Dpb11, RPA (RPA70, RPA32, RPA14) ; chez la levure *S. cerevisiae*, l'ADN polymérase ε joue un rôle structurant au moment de l'initiation. Une fois le complexe de pré-initiation assemblé, le complexe MCM2-7 devient une hélicase mâture, capable de dénaturer le duplex et d'exposer la matrice ADN simple brin nécessaire au chargement des enzymes catalysant la synthèse d'ADN. Une fois l'hélicase activée, certains des facteurs ayant participé à l'assemblage du complexe de pré-initiation se dissocient de la chromatine.

L'ordre d'intervention des kinases DDK et CDK, la séquence d'assemblage et la nature des interactions protéiques diffèrent probablement d'un modèle eucaryote à un autre. En particulier, l'ADN polymérase ε est nécessaire à l'activation du complexe de pré-réplication

[7] D'après Walter, J.C., and Araki, H. Ibid. Activation of Pre-replication Complexes. pp. 89-104.

chez la levure *S. cerevisiae* alors qu'elle est vraisemblablement recrutée après la dénaturation du duplex ADN chez le xénope.

Dénaturation de l'ADN et assemblage de la fourche de réplication [8]

L'étude de la réplication de l'ADN du virus eucaryote SV40 a permis d'élaborer un modèle rendant compte de la séquence des évènements qui suit la dénaturation de l'ADN par l'hélicase et marque la transition de la phase d'initiation à la phase d'élongation (Waga and Stillman, 1994). Chez le virus SV40, la reconnaissance de l'origine, le remodelage et la dénaturation du duplex ADN sont accomplies par une seule et même protéine, la protéine Tag (T antigen) (Waga and Stillman, 1994). Par contre, l'établissement de la fourche de réplication nécessite que le virus détourne à son profit la machinerie de réplication cellulaire. Par conséquent, certaines des observations issues de l'étude de la réplication du virus SV40 restent probablement valables pour décrire les évènements se déroulant après la dénaturation du duplex ADN dans le cadre de la réplication du chromosome cellulaire (Waga and Stillman, 1998).

Après que l'hélicase a dénaturé le duplex ADN, la matrice simple brin est aussitôt recouverte par le complexe RPA. Ensuite, les complexes MCM2-7 et RPA orchestrent le recrutement de l'ADN polymérase α, l'enzyme qui accomplit les évènements d'initiation de la synthèse d'ADN, au niveau de la matrice simple brin. L'ADN polymérase α synthétise alors une amorce ARN, par l'intermédiaire de son activité ARN polymérase, puis ajoute à cette amorce ARN une vingtaine de désoxyribonucléotides pour former un ADN initiateur. Dans la foulée, le complexe RFC procède au chargement du PCNA, le facteur de processivité des ADN polymérases réplicatives. Concrètement, le complexe RFC ouvre la structure annulaire du PCNA selon un mécanisme dépendant de la fixation d'une molécule d'ATP puis il adresse

[8] D'après Kelly, T.J., and Stillman, B. Ibid. Duplication of DNA in Eukaryotic Cells. pp. 1-29.

le PCNA à la matrice d'ADN au niveau de l'ADN initiateur par l'intermédiaire de la protéine RPA (Indiani and O'Donnell, 2006). La fermeture du PCNA autour du duplex formé par la matrice et l'ADN initiateur est dépendante de l'hydrolyse par le RFC de la molécule d'ATP, laquelle est stimulée par l'interaction du RFC avec l'ADN double-brin (Indiani and O'Donnell, 2006). Le chargement du PCNA précède la dissociation de l'ADN polymérase α et l'ancrage de l'ADN polymérase δ au niveau de l'ADN via le PCNA.

Sachant que l'initiation de la réplication du chromosome cellulaire eucaryote requiert un complexe de pré-réplication, il est certain que certains facteurs participant à l'activation de ce complexe, dispensable dans le cadre de la réplication du virus SV40, concourent au recrutement des polymérases. En particulier, les protéines Mcm10 et Cdc45 pourraient être impliquées dans le recrutement de l'ADN polymérase α – ADN primase (Aparicio et al., 1999; Ricke and Bielinsky, 2004), l'enzyme qui a la charge d'initier la synthèse des fragments d'ADN.

Dynamique de la fourche de réplication [9]

Les génomes eucaryotes codent trois ADN polymérases : l'ADN polymérase α, une enzyme qui contient aussi une activité ADN primase, l'ADN polymérase δ et l'ADN polymérase ε. Des données récentes ont permis de démontrer que l'ADN polymérase δ réplique le brin retardé alors que l'ADN polymérase ε réplique le brin continu (Nick McElhinny et al., 2008; Pursell et al., 2007). A l'image du réplisome bactérien, la fourche de réplication eucaryote est une structure dynamique au niveau de laquelle s'assemblent puis se dissocient les protéines impliquées dans la synthèse coordonnée des deux brins.

De la même façon que chez les bactéries, la synthèse du brin discontinu se déroule en plusieurs étapes : initiation par l'ADN primase ; élongation limitée de l'amorce ARN par

[9] D'après Burgers, P.M.J., and Seo, Y.-S. Ibid. Eukaryotic DNA Replication Forks. pp. 105-120.

l'ADN polymérase α – primase ; transfert de l'extrémité de l'amorce ARN-ADN de la Pol α à l'ADN polymérase δ ; élongation par la Pol δ ; maturation des fragments d'Okazaki. Chaque transition repose nécessairement sur des interactions spécifiques entre les protéines situées au niveau de la fourche de réplication, mais la nature de ces interactions fonctionnelles n'est pas clairement établie. En particulier, les protéines Cdc45 ou Mcm10 pourraient être impliquées dans le recrutement cyclique de la Pol α pour initier les fragments d'Okazaki (Aparicio et al., 1999; Ricke and Bielinsky, 2004).

Des données récentes suggèrent que le complexe GINS pourrait jouer un rôle de pont moléculaire au niveau de la fourche de réplication (Gambus et al., 2006; Labib and Gambus, 2007). D'autres données appuient au contraire l'idée que le complexe GINS pourrait stimuler l'activité d'autres facteurs de la réplication. En particulier, l'apparent manque d'activité hélicase du complexe MCM2-7 *in vitro* a jeté des doutes sur la nature exacte de l'hélicase réplicative. Aussi, le complexe GINS a été proposé comme un possible facteur auxiliaire du complexe MCM2-7 (Aparicio et al., 2006; Boskovic et al., 2007; Moyer et al., 2006; Pacek et al., 2006). Néanmoins, des données récentes démontrent que le complexe MCM2-7 possède bien une activité hélicase *in vitro* (Bochman and Schwacha, 2008).

Lorsque l'ADN polymérase officiant sur le brin discontinu rencontre l'amorce ARN du fragment d'Okazaki précédent, elle poursuit son action jusqu'à se trouver au contact du fragment d'ADN précédent. Ainsi, l'ADN polymérase déplace la partie du fragment d'Okazaki contenant la partie du fragment correspondant à l'amorce ARN et l'expose à l'action de nucléases spécifiques qui vont dégrader la partie saillante du fragment d'Okazaki. Chez les eucaryotes, un arsenal important de protéines semble être associé au processus de maturation des fragments d'Okazaki : Pol δ ; FEN-1 ; Dna2 ; Pif1 ; RPA. L'intervention de l'une ou l'autre des protéines dépend de la mobilisation ou non du complexe RPA laquelle est

fonction de la longueur de la structure générée par le déplacement de brin opéré par la Pol δ (pour une revue récente, voir (Burgers, 2008)).

Mécanisme de la réplication de l'ADN chez les archées

Comparativement aux études menées sur la réplication chez les bactéries et les eucaryotes, l'étude de la réplication chez les archées est jeune. Aussi, la quantité de données accumulées à l'heure actuelle est relativement limitée. Néanmoins, sachant que la majorité des protéines impliquées dans la réplication de l'ADN chez les archées sont homologues à celles décrites chez les eucaryotes (**Tableau 3**), le modèle élaboré à partir des organismes eucaryotes fournit une base de travail solide. En fait, les études de la réplication de l'ADN chez les archées se sont principalement focalisées sur la caractérisation biochimique et fonctionnelle des acteurs moléculaires, sur l'interprétation des structures cristallines et, dans une moindre mesure, sur la recherche et l'analyse des interactions fonctionnelles entre protéines. En outre, la majorité des travaux concernent l'étude de protéines impliquées dans l'initiation de la réplication. Les polymérases et d'autres protéines intervenant au niveau de l'élongation ont également fait l'objet de quelques investigations mais la nature des interactions et leur effet sur l'activité respective des partenaires protéiques demeurent énigmatiques. Enfin, le processus de la terminaison de la réplication de l'ADN chez les archées demeure, à ce jour, totalement inconnu.

Initiation de la réplication : mécanismes et régulation

La quasi-totalité des génomes d'Archaea arborent au moins un gène codant une protéine présentant des similarités avec les protéines eucaryotes Orc1 et Cdc6 : la protéine Cdc6/Orc1 (voir Chapitre IV). Sur la base de cette double homologie, il a été proposé que Cdc6/Orc1 ait

Tableau 3 : Protéines de la réplication de l'ADN chez les archées

	Crenarchaeota	Euryarchaeota	Thaumarchaeota
Initiation			
Protéine initiatrice	Cdc6/Orc1	Cdc6/Orc1	Cdc6/Orc1
Hélicase	MCM	MCM	MCM
Chargeur de l'hélicase ?	WhiP [1]	-	-
Assemblage de la fourche	GINS (Gins15 et Gins23)	(GINS) [2]	GINS (Gins15 et Gins23)
Elongation			
Hélicase	MCM (3' vers 5')	MCM (3' vers 5')	MCM (3' vers 5')
ADN primase	PriSL	PriSL	PriSL
Protéine fixant l'ADN simple brin	SSB (RPA) [3]	RPA (SSB) [4]	RPA SSB
Topoisomérase	Topo VI	Topo VI [5] (ADN gyrase) [5]	Topo VI Topo IB [6]
Polymérases réplicatives	PolB (PolD) [3]	PolB (Famille B) PolD (Famille D)	PolB (Famille B) PolD (Famille D)
Facteur de processivité	PCNA	PCNA	PCNA
Protéine chargeant le facteur de processivité	RFC	RFC	RFC
Maturation des fragments d'Okazaki			
Polymérase	?	?	?
Nucléase ARN	RNase HII ?	RNase HII ?	RNase HII ?
Endonucléase ADN	FEN-1	FEN-1 (Dna2) [7]	FEN-1
Hélicase	-	(Dna2)	-
ADN ligase	ATP-ligase	ATP-ligase [8]	ATP-ligase

[1] La protéine WhiP n'est pas conservée chez les Crenarchaeota (voir Chapitre IV)
[2] Les gènes codant les protéines Gins15 et Gins23 semblent ne pas avoir été conservés chez l'ensemble des Euryarchaeota (voir Chapitre IV)
[3] Le gène codant la PolD n'est présent que dans le génome de *Korarchaeum cryptofilum* (voir Chapitre IV)
[4] La protéine SSB est présente uniquement dans certains taxons au sein du phylum Euryarchaeota (voir Chapitre IV)
[5] L'ADN topoisomérase VI est absente des Thermoplasmatales, elle est remplacée par une ADN gyrase (voir Chapitre IV)
[6] Céline Brochier, Simonetta Gribaldo, Patrick Forterre ; résultats non publiés
[7] Seuls les génomes des Halobacteriales possèdent de vrais homologues de la protéine Dna2 eucaryote
[8] La nature du co-facteur de l'ADN ligase n'est pas conservée

pour fonction d'une part de reconnaître l'origine de réplication, d'autre part d'assister le chargement de l'hélicase réplicative MCM (Liu et al., 2000).

Très tôt après la caractérisation expérimentale de la première origine de réplication de l'ADN chez l'archée hyperthermophile *Pyrococcus abyssi* (Myllykallio et al., 2000), des expériences d'immunoprécipitation de la chromatine ont permis de montrer que la protéine Cdc6/Orc1 se fixe *in vivo* au niveau de la région chromosomique contenant l'origine *oriC* (Matsunaga et al., 2001). Par la suite, une analyse génétique a conduit à l'identification d'une séquence de réplication autonome chez *Halobacterium* sp. NRC1 (Berquist and DasSarma, 2003). Enfin, trois origines de réplication ont été décrites chez *Sulfolobus solfataricus* (*Sso*) et *S. acidocaldarius* (Lundgren et al., 2004; Robinson et al., 2004).

La caractérisation des sites de fixation des trois protéines Cdc6/Orc1 codées par le génome de *S. solfataricus* a permis de mettre en évidence la présence d'éléments de séquence inversés répétés, appelés ORB (Origin Recognition Boxes), au niveau de l'origine *SsooriC1* (Robinson et al., 2004). Ces éléments de séquence sont retrouvés au niveau de l'origine probable d'un grand nombre de génomes d'Archaea (Grainge et al., 2006; Robinson et al., 2004). Par ailleurs des éléments de séquence plus petits, appelés mini-ORB (mORB) ont été décrits au niveau de l'origine *SsooriC2* (Robinson et al., 2004). Enfin, la troisième origine de réplication de *S. solfataricus* (*SsooriC3*) contient de courts éléments répétés inversés, différents de ceux trouvés au niveau des deux autres origines, mais ceux-ci ne correspondent pas au site de fixation des trois protéines initiatrices *Sso*Cdc6-1, *Sso*Cdc6-2 et *Sso*Cdc6-3 au niveau de cette origine (Robinson et al., 2007). Par la suite, des éléments mORB ont été observés au niveau de l'origine de réplication prédite pour le génome de *Methano-thermobacter thermautotrophicus* (*Mth*) et il a été démontré que les deux protéines initiatrices *Mth*Cdc6-1 et *Mth*Cdc6-2 se fixent à ce motif de manière spécifique *in vitro* (Capaldi and Berger, 2004). Néanmoins, des analyses biochimiques plus récentes suggèrent que la

séquence nécessaire à la fixation de la protéine *Mth*Cdc6-1 pourrait être légèrement plus étendue que l'élément mORB précédemment décrit (Majernik and Chong, 2008). Par ailleurs, la fixation de la protéine Cdc6/Orc1 à l'origine de réplication est indépendante de l'ATP contrairement à la fixation du complexe ORC eucaryote (Robinson et al., 2004).

La position d'un certain nombre d'origines de réplication n'ont été prédites que par des analyses de biais nucléotidiques mais la localisation doit être confirmée expérimentalement ; la position des origines d'autres génomes échappent totalement aux analyses prédictives actuelles (Zhang and Zhang, 2005). Aussi, il n'est pas possible de dresser des généralités concernant les sites de reconnaissance de(s) protéine(s) initiatrice(s) de la réplication au niveau de(s) origine(s). En outre, les données déjà recueillies suggèrent que les éléments de séquence reconnus par les protéines initiatrices ne correspondent pas nécessairement à des motifs ORB ou mORB, ce qui suggère que ces motifs ne sont pas conservés de manière stricte (Grainge et al., 2006; Robinson et al., 2007). En revanche, les régions contenant des origines de réplication identifiées chez les Archaea se distinguent toujours par leur richesse en nucléotides AT, à l'image de ce qui est observé dans les génomes bactériens et eucaryotes. De manière remarquable, des données cristallographiques récentes, obtenues en parallèle par deux groupes, suggèrent que les motifs de reconnaissance des protéines initiatrices sont plutôt des motifs structuraux (Alderton, 2007; Dueber et al., 2007; Gaudier et al., 2007; Georgescu and O'Donnell, 2007; Tada et al., 2008).

Le mode de chargement de l'hélicase MCM au niveau de l'origine de réplication chez les Archaea demeure inconnu. Aucun homologue des protéines Cdt1, MCM8 ou MCM9 n'a été identifié dans les génomes d'Archaea. Aussi, la protéine Cdc6/Orc1 apparaît comme le candidat le plus probable pour le rôle de facteur de chargement de l'hélicase MCM, d'autant que la protéine Cdc6 présente des similarités de séquences manifestes avec les sous-unités du complexe RFC (Perkins and Diffley, 1998). En plus, de nombreuses études ont montré que la

protéine Cdc6/Orc1 module l'activité hélicase du complexe MCM (De Felice et al., 2004; Haugland et al., 2006; Kasiviswanathan et al., 2005; Shin et al., 2003b).

Néanmoins, les données obtenues à partir de modèles d'étude différents suggèrent que le mode de chargement et d'activation de l'hélicase MCM pourrait diverger d'une archée à une autre bien que les facteurs initiateurs soient homologues. Cette éventualité émerge de la différence observée entre les activités intrinsèques des complexes MCM purifiées à partir d'organismes archéens différents. Les complexes MCM de *S. solfataricus*, *Archaeoglobus fulgidus* et *M. thermautotrophicus* possèdent une forte activité intrinsèque et il est possible que ces protéines soient chargées sous une forme active (Grainge et al., 2003; Kelman et al., 1999a; McGeoch et al., 2005). Au contraire, les complexes MCM de *Thermoplasma acidophilum* et *Pyrococcus furiosus* ont une faible activité intrinsèque ce qui indique que l'hélicase MCM doit être activée ou stimulée par un facteur auxiliaire (Haugland et al., 2006; Yoshimochi et al., 2008).

Par ailleurs, les modalités de recrutement de l'hélicase MCM semblent différentes selon les espèces. Ces différences pourraient être liées à la différence entre le nombre de protéines Cdc6/Orc1 entre ces espèces ou bien à une divergence fonctionnelle de certains facteurs de réplication (voir Chapitre IV). Par exemple, la présence de la protéine Cdc6/Orc1 au niveau de l'origine de réplication ne semble pas suffisante pour promouvoir *in vivo* le chargement du complexe MCM chez *Pyrococcus abyssi* (Matsunaga et al., 2001). De manière intéressante, il a été montré que le complexe GINS se fixe au niveau de l'origine chez *P. furiosus*, ce qui suggère qu'il pourrait assister le chargement de l'hélicase (Yoshimochi et al., 2008). D'un autre côté, il semble que la protéine Cdc6-2 de *M. thermautotrophicum* dirige le chargement de l'hélicase chez cet organisme (Shin et al., 2008). De façon intéressante, l'architecture moléculaire du complexe GINS chez cet organisme pourrait différer de celle observée chez *P. furiosus* (voir Chapitre IV).

Dynamique de la fourche de réplication

La nature des interactions fonctionnelles au niveau de la fourche de réplication chez les Archaea n'est pas clairement établie. En outre, les données obtenues à partir de différents modèles d'étude font apparaître certaines disparités qui, à ce jour, ne sont pas résolues.

Des analyses récentes réalisées chez *Sulfolobus solfataricus* suggèrent que le complexe GINS forme un pont moléculaire entre le brin retardé et le brin continu afin de coordonner la synthèse des amorces ARN sur le brin retardé avec l'ouverture progressive de la double-hélice d'ADN (Marinsek et al., 2006). En revanche, les données obtenues chez *Pyrococcus furiosus* (*Pfu*) ne permettent pas de déterminer si un assemblage macromoléculaire comparable prend place au niveau de la fourche de réplication lors de la duplication du chromosome de cet organisme (Yoshimochi et al., 2008). En fait, ces analyses suggèrent plutôt que le complexe *Pfu*GINS stimule l'activité hélicase du complexe *Pfu*MCM mais aucune interaction physique stable n'a pu être mise en évidence (Yoshimochi et al., 2008). A l'inverse, *Sso*GINS ne semble pas stimuler l'activité de *Sso*MCM. Aussi, la fonction du complexe GINS archéen n'est, à l'image du complexe GINS eucaryote, pas clairement établie (voir aussi Chapitre IV).

Chacun des fragments d'Okazaki est initié par l'ADN primase laquelle cède sa place à une ADN polymérase réplicative qui va utiliser l'amorce produite pour allonger le brin naissant. Des données obtenues chez *S. solfataricus* soutiennent l'idée que le complexe RFC orchestre ce transfert entre les deux polymérases (Wu et al., 2007). En effet, il a été montré que la petite sous-unité du complexe RFC restreint l'activité de l'ADN primase et favorise son décrochage de la matrice (Wu et al., 2007). En retour, l'ADN primase stimule l'activité ATPase du complexe RFC, donc encourage le chargement du PCNA dirigé par RFC (Indiani and O'Donnell, 2006).

Tous les génomes d'Archaea arborent plusieurs gènes codant des ADN polymérases. A l'exception de *Korarchaeum cryptofilum*, toutes les Crenarchaeota possèdent plusieurs gènes

pour des ADN polymérases de la famille B (PolB) (voir Chapitre IV). Le rôle de ces ADN polymérases au niveau de la fourche de réplication n'a cependant fait l'objet d'aucune étude. Les génomes de tous les autres organismes (Euryarchaeota, Thaumarchaeota, *K. cryptofilum*) arborent des gènes codant une ou plusieurs PolB mais également des gènes codant une ADN polymérase singulière, propre aux Archaea, appelée PolD. Aussi, il est probable que le génome de ces organismes soit répliqué à l'aide de deux ADN polymérases, une situation comparable à celle des eucaryotes. En revanche, les données actuelles ne permettent pas de dire quel est le rôle joué par la PolB et la PolD au cours de la réplication. L'analyse des propriétés respectives des PolB et PolD de *Pyrococcus abyssi* suggèrent que la PolB assure la synthèse du brin continu tandis que la PolD réalise la synthèse du brin discontinu (Henneke et al., 2005). En outre, il a été proposé que la PolD joue un rôle analogue à celui de la Pol α eucaryote en convertissant l'amorce ARN en amorce ARN-ADN avant de se faire déplacer par la PolB sur le brin continu (Rouillon et al., 2007). Ce modèle nécessite d'être corroboré par des analyses menées chez d'autres modèles, d'autant que les modalités de l'interaction fonctionnelle entre le PCNA et la PolD diffèrent entre espèces très proches (Tori et al., 2007).

Avant l'intervention de l'ADN ligase, les amorces ARN synthétisée par l'ADN primase doivent être éliminées pour obtenir une molécule d'ADN homogène. Des expériences menées *in vitro* à partir des enzymes RNase HII et FEN-1 de *Pyrococcus furiosus* suggèrent, selon les auteurs, que ces deux enzymes coopèrent lors du processus de maturation des fragments d'Okazaki (Sato et al., 2003). L'impossibilité d'obtenir des mutants pour le gène codant la protéine FEN-1 chez *Halobacterium* sp. NRC1 suggèrent en effet que ce gène est essentiel (Berquist et al., 2007). Des expériences analogues réalisées chez *Haloferax volcanii* montrent au contraire que ce gène peut-être inactivé ; le phénotype du mutant obtenu appuie néanmoins l'idée selon laquelle FEN-1 joue un rôle essentiel au cours de la réplication (Meslet-Cladiere et al., 2007). En revanche, les gènes codant les deux RNases H chez *H. volcanii* — ce génome

contient, à l'image des génomes des autres archées halophiles, un gène codant une RNase HI, probablement d'origine bactérienne, et un gène codant une RNase HII — peuvent être inactivés sans affecter de manière notable la croissance cellulaire (Meslet-Cladiere et al., 2007). Aussi, les données génétiques suggèrent que la RNase HII n'est pas directement impliquée dans le processus de maturation des fragments d'Okazaki. Il a été proposé que la protéine RNase HII intervienne plutôt dans l'excision des ribonucléotides incorporés par erreur dans l'ADN en collaboration avec FEN-1 (Meslet-Cladiere et al., 2007; Rydberg and Game, 2002)

Origine et évolution de la machinerie de réplication

De manière globale, la comparaison des protéines impliquées dans la réplication de l'ADN dans les trois domaines cellulaires du vivant fait apparaître des différences inter- et intradomaines. D'une part, les machineries de réplication bactérienne, archéenne et eucaryote sont différentes les unes des autres (différences interdomaines) **(Tableau 4)**. D'autre part, la comparaison de la composition de la machinerie de réplication indique que l'appareil de réplication diffère entre différents phylums d'un même domaine (différences intradomaines). De manière générale, la machinerie de réplication d'un domaine cellulaire comprend un cœur invariable de protéines auquel s'ajoutent des facteurs, le plus souvent impliquées dans des mécanismes de régulation, qui varient d'un phylum à un autre ; pour une discussion argumentée concernant l'appareil de réplication chez les Archaea, voir le Chapitre IV. Les appareils de réplication bactérien et archéen ont en commun la propriété d'être relativement compacts, mais la plupart des protéines impliquées ne sont pas orthologues. A l'inverse, la majorité des protéines participant à la réplication de l'ADN chez les archées et les eucaryotes sont orthologues, ce qui suggère que ces protéines ont été héritées d'un ancêtre commun.

Tableau 4 : Protéines de la réplication de l'ADN dans les trois domaines cellulaires du vivant

	Bacteria*	Archaea*			Eukarya*
		Crenarchaeota	Euryarchaeota	Thaumarchaeota	
Initiation					
Protéine initiatrice	DnaA	Cdc6/Orc1	Cdc6/Orc1	Cdc6/Orc1	ORC1-6, Cdc6
Hélicase	DnaB (5′ vers 3′)	MCM (3′ vers 5′)	MCM (3′ vers 5′)	MCM (3′ vers 5′)	MCM2-7 (3′ vers 5′)
Chargeur de l'hélicase	DnaC	WhiP	-	-	Cdt1
Assemblage de la fourche	-	GINS	(GINS)	GINS	GINS, Mcm10, Cdc45, Sld2, Sld3, Dpb11
Kinases	-	-	-	-	Cdc7-Dbf4, CDK-cyclin
Elongation					
Synthèse ADN					
Hélicase	DnaB	MCM	MCM	MCM	MCM2-7
ADN primase	DnaG	PriSL	PriSL	PriSL	Complexe Pol α-PriSL
Protéine fixant l'ADN simple brin	SSB	SSB (RPA)	RPA (SSB)	RPA SSB	RPA
Topoisomérase	Topo IV ADN gyrase	Topo VI	Topo VI (ADN gyrase)	Topo VI Topo IB	Topo IB Topo II
Polymérases réplicatives	Pol III PolC	PolB (PolD)	PolB PolD	PolB PolD	Pol δ Pol ε
Facteur de processivité	Sous-unité β	PCNA	PCNA	PCNA	PCNA
Protéine chargeant le facteur de processivité	Complexe γ	RFC	RFC	RFC	RFC
Maturation des fragments d'Okazaki					
Polymérase	Pol I	?	?	?	Pol δ
Nucléase ARN	Pol I (activité exonucléase 5′-3′) RNase HII ?	RNase HII ?	RNase HII ?	RNase HII ?	RNase HII ?
Endonucléase ADN	Pol I (activité exonucléase 5′-3′)	FEN-1	FEN-1 (Dna2)	FEN-1	FEN-1 Dna2
Hélicase	-	-	(Dna2)	-	Dna2, Pif1
ADN ligase	NAD-ligase	ATP-ligase	ATP-ligase	ATP-ligase	ATP-ligase

* Voir les tableaux 1 à 3 pour de plus amples informations

Par ailleurs, certains composants des machineries de réplication bactérienne, archéenne et eucaryote sont homologues, bien que les similarités de séquence soient modestes dans certains cas. En particulier, le facteur de processivité et la protéine assurant son chargement sont conservés à l'échelle des trois domaines. Parmi les autres protéines impliquées dans la réplication, un certain nombre ne sont pas homologues (dérivés d'un ancêtre commun) mais arborent des domaines homologues, ce qui suggère qu'ils pourraient malgré tout dériver d'un patron protéique commun très ancien. Par exemple, les protéines initiatrices de la réplication des trois domaines cellulaires du vivant (DnaA ; Cdc6/Orc1 ; ORC1-6 et Cdc6) entrent dans ce cas de figure. En effet, ces protéines appartiennent toutes à la superfamille des ATPases AAA^+ mais elles sont issues de familles différentes (Leipe et al., 1999; Neuwald et al., 1999). Un autre exemple classique correspond aux hélicases réplicatives. Les protéines DnaB et MCM sont des ATPases et, l'une et l'autre sont membres des NTPases (nucléotide triphosphatase) contenant un repliement P-loop (Leipe et al., 1999). Néanmoins, la topologie structurale de ces deux enzymes est totalement différente dans la mesure où DnaB présente un repliement de type RecA (superfamille IV) alors que la protéine MCM présente un repliement AAA^+ (superfamille VI) (Berger, 2008; Leipe et al., 1999). Le dernier exemple est celui de la protéine fixant l'ADN simple brin. Dans les trois domaines, le motif de fixation à l'ADN correspond à un repliement OB (oligosaccharide/oligonucleotide/oligopeptide binding fold), un motif ancien que l'on observe dans nombre de familles protéiques fonctionnellement différentes (Leipe et al., 1999; Murzin, 1993).

De manière remarquable, les protéines ATPases de la famille AAA^+, l'ATP (fixation et hydrolyse) et l'ADN jouent un rôle régulateur fondamental dans le processus de réplication dans les trois domaines cellulaires du vivant. Qu'il s'agisse de la protéine initiatrice de la réplication, de l'hélicase ou de la protéine opérant le chargement du facteur de processivité

des ADN polymérases, toutes reposent sur l'énergie emmagasinée dans la molécule d'ATP pour fonctionner en interaction avec l'ADN.

En revanche, et de manière tout aussi remarquable, les enzymes qui polymérisent les précurseurs nucléotidiques sont incroyablement diverses. En effet, les bactéries possèdent des ADN polymérases réplicatives et une ADN primase caractéristiques qui ne sont absolument pas apparentées à celles trouvées chez les archées et les eucaryotes. En revanche, les archées et les eucaryotes synthétisent leur génome en utilisant des protéines présentant, pour la plupart, des liens évolutifs clairs.

Les bactéries possèdent une ou deux ADN polymérases réplicatives caractéristiques appartenant à la famille C (Pol III et PolC). Cette famille d'ADN polymérases est uniquement présente chez les bactériens et les bactériovirus (Dervyn et al., 2001). En outre, les bactéries possèdent une ADN primase monomérique (DnaG) dont le site catalytique présente un repliement de type Toprim que l'on retrouve notamment dans les ADN topoisomérases (Aravind et al., 1998; Keck et al., 2000). Une protéine de type DnaG est également présente chez les Archaea mais elle n'est pas essentielle (Le Breton et al., 2007). En outre, elle ne participe vraisemblablement pas à la réplication mais semble plutôt liée à l'exosome (Evguenieva-Hackenberg et al., 2003; Farhoud et al., 2005).

De manière intéressante, les sous-unités catalytiques des ADN polymérases eucaryotes (Pol α, Pol δ, Pol ε) et l'une des ADN polymérases des archées (Pol B) appartiennent à la famille des ADN polymérases B et sont apparentées. D'autre part, la plupart des archées possèdent une ADN polymérase hétérodimérique unique (Pol D) appartenant à une famille à part, la famille D (Ishino et al., 1998). De manière remarquable, la sous-unité exonuclease de la Pol D (DP1) et les sous-unités régulatrices des Pol α, Pol δ, Pol ε (communément appelées sous-unités B) sont également apparentées (Makiniemi et al., 1999), ce qui suggère que cette

sous-unité régulatrice était déjà présente chez l'ancêtre commun entre les Archaea et les Eukarya.

En ce qui concerne l'ADN primase, les archées et les eucaryotes possèdent l'un comme l'autre une ADN primase hétérodimérique dont les sous-unités catalytique et régulatrices sont clairement apparentées. En particulier, le site catalytique de l'ADN primase, contenue dans la petite sous-unité PriS, possède des similarités avec les ADN polymérases de la famille X — impliquées dans la réplication, la réparation et la recombinaison —, ce qui les distingue nettement des ADN primases bactériennes (Lao-Sirieix et al., 2005).

Pour conclure, le mécanisme de base de la réplication de l'ADN est universel ; certains facteurs participant à ce processus le sont aussi. De manière remarquable, les machineries de réplication des eucaryotes et des archées partagent un certain nombre de protéines orthologues, dérivées d'un ancêtre commun. En revanche, la composition protéique du réplisome bactérien diffère fondamentalement du patron archéo-eucaryotique. De manière intéressante, les ATPases occupent une place centrale dans le processus de réplication de l'ADN dans les trois domaines cellulaires du vivant. En revanche, les organismes de chacun des trois domaines cellulaires disposent d'ADN polymérases singulières pour répliquer leur génome. Enfin, les protéines impliquées dans la régulation de l'initiation varient fortement d'un phylum à l'autre, voire d'une famille d'organismes à l'autre, ce qui suggère que les stratégies de régulation adoptées ne sont pas conservées.

De manière remarquable, des résultats récents obtenus chez la levure *S. cerevisiae* suggèrent que des protéines impliquées dans la biogenèse des ribosomes (Noc3p et Yph1p), conservées à l'échelle des eucaryotes, interviennent au niveau de l'initiation de la réplication de l'ADN (Du and Stillman, 2002; Zhang et al., 2002). Le rôle de la protéine Noc3 n'apparaît néanmoins pas conservé chez la levure *Schizosaccharomyces pombe* chez laquelle il a été montré que Noc3p n'est pas essentielle pour la réplication (Houchens et al., 2008). Par

ailleurs, des résultats obtenus chez l'homme indiquent que le facteur de transcription MYC interagit avec les facteurs du complexe pré-réplicatif (ORC2, MCM2-7, Cdc6, Cdt1) et contrôle l'activité de l'origine de réplication indépendamment de son rôle de facteur de transcription (Dominguez-Sola et al., 2007). De manière intéressante, la protéine MYC régulerait également la traduction fournissant un lien entre deux processus cellulaires fondamentaux (Cole and Cowling, 2008). Par conséquent, il est possible que la réplication de l'ADN soit aussi régulée par des facteurs extérieurs, intervenant dans d'autres processus cellulaires, dans le but de mieux ajuster la synthèse d'ADN aux conditions physiologiques de la cellule. Pourtant, la réplication de l'ADN est souvent étudiée de manière isolée sans prendre en conditions ces possibles influences extérieures. La récente mise en évidence de liens génétiques entre la réplication et la glycolyse chez une bactérie (Janniere et al., 2007) suggère que ces systèmes de surveillance de l'état métabolique sont probablement conservés — quoique pas nécessairement universels. Aussi, il conviendrait sans doute de leur accorder une attention plus importante que celle qu'ils reçoivent à l'heure actuelle.

Analyse comparative des génomes

Parallèlement à l'accumulation des séquences de génomes dans les bases de données est née une science, appelée génomique comparative, destinée à tirer le maximum de profit des informations contenues dans les génomes en les comparant les uns aux autres. En effet, la première des tâches qui suit l'achèvement du séquençage d'un nouveau génome est la procédure d'annotation. Celle-ci consiste à délimiter les régions codantes du génome (protéines et ARNs) et à attribuer autant que faire se peut une fonction aux protéines codées par ce génome. Il est techniquement difficile d'envisager de recourir à une approche expérimentale pour assigner une fonction à chacun des gènes de chaque nouveau génome séquencé. Aussi, des méthodes informatiques ont été mises au point pour transférer l'information gagnée par l'étude d'un gène ou d'une protéine chez un organisme lambda aux autres organismes dont le génome est séquencé. L'approche la plus directe se base sur la recherche de similarités entre protéines ou domaines protéiques à l'aide de logiciels dédiés. Une classification naturelle des gènes basée sur les relations de parenté entre séquences a ensuite été développée afin d'élargir le champ d'application de la génomique comparative à l'histoire évolutive des génomes (Tatusov et al., 1997). Dans la continuité, des méthodes de génomique comparative exploitant l'information relative au contexte génomique ont vu le jour. Elles s'appuient sur l'analyse des fusions de gènes, de l'environnement des gènes, de la distribution phylogénétique pour établir des liens fonctionnels entre deux ou un groupe de gènes (pour des revues, voir (Gabaldon and Huynen, 2004; Galperin and Koonin, 2000; Huynen et al., 2000; Marcotte, 2000)). Elles seront brièvement décrites dans les pages qui suivent.

Concept d'orthologie

L'objectif principal de la génomique comparative est d'identifier dans différents génomes les gènes qui codent des protéines qui accomplissent les mêmes fonctions dans la cellule. Cette identification repose principalement sur un critère de similarité de séquence. Lorsque deux protéines ont des séquences assez similaires, il y a une probabilité relativement grande que ces deux protéines dérivent d'une protéine ancestrale commune, c'est-à-dire que ces protéines soient homologues. Or, les protéines homologues remplissent souvent des fonctions similaires ou proches au sein de la cellule. Aussi, identifier, en comparant différents génomes, les gènes codant des protéines homologues permet d'inférer la fonction probable d'une protéine en se basant sur les connaissances acquises par la caractérisation de l'un de ses homologues.

Les relations d'homologie entre gènes sont principalement de deux ordres : soit les gènes sont orthologues, soit ils sont paralogues. Les gènes orthologues sont des gènes d'espèces différentes qui ont évolué par spéciation à partir d'un gène ancestral présent chez le dernier ancêtre commun desdites espèces. Les gènes paralogues sont des gènes qui ont évolué à partir d'un gène ancestral via un évènement de duplication. Généralement, les gènes orthologues conservent une fonction identique au cours de l'évolution, alors que les gènes paralogues développent des fonctions nouvelles, quoique liées à la fonction originale. Afin de faciliter l'identification des gènes orthologues, une base de données dans laquelle les gènes sont répartis suivant des ensembles de groupes orthologues (COG, Cluster of Orthologous Groups) a été créée (Tatusov et al., 1997) ; l'utilité de cette base de données a notamment été éprouvée lors de l'annotation de deux génomes d'Archaea (Natale et al., 2000; Slesarev et al., 2002). Depuis sa création, la base de données COG a été peu réactualisée mais des bases de données annexes, centrées sur les génomes eucaryotes (KOG, euKaryotic Orthologous Groups) et archéens (arCOG, archaeal COG), ont été développées (Makarova et al., 2007; Tatusov et al., 2003) ; la base de données arCOG a, en particulier, été mise à profit au cours

du processus d'annotation du génome de l'archée hyperthermophile Candidatus *Korarchaeum cryptofilum* (Elkins et al., 2008).

Analyse du contexte génomique

Les méthodes d'analyse du contexte génomique reposent sur l'étude des relations entre gènes au sein d'un génome ou de génomes différents **(Figure 9)**. Ces méthodes ont pour objectif de prédire la catégorie fonctionnelle d'un gène de fonction inconnue ou d'inférer des interactions fonctionnelles entre protéines. Ces approches méthodologiques ne reposant pas sur le critère de similarité, elles permettent d'obtenir des informations complémentaires à celles obtenues par l'intermédiaire des méthodes basées sur la recherche d'homologie de séquence.

Environnement des gènes

Cette méthode repose sur l'analyse de l'organisation des gènes dans les génomes et consiste à rechercher des associations conservées de gènes de type opéron ou divergon **(Figure 9a)**. Un opéron est un groupe de gènes qui partagent un même promoteur et qui sont transcrits en un ARN messager polycistronique. Les protéines codées par des gènes organisés en opéron participent généralement à une même voie métabolique ou interagissent les unes avec les autres au sein d'un complexe protéique (Dandekar et al., 1998; Overbeek et al., 1999). Le regroupement de gènes en opérons pourrait optimiser leur co-régulation (Hershberg et al., 2005; Price et al., 2005) ou faciliter l'assemblage des produits des gènes dans des complexes macro-moléculaires (Glansdorff, 1999). Un divergon correspond à une association divergente entre un gène codant un élément régulateur (e.g., un facteur de transcription) et un gène ou un groupe de gènes dont l'expression est sous le contrôle de ce régulateur (Korbel et al., 2004).

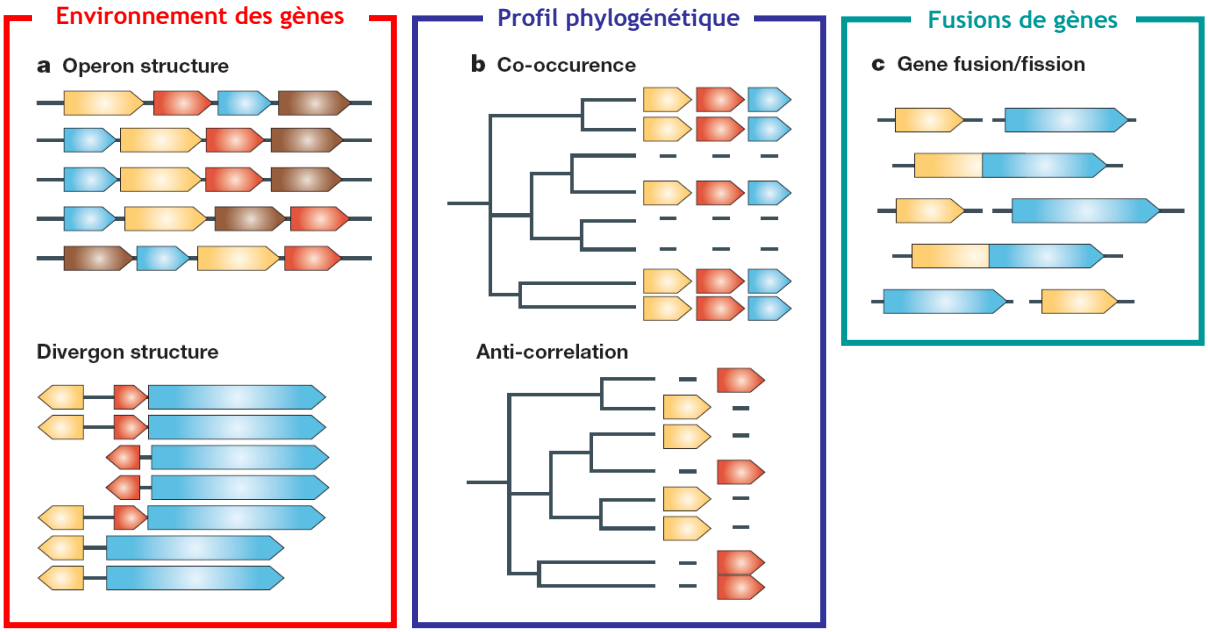

Figure 9 : Méthodes de génomique comparative basées sur l'analyse du contexte. (a) La méthode d'analyse de l'environnement des gènes repose sur l'identification d'associations conservées de gènes en structures de type opéron (haut) ou divergon (bas). (b) La méthode de profil phylogénétique repose sur l'analyse de la distribution des gènes dans les génomes. Le profil de co-occurrence est caractéristique des gènes ayant une trajectoire évolutive commune. Le profil d'anti-corrélation est associé à un cas de déplacement d'un gène par un gène non orthologue. (c) La méthode d'analyse de fusions de gènes repose sur l'identification de cas où deux phases ouvertes de lecture distinctes dans certains génomes forment une phase ouverte de lecture commune dans un ou quelques autres génomes. D'après (Ettema et al., 2005; Korbel et al., 2004)

Profil phylogénétique

Cette méthode repose sur l'analyse de la distribution des gènes dans les génomes **(Figure 9b)**. La méthode dite de co-occurrence s'appuie sur le postulat que des gènes qui sont fonctionnellement liés les uns aux autres suivent une trajectoire évolutive identique (Pellegrini et al., 1999). Autrement dit, si l'on compare différents génomes ces gènes auront tendance à être soit tous présents, soit tous absents d'un génome donné à un autre (Pellegrini et al., 1999). La méthode dite d'anti-corrélation repose sur la constatation que certaines paires de gènes ont des profils phylogénétiques complémentaires ou partiellement complémentaires (Galperin and Koonin, 2000). Autrement dit, la répartition de certaines paires de gènes au sein des génomes est telle que la présence combinée des deux gènes dans un même génome semble incompatible. Une telle distribution témoigne du déplacement d'un gène par un gène non

orthologue, c'est-à-dire qu'une protéine assurant une fonction donnée est remplacée par une protéine de même fonction ne présentant aucune relation de parenté avec la protéine originelle (Galperin and Koonin, 2000; Koonin et al., 1996). Par exemple, l'inventaire des gènes métaboliques contenus dans un génome révèle parfois l'absence d'un ou de plusieurs gènes codant des enzymes intervenant au sein d'une voie métabolique. La méthode d'anti-corrélation peut permettre de proposer un gène candidat assurant cette fonction 'manquante' (Galperin et al., 2000; Makarova and Koonin, 2003b). Des expériences biochimiques peuvent ensuite être mises en œuvre afin de tester la pertinence biologique de la prédiction effectuée (Galperin et al., 2000; Siebers et al., 2001).

Gènes fusionnés

La méthode d'analyse de fusions de gènes repose sur l'identification de cas où deux phases ouvertes de lecture, distinctes dans certains génomes, forment une phase ouverte de lecture commune dans un ou quelques autres génomes (Marcotte et al., 1999) **(Figure 9c)**. La protéine de fusion codée par le gène unique symbolise l'interaction fonctionnelle entre les deux protéines. Ce type d'association concerne principalement les gènes codant des protéines impliquées dans les voies métaboliques (Enright et al., 1999; Tsoka and Ouzounis, 2000).

Présentation de la thématique de recherche

Dans les trois domaines cellulaires du vivant (Archaea, Bacteria, Eukarya), l'ADN est le support physique de l'hérédité. Chez l'ensemble des êtres vivants, la réplication du matériel génétique est une étape cruciale du cycle cellulaire car elle assure la continuité du programme génétique au fil des générations tout en ouvrant la voie à son évolution. En outre, la duplication du matériel génétique fait l'objet d'une régulation très fine car les altérations chromosomiques compromettent l'intégrité du génome et mettent en péril la continuité cellulaire.

De nos jours, la plupart des processus cellulaires sont étudiés à travers le prisme de réactions reconstituées dans un tube à essais afin de moduler à l'envi les conditions d'expérimentation. A ce titre, l'élaboration d'un système de réplication de l'ADN *in vitro* à partir de composants protéiques purifiés a joué un rôle décisif dans la compréhension des mécanismes moléculaires qui se déroulent lors de la réplication de l'ADN chez les bactéries. Néanmoins, le nombre important de facteurs impliqués rend souvent cette approche délicate lorsqu'il s'agit d'un modèle d'étude eucaryote. En ce qui concerne un grand nombre de processus cellulaires, le système archéen représente une version simplifiée par rapport au système eucaryote correspondant. Aussi, les modèles d'études Archaea se sont rapidement imposés comme des modèles alternatifs à l'étude de certains processus cellulaires fondamentaux eucaryotes, dont le mécanisme de réplication de l'ADN.

Or, il n'existe pas à ce jour de système permettant de reproduire *in vitro* la synthèse d'une molécule d'ADN à partir de protéines d'Archaea purifiées. Disposer d'un tel outil expérimental représenterait une avancée technique décisive car cela pourrait ensuite permettre

d'élucider l'ensemble des interactions fondamentales au sein de la machinerie de réplication archéenne et, par extension, eucaryote. Mon travail de thèse s'inscrivait dans le projet commun des laboratoires du professeur Patrick Forterre à Orsay et du professeur Yoshizumi Ishino à Fukuoka (Japon) d'élaborer un système de réplication de l'ADN *in vitro* chez les Archaea. Le laboratoire du professeur Ishino disposait déjà de la majorité des protéines impliquées dans la phase d'élongation de la réplication chez *P. furiosus* à l'état purifié avant mon arrivée au Japon. En revanche, le laboratoire ne parvenait pas à purifier la protéine initiatrice de la réplication directement à partir de la fraction soluble. Mon objectif premier a donc été d'optimiser les conditions d'expression d'une forme soluble de la protéine *Pfu*Cdc6/Orc1 chez *E. coli* puis de mettre au point une méthode de purification simple et robuste à partir de cette forme soluble. Dans un second temps, l'étape d'initiation de la réplication pourrait être examinée *in vitro*.

La protéine *Pfu*Cdc6/Orc1 étant hautement instable en solution, la plus grande partie de mon travail s'est concentré sur la recherche d'une condition permettant d'accroître la stabilité de la protéine. Par ailleurs, j'ai purifié la protéine *Pfu*Cdc6/Orc1 exprimée chez la levure *Pichia pastoris* à partir des corps d'inclusion. Avant mon départ du Japon, j'ai initié une recherche d'interactions physiques par la méthode de résonance du plasmon de surface entre la protéine *Pfu*Cdc6/Orc1 purifiée en conditions dénaturantes et l'ensemble des protéines disponibles sur puce dans le laboratoire du professeur Ishino. Les protéines impliquées dans la réplication disponibles à l'état purifié dans le laboratoire ont également été testés contre la protéine *Pfu*Cdc6/Orc1 fixée sur puce. L'ensemble de ces travaux sont présentés dans le Chapitre I.

Parallèlement à cette approche expérimentale, j'ai réalisé une analyse comparative du contexte génomique de l'ensemble des gènes de la réplication dans les génomes d'Archaea.

Les génomes sont sans cesse remaniés par des réarrangements chromosomiques. En dépit de ce remodelage permanent des associations de gènes appartenant à une catégorie fonctionnelle commune peuvent être observées dans les génomes des archées et des bactéries (Dandekar et al., 1998; Wolf et al., 2001). Lorsqu'une association de gènes est conservée à travers de nombreux génomes, il ne s'agit probablement pas d'une coïncidence. Cela signifie sans doute que des contraintes évolutives maintiennent cette organisation au fil des générations malgré les remaniements du génome car la cellule tire un avantage de cet arrangement particulier des gènes, par exemple dans la co-régulation de l'expression de protéines fonctionnellement liées. Le corollaire de cette assertion est qu'il est possible d'analyser le contexte génomique des gènes pour prédire des interactions fonctionnelles entre les protéines qu'ils codent.

Les études de génomique comparative consistent généralement en une analyse globale du génome (pour des exemples, voir (Graham et al., 2000; Makarova et al., 1999; Makarova et al., 2007; Wolf et al., 2001)). De fait, les conclusions de ces études sont le plus souvent générales et rares sont celles qui s'inscrivent dans une démarche précise (pour des exemples voir, (Gao and Gupta, 2007; Koonin et al., 2001)). Nous avons choisi une approche innovante consistant à se concentrer sur l'analyse d'une classe fonctionnelle particulière de gènes. Aussi, nous avons initié une analyse comparative du contexte génomique de l'ensemble des gènes de la réplication dans vingt-sept génomes d'Archaea. Cet examen avait pour objectif de mettre en évidence des associations conservées impliquant des gènes de la réplication à partir desquelles nous pourrions d'une part suggérer de nouvelles interactions fonctionnelles, d'autre part proposer de probables nouvelles protéines de la réplication. Les résultats de cette étude sont présentés sous la forme d'un article scientifique, publié dans *Genome Biology* en 2008, dans le Chapitre II.

Cette analyse nous a permis de mettre en évidence des associations de gènes très conservés impliquant des gènes de la réplication. De manière inattendue, cette étude nous a

permis d'identifier une association très conservée entre des gènes de la réplication et des gènes liés au ribosome. Cette organisation suggère l'existence d'un mécanisme de couplage entre la réplication de l'ADN et la traduction. Aussi, nous avons initié une recherche bibliographique exhaustive destinée à recueillir des informations appuyant cette hypothèse. De manière remarquable, des données expérimentales obtenues chez des modèles bactériens et eucaryotes appuient cette idée. Cette étude bibliographique, au format minirevue, se trouve à la suite de l'article scientifique dans le Chapitre II.

Afin de confirmer la pertinence biologique des interactions prédites par l'analyse des génomes d'Archaea, j'ai dans un premier temps recherché des interactions physiques par co-immunoprécipation ou par la méthode de résonance du plasmon de surface, en utilisant les protéines disponibles dans le laboratoire du professeur Ishino. Dans un second temps, l'ensemble des gènes se trouvant dans les associations génétiques mises en évidence durant l'analyse du contexte génomique ont été clonés dans des vecteurs d'expression. Enfin, j'ai optimisé une méthode de criblage des interactions physiques basée sur une co-expression des candidats protéiques d'interaction suivie d'une co-purification à partir de la fraction protéique thermostable pour faciliter l'analyse des interactions physiques.

Au cours de l'analyse comparative du contexte génomique, tous les gènes potentiellement impliquées dans la réplication de l'ADN contenus dans les génomes d'Archaea ont été répertoriés. Cet inventaire a ensuite été mis à jour au fil de la mise à disposition des nouvelles séquences de génomes d'Archaea. Le contenu des génomes s'est révélé être particulièrement hétérogène d'un groupe taxonomique à un autre. Aussi, nous avons cherché à savoir si cette disparité pouvait avoir une signification évolutive.

En se basant sur les données les plus récentes concernant la phylogénie des Archaea (Brochier-Armanet et al., 2008), nous avons analysé la distribution de chacun des gènes de la réplication dans les génomes d'Archaea. Pour la plupart des gènes, la distribution phylétique donne une idée de leur histoire évolutive respective. Aussi, nous avons interprété la distribution de chacun des gènes de la réplication selon une logique de parcimonie afin de reconstruire la composition probable de la machinerie de réplication de l'ADN chez le dernier ancêtre commun des Archaea.

Ce manuscrit de thèse est donc divisé en quatre chapitres :

- le premier chapitre a trait à l'étude de la protéine initiatrice de la réplication du chromosome de *Pyrococcus furiosus* ;
- le deuxième chapitre porte sur l'analyse comparative du contexte génomique des gènes de la réplication dans les génomes d'Archaea ;
- le troisième chapitre concerne la recherche d'interactions physiques entre protéines en rapport avec les prédictions de l'analyse comparative du contexte génomique ;
- le quatrième chapitre s'intéresse à l'analyse du profil phylétique des gènes de la réplication et à l'évolution de la machinerie de réplication de l'ADN chez les Archaea.

RESULTATS & DISCUSSION

Chapitre I : Etude de la protéine initiatrice de la réplication du chromosome de *Pyrococcus furiosus* (*Pfu*) : *Pfu*Cdc6/Orc1

Chapitre II : Analyse comparative du contexte génomique des gènes de la réplication dans les génomes d'Archaea

Chapitre III : Recherche d'interactions physiques entre proteines

Chapitre IV : Analyse phylétique des gènes de la réplication et évolutionde la machinerie de réplication chez les Archaea

Chapitre I

Etude de la protéine initiatrice de la réplication du chromosome de *Pyrococcus furiosus* (*Pfu*) : *Pfu*Cdc6/Orc1

Chapitre I : Etude de la protéine initiatrice de la réplication du chromosome de *Pyrococcus furiosus* (*Pfu*) : *Pfu*Cdc6/Orc1

Introduction

La première étape de l'élaboration d'un système de réplication de l'ADN *in vitro* repose sur la nécessité de purifier la protéine initiatrice de la réplication. La protéine *Pfu*Cdc6/Orc1 est majoritairement exprimée sous une forme insoluble chez *E. coli* ce qui rend sa purification à partir de la phase soluble relativement délicate. En outre, les protéines initiatrices de la réplication sont affines à l'ADN, généralement prônes à l'agrégation et instables. L'ensemble de ces difficultés peut-être contourné en procédant à une purification en conditions dénaturantes car les interactions protéines-ADN sont abolies et les agrégats résorbés dans de telles conditions. Néanmoins, la dénaturation représente une étape critique car il est nécessaire de trouver des conditions dans lesquelles la protéine peut être renaturée. Or, il n'est pas possible de garantir qu'à l'issue du processus de dénaturation-renaturation, la protéine se trouve dans une conformation analogue à celle qui était la sienne avant qu'elle ne soit dénaturée. En outre, il est possible qu'à l'issue de cette étape la protéine soit partiellement inactivée. Or, il est primordial de s'assurer que la protéine initiatrice de la réplication de l'ADN soit pleinement active si l'on souhaite pouvoir étudier l'ensemble de ses propriétés fonctionnelles. L'approche la plus naturelle pour obtenir une protéine pleinement active consiste probablement à éviter de la modifier, donc à éviter autant que possible de la dénaturer. Aussi, mon travail s'est focalisé sur le moyen d'optimiser les conditions

d'expression d'une forme soluble de la protéine *Pfu*Cdc6/Orc1 chez *E. coli* à partir de laquelle une purification simple et directe pouvait être mise en œuvre.

Chez les Archaea, la protéine Cdc6/Orc1 occupe de part sa fonction d'initiateur de la réplication de l'ADN un rôle primordial dans un processus fondamental du cycle cellulaire : la duplication du matériel génétique. Deux études structurales récentes ont permis de dévoiler la manière dont la protéine Cdc6/Orc1 fixe et déforme l'origine de réplication *oriC* au moment de l'initiation (Dueber et al., 2007; Gaudier et al., 2007). Cependant, le mode de régulation de l'activité initiatrice de la protéine et la nature de ses partenaires protéiques ne sont, à ce jour, pas clairement établis. Dans le but de mieux délimiter l'étendue des activités de la protéine *Pfu*Cdc6/Orc1, j'ai recherché des interactions physiques entre la protéine *Pfu*Cdc6/Orc1 et diverses protéines impliquées dans la réplication ou la réparation de l'ADN chez *P. furiosus* à l'aide de la technique de la résonance du plasmon de surface.

Résultats

Clonage

Le gène codant la protéine *Pfu*Cdc6/Orc1, protéine initiant la réplication de l'ADN chez *P. furiosus*, a été amplifié par PCR et inséré dans le vecteur de clonage pGEM®-T easy. Les sept inserts qui ont été séquencés montraient une mutation au niveau de la position nucléotidique 113 du gène *cdc6/orc1* (une cytosine remplaçait l'adénine du gène sauvage). Le séquençage de la région du génome concernée a permis de montrer que la mutation observée n'est pas la conséquence du processus de clonage mais correspond à une mutation du génome de la souche de *P. furiosus* conservée dans le laboratoire du LBMGE à Orsay. L'amplification par PCR et le clonage du gène *cdc6/orc1* à partir d'une extraction d'ADN génomique préparée à partir d'une culture de la souche de *P. furiosus* conservée dans le laboratoire du professeur

Ishino à Fukuoka a révélé que la mutation était absente, ce qui confirme que la mutation observée dans les produits d'amplification clonés a eu lieu *in vivo*. La mutagenèse dirigée par PCR du gène *cdc6/orc1* a été réalisée dans le vecteur de clonage, l'opération se révélant impossible dans le vecteur d'expression, certainement en raison de la toxicité de la protéine Cdc6/Orc1 pour la bactérie *E. coli*. Des difficultés ont également été rencontrées pour préparer des extraits plasmidiques de la construction pGEM®-T [PF0017].

Recherche des conditions optimales d'expression d'une forme soluble de PfuCdc6/Orc1

La protéine *Pfu*Cdc6/Orc1 se montrant majoritairement insoluble dans des conditions standards d'expression chez *E. coli*, une grande partie de mon travail a été consacrée à la recherche de conditions permettant d'optimiser l'expression d'une forme soluble de la protéine *Pfu*Cdc6/Orc1 à partir de laquelle une purification simple pouvait être entreprise. Dans un deuxième temps, mon travail a consisté à rechercher le moyen de séparer la forme soluble de la protéine des acides nucléiques de l'hôte bactérien. En parallèle, j'ai cherché à stabiliser la forme soluble de *Pfu*Cdc6/Orc1 au cours du processus de purification par chromatographie d'affinité en variant la composition des tampons de lyse.

Culture d'expression. Les essais préliminaires d'expression de la protéine ont montré que la souche d'expression la plus adaptée est la souche Rosetta(DE3)pLysS et les essais d'expression suivants ont été réalisés avec cette souche.

En premier lieu, l'influence de la température d'expression sur la solubilité de la protéine Cdc6/Orc1 a été examinée. Contrairement aux recommandations d'usage qui préconisent d'abaisser la température d'expression pour réduire la formation de corps d'inclusion (pour une revue récente, voir (Sorensen and Mortensen, 2005)), les cultures d'expression ont été réalisées à haute température (45°C), cette approche ayant été utilisée

avec succès pour l'expression chez *E. coli* de protéines d'hyperthermophiles (Koma et al., 2006). La croissance cellulaire de cette souche bactérienne se poursuit lentement mais de manière régulière à cette température conformément aux observations faites par Koma et collaborateurs, ce qui indique que la souche Rosetta(DE3)pLysS se prête bien à la culture à haute température. De façon remarquable, la culture à haute température n'est pas compatible avec toutes les souches bactériennes comme le suggèrent d'autres résultats expérimentaux (Kube et al., 2006; Ron and Davis, 1971). La culture d'expression à haute température a semblé un temps prometteuse pour accroître la solubilité de *Pfu*Cdc6/Orc1 mais elle a été abandonnée pour deux raisons. D'une part, l'expression à haute température entraîne la production de protéines de choc thermique par la cellule hôte qui s'ajoutent au pool de protéines thermostables, ce qui limite l'intérêt de l'étape de dénaturation thermique de l'extrait protéique soluble qui permet d'éliminer la majeure partie des protéines thermostables bactériennes. D'autre part, des résultats comparables en termes de solubilité apparente de la protéine ont pu être atteints en cultivant à 37°C.

En réalité, le paramètre décisif pour l'expression quantitative d'une forme soluble de la protéine *Pfu*Cdc6/Orc1 chez *E. coli* semble être la durée de l'incubation consécutive à l'induction de l'expression protéique, car la culture bactérienne manifeste des difficultés de croissance de manière précoce. Les conditions optimales d'expression d'une forme soluble de la protéine *Pfu*Cdc6/Orc1 dans la souche Rosetta correspondent à une expression de la protéine sur une période de 4 heures à 37°C. Les quantités de protéine soluble obtenues de la sorte sont appréciables mais la part soluble de la protéine *Pfu*Cdc6/Orc1 est contaminée par de l'ADN et montre des signes d'instabilité.

Elimination de l'ADN. Différentes approches ont été exploitées pour tenter d'éliminer l'ADN contaminant la fraction soluble de la protéine *Pfu*Cdc6/Orc1 : traitement à la DNaseI

(20 mg/l) ; précipitation à la polyéthylènimine ; précipitation au sulfate d'ammonium suivie d'une chromatographie hydrophobe. Toutes ces approches se sont révélées infructueuses mais toutes semblent suggérer que la stabilité de la protéine *Pfu*Cdc6/Orc1 est consubstantielle à la présence d'ADN car soit la protéine devient thermolabile après traitement à la DNase, soit la protéine co-précipite avec l'ADN. Aussi, des expériences pour réduire les interactions entre la protéine *Pfu*Cdc6/Orc1 et l'ADN et stabiliser la protéine en solution ont été menées en jouant sur la composition des tampons de lyse utilisés. Par ailleurs, le comportement erratique de la protéine pouvant être attribuable à l'interférence de protéines bactériennes, des tentatives de purification par chromatographie d'affinité ont été entreprises dans l'espoir que la protéine *Pfu*Cdc6/Orc1 puisse être séparée de l'ADN à une étape ultérieure.

Tampons de lyse et chromatographie d'affinité IMAC. Les culots cellulaires ont été repris dans divers tampons afin d'accroître la stabilité de la protéine *Pfu*Cdc6/Orc1 en solution car la protéine a tendance à précipiter sur la matrice Ni Sepharose™ utilisée pour la chromatographie d'affinité. Les premiers essais de purification ont montré que la protéine Cdc6/Orc1 est thermolabile lorsque le tampon de lyse contient du chlorure de sodium alors qu'elle est thermostable en absence de ce sel. Paradoxalement, les essais de précipitation des acides nucléiques à la polyéthylènimine suggèrent que la protéine Cdc6/Orc1 co-précipite moins en présence de concentrations élevées en chlorure de sodium. Ces observations suggèrent que la protéine a une très forte affinité pour l'ADN et qu'il est très difficile de l'en séparer. L'effet déterminant de la concentration en chlorure de sodium sur la solubilité d'un des homologues Cdc6/Orc1 de l'archée *Archaeoglobus fulgidus* (*Afu*) avait été soulignée par d'autres auteurs (Grainge et al., 2003). Ces auteurs mentionnent également que le positionnement de l'étiquette en position N-terminale a accru la solubilité de la protéine *Afu*Cdc6/Orc1 (Grainge

et al., 2003), mais cette opération n'a eu aucun effet notable dans le cas de la protéine *Pfu*Cdc6/Orc1.

La concentration intracellulaire en potassium est de l'ordre de 500 à 600 mM chez *P. furiosus* (Scholz et al., 1992). L'ion potassium joue également un rôle important chez la bactérie *E. coli* en tant qu'osmolyte, de même que l'ion glutamate (Richey et al., 1987). Afin d'essayer de se rapprocher des conditions physiologiques de *P. furiosus*, et parce qu'il a été précédemment montré que la présence de glutamate de potassium peut accroître la thermostabilité ou stimuler l'activité des protéines d'Archaea (Armengaud et al., 2003; Hethke et al., 1999), l'effet du glutamate de potassium (400 mM) a été examiné. Néanmoins, la présence de ce sel n'a pas affecté de manière notoire la stabilité de la protéine *Pfu*Cdc6/Orc1 au cours de la purification. La présence de glutamate de potassium a même pu avoir un effet contre-productif dans la mesure où la présence de ce sel a été montré comme favorisant les interactions ADN-protéine *in vitro* (Leirmo et al., 1987).

La protéine *Pfu*Cdc6/Orc1 semble hautement labile et sujette à l'agrégation, une propriété déjà observée dans le cas de la protéine DnaA (Fuller and Kornberg, 1983). Cette tendance à l'agrégation semble donc être une caractéristique commune des protéines initiatrices de la réplication. Récemment, des auteurs ont rapporté que les tampons utilisés de façon courante pour les protocoles de purification par chromatographie IMAC (e.g., Tris) ne conviennent pas à la purification de la protéine DnaA et que les tampons à base de HEPES doivent leur être préférés (Zawilak-Pawlik et al., 2006). En outre, ces auteurs ont montré que la présence de glutamate de potassium est critique et affecte de manière importante le comportement de la protéine DnaA en solution. Aussi, un essai avec un tampon à base d'HEPES contenant du glutamate de potassium a été entrepris mais aucune variation significative sur la thermostabilité de la forme soluble de la protéine *Pfu*Cdc6/Orc1 n'a été notée par rapport au tampon à base de Tris.

L'association de la protéine *Pfu*Cdc6/Orc1 avec la membrane plasmidique a également été envisagée comme une source possible de l'instabilité de la protéine en solution sachant que la protéine DnaA et le complexe ORC interagissent *in vivo* et *in vitro* avec les phospholipides (Hwang et al., 1990; Lee et al., 2002; Sekimizu et al., 1988; Xia and Dowhan, 1995). Néanmoins, l'utilisation d'un détergent non ionique (Triton X100) ou zwitterionique (CHAPS) dans le tampon de lyse n'a pas eu d'effet notable sur la stabilité de la protéine en solution.

Recherche d'interactions physiques

La protéine *Pfu*Cdc6/Orc1 utilisée pour la recherche d'interactions a été purifiée en conditions dénaturantes **(Figure I-1)** car les tentatives de purification à partir de la fraction soluble se révélaient infructueuses. Contrairement aux réserves évoquées ci-dessus dans le cadre de l'objectif d'élaboration d'un système de réplication de l'ADN *in vitro*, l'incertitude concernant la conformation tridimensionnelle et l'activité de la protéine purifiée a été jugée acceptable dans le cadre d'une recherche d'interaction entre protéines. Aussi, une interaction a été recherchée entre la protéine *Pfu*Cdc6/Orc1 et toutes les protéines pour lesquelles une puce était disponible au sein du laboratoire du professeur Ishino : le PCNA (PF0983) (Kiyonari et al., 2006), la sous-unité Gins23 (PF0483) (Yoshimochi et al., 2008), l'exonucléase RecJ (PF2055) (Imamura et Ishino, résultats non publiés), l'ADN ligase (PF1635) (Kiyonari et al., 2006), l'hélicase Hel308a/Hjm (PF0677) (Fujikane et al., 2006), la DNase I (PF2046) (Tori et Ishino, résultats non publiés). Aucune interaction n'a pu être observée entre *Pfu*Cdc6/Orc1 et chacune des protéines susnommées. En dehors de la protéine Gins23 qui participe à l'initiation de la réplication, toutes les autres protéines interviennent soit à un autre niveau (PCNA, ADN ligase), soit dans d'autres registres (RecJ, Hel308a/Hjm, DNaseI). L'absence d'interaction entre chacune de ces protéines et la protéine *Pfu*Cdc6/Orc1

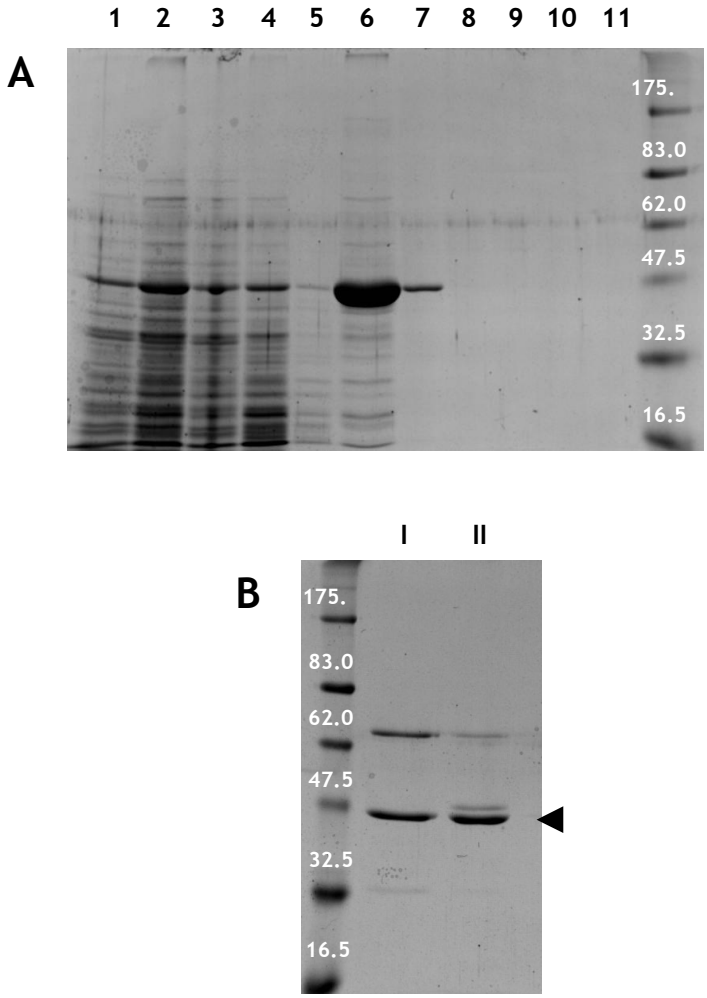

Figure I-1 : Purification de la protéine *Pfu*Cdc6/Orc1 en conditions dénaturantes. (A) La protéine *Pfu*Cdc6/Orc1 fusionnée à une étiquette hexahistidine est surproduite dans des cellules recombinantes de levure (voir Annexes, Protocole 1). Après disruption des cellules, les corps d'inclusion sont recueillis par centrifugation et resuspendus dans un tampon dénaturant. Ensuite, la protéine est purifiée en conditions dénaturantes par chromatographie d'affinité (voir Annexes, Protocole 2). Dépôts : 1. Matériel déposé sur la résine ; 2. Matériel non retenu 1 ; 3. Matériel non retenu 2 (après chargement du matériel non retenu, voir protocole) ; 4. Lavage en présence de 5 mM imidazole ; 5. Lavage en présence de 20 mM imidazole ; 6 à 11 : Elution en présence de 300 mM imidazole. (B) La protéine purifiée est renaturée par la méthode de dilution rapide avant d'être concentrée par chromatographie d'affinité. Puis, l'étiquette hexahistidine est clivée *in situ* par action de la thrombine et les produits de la réaction sont recueillis. La position de la protéine sans étiquette est indiquée par la tête de flèche. La bande située légèrement au dessus correspond à un reste de protéine avec étiquette. La bande du haut correspond à la thrombine. I : Fraction recueillie après retrait du bouchon obturant la colonne ; II : Fraction recueillie après élution en présence de 300 mM imidazole. Le marqueur moléculaire utilisé est le marqueur protéique pré-coloré (6-175 kDa) de chez New England Biolabs.

est donc conforme aux attentes. En revanche, le complexe GINS étant impliqué dans l'initiation de la réplication, la protéine *Pfu*Cdc6/Orc1 aurait pu interagir avec la protéine *Pfu*Gins23. Le fait qu'aucune interaction n'ait été observée entre ces deux protéines est conforme aux résultats d'une analyse par double-hybride qui suggèrent que la protéine *Pfu*Cdc6/Orc1 interagit avec la protéine *Pfu*Gins15 mais pas avec la protéine *Pfu*Gins23 (Yoshimochi et al., 2008). L'existence d'une interaction physique entre les protéines *Pfu*Cdc6/Orc1 et *Pfu*Gins15 n'a pas pu être attestée par la technique de la résonance du plasmon de surface car la protéine Gins15 est insoluble en l'absence de la sous-unité Gins23. Ces résultats suggèrent néanmoins que la protéine *Pfu*Cdc6/Orc1 interagirait spécifiquement seulement avec l'une des deux sous-unités du complexe GINS. Par ailleurs, les analyses préliminaires effectuées avec la protéine *Pfu*Cdc6/Orc1 immobilisée sur une puce suggèrent que la protéine *Pfu*MCM interagit de façon spécifique avec la protéine *Pfu*Cdc6/Orc1 **(Figure I-2A)**. Enfin, les mesures réalisées suggèrent que la protéine *Pfu*Cdc6/Orc1 interagit très faiblement avec elle-même en absence **(Figure I-2B)** ou en présence d'ATP (non montré).

Discussion

La protéine initiatrice de la réplication du chromosome de *P. furiosus Pfu*Cdc6/Orc1 a été choisie comme modèle dans la voie de l'élaboration du premier système de réplication *in vitro* de l'ADN chez les Archaea. Cependant, la protéine *Pfu*Cdc6/Orc1 est majoritairement exprimée sous une forme insoluble chez *E. coli*. Purifier une protéine à partir de la phase insoluble repose sur une étape de dénaturation/renaturation, une opération dont le résultat relativement aléatoire ne garantit pas que la protéine renaturée se trouve dans une configuration semblable à son état natif. En outre, la protéine peut être partiellement inactivée au cours de ce processus. Une incertitude de cet ordre s'accorde *a priori* mal avec l'étude des

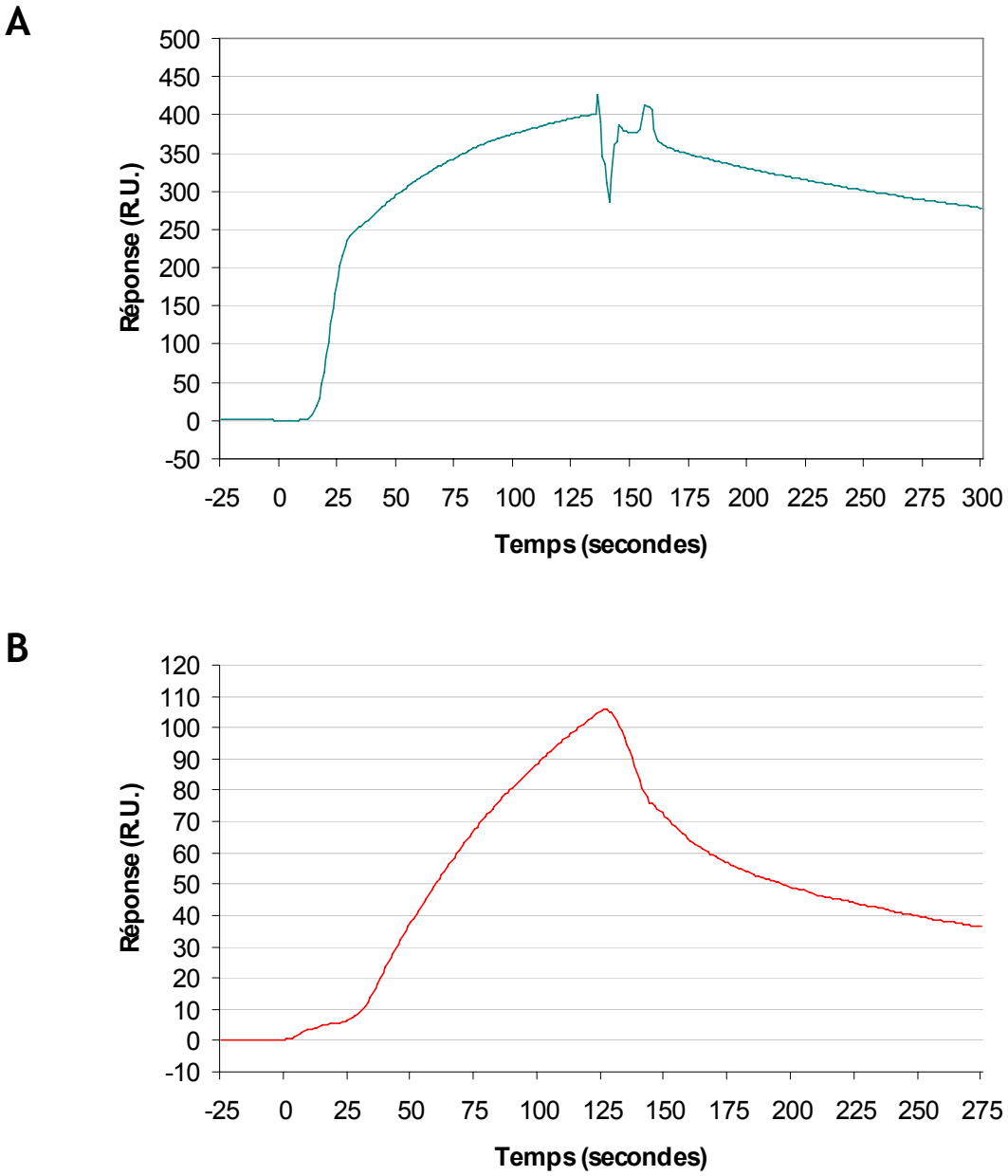

Figure I-2 : Recherche d'interactions physiques par la méthode de mesure de résonance du plasmon de surface. (A) Recherche d'interactions entre la protéine *Pfu*Cdc6/Orc1 et la protéine *Pfu*MCM. (B) Recherche d'interactions de la protéine *Pfu*Cdc6/Orc1 avec elle-même en absence d'ATP. ; le profil obtenu en présence d'ATP est semblable L'injection de l'analyte a lieu au temps t = 0 seconde. La réponse est donnée en unités de résonance (R.U.).

propriétés fonctionnelles de la protéine initiant un mécanisme moléculaire aussi fin et subtil que la réplication de l'ADN. Par conséquent, une grande partie de mon travail a consisté d'une part à optimiser l'expression d'une forme soluble de la protéine *Pfu*Cdc6/Orc1, d'autre part à mettre au point un processus de purification simple et reproductible à partir de cette forme soluble.

La protéine *Pfu*Cdc6/Orc1 est majoritairement exprimée sous une forme insoluble. Les recherches menées dans le but d'optimiser la production d'une forme soluble de la protéine *Pfu*Cdc6/Orc1 ont permis de montrer que la durée d'expression protéique doit être brève car la protéine est toxique pour *E. coli*. Aussi, des quantités substantielles de la forme soluble ont pu être exprimées en respectant ces conditions. Néanmoins, la forme soluble de la protéine *Pfu*Cdc6/Orc1 est particulièrement instable et a tendance à s'agréger ce qui entrave le processus de purification. La formation d'agrégats est une caractéristique que partage la protéine DnaA de la bactérie *E. coli* (Fuller and Kornberg, 1983; Hwang et al., 1990). D'autre part, il a été montré que DnaA interagit avec les phospholipides membranaires (Hwang et al., 1990; Sekimizu et al., 1988; Xia and Dowhan, 1995) ; des observations similaires ont été faites avec le complexe ORC de la levure *Saccharomyces cerevisiae* (Lee et al., 2002). Chacune de ces deux caractéristiques a été considérée, au même titre que l'affinité de la protéine Cdc6/Orc1 pour l'ADN, comme une source possible d'instabilité en solution. Par conséquent, différentes approches, dont une a fait preuve d'efficacité dans le cas de la protéine DnaA, ont été menées pour tenter d'y remédier (Zawilak-Pawlik et al., 2006). Ces tentatives ont toutes été infructueuses ce qui suggère que l'instabilité de la forme soluble de la protéine *Pfu*Cdc6/Orc1 représente un obstacle majeur dans la mise au point d'un processus de purification. En fait, il est possible que la fraction soluble de la protéine soit un mélange entre des espèces correctement repliées et actives et des espèces partiellement ou improprement repliées et donc inactives (Gonzalez-Montalban et al., 2007). Les espèces protéiques mal

repliées pourraient s'agréger autour d'espèces correctement repliées au cours du processus de purification et déstabiliser l'ensemble des espèces solubles. A ce titre, la purification de la protéine *Pfu*Cdc6/Orc1 par la voie dénaturante ne constitue pas une alternative satisfaisante dans la mesure où la protéine purifiée par cette approche est aussi très instable en solution. En revanche, il semble que la protéine purifiée *Pfu*Cdc6/Orc1 purifiée selon cette approche soit compatible avec la recherche de partenaires protéiques puisque deux interactions ont ainsi pu être observées : d'une part l'interaction de la protéine Cdc6/Orc1 avec elle-même, d'autre part l'interaction entre la protéine MCM avec Cdc6/Orc1. D'autre part, il est possible que dans certains cas, le fait qu'aucune interaction n'aie été observée soit une information pertinente qu'il conviendrait d'étayer par des investigations plus poussées.

La première information concerne l'absence d'interaction entre les protéines *Pfu*Cdc6/Orc1 et *Pfu*Gins23. Cette absence d'interaction entre *Pfu*Cdc6/Orc1 et *Pfu*Gins23 mériterait d'être confirmée en réalisant la mesure réciproque — l'interaction entre la protéine *Pfu*Gins23 et la protéine *Pfu*Cdc6/Orc1 immobilisée sur puce —, car il se pourrait que chacune des sous-unités du complexe GINS interagisse avec des partenaires protéiques spécifiques. En effet, lors de la caractérisation des complexes GINS chez *S. solfataricus* et *P. furiosus*, il a été souligné que seule la sous-unité Gins23 interagit spécifiquement avec l'hélicase MCM (Marinsek et al., 2006; Yoshimochi et al., 2008). En outre, les résultats d'une analyse double-hybride suggèrent que la protéine *Pfu*Gins15, mais pas la protéine *Pfu*Gins23, interagit avec la protéine *Pfu*Cdc6/Orc1 (Yoshimochi et al., 2008). L'interaction physique entre les protéines *Pfu*Cdc6/Orc1 et *Pfu*Gins15 n'a malheureusement pas pu être confirmée car cette dernière est insoluble. Aussi, l'ensemble de ces résultats suggère que la sous-unité *Pfu*Gins15 interagit spécifiquement avec la protéine *Pfu*Cdc6/Orc1 alors que la sous-unité *Pfu*Gins23 interagit spécifiquement avec la protéine *Pfu*MCM. Etant donné que le complexe GINS interagit d'une part avec la protéine Cdc6/Orc1, d'autre part avec la protéine MCM, il

représente un candidat sérieux pour le rôle de facteur de chargement de l'hélicase réplicative MCM. Néanmoins, aucun homologue *gins23* n'étant identifiable dans certains génomes d'euryarchées, il est possible que les propriétés fonctionnelles du complexe GINS diffèrent selon les lignées archéennes considérées (voir Chapitre IV).

D'autre part, les mesures réalisées suggèrent que l'hélicase réplicative *Pfu*MCM interagit avec la protéine initiatrice *Pfu*Cdc6/Orc1 fixée sur puce, une observation en accord avec les données obtenues chez d'autres organismes modèles. Celles-ci appuient en effet l'idée selon laquelle la protéine initiatrice pourrait jouer un rôle dans le chargement de l'hélicase (De Felice et al., 2004; Shin et al., 2008) ou en moduler l'activité (Haugland et al., 2006; Shin et al., 2003a). Néanmoins, aucune interaction entre les gènes *cdc6/orc1* et *mcm* de *P. furiosus* n'a été détectée par une analyse en double hybride (Miho Kawashima, Fujihiko Matsunaga, résultats non publiés). En outre, des données préliminaires obtenues chez *P. abyssi* suggèrent que la fixation de la protéine Cdc6/Orc1 à l'origine de réplication n'est pas suffisante pour que la protéine MCM soit chargée. En effet, après traitement à la puromycine la protéine MCM n'est plus associée à l'origine alors que la protéine Cdc6/Orc1 reste en place (Matsunaga et al., 2001). D'autre part, les données obtenues chez *P. furiosus* suggèrent que le complexe GINS pourrait assister le chargement de l'hélicase MCM au niveau de l'origine de réplication marquée par la protéine initiatrice Cdc6/Orc1 (Yoshimochi et al., 2008). De manière intéressante, les données obtenues après traitement à la puromycine appuient aussi cette hypothèse si l'on imagine que la fixation du complexe GINS est compromise dans ces conditions. L'ensemble des données obtenues chez *Pyrococcus* quant à l'interaction de la protéine Cdc6/Orc1 avec la protéine MCM semblent contradictoires. D'autres études devront donc être menées afin de savoir s'il est possible de reconcilier l'ensemble des données en un modèle cohérent.

Conclusions & Perspectives

La route vers l'élaboration d'un système de réplication *in vitro* s'est donc vite obscurcie en raison des obstacles rencontrés pour obtenir une forme soluble stable des protéines initiatrices de la réplication de l'ADN des archées hyperthermophiles choisies comme modèles d'étude. Il est possible qu'une forme soluble et fonctionnelle de la protéine *Pfu*Cdc6/Orc1 puisse un jour être obtenue si le facteur déterminant de la stabilité de la protéine est élucidé. La co-expression de la protéine *Pfu*Cdc6/Orc1 avec un partenaire protéique d'interaction permettrait peut-être d'accroître la stabilité de *Pfu*Cdc6/Orc1 en solution. La liste des candidats possibles comprend entre autres la protéine Gins15 ou le complexe GINS, la sous-unité DP1 ou la PolD, le complexe MCM. Cependant, les difficultés techniques rencontrées pour obtenir une forme soluble et stable amènent à se demander si l'utilisation d'un système hétérologue pour l'expression *in vivo* de protéines codées par des gènes d'archées hyperthermophiles est pertinente dans le cas de protéines qui occupent, de part leur fonction, une place centrale dans le cycle cellulaire, car elles interfèrent probablement avec la machinerie de traitement de l'information de l'hôte d'expression.

Plusieurs alternatives au système hétérologue d'expression *in vivo* sont envisageables. La première consisterait à purifier la protéine *Pfu*Cdc6/Orc1 à partir d'un culot de cellules de *P. furiosus*. Cette approche est concevable dans la mesure où la protéine Cdc6/Orc1 est très abondante dans la cellule, comme l'indiquent des analyses quantitatives réalisées chez *P. abyssi* (Matsunaga et al., 2001). Néanmoins, il est possible que la contamination par les acides nucléiques ou les phospholipides membranaires soit une source de complications. La seconde alternative consisterait à recourir à un vecteur d'expression récemment développé pour la synthèse de protéines *in vivo* chez l'euryarchée hyperthermophile *Thermococcus kodakaraensis* (Santangelo et al., 2008). Le gène codant la protéine *Pfu*Cdc6/Orc1 pourrait être placé sous le contrôle soit d'un promoteur inductible fort — mais en cas d'expression

trop abondante la cellule pourrait mourir —, soit du promoteur naturel pour éviter l'effet cytotoxique. La troisième alternative, plus radicale et plus coûteuse, serait de synthétiser la protéine par traduction *in vitro* en utilisant un extrait acellulaire bactérien ou eucaryote (Endo and Sawasaki, 2006). Néanmoins, il n'est pas possible d'effectuer une traduction à haute température à partir de tels systèmes, circonstance qui pourrait être nécessaire pour que la protéine acquière une configuration native en l'absence de protéines chaperons. Cette condition pourrait être satisfaite en employant un extrait acellulaire de *T. kodakaraensis* mais cette approche ne permet pas de produire des quantités importantes de protéine (Endoh et al., 2008).

Une orientation moins périlleuse consisterait à se baser sur une protéine purifiée en conditions dénaturantes en faisant abstraction de l'incertitude concernant la conformation et l'intégrité fonctionnelle de la protéine à l'issue de la renaturation. En effet, bien que délicate, cette approche a permis, dans le cas de la protéine DnaA, d'obtenir une forme purifiée active (Sekimizu et al., 1988). Néanmoins, des expériences préliminaires menées dans le laboratoire du professeur Ishino avec la protéine *Pfu*Cdc6/Orc1 purifiée à partir de la fraction insoluble (*Pfu*Cdc6/Orc1_DC) suggèrent que la méthode de purification utilisée (voir Matériels et Méthodes) n'est pas pleinement satisfaisante. Premièrement, la protéine est fortement instable en solution et semble s'agréger de manière dépendante de la température (quelques dizaines d'heures à 4°C, quelques jours à 25°C). Ces agrégats peuvent être résorbés en chauffant une dizaine de minutes à 80°C mais ces observations soulignent l'instabilité de la protéine, un paramètre qui demeure non résolu jusqu'à aujourd'hui dans le cas de la protéine *Pfu*Cdc6/Orc1. Deuxièmement, des expériences de fixation à l'ADN montrent que la protéine *Pfu*Cdc6/Orc1_DC reconnaît spécifiquement les motifs ORB (Origin Recognition Box) à l'échelle d'un oligonucléotide mais perd sa spécificité lorsque le motif ORB se trouve au sein d'un fragment de 500 paires de bases (Fujihiko Matsunaga, communication personnelle). Ce

résultat est difficile à interpréter car il peut traduire le fait que la protéine *Pfu*Cdc6/Orc1 est partiellement inactivée car elle a été imparfaitement renaturée ou bien signifier que la protéine Cdc6/Orc1 n'est pas capable de reconnaître seule le motif ORB lorsque celui-ci se trouve au sein d'une séquence nucléotidique de l'ordre de quelques centaines de paires de bases. En revanche, des expériences d'immunoprécipitation de la chromatine couplée à des analyses sur puce d'ADN montrent que la protéine Cdc6/Orc1 affiche *in vivo* une préférence marquée pour l'origine de réplication à l'échelle du génome (Matsunaga et al., 2007).

A défaut de pouvoir assembler dans l'immédiat les protéines dans le tube à essai, la recherche systématique d'interactions peut apparaître comme une approche complémentaire intéressante pour mieux délimiter le domaine d'activité de la protéine Cdc6/Orc1. Dans cette optique, la recherche de partenaires protéiques pourrait être réalisée grâce à la technique de la résonance du plasmon de surface, mais cette approche nécessite de disposer d'une protéine purifiée donc de dresser une liste de ligands potentiels. La recherche des partenaires de la protéine Cdc6/Orc1 pourrait aussi se faire à l'aveugle directement à partir d'extraits cellulaires de *P. furiosus*, en procédant par exemple à une co-purification ou à une co-immunoprécipitation couplées à une analyse par spectrométrie de masse. Le criblage d'une banque génomique par double-hybride en utilisant la protéine Cdc6/Orc1 comme appât est également une option envisageable, mais cette technique engendre beaucoup de faux positifs. Néanmoins, ces méthodes à l'aveugle présentent l'avantage de s'affranchir de la part de subjectivité qui anime certains projets scientifiques et de laisser la place à l'imprévu, imprévu qui peut se matérialiser sous la forme de données inattendues qui bousculent certaines idées, mettent à jour certaines facettes inexplorées d'un mécanisme, révèlent des connections cellulaires insoupçonnées.

Le chapitre suivant aborde les résultats d'une analyse comparative des génomes d'Archaea centrée sur l'étude du contexte génomique des gènes de la réplication. Cette étude

avait pour objectif premier la recherche d'associations conservées entre des gènes de la réplication à partir desquelles de nouvelles interactions fonctionnelles entre constituants de la machinerie de réplication pourraient être prédites (Dandekar et al., 1998; Galperin and Koonin, 2000; Marcotte, 2000). De manière inattendue, cette étude nous a permis de révéler que certains gènes de la réplication sont associés de manière récurrente à des gènes de la traduction en une structure de type opéron, ce qui suggère que ces deux processus font l'objet d'une co-régulation.

Chapitre II

Analyse comparative du contexte génomique des gènes de la réplication dans les génomes d'Archaea

Article publié dans *Genome Biology* en 2008
Genomic context analysis in Archaea suggest previously unrecognized links between DNA replication and translation

Manuscrit soumis au journal Cell
When DNA replication and protein synthesis come together

Chapitre II : Analyse comparative du contexte génomique des gènes de la réplication dans les génomes d'Archaea

Chez les Archaea, la majorité des gènes de la réplication ont été identifiés par homologie avec les gènes codant les protéines impliquées dans la réplication de l'ADN chez les eucaryotes (voir Introduction). Globalement, la machinerie de réplication chez les archées actuelles ressemble à une version simplifiée de la machinerie de réplication de la plupart des eucaryotes (Lao-Sirieix et al., 2007) ; comparable à celle des organismes eucaryotes ayant adopté un mode de vie parasitaire (Morrison et al., 2007). Néanmoins, rien n'indique que l'ensemble des gènes de la réplication ont été identifiés dans les génomes archéens et eucaryotes, comme en témoigne la récente caractérisation du complexe GINS chez les eucaryotes puis les archées (Kanemaki et al., 2003; Kubota et al., 2003; Makarova and Koonin, 2005; Makarova et al., 2005; Marinsek et al., 2006; Takayama et al., 2003). En outre, certains gènes de la réplication sont inidentifiables dans certains génomes d'Archaea (Fitz-Gibbon et al., 2002; Slesarev et al., 2002; White, 2003; Yamashiro et al., 2006). Premièrement, certains gènes de fonction inconnue présents dans des génomes d'Archaea pourraient se révéler être des homologues distants de certains gènes de réplication eucaryotes (Robinson and Bell, 2007). Deuxièmement, il est possible que de nouvelles protéines de la réplication, exclusives des Archaea, soient découvertes à l'avenir : soit qu'ils s'agissent d'innovations archéennes — à l'image de la PolD (Ishino et al., 1998; Uemori et al., 1997) —, soit que le gène correspondant ait été perdu ou remplacé par un gène non orthologue dans la lignée conduisant aux eucaryotes après la divergence entre Archaea et Eucarya (Koonin et al., 1996).

La génomique comparative s'est imposée comme une approche précieuse — complémentaire des méthodes basées sur la recherche d'homologie — pour l'annotation des génomes et pour approfondir notre connaissance de l'évolution et de la physiologie des organismes, y compris chez les Archaea (pour des revues, voir (Ettema et al., 2005; Makarova and Koonin, 2003a, 2005)). En ce qui concerne la réplication, la génomique comparative a notamment permis de conduire à l'identification d'une nouvelle classe de thymidylate synthase (Myllykallio et al., 2002), des homologues archéens des protéines Gins (Makarova and Koonin, 2005; Makarova et al., 2005) et d'un homologue probable de la protéine eucaryote Cdt1 (Robinson and Bell, 2007). Par ailleurs, des associations entre gènes de la réplication ont déjà été décrites de manière ponctuelle (Myllykallio et al., 2000; Robinson et al., 2004) mais rarement de façon systématique. Aussi, nous avons décidé d'analyser le contexte génomique de tous les gènes de la réplication afin de dégager d'éventuelles tendances quant à la disposition réciproque de ces gènes au sein des génomes d'Archaea. La présence d'associations conservées de gènes impliquant des gènes de la réplication devaient permettre de :

1) prédire de nouvelles interactions fonctionnelles entre protéines de la réplication lorsque les gènes correspondants sont contigus dans de nombreux génomes ;
2) identifier de probables nouvelles protéines de la réplication sous la forme de gènes de fonction inconnue situés de manière récurrente à côté d'un gène de réplication.

Le chapitre qui s'ouvre dans les pages suivantes s'articulera autour des éléments développés dans un article scientifique et une courte revue :

- L'article scientifique concerne les résultats de l'analyse comparative des génomes d'Archaea centrée sur l'étude du contexte génomique des gènes de la réplication. Cette analyse nous a permis d'identifier des associations conservées entre gènes de la

réplication à partir desquelles nous avons inféré de nouvelles connections fonctionnelles au sein du réplisome. D'autre part, les résultats de ce travail suggèrent qu'il existe des interactions fonctionnelles entre la réplication et d'autres processus cellulaires de traitement de l'information génétique. En particulier, cette approche ciblée nous a permis de mettre en évidence une association conservée entre des gènes de la réplication et des gènes liés au ribosome ou à la traduction qui n'est pas répertoriée dans la base de données STRING (von Mering et al., 2007) — une base de données dédiée à la recherche d'interactions entre gènes ou protéines. Ce regroupement de gènes en une structure de type opéron suggère selon nous que la réplication de l'ADN et la synthèse des protéines pourraient faire l'objet d'une co-régulation.

- La courte revue vient dans le prolongement de l'article scientifique. Dans cette revue, nous développons et mettons en perspective les interprétations esquissées dans la discussion de l'article à la lumière des connaissances actuelles concernant le couplage fonctionnel entre la réplication et la traduction. En particulier, nous dressons un panorama succinct des travaux réalisés chez les bactéries sur le rôle du (p)ppGpp dans la régulation coordonnée entre la réplication et la traduction. Les études concernant le rôle des petites protéines fixant le GTP de la famille Obg sont également évoquées. Puis, nous présentons les quelques données expérimentales suggérant l'existence d'un couplage fonctionnel entre réplication, traduction et biogenèse du ribosome chez les eucaryotes. Enfin, nous mettons en perspective les résultats et les interprétations issues de l'analyse comparative des génomes d'Archaea à la lumière des données bibliographiques existantes. Pour conclure, nous proposons un modèle pour un système permettant de réguler de façon dynamique la réplication et la traduction chez les Archaea et les Eucarya. Ce modèle se base principalement sur les associations de gènes mises en évidence par l'analyse comparative des génomes d'Archaea.

Les approches de génomique comparative n'ont qu'une valeur prédictive et les hypothèses formulées dans le cadre de cette étude du contexte génomique demandent à être étayées par des données expérimentales. La pertinence biologique de deux des interactions prédites par l'analyse du contexte génomique a été mise à l'épreuve en réalisant des expériences *in vitro* ; celles-ci seront décrites et discutées dans le chapitre suivant.

Au cours de la phase préparatoire de l'analyse du contexte génomique, un inventaire de l'ensemble des gènes codant des protéines de la réplication dans les génomes d'Archaea entièrement séquencés a été dressé, puis mis à jour au gré de la mise à disposition de nouvelles données. L'analyse de la distribution phylétique de ces gènes au sein du domaine Archaea donne une idée approximative de l'histoire évolutive de ces gènes (duplication, perte). Cette analyse sommaire nous a néanmoins permis de reconstruire la composition protéique probable de la machinerie de réplication du dernier ancêtre commun des Archaea. Les résultats de cette analyse seront abordés dans le chapitre IV.

Article

Genomic context analysis in Archaea suggest previously unrecognized links between DNA replication and translation

Jonathan Berthon, Diego Cortez, Patrick Forterre

Genome biology 9(4), R71 (2008)

Research

Open Access

Genomic context analysis in Archaea suggests previously unrecognized links between DNA replication and translation

Jonathan Berthon*†, Diego Cortez‡ and Patrick Forterre*‡

Addresses: *Univ. Paris-Sud 11, CNRS, UMR8621, Institut de Génétique et Microbiologie, 91405 Orsay CEDEX, France. †Laboratory of Protein Chemistry and Engineering, Department of Genetic Resources Technology, Faculty of Agriculture, Kyushu University, 6-10-1 Hakozaki, Higashi-ku, Fukuoka-shi, Fukuoka 812-8581, Japan. ‡Institut Pasteur, rue Dr. Roux, 75724 Paris CEDEX 15, France.

Correspondence: Jonathan Berthon. Email: jonathan.berthon@igmors.u-psud.fr. Patrick Forterre. Email: patrick.forterre@igmors.u-psud.fr

Published: 9 April 2008

Genome ***Biology*** 2008, **9:**R71 (doi:10.1186/gb-2008-9-4-r71)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2008/9/4/R71

Received: 21 December 2007
Revised: 22 February 2008
Accepted: 9 April 2008

Abstract

Background: Comparative analysis of genomes is valuable to explore evolution of genomes, deduce gene functions, or predict functional linking between proteins. Here, we have systematically analyzed the genomic environment of all known DNA replication genes in 27 archaeal genomes to infer new connections for DNA replication proteins from conserved genomic associations.

Results: Two distinct sets of DNA replication genes frequently co-localize in archaeal genomes: the first includes the genes for PCNA, the small subunit of the DNA primase (PriS), and Gins15; the second comprises the genes for MCM and Gins23. Other genomic associations of genes encoding proteins involved in informational processes that may be functionally relevant at the cellular level have also been noted; in particular, the association between the genes for PCNA, transcription factor S, and NudF. Surprisingly, a conserved cluster of genes coding for proteins involved in translation or ribosome biogenesis (S27E, L44E, aIF-2 alpha, Nop10) is almost systematically contiguous to the group of genes coding for PCNA, PriS, and Gins15. The functional relevance of this cluster encoding proteins conserved in Archaea and Eukarya is strongly supported by statistical analysis. Interestingly, the gene encoding the S27E protein, also known as metallopanstimulin 1 (MPS-1) in human, is overexpressed in multiple cancer cell lines.

Conclusion: Our genome context analysis suggests specific functional interactions for proteins involved in DNA replication between each other or with proteins involved in DNA repair or transcription. Furthermore, it suggests a previously unrecognized regulatory network coupling DNA replication and translation in Archaea that may also exist in Eukarya.

Background

Alignment of prokaryotic genomes revealed that synteny is globally weak, indicating that bacterial and archaeal chromosomes experience continuous remodeling [1-3]. A few operons encoding physically interacting proteins involved in fundamental processes have been preserved between Archaea and Bacteria in the course of evolution (for example, operons encoding ribosomal proteins, RNA polymerase subunits, or ATP synthase subunits) [1-3]. Most gene strings are only conserved in closely related genomes or exhibit a patchy distribution among genomes in one large group of organisms (for example, in Archaea). Therefore, gene associations that are

conserved between distantly related organisms should confer some selective advantage. The co-localization of a particular group of genes may optimize their co-regulation at the transcriptional level [4,5] or facilitate the assembly of their products in large protein complexes [6]. A corollary of this statement is that characterization of evolutionarily conserved gene clusters can be used to infer functional linkage of proteins (that is, physical interaction or participation in a common structural complex, metabolic pathway, or biological process). Various comparative genomics methods that exploit gene context are commonly used. These approaches analyze protein and domain fusion or gene neighborhood (groups of genes found in putative operons or divergently transcribed gene pairs) to predict functions for, and interactions between, the encoded proteins (reviewed in [2,7-10]). A dramatic example of a discovery based on genome context analysis is the identification in Archaea and Bacteria of proteins associated with the specific DNA repeats known as CRISPR [11]. These *cas* proteins (for CRISPR associated proteins), which were first proposed to be members of a putative DNA repair system [12], are probable actors in a nucleic-acid based 'immunity' system [13]. Comparative analysis of genomes has been especially helpful in Archaea for functional prediction of uncharacterized proteins in the absence of genetic studies (reviewed in [14,15]). For instance, this strategy has allowed the computational prediction and subsequent experimental confirmation of the archaeal exosome [16,17] and of novel proteins associated with the Mre11/Rad50 complex [18,19].

Many putative DNA replication proteins have been identified in archaeal genomes by similarities with their eukaryotic counterparts known experimentally to be involved in DNA replication (for a review, see [20]). Most of these proteins have now been purified from one or more Archaea and characterized to various extents *in vitro* (reviewed in [20]). Several examples of physical and/or functional interactions between archaeal DNA replication proteins have now emerged from biochemical studies (reviewed in [20]), supporting the idea that these proteins are indeed working together at the replication fork. A few clusters of genes encoding DNA replication proteins have been previously reported in *Pyrococcus* and *Sulfolobus* genomes [21-24]; in one case, the gene association correlates with protein physical interaction [24]. This suggests that systematic identification of clusters of genes encoding DNA replication proteins in the expanding collection of archaeal genomes could identify gene associations connecting genome organization to functional interactions of proteins that could be relevant *in vivo*. More importantly, comparative genomic analyses could be used to determine the most significant interactions, that is, those that appear to be recurrent in the genomes of evolutionarily diverse Archaea.

Here, we have performed a systematic genome context analysis of genes encoding DNA replication proteins in 27 completely sequenced archaeal genomes. Our results show that a subset of genes encoding DNA replication proteins often co-localize, that is, these genes are arranged in operon-like structures (contiguous or adjacent genes in the same transcriptional orientation) that are preserved between distant lineages (as for the majority of the cases discussed here), or they lie in a common chromosomal region less than 5 kilobases away from each other. Some of these associations are conserved between distant lineages, indicating that they reflect a functional and possibly a physical interaction between the gene products. In particular, we identified two conserved genomic associations of DNA replication genes that suggest a functional connection between the PCNA, the DNA primase and the MCM helicase via the GINS complex. We also observed that the gene for PCNA is linked to the gene coding for the transcription factor S (TFS) in 12 out of the 27 analyzed genomes, as well as to a gene encoding the ADP-ribose pyrophosphatase NudF in 8 genomes, pointing toward the existence of cross-talk between DNA replication, DNA repair, and transcription. In addition, we noticed that the gene encoding the initiator protein Cdc6 is usually adjacent to a predicted origin of replication, sometimes together with or close to the gene coding for the small subunit of DNA polymerase (Pol)D (DP1) in euryarchaeal genomes, suggesting that PolD may be recruited by Cdc6 at the origin of replication. Moreover, some proteins without clear functional assignments (an oligonucleotide/oligosaccharide-binding (OB)-fold containing protein, a recently described new GTPase, DnaG) are encoded by genes that co-localize with DNA replication genes, suggesting that they may be involved in DNA transaction processes. Surprisingly, our analysis also reveals a widely conserved clustering of a particular set of genes coding for DNA replication proteins (Gins15, PCNA and/or the DNA primase small subunit (PriS)) with a special set of genes encoding proteins related to the ribosome (L44E, S27E, aIF-2 alpha, Nop10). This cluster is strongly supported by a statistical analysis based on the actual distribution of gene clusters in the set of genomes analyzed in this study, suggesting the existence of a previously unrecognized regulatory network coupling DNA replication and translation in Archaea.

Results and discussion

Systematic identification of DNA replication genes in archaeal genomes

We have performed an exhaustive search of all known putative DNA replication genes in the 27 archaeal genomes available at the NCBI [25] as of 10 April 2006. These genomes include 5 genomes of Crenarchaea and 22 genomes of Euryarchaea, and are distributed among 13 different archaeal orders (Figure 1). Our list of DNA replication genes includes all genes coding for archaeal proteins or subunits of complexes corresponding to eukaryotic homologs known to be involved in DNA replication: the initiation factor Cdc6 (Orc1); PolB; the helicase MCM; the sliding clamp PCNA; the clamp-loader replication factor C (RFC); the DNA primase; the single-

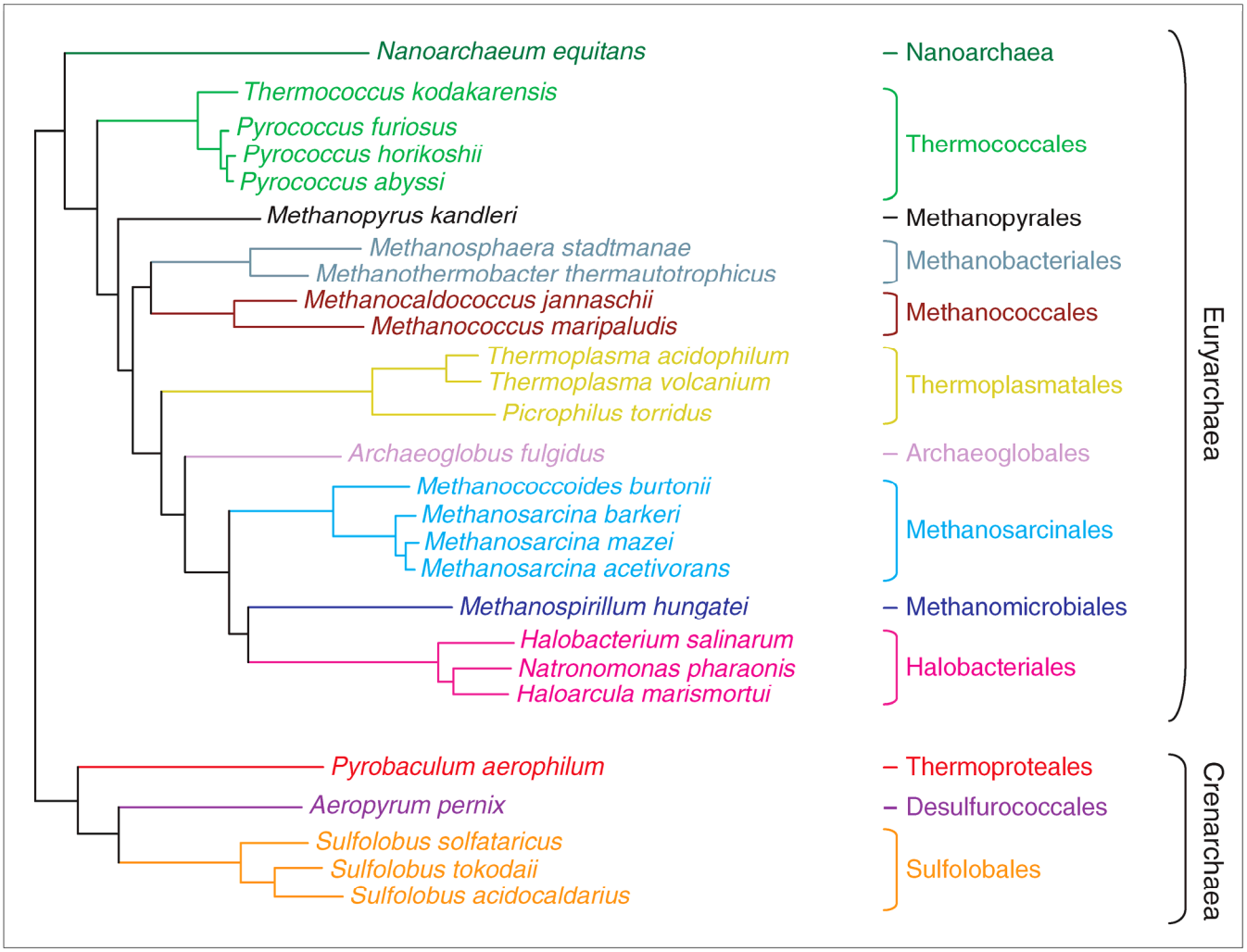

Figure 1
Phylogeny of the Archaea whose genomes have been analyzed in this study. This unrooted tree (kindly provided by Céline Brochier) is based on the concatenation of archaeal ribosomal proteins (see [73] for details). The parasitic archaeon *N. equitans* is placed with Euryarchaeota in accordance with the hypothesis that it likely represents a fast-evolving euryarchaeal lineage [34].

stranded binding protein RPA (or SSB in Crenarchaea); the DNA ligase; the RNase HII; the flap endonuclease FEN-1; and the two Gins subunits (Gins15 and Gins23). We have added to this list PolD (absent from hyperthermophilic Crenarchaea), since its genes are located close to the replication origin in Thermococcales [22] and because this enzyme is essential for *Halobacterium* sp. NRC-1 survival according to recent genetic data [26]. We have also included in our list the DNA topoisomerase VI (Topo VI) since this enzyme is the only DNA topoisomerase known in Archaea that can relax positive superturns, an essential function for DNA replication [27]. First, the 27 archaeal genomes available at the NCBI were searched to retrieve the entries of all the annotated DNA replication proteins (see Materials and methods) encoded by these genomes. Then, systematic BLASTP searches were carried out with several seeds for each protein in order to verify the annotations and to look for missing proteins (see Materials and methods); Additional data file 1 provides a table listing all putative DNA replication proteins identified and used in our analysis.

DNA replication proteins are encoded by a set of genes that is present in all archaeal genomes (sometimes with several paralogues), with the exception of PolD, which is absent in hyperthermophilic Crenarchaea; Gins23, which has only been detected in Crenarchaea and Thermococcales; RPA, which is absent in hyperthermophilic Crenarchaea; and the crenarchaeal SSB, which is currently restricted to Crenarchaea and Thermoplasmatales. We noticed a few interesting instances of missing DNA replication genes. In particular, we and others failed to detect a RPA or a SSB homolog in *Pyrobaculum aerophilum* [28,29] and this study) and a Cdc6/Orc1 homolog in *Methanopyrus kandleri* ([30,31] and this study). On the other hand, we retrieved a Cdc6-like homolog that is related to the putative origin initiator protein of *Methanocaldococcus jannaschii* [32] in the genome of *Methanococcus*

maripaludis. Moreover, we detected only one primase gene in *Nanoarchaeum equitans*; alignment of the amino acid sequence of *N. equitans* primase with other members of the archaeo-eukaryotic primase superfamily shows that it corresponds to the fusion of the amino-terminal region of the small subunit with the carboxy-terminal region of the large subunit [33]. Thus, the primase of *N. equitans* could be an interesting model to study the mechanism of action of this protein *in vitro*. Finally, the genome of *Methanococcoides burtonii* does not harbor any identifiable gene encoding the small non-catalytic subunit of PolD (DP1), whilst the gene encoding the large catalytic subunit (DP2) is present. It would be of particular interest to get insight into the functional properties of the *M. burtonii* PolD to unravel whether or not a core version of PolD exhibits the expected features, given that the interaction between the two subunits has been shown to be essential for full enzymatic activities of the canonical form [21].

Genes encoding subunits of heteromultimeric DNA replication proteins rarely associate

Several DNA replication factors are formed by the association of two or more different protein subunits (that is, these DNA replication factors are heteromultimeric proteins), including RFC (RFC-s and RFC-l), primase (PriS and PriL), the PolD holoenzyme (DP1 and DP2), and Topo VI (A and B subunits). We did not detect any obvious trend of association for the genes encoding different subunits of heteromultimeric proteins among archaeal genomes, except for the genes encoding the Topo VI subunits and the genes for the RFC subunits. The genes encoding the two subunits of Topo VI are contiguous in all Archaea, except for *N. equitans*, Methanococcales, *Archaeoglobus fulgidus* and *Methanopyrus kandleri*, whereas the genes encoding the large and small subunits of RFC co-localize in Crenarchaea, Thermococcales, Methanobacteriales and *M. kandleri* (see Additional data file 2 for illustrations). Interestingly, the genes encoding the two subunits of Topo VI are contiguous to the genes encoding the two subunits of DNA gyrase (of bacterial origin) in all halophilic Archaea and in Methanosarcinales, suggesting a co-regulation of the two type II DNA topoisomerases that was selected after the transfer of the bacterial enzyme into its archaeal host. The genes encoding the two subunits of PolD are adjacent in Thermococcales only, and those for the two subunits of DNA primase co-localize in Thermococcales and Methanobacteriales; the primase genes are fused in *N. equitans* as previously mentioned (Additional data file 2). The genes encoding the three subunits of the heterotrimeric RPA found in Thermococcales (RPA41, RPA32, and RPA14) are clustered in the four completely sequenced genomes presently known, whereas the genes encoding RPA homologs present in other euryarchaeal genomes never associate. Finally, the genes encoding the two Gins proteins in Crenarchaea and Thermococcales are never adjacent. The tendency for genes encoding different subunits of DNA replication factors to co-localize is, therefore, very different from one gene to the other, a first indication that the observed gene associations are not random.

In the course of this work, we noticed that co-localization of DNA replication genes - encoding different subunits of heteromultimeric proteins (see above) or encoding different proteins (see below) - are more frequent in some genomes than in others. They are especially rare in *N. equitans* since all the gene strings that are conserved in all other archaeal genomes are disrupted in this archaeon. It is likely that these disruptions are due to extensive genome rearrangements that occurred in this species because *N. equitans* is a parasitic organism that has adapted to its lifestyle by extensive genome reduction, including the split of several genes [15,34]. At the other end of the spectrum, we observed that the clustering of DNA replication genes occurs very frequently in Thermococcales. Indeed, all genes encoding different subunits of heteromultimeric DNA replication proteins are contiguous in this lineage, except those encoding the two subunits of the archaeal GINS complex.

Conserved gene clusters suggest functional linkage between PCNA, DNA primase, GINS, and MCM

Since DNA replication proteins should interact physically and/or functionally in the replication factory, one can expect that genes encoding different DNA replication proteins sometimes co-localize in archaeal genomes, as a blueprint for these interactions. Such DNA replication islands were previously observed in the vicinity of the *Pyrococcus abyssi* chromosomal replication origin (*oriC*), where the gene encoding Cdc6 lies together with those encoding DP1, DP2, RFC-s, and RFC-l [22]; and at the *cdc6-2* locus in *Sulfolobus solfataricus*, where the genes encoding RFC-s, RFC-l, Cdc6-2, Gins23, and

Figure 2 *(see following page)*
Conserved genomic context of three DNA replication genes in archaeal genomes. This figure highlights the genome context of three DNA replication genes that recurrently associate with a particular set of genes in archaeal genomes (for a detailed picture of the genome context of all DNA replication genes examined in this study see Additional data file 2). **(a)** The gene encoding Gins15 is linked to the gene coding for PCNA and to the gene for the small subunit of the primase in all crenarchaeal genomes, whereas it is alternatively linked to one of these two genes in most euryarchaeal genomes. **(b)** The gene for the PCNA associates with the genes encoding the small or the large subunit of the DNA primase. It is also frequently linked to the gene encoding TFS and/or to the gene coding for the ADP-ribose pyrophosphatase NudF. **(c)** The gene encoding the MCM helicase is contiguous to the gene for Gins23 and/or to the gene for the beta subunit of the initiation factor aIF-2 in several archaeal genomes. Orthologous genes are indicated in the same color. Each gene is denoted by the name of the protein it encodes (see the key at the bottom). Species or cell lineages that have the same genomic environment are listed and the number of corresponding genomes is given in parentheses. White arrows correspond to additional functionally unrelated genes. Genes are not shown to scale.

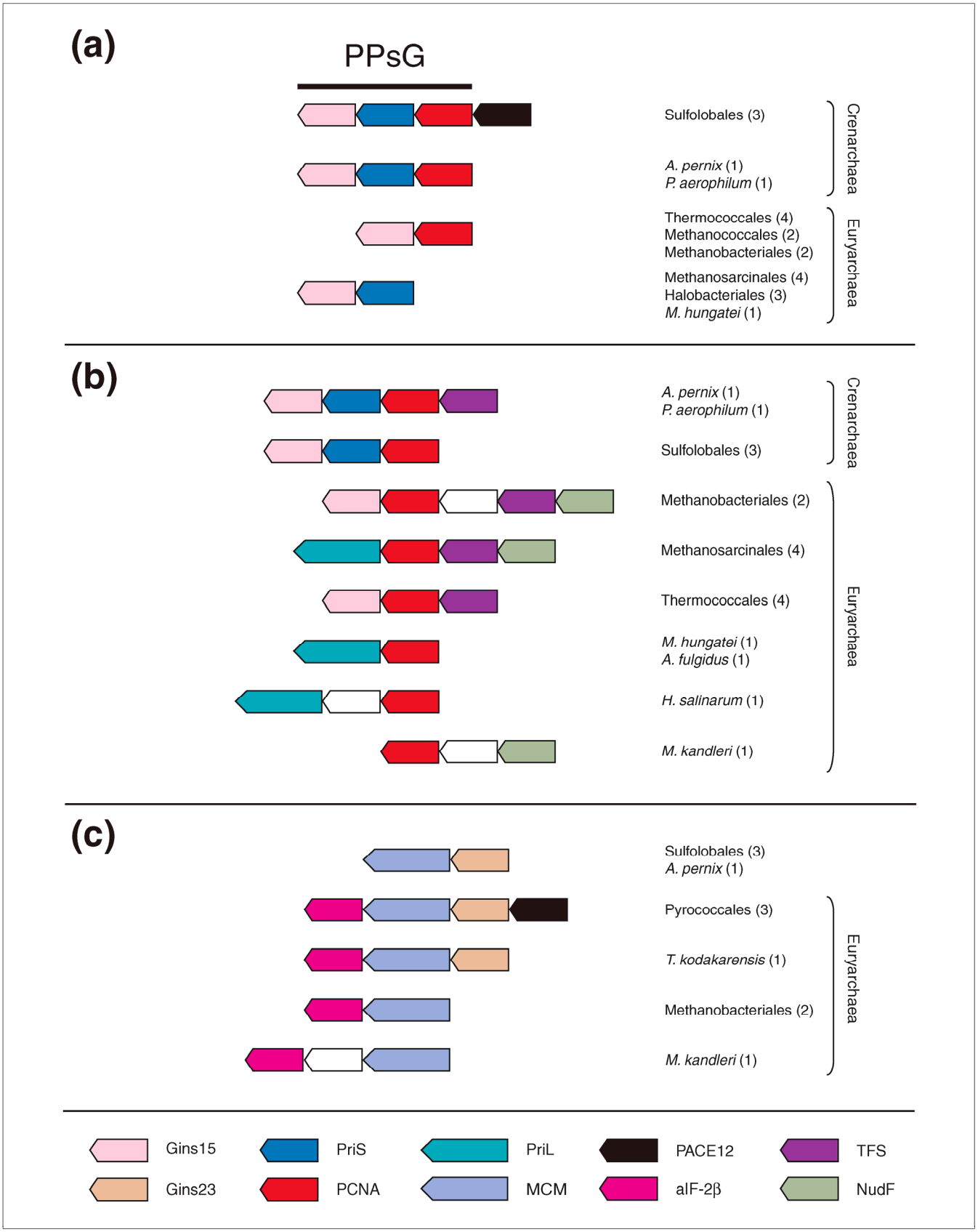

Figure 2 *(see legend on previous page)*

MCM are situated [23,24]. We have detected several new DNA replication islands in our analysis. The association of the genes encoding PCNA, PriS, and Gins15 (hereafter called the PPsG cluster), previously observed by others [14,24], is the most conserved clustering. The full PPsG cluster is not conserved across the entire archaeal domain since the three corresponding genes are adjacent only in crenarchaeal genomes, but the gene encoding Gins15 is contiguous to either the gene for PCNA or the gene for PriS in most euryarchaeal genomes, strongly suggesting that Gins15, PCNA, and PriS functionally associate (Figure 2a). Hence, the genes encoding Gins15 and PCNA are direct neighbors in the four Thermococcales, in two Methanococcales, and in two Methanobacteriales, whereas the genes encoding Gins15 and PriS are adjacent in Methanosarcinales (four species) and in halophilic Archaea (three species). Interestingly, while the gene encoding PCNA is an immediate neighbor of PriS in the PPsG cluster, it co-localizes with the gene encoding the other primase subunit, PriL, in the four Methanosarcinales, in *A. fulgidus*, *Haloarcula marismortui*, and *Halobacterium salinarum* (Figure 2b). In summary, the gene encoding Gins15 is associated with the genes encoding PriS and PCNA (Crenarchaea) or contiguous to one of these two genes (Euryarchaea), whilst the gene coding for PCNA is linked either to the gene encoding PriS (Crenarchaea) or to the gene coding for PriL (Euryarchaea) (Figure 2a,b). This suggests that PCNA could interact with the two primase subunits, whereas Gins15 could interact directly with PCNA and PriS. Finally, the gene encoding Gins23, which has been detected only in Crenarchaea and Thermococcales, neighbors the gene encoding MCM in all these Archaea, except in *P. aerophilum* (Figure 2c).

Altogether, these observations suggest the existence of a core of DNA replication factors, including the PCNA clamp, the DNA primase, the GINS complex, and the helicase MCM, that should be tightly associated with the replication factory during the elongation step of DNA replication. Bell and colleagues [24] have demonstrated by two-hybrid analysis in yeast and immunoprecipitation that the two *Sulfolobus* Gins proteins indeed form a complex that interacts with MCM and the two subunits of the DNA primase. They have suggested that this complex could provide a mechanism to couple the progression of the MCM helicase on the leading strand with priming events on the lagging strand [24]. Our genome context analysis further suggests that PCNA could interact with the GINS complex (via Gins15) and with each of the two subunits of the DNA primase. However, no interaction between PCNA and any of the Gins subunits has been detected by Bell and colleagues [24]. Similarly, no interaction between PCNA and the DNA primase has ever been reported in Archaea, despite the recurrent association of their genes in archaeal genomes. But, it should be noted that the gene for PCNA and the gene for PriS are probably co-transcribed [35], thus strengthening our predictions.

A specific link between PCNA and DNA primase

We noticed that the gene encoding PCNA is often associated with one or two of the genes coding for the subunits of the DNA primase. This linking is especially conserved since it occurs both in the PPsG cluster and in additional contexts. Hence, the gene for PCNA is adjacent to the gene encoding the large subunit of the DNA primase in *A. fulgidus*, *M. hungatei*, *H. salinarum*, *H. marismortui*, and Methanosarcinales (Figure 2b). Besides the likely association of these two factors at the replication fork, an interesting hypothesis is that it could also reflect the involvement of the archaeal primase in DNA repair, since the PCNA clamp is an accessory factor of many DNA repair proteins. It has been previously suggested that archaeal DNA primase may be involved in DNA repair processes as a translesion DNA polymerase, since most archaeal genomes lack genes encoding DNA polymerases of the X or Y families, which are the major translesion DNA polymerases in bacteria or eukaryotes [36]. The DNA primases from *Pyrococcus furiosus* and *S. solfataricus* are indeed able to synthesize DNA strands *in vitro* (reviewed in [36]) and a translesion synthesis activity has been recently detected in fractions containing the DNA primase in partially purified *P. furiosus* cell extracts [37]. Finally, the catalytic site of the archaeal primase exhibits some structural similarities with the repair DNA polymerase of the X family (reviewed in [36]). Therefore, it is tempting to speculate that PCNA contacts the DNA primase during DNA repair transactions and that the genomic association highlighted in this work is functionally relevant.

Interactions between DNA replication and DNA repair

In the course of this analysis, we detected many genomic associations of DNA replication genes with genes coding for archaeal homologs of DNA repair/recombination proteins from Eukarya (XPF, RadA, RadB, Mre11, Rad50) or from Bacteria (PolX, RecJ, Endo III, Endo IV, Endo V, UvrABC). We also found associations between genes for DNA replication proteins and specific archaeal proteins that have been characterized biochemically and predicted to be involved in the repair of stalled replication forks by recombination/repair (the helicase Hel308a/Hjm, a RecQ analogue; the nuclease/helicase Hef; and the Holliday junction resolvase Hjc). All these observations suggest that several DNA replication proteins are also involved in base excision repair, in nucleotide excision repair, or in the repair of stalled replication forks. They are described and discussed in Additional data file 3.

Functional connection of DNA replication, transcription, and DNA repair processes via the TFS and NudF proteins?

We observed an unexpected conserved association between the genes coding for PCNA and TFS. These two genes are neighbors in both crenarchaeal (*P. aerophilum*, *Aeropyrum pernix*) and euryarchaeal genomes (Thermococcales, Methanobacteriales and Methanosarcinales) (Figure 2b). In *P. aerophilum* and *A. pernix*, the gene coding for TFS is located just upstream of the PPsG cluster, whereas it forms a cluster with

the genes coding for PCNA and Gins15 in Thermococcales and Methanobacteriales, and with those encoding PCNA and PriL in Methanosarcinales (Figure 2b).

In summary, the gene for PCNA is linked to the gene coding for TFS in 12 out of the 27 analyzed genomes. Although, this gene pairing is not supported by statistical analyses since two genes clusters are frequently conserved across genomes (Additional data file 4), it cannot be a chance occurrence (see below in the Statistical analyses section). Furthermore, it is remarkable that these two genes are associated in both crenarchaeal and euryarchaeal genomes representing four different orders. In our opinion, this conservation pattern indicates that this gene pairing is not coincidental, pointing towards the existence of cross-talk between replication and transcription processes and indicating that TFS and PCNA may be part of this connection. The archaeal protein TFS is homologous to the carboxy-terminal domain of the eukaryotic transcription factor TFIIS and to one of the small subunits of the three eukaryotic RNA polymerases [38]. TFS is also a functional analogue of the bacterial GreA/GreB proteins. When an RNA polymerase is blocked by a DNA lesion, all these proteins can activate an intrinsic 3' to 5' RNase activity of the RNA polymerase, allowing degradation of the mRNA and re-initiation of transcription [39]. It has been shown *in vitro* that misincorporation of non-templated nucleotide is reduced in the presence of archaeal TFS and that TFS helps the elongation complex to bypass a variety of obstacles in front of transcription forks [39]. One possibility, suggested by our genome context analysis, is that TFS recruits DNA repair proteins via PCNA when a DNA replication fork encounters a transcription fork blocked by a DNA lesion. In agreement with a direct role of TFS in controlling genome stability, *M. kandleri*, which is the only archaeon lacking TFS, exhibits a high frequency of gene rearrangement (fusion, splitting) and gene capture, whereas its RNA polymerase has evolved more rapidly than other archaeal RNA polymerases [40].

Interestingly, the gene coding for TFS co-localizes in several euryarchaeal genomes with a gene encoding a protein belonging to the Nudix phosphohydrolase superfamily (Nudix stands for Nucleoside diphosphate linked to another moiety, X). Nudix proteins, which are found in the three domains of life, hydrolyze a wide range of organic pyrophosphates, including nucleoside di- and triphosphates, dinucleoside polyphosphate, and nucleotide sugars; some superfamily members have the ability to degrade damaged nucleotides (reviewed in [41]). We noticed that the Nudix hydrolase encoded by the gene that is arranged in tandem with the gene coding for TFS has been characterized as an ADP-ribose pyrophosphatase in *M. jannaschii* [42]. Therefore, we suggest that every Nudix gene that is linked to a TFS gene in archaeal genomes likely encodes a protein with a similar function (hereafter called NudF protein according to the nomenclature found in [41]). The clustering between the genes encoding TFS and NudF was previously noticed by Dandekar and co-workers [2] (the NudF protein is mentioned by the name 'MutT-like' in this article), who proposed a physical interaction between the two proteins using structural modeling data. The genes encoding NudF and TFS co-localize with those encoding PCNA and PriL in Methanosarcinales, and with those encoding PCNA and Gins15 in Methanobacteriales (Figure 2b). Remarkably, in *M. kandleri*, which does not contain any TFS homolog, the gene for NudF co-localizes with the PCNA gene (Figure 2b). All these observations suggest that, together with TFS, NudF could be associated at the replication forks with the core of proteins previously identified through the PPsG cluster. The role of NudF could be to hydrolyze damaged nucleotides, in order to prevent their incorporation by DNA or RNA polymerases. However, considering that NudF is an ADP-ribose pyrophosphatase [42], an attractive alternative hypothesis is that NudF participates in a network of activities that regulate DNA replication/repair via ADP-ribosylation. In eukaryotes, several DNA replication factors, such as PCNA, primase and DNA polymerases, are indeed poly-ADP-ribosylated in response to DNA damage in order to prevent transcription or replication of damaged DNA [43]. Moreover, transient inhibition of DNA replication following DNA damage has been noticed in *P. abyssi* [44]. In Archaea, poly-ADP-ribosylation like reactions have been reported in *S. solfataricus*, and the chromosomal protein Sso7d, which is restricted to Sulfolobales, has been identified as a putative substrate [45]. Interestingly, Sso7d has been recently shown to promote the repair of thymine dimers *in vitro* after photoinduction [46]. If some archaeal proteins involved in DNA replication or transcription are also inhibited by ADP-ribosylation following DNA damage (something that has to be tested), the role of NudF could be, once DNA damage has been repaired, to facilitate replication and/or transcription restart by metabolizing the free ADP-ribose released during degradation of ADP-ribose polymers.

Genomic contexts of the *cdc6* gene suggest specific interactions at the replication origin

Besides the DNA replication genes that belong to the PPsG cluster, the gene that co-localizes more frequently with other DNA replication genes is *cdc6*. Our analysis suggests a loose connection between the initiator protein Cdc6 and the clamp loader RFC, the helicase MCM and DNA polymerases (either B or D), respectively. Hence, the gene encoding Cdc6 is located in the vicinity of the genes encoding RFC-s1 and RFC-l in *P. aerophilum*; RFC-s in *H. salinarum*; MCM and DP2 in *M. maripaludis*; and DP1 in *H. salinarum*, *H. marismortui*, *Methanothermobacter thermautotrophicus*, and *Methanosphaera stadtmanae* (Additional data file 2). Remarkably, all these proteins should be recruited at the replication origin for the initiation of DNA replication. In addition, the genes that are located in the vicinity of the *cdc6* gene in the genomes of *P. aerophilum*, Halobacteria and methanogens correspond to those that form the replication islands of *Pyrococcus* or *Sulfolobus* (Additional data file 2). Since the gene encoding Cdc6 is frequently associated with a predicted replication origin

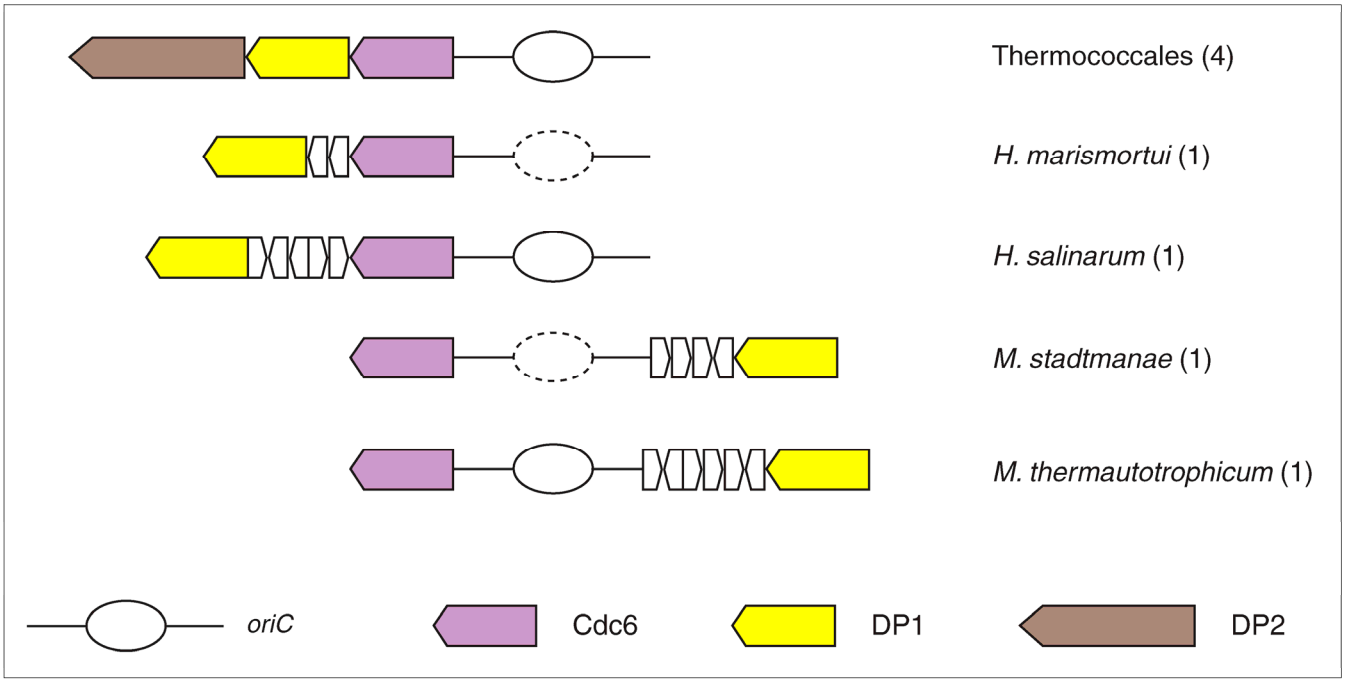

Figure 3
Replication origin is adjacent to *cdc6*, and close to gene for DP1 in several euryarchaeal genomes. Orthologous genes are indicated in the same color. Each gene is denoted by the name of the protein it encodes (see the key at the bottom). The origins of replication (*oriC*) are shown as bubble-shaped replication intermediate sketches; solid lines are used when the origin has been identified experimentally, and broken lines are used when the origin has been predicted with *in silico* analyses. Species or cell lineages that have the same genomic environment are listed and the number of corresponding genomes is given in parentheses. White arrows correspond to additional functionally unrelated genes. Genes are not shown to scale.

[22,23,47], co-localization of the *cdc6* gene with various DNA replication genes in the vicinity of *oriC* could help the recruitment of DNA replication proteins to build new DNA replication factories at the origin of replication. Among the various gene associations of *cdc6* with other DNA replication genes, the most recurrent is the linkage with the gene encoding the small subunit of PolD. First noticed in *M. thermautotrophicus*, *P. furiosus* and *P. horikoshii* [48], this association turns out to be conserved in all Thermococcales, Halobacteriales, and Methanosarcinales (Figure 3), suggesting that PolD may be recruited by Cdc6 to *oriC* via its small subunit DP1. Interestingly, we recently noticed the presence of an origin recognition box (ORB) and mini-ORB repeats in the gene encoding the DP1 subunit of the four Thermococcales [49]. This suggests that the small subunit of PolD indeed plays a specific role, which remains to be explored in the initiation of DNA replication in Euryarchaeota.

Identification of new putative DNA replication proteins

We hoped that genome context analysis could help to identify new putative DNA replication proteins in archaeal genomes via the recurrent association of uncharacterized open reading frames to genes encoding already known DNA replication proteins. As previously observed by others [50], and further confirmed by the present analysis, most euryarchaeal genomes (that is, Methanosarcinales, Thermoplasmatales, Halobacteriales, *A. fulgidus*, *M. maripaludis*, and *M. hungatei*) harbor a gene that encodes an OB fold-containing protein without assigned function that is distantly related to the RPA32 subunit of Thermococcales (COG3390). Interestingly, in most euryarchaeal genomes, the gene belonging to COG3390 is arranged in tandem with a gene encoding a RPA41 homolog (which nearly always contains a Zn-finger domain) suggesting that the two gene products functionally associate ([50] and this study; Additional data file 2). Two copies of this RPA41-COG3390 encoding gene cluster are present in Methanosarcinales and Halobacteriales, indicating that the association of the two genes was maintained in both copies after a duplication event that probably occurred before the divergence of these two archaeal lineages. It is tempting to speculate that this RPA32-related protein is a novel single-stranded binding protein that cooperates with RPA in DNA transactions in some euryarchaea.

Another interesting candidate is a protein that we previously identified as PACE12 in a list of proteins from Archaea conserved in Eukarya [51]. Interestingly, the gene encoding PACE12 is located just upstream of the PPsG DNA replication cluster in all Sulfolobales and of the genes encoding MCM and

Gins23 in the three *Pyrococcus* species (Figure 2a,c). This suggests that PACE12 could be involved in the network connecting these two clusters. Furthermore, the gene encoding the protein PACE12 co-localizes with the gene encoding DP2 in all Thermoplasmatales (they are both transcribed in the same direction), strengthening the link between PACE12 and DNA replication (Additional data file 2). The PACE12 protein has now been identified as the prototype of a new family of GTPases, the GPN-loop GTPases [52]. Three paralogues of PACE12 are present in eukaryotes and all of them are essential in yeast [53]. One of the human homologs, the protein XAB1 (or MBD*in*), has been shown to be a partner of two proteins: XPA involved in nucleotide excision repair [54] and MBD2, a component of the MeCP1 large protein complex that represses transcription of densely methylated genes [55]. Such observations, together with our genomic context analysis, strengthens the idea that these GTPases are involved in informational mechanisms at the DNA level, possibly related to DNA replication/repair and conserved from Archaea to human.

Finally, our analysis suggests that the archaeal homologs of the bacterial primase DnaG may be involved in DNA replication/repair in Archaea since the gene encoding DnaG is adjacent to the gene encoding PolB3 in the three crenarchaeal lineages investigated and is located in the vicinity of a gene encoding a RPA in almost all Methanosarcinales (Additional data file 2). Furthermore, the gene encoding the archaeal DnaG is located beside the gene encoding PACE12 in *Picrophilus torridus*. The archaeal DnaG-like protein associates with archaeal exosome components in *S. solfataricus* [17] and in *M. thermautotrophicus* [56]. It is usually assumed, therefore, that this protein is not involved in archaeal DNA replication, in agreement with the presence in all Archaea of a eukaryotic-like primase. Our observation nevertheless suggests that DnaG could have diverse roles, one of them being associated with DNA replication or possibly DNA repair.

Association of DNA replication genes with translation genes

Surprisingly, we found that the DNA replication genes of the PPsG cluster (in crenarchaeal genomes) or its subsets (in euryarchaeal genomes) are frequently contiguous to a set of genes encoding proteins involved in translation. This association forms a supercluster grouping in the same orientation as the genes of the PPsG cluster and a highly conserved cluster of four genes encoding, in order, the ribosomal proteins L44E and S27E, the alpha subunit of the initiation factor aIF-2, and the protein Nop10 (involved in rRNA processing) (hereafter called the LSIN cluster). The complete LSIN cluster is conserved in all Crenarchaea and nearly all Euryarchaea (Figure 4). Surprisingly, despite the nearly systematic conservation of the LSIN cluster in all archaeal lineages, we did not find any publication reporting a direct link between S27E, L44E, aIF-2, and Nop10. A genetic study in yeast pointing toward a role of S27E in rRNA maturation attracted our attention given that Nop10 is involved in this process [57,58]. However, the association of genes coding for S27E, L44E, aIF-2 alpha, and Nop10 is so highly conserved that a link between these four proteins is to be expected. For instance, they could participate in a mechanism coupling ribosome biogenesis to translation, but establishing a functional connection would require further evidence. In euryarchaeal genomes, the gene encoding Nop10 is almost always associated with an additional gene coding for a putative ATPase with no orthologues in crenarchaea and *N. equitans* (COG2047). Therefore, this protein may interact with Nop10, maybe as a regulator given its predicted function.

The genes of the PPsG and LSIN clusters are always organized in the same order and all transcribed in the same direction (Figure 4). This PPsG-LSIN supercluster is complete in all Crenarchaea and nearly complete in Methanobacteriales (with only the gene encoding PriS missing), Methanosarcinales and Methanomicrobiales (with only the gene encoding PCNA missing). Subsets of the PPsG-LSIN supercluster, still consisting of an association between DNA replication and translation protein-encoding genes, are present in *M. kandleri* (G-LSIN), in Methanococcales (PG-LS) and *A. fulgidus* (G–LS). Interestingly, the genes encoding L44E and S27E (LS cluster) are located close to the gene encoding PolB in Thermococcales, whereas the gene encoding Nop10 (N) is close to the gene encoding MCM in *N. equitans*, indicating that the translation proteins encoded by the genes of the LSIN cluster are somehow linked to DNA replication (Additional data file 2).

The archaeal translation initiation factor IF-2 is composed of three subunits, but the three corresponding genes are never adjacent in archaeal genomes. Since the gene encoding the alpha subunit belongs to a conserved operon structure grouping genes encoding DNA replication and translation proteins (Figure 4), we examined the surroundings of the genes encoding the beta and gamma subunits to detect any recurrent gene pairing. Interestingly, the gene for the beta subunit is also associated with DNA replication genes in archaeal genomes since it is adjacent to the gene encoding the replicative helicase MCM (*M. kandleri*, *M. thermautotrophicum*) or forms a cluster together with the genes encoding MCM and Gins23 in the four Thermococcales (Figure 2c). In contrast, the gene coding for the gamma subunit is not linked to DNA replication genes (data not shown). The association of the gene coding for the beta subunit of the initiation factor aIF-2 is not supported by our numerical analysis (Additional data file 4), indicating that this gene pairing may not be significant, although our numerical analysis clearly shows that this association cannot be considered as a chance occurrence (see below). Furthermore, we believe that the presence of DNA replication genes in the vicinity of two of the genes encoding the subunits of the initiation factor aIF-2 is noteworthy. In eukaryotes, eIF-2 is a major target for protein synthesis regulation since its phosphorylation inhibits translation at the ini-

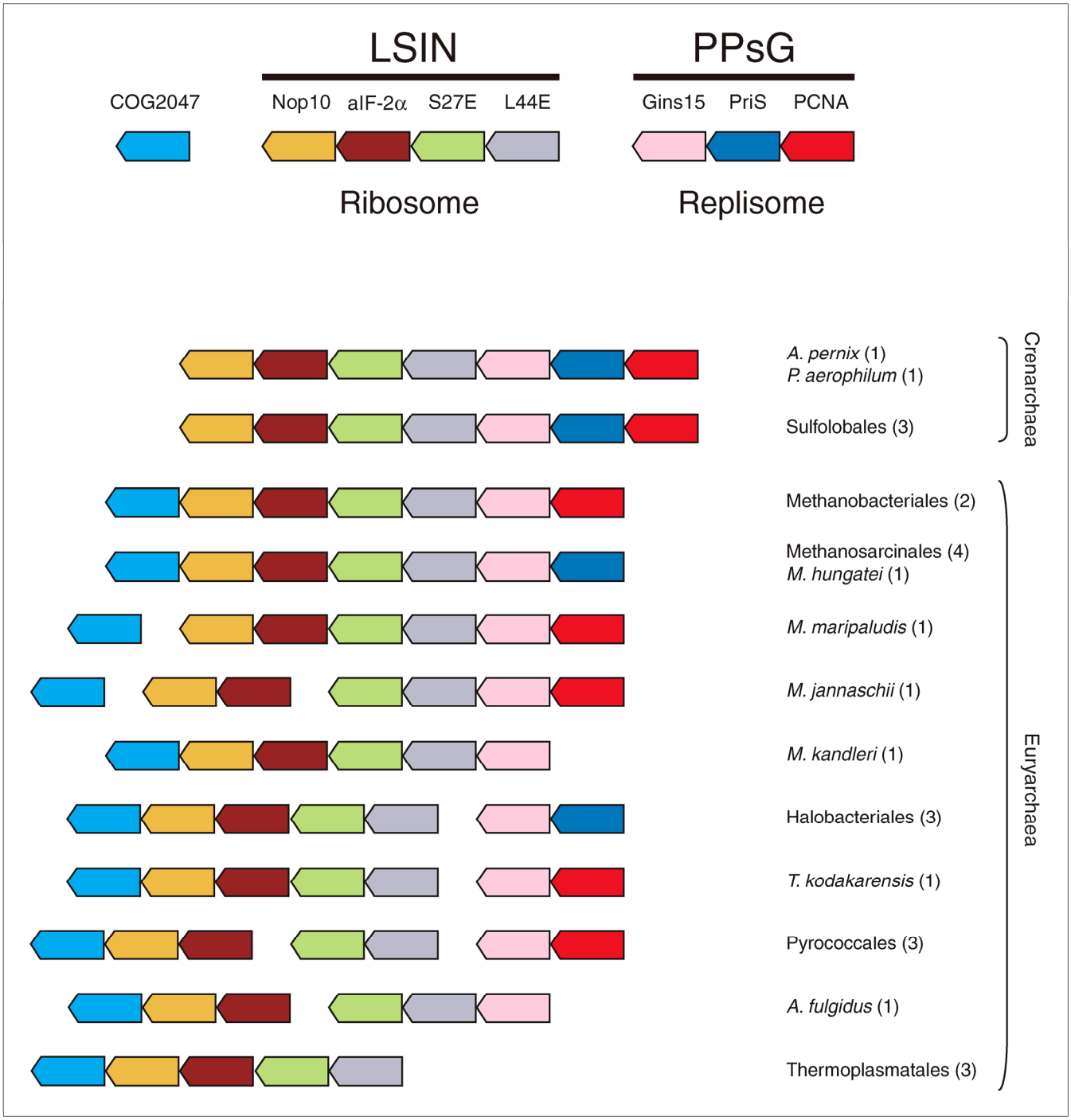

Figure 4
Clustering of DNA replication and ribosome-associated genes in archaeal genomes. Orthologous genes are indicated in the same color. Each gene is denoted by the name of the protein it encodes (see the key at the top). COG2047 encodes an uncharacterized protein of the ATP-grasp superfamily; this COG is absent from Crenarchaea and *N. equitans*. Species or cell lineages that have the same genomic environment are listed and the number of corresponding genomes is given in parentheses. Genes are not shown to scale.

tiation step; notably, it has been shown that phosphorylation of the alpha subunit of eIF-2 leads to apoptosis in stress conditions [59]. A recent *in vitro* study has reported that aIF-2 alpha is phosphorylated in a similar fashion to eIF-2 alpha, suggesting the existence of a phosphorylation pathway in the regulation of protein synthesis in Archaea [60]. Our genome context analysis suggests that aIF-2 may associate with both MCM and the gene products of the PPsG cluster via its beta and alpha subunits, respectively (Figures 2c and 4). Given the homology between the translational processes in Archaea and

eukaryotes, we speculate that a/eIF-2 could be involved in a mechanism that couples the rate of protein synthesis to the regulation of replication, possibly at the elongation step. Some of the partners of aIF-2 in this process may be among the proteins that are encoded by the genes that associate with the gene coding for aIF-2α in the PPsG-LSIN supercluster.

In the course of performing literature mining regarding these proteins, we focused our attention on S27E since this protein exhibits various extra-ribosomal functions. In human, the gene for this ribosomal protein was originally isolated in a screen for growth factor-induced genes and its product called metallopanstimulin (MPS-1) because it was identified as a metalloprotein expressed in a wide spectrum of proliferating tissues [61]. S27E (MPS-1) is considered as an oncogene and a potential target for cancer therapy because it is highly expressed in actively proliferating cells and cancer cell lines and seems to play a role in progression towards malignancy [62]. Wang and co-workers [62] have recently shown that inactivation of MPS-1 inhibits growth and tumorigenesis and leads to an increase of spontaneous apoptosis in gastric cancer cells. These authors stressed that understanding the mechanism of action of S27E in tumorigenesis "is of paramount interest in the target design for medical intervention in malignant tumor formation" [62]. Interestingly, eukaryotic S27E binds single-stranded as well as double-stranded DNA, with specific binding to the cyclic-AMP responsive element sequence [63]. Several data obtained in eukaryotes indeed suggest that, in addition to its role in the ribosome, S27E may deal with RNA or DNA transaction processes. Hence, S27A mutants in *Arabidopsis thaliana* (S27A is homologous to archaeal S27E) are impaired in the elimination of damaged transcripts after a genotoxic stress, suggesting that S27A is involved in mRNA turnover [64]. Of note, computational analysis showed that S27A from *A. thaliana* exhibits a motif in common with transcriptions factors known to have roles in DNA repair [64]. Thus, S27E may deal with translation as well as ribosome biogenesis, transcription, and DNA repair.

Two main hypotheses can be put forward to explain the genomic association of genes encoding proteins involved in DNA replication and genes coding for proteins involved in translation. First, replication proteins encoded by the PPsG cluster or the translation proteins encoded by the LSIN cluster could have evolved a completely new function, thus harboring two different activities, one in translation and another in replication (moonlighting proteins; for a recent review see [65]); the same property (for example, nucleic acid binding ability) could be used to interact with RNA in a ribosome context and to deal with DNA in a chromosome background. The proteins of the LSIN-PPsG cluster might, therefore, be involved in both translation and DNA replication, independently of any connection between these two processes. A second hypothesis is that the PPsG-LSIN cluster reflects some specific regulatory network coupling DNA replication and translation. The latter hypothesis is more appealing to us than the former since it might be logical to couple ribosome biogenesis and DNA replication to maintain the balance between the amount of DNA and proteins in the cell at different times of the cell cycle. This hypothesis was first proposed by Du and Stillman [66], who reported in yeast that ORC (origin recognition complex) and MCM associate in a complex with proteins involved in ribosome biosynthesis, suggesting potential links between cell proliferation, ribosome biogenesis, and DNA replication. Actually, mounting evidence in eukaryotes points toward a link between ribosome biogenesis and the cell cycle (reviewed in [67]). The existence of a coupling between DNA replication and translation could also possibly explain why the MCM protein of the archaeon *P. abyssi* binds preferentially to the ribosomal operon in stationary phase [49]. Thus, unsuspected links between DNA replication and ribosome biogenesis are emerging piecemeal from biochemical and genetic studies in Archaea and eukaryotes.

Statistical analysis of genome context supports the cluster of DNA replication and translation genes

In order to evaluate the statistical significance of the various genes associations that we have detected in this analysis, we first determined the probability of finding by chance groups of two, three, and so on contiguous genes in a set of 26 randomly shuffled genomes (starting from the genome of *S. acidocaldarius* whose size (2,329 genes) is close to the average size of archaeal genomes). As intuitively expected, we determined that the probability of finding that two neighboring genes in *S. acidocaldarius* are still neighbors in any of the 26 randomized *S. acidocaldarius* genomes is very low (Additional data file 4). For instance, the probabilities of finding that two neighboring genes are still neighbors in two or three randomized genomes is 0.23% and 0.04%, respectively.

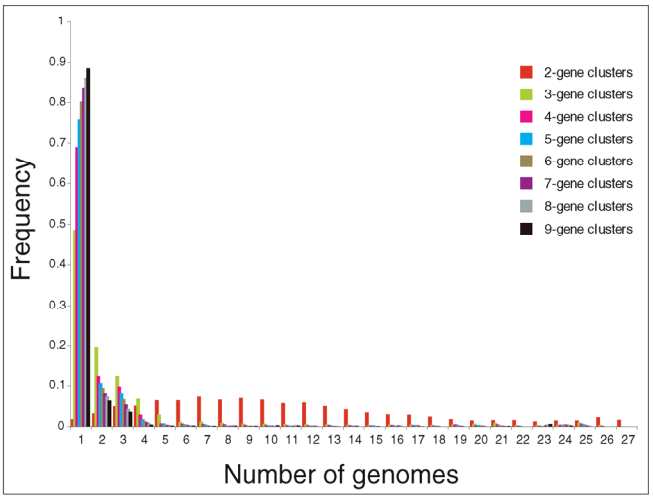

Figure 5
Gene clusters conservation in 27 archaeal genomes. Gene clusters of 2 to 9 genes were searched in 27 archaeal genomes. Two-gene clusters are rather abundant in archaeal genomes; clusters of more than two genes appear mainly in four or fewer genomes.

Accordingly, if two or more genes are located close to each other in the genomes of more than two different species, this cannot be by chance. Two alternatives can be proposed to explain why co-localization of some genes are conserved in several species: these genes were adjacent in the genome of the ancestor of these species and have not yet been separated by chromosome recombination; or there is a selection pressure that favors organisms in which these genes are associated, either by maintaining an association already present in the ancestor of the two genomes or favoring their recurrent association. The distribution of gene clusters in present-day genomes should be the result of a combination of these two alternatives. One can reason that gene clusters maintained only by chance (genes not yet separated by recombination) disappear, on average, more rapidly in the course of evolution than those maintained by selection pressure. In that case, clusters under positive selection pressure should be essentially those present in the highest number of genomes. To perform a quantitative analysis that could be amenable to statistical analysis, we first determined the distribution pattern of gene clusters in our dataset of 27 genomes (see Materials and methods). As shown in Figure 5, we observed that nearly all two-gene clusters (red bar) are present in more than two genomes (up to 27), with a very broad distribution, indicating that genes pairs have been significantly conserved during the evolution of archaeal genomes. In contrast, clusters of three or more genes are much less conserved (most of them being present in from one to four genomes for triplets (green bar) and in one to three genomes for longer clusters). We then calculated a prevalence index based on presence-absence for all gene clusters analyzed, and determined the cumulative index frequency curves for clusters of the same size (see Additional data file 4 for a diagram of curves obtained with clusters of two genes and for clusters of more than two genes). In a traditional statistical approach (one-tailed test), one would consider that a cluster is significant (under positive selection pressure) if its index frequency is located in the portion of the curve corresponding to the 5% less frequent clusters (that is, the very few clusters present in the highest number of genomes), thus strongly deviating from the average of the distribution. The index frequency will hereafter be called the frequency score of this particular cluster. This approach is conservative since it implies that only 5% of the gene clusters in the complete dataset are the result of functional constraints. However, even within a 5% threshold, we found that 13 of the 32 clusters tested in our statistical analysis were supported. These include the supercluster LSIN-PPsG grouping DNA replication and translation genes and most clusters derived from this supercluster (including the PPsG cluster; Additional data file 4). The supercluster LSIN-PPsG itself is highly significant since its frequency score is 2%.

Many potentially interesting gene clusters detected in our analysis (in particular most two-gene clusters) are not statistically supported by a 5% standard threshold. For instance, although the cluster between TFS and PCNA is present in 10 genomes of both Euryarchaea and Crenarchaea (so is probably biologically relevant (see below)), its frequency score is not significant (33%). However, this is also the case for gene associations whose biological relevance has been validated experimentally. For instance, the frequency score of the cluster of genes coding for MCM and Gins23 is not significant (55%) despite the functional relevance of this cluster, as indicated by the work of Bell and colleagues [24]. This again emphasizes that clusters with frequency scores above the 5% threshold are a mixture of clusters maintained only by chance and clusters under selection pressure; there is no easy way to discriminate between them. The best approximation is to consider that clusters under selection pressure are those conserved in genomes from species belonging to different archaeal orders; even more constrained are those conserved across different phyla.

Conclusion

We have identified, through our genome context analysis of archaeal DNA replication genes, several conserved gene associations that have escaped previous global screening, and from which new functional connections have been inferred. Most of these gene clusters are conserved in distantly related archaeal genomes, indicating that these clusterings are not merely coincidental but probably functionally relevant, and that they should be under selection pressure to optimize functional interactions between the encoded proteins (for instance, via transcriptional co-regulation) and/or to facilitate the formation of specific protein sub-complexes. In particular, we predict that the PCNA clamp, the DNA primase, and the helicase MCM are functionally connected via the GINS complex in the replication factory and that Cdc6 may interact with DP1 at *oriC* for the initiation of DNA replication. We also speculate the existence of cross-talk between DNA replication, DNA repair, and transcription in which PCNA, TFS, and the ADP-pyrophosphatase NudF may be involved. Moreover, we suggest that three proteins without clear functional assignations (an OB-fold containing protein, a recently described new GTPase, DnaG) may take part in informational processes at the DNA level.

Finally, and unexpectedly, we discovered that the genes coding for a particular set of proteins (Gins15, PCNA and/or PriS) are almost systematically arranged in an operon-like structure with a conserved cluster of genes coding for ribosome-related proteins (S27E, L44E, aIF-2α, and Nop10), suggesting the existence of a functional coupling between DNA replication and translation in Archaea. The biological relevance of this association is strongly supported by a statistical analysis of the gene cluster distribution in the 27 archaeal genomes of our dataset. Most of the genes belonging to this particular cluster have eukaryotic homologs but are absent from bacteria; thus, we anticipate that DNA replication and translation may be co-regulated by a mechanism conserved from Archaea to human. The nature of these connections remains to be

deciphered but the gene cluster highlighted in this study may be a benchmark for future experimental studies aiming to address this fundamental issue.

Materials and methods

Identification of DNA replication genes in archaeal genomes

A list of 12 factors - corresponding to both monomeric and heteromultimeric proteins - likely to be involved in DNA replication was drawn up. This list contains: the initiation factor Cdc6/Orc1; PolB1, PolB2, and PolB3; the small and large subunits of PolD (DP1 and DP2); the helicase MCM; the sliding clamp PCNA; the small and large subunits of the clamp-loader RFC (RFC-s and RFC-l); the small and large subunits of the DNA primase (PriS and PriL); the single-stranded binding protein (RPA or SSB); DNA ligase; the two subunits of Topo VI (Topo VIA and Topo VIB); RNase HII; the flap endonuclease FEN-1; and the two Gins subunits (Gins15 and Gins23) of the GINS complex. The accession numbers of these proteins or protein subunits were retrieved from 27 complete archaeal genomes (*A. pernix*; *P. aerophilum*; the three Sulfolobales, *S. acidocaldarius*, *S. solfataricus*, and *S. tokodaii*; *N. equitans*; *A. fulgidus*; the three Halobacteriales *H. marismortui*, *H. salinarum*, and *Natronomonas pharaonis*; the two Methanobacteriales *M. thermautotrophicus* and *M. stadtmanae*; the two Methanococcales *M. jannaschii* and *M. maripaludis*; *M. kandleri*; the four Methanosarcinales *M. burtonii*, *Methanosarcina acetivorans*, *M. barkeri*, and *M. mazei*; *Methanospirillum hungatei*; the four Thermococcales *P. abyssi*, *P. furiosus*, *P. horikoshii*, and *Thermococcus kodakaraensis*; and the three Thermoplasmatales *Picrophilus torridus*, *Thermoplasma acidophilum*, and *T. volcanium*) by means of BLASTP or PSI-BLAST [68] performed at the NCBI [25] using *P. abyssi* and *S. solfataricus* homologs and, if available, sequences of biochemically characterized proteins as references. All the proteins of the above list were assigned to clusters of orthologous groups (COGs) [69,70] using the COG guess tool from the LBMGE Genomics ToolBox [71] in order to confirm their annotation. Complete archaeal genomes were searched using BLASTP for each class of proteins with various seeds as bait in order to look for misannotated proteins or to uncover overlooked homologs. Finally, BLASTN searches were achieved at the NCBI against the non-redundant archaeal nucleotide sequences database to identify missing open reading frames using closest relative homolog as a query.

Genome context analysis of DNA replication genes

The genomic context of DNA replication genes were visualized with Genomapper. All genomic contexts were scrutinized manually since the conserved cluster of genes encoding PCNA, PriS and Gins15 was not detected with an automated tool such as STRING [72], likely because sequence similarities are weak between Gins protein family members [24]. In addition, evolutionarily conserved gene neighborhoods turned out to be of valuable importance to identify the archaeal Gins homologs that escaped PSI-BLAST searches. A window encompassing the target gene, the five upstream and five downstream flanking genes was considered during all the genomic environment analysis process. The protein encoded by the genes enclosed in the delimited genomic region were identified using Genome guts [71], assigned to a COG using COG guess, and BLASTP searches against the non-redundant archaeal proteins database were carried out at the NCBI so as to validate their annotation. The surroundings of DNA replication genes that are located on extrachromosomic elements were not inspected since the LBMGE genomes database does not contain archaeal plasmid sequences.

Statistical analyses

Gene cluster conservation in randomized genomes

We chose the *S. acidocaldarius* genome, whose 2,329 genes approximate the average gene content in completely sequenced archaeal genomes, as reference. We generated 26 random genomes as follows: all genes of the *S. acidocaldarius* genome were position-exchanged for another gene chosen randomly from the genome; starting with gene number one and then sequentially applying the same process to all other genes. We then counted the number of times clusters of two to nine genes, present in the genome of *S. acidocaldarius*, remained together in the 26 randomized genomes. For all clusters we calculated a prevalence index based on their presence and absence. That is, every time a gene cluster was indeed present in a randomized genome the prevalence gene index gained one point, otherwise it lost one point. This approach allowed us to calculate the probabilities of having gene clusters by chance only (data not shown). These probabilities were lower than 0.01%, except for clusters of two or three genes in two genomes (see text).

Gene cluster conservation in complete archaeal genomes

To establish if the conservation of the gene clusters characterized in this work was statistically significant, we decided to determine the global gene cluster conservation among the 27 archaeal genomes we used for genome context analysis. A genome was chosen randomly, and from this genome a gene was taken randomly. This gene was then BLAST searched (E-value 0.01) against all other 26 genomes. The same BLAST search (E-value 0.01) was performed for its two neighboring genes (the first upstream and the first downstream). Every time the gene appeared with at least one of the same flanking genes in another genome, the prevalence gene index gained one point, otherwise the index lost one point. The whole operation was repeated 10,000 times. We repeated the same process for gene clusters of three to nine genes. We thus ended up with 10,000 prevalence indexes for each size of gene cluster, from which we constructed frequency distributions (examples of these distributions can be found in Additional data file 4). At the same time we determined the prevalence indexes of 32 representative clusters containing DNA replication genes and/or translation genes (indexes are shown in Additional

data file 4). We then performed a one-tailed test to settle the significance of our clusters; we simply located the prevalence indexes of our 32 clusters in the frequency distributions (frequency score). The indexes were considered statistically supported when they were present in the 5% or less area of the right part of the distributions (examples can be found in Additional data file 4); this area of the distributions contains those very few clusters highly conserved in archaeal genomes.

Abbreviations

COG, cluster of orthologous groups; DP1, PolD small subunit; DP2, PolD large subunit; LSIN, L44E S27E aIF-2 alpha Nop10; MPS-1, metallopanstimulin 1; OB, oligonucleotide/oligosaccharide-binding; PACE, proteins of Archaea conserved in Eukarya; Pol, DNA polymerase; PPsG, PCNA PriS Gins15; PriL, DNA primase large subunit; PriS, DNA primase small subunit; RFC-l, replication factor C large subunit; RFC-s, replication factor C small subunit; TFS, transcription factor S; Topo, topoisomerase.

Authors' contributions

PF initiated the study. JB performed genome context analysis. DC carried out statistical analyses and simulations, and helped to interpret numerical analysis. PF and JB interpreted the data and wrote the paper.

Additional data files

The following additional data are available. Additional data file 1 contains a table listing the DNA replication factors encoded by archaeal genomes analyzed in this work. Additional data file 2 contains several figures showing the genomic context of all the archaeal DNA replication genes analyzed in this study. Additional data file 3 contains a description of and discussion about genomic associations of DNA replication genes with genes coding for archaeal homologs of DNA repair/recombination proteins. Additional data file 4 contains a table with the prevalence indexes of gene clusters and two sketches illustrating frequency distributions of clusters of two genes and clusters of more than two genes.

Acknowledgements

This work was supported by a grant from the Japan Society from Promotion of Sciences to JB and by funds from the Human Frontier Science Program and Association pour la Recherche contre le Cancer to PF.

References

1. Mushegian AR, Koonin EV: **Gene order is not conserved in bacterial evolution.** *Trends Genet* 1996, **12:**289-290.
2. Dandekar T, Snel B, Huynen M, Bork P: **Conservation of gene order: a fingerprint of proteins that physically interact.** *Trends Biochem Sci* 1998, **23:**324-328.
3. Wolf YI, Rogozin IB, Kondrashov AS, Koonin EV: **Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context.** *Genome Res* 2001, **11:**356-372.
4. Hershberg R, Yeger-Lotem E, Margalit H: **Chromosomal organization is shaped by the transcription regulatory network.** *Trends Genet* 2005, **21:**138-142.
5. Price MN, Huang KH, Arkin AP, Alm EJ: **Operon formation is driven by co-regulation and not by horizontal gene transfer.** *Genome Res* 2005, **15:**809-819.
6. Glansdorff N: **On the origin of operons and their possible role in evolution toward thermophily.** *J Mol Evol* 1999, **49:**432-438.
7. Huynen M, Snel B, Lathe W, Bork P: **Exploitation of gene context.** *Curr Opin Struct Biol* 2000, **10:**366-370.
8. Marcotte EM: **Computational genetics: finding protein function by nonhomology methods.** *Curr Opin Struct Biol* 2000, **10:**359-365.
9. Galperin MY, Koonin EV: **Who's your neighbor? New computational approaches for functional genomics.** *Nat Biotechnol* 2000, **18:**609-613.
10. Korbel JO, Jensen LJ, von Mering C, Bork P: **Analysis of genomic context: prediction of functional associations from conserved bidirectionally transcribed gene pairs.** *Nat Biotechnol* 2004, **22:**911-917.
11. Jansen R, Embden JD, Gaastra W, Schouls LM: **Identification of genes that are associated with DNA repeats in prokaryotes.** *Mol Microbiol* 2002, **43:**1565-1575.
12. Makarova KS, Aravind L, Grishin NV, Rogozin IB, Koonin EV: **A DNA repair system specific for thermophilic Archaea and bacteria predicted by genomic context analysis.** *Nucleic Acids Res* 2002, **30:**482-496.
13. Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, Romero DA, Horvath P: **CRISPR provides acquired resistance against viruses in prokaryotes.** *Science* 2007, **315:**1709-1712.
14. Makarova KS, Koonin EV: **Comparative genomics of Archaea: how much have we learned in six years, and what's next?** *Genome Biol* 2003, **4:**115.
15. Makarova KS, Koonin EV: **Evolutionary and functional genomics of the Archaea.** *Curr Opin Microbiol* 2005, **8:**586-594.
16. Koonin EV, Wolf YI, Aravind L: **Prediction of the archaeal exosome and its connections with the proteasome and the translation and transcription machineries by a comparative-genomic approach.** *Genome Res* 2001, **11:**240-252.
17. Evguenieva-Hackenberg E, Walter P, Hochleitner E, Lottspeich F, Klug G: **An exosome-like complex in *Sulfolobus solfataricus*.** *EMBO Rep* 2003, **4:**889-893.
18. Constantinesco F, Forterre P, Elie C: **NurA, a novel 5'-3' nuclease gene linked to rad50 and mre11 homologs of thermophilic Archaea.** *EMBO Rep* 2002, **3:**537-542.
19. Constantinesco F, Forterre P, Koonin EV, Aravind L, Elie C: **A bipolar DNA helicase gene, herA, clusters with rad50, mre11 and nurA genes in thermophilic archaea.** *Nucleic Acids Res* 2004, **32:**1439-1447.
20. Barry ER, Bell SD: **DNA replication in the archaea.** *Microbiol Mol Biol Rev* 2006, **70:**876-887.
21. Uemori T, Sato Y, Kato I, Doi H, Ishino Y: **A novel DNA polymerase in the hyperthermophilic archaeon, *Pyrococcus furiosus*: gene cloning, expression, and characterization.** *Genes Cells* 1997, **2:**499-512.
22. Myllykallio H, Lopez P, Lopez-Garcia P, Heilig R, Saurin W, Zivanovic Y, Philippe H, Forterre P: **Bacterial mode of replication with eukaryotic-like machinery in a hyperthermophilic archaeon.** *Science* 2000, **288:**2212-2215.
23. Robinson NP, Dionne I, Lundgren M, Marsh VL, Bernander R, Bell SD: **Identification of two origins of replication in the single chromosome of the archaeon *Sulfolobus solfataricus*.** *Cell* 2004, **116:**25-38.
24. Marinsek N, Barry ER, Makarova KS, Dionne I, Koonin EV, Bell SD: **GINS, a central nexus in the archaeal DNA replication fork.** *EMBO Rep* 2006, **7:**539-545.
25. **National Center for Biotechnology Information** [http://www.ncbi.nlm.nih.gov/]
26. Berquist B, DasSarma P, DasSarma S: **Essential and non-essential DNA replication genes in the model halophilic Archaeon, *Halobacterium* sp. NRC-1.** *BMC Genet* 2007, **8:**31.
27. Gadelle D, Filee J, Buhler C, Forterre P: **Phylogenomics of type II DNA topoisomerases.** *Bioessays* 2003, **25:**232-242.
28. Fitz-Gibbon ST, Ladner H, Kim UJ, Stetter KO, Simon MI, Miller JH: **Genome sequence of the hyperthermophilic crenarchaeon *Pyrobaculum aerophilum*.** *Proc Natl Acad Sci USA* 2002, **99:**984-989.
29. White MF: **Archaeal DNA repair: paradigms and puzzles.** *Bio-*

chem Soc Trans 2003, **31:**690-693.

30. Slesarev AI, Mezhevaya KV, Makarova KS, Polushin NN, Shcherbinina OV, Shakhova VV, Belova GI, Aravind L, Natale DA, Rogozin IB, Tatusov RL, Wolf YI, Stetter KO, Malykh AG, Koonin EV, Kozyavkin SA: **The complete genome of hyperthermophile *Methanopyrus kandleri* AV19 and monophyly of archaeal methanogens.** *Proc Natl Acad Sci USA* 2002, **99:**4644-4649.
31. Yamashiro K, Yokobori S, Oshima T, Yamagishi A: **Structural analysis of the plasmid pTA1 isolated from the thermoacidophilic archaeon *Thermoplasma acidophilum*.** *Extremophiles* 2006, **10:**327-335.
32. Zhang R, Zhang CT: **Identification of replication origins in the genome of the methanogenic archaeon, *Methanocaldococcus jannaschii*.** *Extremophiles* 2004, **8:**253-258.
33. Iyer LM, Koonin EV, Leipe DD, Aravind L: **Origin and evolution of the archaeo-eukaryotic primase superfamily and related palm-domain proteins: structural insights and new members.** *Nucleic Acids Res* 2005, **33:**3875-3896.
34. Brochier C, Gribaldo S, Zivanovic Y, Confalonieri F, Forterre P: **Nanoarchaea: representatives of a novel archaeal phylum or a fast-evolving euryarchaeal lineage related to Thermococcales?** *Genome Biol* 2005, **6:**R42.
35. Lundgren M, Bernander R: **Genome-wide transcription map of an archaeal cell cycle.** *Proc Natl Acad Sci USA* 2007, **104:**2939-2944.
36. Lao-Sirieix SH, Pellegrini L, Bell SD: **The promiscuous primase.** *Trends Genet* 2005, **21:**568-572.
37. Ishino S, Ishino Y: **Comprehensive search for DNA polymerase in the hyperthermophilic archaeon, *Pyrococcus furiosus*.** *Nucleosides Nucleotides Nucleic Acids* 2006, **25:**681-691.
38. Hausner W, Lange U, Musfeldt M: **Transcription factor S, a cleavage induction factor of the archaeal RNA polymerase.** *J Biol Chem* 2000, **275:**12393-12399.
39. Lange U, Hausner W: **Transcriptional fidelity and proofreading in Archaea and implications for the mechanism of TFS-induced RNA cleavage.** *Mol Microbiol* 2004, **52:**1133-1143.
40. Brochier C, Forterre P, Gribaldo S: **Archaeal phylogeny based on proteins of the transcription and translation machineries: tackling the *Methanopyrus kandleri* paradox.** *Genome Biol* 2004, **5:**R17.
41. McLennan AG: **The Nudix hydrolase superfamily.** *Cell Mol Life Sci* 2006, **63:**123-143.
42. Sheikh S, O'Handley SF, Dunn CA, Bessman MJ: **Identification and characterization of the Nudix hydrolase from the Archaeon, *Methanococcus jannaschii*, as a highly specific ADP-ribose pyrophosphatase.** *J Biol Chem* 1998, **273:**20924-20928.
43. D'Amours D, Desnoyers S, D'Silva I, Poirier GG: **Poly(ADP-ribosyl)ation reactions in the regulation of nuclear functions.** *Biochem J* 1999, **342:**249-268.
44. Jolivet E, Matsunaga F, Ishino Y, Forterre P, Prieur D, Myllykallio H: **Physiological responses of the hyperthermophilic archaeon "*Pyrococcus abyssi*" to DNA damage caused by ionizing radiation.** *J Bacteriol* 2003, **185:**3958-3961.
45. Faraone-Mennella MR, Farina B: **In the thermophilic archaeon *Sulfolobus solfataricus* a DNA-binding protein is *in vitro* (Adpribosyl)ated.** *Biochem Biophys Res Commun* 1995, **208:**55-62.
46. Tashiro R, Wang AH, Sugiyama H: **Photoreactivation of DNA by an archaeal nucleoprotein Sso7d.** *Proc Natl Acad Sci USA* 2006, **103:**16655-16659.
47. Lundgren M, Andersson A, Chen L, Nilsson P, Bernander R: **Three replication origins in *Sulfolobus* species: synchronous initiation of chromosome replication and asynchronous termination.** *Proc Natl Acad Sci USA* 2004, **101:**7046-7051.
48. Lopez P, Philippe H, Myllykallio H, Forterre P: **Identification of putative chromosomal origins of replication in Archaea.** *Mol Microbiol* 1999, **32:**883-886.
49. Matsunaga F, Glatigny A, Mucchielli-Giorgi MH, Agier N, Delacroix H, Marisa L, Durosay P, Ishino Y, Aggerbeck L, Forterre P: **Genome-wide and biochemical analyses of DNA-binding activity of Cdc6/Orc1 and Mcm proteins in *Pyrococcus* sp.** *Nucleic Acids Res* 2007, **35:**3214-3222.
50. Komori K, Ishino Y: **Replication protein A in *Pyrococcus furiosus* is involved in homologous DNA recombination.** *J Biol Chem* 2001, **276:**25654-25660.
51. Matte-Tailliez O, Zivanovic Y, Forterre P: **Mining archaeal proteomes for eukaryotic proteins with novel functions: the PACE case.** *Trends Genet* 2000, **16:**533-536.
52. Gras S, Chaumont V, Fernandez B, Carpentier P, Charrier-Savournin F, Schmitt S, Pineau C, Flament D, Hecker A, Forterre P, Armengaud J, Housset D: **Structural insights into a new homodimeric self-activated GTPase family.** *EMBO Rep* 2007, **8:**569-575.
53. **The Munich Information center for Protein Sequences Comprehensive Yeast Genome Database** [http://mips.gsf.de/genre/proj/yeast]
54. Nitta M, Saijo M, Kodo N, Matsuda T, Nakatsu Y, Tamai H, Tanaka K: **A novel cytoplasmic GTPase XAB1 interacts with DNA repair protein XPA.** *Nucleic Acids Res* 2000, **28:**4212-4218.
55. Lembo F, Pero R, Angrisano T, Vitiello C, Iuliano R, Bruni CB, Chiariotti L: **MBDin, a novel MBD2-interacting protein, relieves MBD2 repression potential and reactivates transcription from methylated promoters.** *Mol Cell Biol* 2003, **23:**1656-1665.
56. Farhoud MH, Wessels HJ, Steenbakkers PJ, Mattijssen S, Wevers RA, van Engelen BG, Jetten MS, Smeitink JA, van den Heuvel LP, Keltjens JT: **Protein complexes in the archaeon *Methanothermobacter thermautotrophicus* analyzed by blue native/SDS-PAGE and mass spectrometry.** *Mol Cell Proteomics* 2005, **4:**1653-1663.
57. Baudin-Baillieu A, Tollervey D, Cullin C, Lacroute F: **Functional analysis of Rrp7p, an essential yeast protein involved in pre-rRNA processing and ribosome assembly.** *Mol Cell Biol* 1997, **17:**5023-5032.
58. Hamma T, Reichow SL, Varani G, Ferré-D'Amaré AR: **The Cbf5-Nop10 complex is a molecular bracket that organizes box H/ACA RNPs.** *Nat Struct Mol Biol* 2005, **12:**1101-1107.
59. Scheuner D, Patel R, Wang F, Lee K, Kumar K, Wu J, Nilsson A, Karin M, Kaufman RJ: **Double-stranded RNA-dependent protein kinase phosphorylation of the alpha-subunit of eukaryotic translation initiation factor 2 mediates apoptosis.** *J Biol Chem* 2006, **281:**21458-21468.
60. Tahara M, Ohsawa A, Saito S, Kimura M: ***In vitro* phosphorylation of initiation factor 2 alpha (aIF2 alpha) from hyperthermophilic archaeon *Pyrococcus horikoshii* OT3.** *J Biochem* 2004, **135:**479-485.
61. Fernandez-Pol JA, Klos DJ, Hamilton PD: **A growth factor-inducible gene encodes a novel nuclear protein with zinc finger structure.** *J Biol Chem* 1993, **268:**21198-21204.
62. Wang YW, Qu Y, Li JF, Chen XH, Liu BY, Gu QL, Zhu ZG: ***In vitro* and *in vivo* evidence of metallopanstimulin-1 in gastric cancer progression and tumorigenicity.** *Clin Cancer Res* 2006, **12:**4965-4973.
63. Fernandez-Pol JA, Klos DJ, Hamilton PD: **Metallopanstimulin gene product produced in a baculovirus expression system is a nuclear phosphoprotein that binds to DNA.** *Cell Growth Differ* 1994, **5:**811-825.
64. Revenkova E, Masson J, Koncz C, Afsar K, Jakovleva L, Paszkowski J: **Involvement of *Arabidopsis thaliana* ribosomal protein S27 in mRNA degradation triggered by genotoxic stress.** *EMBO J* 1999, **18:**490-499.
65. Jeffery CJ: **Mass spectrometry and the search for moonlighting proteins.** *Mass Spectrom Rev* 2005, **24:**772-782.
66. Du YC, Stillman B: **Yph1p, an ORC-interacting protein: potential links between cell proliferation control, DNA replication, and ribosome biogenesis.** *Cell* 2002, **109:**835-848.
67. Dez C, Tollervey D: **Ribosome synthesis meets the cell cycle.** *Curr Opin Microbiol* 2004, **7:**631-637.
68. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-3402.
69. Tatusov RL, Koonin EV, Lipman DJ: **A genomic perspective on protein families.** *Science* 1997, **278:**631-637.
70. Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND, Koonin EV: **The COG database: new developments in phylogenetic classification of proteins from complete genomes.** *Nucleic Acids Res* 2001, **29:**22-28.
71. **Laboratoire de Biologie Moléculaire du Gène chezles Extrêmophiles Genomics ToolBox** [http://www-archbac.u-psud.fr/genomics/GenomicsToolBox.html]
72. von Mering C, Jensen LJ, Snel B, Hooper SD, Krupp M, Foglierini M, Jouffre N, Huynen MA, Bork P: **STRING: known and predicted protein-protein associations, integrated and transferred across organisms.** *Nucleic Acids Res* 2005, **33(Database issue):**D433-D437.
73. Brochier C, Forterre P, Gribaldo S: **An emerging phylogenetic core of Archaea: phylogenies of transcription and translation machineries converge following addition of new genome**

sequences. *BMC Evol Biol* 2005, **5:**36.
74. Perler FB: **InBase: the Intein Database.** *Nucleic Acids Res* 2002, **30:**383-384.
75. Waters E, Hohn MJ, Ahel I, Graham DE, Adams MD, Barnstead M, Beeson KY, Bibbs L, Bolanos R, Keller M, Kretz K, Lin X, Mathur E, Ni J, Podar M, Richardson T, Sutton GG, Simon M, Soll D, Stetter KO, Short JM, Noordewier M: **The genome of *Nanoarchaeum equitans*: insights into early archaeal evolution and derived parasitism.** *Proc Natl Acad Sci USA* 2003, **100:**12984-12988.
76. Kelman Z, Pietrokovski S, Hurwitz J: **Isolation and characterization of a split B-type DNA polymerase from the archaeon *Methanobacterium thermoautotrophicum* deltaH.** *J Biol Chem* 1999, **274:**28751-28761.
77. Robbins JB, McKinney MC, Guzman CE, Sriratana B, Fitz-Gibbon S, Ha T, Cann IK: **The euryarchaeota, nature's medium for engineering of single-stranded DNA-binding proteins.** *J Biol Chem* 2005, **280:**15325-15339.
78. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** *Nucleic Acids Res* 1994, **22:**4673-4680.
79. Corpet F: **Multiple sequence alignment with hierarchical clustering.** *Nucleic Acids Res* 1988, **16:**10881-10890.
80. Ramadan K, Shevelev I, Hübscher U: **The DNA-polymerase-X family: controllers of DNA quality?** *Nat Rev Mol Cell Biol* 2004, **5:**1038-1043.
81. Komori K, Hidaka M, Horiuchi T, Fujikane R, Shinagawa H, Ishino Y: **Cooperation of the N-terminal helicase and C-terminal endonuclease activities of Archaeal Hef protein in processing stalled replication forks.** *J Biol Chem* 2004, **279:**53175-53185.
82. Fujikane R, Shinagawa H, Ishino Y: **The archaeal Hjm helicase has recQ-like functions, and may be involved in repair of stalled replication fork.** *Genes Cells* 2006, **11:**99-110.
83. Guy CP, Bolt EL: **Archaeal Hel308 helicase targets replication forks *in vivo* and *in vitro* and unwinds lagging strands.** *Nucleic Acids Res* 2005, **33:**3678-3690.
84. Roberts JA, White MF: **DNA end-directed and processive nuclease activities of the archaeal XPF enzyme.** *Nucleic Acids Res* 2005, **33:**6662-6670.
85. Hayashi I, Morikawa K, Ishino Y: **Specific interaction between DNA polymerase II (PolD) and RadB, a Rad51/Dmc1 homolog, in *Pyrococcus furiosus*.** *Nucleic Acids Res* 1999, **27:**4695-4702.
86. Komori K, Sakae S, Fujikane R, Morikawa K, Shinagawa H, Ishino Y: **Biochemical characterization of the hjc holliday junction resolvase of *Pyrococcus furiosus*.** *Nucleic Acids Res* 2000, **28:**4544-4551.
87. Dorazi R, Parker JL, White MF: **PCNA activates the Holliday junction endonuclease Hjc.** *J Mol Biol* 2006, **364:**243-247.

Minirevue

When DNA replication and protein synthesis come together

Jonathan Berthon, Ryosuke Fujikane, Patrick Forterre

Manuscrit soumis au journal *Cell*

When DNA replication and protein synthesis come together

Jonathan Berthon, Ryosuke Fujikane and Patrick Forterre

Univ. Paris-Sud 11, CNRS UMR 8621, Institut de Génétique et Microbiologie, 91405 ORSAY CEDEX, France.

Several recent independent lines of evidence support the existence of mechanisms coupling protein synthesis and DNA replication in the three domains of life. Investigation of these mechanisms should reveal critical cellular regulatory networks whose importance has been underestimated until now.

Although very fruitful in unraveling the major mechanisms that allow living organisms to thrive on our planet, the study of molecular systems via the reductionist approach led to the building of quite artificial barriers between fundamental molecular systems such as DNA replication, DNA repair, transcription, and translation. Some of these barriers have already faded away; for instance, the connection between replication, repair and transcription is now well recognized and actively investigated. More recently, another major barrier—that between DNA replication and translation—has started to crumble, following an amazing number of unexpected results obtained in various model organisms from the three domains of life. These results strongly suggest that mechanisms coupling protein and DNA syntheses are widespread and possibly ubiquitous in cellular organisms. Here, we briefly review recent discoveries of novel regulatory systems and/or proteins possibly coupling DNA replication and protein synthesis in Bacteria and Eukarya. We then turn to our recent observation of a widely

conserved gene association that led us to hypothesize the existence of a novel regulatory network connecting DNA and protein syntheses in Archaea (Berthon et al., 2008). Two ribosomal protein encoding genes involved in this network have eukaryotic homologues that have been recently identified as novel oncogenes. We speculate on the possible molecular mechanisms that could explain these observations and are worth of future experimental investigations.

The stringent response and DNA replication in Bacteria

The stringent response is one of the best-studied regulatory networks in Bacteria. It has been known for a long time that amino acid starvation elicits a sharp increase of intracellular (p)ppGpp concentration which results in a rapid inhibition of rRNA genes transcription and protein synthesis. The last year, it turned out that activation of the stringent response in *Bacillus subtilis* also blocks the elongation step of DNA replication (Wang et al., 2007). DNA elongation is rapidly arrested, within minutes, following amino acid starvation. This arrest is mediated by the increase of the alarmone (p)ppGpp, which directly inhibits *in vitro* the bacterial primase DnaG (Wang et al., 2007). This mechanism could prevent the disruption of replication forks that could otherwise occur due to the reduction of the dNTP pool induced by starvation. Importantly, inhibition of DNA replication mediated by the stringent response does not involve the recruitment of RecA (a potential source of genome instability) and is reversed by the addition of excess nutrients (Wang et al., 2007). This means that arrested replication forks remain ready to resume replication, with minimal disturbance to the genome, as soon as the conditions become again suitable for growth. The mechanism that prevents the recruitment of RecA to replication forks arrested by induction of the stringent response is unknown at present (Wang et al., 2007).

Recently, it was shown that the accumulation of (p)ppGpp during the stringent response also inhibits the initiation step of DNA replication in *Caulobacter crescentus* by promoting, via an unknown mechanism, the degradation of the initiator protein DnaA (Lesley and Shapiro, 2008). All bacterial complete genomes harbor genes coding for the primase DnaG, DnaA and most of them also encode the RelA-SpoT proteins whose synthase and hydrolase activities control the level of intracellular (p)ppGpp. However, the stringent response apparently does not interfere with DNA replication in *E. coli*. It will now be important to determine if the mechanism coupling the initiation and elongation steps of DNA replication with protein synthesis via the stringent response is widespread or not in Bacteria.

The bacterial Ogb proteins are both involved in ribosome biogenesis and DNA replication

Another mechanism potentially coupling DNA synthesis and translation in Bacteria (and possibly in the other two domains of life) has recently emerged from studies of essential and abundant GTPases named Ogb or CgtA in Bacteria ((Foti et al., 2007) and references therein). These proteins probably act as sensors of the GTP/GDP intracellular pool. In *E. coli*, Obg/CgtA is associated with the 50S precursor of the large ribosome subunit in exponential phase, but dissociates in stationary phase or during the stringent response (Jiang et al., 2007). The Obg/CgtA protein thus appears to be involved in normal ribosome assembly but also to act as an inhibitor of the stringent response by promoting SpoT dependent hydrolysis of (p)ppGpp in normal growth conditions ((Jiang et al., 2007) and references therein). Obg/CgtA thus appears to be important to optimize the translational capacity of the cell. However, the major phenotype observed in *E. coli* cells depleted of Obg/CgtA (dubbed ObgE in this organism) corresponds to an abnormal pattern of chromosome segregation and cell division (Foti et al., 2007). Although DNA replication appears to occur normally in these cells, Foti

and co-workers have previously isolated an *obgE* mutant which is hypersensitive to hydroxyurea (HU), an inhibitor of ribonucleotide reductase (Foti et al., 2005). This suggests that Obg/CgtA may be required to prevent DNA replication fork collapse when the intracellular pool of dNTP is low. Wang and co-workers have thus suggested that Obg/CgtA could be one of the factors involved in a mechanism that stabilized replication forks arrested by increasing (p)ppGpp concentrations (Wang et al., 2007). In any case, Obg/CgtA is a good candidate to link cell cycle events with translation ability in Bacteria via its association with the ribosome, the chromosome, and/or the replication forks.

Complexes grouping DNA replication and ribosome biogenesis proteins in Eukarya.

A few studies have now provided direct evidence for the existence of proteins (or protein complexes) that could interact with both the ribosome and DNA replication proteins in eucaryotes too. Using a genetic screen to detect previously unknown initiation proteins for DNA replication in *S. cerevisiae*, Zhang et al. (Zhang et al., 2002) identified Noc3p, a protein previously found to be required for ribosome biogenesis, as a potential partner of the Origin of Replication Complex (ORC) and of the replicative helicase MCM. They found that Noc3p is associated to chromatin and directly bounds to origin recognition elements during cell cycle, and thus hypothesized that Noc3p is a multifunctional protein that could play a role in coordinating cellular division with cell growth (Zhang et al., 2002).

The same year, Du and Stillman (Du and Stillman, 2002) identified additional yeast proteins involved in ribosome biogenesis as putative partners of ORC using a co-immunoprecipitation approach. In particular, they showed that Yph1p (yeast pescadillo homolog), the level of which correlates with the rate of proliferation in response to energy sources, links ORC to a large complex of proteins (Du and Stillman, 2002). Interestingly, this Yph1p-associated complex includes replication proteins such as MCM and ORC, but also proteins involved in

the biogenesis of the 60S ribosomal subunit such as Nog1p (an Obg protein) and Erb1p (BOP1 in human), and several large subunit ribosomal proteins (Du and Stillman, 2002). Moreover, Yph1p is required for ribosome biogenesis and is essential for cells to exit from the G_0 state and initiate cell proliferation (Du and Stillman, 2002). Two years later, it was shown that yeast Rrb1p (called GRWD in human), another protein involved in early ribosome assembly, also interacts with Yph1p and Orc6p (Killian et al., 2004). Transient depletion of the homologous proteins in human cells results in an increase of cells with abnormal mitoses (Killian et al., 2004). Killian and co-workers thus proposed that alteration of proteins linking ribosome biogenesis and DNA replication might directly cause chromosome instability and formation of tumors (Killian et al., 2004). Taken together, all these data strongly suggest the existence of a large protein network that connects ribosome biogenesis, DNA replication, and chromosome segregation in Eukarya. However, the precise molecular mechanism(s) by which this network operates remains totally unknown at present.

A potential link between DNA replication and translation in Archaea

We have recently obtained indirect evidence for the existence of a mechanism that could link protein synthesis to DNA replication by analyzing the genomic environment of genes encoding DNA replication proteins in Archaea (Berthon et al., 2008). We observed a large cluster of seven consecutive genes encoding both DNA replication and translation proteins, which is conserved in several genomes of Crenarchaeota and partly conserved in most other archaeal genomes (including very distantly related organisms) (Berthon et al., 2008). This cluster includes the genes encoding PCNA (the clamp that tightly tethers several DNA replication and repair proteins to DNA), PriS (the small subunit of the DNA primase), Gins15 (one of the two subunits of the archaeal GINS complex), the ribosomal protein L44E, the ribosomal protein S27E, the alpha subunit of the translation initiation factor aIF-2, and Nop10

(a protein involved in ribosome biogenesis). We called this set of seven genes 'the PPsGLSIN cluster (Berthon et al., 2008). Noteworthy, whereas the gene encoding the alpha subunit of aIF-2 belongs to the PPsGLSIN cluster (that also includes Gins15), the gene encoding the beta subunit is adjacent to the genes encoding Mcm and Gins23 in several archaeal genomes (Berthon et al., 2008). These observations led us to suggest the existence of unknown physical and/or functional relationships between specific sets of archaeal proteins involved in ribosome biogenesis and translation with proteins involved in DNA replication (Berthon et al., 2008). According to the observed gene associations, a large complex containing several replication factors (PCNA, DNA primase, GINS and MCM) and ribosome-associated proteins (L44E, S27E, aIF-2 and Nop10) would assemble in the cell under specific conditions.

The presence of MCM in such a protein complex could explain a puzzling observation made recently by Matsunaga and co-workers (Matsunaga et al., 2007). These authors reported that MCM from the archaeon *P. abyssi* binds preferentially to the replication origin in exponential growth phase, but to the ribosomal operon in stationary phase, as attested by ChI-chip analysis (Matsunaga et al., 2007). One could speculate that binding of MCM to the rDNA operon or to the ribosome factory inhibits ribosome biogenesis in stationary phase and/or that MCM is sequestered in stationary phase by stalled ribosome factories, preventing its participation to the replication initiation process. This hypothetical scenario would explain why inhibition of protein synthesis by puromycin in *P. abyssi* does not lead to a general decrease in the cellular amount of MCM but removes MCM from the replication origin (Matsunaga et al., 2001). Both *in silico* and experimental observations thus suggest a link between MCM and ribosome biogenesis in Archaea (Berthon et al., 2008; Matsunaga et al., 2001; Matsunaga et al., 2007), which is reminiscent of the physical association between MCM2-7 and the Yhp1 complex in Eukarya (Du and Stillman, 2002).

Is this putative functional coupling linked to cancer formation in Eukarya?

Human homologues of S27E and L44E are oncoproteins

Interestingly, both MCM and all proteins encoded by the PPsGLSIN cluster have homologues in Eukarya. This suggests that the putative underlying regulatory mechanism might be present in both Archaea and Eukarya. The occurrence of aIF-2 in this protein network is especially relevant since IF-2 is a major target for protein synthesis regulation in eukaryotes (i.e. phosphorylation of IF-2 inhibits translation at the initiation step). Moreover, the presence of Nop10, a key component of the rRNA maturation apparatus in both Archaea and Eukarya (see (Berthon et al., 2008) for references), is reminiscent of the connection between DNA replication and ribosome biogenesis. Finally, and most interestingly, the human homologues of the two archaeal ribosomal proteins S27E (known as MPS-1 or RPS27; called thereafter MPS1/RPS27) and L44E (RPL36A in human) are both involved in the control of cell growth and are overexpressed in many tumor cell lines (Kim et al., 2004; Wang et al., 2006). However, the mechanism of action of each of these two ribosomal proteins in cancer formation remains elusive (Kim et al., 2004; Wang et al., 2006). Considering that the gene encoding the archaeal homologues of MPS-1/RPS27 and RPL36A are adjacent to the gene encoding Nop10 and to genes coding for DNA replication proteins (Berthon et al., 2008), a role in coupling ribosome biogenesis and DNA replication could account for the association of MPS-1/RPS27 and RPL36A with cancer formation. Interestingly, MPS-1/RPS27 was characterized biochemically as a nuclear zinc-finger phosphoprotein that binds to DNA (Fernandez-Pol et al., 1994). It is therefore tempting to speculate that MPS-1/RPS27 binds to the replication fork and/or to regulatory sequences involved in the initiation of DNA replication.

A second human orthologue of the archaeal ribosomal protein S27E under the control of p53

Excitingly, two research groups have recently discovered and analyzed a second human homologue of RPS27 (with only three amino acid differences with MPS-1/RPS27) in screening by chip-profiling experiments for new p53 regulated genes (He and Sun, 2007; Li et al., 2007). Unlike its paralogue MPS-1/RPS27, the expression of this new protein (dubbed RPS27L, for RPS27-like) is induced by p53 in multiple cancer cell models (He and Sun, 2007). RPS27L is a nuclear protein that seems to modulate the DNA-damage-p53 response, since depletion of RPS27L results in deficiency in DNA damage checkpoints and finally leads to apoptosis (Li et al., 2007). After treatment with a DNA-damaging agent, RPS27L appear to be recruited to a subset of DNA breaks where it forms foci (Li et al., 2007). The response of cells lacking RPS27L to the genotoxic agent adriamycin—a drug which produces covalent Topo II DNA complexes (cleavable complexes), potentially disrupting replication fork—suggests that RPS27L prevents replication forks from moving through damaged DNA, so as to preserve genome integrity. Indeed, whereas DNA synthesis decreases in normal cells treated with adriamycin, this diminution is partially rescued in cells lacking RPS27L and cells enter S phase prematurely (Li et al., 2007). The genomic context of the archaeal homologues of *RPS27* in archaeal genomes suggests that the human RPS27L protein could directly interact with protein(s) of the replication factory to inhibit DNA synthesis in DNA-damaged cells.

Coupling DNA replication and translation in Archaea and Eukarya: an hypothetical model

The genomic contexts of the genes encoding the archaeal homologues of RPS27 (S27E) and RPL36A (L44E) suggest some hypotheses on their mechanisms of action that will be worth testing. Indeed, the association of the genes encoding S27E, L44E and Gins15 is highly

conserved among archaeal genomes, strongly suggesting that these two ribosomal proteins may interact with the GINS complex. On the other hand, the GINS complex probably interacts with the DNA primase and MCM in both Archaea and Eukarya (reviewed in (Labib and Gambus, 2007)). One could speculate that the GINS complex thus links L44E and S27E to the replisome via the DNA primase and MCM in a dynamic manner. For example, depending on physiological conditions, L44E, S27E and GINS may regulate the loading of MCM either to the replication origin or to the ribosome factory. It will now be important to test such hypotheses by looking directly at the effect of L44E and S27E (and their human counterparts RPS36A and RPS27/MPS-1) on archaeal (and eukaryal) DNA replication proteins in various model systems.

Three domains to play with

It is remarkable that indications for a tight coupling of DNA replication and protein synthesis were obtained more or less simultaneously and independently by researchers working on different cellular domains. This suggests that much more has to be learnt from comparative biochemistry. Of course, the precise mechanisms of the coupling should be different in each domain (most proteins of the eukaryotic Yph1 complex have no archaeal homologues and the stringent response as we know it is specific for bacteria, see below), but common themes or even common mechanisms could exist. The study of the GTPases Obg/Nog1 could be especially rewarding since these proteins are indeed universal (they are also present in Archaea). For instance it would be important to determine if Nog1p and its homologues control DNA replication by sensing the nucleotide pool in Archaea and Eukarya, as hypothesized for Obg in Bacteria; such mechanism has indeed been documented in yeast, although the protein(s) involved are unknown (Koc et al., 2004). The bacterial type stringent response is absent in Archaea and most Eukarya (except plants), which do not harbor neither

RelA-SpoT homologues nor (p)ppGpp. However, several lines of evidence suggest that mechanisms analogous to the bacterial stringent response exist in these two domains of life. In particular, it has been shown that pseudomonic acid (an antibiotic that prevents the charging of tRNA, mimicking the effect of amino acid starvation) inhibits rRNA synthesis in the archaeon *Sulfolobus acidocaldarius* (Cellini et al., 2004). It would be now really exciting to identify alarmones and/or proteins involved in the archaeal stringent response and to determine if their induction also inhibits DNA replication. Considering the close similarity of the translation and DNA replication systems in Archaea and Eukarya, archaeal cells could be excellent models to identify proteins that could be operational in the eukaryotic stringent response and to investigate possible connections between this response and DNA replication. The eukaryotic Obg-like proteins and/or complexes containing ribosome biogenesis factors together with replication protein (Du and Stillman, 2002; Zhang et al., 2002), as well as the proteins encoded by the PPsSLSIN cluster, uncovered by comparative genomics (Berthon et al., 2008), are good candidates to participate in such critical cellular regulatory networks. More studies in these areas will be certainly rewarding. In fact, the realization that a strong connection probably exists between DNA replication and translation is timely at the dawn of the system biology era, when the attention of biologists is now more focus on the integration of molecular systems in the whole cell.

Selected reading

Berthon, J., Cortez, D., and Forterre, P. (2008). Genomic context analysis in Archaea suggests previously unrecognized links between DNA replication and translation. Genome biology *9*, R71.

Cellini, A., Scoarughi, G.L., Poggiali, P., Santino, I., Sessa, R., Donini, P., and Cimmino, C. (2004). Stringent control in the archaeal genus Sulfolobus. Research in microbiology *155*, 98-104.

Du, Y.C., and Stillman, B. (2002). Yph1p, an ORC-interacting protein: potential links between cell proliferation control, DNA replication, and ribosome biogenesis. Cell *109*, 835-848.

Fernandez-Pol, J.A., Klos, D.J., and Hamilton, P.D. (1994). Metallopanstimulin gene product produced in a baculovirus expression system is a nuclear phosphoprotein that binds to DNA. Cell Growth Differ *5*, 811-825.

Foti, J.J., Persky, N.S., Ferullo, D.J., and Lovett, S.T. (2007). Chromosome segregation control by Escherichia coli ObgE GTPase. Molecular microbiology *65*, 569-581.

Foti, J.J., Schienda, J., Sutera, V.A., Jr., and Lovett, S.T. (2005). A bacterial G protein-mediated response to replication arrest. Molecular cell *17*, 549-560.

He, H., and Sun, Y. (2007). Ribosomal protein S27L is a direct p53 target that regulates apoptosis. Oncogene *26*, 2707-2716.

Jiang, M., Sullivan, S.M., Wout, P.K., and Maddock, J.R. (2007). G-protein control of the ribosome-associated stress response protein SpoT. J Bacteriol *189*, 6140-6147.

Killian, A., Le Meur, N., Sesboue, R., Bourguignon, J., Bougeard, G., Gautherot, J., Bastard, C., Frebourg, T., and Flaman, J.M. (2004). Inactivation of the RRB1-Pescadillo pathway involved in ribosome biogenesis induces chromosomal instability. Oncogene *23*, 8597-8602.

Kim, J.H., You, K.R., Kim, I.H., Cho, B.H., Kim, C.Y., and Kim, D.G. (2004). Over-expression of the ribosomal protein L36a gene is associated with cellular proliferation in hepatocellular carcinoma. Hepatology *39*, 129-138.

Koc, A., Wheeler, L.J., Mathews, C.K., and Merrill, G.F. (2004). Hydroxyurea arrests DNA replication by a mechanism that preserves basal dNTP pools. The Journal of biological chemistry *279*, 223-230.

Labib, K., and Gambus, A. (2007). A key role for the GINS complex at DNA replication forks. Trends Cell Biol.

Lesley, J.A., and Shapiro, L. (2008). SpoT Regulates DnaA Stability and the Initiation of DNA Replication in Carbon Starved Caulobacter crescentus. J Bacteriol.

Li, J., Tan, J., Zhuang, L., Banerjee, B., Yang, X., Chau, J.F., Lee, P.L., Hande, M.P., Li, B., and Yu, Q. (2007). Ribosomal protein S27-like, a p53-inducible modulator of cell fate in response to genotoxic stress. Cancer research *67*, 11317-11326.

Matsunaga, F., Forterre, P., Ishino, Y., and Myllykallio, H. (2001). In vivo interactions of archaeal Cdc6/Orc1 and minichromosome maintenance proteins with the replication origin. Proceedings of the National Academy of Sciences of the United States of America *98*, 11152-11157.

Matsunaga, F., Glatigny, A., Mucchielli-Giorgi, M.H., Agier, N., Delacroix, H., Marisa, L., Durosay, P., Ishino, Y., Aggerbeck, L., and Forterre, P. (2007). Genomewide and biochemical analyses of DNA-binding activity of Cdc6/Orc1 and Mcm proteins in Pyrococcus sp. Nucleic Acids Res *35*, 3214-3222.

Wang, J.D., Sanders, G.M., and Grossman, A.D. (2007). Nutritional control of elongation of DNA replication by (p)ppGpp. Cell *128*, 865-875.

Wang, Y.W., Qu, Y., Li, J.F., Chen, X.H., Liu, B.Y., Gu, Q.L., and Zhu, Z.G. (2006). In vitro and in vivo evidence of metallopanstimulin-1 in gastric cancer progression and tumorigenicity. Clin Cancer Res *12*, 4965-4973.

Zhang, Y., Yu, Z., Fu, X., and Liang, C. (2002). Noc3p, a bHLH protein, plays an integral role in the initiation of DNA replication in budding yeast. Cell *109*, 849-860.

Chapitre III

Recherche d'interactions physiques entre proteines

Chapitre III : Recherche d'interactions physiques entre protéines

Introduction

Dans le chapitre précédent ont été présentés les résultats d'une analyse comparative du contexte génomique des gènes de la réplication chez les Archaea. Les principaux résultats de cette étude sont les suivants :

- quelques gènes de la réplication sont associés en des structures de type opéron conservées entre des organismes archéens éloignés, ce qui suggère que les produits d'expression de ces gènes interagissent physiquement, voire fonctionnellement au sein du réplisome ;
- certains gènes de la réplication sont liés de manière fréquente à des gènes codant des protéines impliquées dans d'autres processus cellulaires de transactions de l'information génétique (traduction, transcription, réparation) dans quelques génomes d'Archaea, ce qui pourrait signifier qu'il existe des voies de communication entre la réplication et chacun de ces différents mécanismes afin de coordonner leur action ;
- une structure de type opéron comprenant, sous sa forme entière, trois gènes de la réplication et quatre gènes codant des protéines associées à la traduction ou à la biogenèse des ribosomes est conservée, tout ou partie, dans la quasi-totalité des génomes d'Archaea. Cette association génomique suggère l'existence d'un mécanisme de couplage entre la réplication de l'ADN et la synthèse des protéines.

Le principe de base de ces interprétations est que des gènes associés de manière récurrente dans des génomes éloignés codent des protéines qui, généralement, interagissent de manière physique, voire fonctionnelle (Dandekar et al., 1998; Overbeek et al., 1999). Les hypothèses énoncées peuvent donc être aisément confirmées ou infirmées par une approche expérimentale adéquate. Aussi, nous avons entrepris quelques expériences *in vitro* afin d'étayer les hypothèses émises à partir de l'interprétation des données obtenues *in silico*.

Dans un premier temps, mon travail s'est porté sur la recherche d'une interaction physique entre l'hélicase réplicative MCM et la sous-unité béta du facteur d'initiation de la traduction aIF-2 (aIF-2 beta) par co-immunoprécipitation. Une interaction physique entre le PCNA et l'ADN primase a également été testée à l'aide de la technique de résonance du plasmon de surface. Dans un second temps, la totalité des gènes pour lesquels des associations génétiques récurrentes ont été mentionnées et discutées dans le cadre de l'analyse comparative du contexte génomique, ainsi que d'autres cibles biologiques d'intérêt, ont été clonés dans des vecteurs d'expression dans le but de procéder à un criblage des interactions par une méthode de co-purification. Enfin, le protocole de co-purification (transformation, co-expression et purification) a été éprouvé et optimisé avec les sous-unités du complexe GINS afin de disposer d'une méthode fiable et reproductible.

Résultats

Recherche d'interaction par co-immunoprécipitation

Le gène codant la sous-unité beta du facteur d'initiation de la traduction aIF-2 (aIF-2β) a été cloné dans un vecteur d'expression bactérien puis la protéine recombinante a été exprimée chez *E. coli* avant d'être purifiée par chromatographie d'affinité suivant la méthode décrite par Tahara et collaborateurs (Tahara et al., 2004). Le gène codant la protéine aIF-2β étant

contigu au gène codant la protéine MCM dans plusieurs génomes d'Euryarchaeota dont celui de *P. furiosus* (voir article, Chapitre II), l'interaction entre les deux protéines a été testée. Les deux protéines ont été mélangées en proportions équimolaires et incubées à haute température (70°C) avant que la protéine MCM soit immunoprécipitée et le signal révélé par chimio-luminescence. Aucune interaction entre les protéines aIF-2β et MCM n'a pu être détectée dans les conditions choisies, ce qui suggère qu'elles n'interagissent pas physiquement l'une avec l'autre. Néanmoins, il est possible que les conditions utilisées soient défavorables, soit que certains paramètres demandent à être optimisés (tampon, température), soit que cette interaction ne soit possible que si certaines circonstances sont remplies : i) modification chimique de l'une des deux protéines (phosphorylation, acétylation) ; ii) changement de conformation ; iii) présence d'un cofacteur (ATP, GTP) ; iv) intervention d'une protéine auxiliaire ; v) présence des autres sous-unités (α et γ) du facteur d'initiation aIF-2.

Recherche d'interactions par la technique de résonance du plasmon de surface

Le gène codant le PCNA est souvent associé au(x) gène(s) codant l'ADN primase dans les génomes d'Archaea (voir article, Chapitre II). Aussi, l'interaction physique entre les deux protéines a été examinée à l'aide de la technique de résonance du plasmon de surface. Une préparation de l'ADN primase (hétérodimère PriSL) de *P. furiosus* (Liu et al., 2001) a été injectée au dessus d'une puce à la surface de laquelle la protéine *Pfu*PCNA était fixée (Kiyonari et al., 2006). Aucune interaction n'a été observée dans les conditions de mesure. L'absence d'interactions est plutôt surprenante sachant que l'association entre le gène codant le PCNA et le(s) gènes codant la primase est assez bien conservée entre des génomes éloignés (voir article, Chapitre II). En outre, les résultats d'une analyse de l'expression génétique au cours du cycle cellulaire chez *S. solfataricus* suggèrent que les gènes codant le PCNA et PriS sont probablement co-transcrits chez cet organisme ce qui accrédite d'autant plus l'idée que

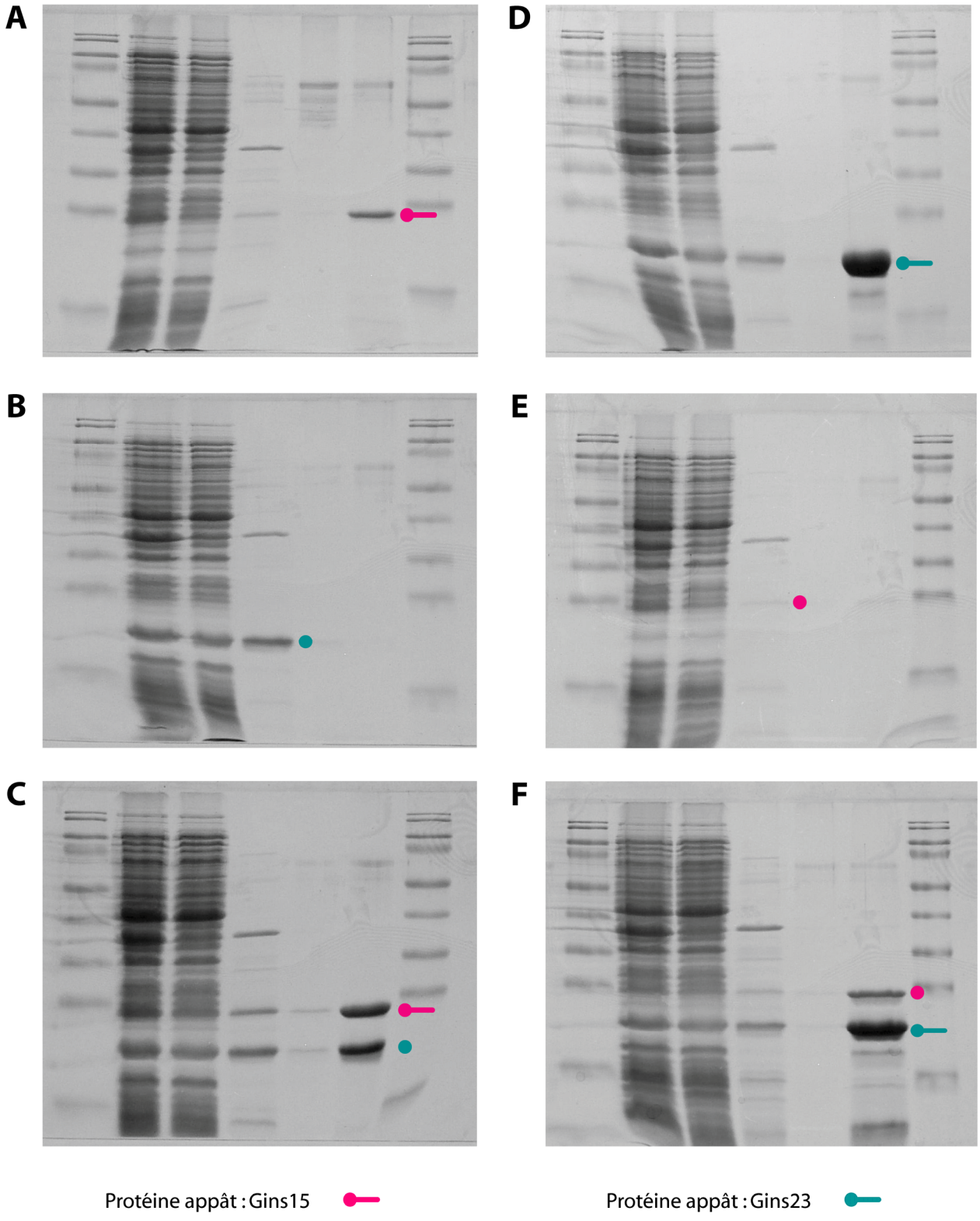

Figure III-1 : Co-purification des sous-unités Gins15 et Gins23 sur résine d'affinité Ni-NTA. (A) Expression et purification témoin de la protéine appât *Pfu*Gins15 (magenta) fusionnée à une étiquette hexahistidine. (B) Expression et purification témoin de la protéine cible *Pfu*Gins23 (vert) fusionnée à une étiquette streptavidine ; la protéine n'est pas retenue sur la résine Ni-NTA. (C) Co-expression et co-purification des protéines *Pfu*Gins15 et *Pfu*Gins23. (D) Expression et purification témoin de la protéine appât *Pfu*Gins23 fusionnée à une étiquette hexahistidine. (E) Expression et purification témoin de la protéine cible *Pfu*Gins15 fusionnée à une étiquette streptavidine ; la protéine n'est pas retenue sur la résine Ni-NTA. (F) Co-expression et co-purification des protéines *Pfu*Gins23 et *Pfu*Gins15.

ces deux protéines interagissent (Lundgren and Bernander, 2007). Le fait qu'aucune interaction n'ait pu être observée entre le PCNA et l'hétérodimère PriSL pourrait s'expliquer de plusieurs façons : i) le PCNA est capable d'interagir avec chacune des deux sous-unités de l'ADN primase dans leur état monomérique mais pas lorsque celles-ci sont associées en un complexe car la surface d'interaction entre le PCNA et chacune des sous-unités n'est pas accessible à l'état hétérodimère PriSL ; ii) le PCNA ou l'ADN primase doit préalablement être la cible d'une modification chimique ; iii) l'interaction entre les deux protéines est subordonnée à la réception d'un stimulus extracellulaire ou à l'apparition d'une lésion dans l'ADN.

Recherche d'interactions par co-purification

Une série de gènes (voir Matériels et Méthodes) ont été clonés dans des vecteurs bactériens dans le but de rechercher des interactions physiques entre protéines deux à deux. Le principe de la méthode consiste à co-exprimer chez *E coli* une protéine fusionnée à une étiquette hexahistidine en présence d'un partenaire protéique potentiel puis de vérifier si l'analyte co-purifie avec la protéine immobilisée sur la matrice d'affinité Ni-NTA. Les paramètres de purification, en particulier la concentration en imidazole du tampon de lavage, ont été optimisés avec le couple modèle Gins15-Gins23. L'interaction entre les deux protéines est détectée indépendamment de la nature de l'analyte, ce qui indique que l'interaction est réciproque **(Figure III-1)**. De manière remarquable, les sous-unités Gins sont obtenues en proportions stœchiométriques uniquement lorsque la protéine Gins15 est immobilisée sur la colonne d'affinité. Ceci s'explique par le fait que la quantité de forme soluble de la protéine Gins15 correspond au facteur limitant ; en l'absence de son partenaire Gins23 la protéine Gins15 est insoluble (Marinsek et al., 2006; Yoshimochi et al., 2008).

Conclusions & Perspectives

Deux des interactions prédites par l'analyse du contexte génomique ont été éprouvées expérimentalement mais aucune interaction physique n'a pu être détectée avec les approches choisies. Afin de pouvoir analyser un grand nombre d'interactions physiques sans avoir à disposer de protéines préalablement purifiées, nous avons opté pour un criblage selon une méthode de co-purification sur une résine d'affinité réalisée dans un tube à centrifuger de 2 ml initialement développée dans le laboratoire de Génomique Structurale dirigé par Herman van Tilbeurgh. Le protocole a été optimisé avec les sous-unités Gins15 et Gins23 du complexe GINS. La technique peut désormais être mise en œuvre pour analyser de façon systématique des interactions entre de nombreux partenaires protéiques potentiels.

Chapitre IV

Analyse phylétique des gènes de la réplication et évolution de la machinerie de réplication chez les Archaea

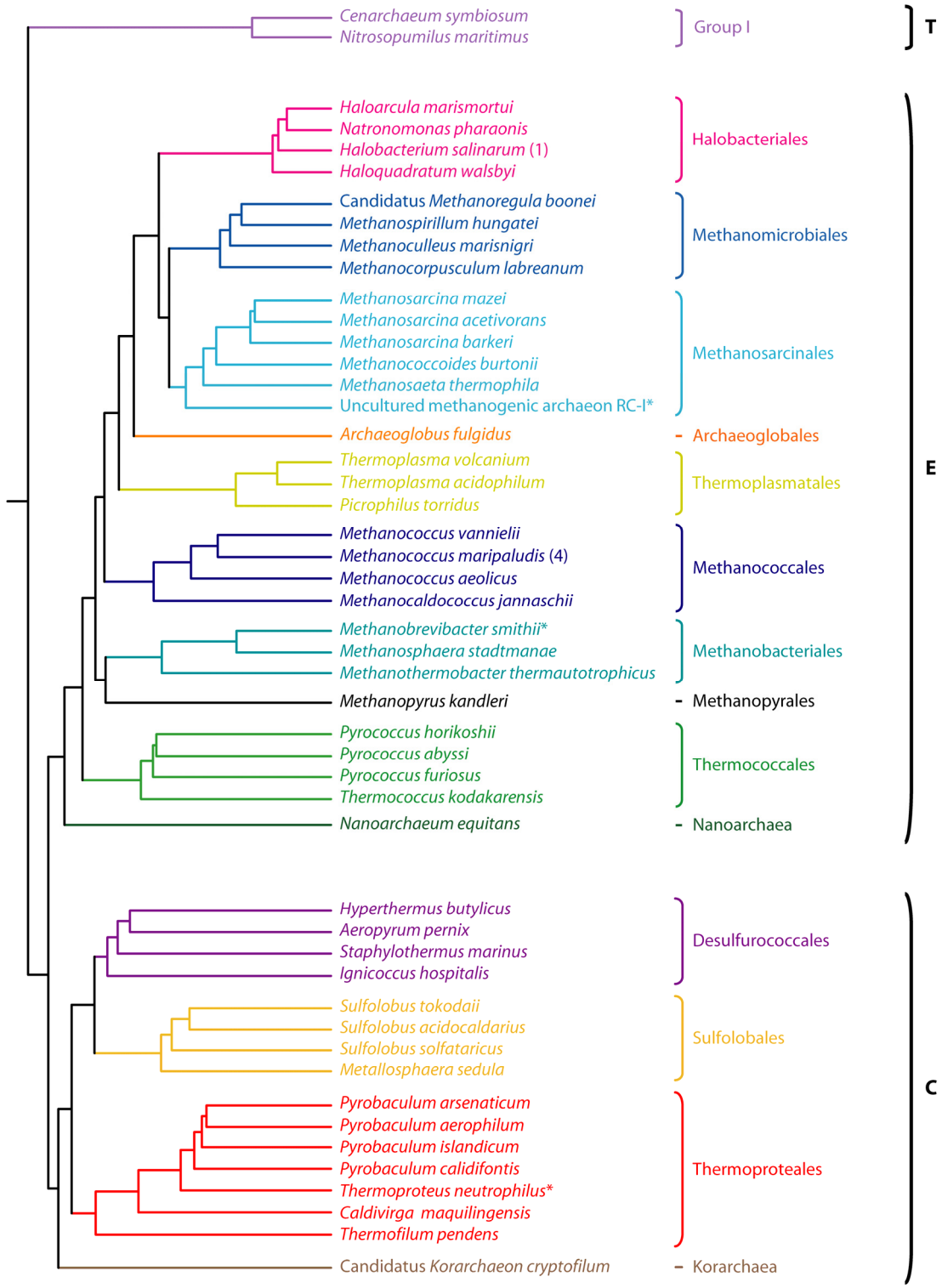

Figure IV-1 : Phylogénie consensuelle des Archaea. Cette phylogénie des Archaea correspond à une compilation d'analyses récentes basées sur différents marqueurs moléculaires (à l'exception des organismes signalés par un astérisque). Ces analyses soutiennent l'existence de trois phylums principaux au sein du domaine Archaea : les Thaumarchaeota, les Crenarchaeota et les Euryarchaeota (Brochier-Armanet et al., 2008). La position du méthanogène non cultivé isolé à partir d'une rizière 'Uncultured methanogenic archaeon RC-I' (Methanosarcinales) est basé sur l'analyse de l'ARNr 16S (Erkel et al., 2006). Les positions de *Thermoproteus neutrophilus* (Thermoproteales) et *Methanobrevibacter smithii* (Methanobacteriales) ont été inférées respectivement à partir de la classification taxonomique et du contenu génomique (Samuel et al., 2007) ; elles sont donc provisoires.
C : Crenarchaeota ; E : Euryarchaeota ; T : Thaumarchaeota.

Chapitre IV : Analyse du profil phylétique des gènes de la réplication et évolution de la machinerie de réplication chez les Archaea

Introduction

La génomique comparative a permis de mettre en évidence l'existence d'un ensemble de gènes qui caractérisent les Archaea et à partir desquels l'histoire évolutive de ce domaine peut être retracée (Forterre et al., 2007a; Gribaldo and Brochier-Armanet, 2006). Les archées dont le génome a été complètement séquencé se repartissent probablement en trois phylums distincts : les Crenarchaeota, les Euryarchaeota et les Thaumarchaeota (Brochier-Armanet et al., 2008). A l'image des deux génomes séquencés des Thaumarchaeota, le génome de l'archée hyperthermophile Candidatus *Korarchaeum cryptofilum* présente un contenu composite de gènes typiquement crénarchéen et de gènes typiquement euryarchéen, mais cet organisme semble se positionner à la base des Crenarchaeota (Elkins et al., 2008). En se basant sur cette représentation récente de la phylogénie des Archaea **(Figure IV-1)**, nous avons analysé la distribution de chacun des gènes de la réplication dans cinquante-deux génomes d'Archaea. (Les génomes des quatre souches de *M. maripaludis* ont été analysés mais seul le contenu de la souche S2 est indiqué pour plus de clarté car ces génomes ont un contenu identique dans la plupart des cas examinés (à l'exception notable du nombre de copies du gène *mcm*, voir ci-après). Les génomes des souches des organismes *Halobacterium* sp. NRC1 (Ng et al., 2000) et *H. salinarum* R1 (Pfeiffer et al., 2008a) ont également été analysés individuellement. Pour autant, seul le contenu du génome de l'archée *H. salinarum* R1 est présenté ci-après dans la mesure où i) les deux génomes sont virtuellement identiques

(Pfeiffer et al., 2008a) — même si cet avis fait l'objet d'une controverse (Ng et al., 2000; Pfeiffer et al., 2008b) ; ii) l'annotation du génome de la souche R1, qui a été en partie vérifiée par des analyses protéomiques, apparaît plus fiable.) La distribution des gènes de la réplication chez les Archaea est illustrée par la suite sous la forme de profils phylétiques s'appuyant sur la représentation à trois phylums. Dans la plupart des cas, la distribution des gènes à travers les différents groupes taxonomiques donne une idée de leur histoire évolutive respective. Aussi, nous avons cherché à interpréter le profil phylétique de chacun des gènes de la réplication selon une logique de parcimonie afin de reconstruire la composition probable de la machinerie de réplication de l'ADN chez le dernier ancêtre commun des Archaea.

Résultats & Discussion

Cdc6/Orc1

Le nombre de gènes *cdc6/orc1* varie de manière notable d'un groupe d'organismes à un autre ; alors que certains organismes présentent un seul homologue, d'autres en possèdent dix-sept, avec une moyenne qui se situe entre deux et trois copies du gène. La position et le nombre d'origines de réplication sont difficiles à prédire pour un certain nombre de génomes (pour une revue récente, voir (Zhang and Zhang, 2005)). En outre, le nombre d'origines de réplication ayant fait l'objet d'une confirmation expérimentale est restreint. Il n'est donc pas possible de savoir s'il existe une corrélation entre le nombre de gènes *cdc6/orc1* et le nombre d'origines de réplication dans les génomes d'Archaea.

La plupart des génomes de Crenarchaeota présente de deux à trois homologues du gène *cdc6/orc1* à l'exception des organismes de la famille des Thermoproteaceae (*Caldivirga, Pyrobaculum*, *Thermoproteus*). Chez ces organismes, seul un gène *cdc6/orc1* est identifiable, alors que l'analyse du génome de *Thermofilum pendens*, un organisme apparenté, révèle la présence de trois homologues **(Figure IV-2)**. Sachant que les génomes des organismes

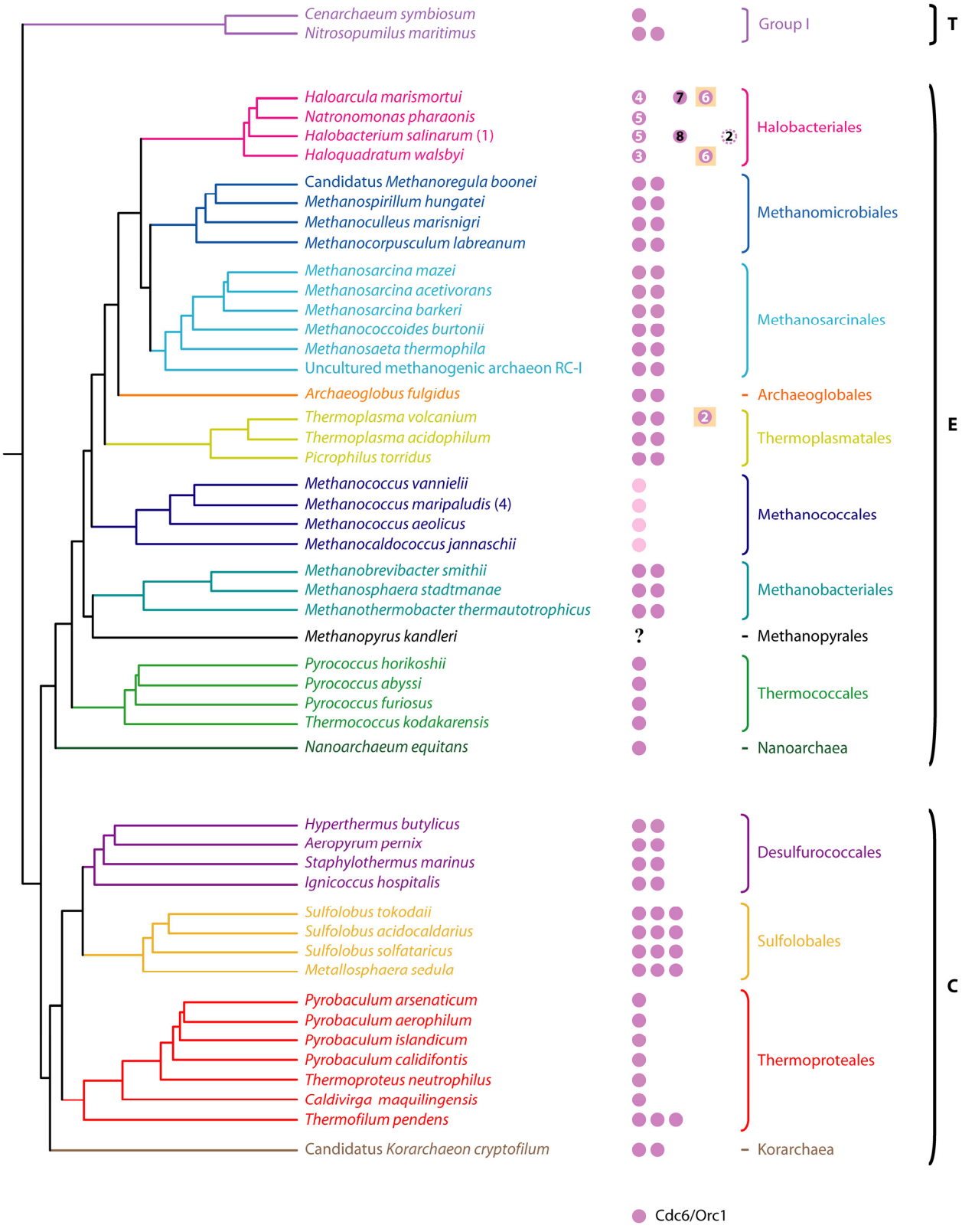

Figure IV-2 : Profil phylétique du gène codant la protéine Cdc6/Orc1 chez les Archaea. La présence d'un homologue Cdc6/Orc1 codé par un gène chromosomique est signalée par un disque couleur lavande. Au-delà de deux (E), le nombre d'homologues est indiqué par un chiffre de couleur blanche pour les gènes chromosomiques, noire pour les gènes plasmidiques (Halobacteriales). Les gènes qui pourraient avoir une origine extra-chromosomique se distinguent par le carré orange clair situé en arrière plan. Les deux pseudogènes plasmidiques du génome d'*Halobacterium salinarum* R1 sont indiqués par un cercle en pointillé. Les protéines Cdc6/Orc1 divergentes des Methanococcales sont signalées par un disque rose bonbon. Le point d'interrogation indique que la protéine initiatrice de la réplication de *Methanopyrus kandleri* demeure non identifiée.

appartenant à l'ordre des Sulfolobales et des Desulfurococcales présentent respectivement trois et deux gènes *cdc6/orc1*, l'ancêtre commun de ces trois ordres possédait probablement déjà deux gènes *cdc6/orc1*. En outre, le génome de *K. cryptofilum*, dont l'affiliation n'est pas clairement établie, mais qui semble correspondre à une lignée branchant à la base des Crenarchaeota, possède également deux gènes *cdc6/orc1* **(Figure IV-2)**. Par conséquent, il est vraisemblable que ces deux gènes *cdc6/orc1* aient été hérités de l'ancêtre commun des Crenarchaeota (étant entendu que *K. cryptofilum* appartient à ce phylum). En outre, cela suggère que la présence d'une seule copie du gène *cdc6/orc1* chez les Thermoproteaceae résulte d'une perte, vraisemblablement chez l'ancêtre commun de ce groupe taxonomique.

Chez les Euryarchaeota, les génomes arborent le plus souvent un ou deux gènes *cdc6/orc1* à l'exception des organismes de l'ordre des Halobacteriales chez lesquels les gènes *cdc6/orc1* sont particulièrement nombreux et divergents (voir, (Berquist et al., 2007)). Néanmoins, il semble que certains de ces gènes ne soient pas essentiels dans des conditions normales de croissance comme le suggèrent les résultats d'une analyse génétique chez *Halobacterium* sp. (Berquist et al., 2007). Parmi ces gènes non essentiels, certains pourraient coder des protéines mutées tandis que d'autres pourraient correspondre à des protéines initiatrices alternatives. Ces protéines initiatrices alternatives pourraient intervenir dans différentes conditions de croissance, par exemple selon la variation de la température du milieu, à la manière de ce qui a été proposé pour la transcription des différents opérons ribosomiques chez *H. marismortui* (Lopez-Lopez et al., 2007). L'origine de ces gènes *cdc6/orc1* surnuméraires n'est pas claire mais certains homologues se trouvent dans des régions du génome correspondant à des éléments intégrés (Diego Cortez, communication personnelle) ou à proximité d'un gène codant une transposase, ce qui suggère qu'ils pourraient avoir une origine extra-chromosomique. Parmi ces nombreux gènes *cdc6/orc1*, certains sont tronqués par rapport aux autres homologues archéens, en particulier dans le

génome de *H. walsbyi* chez lequel on observe respectivement quatre et deux phases ouvertes de lecture présentant des similarités avec le gène *cdc6/orc1* dans deux régions génomiques qui sont sans doute d'origine extra-chromosomique. Par ailleurs, un grand nombre de gènes *cdc6/orc1* sont localisés sur des plasmides d'Archaea, en particulier chez les archées halophiles (Baliga et al., 2004; Ng et al., 2000; Norais et al., 2007) mais aussi chez *M. thermautotrophicum* et *T. acidophilum* (Smith et al., 1997; Yamashiro et al., 2006). Par conséquent, les plasmides ont pu contribuer à la dispersion et à la duplication des gènes *cdc6/orc1* dans certains génomes d'Archaea. Cette hypothèse semble appuyée par le fait que deux gènes dans le génome de *T. volcanium* codent des protéines de type Cdc6 divergentes présentant une forte similarité avec l'homologue Cdc6 codé par l'ORF1 du plasmide pTA1 de *T. acidophilum*.

L'annotation des génomes de *M. jannaschii*, *M. kandleri* et *M. maripaludis* n'avait révélé aucun homologue patent de la protéine Cdc6/Orc1 (Bult et al., 1996; Hendrickson et al., 2004; Slesarev et al., 2002). Des éléments concordants permirent néanmoins de désigner un gène codant un homologue putatif de la protéine Cdc6 dans le génome de *M. jannaschii* (Liu et al., 2000; Zhang and Zhang, 2004), puis dans celui de *M. maripaludis* (voir, (Lundgren and Bernander, 2005)). En revanche, le gène codant la protéine initiatrice de réplication de l'ADN chez *M. kandleri* demeure à ce jour non identifiée. Les homologues Cdc6/Orc1 de *M. jannaschii* et *M. maripaludis* sont clairement apparentés et ils divergent fortement par rapport à la forme canonique de la protéine Cdc6/Orc1. De manière intéressante, l'analyse du profil phylétique du gène *cdc6/orc1* dans les génomes d'Archaea indique que la présence d'homologues Cdc6/Orc1 divergents concerne l'ensemble des organismes appartenant à l'ordre des Methanococcales **(Figure IV-2)** ; les gènes correspondant sont improprement annotés comme codant des régulateurs transcriptionnels chez l'ensemble de ces organismes. Par ailleurs, ces protéines forment un groupe

extrêmement cohérent, les séquences présentant des similarités supérieures à 68%, ce qui suggère que cette forme divergente était déjà présente chez le dernier ancêtre commun des Methanococcales. Ces formes divergentes de la protéine Cdc6/Orc1 présentent néanmoins des caractéristiques communes avec la forme canonique : i) un module AAA^+ avec une activité ATPase en position N-terminale, ii) un motif de fixation à l'ADN de type wHTH ('winged helix-turn-helix') en position C-terminale (Dueber et al., 2007; Gaudier et al., 2007; Liu et al., 2000; Singleton et al., 2004).

L'absence d'un gène codant une protéine de type Cdc6 dans le génome de *M. kandleri* est troublante sachant l'indispensabilité d'une telle protéine dans le cycle cellulaire. Les gènes codant les protéines initiatrices de la réplication ayant été mal annotés dans l'ensemble des génomes des Methanococcales, il pourrait en être de même chez *M. kandleri*. Aussi, le ou les homologues Cdc6/Orc1 chez *M. kandleri* auraient pu correspondre à des protéines présentant un motif wHTH en position C-terminale et improprement annotées comme facteurs de transcription, mais une recherche de ce type dans le génome de *M. kandleri* s'est révélée infructueuse. Par ailleurs, une recherche par TBLASTN n'a pas non plus permis de détecter un homologue Cdc6/Orc1 dans le génome de *M. kandleri*. Le gène codant la protéine initiatrice de la réplication chez cet organisme demeure donc non identifiée à ce jour.

Enfin, chez les Thaumarchaeota, on observe un gène *cdc6/orc1* chez *C. symbiosum* alors que le génome de *N. maritimus* en arbore deux. En l'absence d'un nombre plus conséquent de génomes de Thaumarchaeota, il n'est possible de statuer quant au nombre de copies du gène *cdc6/orc1* présente chez l'ancêtre de ce groupe. De la même façon, il est difficile d'inférer le nombre de copies du gène *cdc6/orc1* chez le dernier ancêtre commun des Archaea. Ce dernier possédait nécessairement au moins un homologue Cdc6/Orc1 et une origine de réplication. Le séquençage de nouvelles espèces de Thaumarchaeota devrait apporter un éclairage intéressant sur cette question.

GINS

Le complexe GINS a été caractérisé biochimiquement chez les archées *Sulfolobus solfataricus* (*Sso*) et *Pyrococcus furiosus* (*Pfu*) et dans chacune de ces deux études il a été montré que les protéines Gins15 et Gins23 s'assemblent en un complexe hétérotétramérique (Marinsek et al., 2006; Yoshimochi et al., 2008). Néanmoins, à la manière de ce qui a été observé chez les eucaryotes (pour des revues récentes, voir (Aparicio et al., 2006; Labib and Gambus, 2007)), le rôle du complexe GINS au cours de la réplication chez les Archaea n'est pas clairement établi. Ce complexe pourrait former un pont moléculaire entre les brins direct et retardé pour coordonner leur synthèse (Marinsek et al., 2006) ou bien il pourrait correspondre à un facteur auxiliaire de l'hélicase réplicative MCM (Yoshimochi et al., 2008). D'autre part, l'analyse des génomes d'archées suggère que la stœchiométrie du complexe GINS n'est peut-être pas conservée chez toutes les archées. En effet, si le gène codant la protéine Gins15 est bien représenté dans tous les génomes d'archées séquencés, celui codant la protéine Gins23 présente un profil phylétique singulier qui éclaire l'histoire évolutive de cette famille protéique et suggère que le rôle du complexe GINS pourrait différer selon les lignées archéennes considérées **(Figure IV-3)**.

En effet, le gène pour la protéine Gins23 est présent dans tous les génomes de Crenarchaeota, dans les génomes de *C. symbiosum*, de *N. maritimus* et dans les génomes des Thermococcales ; autrement dit ce gène est absent dans tous les génomes des Euryarchaeota à l'exception des Thermococcales **(Figure IV-3)**. Toutefois il est possible que tous les génomes d'archées possèdent bien un gène codant la protéine Gins23 si l'on suppose que la séquence de la majorité des homologues présents chez les Euryarchaeota a fortement divergée, rendant la recherche d'homologues par similarité de séquence infructueuse. Cette hypothèse est défendable sachant que la détection des homologues des protéines Gins23 peut être rendue difficile en raison de la permutation de domaine qui a eu lieu lors de l'évolution de ces

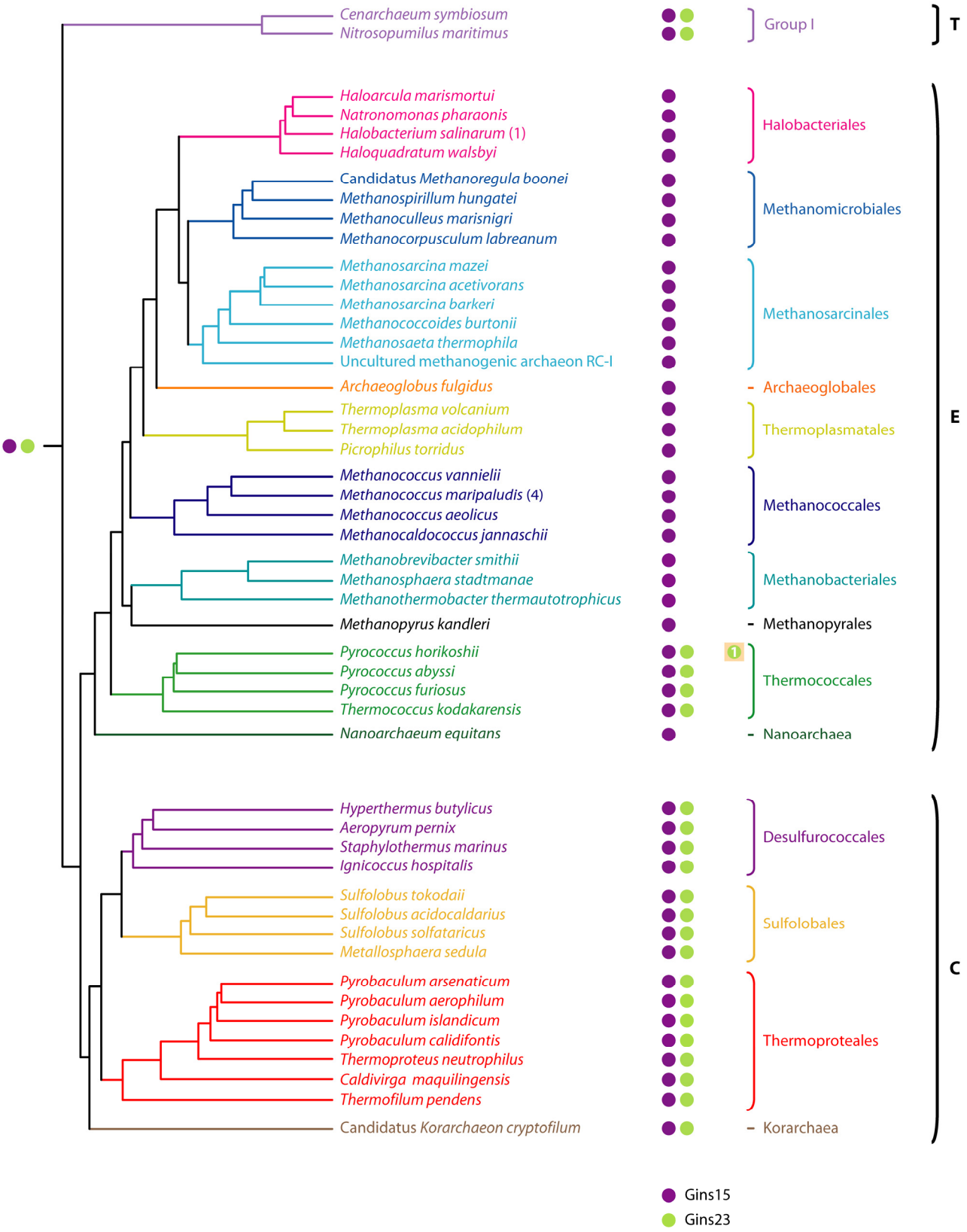

Figure IV-3 : Profil phylétique des gènes codant les sous-unités Gins15 et Gins23 du complexe GINS chez les Archaea. La présence d'un homologue Gins15 est signalée par un disque couleur prune. La présence d'un homologue Gins23 est indiquée par un disque couleur citron vert. Le gène codant la protéine Gins23 est absent des génomes d'Euryarchaeota (E) à l'exception des Thermococcales. Le deuxième homologue Gins23 chez *Pyrococcus horikoshii* pourrait avoir une origine extra-chromosomique.

protéines (voir, (Kamada et al., 2007; Makarova et al., 2005; Marinsek et al., 2006)). D'ailleurs, certains homologues eucaryotes du complexe GINS n'ont pu être clairement identifiés qu'après co-purification avec des partenaires cellulaires (Moyer et al., 2006; Takayama et al., 2003).

Indépendamment de la validité de l'un des deux scénarios précédents, la répartition observée des gènes *gins15* et *gins23* dans les génomes modernes suggère que les deux gènes *gins* étaient présents chez l'ancêtre commun des Archaea **(Figure IV-3)**, voire chez l'ancêtre commun des Eucarya et des Archaea, conformément à l'un des scénarios évolutifs précédemment proposés (Makarova et al., 2005). Si l'on considère l'absence du gène *gins23* dans la majorité des génomes d'Euryarchaeota comme avérée, cela signifie que le gène *gins23* aura été perdu après l'évènement de spéciation à partir duquel la lignée conduisant aux Thermococcales s'est individualisée ; le gène aura aussi été perdu indépendamment chez *N. equitans* si cet organisme représente bien une lignée proche des Thermococcales ayant évolué rapidement du fait de son style de vie parasitique (Brochier et al., 2005). Toujours selon l'hypothèse de la perte du gène *gins23* chez certaines Euryarchaeota, cette répartition des gènes suppose que la composition du complexe GINS diffère selon les lignées dans les organismes archéens contemporains. En outre, ceci pourrait impliquer que les propriétés du complexe GINS sont distinctes selon l'archée considérée. De façon remarquable, même si les caractérisations biochimiques des complexes *Sso*GINS et *Pfu*GINS ont mis en évidence des propriétés communes elles ont également mis en lumière un certain nombre de différences (architecture, stabilité des interactions physiques, nature des interactions fonctionnelles, nature des partenaires protéiques) qui pourraient être interprétées comme les signes d'une possible divergence fonctionnelle ancienne du complexe GINS chez les Archaea. De manière notable, le génome de *Pyrococcus horikoshii* arbore deux gènes *gins23* : le premier est contigu au gène *mcm* (voir discussion ci-après), le second se trouve à l'intérieur d'un élément

intégré (Diego Cortez, communication personnelle), ce qui suggère que cette copie provient d'un transfert horizontal.

Architecture. Les études biochimiques du complexe GINS chez *S. solfataricus* et *P. furiosus* ont en effet montré que les interactions protéiques entre les différentes sous-unités sont différentes d'un organisme à l'autre, ce qui suggère que l'architecture moléculaire de ces complexes pourrait être dissemblable. Dans le cas du complexe GINS de *S. solfataricus*, les analyses en double-hybride indiquent que la protéine *Sso*Gins15 interagit avec elle-même et avec la protéine *Sso*Gins23. En revanche, la protéine *Sso*Gins23 ne semble pas, selon la même approche, être capable d'interagir avec elle-même (Marinsek et al., 2006). Conformément aux résultats obtenus chez *S. solfataricus*, les analyses en double-hybride indiquent que les protéines *Pfu*Gins15 et *Pfu*Gins23 interagissent l'une avec l'autre. En revanche, contrairement à ce qui est observé pour *Sso*Gins23, la protéine *Pfu*Gins23 interagit fortement avec elle-même et forme un dimère en solution (Yoshimochi et al., 2008). Ces différences suggèrent que l'architecture moléculaire des complexes *Sso*GINS et *Pfu*GINS est différente. Cette divergence architecturale pourrait avoir des implications fonctionnelles si elle affecte les interfaces moléculaires de ces complexes, c'est-à-dire que la nature des partenaires protéiques pourrait différer d'un complexe à l'autre. La résolution de la structure de chacun de ces deux complexes permettrait de savoir si cette différence architecturale est réelle ou non.

Stabilité des interactions physiques. Chez tous les génomes qui codent une protéine Gins23 identifiable, à l'exception des génomes des organismes du genre *Pyrobaculum*, le gène *gins23* forme un opéron avec le gène *mcm* ce qui suggère que les produits de ces deux gènes interagissent physiquement, voire fonctionnellement. Les expériences de double-hybride menées chez la crenarchée *Sulfolobus solfataricus* d'une part et l'euryarchée *Pyrococcus*

furiosus d'autre part montrent effectivement que ces deux protéines interagissent physiquement l'une avec l'autre (Marinsek et al., 2006; Yoshimochi et al., 2008). Néanmoins, la nature de cette interaction est différente selon l'organisme considéré : elle est stable chez *S. solfataricus* alors qu'elle semble transitoire chez *P. furiosus* (Marinsek et al., 2006; Yoshimochi et al., 2008).

Interaction fonctionnelle GINS-MCM. L'association entre les gènes *gins23* et *mcm* dans un grand nombre de génomes d'Archaea suggère en outre que les produits d'expression de ces deux gènes interagissent fonctionnellement. Aussi, il est possible que chez ces archées le complexe GINS assiste l'hélicase MCM au niveau de la fourche de réplication, un rôle similaire à celui qui a été proposé pour le complexe GINS chez les eucaryotes (voir, (Aparicio et al., 2006; Labib and Gambus, 2007)). Néanmoins, les données expérimentales obtenues chez les deux archées modèles ne sont pas concordantes quant à l'effet du complexe GINS sur l'activité hélicase de la protéine MCM. En effet, il semble que l'activité hélicase de la protéine MCM soit stimulée en présence du complexe GINS chez *Pyrococcus furiosus*, alors qu'aucune stimulation n'a été observée chez *Sulfolobus solfataricus* (Marinsek et al., 2006; Yoshimochi et al., 2008). En outre, l'activité hélicase de *Pfu*MCM est très faible en absence de *Pfu*GINS alors que *Sso*MCM possède une forte activité hélicase intrinsèque.

D'autre part, contrairement aux résultats obtenus chez *S. solfataricus*, il semble que l'interaction entre le complexe GINS et la protéine MCM chez *P. furiosus* soit transitoire dans la mesure où le complexe GINS et la protéine MCM sont éluées séparément à l'issue d'une chromatographie sur tamis moléculaire (Yoshimochi et al., 2008). Deux modèles ont été proposés pour rendre compte de ces observations (Yoshimochi et al., 2008). Le premier modèle stipule que le complexe GINS aide au recrutement de l'hélicase MCM et induit un changement de conformation du MCM — passage d'une forme inactive à une forme active —

mais ne participe pas à la phase d'élongation. Le second modèle avance que le complexe GINS participe au chargement de l'hélicase MCM au niveau de l'origine de réplication et assiste de manière durable le MCM y compris durant la phase d'élongation pour stimuler son activité hélicase. Pour expliquer l'absence d'interaction directe entre les deux molécules, les auteurs suggèrent l'existence d'un intermédiaire protéique qui stabilise l'interaction, un rôle assuré chez les eucaryotes par la protéine Cdc45, mais ce facteur n'a pas été identifié chez les Archaea.

Nature des partenaires protéiques. Les données obtenues chez *S. solfataricus* indiquent que le complexe *Sso*GINS interagit avec de nombreuses protéines du réplisome. En effet, le complexe *Sso*GINS interagit via sa sous-unité Gins23 d'une part avec l'hélicase MCM, d'autre part avec chacune des deux sous-unités de l'ADN primase ce qui a été interprété comme suggérant que le complexe *Sso*GINS forme un pont moléculaire entre le brin retardé et le brin continu afin de coordonner la synthèse des amorces ARN sur le brin retardé avec l'ouverture progressive de la double-hélice d'ADN (Marinsek et al., 2006). A l'heure actuelle, les données obtenues chez *P. furiosus* ne permettent pas de déterminer si un assemblage macromoléculaire comparable prend place au niveau de la fourche de réplication lors de la duplication du chromosome chez cet organisme (Yoshimochi et al., 2008).

Spectre d'action du complexe GINS. Les données de double-hybride réalisées avec les protéines Gins de *P. furiosus* indiquent que la protéine *Pfu*Gins15 est capable d'interagir avec la protéine initiatrice de la réplication *Pfu*Cdc6/Orc1 et que le complexe *Pfu*GINS peut se fixer au niveau de la région du génome au niveau de laquelle se situe l'origine de la réplication *oriC*. Cette observation suggère que le complexe GINS pourrait intervenir dans le recrutement de l'hélicase MCM lors de la phase d'initiation de la réplication (Yoshimochi et

al., 2008), un rôle tenu par la protéine Cdt1 chez certains eucaryotes (métazoaires et ascomycètes) (Kearsey and Cotterill, 2003). Sachant qu'il a été proposé que la protéine WhiP, présente chez certaines crenarchées (voir ci-après), pourrait être un analogue fonctionnel de la protéine Cdt1 (Robinson and Bell, 2007), il est possible que, chez les Archaea, le spectre d'action du complexe GINS diffère suivant qu'un homologue WhiP est présent ou absent. A ce titre, il serait important de vérifier quels sont les attributs fonctionnels de la protéine Gins15 chez les organismes archéens qui n'encodent pas la protéine Gins23 afin de savoir si un complexe GINS formé par la seule protéine Gins15 intervient uniquement au niveau de la formation du complexe d'initiation ou si ses fonctions s'étendent à la phase d'élongation (stimulation de l'activité hélicase, coordination de la synthèse des deux brins d'ADN). En effet, si le rôle des complexes *Sso*GINS et *Pfu*GINS semblent différents, les expériences de double-hybride ont néanmoins montré que, dans les deux cas, seule la protéine Gins23 est capable d'interagir physiquement avec la protéine MCM (Marinsek et al., 2006; Yoshimochi et al., 2008). Aussi, cela suggère qu'un assemblage protéique formé de la seule protéine Gins15 n'est probablement pas en mesure d'interagir physiquement avec la protéine MCM et donc d'influencer l'activité enzymatique de l'hélicase réplicative. De façon intéressante, alors que les résultats obtenus chez *P. furiosus* suggèrent que l'activité hélicase du complexe MCM est dépendante du complexe GINS (Yoshimochi et al., 2008), les caractérisations des complexes MCM chez les Euryarchaeota *M. thermautotrophicum* (*Mth*), *A. fulgidus* (*Afu*) et *T. acidophilum* (*Tac*) ont montré que *Mth*MCM, *Afu*MCM et *Tac*MCM possèdent une forte activité hélicase en l'absence d'un facteur auxiliaire (Grainge et al., 2003; Haugland et al., 2006; Kelman et al., 1999a). Sachant que les génomes de ces trois organismes n'arborent pas le gène codant la protéine Gins23 **(Figure IV-3)**, il est possible que la fonction du complexe GINS se restreigne à la phase d'initiation de la réplication chez ces organismes et, par extrapolation, chez l'ensemble des Euryarchaeota qui ne codent que la protéine Gins15.

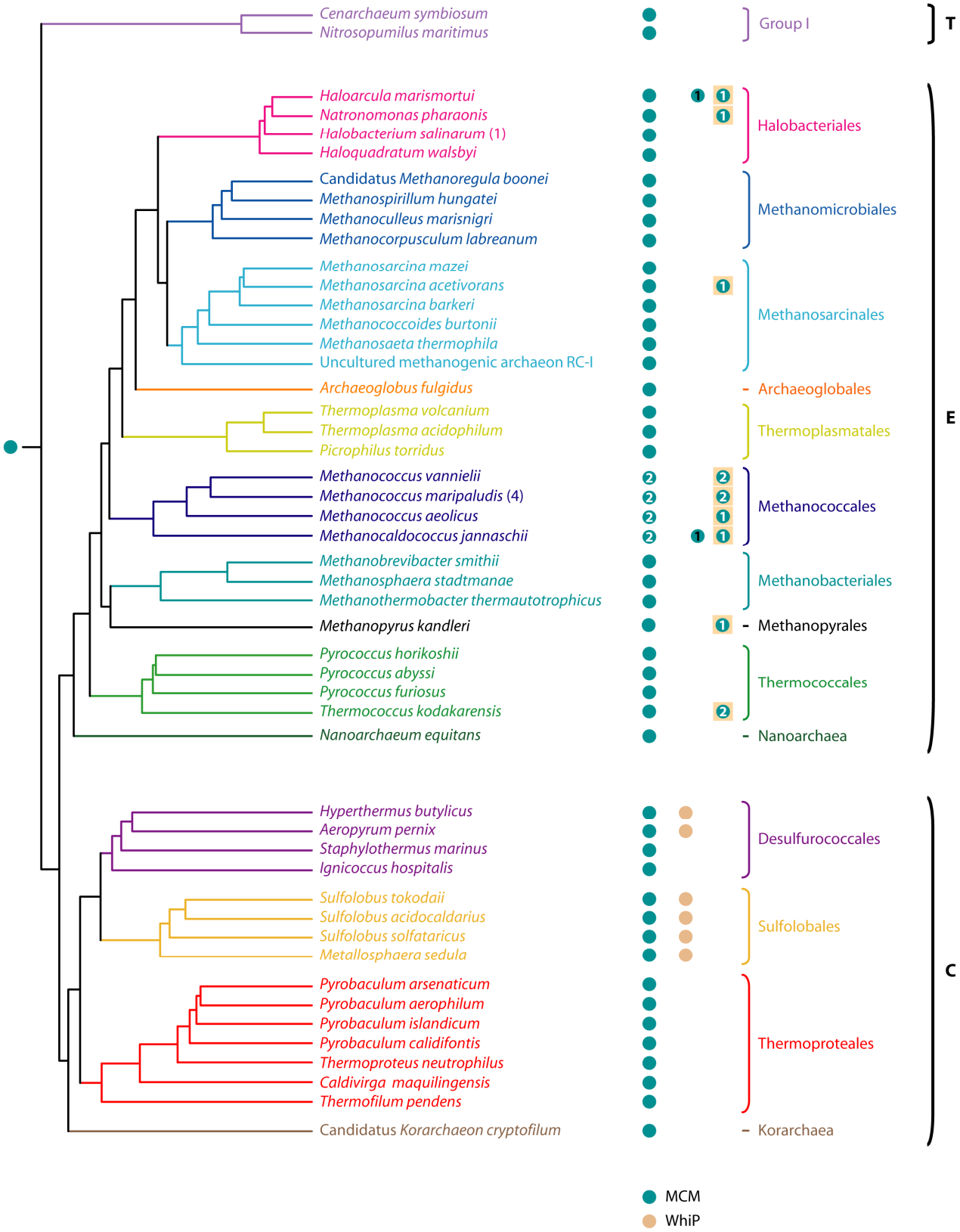

Figure IV-4 : Profil phylétique des gènes codant le MCM et la protéine WhiP chez les Archaea. La présence d'un homologue MCM est signalée par un disque couleur vert. La présence d'un homologue WhiP est indiquée par un disque couleur bisque. Au-delà de un, le nombre de gènes *mcm* chromosomique est indiqué par un chiffre de couleur blanche. Les gènes plasmidiques sont indiqués par un chiffre de couleur noire. Les gènes qui pourraient avoir une origine extra-chromosomique se distinguent par le carré au fond orange clair situé en arrière plan ; leur nombre est indiqué en blanc.

Pour conclure, les données biochimiques concernant le rôle du complexe GINS chez les Archaea sont discordantes — de manière analogue aux résultats des études sur le complexe GINS chez les organismes eucaryotes modèles (voir, (Aparicio et al., 2006; Labib and Gambus, 2007)). De façon remarquable, les gènes *gins15* et *gins23*, qui étaient vraisemblablement présents chez le dernier ancêtre commun des Archaea, ont suivi une trajectoire évolutive différente. Or, il est possible que cette histoire évolutive explique les divergences fonctionnelles observées chez les organismes actuels.

MCM

Dans la majorité des génomes d'Archaea, on dénombre un seul gène *mcm*. Cependant, dans le génome de certains organismes plusieurs paralogues sont présents **(Figure IV-4)**. Par exemple, on trouve trois gènes *mcm* dans le génome de *T. kodakaraensis* dont deux sont probablement d'origine virale (Fukui et al., 2005). On trouve également trois gènes *mcm* dans le génome de l'halophile *H. marismortui* dont un se trouve sur un plasmide. La présence de plusieurs gènes *mcm* est aussi une caractéristique des génomes des organismes appartenant à l'ordre des Methanococcales dans lesquels on peut détecter au moins trois gènes *mcm*. Ces copies supplémentaires ont probablement été introduites par des transferts horizontaux comme le suggèrent d'une part l'analyse du contexte génomique (présence d'un gène codant une transposase à proximité), d'autre part une analyse informatique destinée à déceler les gènes étrangers d'un génome, c'est-à-dire compris dans des éléments intégrés (Diego Cortez, communication personnelle).

Le génome de *M. jannaschii* présente quatre gènes *mcm* ; l'un d'entre eux se trouve sur un plasmide. Parmi les trois gènes chromosomiques, un, voire deux d'entre eux pourraient avoir une origine virale. Le génome de *M. aeolicus* arbore trois gènes *mcm* ; l'un d'entre eux semble avoir une origine virale. Dans les génomes de *M. maripaludis* (souches C5, C7 et S2)

et de *M. vannielii*, deux des quatre gènes *mcm* présents sur le chromosome sont probablement d'origine extra-chromosomique. Enfin, le génome de *M. maripaludis* C6 arbore huit gènes *mcm* dont au moins quatre ont une origine virale probable. Le rôle joué dans la cellule par ces nombreux homologues MCM n'est pas connu mais il a été proposé que certaines des sous-unités MCM surnuméraires sont des versions inactivées et que ces formes mutantes pourraient prendre part à la formation du complexe MCM et en moduler l'activité hélicase (McGeoch and Bell, 2008).

WhiP

Le gène codant la protéine WhiP a récemment été identifié chez les Archaea en comparant l'organisation des gènes situés au voisinage des origines de réplication des Crenarchaeota *Aeropyrum pernix* et *Sulfolobus solfataricus* (Robinson and Bell, 2007). Ce gène code une protéine présentant successivement un domaine wHTH, un domaine central sans motif reconnaissable et un nouveau domaine wHTH. Cet arrangement rappelle l'architecture de la protéine RepA, une protéine qui initie la réplication de certains plasmides bactériens. En outre, un gène codant une protéine présentant de fortes similarités avec la protéine CopG — une protéine impliquée dans le contrôle du nombre de copies plasmidiques — se trouve dans cette même région du génome (Robinson and Bell, 2007). Aussi, il a été suggéré que ce gène, l'origine de réplication qu'il côtoie et les gènes environnants, dont le gène codant la protéine WhiP, soient d'origine extra-chromosomique (Robinson and Bell, 2007). Sachant que la protéine eucaryote Cdt1 pourrait dériver d'une protéine archéenne possédant un motif wHTH (Iyer and Aravind, 2006), il a été proposée que la protéine WhiP est un analogue fonctionnel de la protéine eucaryote Cdt1 chez les Archaea. Par conséquent, il a été avancé que la protéine WhiP assiste le chargement de l'hélicase MCM chez les Archaea (McGeoch and Bell, 2008; Robinson and Bell, 2007), tout comme Cdt1 assiste le chargement du MCM chez les

métazoaires et les ascomycètes (pour une revue, voir (Kearsey and Cotterill, 2003)). Néanmoins, la distribution du gène codant la protéine WhiP dans les génomes d'Archaea semble se restreindre aux Sulfolobales et à deux organismes de l'ordre des Desulforococcales (*A. pernix* et *H. butylicus*) **(Figure IV-4)**, ce qui suggère que cette protéine n'est pas conservée chez les Archaea ou a été perdue à plusieurs reprises au cours de l'évolution cellulaire comme Cdt1 (Iyer and Aravind, 2006). De fait, cette protéine n'est probablement pas un composant essentiel du réplisome archéen et le rôle joué par la protéine WhiP dans les organismes concernés est sans doute assuré par une autre protéine dans les autres organismes. En réalité, la protéine WhiP pourrait être une acquisition récente des Sulfolobales. Par exemple, un élément extra-chromosomique a pu être intégré dans le génome de l'ancêtre commun des Sulfolobales avant d'être transféré vers le génome de l'ancêtre commun des organismes *A. pernix* et *H. butylicus*. Une autre hypothèse envisageable est que ce transfert ait eu lieu chez l'ancêtre commun des Sulfolobales et des Desulfurococcales et que le gène codant la protéine WhiP ait été perdu indépendamment chez *Staphylothermus marinus* et *Ignicoccus hospitalis*.

SSB/RPA

L'analyse de la distribution des gènes codant la protéine de fixation de l'ADN simple brin confirme que la nature de cette protéine est différente entre les deux phylums principaux **(Figure IV-5)**. Dans les deux cas, le motif de fixation à l'ADN correspond à un repliement OB (oligosaccharide/oligonucleotide/oligopeptide binding fold) dont la structure ressemble à celle de l'homologue eucaryote mais dont l'architecture globale varie d'un phylum à l'autre (White, 2003). En effet, chez les Euryarchaeota la protéine fixant le simple brin a une architecture qui ressemble à la protéine RPA eucaryote (plusieurs OB-fold, présence fréquente d'un motif à doigt à zinc) tandis que chez les Crenarchaeota la protéine fixant le

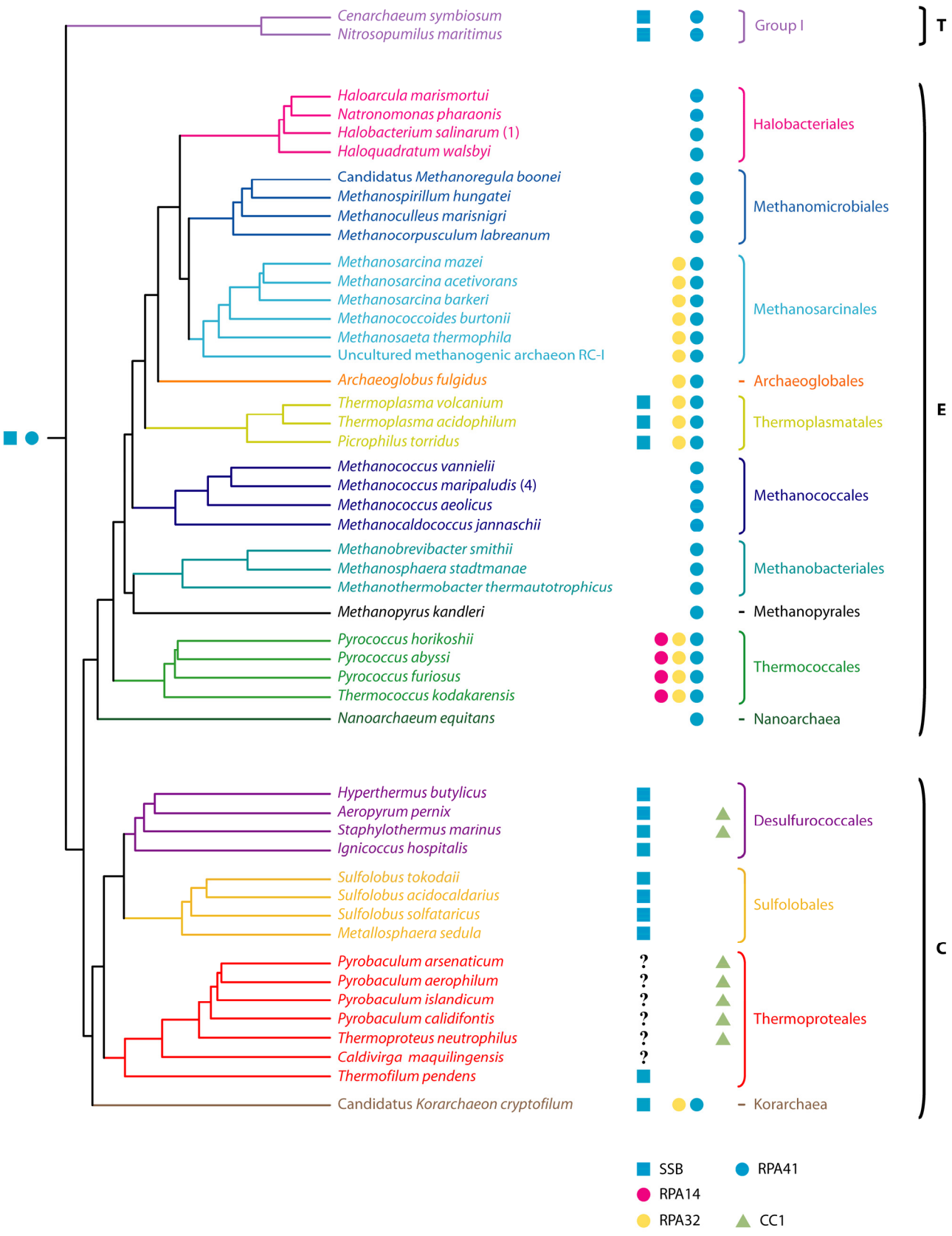

Figure IV-5 : Profil phylétique des gènes codant les protéines de type RPA ou SSB chez les Archaea. La présence d'un homologue SSB est signalée par un carré cyan. Les homologues de type RPA14, RPA32 ou RPA41 sont indiqués par des disques couleur magenta, jaune et cyan, respectivement. Les points d'interrogation signifient que la nature de la protéine fixant le simple brin lors de la réplication demeure inconnue. La distribution de la protéine CC1 est illustrée par le triangle couleur vert. Le nombre d'homologues RPA41 et RPA32 n'est pas précisé car l'identification des protéines est particulièrement difficile.

simple brin a une architecture qui rappelle la protéine SSB bactérienne (1 seul OB-fold, pas de motif à doigt à zinc) (pour une revue, voir (White, 2003)). De manière intéressante, le génome de *K. cryptofilum* arbore un gène codant une protéine de type SSB et deux gènes codant des protéines de type RPA dont une présente un motif à doigt à zinc. Chez les Thaumarchaeota, on observe aussi un gène codant une protéine de type SSB et un gène codant une protéine de type RPA mais cette dernière ne présente pas de motif doigt à zinc.

Une analyse plus détaillée du profil phylétique des gènes *ssb/rpa* révèle certaines particularités au sein de chacun des deux phylums principaux. Chez les Euryarchaeota, on observe d'une part que le nombre de protéines de type RPA (RPA41, RPA32, RPA14) est différent d'un ordre à un autre mais également que l'architecture globale (nombre de motifs 'OB-fold') varie entre les différents groupes taxonomiques (Robbins et al., 2005). Il est donc probable que la structure quaternaire de ces protéines et le mode de fixation de l'ADN simple brin varie d'un groupe d'organismes à un autre (Robbins et al., 2005). Par ailleurs, on peut détecter la présence d'un gène codant une protéine de type SSB chez tous les Thermoplasmatales (White, 2003). Ce gène provient vraisemblablement d'un transfert de gène en provenance d'un organisme de l'ordre des Sulfolobales car ces organismes partagent les mêmes niches écologiques (Ruepp et al., 2000). Sachant que le gène se trouve dans le génome de tous les organismes appartenant à l'ordre des Thermoplasmatales, ce transfert a du avoir lieu chez l'ancêtre commun de ce groupe.

En ce qui concerne les Crenarchaeota (*K. cryptofilum* excepté), il est intéressant de remarquer que le gène codant la protéine SSB n'est pas identifiable dans le génome des organismes appartenant aux genres *Caldivirga*, *Thermoproteus* et *Pyrobaculum* (famille des Thermoproteaceae). En revanche, on peut identifier une copie du gène *ssb* dans le génome de *T. pendens* qui appartient à l'ordre des Thermoproteales, au même titre que les organismes de la famille des Thermoproteaceae **(Figure IV-5)**. Cela suggère que le gène codant la protéine

de type SSB était probablement présent chez l'ancêtre commun des Crenarchaeota. Deux hypothèses peuvent être proposées pour expliquer le fait qu'aucun gène *ssb* n'est identifiable chez les espèces de la famille des Thermoproteaceae : i) le gène est bien présent mais la séquence a fortement divergée et la recherche de similarité est infructueuse ; ii) le gène *ssb* a été remplacé par un gène codant une protéine non homologue assurant des fonctions semblables à celle de la protéine de type SSB, c'est-à-dire qu'il y a eu un déplacement du gène originel par un gène non orthologue (Koonin et al., 1996). Dans ce cas, le déplacement non orthologue a probablement eu lieu chez l'ancêtre commun des Thermoproteaceae étant donné que le gène *ssb* est présent chez *T. pendens*.

Une approche biochimique visant à purifier et identifier une protéine fixant l'ADN simple brin chez *T. tenax* (famille des Thermoproteaceae) a mené à la caractérisation d'une protéine (CC1), mais celle-ci est probablement impliquée dans l'organisation de la chromatine (Luo et al., 2007) et non dans la fixation des intermédiaires simple-brin engendrés au fur et à mesure de la progression de l'hélicase réplicative. En outre, le gène codant la protéine CC1 n'est pas présent dans le génome de *C. maquilingensis* **(Figure IV-5)**. Par conséquent, il est probable que la protéine de fixation du simple brin intervenant dans le cadre de la réplication demeure inconnue chez l'ensemble des Thermoproteaceae.

Dans son ensemble, le profil phylétique des gènes codant les protéines SSB et RPA est plutôt confus et difficile à interpréter. La variabilité observée en ce qui concerne la protéine de fixation du simple-brin indique que l'architecture de cette protéine est extrêmement plastique. A ce titre, il a été proposé que l'architecture de type RPA a évolué à partir d'une protéine de type SSB par duplication du domaine 'OB-fold' et que la diversité structurale des protéines RPA des Euryarchaeota serait la conséquence d'évènements de recombinaisons et de fusion (Chedin et al., 1998; Lin et al., 2008). De manière intéressante, deux protéines de type SSB ont récemment été identifiées chez l'homme et l'une d'entre elles (hSSB1) joue un

rôle crucial dans l'intégrité du génome (Richard et al., 2008). En outre, le phénotype des cellules après inactivation de l'expression de hSSB1 (hypersensitivité aux radiations, défaut d'activation des points de contrôle, instabilité génomique accrue) suggère que la protéine hSSB1 est aussi importante que la protéine RPA pour le métabolisme cellulaire (Richard et al., 2008). Ces protéines possèdent des homologues dans plusieurs royaumes eucaryotes et sont apparentées aux protéines SSB et RPA des Archaea. Par conséquent, les gènes codant les protéines SSB et RPA pourraient avoir été hérités du dernier ancêtre commun des Eucarya et des Archaea avant que l'un, voire les deux gènes ne soient perdus selon les lignées **(Figure IV-5)**.

ADN primase

Le profil phylétique des gènes codant la petite et la grande sous-unité de l'ADN primase est sans ambiguïté (non montré). Tous les génomes d'Archaea possèdent les gènes *priA* et *priB* codant respectivement la petite et la grande sous-unité de l'ADN primase ; seul le génome de *N. equitans* arbore un gène consistant en une fusion entre les deux gènes. Aussi, ce profil indique que le dernier ancêtre commun des Archaea possédait une ADN primase hétérodimérique semblable à celle que l'on trouve chez l'ensemble des Archaea, si l'on excepte la forme monomérique présente chez *N. equitans* qui représente probablement une forme dérivée.

PCNA

Tous les génomes d'Archaea possèdent au moins un gène codant le PCNA. Le génome de *T. kodakaraensis* mis à part, tous les génomes d'Euryarchaeota arborent un seul gène codant le PCNA **(Figure IV-6)**. La copie supplémentaire présente dans le génome de *T. kodakaraensis* est située dans une région qui semble avoir une origine extra-chromosomique, probablement

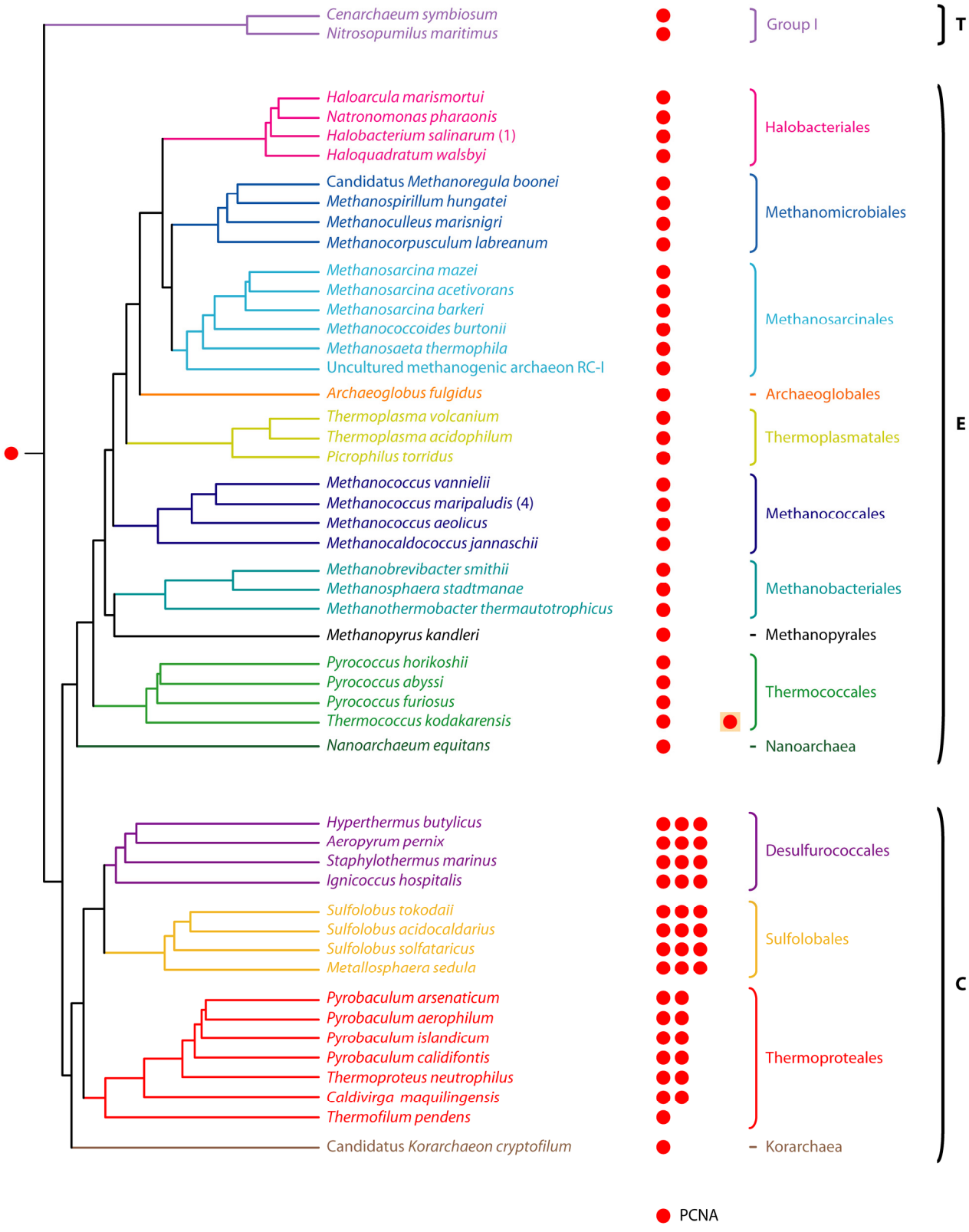

Figure IV-6 : Profil phylétique des gènes codant le PCNA chez les Archaea. La présence d'un homologue PCNA est signalée par un disque rouge. Le gène de *T. kodakaraensis* qui pourrait avoir une origine virale est indiqué par le carré au fond orange clair situé en arrière plan.

virale (Fukui et al., 2005). La présence de plusieurs paralogues du gène codant le PCNA est une caractéristique de la quasi-totalité des génomes des Crenarchaeota car seuls les génomes de *K. cryptofilum* et *T. pendens* ne possèdent qu'une copie de ce gène **(Figure IV-6)**. En effet, tous les autres organismes appartenant à l'ordre des Thermoproteales comptent deux copies du gène codant le PCNA alors que les organismes de l'ordre des Sulfolobales et des Desulfurococcales arborent trois copies de ce gène. En l'absence d'arbres phylogénétiques, il n'est pas possible de déterminer à quel niveau ont eu lieu les duplications de gènes. Néanmoins, sachant que les génomes des Thaumarchaeota et *K. cryptofilum* ne présentent qu'un seul gène pour le PCNA — comme chez les Euryarchaeota —, il est vraisemblable que le dernier ancêtre commun des Archaea ne possédait qu'un seul gène pour le PCNA et que les duplications de ce gène ont eu lieu chez les Crenarchaeota après l'émergence du groupe représenté par *K. cryptofilum*.

RFC

Dans la majorité des génomes d'Archaea, on trouve un gène codant la petite-sous unité du RFC (*rfcS*) et un gène codant la grande sous-unité du RFC (*rfcL*). On dénote cependant deux gènes *rfcS* dans les génomes de la majorité des organismes appartenant aux ordres des Methanosarcinales, des Methanomicrobiales et des Halobacteriales **(Figure IV-7)**. Ces trois ordres étant affiliés, il est probable que la duplication du gène ait eu lieu chez l'ancêtre commun de ces trois ordres. Cette hypothèse est renforcée par le fait que la sous-unité RFC-s2 codée chez ces organismes possède systématiquement une insertion d'au moins vingt acides aminés entre les motifs RFC IV et V (Cann et al., 2001; Chen et al., 2005b; Cullmann et al., 1995). Contrairement aux autres Methanosarcinales, *M. thermophila* ne présente qu'un seul gène *rfcS* ce qui suggère que la seconde copie héritée de l'ancêtre commun des Methanosarcinales a été perdue chez cet organisme. De manière intéressante, le génome de

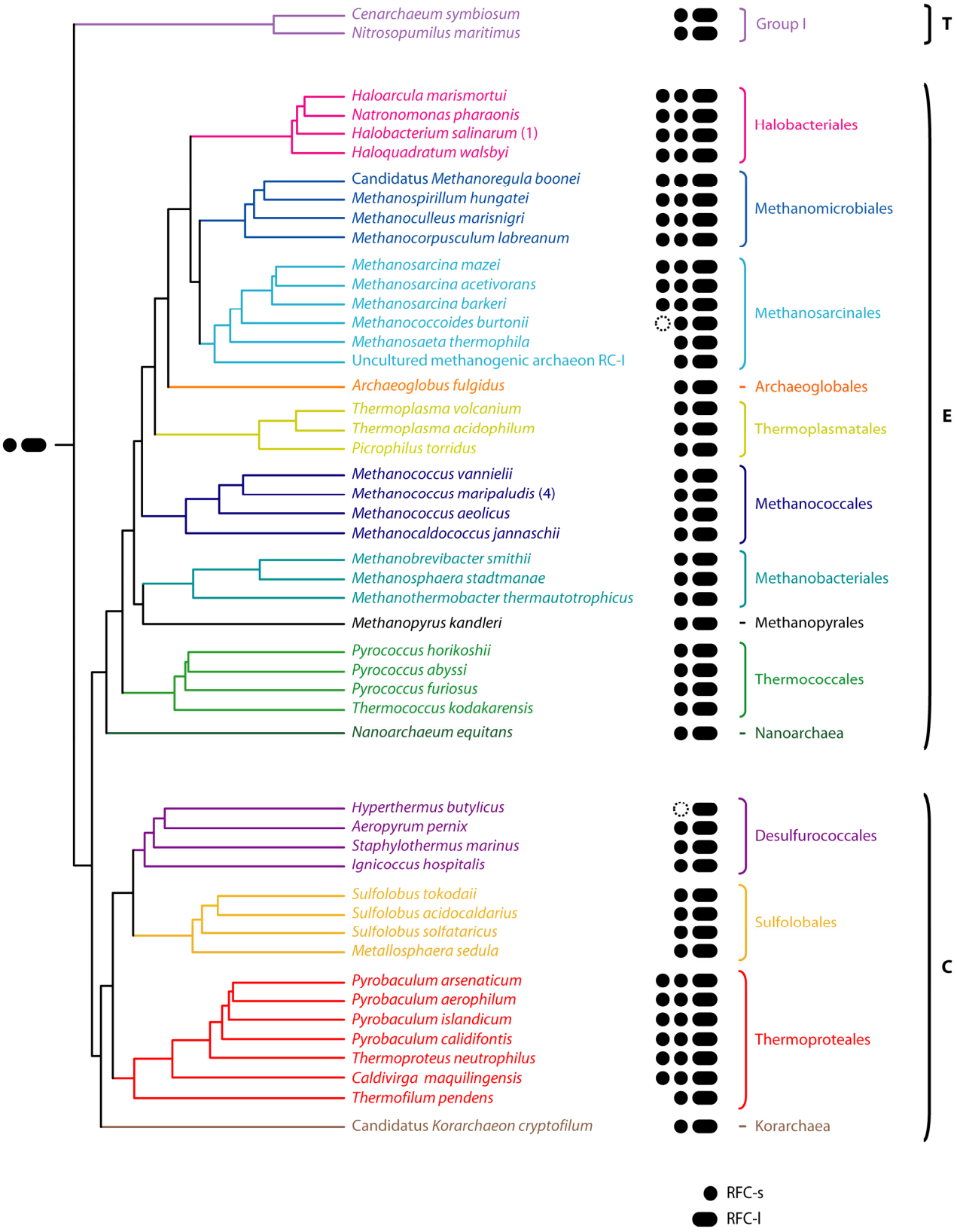

Figure IV-7 : Profil phylétique des gènes codant la petite et la grande sous-unité du RFC chez les Archaea. La présence d'un homologue RFC-s est signalée par un disque noir, celle d'un homologue RFC-l par un rectangle arrondi noir. Les pseudogènes *rfcS* des génomes de *Methanococcoides burtonii* et *Hyperthermus butylicus* sont mentionnés par des cercles noirs pointillés.

M. burtonii possède encore la trace de la présence du deuxième gène *rfcS* car un fragment d'ADN annoté comme pseudogène présente des similarités évidentes avec le gène *rfcS* (non montré). En outre, ce pseudogène se trouve dans un contexte génomique qui ressemble légèrement (présence d'un gène codant une histone) à celui qu'occupe le gène *rfcS1* dans les génomes des organismes du genre *Methanosarcina*.

Au sein du phylum des Crenarchaeota, les génomes des organismes de l'ordre des Thermoproteales portent tous deux copies du gène *rfcS* (*rfcS1* et *rfcS2*) à l'exception de *T. pendens* qui n'en possède qu'un **(Figure IV-7)**. L'évènement de duplication du gène *rfcS* chez ces organismes est vraisemblablement récent (Fitz-Gibbon et al., 2002) ; la similarité au niveau protéique entre les deux RFC-s est en effet très élevée (celle-ci est comprise entre 64% et 74%). Le fait que le génome de *T. pendens* ne comporte qu'une seule copie du gène indique que la duplication a probablement eu lieu après la séparation entre la branche conduisant à *T. pendens* et la branche conduisant à la famille des Thermoproteaceae, qui regroupe les espèces *Pyrobaculum*, *Thermoproteus* et *Caldivirga*. De manière intéressante, le gène *rfcS1* est systématiquement associé au gène *rfcL* dans les organismes concernés alors que le gène *rfcS2* se trouve ailleurs dans le génome. L'association entre les gènes *rfcS1* et *rfcL* suggère l'existence d'un lien privilégié entre les sous-unités RFC-s1 et RFC-l, lesquelles pourraient former la structure principale du complexe RFC. Le rôle exact joué par chacune des deux petites sous-unités au sein du RFC n'est pas connu, mais il a été suggéré que la sous-unité RFC-s2 pourrait intervenir en tant qu'unité régulatrice du complexe RFC chez *Methanosarcina acetivorans* (Chen et al., 2005b).

Aussi bien chez les Euryarchaeota que chez les Crenarchaeota cette copie supplémentaire du gène *rfcS* code une protéine qui semble fonctionnelle. Dans les deux cas, la lysine (K) du motif Walker A (fixation de l'ATP) et l'arginine R du motif SRC, les deux résidus critiques pour l'hydrolyse de l'ATP, sont présents. En revanche, les protéines RFC-s2

codées dans les génomes d'Euryarchaeota présentent un motif RFC V qui diffère légèrement de la séquence canonique. Néanmoins, les résultats d'une étude biochimique suggèrent que la sous-unité RFC-s2 participe à la formation du complexe RFC chez *M. acetivorans* (Chen et al., 2005b). Il est fort probable qu'il en soit de même avec les protéines RFC-s2 homologues présentes chez les Methanosarcinales, les Methanomicrobiales et les Halobacteriales.

Enfin, et de manière surprenante, aucun gène *rfcS* n'est identifiable dans le génome de l'hyperthermophile *H. butylicus*. En fait, le gène *rfcS* est annoté comme pseudogène (Hbut_0905) dans la base de données du NCBI, la phase ouverte de lecture étant interrompue par un déplacement du cadre de lecture. Ce pseudogène *rfcS* est adjacent au gène *rfcL* comme dans l'ensemble des autres génomes de Crenarchaeota. Etant donné le rôle crucial que joue le RFC-s dans l'hydrolyse de l'ATP et donc dans le chargement du PCNA, il est difficile d'imaginer comment la grande sous-unité du RFC pourrait accomplir seule les fonctions de chargement du PCNA chez cet organisme. Aussi, il est possible qu'une erreur de séquençage soit à l'origine de cette modification du cadre de lecture. Cette incertitude pourrait être levée en séquençant à nouveau cette région du génome, en vérifiant si ce pseudogène est transcrit ou en analysant le protéome de cet organisme.

ADN polymérases

ADN polymérases de type B. Tous les génomes d'Archaea séquencés arborent au moins un gène codant une ADN polymérase de la famille B. On distingue classiquement trois sous-familles dénommées PolB1, PolB2, PolB3 (Edgell et al., 1997). Les ADN polymérases B1 et B3 sont sans doute impliquées dans la réplication de l'ADN car elles possèdent toutes les signatures caractéristiques des ADN polymérases réplicatives ; en revanche, certains motifs conservés sont absents ou mutés dans les séquences des PolB2 (Bohlke et al., 2002). En particulier, deux résidus critiques pour l'activité polymérase situés au sein du motif qui est

impliqué dans la coordination de l'ion magnésium sont mutés chez les PolB2 (Bohlke et al., 2002; Rogozin et al., 2008). Il est donc peu probable que les PolB2 aient un rôle catalytique dans la duplication du matériel génétique (Bohlke et al., 2002; Rogozin et al., 2008). Récemment, il a été avancé que les PolB2 pourraient être impliquées dans la réplication en tant qu'unité structurale au sein de la forme holenzyme d'une ADN polymérase (Rogozin et al., 2008). De manière remarquable, les génomes de *P. aerophilum*, *P. arsenaticum*, *P. calidifontis* et *T. neutrophilus* codent une ADN polymérase présentant une architecture comparable à celle des PolB2, notamment l'absence de deux domaines exonucléase et d'une séquence jouant un rôle dans la coordination entre les activités de synthèse et de dégradation des PolBs (Sartori and Jiricny, 2003). Chez *P. aerophilum* cette ADN polymérase (PAE1113) a également été dénommé PolB2 alors que, contrairement aux PolB2, cette protéine possède les résidus impliqués dans l'activité catalytique (Sartori and Jiricny, 2003). En fait, la protéine PAE1113 présente une activité polymérase de déplacement de brin et il a été proposé qu'elle soit un homologue fonctionnel de l'ADN polymérase β (Sartori and Jiricny, 2003). Nous proposons de baptiser ces ADN polymérases B PolB4 afin de les distinguer des PolB2.

La PolB2 est distribuée de manière non uniforme dans un certain nombre de génomes de Crenarchaeota et d'Euryarchaeota représentant différentes lignées **(Figure IV-8)** et son origine ne peut être déterminée avec certitude (Rogozin et al., 2008). La PolB4 est uniquement présente chez *P. aerophilum*, *P. arsenaticum*, *P. calidifontis* et *T. neutrophilus*, quatre organismes apparentés appartenant à l'ordre des Thermoproteaceae mais le gène correspondant est absent du génome de *P. islandicum* **(Figure IV-8)**. L'origine la plus probable de cette ADN polymérase est une origine extra-chromosomique (transfert horizontal) car les gènes correspondants se trouvent au sein d'éléments intégrés dans ces génomes (Diego Cortez, communication personnelle).

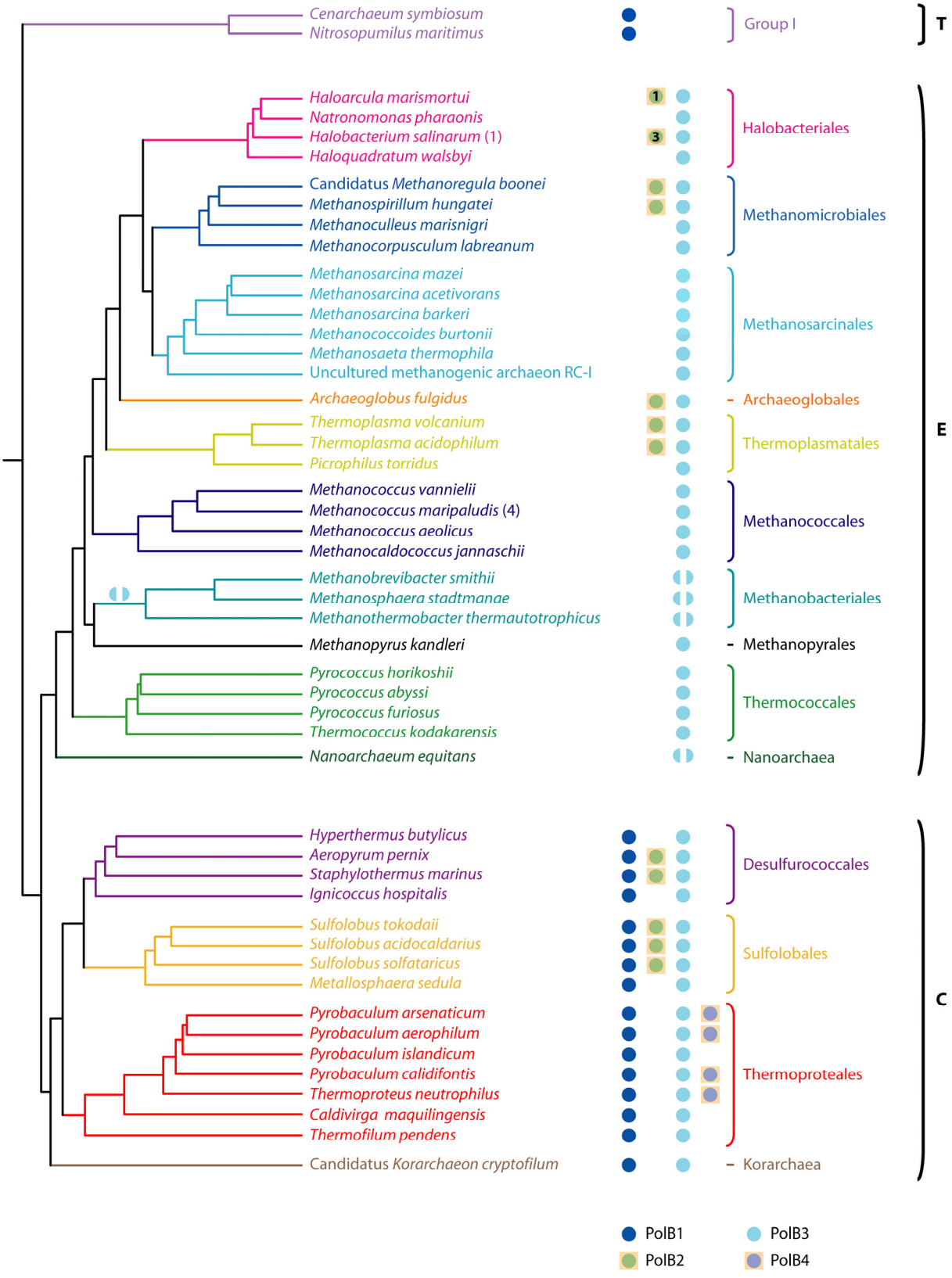

Figure IV-8 : Profil phylétique des gènes codant les ADN polymérases de la famille B chez les Archaea. La présence des ADN polymérases de type B1, B2, B3 ou B4 sont signalées par un disque couleur bleu foncé, vert, bleu clair et mauve, respectivement. Les disques scindés en deux (PolB3) indiquent que la polymérase est codée par deux gènes apparus suite à un évènement de scission du gène originel. Les PolB2 et PolB4 pourraient avoir une origine extrachromosomique (carré orange en arrière plan).

La PolB1 est présente dans tous les génomes de Crenarchaeota et dans le génome des Thaumarchaeota **(Figure IV-8)** ; autrement dit, la PolB1 est absente de tous les génomes d'Euryarcheaea. La PolB3 est quant à elle représentée dans tous les génomes d'Archaea à l'exception des Thaumarchaeota **(Figure IV-8)**. Trois hypothèses également parcimonieuses peuvent être avancées concernant la présence de la PolB1 et de la PolB3 chez le dernier ancêtre commun : i) la PolB1 et la PolB3 étaient déjà présentes, la PolB3 a été perdue chez les Thaumarchaeota et la PolB1 a été perdue chez l'ancêtre commun des Euryarchaeota ; ii) seule la PolB1 était présente, ce qui implique que la PolB1 a été perdue chez l'ancêtre commun des Euryarchaeota et que la PolB3 est apparue chez l'ancêtre commun du groupe comprenant les Euryarchaeota et les Crenarchaeota ; iii) seule la PolB3 était présente, ce qui suppose que la PolB1 a été gagnée indépendamment chez les Thaumarchaeota et chez l'ancêtre commun des Crenarchaeota. Par conséquent, la distribution phylogénétique des gènes codant des ADN polymérases de la famille B dans le domaine Archaea suggère qu'au moins une ADN polymérase de type B se trouvait chez le dernier ancêtre commun des archées.

Chez les Methanobacteriales, la PolB3 est codée par deux gènes et la polymérase n'est fonctionnelle que sous sa forme dimérique (Kelman et al., 1999b). Ces deux gènes sont très certainement issus d'un évènement de scission du gène entier tel que l'on observe dans les autres génomes d'Archaea. Une alternative moins parcimonieuse consisterait à supposer que cette forme scindée est ancestrale et que plusieurs évènements de fusion indépendants sont à l'origine de la répartition actuelle entre forme scindée et forme entière. En fait, sachant que la forme scindée est présente chez l'ensemble des Methanobacteriales l'évènement de scission du gène entier a probablement eu lieu chez l'ancêtre de ce groupe. Par ailleurs, la PolB3 de *N. equitans* est également codée par deux gènes (Waters et al., 2003), mais la forme mature de la protéine est monomérique car les deux segments polypeptidiques sont réunis par un épissage protéique en trans impliquant les deux portions d'une mini-intéine (Choi et al., 2006).

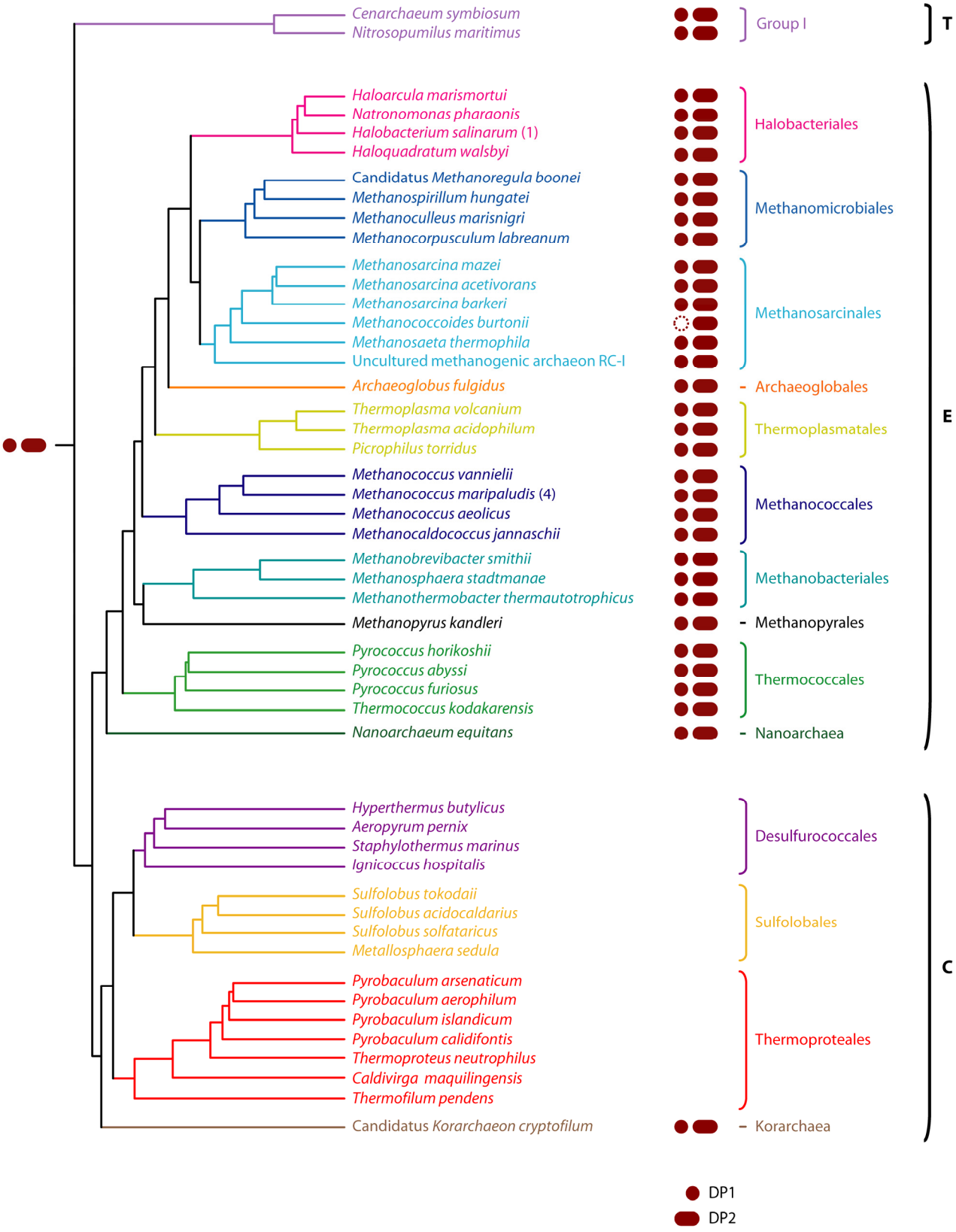

Figure IV-9 : Profil phylétique des gènes codant la petite et la grande sous-unité de l'ADN polymérase de la famille D chez les Archaea. La présence de la petite sous-unité (DP1) et de la grande sous-unité (DP2) de l'ADN polymérase D sont signalées respectivement par un disque et un rectangle arrondi couleur cramoisi. Le pseudogène *dp1* de *Methanococcoides burtonii* est indiqué par un cercle en pointillé.

ADN polymérase de type D. A l'exception du génome de l'archée psychrophile *Methanococcoides burtonii*, tous les génomes d'Euryarchaeota présentent les gènes *dp1* et *dp2* codant respectivement la petite sous-unité (DP1) et la grande sous-unité (DP2) de l'ADN polymérase D. Les deux gènes sont également présents chez les Thaumarchaeota et chez *K. cryptofilum.* Ils sont en revanche systématiquement absents de tous les autres génomes des Crenarchaeota **(Figure IV-9)**. De façon remarquable, seul le gène codant la grande sous-unité de l'ADN polymérase D (DP2) est annoté dans le génome de *M. burtonii*, le gène codant la petite sous-unité (DP1) étant mentionné comme pseudogène. L'activité catalytique de l'ADN polymérase est portée par la grande sous-unité donc il est *a priori* concevable que la protéine DP2 soit suffisante pour que l'ADN polymérase D de *M. burtonii* soit fonctionnelle (Cann et al., 1998; Shen et al., 2001). Néanmoins, il conviendrait de vérifier qu'aucun ARN messager n'est effectivement transcrit à partir de la région du génome de *M. burtonii* dans laquelle se trouve le pseudogène *dp1*. Dans l'affirmative, il serait intéresser de vérifier si une ADN polymérase D formée de la seule grande sous-unité présente bien les propriétés attendues, sachant qu'il a été montré chez que l'association entre les deux sous-unités est cruciale pour que l'activité de polymérisation soit élevée (Ishino et al., 1998; Uemori et al., 1997).

Cette exception mis à part, le profil phylétique des gènes *dp1* et *dp2* dans les génomes d'Archaea permet de déduire, indépendamment de la position phylogénétique de *C. symbiosum*, *N. maritimus* et *K. cryptofilum* vis-à-vis des Crenarchaeota, que les deux sous-unités de l'ADN polymérase D étaient présentes chez le dernier ancêtre commun des archées et que les gènes *dp1* et *dp2* ont très vraisemblablement été perdus dans la branche conduisant à l'ancêtre du groupe Desulfurococcales-Sulfolobales-Thermoproteales **(Figure IV-9)**.

Par conséquent, au vue de la distribution du gène codant l'ADN polymérase B et de la distribution des gènes codant l'ADN polymérase D **(Figures IV-8 et IV-9)**, le dernier ancêtre

commun des archées possédait vraisemblablement à la fois une ADN polymérase de type D et au moins une ADN polymérase de type B. La connaissance du rôle respectif de chacune des ADN polymérases dans la réplication du génome de l'ancêtre est malheureusement inaccessible. Néanmoins, la répartition de la PoD et des différentes sous-familles PolB chez les organismes actuels permettraient d'inférer leur rôle chez l'ancêtre. En effet, la PolB1 et la PolD sont conjointement présentes chez les Thaumarchaeota tandis que la PolB3 et la PolD sont présentes chez l'ensemble des Euryarchaeota. Il est donc envisageable que les expériences menées chez un modèle Thaumarchaeota d'une part et un modèle Euryarchaeota d'autre part permettent à l'avenir de déterminer la fonction respective de ces ADN polymérases dans la réplication du génome des organismes actuels et par extrapolation de supposer quel était leur rôle respectif chez l'ancêtre commun des Archaea.

A l'heure actuelle, les informations concernant le rôle respectif des différentes ADN polymérases lors de la réplication de l'ADN chez les Archaea sont limitées. Il a été montré que l'aphidicoline, un inhibiteur spécifique des ADN polymérases de type α, inhibe la synthèse d'ADN chez *Halobacterium*, ce qui indique qu'une ADN polymérase de la famille B participe à la réplication de l'ADN chez cet organisme (Forterre et al., 1984). Récemment, une étude génétique a montré que la PolB3 et la PolD, mais pas la PolB2, sont essentielles chez l'euryarchée *Halobacterium* sp. NRC1, ce qui suggère que la PolB3 et la PolD participent conjointement à la réplication du génome (Berquist et al., 2007). En fait, le problème du rôle respectif des ADN polymérases dans la réplication n'a vraiment été abordé que chez l'euryarchée modèle *P. abyssi*. L'étude des propriétés des ADN polymérases PolB3 et PolD, en particulier la capacité à tolérer une amorce ARN, a conduit à la proposition d'un modèle quant au rôle respectif de chacune des deux ADN polymérases. En effet, la PolD s'accommode d'une amorce ARN tandis que la PolB3 ne tolère que des amorces ADN (Henneke et al., 2005). Aussi, il a été proposé que la PolD serait impliquée dans l'extension

des amorces ARN produites par l'ADN primase, dans la maturation des fragments d'Okazaki, voire dans la synthèse du brin retardé (Henneke et al., 2005). La PolB3 étant capable de déplacer la PolD du PCNA, elle pourrait remplacer la PolD une fois les amorces ARN rendues compatibles (Rouillon et al., 2007). La PolB3 étant particulièrement processive en présence du PCNA, il est probable qu'elle assure la synthèse du brin direct, voire celle du brin retardé si elle est déplace la PolD sur les deux brins (Henneke et al., 2005; Rouillon et al., 2007). Il serait intéressant de savoir si les propriétés respectives de la PolD et de la PolB1 isolées à partir d'une thaumarchée sont comparables à celles des PolD et PolB3 de l'euryarchée *P. abyssi*. Le cas échéant, la distribution des rôles entre les deux familles d'ADN polymérases durant la réplication de l'ADN pourrait avoir été conservée depuis l'époque du dernier ancêtre commun des Archaea. En outre, il serait intéressant d'étudier la tolérance respective des PolB1 et PolB3 d'un organisme Crenarchaeota vis-à-vis d'une amorce ARN pour savoir si cette incapacité à réaliser l'extension d'une amorce ARN est caractéristique des ADN polymérases de la famille B. Si tel est le cas, alors se pose la question de savoir comment l'amorce ARN synthétisée par l'ADN primase est transformée en un substrat compatible chez l'ensemble des Crenarchaeota qui possèdent uniquement une PolB1 et une PolB3 pour répliquer leur ADN. Est-ce que l'ADN primase, qui possède une double activité ARN-ADN polymérase, joue un rôle comparable au complexe Polα–primase présent chez les eucaryotes (Lao-Sirieix and Bell, 2004; Liu et al., 2001) ? Est-ce qu'une autre ADN polymérase est recrutée pour remplir cette fonction ? Est-ce qu'un facteur de réplication non encore identifié module la capacité de l'une ou l'autre des PolB à s'accommoder d'une amorce ARN ?

La détermination du rôle respectif des ADN polymérases durant la réplication anima de la même façon les études menées sur les modèles eucaryotes. Dernièrement, l'utilisation de mutants des ADN polymérases δ et ε enclins à introduire des erreurs lors de la synthèse de

l'ADN a permis de démontrer que l'ADN polymérase δ réplique le brin retardé alors que l'ADN polymérase ε réplique le brin continu (Nick McElhinny et al., 2008; Pursell et al., 2007). Une stratégie similaire pourrait permettre à l'avenir d'élucider le rôle respectif des PolB1, PolB3 et PolD durant la réplication de l'ADN chez les Archaea.

RNase HII/FEN-1/ADN ligase

La distribution des gènes codant pour la RNase HII, l'endonucléase FEN-1 et l'ADN ligase indique que la machinerie de réplication du dernier ancêtre commun des Archaea possédait chacune de ces trois protéines (non montré). On dénote juste des cas isolés d'organismes chez lesquels une duplication a eu lieu. Par exemple, on dénombre deux gènes codant l'ADN ligase et deux gènes codant l'endonucléase FEN-1 dans le génome de *T. pendens*. De manière intéressante l'un des deux gènes codant une des deux ADN ligases est adjacent à l'un des deux gènes codant l'une des deux endonucléases, ce qui suggère que les deux protéines interagissent. En dehors du génome de *K. cryptofilum* dans lequel on dénombre trois gènes pour la RNase HII, la totalité des génomes d'Archaea contient un seul gène *rnhB*. Enfin, certains génomes contiennent également un, voire plusieurs gènes codant une ribonucléase HI. Il a été montré que ces protéines sont capables de dégrader des duplex ADN-ARN (Ohtani et al., 2004a, b). La protéine RNase HI de *S. tokodaii* possède en outre la particularité de dégrader des duplex ARN-ARN (Ohtani et al., 2004a). Ces enzymes sont donc susceptibles d'intervenir au même titre que la RNase HII dans la maturation des fragments d'Okazaki (Ohtani et al., 2004a, b).

ADN topoisomérases

Tous les génomes d'Archaea présentent les gènes *top6A* et *top6B* qui codent respectivement pour les sous-unités A et B de l'ADN topoisomérase VI (Topo VI), à l'exception notable des

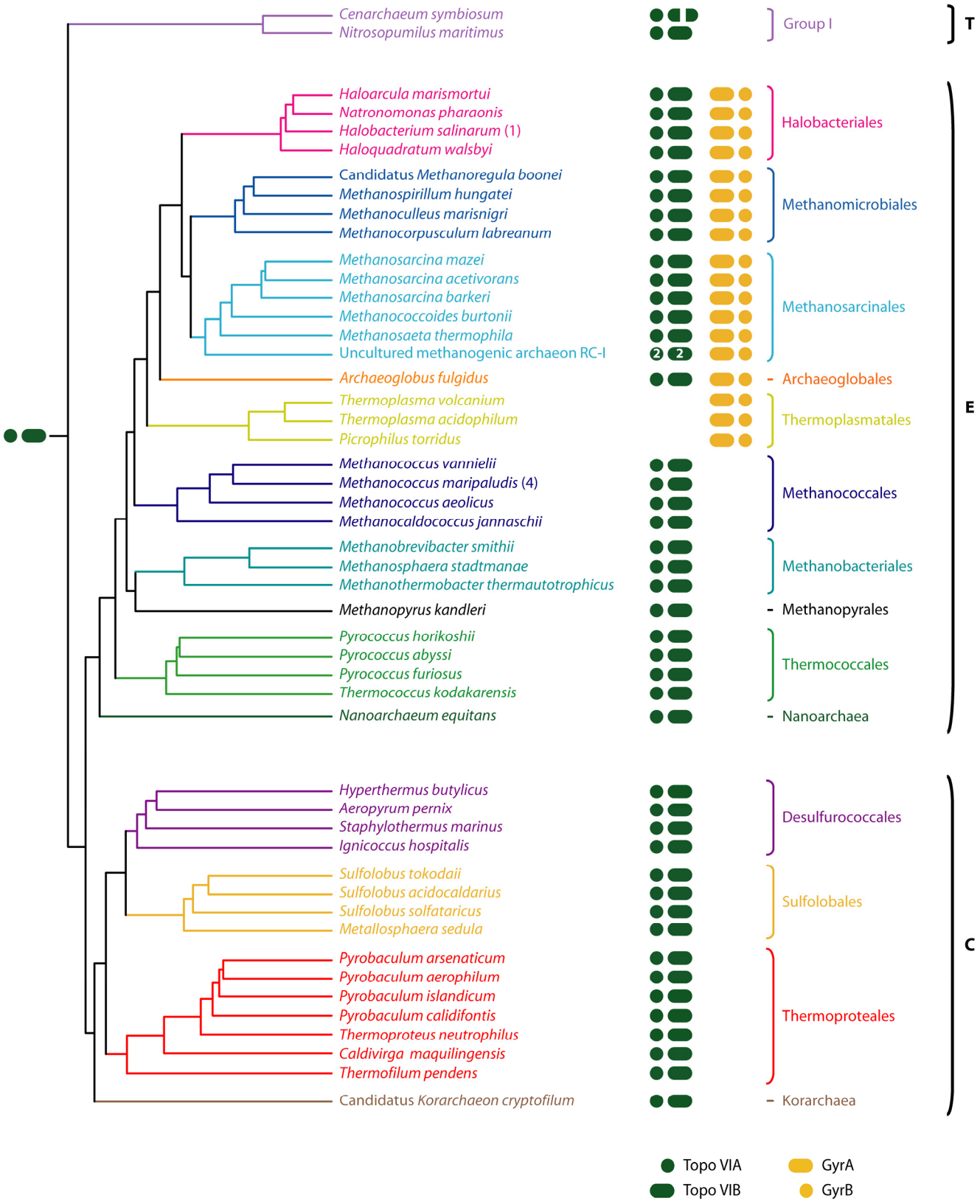

Figure IV-10 : Profil phylétique des gènes codant les deux sous-unités de la topoisomérase VI et des gènes codant les deux sous-unités de l'ADN gyrase chez les Archaea. La présence des sous-unités A (Topo VIA) et B (Topo VIB) de l'ADN topoisomérase VI sont signalées respectivement par un disque et un rectangle arrondi couleur vert foncé. Au-delà de un, le nombre d'homologues est indiqué par un chiffre de couleur blanche. La scission du gène codant la TopoVIB de *Cenarchaeum symbiosum* est signalée par les deux demi-rectangles arrondis. La présence des sous-unités A (GyrA) et B (GyrB) de l'ADN gyrase sont indiquées respectivement par un rectangle arrondi et un disque couleur orange.

génomes des Thermoplasmatales (Kawashima et al., 2000; Ruepp et al., 2000). Chez cet ordre, une ADN gyrase d'origine bactérienne se substitue à la Topo VI (Gadelle et al., 2003). Par ailleurs, on trouve conjointement une ADN gyrase et une Topo VI chez un certain nombre d'Euryarchaeota appartenant à des ordres apparentés. En effet, en sus des Thermoplasmatales, on détecte la présence des gènes *gyrA* et *gyrB* dans tous les génomes des organismes appartenant aux ordres Archaeoglobales, Methanosarcinales, Methanomicrobiales et Halobacteriales **(Figure IV-10)**. En outre, ces deux gènes sont présents dans le génome de l'archée méthanogène non cultivée provenant d'une rizière. D'une part, ce profil suggère que cette archée méthanogène est apparentée aux méthanogènes de la classe II (Bapteste et al., 2005). D'autre part, cela suggère que le transfert du gène codant l'ADN gyrase a eu lieu chez l'ancêtre commun du groupe comprenant les Thermoplasmatales, les Archaeoglobales, les Methanosarcinales, les Methanomicrobiales et les Halobactériales **(Figure IV-10)**. Néanmoins, les analyses phylogénétiques suggèrent au contraire que les gènes codant les deux sous-unités de l'ADN gyrase ont été acquis via plusieurs transferts horizontaux de gènes en provenance des bactéries et non grâce à un transfert unique chez l'ancêtre commun (Forterre et al., 2007b). Toutefois, il convient de noter que cette phylogénie est mal résolue ; par conséquent, cette analyse phylogénétique ne permet pas d'exclure l'hypothèse d'un transfert horizontal unique chez l'ancêtre du groupe d'organismes. En outre, l'identité de la bactérie donneuse reste à ce jour indéterminée.

De manière notable, on observe deux copies des gènes *top6A* et *top6B* dans le génome de l'archée méthanogène non cultivée provenant d'une rizière. Dans les deux cas, les gènes *top6A* et *top6B* sont contigus. De manière intéressante, les deux sous-unités VIA d'une part et les deux sous-unités VIB d'autre part sont assez divergentes, ce qui suggère que la duplication des gènes *top6A* et *top6B* est ancienne ou que l'une des Topo VI a rapidement divergé, voire acquis une nouvelle fonction.

Enfin, les Thaumarchaeota possèdent à la différence du reste des Archaea, une Topo IB, une ADN topoisomérase très répandue chez les eucaryotes (Céline Brochier-Armanet, Simonetta Gribaldo, Patrick Forterre ; résultats non publiés). Les analyses phylogénétiques indiquent que le gène codant la Topo IB n'a pas été acquis par un transfert horizontal en provenance des eucaryotes, mais plutôt que ce gène était déjà présent chez l'ancêtre des Archaea et des Eucarya (Céline Brochier-Armanet, Simonetta Gribaldo, Patrick Forterre ; résultats non publiés). Aussi, le gène codant la Topo IB a vraisemblablement été perdu chez l'ancêtre des Crenarchaeota et des Euryarchaeota.

Conclusion & Perspectives

Les conclusions dressées dans ce chapitre reposent uniquement sur l'analyse de la distribution des gènes de la réplication dans les génomes d'Archaea. Dans la plupart des cas, le profil phylétique de chacun des gènes de la réplication donne toutefois une idée de leur histoire évolutive respective. En suivant une logique parcimonieuse, il est souvent possible d'inférer la présence ou l'absence du gène considéré chez le dernier ancêtre commun des archées. Par extra-polation, il est possible de prédire la composition de la machinerie de réplication de l'ADN chez cet ancêtre. Tous les scénarios évolutifs proposés devront être confirmés par des analyses phylogénétiques qui permettront de retracer de manière précise l'histoire évolutive de ces gènes chez les Archaea. Prise dans son ensemble, cette analyse soulève néanmoins certaines questions évolutives intéressantes.

Premièrement, l'analyse de la distribution des gènes de la réplication dans les génomes d'Archaea met en évidence l'existence d'un groupe de gènes conservé à l'échelle du domaine des Archaea. Ces gènes codent pour un noyau protéique comprenant : la protéine initiatrice de la réplication Cdc6/Orc1, la sous-unité Gins15 du complexe GINS, l'hélicase réplicative

MCM, le facteur de processivité PCNA, le complexe RFC qui dirige le chargement du PCNA, l'ADN primase, la RNase HII, FEN-1, l'ADN ligase.

Deuxièmement, cette analyse appuie l'idée selon laquelle la machinerie de réplication chez le dernier ancêtre commun des Archaea était comparable, voire plus élaborée que celle présente chez certains organismes actuels. En particulier, les profils phylétiques des gènes codant le complexe GINS atteste que le dernier ancêtre commun des Archaea possédait déjà les gènes codant les protéines Gins15 et Gins23 et que le gène *gins23* a vraisemblablement été perdu chez les Euryarchaeota, après la séparation de la branche menant aux Thermococcales, et chez *N. equitans*. D'autre part, les profils phylétiques des gènes codant les PolB et PolD suggèrent que cet ancêtre disposait déjà d'une PolD et d'au moins une PolB pour répliquer son génome. Cette observation implique *de facto* que la PolD a été perdue chez l'ensemble des Crenarchaeota à l'exception de *K. cryptofilum*. Enfin, la distribution de la Topo IB et une analyse phylogénétique suggèrent que le dernier ancêtre commun des Archaea possédait cette ADN topoisomérase et qu'elle a été perdue chez l'ensemble des Archaea à l'exception des Thaumarchaeota (Céline Brochier-Armanet, Simonetta Gribaldo, Patrick Forterre ; résultats non publiés).

Troisièmement, certains gènes (par exemple, les gènes *cdc6/orc1*, *mcm* et *rfcS*) sont présents sous la forme de plusieurs paralogues ce qui pose des questions quant à l'origine de ces copies surnuméraires (duplication, origine extra-chromosomique), mais aussi quant au rôle des différents produits d'expression durant la réplication. Il a été récemment suggéré que ces protéines surnuméraires pourraient moduler l'activité de certains facteurs de réplication (McGeoch and Bell, 2008). L'origine de ces copies devrait quant à elle pouvoir être établie en construisant des arbres phylogénétiques.

Quatrièmement, l'analyse comparative des génomes d'archées montre que le répertoire des gènes impliqués dans la réplication de l'ADN varie selon le génome considéré. En effet, il

n'y a pas une version du réplisome archéen mais une multiplicité de variantes construites autour du cœur protéique commun. Cette hétérogénéité de la machinerie de réplication n'est pas une singularité des Archaea car elle s'observe aussi au niveau du réplisome bactérien, en particulier en ce qui concerne les protéines régulatrices de l'initiation (pour une revue, voir (Zakrzewska-Czerwinska et al., 2007)). De la même façon, la machinerie de réplication varie au sein des eucaryotes, certaines protéines n'étant observées que dans certains phylums, soit qu'il s'agisse d'une innovation fonctionnelle comme la protéine geminin, qui n'est présente que chez les métazoaires, soit que la protéine ait été perdue à plusieurs reprises au cours de l'évolution, comme la protéine Cdt1 (Iyer and Aravind, 2006).

Pour conclure, les résultats de cette analyse suggèrent que le dernier ancêtre commun des Archaea possédait déjà une machinerie de réplication relativement élaborée et que certains gènes de la réplication ont été perdus au cours de l'évolution. Cette réduction de la machinerie ancestrale de réplication a été contrebalancée par l'acquisition de nouvelles protéines (WhiP, PolB2) ou l'expansion du nombre de copies de gènes codant des protéines déjà présentes chez l'ancêtre (Cdc6/Orc1, MCM, PCNA). L'origine et le rôle de ces copies surnuméraires ne sont pas clairement établis, mais l'hétérogénéité observée à l'échelle des Archaea témoigne de la grande plasticité de la machinerie de réplication par rapport aux autres machineries informationnelles. Une analyse comparative des protéines ribosomales a montré que la composition du ribosome des Archaea est étonnamment malléable mais cette plasticité se caractérise par la perte de protéines facultatives et non par l'acquisition de protéines nouvelles (Lecompte et al., 2002). Le contenu protéique de l'ARN polymérase est quant à lui particulièrement homogène (voir, (Kwapisz et al., 2008)). Aussi, la plasticité de la machinerie de réplication des Archaea est singulière dans le sens où elle consiste d'une part en la perte de gènes ancestraux, d'autre part en l'acquisition de gènes nouveaux.

Par ailleurs, la perte de protéines ancestrales observée à l'échelle du réplisome suggère que certains systèmes cellulaires des Archaea ont pu évoluer par réduction et non par complexification, selon le schéma communément admis. A ce titre, le ribosome représente l'exemple le plus manifeste d'évolution réductrice chez les Archaea (Lecompte et al., 2002). D'ailleurs, certains auteurs soutiennent que l'ancêtre commun universel était une entité d'une certaine complexité, comparable à celle observée chez les eucaryotes actuels, et que l'émergence des domaines Archaea et Bacteria se caractérise par une évolution minimaliste (Forterre, 1995; Glansdorff et al., 2008; Poole et al., 1999). Ces hypothèses évolutives trouvent écho dans les résultats de deux études comparatives récentes qui suggèrent que la diversification des différents domaines s'est accompagnée d'une réduction du répertoire architectural protéique (Kurland et al., 2007; Wang et al., 2007), en particulier chez les Archaea (Wang et al., 2007).

CONCLUSION GENERALE

Conclusion générale

La réplication de l'ADN est un processus fondamental permettant d'assurer la pérennité des organismes cellulaires des trois domaines du vivant (*Archaea*, *Bacteria*, *Eucarya*). Au cours de ma thèse, j'ai étudié la réplication de l'ADN chez les Archaea, des organismes unicellulaires partageant de nombreux points communs avec les eucaryotes en matière de traitement de l'information génétique. J'ai développé mon étude autour de deux axes de recherche : une approche *in vitro* et une approche *in silico*.

D'une part, j'ai cherché à poser les bases vers l'élaboration du premier système de réplication *in vitro* chez les Archaea, car un tel système offrirait une grande souplesse pour tenter d'élucider la séquence d'évènements qui anime le processus de la réplication. A ce titre, mon travail s'est concentré sur la recherche de conditions d'expression d'une forme soluble de la protéine initiatrice de la réplication chez *Pyrococcus furiosus*. Cet objectif n'a toutefois pas pu être atteint en raison de l'instabilité intrinsèque de cette protéine.

D'autre part, j'ai effectué une analyse comparative du contexte génomique des gènes de la réplication dans les génomes d'Archaea afin de prédire des interactions fonctionnelles. De manière intéressante, cette méthode nous a permis de mettre en évidence des associations conservées entre certains gènes de la réplication, ce qui suggère que les protéines codées par ces gènes interagissent au niveau de la fourche de réplication. En outre, certains gènes de réplication localisent de manière fréquente avec des gènes codant des protéines impliquées dans d'autres processus de traitement de l'information génétique (transcription, traduction, réparation), ce qui pourrait indiquer que des liens fonctionnels existent entre ces différents processus. En particulier, un groupe de trois gènes de réplication forme une structure de type

opéron, très conservée à l'échelle des Archaea, avec quatre gènes associés au ribosome (traduction ou maturation des ribosomes). Cette association suggère selon nous l'existence d'un mécanisme de couplage entre la réplication de l'ADN et la traduction, voire la biogenèse des ribosomes. Une analyse bibliographique axée sur la recherche de liens entre la réplication et la synthèse protéique nous a permis de collecter des informations, issues d'études réalisées sur des modèles bactériens et eucaryotes, appuyant l'existence de tels mécanismes.

Afin d'éprouver la pertinence biologique de ces prédictions, j'ai cloné, avec l'appui technique de collaborateurs, l'ensemble des gènes impliqués dans les associations conservées mises en évidence au cours de l'analyse comparative des génomes. Ensuite, j'ai adapté et optimisé une méthode de criblage des interactions physiques basée sur une co-expression des candidats protéiques d'interaction. Celle-ci est désormais prête à être mise en œuvre pour une exploration globale des interactions fonctionnelles entre protéines.

Finalement, j'ai étudié la distribution de chacun des gènes de la réplication chez les Archaea en me basant sur les données phylogénétiques les plus récentes. L'analyse des profils phylétiques de chacun des gènes selon une logique de parcimonie suggère que la composition de la machinerie de réplication de l'ancêtre commun des Archaea était probablement plus complexe que celle observée chez les organismes archéens contemporains. La construction d'arbres phylogénétiques pour chacun des gènes de la réplication devraient nous permettre de confirmer ou d'infirmer cette hypothèse.

MATERIELS ET METHODES

Etude de la protéine initiatrice de la réplication du chromosome de *Pyrococcus furiosus* (*Pfu*) : *Pfu*Cdc6/Orc1

Analyse du contexte génomique des gènes de la réplication dans les génomes d'Archaea

Recherche d'interactions physiques entre protéines

Etude de la protéine initiatrice de la réplication du chromosome de *Pyrococcus furiosus* (*Pfu*) : *Pfu*Cdc6/Orc1

Clonages, expression protéique et purification

Clonage

Le gène *cdc6/orc1* a été amplifié par PCR à partir d'un extrait d'ADN génomique de *P. furiosus* à l'aide d'amorces gène-spécifiques (voir Annexes, fiche clonage) en utilisant une polymérase à haute-fidélité (*Pyrobest*™, Takara). A l'issue de la réaction, le produit d'amplification a été purifié sur gel d'agarose (Wizard SV Gel and PCR Clean-up System, Promega), soumis à l'action de l'ADN polymérase extraite de *Thermus aquaticus* avant d'être ligaturé avec le vecteur de clonage pGEM®-T easy (Promega), en suivant les recommandations du fournisseur (Promega Technical Manual No. 042). A ce stade, la nature de l'insert cloné a été confirmée par séquençage pour s'assurer que la séquence du fragment cloné est conforme au gène d'intérêt (voir la section Séquençage ci-après). Les sept inserts séquencés présentaient systématiquement une mutation à la position 113 (une cytosine se substitue à une adénine). La construction plasmidique ainsi obtenue a été dénommée pGEM-T [PF0017 a113c].

Mutagenèse dirigée par PCR

La construction plasmidique pGEM-T [PF0017 a113c] a été amplifiée par PCR en présence d'un couple d'amorces mutagènes (voir Annexes, fiche clonage) afin de rétablir la forme

sauvage du gène *cdc6/orc1* en position 113. A l'issue de la réaction, la matrice parentale a été digérée en incubant deux heures à 37°C en présence de l'enzyme de restriction DpnI puis le produit de réaction a été utilisé pour transformer des cellules compétentes *Escherichia coli* JM109 (TaKaRa) selon la méthode de transformation chimique. La séquence de l'insert a été vérifiée pour s'assurer que la mutation était effective. La construction plasmidique ainsi obtenue a été dénommée pGEM-T [PF0017].

Sous-clonage dans un vecteur d'expression

Le gène codant la protéine *Pfu*Cdc6/Orc1 a été excisé de la construction plasmidique pGEM-T [PF0017] par une hydrolyse enzymatique NdeI/XhoI. Le produit de la réaction a alors été purifié sur gel d'agarose avant d'être ligaturé avec le plasmide pET-28a. A ce stade, la nature de l'insert cloné a été confirmée par séquençage pour s'assurer que la séquence du fragment cloné est conforme au gène d'intérêt. La construction plasmidique obtenue (pET-28a [PF0017]) permet l'expression d'une protéine *Pfu*Cdc6/Orc1 équipée d'une étiquette hexahistidine clivable en position N-terminale.

Séquençage

Le fragment d'ADN recombinant contenu dans le vecteur a été amplifié par PCR à partir d'un couple d'amorces *ad hoc* en présence du réactif GenomeLab™ DTCS Quick Start Master Mix (Beckman Coulter), selon les recommandations du fournisseur. Puis, le produit de la réaction de PCR a été purifié par une chromatographie d'exclusion sur une résine Sephadex™ G25 (Amersham Biosciences), concentré par évaporation sous vide et repris dans la solution de dépôt fournie avec le kit. Les échantillons ont alors été transférés dans une plaque de dépôt qui a été chargée dans un séquenceur Beckman-Coulter CEQ 2000 XL (Beckman Coulter).

Les séquences obtenues ont été analysées manuellement à l'aide du logiciel BioEdit (Hall, 1999).

Constitution d'un stock glycérol

La construction plasmidique pET-28a [PF0017] a été utilisée pour transformer des cellules compétentes Rosetta(DE3)pLysS (Novagen) selon une méthode de transformation chimique (Promega Technical Manual No. 042). Le produit de la transformation a été étalé sur un milieu solide LB contenant 30 µg/ml de kanamycine, 34 µg/ml de chloramphenicol et 1% de glucose (afin de réduire la fuite du promoteur T7). Une colonie isolée a été utilisée pour ensemencer un milieu liquide LB contenant 30 µg/ml de kanamycine, 34 µg/ml de chloramphenicol et 1% de glucose, et la culture a été incubée une nuit à 37°C sous agitation modérée. Afin de constituer un stock, un aliquot de la culture de nuit a été mélangé à une solution stérile de glycérol 50% (concentration finale de 8%) et conservé à -80°C.

Recherche des conditions optimales d'expression d'une forme soluble de PfuCdc6

Culture d'expression. La surface du bloc glycérolé a été gratté pour récupérer les quelques cellules nécessaires à l'ensemencement d'un milieu liquide LB contenant 30 µg/ml de kanamycine, 34 µg/ml de chloramphenicol et 1% de glucose. La culture a été incubée une nuit à 37°C sous agitation modérée. La pré-culture de nuit a été utilisée pour ensemencer un milieu LB contenant 30 µg/ml de kanamycine et 34 µg/ml de chloramphenicol à une densité optique à 600 nm (DO_{600}) égale à 0,1. Les cellules ont été incubées à 37°C sous agitation modérée jusqu'à ce que la DO_{600} atteigne 0,6. L'expression protéique a alors été induite par ajout d'isopropyl-1-thio-β-D-galactopyranoside (IPTG) à une concentration finale de 1 mM. Les cellules ont ensuite été placées dans un incubateur à 37°C, 45°C ou 46°C, pendant une durée de 3, 4, 5, 6, 7, 8, ou 24 heures en faisant varier différents paramètres d'ajustement (vitesse

d'agitation, niveau d'oxygénation). A l'issue de la période d'expression, les cellules ont été récoltées par centrifugation (5000 g, 20 minutes, 4°C) puis stockées à -20°C jusqu'à utilisation.

Tampons de lyse. Les culots cellulaires ont été repris dans différents tampons afin d'essayer de stabiliser la protéine lors des étapes subséquentes de purification par chromatographie. La majorité des essais ont été réalisés dans un tampon Tris [20 mM Tris-HCl (pH 8.0) ; 10% glycerol ; 0,5 mM 1,4-Dithiothreitol ; 0.5 mM Ethylenediaminetetraacetic acid] éventuellement supplémenté en chlorure de sodium, en glutamate de potassium (400 mM) ou en Triton X100. Un tampon à base de HEPES [25 mM HEPES-KOH (pH 7.6) ; 1 mM DTT] supplémenté en glutamate de potassium (50 mM, 100 mM, 150 mM ou 200 mM) a également été utilisé.

Sonication. Après resuspension des cellules dans le tampon de lyse, les parois cellulaires ont été rompues sur la glace en délivrant des décharges d'ultrasons de 5 secondes pour une durée cumulée de 10 minutes à l'aide d'un sonicateur programmable (Misonix) ; chaque salve était suivie d'une période de relâche de 10 secondes. Les débris cellulaires ont ensuite été éliminés par l'intermédiaire d'une étape de centrifugation (20000 g, 25 minutes, 4°C).

Elimination des protéines thermolabiles. La fraction protéique soluble débarrassée des débris membranaires a été chauffée pendant 20 minutes à 75°C pour éliminer la majorité des contaminants bactériens. La fraction protéique soluble thermostable a été récupérée après une nouvelle centrifugation (20000 g, 25 minutes, 4°C).

Essais de purification de PfuCdc6/Orc1 à partir de la fraction soluble

Elimination de l'ADN. Différentes approches ont été employées pour tenter d'éliminer les acides nucléiques contaminant la fraction soluble de la protéine *Pfu*Cdc6/Orc1. 1) L'extrait protéique total a été incubé en présence de DNaseI (20 mg/l) durant 1 heure à 4°C ou à 37°C, centrifugé (4000 g, 20 minutes) avant de procéder à l'étape de chauffage. 2) Les acides nucléiques contenus dans la fraction soluble thermostable ont été précipités par ajout de 0,15% de polyethylenimine en présence de concentrations variées de chlorure de sodium (gamme de NaCl entre 0,1 M et 1,2 M ; incrément de 0,1 M). Après centrifugation (20000 g, 25 minutes, 4°C), la fraction soluble a été analysée par SDS-PAGE. 3) Les protéines ont été précipitées en présence de sulfate d'ammonium (80% saturation), resuspendues dans un tampon à faible force ionique avant injection sur une chromatographie d'interaction hydrophobe HiTrap™ Phenyl 1 ml (Amersham Biosciences).

Chromatographie d'affinité IMAC. La fraction protéique a été filtrée sur une membrane 0,22 µm avant d'être injectée sur une colonne d'affinité HisTrap™ HP, préalablement équilibrée, montée sur un appareil de chromatographie FPLC (Amersham Biosciences). A l'issue de l'injection, la colonne a été lavée avec un grand volume de tampon contenant de l'imidazole à faible concentration puis les protéines ont été éluées en augmentant la concentration en imidazole suivant un gradient linéaire.

Recherche d'interactions physiques par la technique de résonance du plasmon de surface

Principe de la méthode

La résonance du plasmon de surface mesure l'interaction entre deux partenaires moléculaires dont l'un est attaché à la surface d'une puce (le ligand) et l'autre est en solution (l'analyte). La technique consiste à faire passer la protéine dont on souhaite connaître la capacité à interagir (l'analyte) au dessus d'une puce sur laquelle le partenaire d'interaction potentiel (le ligand) a été fixé de manière covalente. A l'issue de la mesure, la puce est régénérée grâce à un traitement chimique qui élimine les analytes fixés à la surface de la puce sans endommager le ligand. L'ensemble des opérations et analyses décrites ci-après ont été réalisées sur un modèle Biacore J (Biacore) à une température de 25°C. L'ensemble des mesures ont été réalisées avec un débit constant de 30 µl/min dans un tampon HBS-P [10 mM HEPES (pH 7.4) ; 150 mM NaCl ; 0,005% TWEEN®-20]. A l'issue de la mesure, la puce a été régénérée en lavant avec une solution de NaCl à 2M.

Purification de PfuCdc6/Orc1 en conditions dénaturantes

La protéine *Pfu*Cdc6/Orc1 équipée d'une étiquette hexahistidine clivable en position N-terminale (His6_*Pfu*Cdc6) a été exprimée dans des cellules recombinantes de levure *Pichia pastoris* (EasySelect™ *Pichia* expression kit, Invitrogen), suivant les recommandations du fournisseur (voir Annexes, Protocole 1). La majorité de la protéine se trouvant dans la fraction insoluble, la protéine His6_*Pfu*Cdc6 a été purifiée par une chromatographie d'affinité IMAC (Ni^{+}-NTA agarose, Qiagen) en conditions dénaturantes (voir Annexes, Protocole 2). La protéine a ensuite été renaturée par la méthode de dilution rapide, concentrée sur résine Ni-NTA agarose et l'étiquette a été clivée par la thrombine avant élution. La protéine a alors été

dialysée à température ambiante avant d'être stockée à 25°C pour une utilisation rapide ou congelée dans de l'azote liquide et stockée à -80°C pour conservation.

Fabrication de la puce PfuCdc6

La matrice hydrophile d'une puce CM5 (carboxymethylated dextran) a été activée par l'injection de 190 µl d'un mélange de N-hydroxysuccinimide (NHS) et de N-ethyl-N'-(dimethylaminopropyl)carbodiimide (EDC) (0,2 M/0,05 M) suivant les recommandations du fabricant (Biacore). Puis, la protéine *Pfu*Cdc6 (190 µl d'une solution à 100 µg/ml dans le tampon de couplage à 10 mM acetate (pH 4.5)) a été injectée pendant 6 minutes au dessus de la puce CM5 activée. Enfin, les groupes réactifs demeurant à la surface de la puce ont été neutralisés par ajout de 190 µl d'une solution d'éthanolamine 1M.

Recherche d'interactions avec Cdc6 comme analyte

La protéine *Pfu*Cdc6 (70 µl d'une solution à 1 µM) a été injectée au dessus de chacune des protéines pour lesquelles une puce était disponible au sein du laboratoire du professeur Ishino. Chacun des ligands potentiels testés correspond à une protéine de *P. furiosus* impliquée dans la réplication ou la réparation de l'ADN étudiée dans le laboratoire du professeur Ishino : PCNA (PF0983) (Kiyonari et al., 2006) ; Gins23 (PF0483) (Yoshimochi et al., 2008) ; RecJ (PF2055) (Imamura et Ishino, résultats non publiés) ; ADN ligase (PF1635) (Kiyonari et al., 2006) ; Hel308a/Hjm (PF0677) (Fujikane et al., 2006) ; DNase I (PF2046) (Tori et Ishino, résultats non publiés).

Recherche d'interactions avec Cdc6 comme ligand

La protéine *Pfu*Cdc6 (70 µl d'une solution à 1 µM) seule ou pré-incubée 20 minutes (à 4°C ou à température ambiante) en présence de 500 µM d'ATP a été injectée au dessus de la

surface de la puce *Pfu*Cdc6-CM5. La protéine *Pfu*MCM a également été injectée à la surface de la puce *Pfu*Cdc6-CM5 pour analyse.

Analyse du contexte génomique des gènes de la réplication dans les génomes d'Archaea

Identification des gènes de la réplication dans les génomes d'Archaea

Une liste de 12 facteurs probablement impliqués dans le processus de réplication de l'ADN chez les Archaea a été établie. Cette liste comprend : le facteur d'initiation de la réplication Cdc6/Orc1 ; les ADN polymérases de la famille B (PolB1, PolB2 et PolB3) ; la petite et la grande sous-unité de l'ADN polymérase D (DP1 et DP2) ; l'hélicase réplicative MCM ; le facteur de processivité PCNA ; les petite et grande sous-unités du RFC (RFC-s et RFC-l), le facteur de chargement du PCNA ; la petite et la grande sous-unité de l'ADN primase (PriS et PriL) ; la protéine de fixation à l'ADN simple brin (SSB ou RPA selon les organismes) ; l'ADN ligase ; les deux sous-unités de l'ADN topoisomérase VI (Topo VIA et Topo VIB) ; la RNase HII ; l'endonucléase spécifique des structures 'flap', FEN-1 ; les deux sous-unités Gins (Gins15 et Gins23) du complexe GINS. Les références d'accès de chacune de ces protéines ou sous-unités protéiques ont été extraites des 27 génomes d'Archaea séquencés à la date du 1er Avril 2006 (*Aeropyrum pernix* K1 (Kawarabayasi et al., 1999; Natale et al., 2000; Yamazaki et al., 2006) ; *Pyrobaculum aerophilum* str. IM2 (Fitz-Gibbon et al., 2002) ; les trois Sulfolobales *Sulfolobus acidocaldarius* DSM 639 (Chen et al., 2005a), *S. solfataricus* P2 (She et al., 2001) et *S. tokodaii* str. 7 (Kawarabayasi et al., 2001) ; *Nanoarchaeum equitans* Kin4-M (Waters et al., 2003) ; *Archaeoglobus fulgidus* DSM 4304 (Klenk et al., 1997) ; les trois Halobacteriales *Haloarcula marismortui* ATC 43049 (Baliga et al., 2004), *Halobacterium salinarum* str. NRC-1 (Gruber et al., 2004; Ng et al., 2000) et *Natronomonas pharaonis* DSM 2160 (Falb et al., 2005) ; les deux Methanobacteriales *Methanothermobacter*

thermautotrophicus str. deltaH (Smith et al., 1997) et *Methanosphaera stadtmanae* DSM 3091 (Fricke et al., 2006) ; les deux Methanococcales *Methanocaldococcus jannaschii* DSM 2661 (Bult et al., 1996) et *Methanococcus maripaludis* S2 (Hendrickson et al., 2004) ; *Methanopyrus kandleri* (Slesarev et al., 2002); les quatre Methanosarcinales *Methanococcoides burtonii* DSM 6242, *Methanosarcina acetivorans* C2A (Galagan et al., 2002), *M. barkeri* str. Fusaro (Maeder et al., 2006) et *M. mazei* Go1 (Deppenmeier et al., 2002) ; *Methanospirillum hungatei* JF-1 ; les quatre Thermococcales *Pyrococcus abyssi* GE5 (Cohen et al., 2003), *P. furiosus* DSM 3638, *P. horikoshii* OT3 (Kawarabayasi et al., 1998a, b) et *Thermococcus kodakaraensis* KOD1 (Fukui et al., 2005) ; les trois Thermoplasmatales *Picrophilus torridus* DSM 9790 (Futterer et al., 2004), *Thermoplasma acidophilum* DSM 1728 (Ruepp et al., 2000) et *T. volcanium* GSS1 (Kawashima et al., 2000)) par des recherches BLASTP et PSI-BLAST (Altschul et al., 1997) réalisées sur le site du NCBI (http://blast.ncbi.nlm.nih.gov/Blast.cgi). Les séquences protéiques de référence qui ont été utilisées pour mener ces recherches correspondaient aux homologues de *P. furiosus* et de *S. solfataricus* ; le cas échéant, les séquences de protéines ayant fait l'objet d'une caractérisation biochimique ont été utilisées. Toutes les protéines de la liste ainsi identifiées ont été assignées à un ensemble de groupes orthologues (Cluster of Orthologous Groups, COGs) (Tatusov et al., 1997; Tatusov et al., 2001) à l'aide du programme COG guess (http://www-archbac.u-psud.fr/genomics/GenomicsToolBox.html) afin de confirmer leur annotation. Ensuite, les génomes d'Archaea entièrement séquencés ont été explorés par BLASTP avec différentes séquences de référence en tant qu'appât afin d'identifier les protéines mal annotées ou les homologues ayant échappé à l'analyse précédente et ce pour chacune des classes de protéines. Enfin, des recherches par BLASTN ont été réalisées sur le site du NCBI (http://blast.ncbi.nlm.nih.gov/Blast.cgi) contre la base de données non redondante de séquences nucléotidiques des génomes d'Archaea afin d'identifier d'éventuelles phases

ouvertes de lecture manquantes, en utilisant un homologue du génome le plus apparenté comme référence.

Analyse de l'environnement génomique des gènes de la réplication

Le contexte génomique de chacun des gènes de la réplication a été analysé manuellement avec le programme Genomapper de la boîte à outils génomiques disponible en ligne sur le site du LBMGE (http://www-archbac.u-psud.fr/genomics/GenomicsToolBox.html) parce que l'association conservée entre les gènes codant les protéines Gins15, PriS et PCNA n'était pas répertoriée dans la base de données STRING (von Mering et al., 2007; von Mering et al., 2005) ; probablement parce que les similarités de séquence entre les différents membres de la famille Gins sont faibles (Marinsek et al., 2006). Par ailleurs, les environnements de gènes conservés entre génomes éloignés ont été utiles au cours de l'identification des homologues archéens des protéines Gins qui échappent aux recherches par PSI-BLAST. Pour chacun des gènes étudiés, une fenêtre comprenant le gène ciblé, les cinq gènes situés en amont et les cinq gènes situés en aval a été analysée. L'identité des protéines codées par les 11 gènes compris dans cet intervalle génomique a été extraite à l'aide du programme Genome guts (http://www-archbac.u-psud.fr/genomics/GenomicsToolBox.html). Puis, chacune des séquences protéiques a été assignée à un COG en utilisant le programme COG guess et une recherche par BLASTP a été réalisée sur le site du NCBI afin de valider l'annotation fonctionnelle. Les abords des gènes de réplication situés sur des éléments extra-chromosomiques n'ont pas été inspectés dans la mesure où la base de données génomiques du LBMGE ne contient aucune séquence de plasmides d'Archaea. Parmi les nombreux homologues Cdc6/Orc1 décelées dans les génomes des Halobacteriales, seuls ceux répertoriés dans la base de données SWISS-PROT ont été retenus pour examen (Bairoch and Apweiler, 2000).

Mise à jour du répertoire des gènes de la réplication

Le répertoire de protéines impliquées dans la réplication de l'ADN a ensuite été mis à jour en incluant les génomes d'Archaea rendus publiques après la date du 1er Avril 2006 : Candidatus *Korarchaeum cryptofilum* OPF8 (Elkins et al., 2008) ; *Cenarchaeum symbiosum* A (Hallam et al., 2006) et *Nitrosopumilus maritimus* SCM1 ; les trois Desulfurococcales *Hyperthermus butylicus* DSM 5456 (Brugger et al., 2007), *Ignicoccus hospitalis* KIN4/I et *Staphylothermus marinus* F1 ; *Metallosphaera sedula* DSM 5348 (Auernik et al., 2008) ; les six Thermoproteales *Caldivirga maquilingensis* IC-167, *Pyrobaculum arsenaticum* DSM 13514, *P. calidifontis* JCM 11548, *P. islandicum* DSM 4184, *Thermofilum pendens* Hrk5 (Anderson et al., 2008) et *Thermoproteus neutrophilus* V24Sta ; les deux Halobacteriales *Halobacterium salinarum* R1 (Pfeiffer et al., 2008a) et *Haloquadratum walsbyi* DSM 16790 (Bolhuis et al., 2006) ; *Methanobrevibacter smithii* ATCC 35061 (Samuel et al., 2007) ; les cinq Methanococcales *Methanococcus aeolicus* Nankai-3, *M. maripaludis* (souches C5, C6 et C7) et *M. vannielii* SB ; les trois Methanomicrobiales Candidatus *Methanoregula boonei* 6A8, *Methanocorpusculum labraneum* Z et *Methanoculleus marisnigri* JR1 ; *Methanosaeta thermophila* PT ; le métagénome de l'archée méthanogène non cultivée provenant d'une rizière (souche RC-I) (Erkel et al., 2006).

Recherche d'interactions physiques entre protéines

Recherche d'interaction par co-immunoprécipitation

Clonage de la sous-unité beta du facteur d'initiation aIF-2

Le gène codant la sous-unité beta du facteur d'initiation IF-2 (PF0481) a été amplifié par PCR à partir d'un extrait d'ADN génomique de *P. furiosus* à l'aide d'amorces gène-spécifiques en utilisant une polymérase à haute-fidélité (*Pyrobest*™, Takara). Le produit d'amplification a été ligaturé avec le vecteur de clonage pGEM®-T easy (Promega) et, à l'issue du clonage, la nature de l'insert cloné a été confirmée par séquençage. Ensuite, le gène a été excisé par hydrolyse enzymatique NdeI/XhoI et le fragment obtenu cloné dans le plasmide pET-28a. La séquence de l'insert a alors été vérifiée une nouvelle fois.

Purification de la protéine aIF-2 beta par chromatographie d'affinité IMAC

La sous-unité beta du facteur d'initiation de *P. furiosus* (*Pfu*aIF-2β) a été purifiée par chromatographie d'affinité IMAC suivant la méthode décrite par Tahara et collaborateurs (Tahara et al., 2004). Brièvement, les cellules *E. coli* BL21-Codon Plus® (DE3)-RIL (Stratagene) ayant surproduit une version de la protéine *Pfu*aIF-2 beta modifiée par ajout d'une étiquette hexahistidine en N-terminale, ont été rompues par sonication et les débris cellulaires éliminés par centrifugation. La fraction soluble a alors été chauffée 10 minutes à 70°C, les protéines thermolabiles éliminées par centrifugation et la fraction soluble thermostable — présentant une teinte rosée — a été injectée sur une colonne HisTrap HP

(Amersham Biosciences). Les protéines accrochées à la résine ont été éluées en augmentant la concentration en imidazole de 5 mM à 1 M selon un gradient linéaire.

Recherche d'interaction entre PfuaIF-2β et PfuMCM par co-immunoprécipitation

Les protéines aIF-2β et MCM de *P. furiosus* ont été mélangées en proportions équimolaires (50 pmoles) dans du tampon PBST [10 mM Tris (pH 8.0) ; 150 mM NaCl ; 0,1% TWEEN®-20] et incubées 10 minutes à 70°C. L'anticorps anti-MCM a alors été ajouté et le mélange incubé 1 heure à température ambiante en agitant. Puis, la protéine A sépharose (Amersham Biosciences) a été ajoutée, le mélange incubé 1 heure supplémentaire avant d'être centrifugé (20000 g, 1 minute). Ensuite, les billes de sépharose couplées à la protéine A ont été lavées à cinq reprises avec du PBS, reprises dans du tampon de charge pour SDS-PAGE, ébouillantées, avant d'être chargées sur un gel acrylamide dénaturant. A l'issue de l'électrophorèse, les protéines ont été transférées sur une membrane PVDF (Immun-Blot® PVDF membrane, Bio-Rad) selon la méthode de transfert semi-sec. Les protéines immunoprécipitées ont alors été révélées par chimioluminescence en utilisant un anticorps dirigé contre le peptide tetrahistidine (ECL Plus Western Blotting Detection Reagent, Amersham) et le signal produit détecté à l'aide d'un système d'acquisition d'images (LAS-3000, Fujifilm).

Recherche d'interactions par la technique de co-purification

Clonages des gènes de la réplication de l'ADN, des gènes de la recombinaison, des gènes de la transcription et des gènes associés au ribosome

Amplification par PCR et clonage dans des vecteurs d'expression. Les gènes codant les protéines suivantes : Cdc6/Orc1 (PF0017) ; DP1 (PF0018) ; DP2 (PF0019) ; PriS (PF0110) ;

PriL (PF0011) ; L44E (PF0217) ; S27E (PF0218) ; aIF-2 beta (PF0481) ; MCM (PF0482) ; Gins23 (PF0483) ; Gins15 (PF0982) ; PCNA (PF0983) ; TFS (PF0986) ; aIF-2 alpha (PF1140) ; Nop10 (PF1141) ; COG2047 (PF1142) ; Sir2 (PF1154) ; NudF (PF1590) ; RecJ (PF2055) ont été clonés dans une paire de vecteurs d'expression (pASH et pKHS) dérivés du plasmide pET (Sophie Quevillon-Cheruel, résultats non publiés). Neuf de ces dix-neuf gènes ont été amplifiés et clonés au cours de travaux pratiques réalisés dans le cadre de la formation DEUST sous l'encadrement de Messieurs Robert Aufrère et Gilles Henckès. Les dix gènes restant ont été amplifiés et clonés par mes soins. La méthode suivie dans les deux cas est décrite ci-après. Chacun des gènes a été amplifié par PCR à partir d'un extrait d'ADN génomique de *P. furiosus* à l'aide d'amorces gène-spécifiques (voir Annexes, fiche clonage respective) en utilisant une ADN polymerase à haute fidélité (*Pfu* DNA polymerase, Promega). Le produit de la réaction de PCR a été purifié sur gel d'agarose avant d'être hydrolysé par les enzymes de restriction EagI et NotI (New England Biolabs). Le produit de la digestion a alors été ligaturé avec les vecteurs pASH et pKHS préalablement linéarisés par l'enzyme NotI et soumis à l'action de la phosphatase alcaline (Euromedex). Le produit de ligature a été utilisé pour transformer des cellules JM109 par la méthode d'électroporation. Le clonage n'étant pas directionnel, l'orientation des inserts a été vérifiée à l'aide d'une PCR sur colonie en utilisant une amorce s'hybridant sur le plasmide et une amorce gène-spécifique. Les clones transformants positifs ont été stockés en glycérol 8% à -80°C. La totalité des manipulations décrites ci-après ont été réalisés par mes soins. A ce stade, chacun des clones positifs a été strié sur boîtes, un clone isolé mis en culture et le plasmide extrait pour analyse. Aussi, l'insert de chaque construction plasmidique a été envoyé pour séquençage (Genome Express) et la nature de l'insert a été confirmée par une analyse manuelle à l'aide du logiciel BioEdit (Hall, 1999).

Sous-clonage. Certaines constructions plasmidiques ont été obtenues en sous-clonant l'insert conforme d'une construction plasmidique vers le vecteur d'expression jumeau. Dans ce cas, l'insert a été excisé de la construction plasmidique donneuse par une hydrolyse enzymatique XbaI/NotI puis le fragment d'ADN obtenu purifié sur gel d'agarose avant d'être ligaturé avec le vecteur receveur. La construction plasmidique a alors été clonée chez *E. coli* et la séquence de l'insert vérifiée.

Mutagenèse dirigée par PCR du gène cdc6/orc1. Dans le cas du clonage du gène *cdc6/orc1*, les constructions pASH et pKHS obtenues contenaient une mutation à la position 113 (voir Chapitre I). Pour rétablir le nucléotide trouvé dans la version sauvage, des amorces mutagènes ont été utilisées pour amplifier la totalité des plasmides pASH et pKHS suivant une réaction de PCR inverse. Après digestion des brins parentaux par l'enzyme DpnI, les plasmides mutés ont été utilisés pour transformer des cellules *E. coli* JM109 par la méthode d'électroporation. La séquence de l'insert a de nouveau été vérifiée pour s'assurer que la mutation était effective.

Elimination du fragment codant l'intéine de PfuMCM. Dans un premier temps, la totalité de la séquence codant la protéine *Pfu*MCM a été amplifiée par PCR et clonée dans les vecteurs pASH et pKHS (travail réalisé par les étudiants de licence professionnelle). Dans un second temps, le fragment codant l'intéine a été supprimé des constructions plasmidiques pASH et pKHS au moyen d'une réaction de PCR inverse initiée avec des amorces phosphorylées à leur extrémité 5' s'hybridant respectivement à l'extrémité 3' de la région codant l'extéine 1 et à l'extrémité 5' de la région codant l'extéine 2. A l'issue de la réaction d'amplification, le plasmide a été refermé à l'aide d'une ligature à bouts francs par l'action de l'ADN ligase T4 (Fermentas). Le produit de la réaction a ensuite été utilisé pour transformer des cellules *E.*

coli JM109. La nature de l'insert de chaque construction plasmidique ainsi obtenue a été vérifiée par séquençage.

Optimisation de la technique de co-purification sur résine Ni-NTA

Un protocole de co-purification sur résine Ni-NTA, mis au point dans le laboratoire de Génomique Structurale dirigé par Herman van Tilbeurgh, a été utilisé comme point de départ à ce travail. La souche d'expression BL21-Gold(DE3) (Stratagene) est co-transformée avec 200 ng de chacune des constructions plasmidiques permettant la co-expression des partenaires protéiques potentiels selon la méthode de transformation chimique (voir Annexes, protocole 3). Parallèlement, la souche d'expression est transformée indépendamment avec 200 ng de l'une ou l'autre de ces deux constructions dans le cas des cultures d'expression témoins de chaque protéine. Après une incubation de 45 minutes à 37°C sous agitation vigoureuse (250 rpm) en milieu 2x YT, la culture de transformants est utilisée pour inoculer 10 ml d'un milieu liquide 2x YT contenant le(s) antibiotique(s) approprié(s) et la culture est placée une nuit à 37°C sous agitation modérée (180 rpm à 200 rpm). Le lendemain, une culture de 10 ml de 2x YT est ensemencée à $DO_{600} = 0,1$ et incubée à 37°C jusqu'à ce que la DO_{600} atteigne 1,0. Puis, l'expression protéique est induite par ajout d'IPTG à une concentration finale de 0,5 mM et la culture est prolongée de 3 heures à 37°C. Les cellules sont alors récoltées par centrifugation (5000 g, 10 minutes) puis stockées à -80°C jusqu'à utilisation. Les culots cellulaires sont repris dans du tampon de lyse et les parois cellulaires sont rompues sur la glace par trois décharges d'ultrasons de 15 secondes entrecoupées d'une période de relâche de 30 secondes. Les débris cellulaires sont éliminés par centrifugation (20000 g, 30 minutes, 4°C) puis la fraction protéique soluble est transférée dans un tube à centrifuger de 2 ml avant d'être incubée pendant 15 minutes à 75°C dans un bloc chauffant. Après centrifugation (20000 g, 30 minutes, 4°C), la fraction protéique thermostable est placée au contact de la

résine Ni-NTA (Qiagen) préalablement lavée et équilibrée dans le tampon de fixation et la concentration en imidazole est ajustée à environ 20 mM. L'extrait protéique thermostable est incubé avec la résine pendant au moins 1 heure à température ambiante sur un agitateur rotatif. Puis, la résine est récoltée par centrifugation, lavée à trois reprises dans du tampon de lavage et, à l'issue du dernier lavage, la résine est mélangée à du tampon de charge. L'ensemble des échantillons protéiques récoltés au cours de la purification sont analysés par SDS-PAGE (Laemmli, 1970).

Recherche d'interactions par la technique de résonance du plasmon de surface

L'ADN primase *Pfu*PriSL (Liu et al., 2001) (70 µl d'une solution à 1 µM) a été injectée au dessus d'une puce à la surface de laquelle était covalemment attachée le facteur de processivité *Pfu*PCNA (Kiyonari et al., 2006). La mesure a été réalisée avec un débit constant de 30 µl/min dans un tampon HBS-P [10 mM HEPES (pH 7.4) ; 150 mM NaCl ; 0,005% TWEEN®-20]. A l'issue de la mesure, la puce a été régénérée en lavant avec une solution de NaCl à 2M.

REFERENCES BIBLIOGRAPHIQUES

Références bibliographiques

Aladjem, M.I., Falaschi, A., and Kowalski, D. (2006). Eukaryotic DNA Replication Origins. In DNA replication and human disease, M.L. DePamphilis, ed. (Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory Press), pp. 31-61.

Alderton, G.K. (2007). DNA replication: Shaping up for a new start. Nature reviews *8*, 754-754.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res *25*, 3389-3402.

Anderson, I., Rodriguez, J., Susanti, D., Porat, I., Reich, C., Ulrich, L.E., Elkins, J.G., Mavromatis, K., Lykidis, A., Kim, E.*, et al.* (2008). Genome sequence of Thermofilum pendens reveals an exceptional loss of biosynthetic pathways without genome reduction. J Bacteriol *190*, 2957-2965.

Aparicio, O.M., Stout, A.M., and Bell, S.P. (1999). Differential assembly of Cdc45p and DNA polymerases at early and late origins of DNA replication. Proceedings of the National Academy of Sciences of the United States of America *96*, 9130-9135.

Aparicio, T., Ibarra, A., and Mendez, J. (2006). Cdc45-MCM-GINS, a new power player for DNA replication. Cell division *1*, 18.

Aravind, L., Leipe, D.D., and Koonin, E.V. (1998). Toprim--a conserved catalytic domain in type IA and II topoisomerases, DnaG-type primases, OLD family nucleases and RecR proteins. Nucleic Acids Res *26*, 4205-4213.

Armengaud, J., Fernandez, B., Chaumont, V., Rollin-Genetet, F., Finet, S., Marchetti, C., Myllykallio, H., Vidaud, C., Pellequer, J.L., Gribaldo, S.*, et al.* (2003). Identification, purification, and characterization of an eukaryotic-like phosphopantetheine adenylyltransferase (coenzyme A biosynthetic pathway) in the hyperthermophilic archaeon Pyrococcus abyssi. The Journal of biological chemistry *278*, 31078-31087.

Auernik, K.S., Maezato, Y., Blum, P.H., and Kelly, R.M. (2008). The genome sequence of the metal-mobilizing, extremely thermoacidophilic archaeon Metallosphaera sedula provides insights into bioleaching-associated metabolism. Applied and environmental microbiology *74*, 682-692.

Avery, O.T., MacLeod, C.M., and McCarty, M. (1944). Studies on the chemical nature of the substance inducing transformation of pneumococcal types. Inductions of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. J. Exp. Med. *79*, 137-158.

Bairoch, A., and Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res *28*, 45-48.

Baliga, N.S., Bonneau, R., Facciotti, M.T., Pan, M., Glusman, G., Deutsch, E.W., Shannon, P., Chiu, Y., Weng, R.S., Gan, R.R.*, et al.* (2004). Genome sequence of Haloarcula marismortui: a halophilic archaeon from the Dead Sea. Genome research *14*, 2221-2234.

Bapteste, E., and Boucher, Y. (2008). Lateral gene transfer challenges principles of microbial systematics. Trends in microbiology *16*, 200-207.

Bapteste, E., Brochier, C., and Boucher, Y. (2005). Higher-level classification of the Archaea: evolution of methanogenesis and methanogens. Archaea (Vancouver, B.C *1*, 353-363.

Barns, S.M., Delwiche, C.F., Palmer, J.D., and Pace, N.R. (1996). Perspectives on archaeal diversity, thermophily and monophyly from environmental rRNA sequences. Proceedings of the National Academy of Sciences of the United States of America *93*, 9188-9193.

Bell, S.P., and Stillman, B. (1992). ATP-dependent recognition of eukaryotic origins of DNA replication by a multiprotein complex. Nature *357*, 128-134.

Berger, J.M. (2008). SnapShot: nucleic acid helicases and translocases. Cell *134*, 888-888 e881.

Bergerat, A., de Massy, B., Gadelle, D., Varoutas, P.C., Nicolas, A., and Forterre, P. (1997). An atypical topoisomerase II from Archaea with implications for meiotic recombination. Nature *386*, 414-417.

Bergerat, A., Gadelle, D., and Forterre, P. (1994). Purification of a DNA topoisomerase II from the hyperthermophilic archaeon Sulfolobus shibatae. A thermostable enzyme with both bacterial and eucaryal features. The Journal of biological chemistry *269*, 27663-27669.

Berquist, B., DasSarma, P., and DasSarma, S. (2007). Essential and non-essential DNA replication genes in the model halophilic Archaeon, Halobacterium sp. NRC-1. BMC Genetics *8*, 31.

Berquist, B.R., and DasSarma, S. (2003). An archaeal chromosomal autonomously replicating sequence element from an extreme halophile, Halobacterium sp. strain NRC-1. J Bacteriol *185*, 5959-5966.

Bochman, M.L., and Schwacha, A. (2008). The Mcm2-7 complex has in vitro helicase activity. Molecular cell *31*, 287-293.

Boetius, A., Ravenschlag, K., Schubert, C.J., Rickert, D., Widdel, F., Gieseke, A., Amann, R., Jorgensen, B.B., Witte, U., and Pfannkuche, O. (2000). A marine microbial consortium apparently mediating anaerobic oxidation of methane. Nature *407*, 623-626.

Bohlke, K., Pisani, F.M., Rossi, M., and Antranikian, G. (2002). Archaeal DNA replication: spotlight on a rapidly moving field. Extremophiles *6*, 1-14.

Bolhuis, H., Palm, P., Wende, A., Falb, M., Rampp, M., Rodriguez-Valera, F., Pfeiffer, F., and Oesterhelt, D. (2006). The genome of the square archaeon Haloquadratum walsbyi : life at the limits of water activity. BMC genomics *7*, 169.

Boskovic, J., Coloma, J., Aparicio, T., Zhou, M., Robinson, C.V., Mendez, J., and Montoya, G. (2007). Molecular architecture of the human GINS complex. EMBO Rep *8*, 678-684.

Boucher, Y. (2007). Lipids: Biosynthesis, Function, and Evolution. In Archaea: molecular and cellular biology, R. Cavicchioli, ed. (Washington, DC: ASM Press), pp. 341-353.

Brochier-Armanet, C., Boussau, B., Gribaldo, S., and Forterre, P. (2008). Mesophilic crenarchaeota: proposal for a third archaeal phylum, the Thaumarchaeota. Nature reviews *6*, 245-252.

Brochier, C., Forterre, P., and Gribaldo, S. (2004). Archaeal phylogeny based on proteins of the transcription and translation machineries: tackling the Methanopyrus kandleri paradox. Genome biology *5*, R17.

Brochier, C., Gribaldo, S., Zivanovic, Y., Confalonieri, F., and Forterre, P. (2005). Nanoarchaea: representatives of a novel archaeal phylum or a fast-evolving euryarchaeal lineage related to Thermococcales? Genome biology *6*, R42.

Brugger, K., Chen, L., Stark, M., Zibat, A., Redder, P., Ruepp, A., Awayez, M., She, Q., Garrett, R.A., and Klenk, H.P. (2007). The genome of Hyperthermus butylicus: a sulfur-reducing, peptide fermenting, neutrophilic Crenarchaeote growing up to 108 degrees C. Archaea (Vancouver, B.C *2*, 127-135.

Bult, C.J., White, O., Olsen, G.J., Zhou, L., Fleischmann, R.D., Sutton, G.G., Blake, J.A., FitzGerald, L.M., Clayton, R.A., Gocayne, J.D.*, et al.* (1996). Complete genome sequence of the methanogenic archaeon, Methanococcus jannaschii. Science (New York, N.Y *273*, 1058-1073.

Burgers, P.M. (2008). Polymerase dynamics at the eukaryotic DNA replication fork. J. Biol. Chem., R800062200.

Burgers, P.M.J., and Seo, Y.-S. (2006). Eukaryotic DNA Replication Forks. In DNA replication and human disease, M.L. DePamphilis, ed. (Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory Press), pp. 105-120.

Burggraf, S., Stetter, K.O., Rouviere, P., and Woese, C.R. (1991). Methanopyrus kandleri: an archaeal methanogen unrelated to all other known methanogens. Systematic and applied microbiology *14*, 346-351.

Cairns, J. (1963). The bacterial chromosome and its manner of replication as seen by autoradiography. Journal of molecular biology *6*, 208-213.

Cann, I.K., Ishino, S., Yuasa, M., Daiyasu, H., Toh, H., and Ishino, Y. (2001). Biochemical analysis of replication factor C from the hyperthermophilic archaeon Pyrococcus furiosus. J Bacteriol *183*, 2614-2623.

Cann, I.K., Komori, K., Toh, H., Kanai, S., and Ishino, Y. (1998). A heterodimeric DNA polymerase: evidence that members of Euryarchaeota possess a distinct DNA polymerase. Proceedings of the National Academy of Sciences of the United States of America *95*, 14250-14255.

Capaldi, S.A., and Berger, J.M. (2004). Biochemical characterization of Cdc6/Orc1 binding to the replication origin of the euryarchaeon Methanothermobacter thermoautotrophicus. Nucleic Acids Res *32*, 4821-4832.

Cavicchioli, R., Curmi, P.M., Saunders, N., and Thomas, T. (2003). Pathogenic archaea: do they exist? Bioessays *25*, 1119-1128.

Cavicchioli, R., Demaere, M.Z., and Thomas, T. (2007). Metagenomic studies reveal the critical and wide-ranging ecological importance of uncultivated Archaea: the role of ammonia oxidizers. Bioessays *29*, 11-14.

Chaban, B., Ng, S.Y., and Jarrell, K.F. (2006). Archaeal habitats--from the extreme to the ordinary. Canadian journal of microbiology *52*, 73-116.

Chedin, F., Seitz, E.M., and Kowalczykowski, S.C. (1998). Novel homologs of replication protein A in archaea: implications for the evolution of ssDNA-binding proteins. Trends in biochemical sciences *23*, 273-277.

Chen, L., Brugger, K., Skovgaard, M., Redder, P., She, Q., Torarinsson, E., Greve, B., Awayez, M., Zibat, A., Klenk, H.P., and Garrett, R.A. (2005a). The genome of Sulfolobus acidocaldarius, a model organism of the Crenarchaeota. J Bacteriol *187*, 4992-4999.

Chen, Y.H., Kocherginskaya, S.A., Lin, Y., Sriratana, B., Lagunas, A.M., Robbins, J.B., Mackie, R.I., and Cann, I.K. (2005b). Biochemical and mutational analyses of a unique clamp loader complex in the archaeon Methanosarcina acetivorans. The Journal of biological chemistry *280*, 41852-41863.

Choi, J.J., Nam, K.H., Min, B., Kim, S.J., Soll, D., and Kwon, S.T. (2006). Protein trans-splicing and characterization of a split family B-type DNA polymerase from the hyperthermophilic archaeal parasite Nanoarchaeum equitans. Journal of molecular biology *356*, 1093-1106.

Ciccarelli, F.D., Doerks, T., von Mering, C., Creevey, C.J., Snel, B., and Bork, P. (2006). Toward automatic reconstruction of a highly resolved tree of life. Science (New York, N.Y *311*, 1283-1287.

Cohen, G.N., Barbe, V., Flament, D., Galperin, M., Heilig, R., Lecompte, O., Poch, O., Prieur, D., Querellou, J., Ripp, R.*, et al.* (2003). An integrated analysis of the genome of the hyperthermophilic archaeon Pyrococcus abyssi. Molecular microbiology *47*, 1495-1512.

Cole, M.D., and Cowling, V.H. (2008). Transcription-independent functions of MYC: regulation of translation and DNA replication. Nature reviews *9*, 810-815.

Copeland, H.F. (1938). The kingdoms of organisms. Quart Rev Biol *13*, 383-420.

Corn, J.E., and Berger, J.M. (2006). Regulation of bacterial priming and daughter strand synthesis through helicase-primase interactions. Nucleic Acids Res *34*, 4082-4088.

Cotterill, S., and Kearsey, S.E. (2008). DNAReplication: a database of information and resources for the eukaryotic DNA replication community. Nucl. Acids Res., gkn726.

Courbet, S., Gay, S., Arnoult, N., Wronka, G., Anglana, M., Brison, O., and Debatisse, M. (2008). Replication fork movement sets chromatin loop size and origin choice in mammalian cells. Nature.

Cullmann, G., Fien, K., Kobayashi, R., and Stillman, B. (1995). Characterization of the five replication factor C genes of Saccharomyces cerevisiae. Molecular and cellular biology *15*, 4661-4671.

Cunningham, E.L., and Berger, J.M. (2005). Unraveling the early steps of prokaryotic replication. Current opinion in structural biology *15*, 68-76.

Dagan, T., Artzy-Randrup, Y., and Martin, W. (2008). Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. Proceedings of the National Academy of Sciences of the United States of America *105*, 10039-10044.

Dagan, T., and Martin, W. (2006). The tree of one percent. Genome biology *7*, 118.

Dandekar, T., Snel, B., Huynen, M., and Bork, P. (1998). Conservation of gene order: a fingerprint of proteins that physically interact. Trends in biochemical sciences *23*, 324-328.

Davey, M.J., and O'Donnell, M. (2003). Replicative helicase loaders: ring breakers and ring makers. Curr Biol *13*, R594-596.

Dawson, S., Delong, E., and Pace, N. (2006). Phylogenetic and Ecological Perspectives on Uncultured Crenarchaeota and Korarchaeota. In The Prokaryotes, pp. 281-289.

De Felice, M., Esposito, L., Pucci, B., De Falco, M., Rossi, M., and Pisani, F.M. (2004). A CDC6-like factor from the archaea Sulfolobus solfataricus promotes binding of the mini-chromosome maintenance complex to DNA. The Journal of biological chemistry *279*, 43008-43012.

de la Torre, J.R., Walker, C.B., Ingalls, A.E., Konneke, M., and Stahl, D.A. (2008). Cultivation of a thermophilic ammonia oxidizing archaeon synthesizing crenarchaeol. Environmental microbiology *10*, 810-818.

DeLong, E.F. (1992). Archaea in coastal marine environments. Proceedings of the National Academy of Sciences of the United States of America *89*, 5685-5689.

DeLong, E.F. (2000). Resolving a methane mystery. Nature *407*, 577, 579.

DeLong, E.F., Wu, K.Y., Prezelin, B.B., and Jovine, R.V. (1994). High abundance of Archaea in Antarctic marine picoplankton. Nature *371*, 695-697.

Delsuc, F., Brinkmann, H., and Philippe, H. (2005). Phylogenomics and the reconstruction of the tree of life. Nat Rev Genet *6*, 361-375.

Deppenmeier, U., Johann, A., Hartsch, T., Merkl, R., Schmitz, R.A., Martinez-Arias, R., Henne, A., Wiezer, A., Baumer, S., Jacobi, C*., et al.* (2002). The genome of Methanosarcina mazei: evidence for lateral gene transfer between bacteria and archaea. Journal of molecular microbiology and biotechnology *4*, 453-461.

Dervyn, E., Suski, C., Daniel, R., Bruand, C., Chapuis, J., Errington, J., Janniere, L., and Ehrlich, S.D. (2001). Two essential DNA polymerases at the bacterial replication fork. Science (New York, N.Y *294*, 1716-1719.

Di Giulio, M. (2006). Nanoarchaeum equitans is a living fossil. Journal of theoretical biology *242*, 257-260.

Dominguez-Sola, D., Ying, C.Y., Grandori, C., Ruggiero, L., Chen, B., Li, M., Galloway, D.A., Gu, W., Gautier, J., and Dalla-Favera, R. (2007). Non-transcriptional control of DNA replication by c-Myc. Nature *448*, 445-451.

Doolittle, W.F. (1999). Phylogenetic classification and the universal tree. Science (New York, N.Y *284*, 2124-2129.

Doolittle, W.F. (2000). Uprooting the tree of life. Scientific American *282*, 90-95.

Doolittle, W.F., and Bapteste, E. (2007). Pattern pluralism and the Tree of Life hypothesis. Proceedings of the National Academy of Sciences of the United States of America *104*, 2043-2049.

Doolittle, W.F., and Papke, R.T. (2006). Genomics and the bacterial species problem. Genome biology *7*, 116.

Du, Y.C., and Stillman, B. (2002). Yph1p, an ORC-interacting protein: potential links between cell proliferation control, DNA replication, and ribosome biogenesis. Cell *109*, 835-848.

Dueber, E.L., Corn, J.E., Bell, S.D., and Berger, J.M. (2007). Replication origin recognition and deformation by a heterodimeric archaeal Orc1 complex. Science (New York, N.Y *317*, 1210-1213.

Dutilh, B.E., Snel, B., Ettema, T.J., and Huynen, M.A. (2008). Signature genes as a phylogenomic tool. Molecular biology and evolution *25*, 1659-1667.

Eckburg, P.B., Lepp, P.W., and Relman, D.A. (2003). Archaea and their potential role in human disease. Infection and immunity *71*, 591-596.

Edgell, D.R., and Doolittle, W.F. (1997). Archaea and the origin(s) of DNA replication proteins. Cell *89*, 995-998.

Edgell, D.R., Klenk, H.P., and Doolittle, W.F. (1997). Gene duplications in evolution of archaeal family B DNA polymerases. J Bacteriol *179*, 2632-2640.

Elkins, J.G., Podar, M., Graham, D.E., Makarova, K.S., Wolf, Y., Randau, L., Hedlund, B.P., Brochier-Armanet, C., Kunin, V., Anderson, I.*, et al.* (2008). A korarchaeal genome reveals insights into the evolution of the Archaea. Proceedings of the National Academy of Sciences of the United States of America *105*, 8102-8107.

Endo, Y., and Sawasaki, T. (2006). Cell-free expression systems for eukaryotic protein production. Current opinion in biotechnology *17*, 373-380.

Endoh, T., Kanai, T., and Imanaka, T. (2008). Effective approaches for the production of heterologous proteins using the Thermococcus kodakaraensis-based translation system. Journal of biotechnology *133*, 177-182.

Enright, A.J., Iliopoulos, I., Kyrpides, N.C., and Ouzounis, C.A. (1999). Protein interaction maps for complete genomes based on gene fusion events. Nature *402*, 86-90.

Erkel, C., Kube, M., Reinhardt, R., and Liesack, W. (2006). Genome of Rice Cluster I archaea--the key methane producers in the rice rhizosphere. Science (New York, N.Y *313*, 370-372.

Ettema, T.J., de Vos, W.M., and van der Oost, J. (2005). Discovering novel biology by in silico archaeology. Nature reviews *3*, 859-869.

Evguenieva-Hackenberg, E., Walter, P., Hochleitner, E., Lottspeich, F., and Klug, G. (2003). An exosome-like complex in Sulfolobus solfataricus. EMBO Rep *4*, 889-893.

Falb, M., Pfeiffer, F., Palm, P., Rodewald, K., Hickmann, V., Tittor, J., and Oesterhelt, D. (2005). Living with two extremes: conclusions from the genome sequence of Natronomonas pharaonis. Genome research *15*, 1336-1343.

Falkowski, P.G., Fenchel, T., and Delong, E.F. (2008). The microbial engines that drive Earth's biogeochemical cycles. Science (New York, N.Y *320*, 1034-1039.

Farhoud, M.H., Wessels, H.J., Steenbakkers, P.J., Mattijssen, S., Wevers, R.A., van Engelen, B.G., Jetten, M.S., Smeitink, J.A., van den Heuvel, L.P., and Keltjens, J.T. (2005). Protein complexes in the archaeon Methanothermobacter thermautotrophicus analyzed by blue native/SDS-PAGE and mass spectrometry. Mol Cell Proteomics *4*, 1653-1663.

Ferry, J.G., and Kastead, K.A. (2007). Methanogenesis. In Archaea: molecular and cellular biology, R. Cavicchioli, ed. (Washington, DC: ASM Press), pp. 288-314.

Fitz-Gibbon, S.T., Ladner, H., Kim, U.J., Stetter, K.O., Simon, M.I., and Miller, J.H. (2002). Genome sequence of the hyperthermophilic crenarchaeon Pyrobaculum aerophilum. Proceedings of the National Academy of Sciences of the United States of America *99*, 984-989.

Forterre, P. (1992). Neutral terms. Nature *355*, 305-305.

Forterre, P. (1995). Thermoreduction, a hypothesis for the origin of prokaryotes. Comptes rendus de l'Academie des sciences *318*, 415-422.

Forterre, P., Bouthier De La Tour, C., Philippe, H., and Duguet, M. (2000). Reverse gyrase from hyperthermophiles: probable transfer of a thermoadaptation trait from archaea to bacteria. Trends Genet *16*, 152-154.

Forterre, P., Brochier, C., and Philippe, H. (2002). Evolution of the Archaea. Theoretical population biology *61*, 409-422.

Forterre, P., Elie, C., and Kohiyama, M. (1984). Aphidicolin inhibits growth and DNA synthesis in halophilic archaebacteria. J Bacteriol *159*, 800-802.

Forterre, P., Gribaldo, S., and Brochier-Armanet, C. (2007a). Natural history of the archaeal domain. In Archaea: evolution, physiology, and molecular biology, R.A. Garrett, and H.P. Klenk, eds. (Malden, MA: Blackwell Pub.), pp. 17-28.

Forterre, P., Gribaldo, S., Gadelle, D., and Serre, M.C. (2007b). Origin and evolution of DNA topoisomerases. Biochimie *89*, 427-446.

Fox, G.E., Magrum, L.J., Balch, W.E., Wolfe, R.S., and Woese, C.R. (1977). Classification of Methanogenic Bacteria by 16S Ribosomal RNA Characterization. Proceedings of the National Academy of Sciences *74*, 4537-4541.

Fricke, W.F., Seedorf, H., Henne, A., Kruer, M., Liesegang, H., Hedderich, R., Gottschalk, G., and Thauer, R.K. (2006). The genome sequence of Methanosphaera stadtmanae reveals why this human intestinal archaeon is restricted to methanol and H2 for methane formation and ATP synthesis. J Bacteriol *188*, 642-658.

Fuhrman, J.A., McCallum, K., and Davis, A.A. (1992). Novel major archaebacterial group from marine plankton. Nature *356*, 148-149.

Fujikane, R., Shinagawa, H., and Ishino, Y. (2006). The archaeal Hjm helicase has recQ-like functions, and may be involved in repair of stalled replication fork. Genes Cells *11*, 99-110.

Fukui, T., Atomi, H., Kanai, T., Matsumi, R., Fujiwara, S., and Imanaka, T. (2005). Complete genome sequence of the hyperthermophilic archaeon Thermococcus kodakaraensis KOD1 and comparison with Pyrococcus genomes. Genome research *15*, 352-363.

Fukushima, S., Itaya, M., Kato, H., Ogasawara, N., and Yoshikawa, H. (2007). Reassessment of the in vivo functions of DNA polymerase I and RNase H in bacterial cell growth. J Bacteriol *189*, 8575-8583.

Fuller, R.S., and Kornberg, A. (1983). Purified dnaA protein in initiation of replication at the Escherichia coli chromosomal origin of replication. Proceedings of the National Academy of Sciences of the United States of America *80*, 5817-5821.

Futterer, O., Angelov, A., Liesegang, H., Gottschalk, G., Schleper, C., Schepers, B., Dock, C., Antranikian, G., and Liebl, W. (2004). Genome sequence of Picrophilus torridus and its implications for life around pH 0. Proceedings of the National Academy of Sciences of the United States of America *101*, 9091-9096.

Gabaldon, T., and Huynen, M.A. (2004). Prediction of protein function and pathways in the genome era. Cell Mol Life Sci *61*, 930-944.

Gadelle, D., Filee, J., Buhler, C., and Forterre, P. (2003). Phylogenomics of type II DNA topoisomerases. Bioessays *25*, 232-242.

Galagan, J.E., Nusbaum, C., Roy, A., Endrizzi, M.G., Macdonald, P., FitzHugh, W., Calvo, S., Engels, R., Smirnov, S., Atnoor, D.*, et al.* (2002). The genome of M. acetivorans reveals extensive metabolic and physiological diversity. Genome research *12*, 532-542.

Galperin, M.Y., Aravind, L., and Koonin, E.V. (2000). Aldolases of the DhnA family: a possible solution to the problem of pentose and hexose biosynthesis in archaea. FEMS microbiology letters *183*, 259-264.

Galperin, M.Y., and Koonin, E.V. (2000). Who's your neighbor? New computational approaches for functional genomics. Nature biotechnology *18*, 609-613.

Gambus, A., Jones, R.C., Sanchez-Diaz, A., Kanemaki, M., van Deursen, F., Edmondson, R.D., and Labib, K. (2006). GINS maintains association of Cdc45 with MCM in replisome progression complexes at eukaryotic DNA replication forks. Nature cell biology *8*, 358-366.

Gao, B., and Gupta, R.S. (2007). Phylogenomic analysis of proteins that are distinctive of Archaea and its main subgroups and the origin of methanogenesis. BMC genomics *8*, 86.

Gaudier, M., Schuwirth, B.S., Westcott, S.L., and Wigley, D.B. (2007). Structural basis of DNA replication origin recognition by an ORC protein. Science (New York, N.Y *317*, 1213-1216.

Georgescu, R.E., and O'Donnell, M. (2007). STRUCTURAL BIOLOGY: Getting DNA to Unwind. Science (New York, N.Y *317*, 1181-1182.

Glansdorff, N. (1999). On the origin of operons and their possible role in evolution toward thermophily. Journal of molecular evolution *49*, 432-438.

Glansdorff, N. (2000). About the last common ancestor, the universal life-tree and lateral gene transfer: a reappraisal. Molecular microbiology *38*, 177-185.

Glansdorff, N., Xu, Y., and Labedan, B. (2008). The Last Universal Common Ancestor: emergence, constitution and genetic legacy of an elusive forerunner. Biology direct *3*, 29.

Glover, B.P., and McHenry, C.S. (2001). The DNA polymerase III holoenzyme: an asymmetric dimeric replicative complex with leading and lagging strand polymerases. Cell *105*, 925-934.

Goldenfeld, N., and Woese, C. (2007). Biology's next revolution. Nature *445*, 369-369.

Gonzalez-Montalban, N., Garcia-Fruitos, E., and Villaverde, A. (2007). Recombinant protein solubility - does more mean better? Nature biotechnology *25*, 718-720.

Graham, D.E., Overbeek, R., Olsen, G.J., and Woese, C.R. (2000). An archaeal genomic signature. Proceedings of the National Academy of Sciences of the United States of America *97*, 3304-3308.

Grainge, I., Gaudier, M., Schuwirth, B.S., Westcott, S.L., Sandall, J., Atanassova, N., and Wigley, D.B. (2006). Biochemical analysis of a DNA replication origin in the archaeon Aeropyrum pernix. Journal of molecular biology *363*, 355-369.

Grainge, I., Scaife, S., and Wigley, D.B. (2003). Biochemical analysis of components of the pre-replication complex of Archaeoglobus fulgidus. Nucleic Acids Res *31*, 4888-4898.

Gribaldo, S., and Brochier-Armanet, C. (2006). The origin and evolution of Archaea: a state of the art. Philosophical transactions of the Royal Society of London *361*, 1007-1022.

Gruber, C., Legat, A., Pfaffenhuemer, M., Radax, C., Weidler, G., Busse, H.J., and Stan-Lotter, H. (2004). Halobacterium noricense sp. nov., an archaeal isolate from a bore core of an alpine Permian salt deposit, classification of Halobacterium sp. NRC-1 as a strain of H. salinarum and emended description of H. salinarum. Extremophiles *8*, 431-439.

Haeckel, E.H.P.A. (1866). Generelle Morphologie der Organismen: Allgemeine Grundzüge der Organischen Formen-Wissenschaft, Mechanisch Begründet durch die von Charles Darwin Reformirte Descendenz–Theorie (Berlin: G. Reimer).

Hall, T.A. (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symposium Series *41*, 95-98.

Hallam, S.J., Konstantinidis, K.T., Putnam, N., Schleper, C., Watanabe, Y., Sugahara, J., Preston, C., de la Torre, J., Richardson, P.M., and DeLong, E.F. (2006). Genomic analysis of the uncultivated marine crenarchaeote Cenarchaeum symbiosum. Proceedings of the National Academy of Sciences of the United States of America *103*, 18296-18301.

Hallam, S.J., Putnam, N., Preston, C.M., Detter, J.C., Rokhsar, D., Richardson, P.M., and DeLong, E.F. (2004). Reverse methanogenesis: testing the hypothesis with environmental genomics. Science (New York, N.Y *305*, 1457-1462.

Hatzenpichler, R., Lebedeva, E.V., Spieck, E., Stoecker, K., Richter, A., Daims, H., and Wagner, M. (2008). A moderately thermophilic ammonia-oxidizing crenarchaeote from a hot spring. Proceedings of the National Academy of Sciences of the United States of America *105*, 2134-2139.

Haugland, G.T., Shin, J.H., Birkeland, N.K., and Kelman, Z. (2006). Stimulation of MCM helicase activity by a Cdc6 protein in the archaeon Thermoplasma acidophilum. Nucleic Acids Res *34*, 6337-6344.

Hendrickson, E.L., Kaul, R., Zhou, Y., Bovee, D., Chapman, P., Chung, J., Conway de Macario, E., Dodsworth, J.A., Gillett, W., Graham, D.E., *et al.* (2004). Complete genome sequence of the genetically tractable hydrogenotrophic methanogen Methanococcus maripaludis. J Bacteriol *186*, 6956-6969.

Henneke, G., Flament, D., Hubscher, U., Querellou, J., and Raffin, J.P. (2005). The hyperthermophilic euryarchaeota Pyrococcus abyssi likely requires the two DNA polymerases D and B for DNA replication. Journal of molecular biology *350*, 53-64.

Hershberg, R., Yeger-Lotem, E., and Margalit, H. (2005). Chromosomal organization is shaped by the transcription regulatory network. Trends Genet *21*, 138-142.

Hershey, A.D., and Chase, M. (1952). INDEPENDENT FUNCTIONS OF VIRAL PROTEIN AND NUCLEIC ACID IN GROWTH OF BACTERIOPHAGE. J. Gen. Physiol. *36*, 39-56.

Hethke, C., Bergerat, A., Hausner, W., Forterre, P., and Thomm, M. (1999). Cell-free transcription at 95 degrees: thermostability of transcriptional components and DNA topology requirements of Pyrococcus transcription. Genetics *152*, 1325-1333.

Hinrichs, K.U., Hayes, J.M., Sylva, S.P., Brewer, P.G., and DeLong, E.F. (1999). Methane-consuming archaebacteria in marine sediments. Nature *398*, 802-805.

Houchens, C.R., Perreault, A., Bachand, F., and Kelly, T.J. (2008). Schizosaccharomyces pombe Noc3 is essential for ribosome biogenesis and cell division but not DNA replication. Eukaryotic cell *7*, 1433-1440.

Huber, H., Hohn, M.J., Rachel, R., Fuchs, T., Wimmer, V.C., and Stetter, K.O. (2002). A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont. Nature *417*, 63-67.

Huynen, M., Snel, B., Lathe, W., and Bork, P. (2000). Exploitation of gene context. Current opinion in structural biology *10*, 366-370.

Hwang, D.S., Crooke, E., and Kornberg, A. (1990). Aggregated dnaA protein is dissociated and activated for DNA replication by phospholipase or dnaK protein. The Journal of biological chemistry *265*, 19244-19248.

Imamura, M., Uemori, T., Kato, I., and Ishino, Y. (1995). A non-alpha-like DNA polymerase from the hyperthermophilic archaeon Pyrococcus furiosus. Biological & pharmaceutical bulletin *18*, 1647-1652.

Indiani, C., and O'Donnell, M. (2006). The replication clamp-loading machine at work in the three domains of life. Nature reviews *7*, 751-761.

Ishino, Y., Komori, K., Cann, I.K., and Koga, Y. (1998). A novel DNA polymerase family found in Archaea. J Bacteriol *180*, 2232-2236.

Iyer, L.M., and Aravind, L. (2006). The Evolutionary History of Proteins Involved in Pre-replication Complex Assembly. In DNA replication and human disease, M.L. DePamphilis, ed. (Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory Press), pp. 751-757.

Jain, R., Rivera, M.C., and Lake, J.A. (1999). Horizontal gene transfer among genomes: the complexity hypothesis. Proceedings of the National Academy of Sciences of the United States of America *96*, 3801-3806.

Janniere, L., Canceill, D., Suski, C., Kanga, S., Dalmais, B., Lestini, R., Monnier, A.F., Chapuis, J., Bolotin, A., Titok, M.*, et al.* (2007). Genetic evidence for a link between glycolysis and DNA replication. PLoS ONE *2*, e447.

Jiang, W., McDonald, D., Hope, T.J., and Hunter, T. (1999). Mammalian Cdc7-Dbf4 protein kinase complex is essential for initiation of DNA replication. The EMBO journal *18*, 5703-5713.

Kamada, K., Kubota, Y., Arata, T., Shindo, Y., and Hanaoka, F. (2007). Structure of the human GINS complex and its assembly and functional interface in replication initiation. Nature structural & molecular biology *14*, 388-396.

Kanemaki, M., Sanchez-Diaz, A., Gambus, A., and Labib, K. (2003). Functional proteomic identification of DNA replication proteins by induced proteolysis in vivo. Nature *423*, 720-724.

Karner, M.B., DeLong, E.F., and Karl, D.M. (2001). Archaeal dominance in the mesopelagic zone of the Pacific Ocean. Nature *409*, 507-510.

Kasiviswanathan, R., Shin, J.H., and Kelman, Z. (2005). Interactions between the archaeal Cdc6 and MCM proteins modulate their biochemical properties. Nucleic Acids Res *33*, 4940-4950.

Katscher, F. (2004). The history of the terms prokaryotes and eukaryotes. Protist *155*, 257-263.

Kawarabayasi, Y., Hino, Y., Horikawa, H., Jin-no, K., Takahashi, M., Sekine, M., Baba, S., Ankai, A., Kosugi, H., Hosoyama, A.*, et al.* (2001). Complete genome sequence of an aerobic thermoacidophilic crenarchaeon, Sulfolobus tokodaii strain7. DNA Res *8*, 123-140.

Kawarabayasi, Y., Hino, Y., Horikawa, H., Yamazaki, S., Haikawa, Y., Jin-no, K., Takahashi, M., Sekine, M., Baba, S., Ankai, A.*, et al.* (1999). Complete genome sequence of an aerobic hyper-thermophilic crenarchaeon, Aeropyrum pernix K1. DNA Res *6*, 83-101, 145-152.

Kawarabayasi, Y., Sawada, M., Horikawa, H., Haikawa, Y., Hino, Y., Yamamoto, S., Sekine, M., Baba, S., Kosugi, H., Hosoyama, A.*, et al.* (1998a). Complete sequence and gene

organization of the genome of a hyper-thermophilic archaebacterium, Pyrococcus horikoshii OT3. DNA Res *5*, 55-76.

Kawarabayasi, Y., Sawada, M., Horikawa, H., Haikawa, Y., Hino, Y., Yamamoto, S., Sekine, M., Baba, S., Kosugi, H., Hosoyama, A.*, et al.* (1998b). Complete sequence and gene organization of the genome of a hyper-thermophilic archaebacterium, Pyrococcus horikoshii OT3 (supplement). DNA Res *5*, 147-155.

Kawashima, T., Amano, N., Koike, H., Makino, S., Higuchi, S., Kawashima-Ohya, Y., Watanabe, K., Yamazaki, M., Kanehori, K., Kawamoto, T.*, et al.* (2000). Archaeal adaptation to higher temperatures revealed by genomic sequence of Thermoplasma volcanium. Proceedings of the National Academy of Sciences of the United States of America *97*, 14257-14262.

Kearsey, S.E., and Cotterill, S. (2003). Enigmatic variations: divergent modes of regulating eukaryotic DNA replication. Molecular cell *12*, 1067-1075.

Keck, J.L., Roche, D.D., Lynch, A.S., and Berger, J.M. (2000). Structure of the RNA polymerase domain of E. coli primase. Science (New York, N.Y *287*, 2482-2486.

Kelly, T.J., and Stillman, B. (2006). Duplication of DNA in Eukaryotic Cells. In DNA replication and human disease, M.L. DePamphilis, ed. (Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory Press), pp. 1-29.

Kelman, Z., Lee, J.K., and Hurwitz, J. (1999a). The single minichromosome maintenance protein of Methanobacterium thermoautotrophicum DeltaH contains DNA helicase activity. Proceedings of the National Academy of Sciences of the United States of America *96*, 14783-14788.

Kelman, Z., Pietrokovski, S., and Hurwitz, J. (1999b). Isolation and characterization of a split B-type DNA polymerase from the archaeon Methanobacterium thermoautotrophicum deltaH. The Journal of biological chemistry *274*, 28751-28761.

Kiyonari, S., Takayama, K., Nishida, H., and Ishino, Y. (2006). Identification of a novel binding motif in Pyrococcus furiosus DNA ligase for the functional interaction with proliferating cell nuclear antigen. The Journal of biological chemistry *281*, 28023-28032.

Klenk, H.P., Clayton, R.A., Tomb, J.F., White, O., Nelson, K.E., Ketchum, K.A., Dodson, R.J., Gwinn, M., Hickey, E.K., Peterson, J.D.*, et al.* (1997). The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon Archaeoglobus fulgidus. Nature *390*, 364-370.

Koma, D., Sawai, T., Harayama, S., and Kino, K. (2006). Overexpression of the genes from thermophiles in Escherichia coli by high-temperature cultivation. Appl Microbiol Biotechnol *73*, 172-180.

Konneke, M., Bernhard, A.E., de la Torre, J.R., Walker, C.B., Waterbury, J.B., and Stahl, D.A. (2005). Isolation of an autotrophic ammonia-oxidizing marine archaeon. Nature *437*, 543-546.

Koonin, E.V., Mushegian, A.R., and Bork, P. (1996). Non-orthologous gene displacement. Trends Genet *12*, 334-336.

Koonin, E.V., Wolf, Y.I., and Aravind, L. (2001). Prediction of the archaeal exosome and its connections with the proteasome and the translation and transcription machineries by a comparative-genomic approach. Genome research *11*, 240-252.

Korbel, J.O., Jensen, L.J., von Mering, C., and Bork, P. (2004). Analysis of genomic context: prediction of functional associations from conserved bidirectionally transcribed gene pairs. Nature biotechnology *22*, 911-917.

Kornberg, A., and Baker, T.A. (1992). DNA replication, 2nd edn (New York: Freeman).

Kube, J., Brokamp, C., Machielsen, R., van der Oost, J., and Markl, H. (2006). Influence of temperature on the production of an archaeal thermoactive alcohol dehydrogenase from Pyrococcus furiosus with recombinant Escherichia coli. Extremophiles *10*, 221-227.

Kubota, Y., Takase, Y., Komori, Y., Hashimoto, Y., Arata, T., Kamimura, Y., Araki, H., and Takisawa, H. (2003). A novel ring-like complex of Xenopus proteins essential for the initiation of DNA replication. Genes & development *17*, 1141-1152.

Kurland, C.G., Canback, B., and Berg, O.G. (2003). Horizontal gene transfer: a critical view. Proceedings of the National Academy of Sciences of the United States of America *100*, 9658-9662.

Kurland, C.G., Canback, B., and Berg, O.G. (2007). The origins of modern proteomes. Biochimie *89*, 1454-1463.

Kwapisz, M., Beckouet, F., and Thuriaux, P. (2008). Early evolution of eukaryotic DNA-dependent RNA polymerases. Trends Genet *24*, 211-215.

Labib, K., and Gambus, A. (2007). A key role for the GINS complex at DNA replication forks. Trends Cell Biol.

Laemmli, U.K. (1970). Cleavage of structural proteins during the assembly of the head of bacteriophage T4. Nature *227*, 680-685.

Lake, J.A. (1988). Origin of the eukaryotic nucleus determined by rate-invariant analysis of rRNA sequences. Nature *331*, 184-186.

Langston, L.D., and O'Donnell, M. (2006). DNA replication: keep moving and don't mind the gap. Molecular cell *23*, 155-160.

Lao-Sirieix, S.H., and Bell, S.D. (2004). The heterodimeric primase of the hyperthermophilic archaeon Sulfolobus solfataricus possesses DNA and RNA primase, polymerase and 3'-terminal nucleotidyl transferase activities. Journal of molecular biology *344*, 1251-1263.

Lao-Sirieix, S.H., Marsh, V.L., and Bell, S.D. (2007). DNA Replication and Cell Cycle. In Archaea: molecular and cellular biology, R. Cavicchioli, ed. (Washington, DC: ASM Press), pp. 93-109.

Lao-Sirieix, S.H., Pellegrini, L., and Bell, S.D. (2005). The promiscuous primase. Trends Genet *21*, 568-572.

Le Breton, M., Henneke, G., Norais, C., Flament, D., Myllykallio, H., Querellou, J., and Raffin, J.P. (2007). The Heterodimeric Primase from the Euryarchaeon Pyrococcus abyssi: A Multifunctional Enzyme for Initiation and Repair? Journal of molecular biology.

Lecompte, O., Ripp, R., Thierry, J.C., Moras, D., and Poch, O. (2002). Comparative analysis of ribosomal proteins in complete genomes: an example of reductive evolution at the domain scale. Nucleic Acids Res *30*, 5382-5390.

Lee, J.R., Makise, M., Takenaka, H., Takahashi, N., Yamaguchi, Y., Tsuchiya, T., and Mizushima, T. (2002). Inhibitory effects of acidic phospholipids on the binding of origin-recognition complex to origin DNA. The Biochemical journal *362*, 395-399.

Leininger, S., Urich, T., Schloter, M., Schwark, L., Qi, J., Nicol, G.W., Prosser, J.I., Schuster, S.C., and Schleper, C. (2006). Archaea predominate among ammonia-oxidizing prokaryotes in soils. Nature *442*, 806-809.

Leipe, D.D., Aravind, L., and Koonin, E.V. (1999). Did DNA replication evolve twice independently? Nucleic Acids Res *27*, 3389-3401.

Leirmo, S., Harrison, C., Cayley, D.S., Burgess, R.R., and Record, M.T., Jr. (1987). Replacement of potassium chloride by potassium glutamate dramatically enhances protein-DNA interactions in vitro. Biochemistry *26*, 2095-2101.

Lepp, P.W., Brinig, M.M., Ouverney, C.C., Palm, K., Armitage, G.C., and Relman, D.A. (2004). Methanogenic Archaea and human periodontal disease. Proceedings of the National Academy of Sciences of the United States of America *101*, 6176-6181.

Lin, Y., Lin, L.J., Sriratana, P., Coleman, K., Ha, T., Spies, M., and Cann, I.K. (2008). Engineering of functional replication protein a homologs based on insights into the evolution of oligonucleotide/oligosaccharide-binding folds. J Bacteriol *190*, 5766-5780.

Liu, J., Smith, C.L., DeRyckere, D., DeAngelis, K., Martin, G.S., and Berger, J.M. (2000). Structure and function of Cdc6/Cdc18: implications for origin recognition and checkpoint control. Molecular cell *6*, 637-648.

Liu, L., Komori, K., Ishino, S., Bocquier, A.A., Cann, I.K., Kohda, D., and Ishino, Y. (2001). The archaeal DNA primase: biochemical characterization of the p41-p46 complex from Pyrococcus furiosus. The Journal of biological chemistry *276*, 45484-45490.

Lopez-Lopez, A., Benlloch, S., Bonfa, M., Rodriguez-Valera, F., and Mira, A. (2007). Intragenomic 16S rDNA divergence in Haloarcula marismortui is an adaptation to different temperatures. Journal of molecular evolution *65*, 687-696.

Lopez, P., Philippe, H., Myllykallio, H., and Forterre, P. (1999). Identification of putative chromosomal origins of replication in Archaea. Molecular microbiology *32*, 883-886.

Lundgren, M., Andersson, A., Chen, L., Nilsson, P., and Bernander, R. (2004). Three replication origins in Sulfolobus species: synchronous initiation of chromosome replication and asynchronous termination. Proceedings of the National Academy of Sciences of the United States of America *101*, 7046-7051.

Lundgren, M., and Bernander, R. (2005). Archaeal cell cycle progress. Current opinion in microbiology *8*, 662-668.

Lundgren, M., and Bernander, R. (2007). Genome-wide transcription map of an archaeal cell cycle. Proceedings of the National Academy of Sciences of the United States of America *104*, 2939-2944.

Luo, X., Schwarz-Linek, U., Botting, C.H., Hensel, R., Siebers, B., and White, M.F. (2007). CC1, a novel crenarchaeal DNA binding protein. J Bacteriol *189*, 403-409.

Lutzmann, M., and Mechali, M. (2008). MCM9 binds Cdt1 and is required for the assembly of prereplication complexes. Molecular cell *31*, 190-200.

Machida, Y.J., Hamlin, J.L., and Dutta, A. (2005). Right place, right time, and only once: replication initiation in metazoans. Cell *123*, 13-24.

MacNeill, S.A. (2001). Understanding the enzymology of archaeal DNA replication: progress in form and function. Molecular microbiology *40*, 520-529.

Maeder, D.L., Anderson, I., Brettin, T.S., Bruce, D.C., Gilna, P., Han, C.S., Lapidus, A., Metcalf, W.W., Saunders, E., Tapia, R., and Sowers, K.R. (2006). The Methanosarcina barkeri genome: comparative analysis with Methanosarcina acetivorans and Methanosarcina mazei reveals extensive rearrangement within methanosarcinal genomes. J Bacteriol *188*, 7922-7931.

Maiorano, D., Cuvier, O., Danis, E., and Mechali, M. (2005). MCM8 is an MCM2-7-related protein that functions as a DNA helicase during replication elongation and not initiation. Cell *120*, 315-328.

Majernik, A.I., and Chong, J.P. (2008). A conserved mechanism for replication origin recognition and binding in archaea. The Biochemical journal *409*, 511-518.

Makarova, K.S., Aravind, L., Galperin, M.Y., Grishin, N.V., Tatusov, R.L., Wolf, Y.I., and Koonin, E.V. (1999). Comparative genomics of the Archaea (Euryarchaeota): evolution of conserved protein families, the stable core, and the variable shell. Genome research *9*, 608-628.

Makarova, K.S., and Koonin, E.V. (2003a). Comparative genomics of Archaea: how much have we learned in six years, and what's next? Genome biology *4*, 115.

Makarova, K.S., and Koonin, E.V. (2003b). Filling a gap in the central metabolism of archaea: prediction of a novel aconitase by comparative-genomic analysis. FEMS microbiology letters *227*, 17-23.

Makarova, K.S., and Koonin, E.V. (2005). Evolutionary and functional genomics of the Archaea. Current opinion in microbiology *8*, 586-594.

Makarova, K.S., Sorokin, A.V., Novichkov, P.S., Wolf, Y.I., and Koonin, E.V. (2007). Clusters of orthologous genes for 41 archaeal genomes and implications for evolutionary genomics of archaea. Biology direct *2*, 33.

Makarova, K.S., Wolf, Y.I., Mekhedov, S.L., Mirkin, B.G., and Koonin, E.V. (2005). Ancestral paralogs and pseudoparalogs and their role in the emergence of the eukaryotic cell. Nucleic acids research *33*, 4626-4638.

Makiniemi, M., Pospiech, H., Kilpelainen, S., Jokela, M., Vihinen, M., and Syvaoja, J.E. (1999). A novel family of DNA-polymerase-associated B subunits. Trends in biochemical sciences *24*, 14-16.

Marcotte, E.M. (2000). Computational genetics: finding protein function by nonhomology methods. Current opinion in structural biology *10*, 359-365.

Marcotte, E.M., Pellegrini, M., Ng, H.L., Rice, D.W., Yeates, T.O., and Eisenberg, D. (1999). Detecting protein function and protein-protein interactions from genome sequences. Science (New York, N.Y *285*, 751-753.

Marinsek, N., Barry, E.R., Makarova, K.S., Dionne, I., Koonin, E.V., and Bell, S.D. (2006). GINS, a central nexus in the archaeal DNA replication fork. EMBO Rep *7*, 539-545.

Martin, W. (2004). Pathogenic archaebacteria: do they not exist because archaebacteria use different vitamins? Bioessays *26*, 592-593; author reply 593.

Martin, W. (2005). Archaebacteria (Archaea) and the origin of the eukaryotic nucleus. Current opinion in microbiology *8*, 630-637.

Martin, W., Baross, J., Kelley, D., and Russell, M.J. (2008). Hydrothermal vents and the origin of life. Nature reviews.

Martin, W., and Koonin, E.V. (2006). A positive definition of prokaryotes. Nature *442*, 868-868.

Matsunaga, F., Forterre, P., Ishino, Y., and Myllykallio, H. (2001). In vivo interactions of archaeal Cdc6/Orc1 and minichromosome maintenance proteins with the replication origin. Proceedings of the National Academy of Sciences of the United States of America *98*, 11152-11157.

Matsunaga, F., Glatigny, A., Mucchielli-Giorgi, M.H., Agier, N., Delacroix, H., Marisa, L., Durosay, P., Ishino, Y., Aggerbeck, L., and Forterre, P. (2007). Genomewide and biochemical analyses of DNA-binding activity of Cdc6/Orc1 and Mcm proteins in Pyrococcus sp. Nucleic Acids Res *35*, 3214-3222.

Mayr, E. (1998). Two empires or three? Proceedings of the National Academy of Sciences of the United States of America *95*, 9720-9723.

McGeoch, A.T., and Bell, S.D. (2008). Extra-chromosomal elements and the evolution of cellular DNA replication machineries. Nature reviews *9*, 569-574.

McGeoch, A.T., Trakselis, M.A., Laskey, R.A., and Bell, S.D. (2005). Organization of the archaeal MCM complex on DNA and implications for the helicase mechanism. Nature structural & molecular biology *12*, 756-762.

McInerney, J.O., Cotton, J.A., and Pisani, D. (2008). The prokaryotic tree of life: past, present...and future? Trends Ecol Evol *23*, 276-281.

Meselson, M., and Stahl, F.W. (1958). The Replication of DNA in Escherichia Coli. Proceedings of the National Academy of Sciences of the United States of America *44*, 671-682.

Meslet-Cladiere, L., Norais, C., Kuhn, J., Briffotaux, J., Sloostra, J.W., Ferrari, E., Hubscher, U., Flament, D., and Myllykallio, H. (2007). A novel proteomic approach identifies new interaction partners for proliferating cell nuclear antigen. Journal of molecular biology *372*, 1137-1148.

Morrison, H.G., McArthur, A.G., Gillin, F.D., Aley, S.B., Adam, R.D., Olsen, G.J., Best, A.A., Cande, W.Z., Chen, F., Cipriano, M.J.*, et al.* (2007). Genomic Minimalism in the Early Diverging Intestinal Parasite Giardia lamblia. Science (New York, N.Y *317*, 1921-1926.

Mott, M.L., and Berger, J.M. (2007). DNA replication initiation: mechanisms and regulation in bacteria. Nature reviews *5*, 343-354.

Moyer, S.E., Lewis, P.W., and Botchan, M.R. (2006). Isolation of the Cdc45/Mcm2-7/GINS (CMG) complex, a candidate for the eukaryotic DNA replication fork helicase. Proceedings of the National Academy of Sciences of the United States of America *103*, 10236-10241.

Murzin, A.G. (1993). OB(oligonucleotide/oligosaccharide binding)-fold: common structural and functional solution for non-homologous sequences. The EMBO journal *12*, 861-867.

Myllykallio, H., Lipowski, G., Leduc, D., Filee, J., Forterre, P., and Liebl, U. (2002). An alternative flavin-dependent mechanism for thymidylate synthesis. Science (New York, N.Y *297*, 105-107.

Myllykallio, H., Lopez, P., Lopez-Garcia, P., Heilig, R., Saurin, W., Zivanovic, Y., Philippe, H., and Forterre, P. (2000). Bacterial mode of replication with eukaryotic-like machinery in a hyperthermophilic archaeon. Science (New York, N.Y *288*, 2212-2215.

Natale, D.A., Shankavaram, U.T., Galperin, M.Y., Wolf, Y.I., Aravind, L., and Koonin, E.V. (2000). Towards understanding the first genome sequence of a crenarchaeon by genome

annotation using clusters of orthologous groups of proteins (COGs). Genome biology *1*, RESEARCH0009.

Neuwald, A.F., Aravind, L., Spouge, J.L., and Koonin, E.V. (1999). AAA+: A class of chaperone-like ATPases associated with the assembly, operation, and disassembly of protein complexes. Genome research *9*, 27-43.

Ng, W.V., Kennedy, S.P., Mahairas, G.G., Berquist, B., Pan, M., Shukla, H.D., Lasky, S.R., Baliga, N.S., Thorsson, V., Sbrogna, J.*, et al.* (2000). Genome sequence of Halobacterium species NRC-1. Proceedings of the National Academy of Sciences of the United States of America *97*, 12176-12181.

Nick McElhinny, S.A., Gordenin, D.A., Stith, C.M., Burgers, P.M., and Kunkel, T.A. (2008). Division of labor at the eukaryotic replication fork. Molecular cell *30*, 137-144.

Norais, C., Hawkins, M., Hartman, A.L., Eisen, J.A., Myllykallio, H., and Allers, T. (2007). Genetic and physical mapping of DNA replication origins in Haloferax volcanii. PLoS genetics *3*, e77.

Ohtani, N., Yanagawa, H., Tomita, M., and Itaya, M. (2004a). Cleavage of double-stranded RNA by RNase HI from a thermoacidophilic archaeon, Sulfolobus tokodaii 7. Nucleic Acids Res *32*, 5809-5819.

Ohtani, N., Yanagawa, H., Tomita, M., and Itaya, M. (2004b). Identification of the first archaeal Type 1 RNase H gene from Halobacterium sp. NRC-1: archaeal RNase HI can cleave an RNA-DNA junction. The Biochemical journal *381*, 795-802.

Okazaki, R., Okazaki, T., Sakabe, K., Sugimoto, K., and Sugino, A. (1968). Mechanism of DNA chain growth. I. Possible discontinuity and unusual secondary structure of newly synthesized chains. Proceedings of the National Academy of Sciences of the United States of America *59*, 598-605.

Olsen, G.J., and Woese, C.R. (1996). Lessons from an Archaeal genome: what are we learning from Methanococcus jannaschii? Trends Genet *12*, 377-379.

Orphan, V.J., House, C.H., Hinrichs, K.U., McKeegan, K.D., and DeLong, E.F. (2001). Methane-consuming archaea revealed by directly coupled isotopic and phylogenetic analysis. Science (New York, N.Y *293*, 484-487.

Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G.D., and Maltsev, N. (1999). The use of gene clusters to infer functional coupling. Proceedings of the National Academy of Sciences of the United States of America *96*, 2896-2901.

Pace, N.R. (1997). A molecular view of microbial diversity and the biosphere. Science (New York, N.Y *276*, 734-740.

Pace, N.R. (2006). Time for a change. Nature *441*, 289-289.

Pacek, M., Tutter, A.V., Kubota, Y., Takisawa, H., and Walter, J.C. (2006). Localization of MCM2-7, Cdc45, and GINS to the site of DNA unwinding during eukaryotic DNA replication. Molecular cell *21*, 581-587.

Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D., and Yeates, T.O. (1999). Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proceedings of the National Academy of Sciences of the United States of America *96*, 4285-4288.

Pereto, J., Lopez-Garcia, P., and Moreira, D. (2004). Ancestral lipid biosynthesis and early membrane evolution. Trends in biochemical sciences *29*, 469-477.

Perkins, G., and Diffley, J.F. (1998). Nucleotide-dependent prereplicative complex assembly by Cdc6p, a homolog of eukaryotic and prokaryotic clamp-loaders. Molecular cell *2*, 23-32.

Pfeiffer, F., Schuster, S.C., Broicher, A., Falb, M., Palm, P., Rodewald, K., Ruepp, A., Soppa, J., Tittor, J., and Oesterhelt, D. (2008a). Evolution in the laboratory: the genome of Halobacterium salinarum strain R1 compared to that of strain NRC-1. Genomics *91*, 335-346.

Pfeiffer, F., Schuster, S.C., Broicher, A., Falb, M., Palm, P., Rodewald, K., Ruepp, A., Soppa, J., Tittor, J., and Oesterhelt, D. (2008b). Genome sequences of Halobacterium salinarum: A reply. Genomics *91*, 553-554.

Pomerantz, R.T., and O'Donnell, M. (2007). Replisome mechanics: insights into a twin DNA polymerase machine. Trends in microbiology *15*, 156-164.

Poole, A., Jeffares, D., and Penny, D. (1999). Early evolution: prokaryotes, the new kids on the block. Bioessays *21*, 880-889.

Price, M.N., Huang, K.H., Arkin, A.P., and Alm, E.J. (2005). Operon formation is driven by co-regulation and not by horizontal gene transfer. Genome research *15*, 809-819.

Pursell, Z.F., Isoz, I., Lundstrom, E.-B., Johansson, E., and Kunkel, T.A. (2007). Yeast DNA Polymerase {epsilon} Participates in Leading-Strand DNA Replication. Science (New York, N.Y *317*, 127-130.

Raghoebarsing, A.A., Pol, A., van de Pas-Schoonen, K.T., Smolders, A.J., Ettwig, K.F., Rijpstra, W.I., Schouten, S., Damste, J.S., Op den Camp, H.J., Jetten, M.S., and Strous, M. (2006). A microbial consortium couples anaerobic methane oxidation to denitrification. Nature *440*, 918-921.

Randell, J.C., Bowers, J.L., Rodriguez, H.K., and Bell, S.P. (2006). Sequential ATP hydrolysis by Cdc6 and ORC directs loading of the Mcm2-7 helicase. Molecular cell *21*, 29-39.

Reeve, J.N. (1999). Archaebacteria then ... Archaes now (are there really no archaeal pathogens?). J Bacteriol *181*, 3613-3617.

Reysenbach, A.L., Liu, Y., Banta, A.B., Beveridge, T.J., Kirshtein, J.D., Schouten, S., Tivey, M.K., Von Damm, K.L., and Voytek, M.A. (2006). A ubiquitous thermoacidophilic archaeon from deep-sea hydrothermal vents. Nature *442*, 444-447.

Richard, D.J., Bolderson, E., Cubeddu, L., Wadsworth, R.I., Savage, K., Sharma, G.G., Nicolette, M.L., Tsvetanov, S., McIlwraith, M.J., Pandita, R.K.*, et al.* (2008). Single-stranded DNA-binding protein hSSB1 is critical for genomic stability. Nature *453*, 677-681.

Richey, B., Cayley, D.S., Mossing, M.C., Kolka, C., Anderson, C.F., Farrar, T.C., and Record, M.T., Jr. (1987). Variability of the intracellular ionic environment of Escherichia coli. Differences between in vitro and in vivo effects of ion concentrations on protein-DNA interactions and gene expression. The Journal of biological chemistry *262*, 7157-7164.

Ricke, R.M., and Bielinsky, A.K. (2004). Mcm10 regulates the stability and chromatin association of DNA polymerase-alpha. Molecular cell *16*, 173-185.

Robbins, J.B., McKinney, M.C., Guzman, C.E., Sriratana, B., Fitz-Gibbon, S., Ha, T., and Cann, I.K. (2005). The euryarchaeota, nature's medium for engineering of single-stranded DNA-binding proteins. The Journal of biological chemistry *280*, 15325-15339.

Robinson, N.P., and Bell, S.D. (2007). Extrachromosomal element capture and the evolution of multiple replication origins in archaeal chromosomes. Proceedings of the National Academy of Sciences of the United States of America *104*, 5806-5811.

Robinson, N.P., Blood, K.A., McCallum, S.A., Edwards, P.A., and Bell, S.D. (2007). Sister chromatid junctions in the hyperthermophilic archaeon Sulfolobus solfataricus. The EMBO journal *26*, 816-824.

Robinson, N.P., Dionne, I., Lundgren, M., Marsh, V.L., Bernander, R., and Bell, S.D. (2004). Identification of two origins of replication in the single chromosome of the archaeon Sulfolobus solfataricus. Cell *116*, 25-38.

Rogozin, I.B., Makarova, K.S., Pavlov, Y.I., and Koonin, E.V. (2008). A highly conserved family of inactivated archaeal B family DNA polymerases. Biology direct *3*, 32.

Ron, E.Z., and Davis, B.D. (1971). Growth rate of Escherichia coli at elevated temperatures: limitation by methionine. J Bacteriol *107*, 391-396.

Rouillon, C., Henneke, G., Flament, D., Querellou, J., and Raffin, J.P. (2007). DNA polymerase switching on homotrimeric PCNA at the replication fork of the euryarchaea Pyrococcus abyssi. Journal of molecular biology *369*, 343-355.

Ruepp, A., Graml, W., Santos-Martinez, M.L., Koretke, K.K., Volker, C., Mewes, H.W., Frishman, D., Stocker, S., Lupas, A.N., and Baumeister, W. (2000). The genome sequence of the thermoacidophilic scavenger Thermoplasma acidophilum. Nature *407*, 508-513.

Rydberg, B., and Game, J. (2002). Excision of misincorporated ribonucleotides in DNA by RNase H (type 2) and FEN-1 in cell-free extracts. Proceedings of the National Academy of Sciences of the United States of America *99*, 16654-16659.

Samuel, B.S., Hansen, E.E., Manchester, J.K., Coutinho, P.M., Henrissat, B., Fulton, R., Latreille, P., Kim, K., Wilson, R.K., and Gordon, J.I. (2007). Genomic and metabolic adaptations of Methanobrevibacter smithii to the human gut. Proceedings of the National Academy of Sciences of the United States of America *104*, 10643-10648.

Santangelo, T.J., Cubonova, L., and Reeve, J.N. (2008). Shuttle vector expression in Thermococcus kodakaraensis: contributions of cis elements to protein synthesis in a hyperthermophilic archaeon. Applied and environmental microbiology *74*, 3099-3104.

Sapp, J. (2005). The prokaryote-eukaryote dichotomy: meanings and mythology. Microbiol Mol Biol Rev *69*, 292-305.

Sapp, J. (2006). Two faces of the prokaryote concept. Int Microbiol *9*, 163-172.

Sartori, A.A., and Jiricny, J. (2003). Enzymology of base excision repair in the hyperthermophilic archaeon Pyrobaculum aerophilum. The Journal of biological chemistry *278*, 24563-24576.

Sato, A., Kanai, A., Itaya, M., and Tomita, M. (2003). Cooperative regulation for Okazaki fragment processing by RNase HII and FEN-1 purified from a hyperthermophilic archaeon, Pyrococcus furiosus. Biochemical and biophysical research communications *309*, 247-252.

Schleper, C., Jurgens, G., and Jonuscheit, M. (2005). Genomic studies of uncultivated archaea. Nature reviews *3*, 479-488.

Scholz, S., Sonnenbichler, J., Schafer, W., and Hensel, R. (1992). Di-myo-inositol-1,1'-phosphate: a new inositol phosphate isolated from Pyrococcus woesei. FEBS letters *306*, 239-242.

Sekimizu, K., Yung, B.Y., and Kornberg, A. (1988). The dnaA protein of Escherichia coli. Abundance, improved purification, and membrane binding. The Journal of biological chemistry *263*, 7136-7140.

She, Q., Singh, R.K., Confalonieri, F., Zivanovic, Y., Allard, G., Awayez, M.J., Chan-Weiher, C.C., Clausen, I.G., Curtis, B.A., De Moors, A.*, et al.* (2001). The complete genome of the crenarchaeon Sulfolobus solfataricus P2. Proceedings of the National Academy of Sciences of the United States of America *98*, 7835-7840.

Shen, Y., Musti, K., Hiramoto, M., Kikuchi, H., Kawarabayashi, Y., and Matsui, I. (2001). Invariant Asp-1122 and Asp-1124 are essential residues for polymerization catalysis of family D DNA polymerase from Pyrococcus horikoshii. The Journal of biological chemistry *276*, 27376-27383.

Sheu, Y.J., and Stillman, B. (2006). Cdc7-Dbf4 phosphorylates MCM proteins via a docking site-mediated mechanism to promote S phase progression. Molecular cell *24*, 101-113.

Shin, J.H., Grabowski, B., Kasiviswanathan, R., Bell, S.D., and Kelman, Z. (2003a). Regulation of minichromosome maintenance helicase activity by Cdc6. The Journal of biological chemistry *278*, 38059-38067.

Shin, J.H., Heo, G.Y., and Kelman, Z. (2008). The Methanothermobacter thermautotrophicus Cdc6-2 protein, the putative helicase loader, dissociates the minichromosome maintenance helicase. J Bacteriol *190*, 4091-4094.

Shin, J.H., Jiang, Y., Grabowski, B., Hurwitz, J., and Kelman, Z. (2003b). Substrate requirements for duplex DNA translocation by the eukaryal and archaeal minichromosome maintenance helicases. The Journal of biological chemistry *278*, 49053-49062.

Siebers, B., Brinkmann, H., Dorr, C., Tjaden, B., Lilie, H., van der Oost, J., and Verhees, C.H. (2001). Archaeal fructose-1,6-bisphosphate aldolases constitute a new family of archaeal type class I aldolase. The Journal of biological chemistry *276*, 28710-28718.

Singleton, M.R., Morales, R., Grainge, I., Cook, N., Isupov, M.N., and Wigley, D.B. (2004). Conformational changes induced by nucleotide binding in Cdc6/ORC from Aeropyrum pernix. Journal of molecular biology *343*, 547-557.

Sivaprasad, U., Dutta, A., and Bell, S.P. (2006). Assembly of Pre-replication Complexes. In DNA replication and human disease, M.L. DePamphilis, ed. (Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory Press), pp. 63-88.

Slesarev, A.I., Mezhevaya, K.V., Makarova, K.S., Polushin, N.N., Shcherbinina, O.V., Shakhova, V.V., Belova, G.I., Aravind, L., Natale, D.A., Rogozin, I.B.*, et al.* (2002). The complete genome of hyperthermophile Methanopyrus kandleri AV19 and monophyly of archaeal methanogens. Proceedings of the National Academy of Sciences of the United States of America *99*, 4644-4649.

Smith, D.R., Doucette-Stamm, L.A., Deloughery, C., Lee, H., Dubois, J., Aldredge, T., Bashirzadeh, R., Blakely, D., Cook, R., Gilbert, K.*, et al.* (1997). Complete genome sequence of Methanobacterium thermoautotrophicum deltaH: functional analysis and comparative genomics. J Bacteriol *179*, 7135-7155.

Sorensen, H.P., and Mortensen, K.K. (2005). Soluble expression of recombinant proteins in the cytoplasm of Escherichia coli. Microbial cell factories *4*, 1.

Stanier, R.Y., and Van Niel, C.B. (1962). The concept of a bacterium. Archiv fur Mikrobiologie *42*, 17-35.

Tada, S., Kundu, L.R., and Enomoto, T. (2008). Insight into initiator-DNA interactions: a lesson from the archaeal ORC. Bioessays *30*, 208-211.

Tahara, M., Ohsawa, A., Saito, S., and Kimura, M. (2004). In vitro phosphorylation of initiation factor 2 alpha (aIF2 alpha) from hyperthermophilic archaeon Pyrococcus horikoshii OT3. Journal of biochemistry *135*, 479-485.

Takayama, Y., Kamimura, Y., Okawa, M., Muramatsu, S., Sugino, A., and Araki, H. (2003). GINS, a novel multiprotein complex required for chromosomal DNA replication in budding yeast. Genes & development *17*, 1153-1165.

Tanaka, S., Umemori, T., Hirai, K., Muramatsu, S., Kamimura, Y., and Araki, H. (2007). CDK-dependent phosphorylation of Sld2 and Sld3 initiates DNA replication in budding yeast. Nature *445*, 328-332.

Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Kiryutin, B., Koonin, E.V., Krylov, D.M., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N.*, et al.* (2003). The COG database: an updated version includes eukaryotes. BMC bioinformatics *4*, 41.

Tatusov, R.L., Koonin, E.V., and Lipman, D.J. (1997). A genomic perspective on protein families. Science (New York, N.Y *278*, 631-637.

Tatusov, R.L., Natale, D.A., Garkavtsev, I.V., Tatusova, T.A., Shankavaram, U.T., Rao, B.S., Kiryutin, B., Galperin, M.Y., Fedorova, N.D., and Koonin, E.V. (2001). The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res *29*, 22-28.

Thauer, R.K., and Shima, S. (2006). Biogeochemistry: methane and microbes. Nature *440*, 878-879.

Tori, K., Kimizu, M., Ishino, S., and Ishino, Y. (2007). DNA polymerases BI and D, from the hyperthermophilic archaeon, Pyrococcus furiosus, both bind to PCNA with their C-terminal PIP box motifs. J Bacteriol.

Tsoka, S., and Ouzounis, C.A. (2000). Prediction of protein interactions: metabolic enzymes are frequently involved in gene fusion. Nature genetics *26*, 141-142.

Tye, B.K. (2000). Insights into DNA replication from the third domain of life. Proceedings of the National Academy of Sciences of the United States of America *97*, 2399-2401.

Uemori, T., Sato, Y., Kato, I., Doi, H., and Ishino, Y. (1997). A novel DNA polymerase in the hyperthermophilic archaeon, Pyrococcus furiosus: gene cloning, expression, and characterization. Genes Cells *2*, 499-512.

Ueno, Y., Yamada, K., Yoshida, N., Maruyama, S., and Isozaki, Y. (2006). Evidence from fluid inclusions for microbial methanogenesis in the early Archaean era. Nature *440*, 516-519.

Valentine, D.L. (2007). Adaptations to energy stress dictate the ecology and evolution of the Archaea. Nature reviews *5*, 316-323.

Vas, A., and Leatherwood, J. (2000). Where does DNA replication start in archaea? Genome biology *1*, REVIEWS1020.

Vianna, M.E., Conrads, G., Gomes, B.P., and Horz, H.P. (2006). Identification and quantification of archaea involved in primary endodontic infections. J Clin Microbiol *44*, 1274-1282.

Volkening, M., and Hoffmann, I. (2005). Involvement of human MCM8 in prereplication complex assembly by recruiting hcdc6 to chromatin. Molecular and cellular biology *25*, 1560-1568.

von Mering, C., Jensen, L.J., Kuhn, M., Chaffron, S., Doerks, T., Kruger, B., Snel, B., and Bork, P. (2007). STRING 7--recent developments in the integration and prediction of protein interactions. Nucleic Acids Res *35*, D358-362.

von Mering, C., Jensen, L.J., Snel, B., Hooper, S.D., Krupp, M., Foglierini, M., Jouffre, N., Huynen, M.A., and Bork, P. (2005). STRING: known and predicted protein-protein associations, integrated and transferred across organisms. Nucleic Acids Res *33*, D433-437.

Waga, S., and Stillman, B. (1994). Anatomy of a DNA replication fork revealed by reconstitution of SV40 DNA replication in vitro. Nature *369*, 207-212.

Waga, S., and Stillman, B. (1998). The DNA replication fork in eukaryotic cells. Annual review of biochemistry *67*, 721-751.

Walsh, D.A., and Doolittle, W.F. (2005). The real 'domains' of life. Curr Biol *15*, R237-240.

Walter, J.C., and Araki, H. (2006). Activation of Pre-replication Complexes. In DNA replication and human disease, M.L. DePamphilis, ed. (Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory Press), pp. 89-104.

Wang, M., Yafremava, L.S., Caetano-Anolles, D., Mittenthal, J.E., and Caetano-Anolles, G. (2007). Reductive evolution of architectural repertoires in proteomes and the birth of the tripartite world. Genome research *17*, 1572-1585.

Waters, E., Hohn, M.J., Ahel, I., Graham, D.E., Adams, M.D., Barnstead, M., Beeson, K.Y., Bibbs, L., Bolanos, R., Keller, M.*, et al.* (2003). The genome of Nanoarchaeum equitans: insights into early archaeal evolution and derived parasitism. Proceedings of the National Academy of Sciences of the United States of America *100*, 12984-12988.

Watson, J.D., and Crick, F.H. (1953a). Genetical implications of the structure of deoxyribonucleic acid. Nature *171*, 964-967.

Watson, J.D., and Crick, F.H. (1953b). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature *171*, 737-738.

Wheelis, M.L., Kandler, O., and Woese, C.R. (1992). On the nature of global classification. Proceedings of the National Academy of Sciences of the United States of America *89*, 2930-2934.

White, M.F. (2003). Archaeal DNA repair: paradigms and puzzles. Biochemical Society transactions *31*, 690-693.

Whittaker, R.H. (1969). New concepts of kingdoms or organisms. Evolutionary relations are better represented by new classifications than by the traditional two kingdoms. Science (New York, N.Y *163*, 150-160.

Woese, C.R. (1987). Bacterial evolution. Microbiological reviews *51*, 221-271.

Woese, C.R. (1994). There must be a prokaryote somewhere: microbiology's search for itself. Microbiological reviews *58*, 1-9.

Woese, C.R. (1998). Default taxonomy: Ernst Mayr's view of the microbial world. Proceedings of the National Academy of Sciences of the United States of America *95*, 11043-11046.

Woese, C.R. (2004). A new biology for a new century. Microbiol Mol Biol Rev *68*, 173-186.

Woese, C.R. (2007). The birth of the archaea: a personal retrospective. In Archaea: evolution, physiology, and molecular biology, R.A. Garrett, and H.P. Klenk, eds. (Malden, MA: Blackwell Pub.), pp. 1-15.

Woese, C.R., and Fox, G.E. (1977). Phylogenetic Structure of the Prokaryotic Domain: The Primary Kingdoms. PNAS *74*, 5088-5090.

Woese, C.R., Kandler, O., and Wheelis, M.L. (1990). Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proceedings of the National Academy of Sciences of the United States of America *87*, 4576-4579.

Woese, C.R., Magrum, L.J., and Fox, G.E. (1978). Archaebacteria. Journal of molecular evolution *11*, 245-251.

Wohlschlegel, J.A., Dwyer, B.T., Dhar, S.K., Cvetic, C., Walter, J.C., and Dutta, A. (2000). Inhibition of eukaryotic DNA replication by geminin binding to Cdt1. Science (New York, N.Y *290*, 2309-2312.

Wolf, Y.I., Rogozin, I.B., Grishin, N.V., and Koonin, E.V. (2002). Genome trees and the tree of life. Trends Genet *18*, 472-479.

Wolf, Y.I., Rogozin, I.B., Kondrashov, A.S., and Koonin, E.V. (2001). Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. Genome research *11*, 356-372.

Wu, K., Lai, X., Guo, X., Hu, J., Xiang, X., and Huang, L. (2007). Interplay between primase and replication factor C in the hyperthermophilic archaeon Sulfolobus solfataricus. Molecular microbiology *63*, 826-837.

Xia, W., and Dowhan, W. (1995). In vivo evidence for the involvement of anionic phospholipids in initiation of DNA replication in Escherichia coli. Proceedings of the National Academy of Sciences of the United States of America *92*, 783-787.

Yabuuchi, H., Yamada, Y., Uchida, T., Sunathvanichkul, T., Nakagawa, T., and Masukata, H. (2006). Ordered assembly of Sld3, GINS and Cdc45 is distinctly regulated by DDK and CDK for activation of replication origins. The EMBO journal *25*, 4663-4674.

Yamashiro, K., Yokobori, S., Oshima, T., and Yamagishi, A. (2006). Structural analysis of the plasmid pTA1 isolated from the thermoacidophilic archaeon Thermoplasma acidophilum. Extremophiles *10*, 327-335.

Yamazaki, S., Yamazaki, J., Nishijima, K., Otsuka, R., Mise, M., Ishikawa, H., Sasaki, K., Tago, S., and Isono, K. (2006). Proteome analysis of an aerobic hyperthermophilic crenarchaeon, Aeropyrum pernix K1. Mol Cell Proteomics *5*, 811-823.

Yoshimochi, T., Fujikane, R., Kawanami, M., Matsunaga, F., and Ishino, Y. (2008). The GINS complex from Pyrococcus furiosus stimulates the MCM helicase activity. The Journal of biological chemistry *283*, 1601-1609.

Yuzhakov, A., Kelman, Z., and O'Donnell, M. (1999). Trading places on DNA--a three-point switch underlies primer handoff from primase to the replicative DNA polymerase. Cell *96*, 153-163.

Zakrzewska-Czerwinska, J., Jakimowicz, D., Zawilak-Pawlik, A., and Messer, W. (2007). Regulation of the initiation of chromosomal replication in bacteria. FEMS microbiology reviews *31*, 378-387.

Zawilak-Pawlik, A.M., Kois, A., and Zakrzewska-Czerwinska, J. (2006). A simplified method for purification of recombinant soluble DnaA proteins. Protein expression and purification *48*, 126-133.

Zegerman, P., and Diffley, J.F. (2007). Phosphorylation of Sld2 and Sld3 by cyclin-dependent kinases promotes DNA replication in budding yeast. Nature *445*, 281-285.

Zhang, R., and Zhang, C.T. (2004). Identification of replication origins in the genome of the methanogenic archaeon, Methanocaldococcus jannaschii. Extremophiles *8*, 253-258.

Zhang, R., and Zhang, C.T. (2005). Identification of replication origins in archaeal genomes based on the Z-curve method. Archaea (Vancouver, B.C *1*, 335-346.

Zhang, Y., Yu, Z., Fu, X., and Liang, C. (2002). Noc3p, a bHLH protein, plays an integral role in the initiation of DNA replication in budding yeast. Cell *109*, 849-860.

Zuckerkandl, E., and Pauling, L. (1965). Molecules as documents of evolutionary history. Journal of theoretical biology *8*, 357-366.

ANNEXES

Protocole 1 : Expression de la protéine *Pfu*Cdc6/Orc1 dans des cellules recombinantes de levure

Protocole 2 : Purification de la protéine *Pfu*Cdc6/Orc1

Protocole 3 : Co-purification sur résine Ni-NTA avec une protéine appât fusionnée à une étiquette hexahistidine

Fiches clonage

Protocole 1 : Expression de la protéine *Pfu*Cdc6/Orc1 dans des cellules recombinantes de levure

Expression of the Cdc6/Orc1 protein from *Pyrococcus furiosus* in recombinant yeast cells from the EasySelect Pichia expression system (Invitrogen)

1 Amplification of cell biomass

- Inoculate 2 × 160 ml of MGYH medium (2 Erlenmeyer of 1 liter) with 2 × 0.5 ml (0.5 ml per 160 ml culture) of a glycerol stock of *Pichia pastoris* KM71H AOX1::cdc6^{+} strain
- Incubate at 30°C, 160 rpm until OD_{600nm} is between 2 to 4
- Dilute the starter culture 1/100 into 4 × 2 liters of MGYH medium (4 Erlenmeyer of 5 liters)
- Incubate at 30°C under moderate shaking (160 rpm) until OD600nm is between 2 to 6 (ideally 4)

2 Protein expression

- Harvest cells by centrifugation at 2000 x *g* for 5 minutes at 4°C
- Discard supernatant very carefully
- Resuspend cells in 2 × 1 liter of MMH medium (2 Erlenmeyer of 5 liters)
- Incubate at 30°C for 24 hours under moderate shaking (160 rpm)
- Add 5 ml of 100% methanol to each 1 liter of culture
- Incubate at 30°C for 24 hours under moderate shaking
- Harvest cells by centrifugation at 5000 x *g* for 10 minutes at 20°C
- Resuspend each cell pellet in 30 ml of PBS [137 mM NaCl, 8.1 mM Na_2HPO_4, 2.68 mM KCl, 1.47 mM KH_2PO_4]
- Transfer the resuspended cells into 50 ml conical tubes
- Harvest cells by centrifugation at 5000 x *g* for 10 minutes at 20°C
- Store at -80°C until use

Protocole 2 : Purification de la protéine *Pfu*Cdc6/Orc1

Purification of the Cdc6/Orc1 protein from *Pyrococcus furiosus*

1 Solubilization of protein from inclusion bodies (see Appendix 1 for buffers preparation)

- Heat-treat the cell pellet for 20-30 minutes at 80°C. Gently agitate occasionally.
- Distribute the pellet mixture in 50 ml tubes with about 5-10 ml per tube for better disruption
- Add at least an equal volume of 425 – 600 µm glass beads (Sigma G-8772) and disrupt cells by vortexing at room temperature (max power, at least 30 minutes are required). Confirm the cell breakage by microscopic observation.
- Add 1 ml of chilled Buffer A [50 mM Tris-HCl (pH 8.0), 250 mM NaCl, 10% glycerol, 5 mM ß-mercaptoethanol] per gram of cell pellet.
- Transfer the cell extract and debris to a centrifuge tube via 1-2 intermediate tubes to avoid getting the glass beads.
- Centrifuge at 17,000 *x g* for 10 minutes at 4°C (12,000 rpm for Beckman JA25.50)
- Discard the supernatant and wash the inclusion bodies with 2 ml of Buffer B [2% Triton X100, 10 mM EDTA] pre-cooled on ice per gram of cell pellet (use a spatula to wash the inclusion bodies).
- Centrifuge at 17,000 *x g* for 10 minutes at 4°C
- Wash the inclusion bodies with 2 ml of chilled Buffer A [50 mM Tris-HCl (pH 8.0), 250 mM NaCl, 10% glycerol, 5 mM ß-mercaptoethanol] per gram of cell pellet (use a spatula to wash the inclusion bodies).
- Centrifuge at 17,000 *x g* for 10 minutes at 4°C.
- Add 2 ml Buffer C [50 mM K_2HPO_4/KH_2PO_4, 6 M GuHCl (pH 7.8)] per gram of cell pellet to the inclusion bodies.
- Dissolve the inclusion bodies by incubating the tube at 25°C for 1 hour with gentle agitation.
- Centrifuge at 17,000 *x g* for 10 minutes at 25°C.

In the meantime, prepare the Ni-NTA agarose column (about 1 ml of 50% Ni-NTA agarose slurry (Qiagen) per 10 g of cell pellet)

2 Immobilized Metal Affinity Chromatography (IMAC) purification

- Equilibrate the column with 40 CV of Buffer C [50 mM K_2HPO_4/KH_2PO_4, 6 M GuHCl (pH 7.8)]. Following steps are done at room temperature.
- Apply the supernatant to the Ni-NTA agarose column. Collect flow-through.
- (Apply flow-through to the Ni-NTA agarose column. Collect flow-through)
- Wash column with 10 CV of Buffer C + 5 mM Imidazole [50 mM K_2HPO_4/KH_2PO_4, 6 M GuHCl, 5 mM Imidazole (pH 7.8)]
- Wash column with 40 CV of Buffer C + 20 mM Imidazole [50 mM K_2HPO_4/KH_2PO_4, 6 M GuHCl, 20 mM Imidazole (pH 7.8)]
- Elute *Pfu*Cdc6 with 4 × 5 CV of Buffer C + 300 mM Imidazole [50 mM K_2HPO_4/KH_2PO_4, 6 M GuHCl, 300 mM Imidazole (pH 7.8)]
- Separate GuHCl from protein with 10% TCA precipitation and analyze fractions by SDS-PAGE (see Appendix 2). Save the fraction(s) that contained the protein (usually the first fraction after elution with 300 mM Imidazole)

  Wash the column with 20 CV of Buffer C + 1 M Imidazole [50 mM K_2HPO_4/KH_2PO_4, 6 M GuHCl, 1 M Imidazole (pH 7.8)]
  Store the column in 5 CV of Buffer A (-) [50 mM Tris-HCl (pH 8.0), 250 mM NaCl, 10% glycerol]

3 Folding by the rapid dilution method

- Pour the elicited fractions into 100 volumes of Buffer A [50 mM Tris-HCl (pH 8.0), 250 mM NaCl, 10% glycerol, 5 mM ß-mercaptoethanol]
- Keep the solution at 25°C overnight

4 Concentration of *Pfu*Cdc6 sample by IMAC purification

- The following morning, heat-treat the solution at 80°C for 20 minutes
- Centrifuge at 17,000 g, 10 minutes, 25°C

  Meanwhile, re-equilibrate the column with 40 CV of Buffer A [50 mM Tris-HCl (pH 8.0), 250 mM NaCl, 10% glycerol, 5 mM ß-mercaptoethanol]

- Take the supernatant and pour it onto the column

- Remove tag by thrombin digestion (see Appendix 3) or elute the affinity-bound hexahistidine tagged *Pfu*Cdc6 with 10 × 1 CV of Buffer A + 300 mM Imidazole [50 mM Tris-HCl, 250 mM NaCl, 10% glycerol, 300 mM Imidazole, 5 mM ß-mercaptoethanol (pH 8.0)]
- Analyze fractions by SDS-PAGE

Wash the column with 20 CV of Buffer A + 1 M Imidazole [50 mM Tris-HCl, 250 mM NaCl, 10% glycerol, 5 mM ß-mercaptoethanol, 1 M Imidazole (pH 8.0)]
Store the column in 5 CV of Buffer A (-) [50 mM Tris-HCl (pH 8.0), 250 mM NaCl, 10% glycerol]

5 Final recommendations

- Dialyze the concentrated Cdc6 against Buffer A [50 mM Tris-HCl (pH 8.0), 250 mM NaCl, 10% glycerol, 5 mM ß-mercaptoethanol], overnight at 25°C.
- Quantify the protein concentration and conserve at 25°C; otherwise, flash-freeze with liquid nitrogen

Important note
*Pfu*Cdc6 tends to aggregate; to solubilise these aggregates heat samples at 80°C for 15 minutes and centrifuge at 17,000 *x g* for 10 minutes.

Appendix 1: Buffers preparation

Tampon A (-)

50 mM Tris-HCl (pH 8.0), 250 mM NaCl, 10% Glycérol,	1 l
(5 mM ß-mercaptoethanol)	
1M Tris-HCl (pH 8.0)	50 ml
5M NaCl	50 ml
Glycerol	100 ml

Qsp 1 l with water

Autoclave

Store at RT

Add ß-mercaptoethanol freshly from 14.3M stock solution to prepare Buffer A

Tampon B

2% Triton X100, 10 mM EDTA	50 ml
Triton X100	1 ml
0.5M EDTA (pH 8.0)	1 ml

Qsp 50 ml with water

Foil in aluminium to protect the solution from light

Autoclave

Store at 4°C

Tampon C

50 mM K_2HPO_4/KH_2PO_4, 6M GuHCl (pH 7.8)	for 100 ml
1M K_2HPO_4/KH_2PO_4 (pH 7.8)	5 ml
GuHCl (MW = 95.53 g/mol)	57.318 g

Adjust pH with KOH

Qsp 100 ml with water

Autoclave

Store at RT

Tampon C + 5 mM Imidazole

50 mM K_2HPO_4/KH_2PO_4, 6M GuHCl, 5 mM Imidazole (pH 7.8)	100 ml
1M K_2HPO_4/KH_2PO_4 (pH 7.8)	5 ml
GuHCl (MW = 95.53 g/mol)	57.318 g
Imidazole 1M (pH 8.0)	500 µl

Adjust pH with 0.5-1M KOH

Qsp 100 ml with water

Foil in aluminium to protect the solution from light

Autoclave

Store at 4°C

Tampon C + 20 mM Imidazole

50 mM K_2HPO_4/KH_2PO_4, 6M GuHCl, 20 mM Imidazole (pH 7.8)	100 ml
1M K_2HPO_4/KH_2PO_4 (pH 7.8)	5 ml
GuHCl (MW = 95.53 g/mol)	57.318 g
Imidazole 1M (pH 8.0)	4 ml

Adjust pH with 0.5-1M KOH

Qsp 100 ml with water

Foiled in aluminium to protect the solution from light

Autoclave

Store at 4°C

Tampon C + 300 mM Imidazole

50 mM K_2HPO_4/KH_2PO_4, 6M GuHCl, 300 mM Imidazole (pH 7.8)	100 ml
1M K_2HPO_4/KH_2PO_4 (pH 7.8)	5 ml
GuHCl (MW = 95.53 g/mol)	57.318 g
Imidazole (MW = 68.08 g/mol)	2.024 g

Adjust pH with 0.5-1M KOH

Qsp 100 ml with water

Foil in aluminium to protect the solution from light

Autoclave

Store at 4°C

Appendix 2: TCA precipitation

- Dilute samples to 100 µL (e.g., add 50 µL H_2O to 50 µL protein samples)
- Add 100 µL of 10% TCA
- Leave on ice for 20 minutes
- Centrifuge 14,000 rpm (17,800 g), 15 minutes, 4°C
- Wash pellet 300 µL of ice-cold ethanol; centrifuge 15,000 rpm, 5 minutes, 4°C
- Dry for 5 minutes in a heated vacuum chamber
- Resuspend protein in 50 µL SDS-PAGE loading buffer [62.5 mM Tris-HCl (pH 6.8), 2% SDS, 100 mM DTT, 10 % glycerol, 0.01% bromophenol blue]

Appendix 3: Removal of the N-terminal hexahistidine tag by thrombin

The removal of the tag can be done either during the concentration of Cdc6 on the Ni-NTA column (to obtain only untagged *Pfu*Cdc6), or after the dialysis following the concentration on Ni-NTA column by pouring imidazole-free protein on Ni-NTA column.
In both cases, the protein concentration should be quantified prior to the digestion.

- Load the protein onto the Ni-NTA column
- Cap the bottom of the column (or close the tap)
- Prepare 1-2 unit(s) of thrombin per mg of protein in 1 ml of Buffer A (-) [50 mM Tris-HCl (pH 8.0), 250 mM NaCl, 10% glycerol]
- Load the enzyme reaction mixture onto the column
- Digest overnight
- Recover the untagged *Pfu*Cdc6 by uncapping the column
- Elute the remaining untagged protein with 5 × 1 CV of Buffer A [50 mM Tris-HCl (pH 8.0), 250 mM NaCl, 10% glycerol, 5 mM ß-mercaptoethanol]
- Wash the column with 5 × 1 CV of Buffer A + 300 mM Imidazole [50 mM Tris-HCl, 250 mM NaCl, 10% glycerol, 300 mM Imidazole, 5 mM ß-mercaptoethanol (pH 8.0)]
- Analyze the fractions by SDS-PAGE
- Store the protein in appropriate conditions (see Final recommendations)

Protocole 3 : Co-purification sur résine Ni-NTA avec une protéine appât fusionnée à une étiquette hexahistidine

Small-scale co-purification with His-tagged protein bait

Day 1: Transformation of *E. coli* and preculture

- Transform 50 µL competent cells with 200 ng of plasmid DNA
  - ✓ 50 µL cells with 200 ng pKHS [Bait];
  - ✓ 50 µL cells with 200 ng pASH [Prey];
  - ✓ 50 µL cells with 200 ng pKHS [Bait] and 200 ng pASH [Prey]
- Incubate on ice for 30 minutes
- Heat-shock the cells for 50-60 seconds in a water bath (without agitation) or a thermoblock at exactly 42°C
- Immediately place the tubes on ice for 2 minutes
- Add 1 ml of 2x YT
- Grow for 30-45 minutes at 37°C, 200 rpm
- Inoculate 10 ml of 2x YT containing the appropriate antibiotic(s) with 50 µl (single transformation) or 100 µl (double transformation) of the transformation culture
  For toxic gene (e.g., Cdc6), use 2x YT supplemented with 1% glucose to avoid basal expression of target gene.
- Incubate at 37°C overnight

Day 2: Culture

- Inoculate 10 ml of 2x YT containing the appropriate antibiotic(s) at OD600nm=0.1 with the overnight preculture
- Grow at 37°C until an OD600 of 1.0 is reached
- Induce protein expression with 0.5 mM IPTG (final concentration)
- Incubate at 37°C for 3 hours
- Pellet the cells (5000 g, 10 minutes)
- Store at -80°C overnight

Day 3: Small-scale purification

- Thaw the resuspended cells on ice
- Resuspend the cell pellet in 1 ml lysis buffer [20 mM Tris-HCl pH 7.5; 250 mM NaCl; 5% glycerol; (5 mM ß-mercaptoethanol)]
  - ➲ *Add ß-mercaptoethanol freshly*
- Transfer the resuspended cells to a 2 ml centrifuge tube
- Lyse cells by sonication with a micro-tip sonicator using 2-3 pulses of 15 seconds separated by 30 seconds intervals (keep cells on ice while proceeding)
  - ➲ *Save 80 µl of lysate for SDS-PAGE analysis (FTot)*
- Centrifuge lysate at 20000 g for 30 minutes, at 4°C
  - ➲ *Save 80 µl of supernatant for SDS-PAGE analysis (FS)*
  - ➲ *Transfer the supernatant to a new 2.0 ml microcentrifuge tube*
- Heat the supernatant for 15 minutes at 75°C
- Centrifuge at 20000 g for 25 minutes, at 4°C
  - ➲ *Save 80 µl of supernatant for SDS-PAGE analysis (FS 75°C)*
- During centrifugation, wash Ni-NTA agarose resin (100 µl of Ni-NTA slurry; 50 µl bed volume resin for each sample) 3 times (spin at 20000 g for 1 minute to pellet the resin with 1 ml binding buffer (20 mM Tris-HCl pH 7.5; 250 mM NaCl; 20 mM Imidazole; 5% glycerol; 5 mM ß-mercaptoethanol)) in a 2 ml microcentrifuge tube
- Add the thermostable protein supernatant to the 100 µl of washed Ni-NTA agarose resin (2 ml tube is preferable for efficient mixing)
- Add imidazole to 20 mM final concentration (about 750 µl of thermostable protein supernatant)
- Incubate at room temperature for at least 1 hour with mixing (use a wheel for efficient mixing)
- Pellet resin by centrifuging at 20000 g for 1 minute
  - ➲ *Save 80 µl of supernatant for SDS-PAGE analysis (FT; unbound material)*
- Wash resin 3 times with 500 µl (10 column volumes) of washing buffer (20 mM Tris-HCl pH 7.5; 250 mM NaCl; 50 mM Imidazole; 5% glycerol; 5 mM ß-mercaptoethanol); mix for 5 minutes; pellet the resin; save an aliquot for SDS-PAGE; discard supernatant)
  - ➲ *Save 25 µl of supernatant after each wash for SDS-PAGE analysis (W); pool wash samples*
- Add 80 µl of SDS-PAGE sampling buffer 5X to the resin
  - ➲ *This sample corresponds to bound material (P)*
- Heat all SDS-PAGE samples at 95°C-100°C for 5 minutes
- Load 30 µl of each sample onto an SDS-PAGE gel

Fiches clonages

Cdc6/Orc1

DNA polymerase protein 1 (DP1)

DNA polymerase protein 2 (DP2)

DNA primase small subunit (PriS)

DNA primase large subunit (PriL)

Ribosomal protein L44E

Ribosomal protein S27E

Archaeal initiation factor 2 subunit beta

Mini-chromosome maintenance (MCM)

Go ichi ni san subunit 23 (Gins23)

Go ichi ni san subunit 15 (Gins15)

Proliferating Cell Nuclear Antigen (PCNA)

Transcription factor S (TFS)

Archaeal initiation factor 2 subunit alpha

Nucleolar protein 10 (Nop10)

COG2047

Silent information regulator 2 (Sir2)

NudF

Recombination J (RecJ)

Cdc6/Orc1 (Cell division cycle 6/Origin recognition complex 1)

COG1474: Cdc6-related protein, AAA superfamily ATPase

Facteur d'initiation de la réplication de l'ADN

> PF0017

ATGAACGAAGGTGAACATCAAATAAAGCTTGACGAGCTATTCGAAAAGTTGCTCCGAGCTAGGAAGATAT
TCAAAAACAAAGATGTCCTTAGGCATAGCTATACTCCCAAGGATCTACCTCACAGACATGAGCAAATAGA
AACTCTCGCCCAAATTTTAGTACCAGTTCTCAGAGGAGAAACTCCATCAAACATATTCGTTTATGGGAAG
ACTGGAACTGGAAAGACTGTAACTGTAAAATTTGTAACTGAAGAGCTGAAAAGAATATCTGAAAAATACA
ACATTCCAGTTGATGTGATCTACATTAATTGTGAGATTGTCGATACTCACTATAGAGTTCTTGCTAACAT
AGTTAACTACTTCAAAGATGAGACTGGGATTGAAGTTCCAATGGTAGGTTGGCCTACCGATGAAGTTTAC
GCAAAGCTTAAGCAGGTTATAGATATGAAGGAGAGGTTTGTGATAATTGTGTTGGATGAAATTGACAAGT
TGGTAAAGAAGAGTGGTGATGAGGTTCTCTATTCATTAACAAGAATAAATACTGAACTTAAAAGGGCTAA
AGTGAGTGTAATTGGTATATCAAACGACCTTAAATTTAAAGAGTATCTAGATCCAAGAGTTCTCTCAAGT
TTGAGTGAGGAAGAGGTGGTATTCCCACCCTATGATGCAAATCAGCTTAGGGATATACTGACCCAAAGAG
CTGAAGAGGCCTTTTATCCTGGGGTTTTAGACGAAGGTGTGATTCCCCTCTGTGCAGCATTAGCTGCTAG
AGAGCATGGAGATGCAAGAAAGGCACTTGACCTTCTAAGAGTTGCAGGGGAAATAGCGGAAAGAGAAGGG
GCAAGTAAAGTAACTGAAAAGCATGTTTGGAAAGCCCAGGAAAAGATTGAACAGGACATGATGGAGGAGG
TAATAAAAACTCTACCCCTTCAGTCAAAAGTTCTCCTCTATGCCATAGTTCTTTTGGACGAAAACGGCGA
TTTACCAGCAAATACTGGGGATGTTTACGCTGTTTATAGGGAATTGTGCGAGTACATTGACTTGGAACCT
CTCACCCAAAGAAGGATAAGTGATCTAATTAATGAGCTTGACATGCTTGGAATAATAAATGCAAAAGTTG
TTAGTAAGGGGAGATATGGGAGGACAAAGGAAATAAGGCTTAACGTTACCTCATATAAGATAAGAAATGT
GCTGAGATATGATTACTCTATTCAGCCCCTCCTCACAATTTCCCTTAAGAGTGAGCAGAGGAGGTTGATC
TAA

> Amorces de clonage

Cdc6-Pf-EagI 5'-TTTTCGGCCGATTAATTTAAGAAGGAGATATATATATGAACGAAGGTGAACATCA-3'

Cdc6-Pf-NotI 5'-TTTTTGCGGCCGCGATCAACCTCCTCTGCTCAC-3'

> Amorces mutagènes

PF0017_mut113F 5'-GCTATACTCCCAAGG**A**TCTACCTCACAGAC-3'

PF0017_mut113R 5'-GTCTGTGAGGTAGA**T**CCTTGGGAGTATAGC-3'

DP1 (DNA polymerase Protein 1)

COG1311: Archaeal DNA polymerase II, small subunit/DNA polymerase delta, subunit B

Petite sous-unité de l'ADN polymérase D (activité exonucléase)

> PF0018

ATGGATGAATTTGTAAAATCACTTCTAAAAGCTAACTATCTAATAACTCCCTCTGCCTACTATCTCTTGA
GAGAATACTATGAAAAAGGTGAATTCTCAATTGTGGAGCTGGTAAAATTTGCAAGATCAAGAGAGAGCTA
CATAATTACTGATGCTTTAGCAACAGAATTCCTTAAAGTTAAAGGCCTTGAACCAATTCTTCCAGTGGAA
ACAAAGGGGGGTTTTGTTTCCACTGGAGAGTCCCAAAAAGAGCAGTCTTATGAAGAGTCTTTTGGGACTA
AAGAAGAAATTTCCCAGGAGATTAAAGAAGGAGAGAGTTTTATTTCCACTGGAAGTGAACCACTTGAAGA
GGAGCTCAATAGCATTGGAATTGAGGAAATTGGGGCAAATGAAGAGTTAGTTTCTAATGGAAATGACAAT
GGTGGAGAGGCAATTGTCTTTGACAAATATGGCTATCCAATGGTATATGCTCCAGAAGAAATAGAGGTTG
AGGAGAAGGAGTACTCGAAGTATGAAGATCTGACAATACCCATGAACCCCGACTTCAATTATGTGGAAAT
AAAGGAAGATTATGATGTTGTCTTCGATGTTAGGAATGTAAAGCTGAAGCCTCCTAAGGTAAAGAACGGT
AATGGGAAGGAAGGTGAAATAATTGTTGAAGCTTATGCTTCTCTCTTCAGGAGTAGGTTGAAGAAGTTAA
GGAAAATACTAAGGGAAAATCCTGAATTGGACAATGTTGTTGATATTGGGAAGCTGAAGTATGTGAAGGA
AGATGAAACCGTGACAATAATAGGGCTTGTCAATTCCAAGAGGGAAGTGAATAAAGGATTGATATTTGAA
ATAGAAGATCTCACAGGAAAGGTTAAAGTTTTCTTGCCGAAAGATTCGGAAGATTATAGGGAGGCATTTA
AGGTTCTTCCAGATGCCGTCGTCGCTTTTAAGGGGGTGTATTCAAAGAGGGGAATTTTGTACGCCAACAA
GTTTTACCTTCCAGACGTTCCCCTCTATAGGAGACAAAAGCCTCCACTGGAAGAGAAAGTTTATGCTATT
CTCATAAGTGATATACACGTCGGAAGTAAAGAGTTCTGCGAAAATGCCTTCATAAAGTTCTTAGAGTGGC
TCAATGGAAACGTTGAAACTAAGGAAGAGGAAGAAATCGTGAGTAGGGTTAAGTATCTAATCATTGCAGG
AGATGTTGTTGATGGTGTTGGCGTTTATCCGGGCCAGTATGCCGACTTGACGATTCCAGATATATTCGAC
CAGTATGAGGCCCTCGCAAACCTTCTCTCTCACGTTCCTAAGCACATAACAATGTTCATTGCCCCAGGAA
ACCACGATGCTGCTAGGCAAGCTATTCCCCAACCAGAATTCTACAAAGAGTATGCAAAACCTATATACAA
GCTCAAGAACGCCGTGATAATAAGCAATCCTGCTGTAATAAGACTACATGGTAGGGACTTTCTGATAGCT
CATGGTAGGGGGATAGAGGATGTCGTTGGAAGTGTTCCTGGGTTGACCCATCACAAGCCCGGCCTCCCAA
TGGTTGAACTATTGAAGATGAGGCATGTAGCTCCAATGTTTGGAGGAAAGGTTCCAATAGCTCCTGATCC
AGAAGATTTGCTTGTTATAGAAGAAGTTCCTGATGTAGTTCACATGGGTCACGTTCACGTTTACGATGCG
GTAGTTTATAGGGGAGTTCAGCTGGTTAACTCCGCCACCTGGCAGGCTCAGACCGAGTTCCAGAAGATGG
TGAACATAGTTCCAACGCCTGCAAAGGTTCCCGTTGTTGATATTGATACTGCAAAAGTTGTCAAGGTTTT
GGACTTTAGTGGGTGGTGC**TGA**

> Amorces de clonage

DP1-Pfu-EagI 5'-TTTTCGGCCGATTAATTTAAGAAGGAGATATATATATGGATGAATTTGTAAAATC-3'

DP1-Pfu-NotI 5'-TTTTTGCGGCCGCGCACCACCCACTAAAGTCCA-3'

DP2 (DNA polymerase Protein 2)

COG1933: Archaeal DNA polymerase II, large subunit

Grande sous-unité de l'ADN polymérase D (sous-unité catalytique)

> PF0019

GTGGTGCTGATGGAGCTTCCAAAGGAAATTGAGGAGTATTTTGAGATGCTTCAAAGGGAAATTGACAAAG
CTTACGAGATTGCTAAGAAGGCTAGGAGTCAGGGTAAAGACCCCTCAACCGATGTTGAGATTCCCCAGGC
TACAGACATGGCTGGAAGAGTTGAGAGCTTAGTTGGCCCTCCCGGAGTTGCTCAGAGAATTAGGGAGCTT
TTAAAAGAGTATGATAAGGAAATTGTTGCTTTAAAGATAGTTGATGAGATAATTGAGGGCAAATTTGGTG
ATTTTGGAAGTAAAGAGAAGTACGCTGAACAGGCTGTAAGGACAGCCTTGGCAATATTAACTGAGGGTAT
TGTTTCTGCTCCACTTGAGGGTATAGCTGATGTTAAAATCAAGCGAAACACCTGGGCTGATAACTCTGAA
TACCTCGCCCTTTACTATGCTGGGCCAATTAGGAGTTCTGGTGGAACTGCTCAAGCTCTCAGTGTACTTG
TTGGTGATTACGTTAGGCGAAAGCTTGGCCTTGATAGGTTTAAGCCAAGTGGGAAGCATATAGAGAGAAT
GGTTGAGGAAGTTGACCTCTATCATAGAGCTGTTTCAAGGCTTCAATATCATCCCTCACCTGATGAAGTG
AGATTAGCAATGAGGAATATTCCCATAGAAATCACTGGTGAAGCCACTGACGATGTGGAGGTTTCCCATA
GAGATGTAGAGGGAGTTGAGACAAATCAGCTGAGAGGAGGAGCGATCCTAGTTTTGGCGGAGGGTGTTCT
CCAGAAGGCTAAAAAGCTCGTGAAATACATTGACAAGATGGGGATTGATGGATGGGAGTGGCTTAAAGAG
TTTGTAGAGGCTAAAGAAAAAGGTGAAGAAATCGAAGAGAGTGAAAGTAAAGCCGAGGAGTCAAAAGTTG
AAACAAGGGTGGAGGTAGAGAAGGGATTCTACTACAAGCTCTATGAGAAATTTAGGGCTGAGATTGCCCC
AAGCGAAAAGTATGCAAAGGAAATAATTGGTGGGAGGCCGTTATTCGCTGGACCCTCGGAAAATGGGGGA
TTTAGGCTTAGATATGGTAGAAGTAGGGTGAGTGGATTTGCAACATGGAGCATAAATCCAGCAACAATGG
TTTTGGTTGACGAGTTCTTGGCCATTGGAACTCAAATGAAAACCGAGAGGCCTGGGAAAGGTGCAGTAGT
GACTCCAGCAACAACCGCTGAAGGGCCGATTGTTAAGCTAAAGGATGGGAGTGTTGTTAGGGTTGATGAT
TACAACTTGGCCCTCAAAATAAGGGATGAAGTCGAAGAGATACTTTATTTGGGAGATGCAATCATAGCCT
TTGGAGACTTTGTGGAGAACAATCAAACTCTCCTTCCTGCAAACTATGTAGAGGAGTGGTGGATCCAAGA
GTTCGTAAAGGCCGTTAATGAGGCATATGAAGTTGAGCTTAGACCCTTTGAGGAAAATCCCAGGGAGAGC
GTTGAGGAAGCAGCAGAGTACCTTGAAGTTGACCCAGAATTCTTGGCTAAGATGCTTTACGATCCTCTAA
GGGTTAAGCCTCCCGTGGAGCTAGCCATACACTTCTCGGAAATCCTGGAAATTCCTCTCCACCCATACTA
CACCCTTTATTGGAATACTGTAAATCCTAAAGATGTTGAAAGACTTTGGGGAGTATTAAAAGACAAGGCC
ACCATAGAATGGGGCACTTTCAGAGGTATAAAGTTTGCAAAGAAAATTGAAATTAGCCTGGACGACCTGG
GAAGTCTTAAGAGAACCCTAGAGCTCCTGGGACTTCCTCATACGGTAAGAGAAGGGATTGTAGTGGTTGA
TTATCCGTGGAGTGCAGCTCTTCTCACTCCATTGGGCAATCTTGAATGGGAGTTTAAGGCCAAGCCCTTC
TACACTGTAATAGACATCATTAACGAGAACAATCAGATAAAGCTCAGGGACAGGGGAATAAGCTGGATAG
GGGCAAGAATGGGAAGGCCAGAGAAGGCAAAAGAAAGAAAAATGAAGCCACCTGTTCAAGTCCTCTTCCC
AATTGGCTTGGCAGGGGGTTCTAGCAGAGATATAAAGAAGGCTGCTGAAGAGGGAAAAATAGCTGAAGTT
GAGATTGCTTTCTTCAAGTGTCCGAAGTGTGGCCATGTAGGGCCTGAAACTCTCTGTCCCGAGTGTGGGA
TTAGGAAAGAGTTGATATGGACATGTCCCAAGTGTGGGGCTGAATACACCAATTCCCAGGCTGAGGGGTA
CTCGTATTCATGTCCAAAGTGCAATGTGAAGCTAAAGCCATTCACAAAGAGGAAGATAAAGCCCTCAGAG
CTCTTAAACAGGGCCATGGAAAACGTGAAGGTTTATGGAGTTGACAAGCTTAAGGGCGTAATGGGAATGA
CTTCTGGCTGGAAGATTGCAGAGCCGCTGGAGAAAGGTCTTTTGAGAGCAAAAAATGAAGTTTACGTCTT
TAAGGATGGAACCATAAGATTTGATGCCACAGATGCTCCAATAACTCACTTTAGGCCTAGGGAGATAGGA
GTTTCAGTGGAAAAGCTGAGAGAGCTTGGCTACACCCATGACTTCGAAGGGAAACCTCTGGTGAGTGAAG
ACCAGATAGTTGAGCTTAAGCCCCAAGATGTAATCCTCTCAAAGGAGGCTGGCAAGTACCTCTTAAGAGT
GGCCAGGTTTGTTGATGATCTTCTTGAGAAGTTCTACGGACTTCCCAGGTTCTACAACGCCGAAAAAATG
GAGGATTTAATTGGTCACCTAGTGATAGGATTGGCCCCTCACACTTCAGCCGGAATCGTGGGGAGGATAA
TAGGCTTTGTAGATGCTCTGGTTGGCTACGCTCACCCCTACTTCCATGCGGCCAAGAGAAGGAACTGTGA
TGGAGATGAGGATAGTGTAATGCTACTCCTTGATGCCCTATTGAACTTCTCCAGATACTACCTCCCCGAA
AAAAGAGGAGGAAAAATGGACGCTCCTCTTGTCATAACCACGAGGCTTGATCCAAGAGAGGTGGACAGTG
AAGTGCACAACATGGATGTCGTTAGATACTATCCATTAGAGTTCTATGAAGCAACTTACGAGCTTAAATC
ACCAAAGGAACTTGTGGGAGTTATAGAGAGAGTTGAAGATAGATTAGGAAAGCCTGAAATGTATTACGGA
ATAAAGTTCACCCACGATACCGACGACATAGCTCTAGGACCAAAGATGAGCCTCTACAAGCAGTTGGGAG
ATATGGAGGAGAAAGTGAAGAGGCAATTGACATTGGCAGAGAGAATTAGAGCTGTGGATCAACACTATGT
TGCTGAAACAATCCTCAACTCCCACTTAATTCCCGACTTGAGGGGTAACCTAAGGAGCTTTACTAGACAA
GAATTTCGCTGTGTGAAGTGTAACACAAAGTACAGAAGGCCGCCCTTGGATGGAAAATGCCCAGTCTGTG
GAGGAAAGATAGTGCTGACAGTTAGCAAAGGAGCCATTGAAAAGTACTTGGGGACTGCCAAGATGCTCGT
AGCTAACTACAACGTAAAGCCATATACAAGGCAGAGAATATGCTTGACGGAGAAGGATATTGATTCACTC
TTTGAGTACTTATTCCCAGAAGCCCAGTTAACGCTCATTGTAGATCCAAACGACATCTGTATGAAAATGA

TCAAGGAAAGAACGGGGGAAACAGTTCAAGGAGGCCTGCTTGAGAACTTTAATTCCTCTGGAAATAATGG
GAAGAAAATAGAGAAGAAGGAGAAAAAGGCAAAGGAAAAGCCTAAAAAGAAGAAAGTTATAAGCTTGGAC
GACTTCTTCTCCAAACGC**TGA**

> Amorces de clonage

DP2-Pfu-EagI	5'-TTTTCGGCCGATTAATTTAAGAAGGAGATATATATATG GTGCTGATGGAGCT-3'
DP2-Pfu-NotI	5'-TTTTTGCGGCCGCGCGTTTGGAGAAGAAGTCGTC-3'

PriS (DNA primase small-subunit)

COG1467: Eukaryotic-type DNA primase, catalytic (small) subunit

Petite sous-unité de l'ADN primase (sous-unité catalytique)

> PF0110

ATGCTGATGAGGGAAGTGACAAAGGAGGAAAGGAGCGAATTCTACAGTAAAGAATGGAGTGCAAAGAAAA
TACCAAAGTTCATAGTGGACACTCTAGAAAGTAGAGAATTCGGCTTCGATCATAACGGGGAAGGTCCAAG
TGACAGGAAAAATCAATATTCTGACATAAGAGATTTAGAGGACTACATTAGAGCCACATCCCCCTACGCA
GTATATTCAAGTGTGGCATTTTATGAAAACCCCAGGGAGATGGAAGGGTGGAGAGGAGCTGAGTTAGTTT
TTGACATTGATGCCAAGGATCTCCCCCTAAAGAGGTGCAACCACGAACCTGGGACAGTGTGTCCAATATG
CCTTGAAGATGCAAAAGAGCTAGCTAAAGATACTCTAATAATTCTCAGGGAAGAACTCGGCTTTGAAAAT
ATCCATGTAGTCTACTCCGGAAGAGGATATCACATAAGAATCCTAGATGAATGGGCCCTCCAATTGGACT
CCAAAAGTAGAGAAAGAATTCTTGCCTTTATTTCAGCTAGTGAAATTGAGAACGTTGAAGAATTTAGAAG
ATTTCTACTGGAGAAGAGAGGATGGTTTGTGTTAAAGCATGGCTACCCGAGAGTATTTAGGTTGAGACTG
GGATACTTTATTCTAAGGGTTAACGTACCTCACTTGCTAAGCATTGGAATAAGAAGAAATATTGCAAAGA
AAATTCTAGATCACAAAGAAGAAATATACGAGGGATTTGTAAGGAAGGCAATATTGGCATCTTTTCCAGA
AGGCGTGGGAATTGAAAGCATGGCTAAGCTCTTTGCCCTATCAACTAGATTTTCAAAGGCCTATTTTGAT
GGTAGGGTTACAGTTGATATAAAGAGAATCCTAAGGTTGCCCTCAACACTCCATTCCAAAGTGGGCCTTA
TAGCAACTTATGTTGGAACCAAGGAGAGAGAGGTCATGAAGTTTAATCCATTTAGACATGCAGTGCCAAA
GTTCAGGAAAAAAGAAGTGCGCGAAGCTTATAAACTGTGGAGAGAGTCCTTGGAATATGAA**TAA**

> Amorces de clonage

PriS-Pfu-EagI 5'-TTTTCGGCCGATTAATTTAAGAAGGAGATATATATATG CTGATGAGGGAAGTGAC-3'

PriS-Pfu-NotI 5'-TTTTTGCGGC CGCTTCATATTCCAAGGACTCTCTCC-3'

PriL (DNA primase large subunit)

COG2219: Eukaryotic-type DNA primase, large subunit

Grande sous-unité de l'ADN primase (sous-unité régulatrice)

> PF0111

TTGTTAACTTCCATTCTCCCACCTCCATTAAAGTTATATCGGGGTTCATTTATTTTATCAAACATGCTAG
ACCCATTTAGTGAGAAGGCCAAAGAACTACTAAAAGAATTCGGATCAATGAATGAATTCCTTCAAGCTAT
CCCCTCTCTTGTGGATATAGAGGAAGTCATGAATAGGTTAAAATTTGCAAAAGAATCCGAAATCTCCGAA
GATATTCTGAATATAGAGGATATACGAGATTTAGCAAGCTTTTATGCCCAAATAGGAGCATTAGCTTACT
CCCCATATGGACTGGAATTGGAACTAGTAAAGAAGGCTAATTTGAGAATATATACAGAGAGAATCCGCAG
AAGAAGGAAAATAAGGAGCGATGAAATTGGAATTGAAGTAAAAATAGCAGTTGAATTCCCAGAAAACGAC
ATAAAAACACTTGAAAAAGTCTATGGTGGCCTTCCAGAATACATAGTTTCCCTAAGGGAGTTTTTAGATC
TAGTTCCAGATGAAAAACTCTCCTCTTATTACGTCTATGATGGGAATGTGTATTTAAGGAAGGATGACCT
CTTAAAAGTGTGGAGCAAAGCTTTTGAGAGAAACGTTGAAAAGGCCGTGAATATAATTTACGAAATAAGG
GACGAGCTTCCAGAGTTTTATAGAAGACTTGCAGGAGAGATAAGATCTTTTGCCGAGAAAGAATTTTCAG
ATAAGTTTAGAGAGGTTCAAGCAGGAGAACTAAAACACCATCTATTCCCTCCCTGTGTTAAAAATGCTCT
CAGAGGAGTTCCACAGGGAATGAGGAACTATGCAATAACGGTATTGCTCACGAGCTTTCTAAGCTATGCA
AGGATATGTCCAAATCCTCCCAGGAGAAATGTAAAAATTAGGGACTGCATTAAAGATATGAGGGTAATAA
CCGAGGAAATACTTCCCATAATAATAGAGGCCGGGAACAGATGCTCACCTCCACTATTCGAAGATCAACC
AAACGAAATAAAGAATATATGGTACCACTTGGGCTTTGGATACACTGCAAATCCTACCCTTGAAGACAGC
GGGAACTCAACATGGTACTTTCCCCCTAACTGTGATAAGATAAAGGCAAATGCTCCACAGCTTTGCACTC
CTGACAAGCACTGCAGATACATTAGAAATCCCCTAACATATTATCTAAGGCGTCTTTACTTAGAAGAGAA
GAGGAGGGCCAAGCATGCTGATGAGGGAAGTGACAAAGGAGGAAAGGAGCGAATTCTACAG**TAA**

> Amorces de clonage

PriL-Pfu-EagI 5'-TTTTCGGCCGATTAATTTAAGAAGGAGATATATATATGTTAACTTCCATTCTCC-3'

PriL-Pfu-NotI 5'-TTTTTGCGGCCGCCTGTAGAATTCGCTCCTTTCC-3'

L44E

COG1631: Ribosomal protein L44E

Protéine ribosomale L44E (grande sous-unité du ribosome)

> PF0217

ATGAAAATGAAGTATCCAAAGCAAATAAGGACTTATTGCCCGTTTTGTAAGAAACACACAATCCACAAGG
TAGAAAGAGTAAAAAAGAGGCCGAGGAGTGAGCTTAGTGCAGGTCAGAGAAGGTTCAGGAGGATACTTAA
GGGTTATGGAGGATTCCCAAGGCCCAAGCCAGAGGGCAGAGAAAAGCCAGTTAAAAAGCTAGACTTGAGA
TTCAGATGCACTGAGTGTGGAAAAGCTCACACTAGAGGAAGAGGATTTAGAGTAAAGAAGTTTGAGCTAG
TGGAGGGA**TGA**

> Amorces de clonage

L44E-Pfu-EagI	5'-TTTTCGGCCGATTAATTTAAGAAGGAGATATATATATGAAAATGAAGTATCCAAAGC-3'
L44E-Pfu-NotI	5'-TTTTTGCGGCCGCTCCCTCCACTAGCTCAAACT-3'

S27E

COG2051: Ribosomal protein S27E

Protéine ribosomale S27E (petite sous-unité du ribosome)

> PF0218

ATGGCTAAGCCAATAATTCCAATGCCAAGATCAAGATTTCTAAGAGTGAAGTGCATTGACTGTGGAAATG
AGCAGATAGTATTTAGCCATCCAGCAACTAAAGTAAGATGCCTAATTTGCGGAGCAACTCTTGTTGAGCC
AACAGGTGGAAAGGGAATTGTAAAAGCTAAGATCCTTGAAGTTCTAGAG**TGA**

> Amorces de clonage

S27E-Pfu-EagI	5'-TTTTCGGCCGATTAATTTAAGAAGGAGATATATATATGGCTAAGCCAATAATTC-3'
S27E-Pfu-NotI	5'-TTTTTGCGGCCGCCTCTAGAACTTCAAGGATCTTAG-3'

aIF-2 beta (archaeal Initiation Factor 2 subunit beta)

COG1601: Translation initiation factor 2, beta subunit (eIF-2beta)/eIF-5 N-terminal domain

Sous-unité beta du facteur d'initiation de la traduction aIF-2

> PF0481

ATGGAAATTGACTATTACGATTATGAAAAGCTCCTTGAAAAGGCATACCAAGAGTTACCTGAGAACGTTA
AACACCACAAGTCACGTTTTGAAGTCCCAGGAGCTCTCGTAACTATTGAAGGTAATAAGACTATAATCGA
GAACTTCAAGGATATTGCGGATGCTCTAAACAGAGATCCACAGCACTTGCTCAAGTTCTTGCTTAGAGAA
ATAGCTACAGCTGGAACTCTTGAAGGTAGAAGAGTAGTCCTTCAGGGTAGATTCACGCCATATTTAATAG
CAAACAAGCTAAAGAAGTACATAAAAGAGTATGTTATCTGTCCAGTATGTGGCTCTCCTGATACGAAGAT
AATTAAAAGGGACAGGTTCCACTTCCTTAAGTGTGAAGCTTGTGGTGCAGAAACTCCAATCCAGCATCTC
TAG

> Amorces de clonage

aIF2b-Pfu-EagI 5'-TTTTCGGCCGATTAATTTAAGAAGGAGATATATATATGGAAATTGACTATTACG-3'

aIF2b-Pfu-NotI 5'-TTTTTGCGGCCGCGAGATGCTGGATTGGAGTTTC-3'

MCM (Mini Chromosome Maintenance)

COG1241: Predicted ATPase involved in replication control, Cdc46/Mcm family

Hélicase réplicative

> PF0482

GTGGACAGGGAGGAGATGATTGAGAGATTTGCAAACTTCCTTAGGGAGTATACAGACGAAGATGGTAACC
CAGTATACAGAGGTAAAATAACTGATTTACTTACAATAACACCCAAGAGGTCTGTTGCAATAGACTGGAT
GCACCTAAATTCCTTTGACTCAGAGCTAGCTCATGAAGTTATAGAGAACCCCGAAGAAGGAATAAGTGCC
GCAGAAGATGCAATCCAGATTGTATTACGAGAAGACTTCCAAAGAGAAGACGTGGGAAAAATACACGCAA
GGTTTTATAATTTGCCAGAAACCCTAATGGTCAAAGACATTGGGGCAGAGCACATCAACAAGTTAATTCA
AGTAGAAGGAATCGTGACGAGAGTAGGAGAAATTAAGCCCTTTGTCTCTGTGGCAGTTTTCGTATGTAAG
GACTGCGGTCATGAAATGATAGTGCCTCAAAAACCCTATGAAAGCCTTGAAAAAGTTAAGAAGTGCGAAC
AATGTGGAAGCAAAAATATAGAACTAGATGTTAACAAGAGCTCCTTCGTAAACTTCCAGAGCTTTAGGAT
TCAGGATAGACCAGAAACCCTAAAAGGAGGAGAAATGCCAAGGTTTATCGATGGTATTCTGCTTGATGAC
ATAGTGGATGTAGCCCTCCCAGGAGACAGAGTTATTGTAACAGGAATTTTGAGAGTCGTTCTTGAAAAGA
GAGAGAAAACTCCAATATTTAGAAAAATCCTCGAGGTAAATCACATTGAACCTGTTAGTAAAGAGATACA
AGAATTAGAAATTTCTCCAGAAGAGGAGCAGATAATAAAGGAGCTAGCAAAAAGAAAAGACATAGTAGAT
GCAATAGTTGATTCAATAGCTCCTGCCATATATGGATACAAAGAAGTCAAGAAGGGAATAGCACTTGCCC
TGTTTGGAGGAGTTTCAAGAAAGTTACCTGATGGAACTAGGCTTAGAGGAGATATACATGTCCTCCTGGT
CGGAGACCCAGGAGTTGCAAAGAGCCAGATTTTAAGGTATGTGGCAAACCTCGCTCCTAGGGCCATTTAC
ACTTCAGGAAAAAGTAGTTCCGCAGCAGGTCTTTGTGTAGCTCCCGATTCTTTAGTGGTAGTGAATGACA
AAGTTCAAGAAATAGGAAAGCTAACGGAAGAATGGGGAAGAGAAGTAGGCTTCCTAGAATACTCAAGTGG
GATTTTCTATGCTCCTTACCTGGGAAGAGGAATATCCCTAGATTTAGTAACAGGGAAAGTCAAACCTTCA
GTTGTTAGCAAGGTTTGGAAGTTAAAATCCCCAGAAGAATTAGTTACAATAAAGACCATTACTGGAAAAG
AGATAACAGTAACTCCTGAGACAAAACTTCTGACCTTCAATGGGACACTTGAATGGAAAGAAGCTGGAAA
AATAAAACCTGGAGATTACGTCCTAACGGTTAAAAAGTTACATATCAATGGGAAACAAGAAACTTTAGAT
GAAAAGCTTGCATACAAGCGTGGACTGTCCCTTTCCGATCCTTTGGAGTTCTTTAGTTCATCTGAGAGGA
CAATTTCCGCTTATCTAAAGGGAATATTTGACAAAGTCGGAAGACTCGTCGGAGATACAGCGGTCATTAA
AGTCGATAAAGATATGGCAAAGAGGCTACAGATTTTATTGCTAAGGCTTGGAATAGTTTCCTCGGTAGAT
GAGACAGGAAAAGTCATCATTGGAAGGGAGTACATCCAGAAAATCTTAGGGTACAACGTTAGCGTCGTGA
CCCATGAAGTGGAGCTATTTAGAGAGTTTATAGCTGAAATATCTAAGTTCTATGGAACCAGTGAAGAGGA
TGTCTACAGTTCCCTCCATGAAAAAGGAGAACTCGATATAGGGACAGTTCCAGTAGAGCTCCCAGAGGGC
TTAAGAGAAGAAATAAATCGTGAAAGAGCAACTTACAGTGAACTTGTGAAAATTGCCCAGGAAATAAAAG
ATGAAAAACTCTACAATAAACTTGCGTGGATTTTAAGTGAAGTTACGGAAGAAGAGGCAAAAATTAAGGA
AAAAGTTAACACTCTAAAGGTCATACTCTCCTCAGATTTGATACCAGAAAGAGTAGAATCTGTAAAGATT
ATCAAAAGTCCATACCCCTACGTTTATGACCTTACAGTTGAAGGTTCTCACAGCTTCATAGCAAATGGCT
TTGTAGTCCACAATACTGCTGCAGCAGTTAGGGATGAGTTCACGGGAGGATGGGTTTTGGAGGCGGGAGC
TTTAGTCCTTGCAGATGGGGGTTATGCTCTAATCGACGAGCTCGACAAGATGAGCGACAGGGATAGGAGC
GTGATACATGAAGCCTTAGAACAACAGACAATAAGCATTTCAAAAGCAGGGATTACGGCAACTCTAAACG
CTAGAACTACAGTCATTGCGGCTGCAAATCCGAAACAGGGAAGATTTAATAGAATGAAAAATCCATTCGA
GCAAATTGACCTTCCCCCTACACTTCTAAGTAGATTTGACCTAATATTTGTGTTAATTGATGAGCCCGAT
GACAAAATTGACAGTGAAGTTGCCAGACACATCTTAAGGGTCAGAAGGGGAGAAAGTGAAGTCGTGGCCC
CAAAAATACCTCATGAAATTCTAAGGAAGTACATCGCTTATGCAAGGAAGAATATTCATCCCGTTATAAG
TGAAGAAGCTATGGAAGAGATAGAGAAGTACTATGTGAGAATGAGAAAGAGTGTAAAGAAGACAAAGGGA
GAAGAAGAGGGGATACCACCAATCCCAATAACAGCTAGACAGCTCGAGGCCCTCATTAGATTAAGCGAAG
CTCATGCAAGGATGAGGCTAAGCCCAATAGTAACAAGGGAAGATGCAAGAGAAGCAATAAAACTGATGGA
ATACACGCTAAAGCAAATTGCAATGGATGAGACCGGGCAAATTGACGTGACAATTCTAGAATTAGGTCAG
AGCGCAAGAAAGCTCAGTAAAATAGAGAAAATACTGGATATCATTGAAAAGCTTCAGAAGACCAGCGAAA
GAGGCGCCCACGTTAATGATATCTTAGAAGAAGCAAAGAAAGCAGGAATAGAGAAGCAGGAAGCAAGAGA
AATCCTTGAAAAACTTTTGGAGAAGGGTCAAATATATATGCCAGAGAGTGGTTACTACAAAACCGTC**TGA**

> Amorces de clonage

MCM-Pf-EagI 5′-TTTTCGGCCGATTAATTTAAGAAGGAGATATATATATGGACAGGGAGGAGATGATTG-3′

MCM-Pf-NotI 5′-TTTTTGCGGCCGCGACGGTTTTGTAGTAACCAC-3′

> Amorces pour l'élimination de l'intéine (le fragment d'ADN codant l'intéine est souligné)

PF0482_extein1R	5'-AAGACCTGCTGCGGAACTACTTTTTC-3'
PF0482_extein2F	5'-ACTGCTGCAGCAGTTAGGGATG-3'

Gins23 (Go ichi ni san complex subunit 23)

COG1711: DNA replication initiation complex subunit, GINS family

Sous-unité Gins23 du complexe GINS, homologue archéen des protéines eucaryotes Psf2 et Psf3

> PF0483

ATGTTCACGGGTAAGGTATTGATTCCAGTAAAAGTACTCAAGAAGTTTGAGAATTGGAATGAAGGAGATA
TGATACTGCTAGAAGATTGGAAAGCCAAGGAATTGTGGGAGAGTGGAGTAGTTGAAATAATCGATGAAGC
TGATAAAGTCATAGGAGAGATCGATAGAGTGTTATCAGAAGAAAAGAAAAACCTCCCATTGACTCCAATA
CCAGAGGGACTGTACGAAAAAGCTGAATTTTACATCTATTATCTAGAAAAGTACATCCAAGAGAAGGTCG
ACAACATAGAAACAATACAAACTAAGGTCACAAAGTTAGCAAATCTAAAGAAGAAGTATAAGACTCTGAA
AGAGATAAGATTTAAAAAGATACTAGAGGCTGTGAGGCTTAGACCAAACAGTATGGAAATTCTAGCGAGA
TTATCCCCAGCTGAAAAGAGAATATACCTTGAGATCTCTAAAATAAGGAGAGAGTGGATAGGTGAT**TAG**

> Amorces de clonage

Gins23-Pf-EagI	5'-TTTTCGGCCGATTAATTTAAGAAGGAGATATATATATG TTCACGGGTAAGGTATTG-3'
Gins23-Pf-NotI	5'-TTTTTGCGGCCGCATCACCTATCCACTCTCTC-3'

Gins15 (Go ichi ni san complex subunit 15)

COG1711: DNA replication initiation complex subunit, GINS family

Sous-unité Gins15 du complexe GINS, homologue archéen des protéines eucaryotes Sld5 et Psf1

> PF0982

ATGGATATTGAGGTTCTCAGAAGATTATTGGAGAGAGAACTTTCAAGCGAAGAACTGACTAAAATAGAGG
AAGAATTTTATGACGATTTAGAAAGCTTTAGAAAAGCCTTGGAAATCAATGCCGAGAGACATGAAGAAAG
AGGAGAGGACATTCACAAAAAGCTGTATTTAGCTCAACTATCTTTGGTTAGGAATCTTGTTAGAGAAATA
TTAAGGATTAGGTTGCATAAGATTGTTGATATGGCATTTGAGGGAGTTCCCAGAAATTTAGTTGGAGATG
AAAAGAAAATATACAAGATAATAACAGCTTTCATAAATGGAGAACCTCTTGAAATTGAAACGGCAGGAGA
AGAGAGTATTGAAGTTATTGAAGAGGAAAAAGAAACATCTCCTGGGATAATAGAGGCATATCTTCTTAGA
GTTGATATTCCCAAAATATTGGATGAAAATTTGAGAGAATATGGGCCCTTCAAGGCTGGCGATCTTGTTG
TATTGCCGAAGTCTATTGGCAGGGTACTCATTCAGAGGGATGCCGCGGATAAGGTATTGATACAATTG**TA
A**

> Amorces de clonage

Gins15-Pf-EagI	5'-TTTTCGGCCGATTAATTTAAGAAGGAGATATATATATGGATATTGAGGTTCTCAG-3'
Gins15-Pf-NotI	5'-TTTTTGCGGCCGCCAATTGTATCAATACCTTATC-3'

NOTE :

La phase ouverte de lecture PF0982 présente deux codons d'initiation potentiels (TTG et ATG) séparés par trois triplets. Le gène qui a été amplifié débute au niveau du deuxième codon (ATG).

Le codon TTG se réfère au codon d'initiation retenu dans l'annotation du génome de *Pyrococcus horikoshii*, le premier organisme de l'ordre des Thermococcales dont le génome a été séquencé. Le codon ATG a été retenu dans l'annotation des deux génomes de Thermococcales les plus récents, *P. abyssi* et *Thermococcus kodakaraensis*. Les logiciels de prédiction de gènes ayant fait l'objet de constantes améliorations, le codon ATG a été considéré comme le plus probable et retenu pour le clonage du gène codant la protéine Gins15.

PCNA (Proliferating Cell Nuclear Antigen)

COG0592: DNA polymerase sliding clamp subunit (PCNA homolog)

Facteur de processivité de l'ADN polymérase

> PF0983

ATGCCATTTGAAATCGTATTTGAAGGTGCAAAAGAGTTTGCCCAACTTATAGACACCGCAAGTAAGTTAA
TAGATGAGGCCGCGTTTAAAGTTACAGAAGATGGGATAAGCATGAGGGCCATGGATCCAAGTAGAGTTGT
CCTGATTGACCTAAATCTCCCGTCAAGCATATTTAGCAAATATGAAGTTGTTGAACCAGAAACAATTGGA
GTTAACATGGACCACCTAAAGAAGATCCTAAAGAGAGGTAAAGCAAAGGACACCTTAATACTCAAGAAAG
GAGAGGAAAACTTCTTAGAGATAACAATTCAAGGAACTGCAACAAGAACATTTAGAGTTCCCCTAATAGA
TGTAGAAGAGATGGAAGTTGACCTCCCAGAACTTCCATTCACTGCAAAGGTTGTAGTTCTTGGAGAAGTC
CTAAAAGATGCTGTTAAAGATGCCTCTCTAGTGAGTGACAGCATAAAATTTATTGCCAGGGAAAATGAAT
TTATAATGAAGGCAGAGGGAGAAACCCAGGAAGTTGAGATAAAGCTAACTCTTGAAGATGAGGGATTATT
GGACATCGAGGTTCAAGAGGAGACAAAGAGCGCATATGGAGTCAGCTATCTCTCCGACATGGTTAAAGGA
CTTGGAAAGGCCGATGAAGTTACAATAAAGTTTGGAAATGAAATGCCCATGCAAATGGAGTATTACATTA
GAGATGAAGGAAGACTTACATTCCTACTGGCTCCAAGAGTTGAAGAG**TGA**

> Amorces de clonage

PCNA-Pfu-EagI 5'-TTTTCGGCCGATTAATTTAAGAAGGAGATATATATATGCCATTTGAAATCGTAT-3'

PCNA-Pfu-NotI 5'-TTTTTGCGGCCGCCTCTTCAACTCTTGGAGCCAG-3'

TFS (Transcription Factor S)

COG1594: DNA-directed RNA polymerase, subunit M/Transcription elongation factor TFIIS

Facteur de stimulation de l'activité exonucléase de l'ARN polymérase

> PF0986

ATGGTGAAATTCTGCCCCAAATGTGGGAGTATTATGATACCCGACAGGAGGAGAGGAGTCTTTGTCTGTA
GAAAATGTGGTTATGAGGAACCTATAAATCCTGAGGACACCAAAGCATATAGAAGAACAGAAGAAGTCAA
GCATAGGCCTGATGAAGGAGTAGTTGTAATTGAACAAGAAGTTTCGACTCTTCCAACAGCAAAAGTAACC
TGCCCAAAATGTGGGCATAATGAAGCATGGTGGTGGGAACTTCAAACTAGGGCAGGAGATGAGCCAAGTA
CAATATTCTATAAGTGTAAGAAGTGTGGATACGTATGGAGGAGTTACGAA**TAA**

> Amorces de clonage

TFS-Pfu-EagI	5'-TTTTCGGCCGATTAATTTAAGAAGGAGATATATATATGGTGAAATTCTGCCCC-3'
TFS-Pfu-NotI	5'-TTTTTGCGGCCGCTTCGTAACTCCTCCATACGTATC-3'

aIF-2 alpha (archaeal Initiation Factor 2 subunit alpha)

COG1093: Translation initiation factor 2, alpha subunit (eIF-2alpha)

Sous-unité alpha du facteur d'initiation de la traduction aIF-2

> PF1140

ATGCCCAGGAGAGCAAGGGAGTATCCCGAAGAGGGAGAGCTTGTGGTTGCTACAGTTAAGAGGGTTCACA
ATTATGGAGCATTTCTTGACCTTGACGAGTATCCGGGGAAAGAAGGGTTTATGCACATTAGTGAGGTGGC
CTCAACCTGGGTAAAGAACATTAGAGATTACCTAAGGGAAGGTCAAAAGGTAGTGGCCAAGGTTATAAGG
GTTGACCCAAAGAAGGGTCATATAGACCTCAGCCTAAGGAGGGTAACCCAGCAACAGAGAAAAGCAAAGC
TTCAAGAGTTCAAGAGGGCTCAGAAAGCTGAAAACTTACTAAAGCTTGCCGCGGAGAAATTAGGAAAAGA
CTTTGAAGAGGCTTGGAGAGAGGTATGGGTGCCACTGGAGAACGAGTGGGGCGAGGTTTATGCTGCCTTC
GAAGATGCAGCGAGGAATGGAATAGAAGTTCTCAAAGGCTACGTTCCAGATGAGTGGCTTCCAGTTCTCA
AGGAGATAATTGACAGCTATGTAGAGGTTCCTACTGTAACAATAGATGCCGAGTTTGAAATCACTGTGCC
AAAGCCAAATGGAATTGAGATAATTAAGGAGGCTTTAATAAAGGCAAGGGACAGAGCAAACCAAGAGAAA
GACATCGAAGTTAAGTTCACCTACCTGGGGGCTCCAAGGTACAGGATAGACATTACGGCCCCAGACTACT
ACAAGGCCGAAGAAGTCCTGGAGGATATAGCAGAAGAAATTCTTAGAGTAATAAAGGAGGCTGGTGGCGA
GGCCACACTTCTGAGAAAGGAGAAGAGAATTAGGAAGGTTAAGAAGAGGAAGAAG**TGA**

> Amorces de clonage

aIF2a-Pfu-EagI	5'-TTTTCGGCCGATTAATTTAAGAAGGAGATATATATATGCCCAGGAGAGCAAGG-3'
aIF2a-Pfu-NotI	5'-TTTTTGCGGCCGCCTTCTTCCTCTTCTTAACCTTC-3'

Nop10 (Nucleolar protein 10)

COG2260: Predicted Zn-ribbon RNA-binding protein

Sous-unité protéique du complexe H/ACA, impliqué dans la maturation des ribosomes (pseudouridylation)

> PF1141

ATGAGGTTTAGGATAAGGAAGTGTCCCAAGTGTGGGAGATACACCCTCAAGGAAGTCTGCCCTGTGTGTG
GGGAAAAGACTAAAGTAGCTCATCCACCAAGGTTTTCCCCTGAGGACCCATATGGGGAATATAGGAGAAG
GTGGAAGAGGGAGGTACTCGGAATAGGGAGGAAGGAAAAA**TGA**

> Amorces de clonage

Nop10-Pfu-EagI	5'-TTTTCGGCCGATTAATTTAAGAAGGAGATATATATATGAGGTTTAGGATAAG-3'
Nop10-Pfu-NotI	5'-TTTTTGCGGCCGCTTTTTCCTTCCTCCCTATTCCG-3'

COG2047

COG2047: Uncharacterized protein (ATP-grasp superfamily)

Protéine de fonction inconnue

> PF1142

GTGGAAGAGGGAGGTACTCGGAATAGGGAGGAAGGAAAAATGAAGGAAACCACAATTGTTGTATATGAAA
GGCCCGATATTTATGACCCAATATTTATTGAGGGCCTTCCTGGGATTGGATTAGTTGGAAAGCTTGCAGC
TGAACACTTAATTCAAGAGCTAAAGGCCAAAAAGTTTGCCGAGCTTTACTCTCCTCACTTCATGCACCAG
GTCTTAATTAGAAAGAACTCAGTCGTAGAGCTAATGAAAAACGAATTCTACTACTGGAAAAGCCCTGACG
ATGAGCATAGAGATTTGATAATAGTGACTGGAGATACTCAAGTTCCTCCAACGGACAGCTACGGACACTT
TGAGGTTGCTGGGAAGATGCTTGACTTCGTTCAAGAGTTTGGGACTAGAGAAATAATAACGATGGGAGGC
TATCAAGTTCCTGAAATCCAAGGAGAGCCGAGGGTTCTTGCAGCTGTGACCCATGAGGACTTAATAGAGT
ACTACAAGAGCAAGCTGGAGGGCTGTTCAGTTGAGGTAATTTGGAGAGAAGATGAGGGAGGGGCTATAGT
AGGTGCAGCAGGGCTCCTCCTGGGAATTGGTAAGCTTAGAGGAATGTTCGGCATAAGCTTGCTCGGAGAG
AGCCTTGGATATATAGTTGATGCAAAGGCCGCGAAGGCTGTCCTTTCTGCAGTCACAAAGATACTCGGAC
TGGAGATAGACATGACCGCTTTAGATGAGAGGGCTAAGGAGACTGAGGAGATCTTGAGAAAAGTTGAAGA
AATGCAGAGGGCAATGATGGAGCAAGTCACTCCAAAGTTGCCTCATGAAGAGGAAGACAGGGGATACCTC
TAA

> Amorces de clonage

COG2047-Pfu-EagI	5′-TTTTCGGCCGATTAATTTAAGAAGGAGATATATATATGGAAGAGGGAGGTAC-3′
COG2047-Pfu-NotI	5′-TTTTTGCGGCCGCGAGGTATCCCCTGTCTTCC-3′

Sir2 (Silent information regulator 2)

COG0846: NAD-dependent protein deacetylase, SIR2 family

Protéine déacétylase appartenant à la famille des sirtuins

> PF1154

ATGCCATTTCAACCCTCTAAATATGAGCTCACCGTAAACTTGGTGCAAGGCAATTACATGTTGCATGTTC
TATCATTAAAGGGTGAATATAAAAATATTAGCCCAAAAGAAATTTTAAGGTATCCTTCTACTAATTCTCT
GATGCTAGGTGAAGTGAGCAAAATCTTAGCAAAATCTTCTATGGCAATAGCATTTACTGGAGCGGGCATT
AGTGCTGAGAGTGGTATCCCCACTTTTAGGGGAAAAGATGGACTATGGAGAAAGTATAGGGCTGAGGAGC
TAGCTACTCCTGAAGCTTTTAAGAGAGATCCAAAGCTTGTATGGGAGTTCTACAAGTGGAGAATCAAGAA
GATTTTGGAAGCGAAGCCAAATCCTGCTCACATAGCTTTGGCAGAGCTGGAGAAGATGGGAATAATAAAG
GCCGTGATTACCCAAAACGTTGATGATCTTCACAGGGAAGCAGGGAGCAAGAATGTTATAGAGTTGCACG
GAAATATATTTCGAGTTAAATGCACGAGTTGCTCATATAGAGAATACTTGAAGGAAAGTGACAGAATTGG
GTGGCTACTTTCCCAAGAACTACCTAGGTGTCCCAAGTGTGGCTCTCTTCTAAGGCCCGATGTAGTTTGG
TTTGGCGAGGCTTTGCCCGAAAAAGAGCTAACAACAGCGTTTTCACTAGCTAAAAAAGCTGATGTTGTTC
TCGTTGTAGGAACTAGTGGCGTCGTATACCCTGCCGCATACATACCATATATAGTTAAGGAGAGCGGGGG
TATTGTCGTTGAGATAAACATTGAACCCTCTGCAATCACTCCGATAGCGGACTTCTTCCTTAGGGGCAAA
GCTGGAGAAGTTTTGCCAAAGTTGGTTGAGGAGATCAGGAGGATTTCTAAA**TGA**

> Amorces de clonage

Sir2-Pfu-EagI	5'-TTTTCGGCCGATTAATTTAAGAAGGAGATATATATATGCCATTTCAACCCTCTA-3'
Sir2-Pfu-NotI	5'-TTTTTGCGGCCGCTTTAGAAATCCTCCTGATCTCC-3'

NudF

COG1051: ADP-ribose pyrophosphatase

Hydrolase de la superfamille Nudix (Nucleoside diphosphate linked to another moiety X)

> PF1590

ATGGATAGGTATGTGCTCCTTGTGAAGGCTCCAAAGGAATATGATATTTCAAAATTTAGAGAGGAAGTTG
AAAAAATAGCAAAAGATTATGGGCTTAAAGTAGAGGCACACAAGTGCATTGGAGTTACAGTAGACATAGT
GATTATTCACAATGGCAACATTGTCCTCATAGAGAGGAAAAACGACCCATACAAGGGATATTTGGCACTC
CCAGGAGGCTTTGTTGAGTATGGGGAGAAGGTGGAAGAAGCAGCAATAAGAGAAGCAAAAGAAGAAACTG
GATTAGATGTTAAGTTGTTGAGAGTTGTAGGAGTTTATTCAGATCCCAATCGAGATCCCAGGGGTCATAC
AATAACGGTTGCGTTTTTGGCCATTGGCCTGGGAGAACCAAAGGCTGGAGATGATGCAAAGAAAGTTCAC
CTAATCCCAATTGAAGAAATTGAAAAAATAAAAGCAAAGCTGGCTTTTGACCACGCTAAGATTATAGAAG
ATGCCTTAACGCTAAGA**TAA**

> Amorces de clonage

NudF-Pfu-EagI	5′-TTTTCGGCCGATTAATTTAAGAAGGAGATATATATATGGATAGGTATGTGCTC-3′
NudF-Pfu-NotI	5′-TTTTTGCGGCCGCTCTTAGCGTTAAGGCATCTTC-3′

RecJ (Recombination J)

COG0608: Single-stranded DNA-specific exonuclease RecJ

Exonucléase spécifique de l'ADN simple brin

> PF2055

ATGGATAAGGAGGGTTTTTTGAACAAGGTTAGGGAGGCTGTGGATGTAGTAAAGCTCCACATCGAGTTAG
GTCATACTATAAGGATAATCTCTCATAGGGATGCGGATGGAATAACCTCTGCGGCAATTCTTGCAAAGGC
TTTGGGAAGAGAAGGAGCGAGCTTTCACATTTCGATTGTTAAACAGGTAAGTGAAGATCTTTTAAGAGAA
TTAAAGGATGAAGATTACAAAATCTTCATTTTTTCCGACCTGGGTAGTGGTTCTTTAAGTTTGATAAAAG
AGTATCTTAAGGAAAAAACTGTTATAATCCTTGATCACCATCCTCCGGAAAATGTGAAGTTGGAAGAAAA
GCATATACTTGTTAATCCAGTTCAATTTGGCGCAAATAGCGTTAGGGATCTGAGTGGATCTGGGGTTACA
TACTTCTTTGCAAGGGAGCTAAATGAAAAGAATAGGGACCTTGCTTACATTGCAATAGTGGGAGCAGTTG
GGGATATGCAAGAGAACGATGGAGTTTTCCATGGGATGAACCTTGATATTATTGAAGATGGGAAATCTCT
GGGAATTCTTGAGGTTAAAAAAGAATTGCGCCTGTTTGGTAGGGAAACTAGACCTCTCTATCAAATGCTC
GCATATGCCACAAATCCGGAAATTCCTGAAGTTACTGGAGACGAGAGGAAGGCCATAGAGTGGTTAAAGA
ACAAGGGCTTCAATCCCGAGAAAAAATATTGGGAATTAAGTGAGGAGGAAAAGAAAAAGTTACATGATTT
CCTAATCATTCACATGATCAAGCATGGAGCTGGAAAAGAGGATATAGATAGGCTAATAGGAGACGTTGTT
ATTAGTCCCTTATATCCTGAAGGGGATCCCAGGCACGAGGCTAGAGAATTTGCTACCCTATTAAACGCTA
CAGGCAGGTTAAACTTGGGCAACTTAGGAGTGGCTGTATGTTTGGGAGATGAGGAGGCTTTCAGAAAGGC
CCTAAAGATGGTTGAAGACTACAAGAGGGAGCAAATTGAAGCAAGAAAGTGGCTACTTCAAAATTGGAAC
AGTGAAGTTTGGGAGGGGGATCATGTTTACGTCTTATATGTGGGAAAGAGTATTAGAGATACTCTCGTTG
GAATAGCAGCTAGCATGGCCATCAATGCTGGACTGGCAGATCCTGAAAAGCCGGTTATAGTGTTTGCAGA
TACTGATGAAGATCCAAACCTTCTCAAAGGTTCAGCTAGAACAACTGAAAGGGCTTTAGCTAAGGGTTAC
AATTTGGGAGAAGCTCTTAGGAAAGCGGCTGAGCTAGTGAATGGGGAAGGGGGAGGACACGCGATAGCTG
CAGGTATAAGAATTCCCAGGGCCAGGTTGGCGGAGTTTAGAAAATTAATAGATAAAATCCTTGGAGAACA
GGTGAGCAAAGGTGGAGATAAAAGCGAAAGC**TGA**

> Amorces de clonage

RecJ-Pfu-EagI 5'-TTTTCGGCCGATTAATTTAAGAAGGAGATATATATATGGATAAGGAGGGTTTTTTG-3'

RecJ-Pfu-NotI 5'-TTTTTGCGGCCGCGCTTTCGCTTTTATCTCCACC-3'