Phylogénie et évolution des Archaea, une approche phylogénomique

Celine Petitjean

Thèse de Doctorat de l'Université Paris-Sud

École Doctorale Gènes, Génomes, Cellules (ED 426)

Présentée par :

Céline Petitjean

Pour l'obtention du grade de :

Docteur ès Sciences de l'Université Paris-Sud

Phylogénie et évolution des Archaea, une approche phylogénomique

Thèse soutenue le 27 septembre 2013, devant le jury composé de :

Simonetta GRIBALDO	**Rapporteur**
Vincent DAUBIN	**Rapporteur**
Pierre CAPY	**Examinateur**
Olivier LESPINET	**Examinateur**
David MOREIRA	**Co-directeur de thèse**
Céline BROCHIER-ARMANET	**Co-directeur de thèse**

Résumé

En 1977, Carl Woese sépare les procaryotes en deux grands groupes en proposant une nouvelle classification basée sur des critères phylogénétiques. Les Archaea deviennent ainsi un domaine à part entière aux cotés des Bacteria et des Eucarya. Depuis, la compréhension de ce nouveau groupe et de ses relations avec les deux autres domaines, essentielles pour comprendre l'évolution ancienne du vivant, est largement passée par l'étude de leur phylogénie. Presque 40 ans de recherche sur les archées ont permis de faire évoluer leur image : de bactéries vivant dans des milieux spécialisés, souvent extrêmes, on est passé à un domaine indépendant, très diversifié aussi bien génétiquement, métaboliquement ou encore écologiquement. Ces dernières années la barre symbolique de cent génomes complets d'archées séquencés a été franchie et, parallèlement, les projets génomiques et métagénomiques sur des groupes peu caractérisés ou de nouvelles lignées de haut rang taxonomique (e.g. Nanohaloarchaea, Thaumarchaeota, ARMAN, Aigarchaeota, groupe MGC, groupe II des Euryarchaeota, etc.) se sont multipliés. Tout ceci apporte un matériel sans précédent pour l'étude de l'histoire évolutive et de la diversité des Archaea. Les protéines ribosomiques ont été utilisées de façon courante pour inférer la position phylogénétique des nouvelles lignées d'Archaea. Néanmoins, les phylogénies résultantes ne sont pas complètement résolues, laissant des interrogations concernant d'importantes relations de parenté. La recherche de nouveaux marqueurs est donc cruciale et c'est dans ce contexte que mon projet de thèse s'inscrit.

À partir de l'analyse des génomes de deux Thaumarchaeota et d'une Aigarchaeota, nous avons identifié 200 protéines conservées et bien représentées dans les différents phyla d'archées. Ces protéines sont impliquées dans de nombreux processus cellulaires, ce qui peut apporter un signal phylogénétique complémentaire à celui des marqueurs de type informationnel utilisés par le passé. En plus de confirmer la plupart des relations phylogénétiques inférées à partir de ces derniers (i.e., protéines ribosomiques et sous unités de l'ARN polymérase), l'analyse phylogénétique de ces nouveaux marqueurs apporte un signal permettant une meilleure résolution de la phylogénie des archées et la clarification de certaines relations jusqu'ici confuses.

Un certain nombre de ces nouveaux marqueurs sont aussi présents chez les bactéries. Les relations entre les grands phyla d'archées restant encore non résolues, nous avons utilisé ces protéines pour essayer de placer la racine de l'arbre des Archaea en utilisant comme groupe extérieur les bactéries. Nous avons ainsi pu identifier 38 protéines, parmi les 200 sélectionnées précédemment, ayant un signal phylogénétique suffisamment fiable pour cette étude, auxquelles nous avons ajouté 32 protéines ribosomiques universelles. L'utilisation conjointe de ces données nous a permis de placer la racine entre les Euryarchaeota, d'une part, et un groupe rassemblant les Thaumarchaeota, les Aigarchaeota, les Korarchaeota et les Crenarchaeota, d'autre part. Ce nouvel éclairage sur l'évolution ancienne des archées nous a amené à proposer une révision de leur taxonomie avec, principalement, la création du nouveau phylum "Proteoarchaeota" contenant les quatre phyla actuels que nous proposons de rétrograder en classes : Thaumarchaea, Aigarchaea, Korarchaea et Crenarchaea.

Finalement, l'analyse des protéines codées dans les trois génomes qui ont servi de point de départ de ma thèse nous a permis de générer une masse considérable de données qui ont révélé des traits particuliers ou encore des histoires évolutives inattendues. Un exemple est l'histoire du complexe formé par la chaperonne DnaK et de ses co-chaperonnes GrpE, DnaJ, et DnaJ-Fer chez les Thaumarchaeota, impliquant plusieurs transferts horizontaux entre les trois domaines du vivant.

Abstract

In 1977, Carl Woese proposed a new classification of organisms based on phylogenetic criteria where he divided prokaryotes into two major groups. Thus, Archaea were defined as a new domain, together with Bacteria and Eucarya. Since then, the study of this group and its relationships with the two other domains, essential to understand the early evolution of Life, has been largely done through the investigation of its phylogeny. Almost 40 years of research on the archaea have led to a significant evolution of the knowledge on this group: from considering them as bacteria living in specialized environments, most often extreme ones, to defining them as an independent domain, highly diversified in genetic, metabolic and ecological terms. During the last years, the symbolic barrier of 100 complete archaeal genome sequences has been reached and, simultaneously, many genome projects from poorly-known groups or new high-rank lineages (e.g., Nanohaloarchaea, Thaumarchaeota, ARMAN, Aigarchaeota, MGC, group II Euryarchaeota, etc.) have been launched. All this provides unprecedented information to study the evolutionary history of Archaea. Ribosomal proteins have been used recurrently to infer the phylogenetic position of new archaeal lineages. Nevertheless, the resulting phylogenies are not fully resolved and several important nodes remain uncertain. The identification of new phylogenetic markers is therefore crucial. This represents the framework of my PhD thesis project.

On the basis of the analysis of the genome sequences of two Thaumarchaeota and one Aigarchaeota, we have identified 200 conserved proteins well represented among the different archaeal phyla. These proteins are involved in a number of cellular functions, thus providing a phylogenetic signal complementary to the one obtained from the informational proteins (i.e., ribosomal proteins and RNA polymerase subunits). The phylogenetic analysis of these new markers has led to a better resolution of the archaeal phylogeny, including several relationships that remained unclear.

Several of the new markers are also present in bacteria. Since the relationships among the different archaeal phyla are not yet resolved, we have used those markers to try to place the root of the archaeal phylogeny using the bacterial sequences as outgroup. We have identified 38 proteins among the 200 detected before containing a phylogenetic signal useful for that purpose, to which we have added 32 universal ribosomal proteins. The use of this complete dataset allowed us locating the root between the Euryarchaeota and a large group joining the Thaumarchaeota, Aigarchaeota, Korarchaeota and Crenarchaeota. This new result on the ancient evolutionary history of Archaea has led us to propose a taxonomic revision for this domain, in particular the erection of a new phylum "Proteoarchaeota", containing the current four phyla that we propose to retrograde into classes (Thaumarchaeales, Aigarchaeales, Korarchaeales and Crenarchaeales).

Finally, the analysis of the proteins encoded by the three reference genomes at the origin of this work has generated a large amount of data, which reveals particular traits in certain organisms or unexpected evolutionary histories. One example concerns the evolution in Thaumarchaeota of the protein complex composed of the DnaK chaperon and its co-chaperons GrpE, DnaJ, and DnaJ-Fer, which involves several horizontal gene transfer events among the three domains of Life.

Remerciements

Merci aux membres de mon jury, Simonetta Gribaldo, Vincent Daubin, Pierre Capy et Olivier Lespinet d'avoir accepté d'évaluer mon travail de thèse.

Merci à mes deux directeurs de thèse, David Moreira et Céline Brochier-Armanet de m'avoir fait confiance et de m'avoir permis de mener à bien ce travail.

Merci à Purificación López-García de m'avoir fait confiance et accepté dans son équipe.

Merci à Simonetta Gribaldo de m'avoir permis et poussée à entreprendre cette thèse.

Merci à l'Agence Nationale de la Recherche d'avoir financé ma thèse au travers du projet EVOLDEEP et de l'Investissement d'avenir ANCESTROME pour l'accès à certains clusters de calculs ; à l'Université Paris-sud pour l'accès au cluster de calcul ebio.

Merci à l'École Doctorale GGC de m'avoir acceptée malgré la complexité de mon dossier.

Au début de mon stage de M2, Simonetta m'a dis un jour « tu verras, bientôt le labo sera ta deuxième maison *» ; je ne soupçonnais pas à quel point cela serai vrai.*

Alors merci aussi à :

Simonetta pour ton encadrement, ta confiance et ta passion pour la science et la phylogénie, et aussi de t'être battue pour moi et de m'avoir poussée à rejoindre finalement David et Céline pour ma thèse.

Alexis pour toute l'aide que tu m'as apportée pendant mon M2 et après, et pour toutes nos discussions. Elie, d'avoir été un point de repère en tant que thésard et pour ton aide dans des moments critiques de fin de stage.

Céline, de m'avoir acceptée en thèse, de m'avoir fait confiance. Merci aussi de m'avoir fait partager ta passion, de ta rigueur scientifique, et de ta bienveillance, particulièrement dans des moments compliqués.

Laura, ma co-thésarde du LCB, pour tout ce qu'on a pu partager, scientifiquement et personnellement ; d'être là tout simplement. Rym, pour toutes tes questions, ta passion et notre amitié. Rémi, Sandrine, Boyang de l'équipe GEB, Mélodie, ma stagiaire, et aux membres du LCB, de tous nos échanges.

Toute l'équipe microbio de l'ESE, de votre passion scientifique, d'être aussi soudés et de pouvoir partager autant avec vous.

David M., de m'avoir acceptée en thèse, de m'avoir fait confiance, de partager ton incroyable savoir, de la justesse de ton encadrement, de ta bienveillance, et pour ton humour caustique.

Purificación, de ta droiture, de tes coup de gueules, de toutes nos discussions, d'avoir su créer avec David cette équipe, de ta confiance, et particulièrement, de m'avoir poussé à raciner l'arbre des archées. Philippe, pour toute ton aide, de ta capacité à saisir certaines choses, de ton humour, d'être passé à Linux ; et Hélène, de ta présence autour de notre équipe. Bienvenue à Juliette Rusticule. Ludwig et Paola, de votre amitié et de votre présence. Marie, coloc et co-bureau, de ton amitié, d'être la première thésarde de l'équipe microbio et toujours présente. Charles, mon « grand frère » du labo, d'être aussi adorable, de tes talents d'imitateur, et de ta thèse qui m'a bien inspirée. Estelle, co-bureau et voisine, de ton soutien, de ton aide et de ce qu'on a partagé du coté d'Alesia.

Jonathan, pour répondre à tes remerciements, te supporter a été un immense plaisir et une grande richesse, au labo, à la maison et le reste du temps, mais aussi depuis bientôt 10 ans ; de toutes nos discussions enflammées ou calmes, de ton humour de ton soutien et de tout ce pourquoi je ne trouverai pas les mots, merci. Marianne, de ton aide et de ta gentillesse, de tous les CMD. Je n'ai aucun doute sur le fait que tu finiras ta thèse avec panache. Aurélien, je ne doute pas que ta thèse sera une belle aventure.
Julien, mon co-bureau direct, de toutes nos discussions sur la phylogénie et sur la vie, et pour mendeley. Tous les membres du bureau 208 et de la salle T pour nos passionnantes discussions, à 5h et le reste du temps. Pour ceux que je n'ai pas encore nommés, Boris et Yann ; Vincent, passé rapidement. Aux autres membres de l'ESE, si intéressants à côtoyer ; particulièrement Gwendal, Jacqui, Amandine, Hervé, Lucie, Alodie, Martha… c'est un vrai plaisir que de travailler dans cet environnement.

Aux membres du LBBE, croisés trop peu souvent, mais toujours avec plaisir.

Aux secrétaires, Vanessa, Nathalie et Emmanuelle, qui m'ont largement aidée dans mes nombreux problèmes administratifs.

Tous les profs qui m'ont apporté savoirs, savoir-faire et méthodes. Particulièrement mes profs de bio de collège et lycée, Philippe Kachidian, Pascal Hingamp, Keith Dudley, Michel Termier, Daniel Gautheret, Olivier Lespinet, Dominique De Vienne, Fabrice Confalonieri et Pierre Capy de m'avoir encouragé à poursuivre. De même que Patrick Forterre, Eduardo Rocha et Pierre-Henri Gouyon pour les discussions partagées. Emese et Carl, ça a été un plaisir d'enseigner à vos côté. Pierre, de tes encouragement, de ta disponibilité, de ta bienveillance et de m'avoir accepté au sein de GGC. Olivier et Pascal, mes « parrains », d'avoir accepté ce rôle. Pascal, de ta compréhension et de ta disponibilité quand j'en ai eu besoin.

Marie-Françoise et Patrick, de m'avoir toujours encouragée et soutenue dans mes choix. Aurore et Bérangère, mes merveilleuses petites sœurs, Davi mon cousin adoré, Isabelle et Martine pour être là, toujours.
Mes amis, tous, depuis longtemps et de tous les côtés, qui m'aide à avancer jour après jour. Parmi eux, Claire, David et Doris si importants où que vous soyez ; mes colocs actuels Gregory et Virginie, qui m'avez offert une belle mise au vert pour ma rédaction ; Sarah et Sylvain, toujours là; et tout ceux que je ne peux pas nommer ici mais qui n'en restent pas moins essentiels.
L'AMZ, l'ESP, Cannes et Dragons et Danielle Pauly, & Cie. de m'avoir permis de décompresser et de tenir le coup !

Aux anciens thésards dont j'ai relu les thèses, et à tous les relecteurs de la mienne, Bérangère, David, Jonathan, Laura, Marie, Marianne et Sébastien, MERCI, avec une mention spéciale à Sébastien et Marianne pour en avoir relu la quasi-totalité.

Bestiolus, d'avoir réalisé cette thèse avec moi, sans bug majeur en trois ans.

Sébastien, pour ton incommensurable soutien ces derniers temps, pour ton aide précieuse pour la toute fin, pour ta patience, et pour tout ce que je ne sais comment exprimer. Je ne m'étends pas plus, je mets mon chapeau bleu, et j'arrive.

En hommage posthume à Carl Woese (1928- 2012)

« *Tel est le pouvoir des lettres quand seulement l'ordre en est changé* »

Lucrèce, *De rerum natura*

Sommaire

Table des figures

Introduction

A. Histoire de la découverte des Archaea

1. Les trois domaines du vivant

a. Un marqueur moléculaire pour la classification bactérienne : l'ARN ribosomique

Au début des années 1970, suivant la proposition de Darwin en 1859 (Darwin 1859), classer les organismes vivants selon leur généalogie, puis celle de Zuckerkandl et Pauling en 1965 (Zuckerkandl and Pauling 1965), utiliser les molécules portant l'information génétique pour l'inférence phylogénétique (à savoir les gènes ou les molécules issues de l'expression des gènes (protéines et ARN)), Woese a utilisé les ARN ribosomiques afin de retracer les relations phylogénétiques entre organismes vivants. Des mesures de distances phylogénétiques avaient déjà été construites sur la base de petits polypeptides tels que les hémoglobines ou le cytochrome c, mais pas sur l'ensemble du vivant (Fitch and Margoliash 1967), et à partir de mesures d'hybridation des gènes codant les ARN ribosomiques mais de façon peu précise (Moore and McCarthy 1967; B. Pace and Campbell 1971). En trouvant un moyen d'analyser la séquence assez longue des ARN ribosomiques, Woese a pu appliquer sa méthode d'analyse sur tout organisme vivant, le ribosome et ses composés protéiques et nucléiques étant homologues dans l'ensemble des organismes connus. Pour cela, il a utilisé en premier lieu les molécules d'ARN 5S de la grande sous-unité du ribosome (ARNr 5S) digérées par deux nucléases (nucléase T1 et une nucléase pancréatique). Les oligonucléotides résultants pouvant être séquencés, il a ensuite comparé les séquences et les fréquences de ces oligomères pour définir une distance phylogénétique, « PD value » (« Phylogenetic Distance value »), entre les différents organismes étudiés (Sogin, Sogin, and Woese 1972). Cette distance phylogénétique est un pourcentage correspondant au nombre de changements de bases nécessaires pour passer du catalogue d'oligonucléotides d'une espèce à celui d'une autre espèce. C'est une forme d'approximation du nombre de substitutions entre les séquences de deux organismes. Les premières analyses ont été conduites sur quelques bactéries et ont permis de mettre en évidence l'intérêt de cette méthode. En effet, la distance phylogénétique ainsi mesurée est alors inférieure à 15% entre deux individus d'une même famille, par exemple entre deux entérobactéries, *Escherichia coli* et *Proteus mirabilis* (Sogin, Sogin, and Woese 1972; Carl R Woese et al. 1976). L'intérêt principal de cette méthode était la possibilité de mesurer la distance évolutive entre espèces bactériennes d'une manière objective. Auparavant, seuls des critères phénotypiques (morphologie, tolérance à l'oxygène, capacité à utiliser le glucose, etc.) étaient employés pour

distinguer les espèces bactériennes, mais les microbiologistes étaient bien conscients du fait que les classifications ainsi obtenues n'étaient pas "naturelles", i. e. elles ne reflétaient pas les relations évolutives entre espèces. Ceci avait amené le célèbre microbiologiste américain Roger Stanier à déclarer avec pessimisme que « *the ultimate scientific goal of biological classification cannot be achieved in the case of bacteria* » (Stanier, R. Y., Doudoroff, M. & Adelberg 1963). Le travail de Woese prouva le contraire.

b. Premières phylogénies moléculaires microbiennes et découverte des « Archaebacteria »

En 1977, Woese et ses collaborateurs analysent l'ARN ribosomique de la petite sous-unité du ribosome (ARNr SSU), 16S chez les bactéries et 18S chez les eucaryotes, et calculent un coefficient d'association entre deux catalogues d'oligonucléotides A et B : le S_{AB}. Il correspond à deux fois le nombre d'oligonucléotides (hexamères ou plus longs) communs entre les deux catalogues sur la somme des oligonucléotides totaux des deux catalogues. Ainsi plus deux séquences sont proches, plus le nombre d'oligonucléotides en commun est important et plus le S_{AB} sera élevé (George E Fox et al. 1977; G. E. Fox, Pechman, and Woese 1977) (Figure 1).

Table 1. Association coefficients (S_{AB}) between representative members of the three primary kingdoms

	1	2	3	4	5	6	7	8	9	10	11	12	13
1. *Saccharomyces cerevisiae*, 18S	—	0.29	0.33	0.05	0.06	0.08	0.09	0.11	0.08	0.11	0.11	0.08	0.08
2. *Lemna minor*, 18S	0.29	—	0.36	0.10	0.05	0.06	0.10	0.09	0.11	0.10	0.10	0.13	0.07
3. L cell, 18S	0.33	0.36	—	0.06	0.06	0.07	0.07	0.09	0.06	0.10	0.10	0.09	0.07
4. *Escherichia coli*	0.05	0.10	0.06	—	0.24	0.25	0.28	0.26	0.21	0.11	0.12	0.07	0.12
5. *Chlorobium vibrioforme*	0.06	0.05	0.06	0.24	—	0.22	0.22	0.20	0.19	0.06	0.07	0.06	0.09
6. *Bacillus firmus*	0.08	0.06	0.07	0.25	0.22	—	0.34	0.26	0.20	0.11	0.13	0.06	0.12
7. *Corynebacterium diphtheriae*	0.09	0.10	0.07	0.28	0.22	0.34	—	0.23	0.21	0.12	0.12	0.09	0.10
8. *Aphanocapsa* 6714	0.11	0.09	0.09	0.26	0.20	0.26	0.23	—	0.31	0.11	0.11	0.10	0.10
9. Chloroplast (*Lemna*)	0.08	0.11	0.06	0.21	0.19	0.20	0.21	0.31	—	0.14	0.12	0.10	0.12
10. *Methanobacterium thermoautotrophicum*	0.11	0.10	0.10	0.11	0.06	0.11	0.12	0.11	0.14	—	0.51	0.25	0.30
11. *M. ruminantium* strain M-1	0.11	0.10	0.10	0.12	0.07	0.13	0.12	0.11	0.12	0.51	—	0.25	0.24
12. *Methanobacterium* sp., Cariaco isolate JR-1	0.08	0.13	0.09	0.07	0.06	0.06	0.09	0.10	0.10	0.25	0.25	—	0.32
13. *Methanosarcina barkeri*	0.08	0.07	0.07	0.12	0.09	0.12	0.10	0.10	0.12	0.30	0.24	0.32	—

Figure 1. Table représentant les coefficients d'association (SAB) entre 13 représentants des trois royaumes primaires. Adapté de l'article de 1977 de Woese et Fox, « Phylogenetic structure of the prokaryotic domain: the primary kingdoms »(C R Woese and Fox 1977). Les organismes 1, 2 et 3 sont des eucaryotes, 4 à 9 des bactéries et de 10 à 13 des archées. Les SAB entre organismes du même groupe sont tous supérieurs à 0,19 (en rouge pour les eucaryotes, bleu pour les bactéries et vert pour les archées) alors qu'entre organismes de deux groupes différents, la valeur maximale est de 0,14.

Les résultats de S_{AB} calculés pour chaque couple d'espèces donnent une matrice de caractères numériques, utilisable pour la construction d'un dendrogramme représentant les relations phylogénétiques entre ces espèces. Après avoir analysé l'ARNr SSU de quelques bactéries

seulement, Fox et Woese décident de conduire une étude à plus large échelle, sur 40 bactéries représentant la diversité disponible à ce moment là (Balch et al. 1977). Le résultat est surprenant pour les auteurs, car il montre une divergence bien plus forte entre les deux « bactéries » méthanogènes (*Methanobacterium ruminantium* et *Methanobacterium thermoautotrophicum*) et le reste des bactéries analysées, qu'entre ces dernières. Ils proposent alors qu'une divergence très ancienne a donné naissance à deux lignées bien distinctes, l'une contenant les bactéries méthanogènes, l'autre contenant toutes les autres bactéries. La même année, Woese propose une nouvelle classification des procaryotes (C R Woese and Fox 1977) en créant la notion de « royaume primaire » ou « royaume », qui se réfère aux premières unités phylogénétiques dans l'arbre du vivant, donc les premiers niveaux de classification à utiliser. Ce nouveau mode de classification se place en opposition aux classifications utilisées à ce moment là, à savoir la dichotomie entre procaryotes et eucaryotes et la classification de Whittaker en cinq règnes (Animaux, Plantes, Champignons, Protistes et Monères), qui ne sont pas basées sur des critères phylogénétiques (Whittaker 1969). Fox et Woese vont définir trois royaumes, les « Eubacteria », les « Archaebacteria » et les « Urcaryotes ». Pour la première fois, les microbiologistes disposent d'une classification basée sur un critère objectif et phylogénétique, en accord avec la théorie de l'évolution telle que décrite par Darwin dans son livre *On the Origin of Species* (Darwin 1859), et c'est en cela qu'elle revêt une importance toute particulière.

Dans cette première classification les « archaebactéries » étaient composées uniquement des méthanogènes, puisqu'aucun autre ARNr SSU d'archée n'avait encore été étudié. Néanmoins, Fox et Woese proposent que d'autres bactéries puissent aussi être des « archaebactéries » : certaines bactéries halophiles extrêmes dont les parois cellulaires ne contiennent pas de peptidoglycane, comme chez les méthanogènes. Au cours des années suivantes, Woese et ses collaborateurs analysent d'autres ARNr SSU, et cela leur permet de confirmer la place d'autres méthanogènes et des halophiles extrêmes au sein des « archaebactéries » (Magrum, Luehrsen, and Woese 1978); suivent différents « thermoacidophiles », comme S*ulfolobus acidocaldarius* (C R Woese, Magrum, and Fox 1978; Atoji et al. 1980), et les *Thermoplasma*, classés jusqu'ici parmi les mycoplasmes à cause de leur absence de paroi cellulaire mais que l'analyse de l'ARNr SSU de *Thermoplasma acidophilum* permet de les reclasser parmi les archaebacteria (C R Woese, Magrum, and Fox 1978; C R Woese, Maniloff, and Zablen 1980). Un article de 1980 « *The phylogeny of Prokaryotes* » de Fox et collaborateurs (Atoji et al. 1980) reprend l'ensemble des résultats accumulés pendant les années 70 et propose une phylogénie schématique du vivant (Figure 2) et plusieurs dendrogrammes pour les « archaebactéries » (Figure 3) et différentes familles bactériennes. On peut déjà voir plusieurs des ordres d'archées dans ce dendrogramme, à savoir les *Halobacteriales*, les

Methanomicrobiales, les *Methanococcales*, les *Thermoplasmatales* (représentés par *Thermoplasma acidophilum*) et les Sulfolobales (représentés par *Sulfolobus acidocaldarius*), seuls représentants des futurs Crenarchaeota.

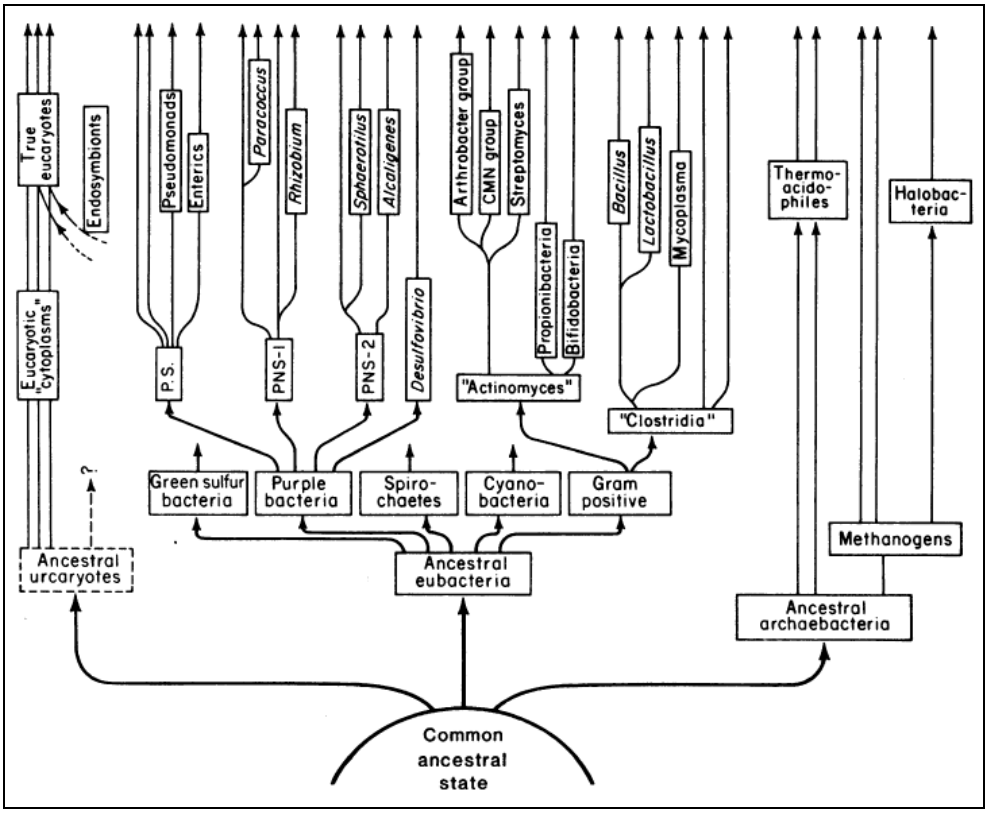


Figure 2. Représentation des grandes lignées du vivant selon Fox et collaborateurs en 1980. Les trois domaines proposés dérivent d'un même ancêtre commun (Atoji et al. 1980).

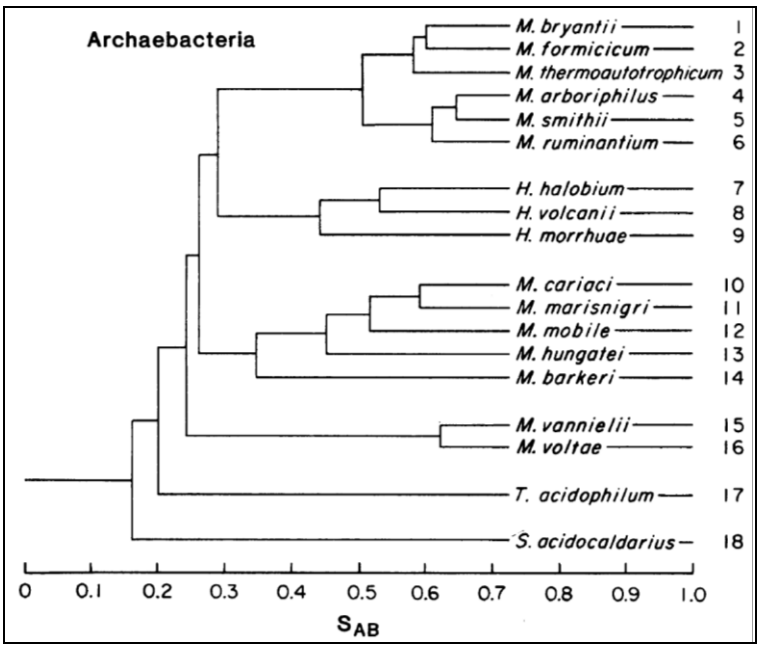


Figure 3. Premier dendrogramme représentant les relations entre les « archaebacteria » selon Fox et collaborateurs en 1980. (Atoji et al. 1980). La racine est placée entre un groupe contenant les méthanogènes (1-6, 10-16), les halophiles (7-9) et Thermoplasma acidophilum (17), et les Sulfolobales, représentés par Sulfolobus acidocaldarius (18). Ces deux groupes correspondent aux futurs phyla des Euryarchaeota et des Crenarchaeota. Au sein du premier groupe, on voit déjà apparaître trois groupes de méthanogènes correspondants aux ordres des Methanobacteriales (1-6), Methanomicrobiales (10-14) et Methanococcales (15-16). (Atoji et al. 1980).

Le début des années 1980 a vu le remplacement des catalogues d'oligonucléotides par les premiers séquençages complets de molécules d'ARNr SSU d'« archaebactéries » (R. Gupta, Lanter, and Woese 1983; Olsen et al. 1985). Il devient donc possible de construire des arbres phylogénétiques du vivant à partir de la séquence primaire de cette molécule (N. R. Pace, Olsen, and Woese 1986) et ainsi de confirmer la phylogénie proposée précédemment sur la base du S_{AB}. La Figure 4 montre un premier arbre non raciné inféré à partir de 950 positions des séquences de l'ARNr SSU de 21 espèces, où l'on peut observer les trois domaines « Eucaryotes », « Eubacteria » et « Archaebacteria ». L'ARNr SSU semblait être un marqueur efficace pour l'analyse de la phylogénie du vivant, et particulièrement des procaryotes. Cette molécule est présente chez tous les organismes vivants connus, sa fonction est très conservée, il est donc facile de trouver son homologue dans une nouvelle espèce, elle est assez longue pour qu'un nombre de positions suffisamment informatives puisse être utilisé et elle est peu transférée entre différents organismes. Ces caractéristiques sont aussi celles des autres ARN ribosomiques (23S et 5S chez les procaryotes), mais l'ARNr 5S, étant plus court, contient moins de signal informatif que l'ARNr SSU et un grand nombre de molécules d'ARNr SSU était déjà séquencé. L'ARNr SSU s'est de fait imposé comme marqueur pour l'étude de la phylogénie des microorganismes. La possibilité d'avoir un seul marqueur pour étudier l'évolution de l'ensemble du vivant est une avancée sans précédent,

particulièrement en ce qui concerne l'évolution des microorganismes ; celle des bactéries et archées bien sûr, mais aussi celle des organismes unicellulaires eucaryotes, considérés comme un seul groupe, par exemple par Whittaker (Whittaker 1969). Le développement de la phylogénie moléculaire a donc ouvert la porte à l'étude de l'évolution et de la phylogénie microbienne et a ainsi donné un nouveau cadre de pensée à la classification microbienne.

Il me semble intéressant d'aborder un point un peu particulier dans cette révolution portée par Woese à propos de la refonte de la classification microbienne. Il apparaît à la lecture des articles qu'il a écrit dans les années 80 que deux idées plutôt contradictoires sont abordées. D'une part, comme je viens de l'expliquer, Woese propose que la classification doive se faire uniquement sur des critères phylogénétiques, la phylogénie des microorganismes se faisant sur la base de caractères moléculaires, et particulièrement sur la séquence des ARN ribosomiques. D'autre part, la proposition du domaine des « Archaebactéries » est suivie d'une description des caractéristiques communes entre les organismes de ce groupe. Ce qui n'est au départ (C R Woese and Fox 1977) qu'une description de ces microorganismes, devient bientôt une recherche de synapomorphies justifiant l'existence d'un groupe taxonomique à part entière. Par exemple, dans l'article « Archaebacteria » de 1978 (C R Woese, Magrum, and Fox 1978) qui est une synthèse des travaux sur ces questions, Woese répertorie quatre grands caractères communs aux archées selon lui : les caractéristiques particulières des ARN ribosomiques et des ARN de transferts, l'absence de parois cellulaires de peptidoglycane, les lipides de membranes contenant des isoprénoïdes et reliés par des liens éthers, et l'habitat « spécialisé » des Archaea ; il se sert de ces caractéristiques pour justifier l'existence du groupe taxonomique « archaebacteria » : *« In summary, enough comparative information now exists pertaining to the archaebacteria to make it possible to characterize the kingdom in a positive way »* (C R Woese, Magrum, and Fox 1978). Il est facilement compréhensible qu'une proposition de changement aussi importante dans la classification microbienne ait fait débat (Wheelis, Kandler, and Woese 1992; Mayr 1990; Mayr 1991; Margulis and Guerrero 1991), et que Woese cherche à répondre à différentes critiques et à justifier la proposition du nouveau groupe « Archaebacteria », mais il est paradoxal d'utiliser ces caractères phénotypiques à homologie discutable comme justification d'une classification qui se veut basée sur les relations phylogénétiques de caractères homologues avérés. Une autre critique peut être formulée à l'encontre de cette description, à propos des environnements « spécialisés » évoqués par Carl Woese. Certes, les archées connues à ce moment-là étaient principalement des extrêmophiles, halophiles et thermoacidophiles, et des méthanogènes. Les milieux dans lesquels se développent ces organismes peuvent nous sembler particuliers, mais ils sont totalement différents entre eux et ils impliquent des adaptations différentes, par exemple, entre les organismes vivant à haute

température et ceux vivant dans des conditions de haute salinité. Leur regroupement n'est le résultat que de leur exotisme à nos yeux, ce qui constitue un argument subjectif et anthropocentré.

c. La nouvelle classification du vivant : « Archaea », « Bacteria », « Eucarya »

La vision tripartite du vivant est renforcée en 1990 quand Woese et ses collaborateurs proposent une nouvelle taxonomie en modifiant les noms des trois royaumes primaires pour les remplacer par trois domaines : « Bacteria », « Eucarya » et « Archaea » (C R Woese, Kandler, and Wheelis 1990). Le premier but de cette modification des noms est d'abord de marquer la séparation entre les Archaea et les Bacteria et ainsi de désacraliser définitivement la dichotomie Procaryotes/Eucaryotes, pour ne pas laisser de doute quant au fait que les Archaea ne sont pas une branche parmi les bactéries. L'autre but de cette proposition est d'insister sur le fait que la classification et la taxonomie doivent se faire selon un « système naturel », que le plus logique est celui de l'histoire évolutive comme l'a proposé Darwin et que, finalement, la phylogénie moléculaire est la méthode la plus efficace pour sa reconstruction : *« The time has come to bring formal taxonomy into line with the natural system emerging from molecular data »* (C R Woese, Kandler, and Wheelis 1990). Un arbre du vivant est ainsi proposé (Figure 4), inféré sur la base des séquences de l'ARNr SSU et raciné entre les bactéries d'une part et les archées et les eucaryotes d'autre part. La racine est placée ainsi d'après les analyses faites à la même époque sur des paralogues présents chez le dernier ancêtre commun, dont chaque copie peut servir de racine à la phylogénie de l'autre (Iwabe et al. 1989). La place de cette racine admet donc un ancêtre commun exclusif entre les archées et les eucaryotes et éloigne les archées des bactéries. Pour les auteurs, c'est un argument de poids pour appuyer la validité du domaine « Archaea » et le séparer du domaine « Bacteria ».

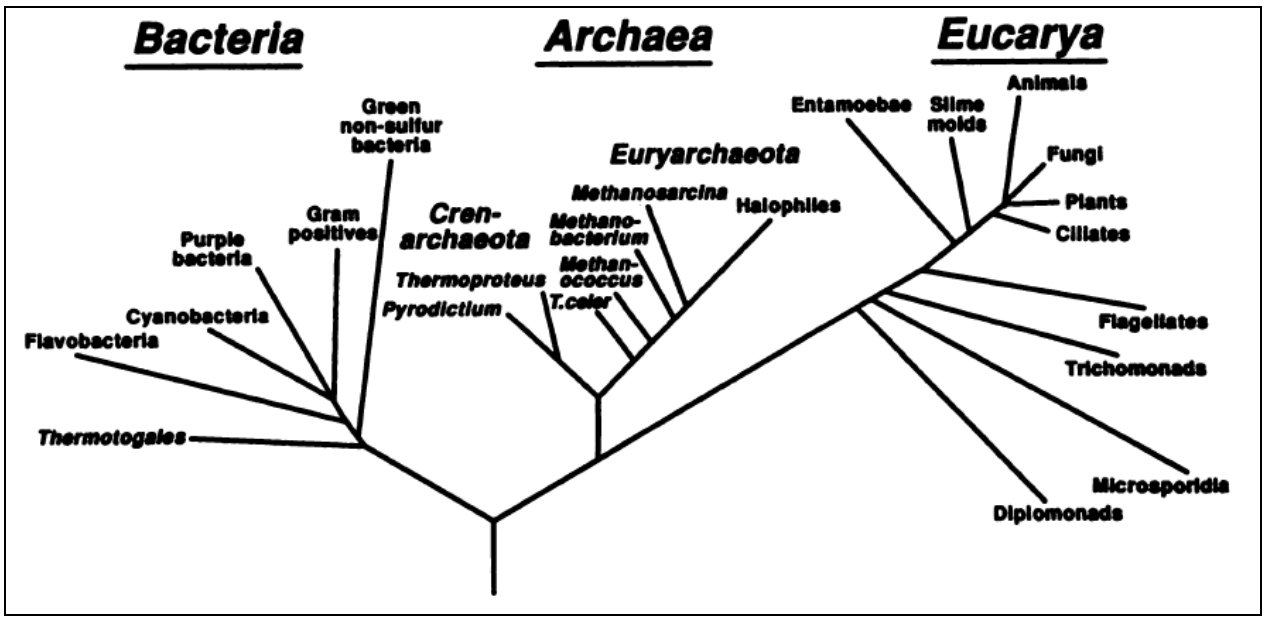


Figure 4. Arbre du vivant représentant les trois domaines Bacteria, Archaea et Eucarya. (Olsen and Woese 1993). On voit bien les deux « royaumes » archéens : Crenarchaeota et Euryarchaeota. La racine de l'arbre est placée entre les bactéries et les archées/eucaryotes d'après l'étude des phylogénies de couples de paralogues anciens.

Le domaine Archaea est composé de deux lignées majeures, appelées « royaumes » dans cette nouvelle classification, les Euryarchaeota et les Crenarchaeota, entre lesquelles se place la racine de la phylogénie des archées. Les Euryarchaeota rassemblent les halophiles extrêmes, trois lignées de méthanogènes, le genre *Archaeoglobus*, le genre *Thermoplasma*, et le groupe des *Thermococcus-Pyrococcus*, deux groupes de thermophiles. Les Euryarchaeota sont donc phénotypiquement très divers, avec des modes de vie très différents, en termes écologiques (thermophiles, acidophiles, halophiles…) et métaboliques (méthanogènes, sulfato-réducteur, hétérotrophes…), ce qui leur a valu leur nom d'Euryarchaeota, « eury » du grec « *εύρύς* » pour « large ». Les Crenarchaeota regroupent uniquement des organismes thermophiles extrêmes et ont un mode de vie similaire, nécessitant du soufre comme source d'énergie. Leur nom vient de « cren » du grec « *κρήνη* » pour « source », rappelant leur ressemblance phénotypique avec l'ancêtre commun des archées, que les auteurs supposent être plus proche de ce groupe là. Cette classification des archées en deux royaumes Euryarchaeota et Crenarchaeota restera la norme dans ses grandes lignes pendant presque 20 ans.

2. A partir des années 1990 : Exploration de la diversité archée.

a. Etudes environnementales

Le début des années 1990 a été marqué par la découverte concomitante par DeLong (DeLong 1992) et Fuhrman et collaborateurs (Fuhrman, McCallum, and Davis 1992), d'archées marines mésophiles et aérobies, phylogénétiquement proches pour certaines des Crenarchaeota (le groupe I), et pour d'autres se plaçant au sein des Euryarchaeota (le groupe II) (Figure 5).

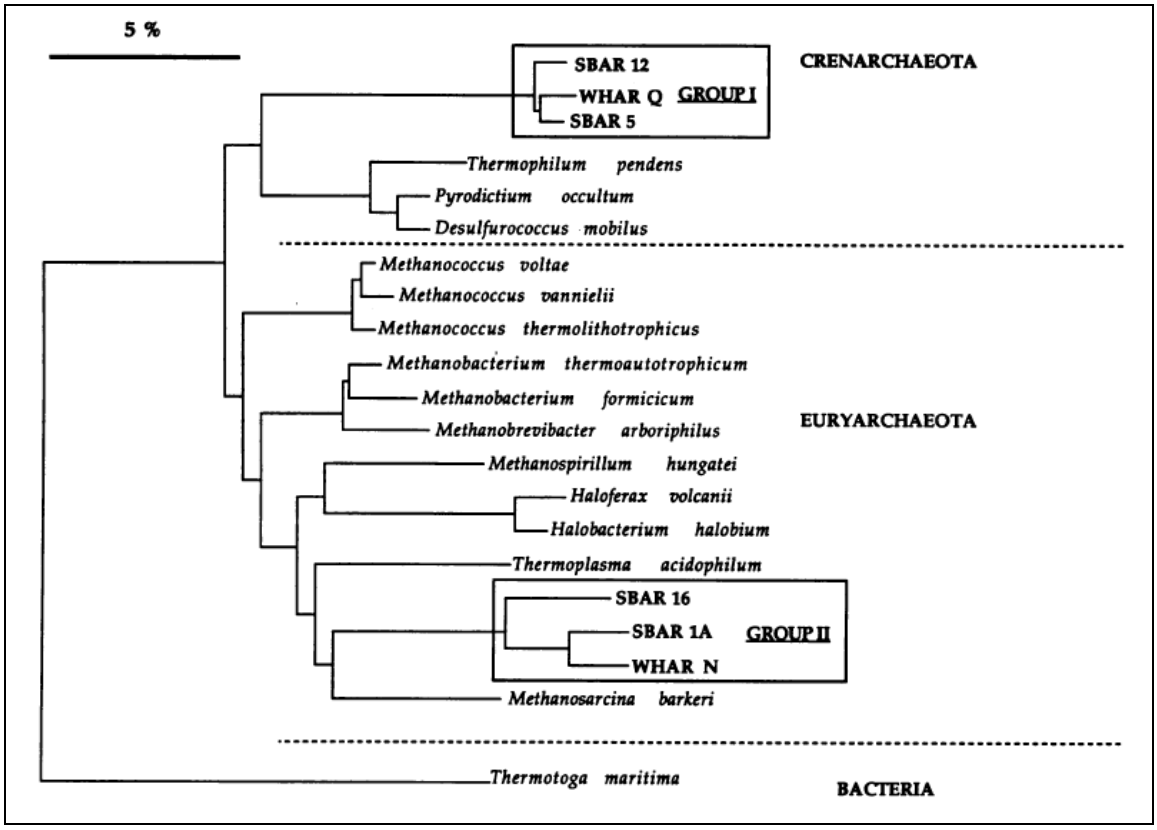


Figure 5. Phylogénie des Archées montrant les groupes I et II. Arbre inféré à partir de l'ARNr SSU d'archées cultivées et de séquences d'espèces non-cultivées d'échantillons marins (DeLong 1992). Le groupe I se place à la base des Crenarchaeota déjà connues, et le groupe II se place à l'intérieur des Euryarchaeota, comme groupe frère des Methanosarcinales. Cet arbre est raciné sur la bactérie Thermotoga maritima.

Cette découverte marque le début des études environnementales pour l'exploration de la diversité des archées. En effet, en ayant mis en évidence que l'ARNr SSU était un bon marqueur pour la phylogénie des organismes, Woese a ouvert la porte à l'étude de la diversité microbienne grâce à une approche moléculaire par amplification par PCR et séquençage des gènes d'ARNr SSU à partir de l'ADN d'échantillons naturels (Olsen et al. 1986).

DeLong, Fuhrman et leurs collaborateurs ont extrait l'ADN de différents échantillons marins de surface, puis amplifié et séquencé l'ARNr SSU et ont ainsi pu avoir un aperçu de la diversité

microbienne présente. Les archées ayant été décrites jusqu'alors comme un domaine d'organismes vivant dans des environnements « spécialisés », elles n'étaient pas attendues dans des échantillons d'eau de mer de surface, oxygénée et plutôt froide, et la première hypothèse avancée par les auteurs de ces deux articles pour expliquer la présence de séquences d'archées dans leurs échantillons est une possible contamination. Il se trouve que des méthanogènes avaient été découverts peu de temps avant dans l'intestin de gros animaux, dont l'homme, et qu'une contamination humaine aurait pu expliquer la présence des archées du groupe II (se plaçant au sein des Euryarchaeota dont font partie les méthanogènes). Cependant, les auteurs vont privilégier une autre explication, celle d'une diversité d'archées vraiment inféodées au milieu planctonique marin, malgré le caractère exceptionnel de cette découverte comme le montre cette phrase de DeLong : « *A more provocative explanation is that group I and group II represent undescribed mesophilic, aerobic members of the archaea* » (DeLong 1992). En effet, ces deux travaux décrivant une même diversité sur des échantillons océaniques provenant de quatre localités géographiques différentes sur les deux côtes des Etats-Unis (DeLong 1992; Fuhrman, McCallum, and Davis 1992) se confirment l'un l'autre et font apparaître une diversité d'archées inconnues et non cultivées assez remarquable. Ces communautés d'archées semblent alors importantes en termes de biomasse dans les océans, bien que moins importantes que les bactéries présentes dans les mêmes milieux. Cette découverte est très importante d'un point de vue écologique puisqu'elle met en évidence une part de la biodiversité océanique ignorée jusqu'alors. Elle est aussi importante du point de vue de la connaissance de la diversité des archées, car elle décrit des groupes totalement inconnus dont le mode de vie mésophile et aérobie défie l'idée selon laquelle les archées vivent exclusivement dans des environnements « spécialisés ». Enfin, du point de vue de leur histoire évolutive, la découverte selon laquelle l'une de ces lignées se place à la base des Crenarchaeota ouvre de nombreuses questions quant à la nature de l'ancêtre commun des archées. Le groupe I sera d'ailleurs proposé comme étant un phylum à part entière en 2008 par Brochier-Armanet et collaborateurs sous le nom de « Thaumarchaeota », du grec « *θαυμασ* » « thaumas » pour « merveilleux » (Brochier-Armanet et al. 2008).

Les études environnementales sur la diversité des archées vont mettre en évidence dans les années suivantes un nombre important de lignées non cultivées vivant, pour certaines, dans des environnements nous semblant moins extrêmes que ceux des archées cultivées et connues (Christa Schleper, Jurgens, and Jonuscheit 2005). À l'instar de l'océan, on trouve des archées dans des sédiments marins, des lacs d'eau douce, des sols, en symbiose avec d'autres organismes, mais aussi dans des milieux très acides (mines acides), ou des tapis microbiens riches en méthane, en association avec des méthanogènes et des bactéries sulfato-réductrices. Initialement basées sur l'amplification et le séquençage de l'ARNr SSU, les études environnementales se sont

progressivement diversifiées en termes de techniques. Le groupe de DeLong a largement contribué à ce développement, par exemple avec l'isolement et l'analyse d'un large fragment génomique de 40 kilobases d'une archée du groupe I, un des premiers exemples de l'application des techniques de métagénomique (Stein et al. 1996). Dans ce travail, les auteurs ont extrait et digéré l'ADN d'un échantillon de plancton marin pour construire une banque de BAC (Bacterial Artifical Chromosome) modifiés pour recevoir des gros inserts d'ADN. Ces inserts pouvant contenir plusieurs gènes, si un marqueur phylogénétique tel que l'ARNr SSU est présent, il est possible, non seulement de savoir à quel groupe taxonomique appartient l'organisme, mais surtout d'étudier les gènes placés autour de ce marqueur et ainsi d'en connaître un peu plus sur les activités métaboliques que cet organisme peut réaliser. L'application de cette méthodologie à des environnements dans lesquels une seule espèce d'archée est présente a permis de reconstruire et de séquencer des génomes presque complets, comme ça a été le cas pour l'organisme *Cenarchaeum symbiosum* en 2006, vivant en symbiose avec une éponge et une bactérie (Hallam et al. 2006). Les auteurs de ce papier (dont DeLong) ont pour cela construit une banque de fosmides, séquencé chaque insert et réassemblé le génome en utilisant les parties chevauchantes des différents fragments. Ce génome a été le premier à être séquencé pour une archée du groupe I.

Il est également possible d'étudier la grande diversité d'archées non cultivées en ayant recours aux observations par FISH (hybridation in situ en fluorescence), en ciblant spécifiquement les ARNr SSU d'archées (Christa Schleper, Jurgens, and Jonuscheit 2005; DeLong 1992). Cette technique a pour avantage de montrer l'abondance de la population ciblée, mais aussi de montrer des cellules vivantes et de prouver que la détection d'ARNr SSU d'archées dans ces milieux n'est pas un artefact lié à des contaminations ou des cellules mortes (Figure 6). C'est ainsi que Karner, DeLong et Karl ont pu montrer en 2001 que les crenarchaeota marines du groupe I étaient très abondantes dans les milieux océaniques, de l'ordre de 20% de la population microbienne planctonique (Karner, DeLong, and Karl 2001).

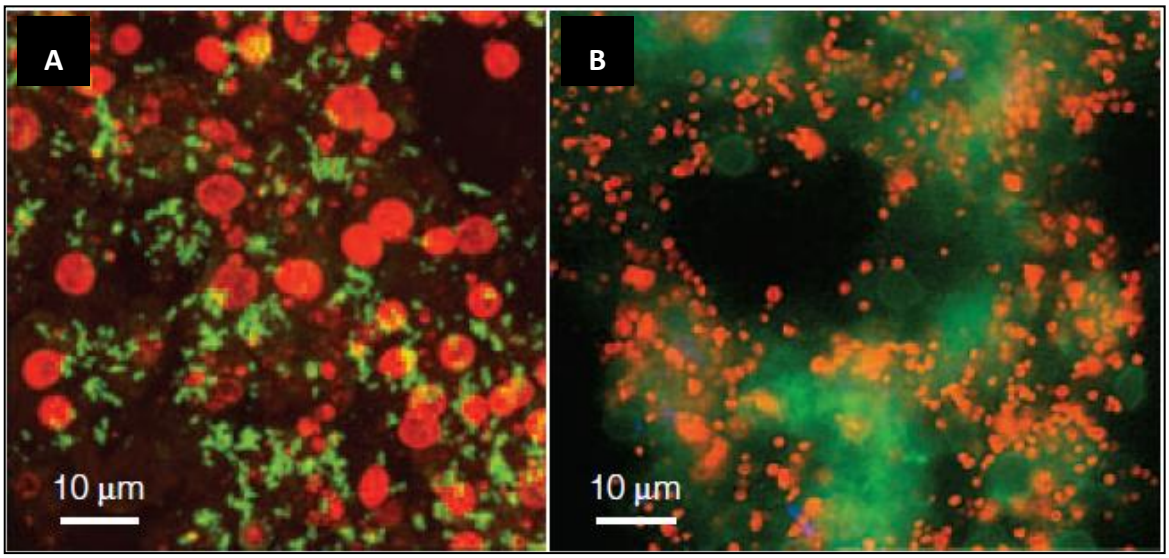


Figure 6. Images d'archées non cultivées en FISH. Issues de l'article Schleper et collaborateurs en 2005 (Christa Schleper, Jurgens, and Jonuscheit 2005). A : cellules de Cenarchaeum symbiosum, archée du groupe I (en vert), et d'Axinella mexicana, une éponge, hôte de C. symbiosum (en rouge). Image de C. Preston, Monterey Bay Aquarium Research Institute, Moss Landing, USA. B : Enrichissement de crenarchaeota du groupe I (en rouge), sur des racines de tomates, et des bactéries associées (en vert). Image de H. Simon, Oregon Health & Science University, Beaverton, USA.

Si la découverte d'archées mésophiles et aérobies fut très importante pour la perception que l'on pouvait avoir de ce domaine, la recherche d'archées dans des milieux extrêmes a aussi apporté son lot de lignées inconnues, certaines se plaçant à proximité de la base de la phylogénie des archées. En 1994, Barns et ses collaborateurs (Barns, Fundyga, Jeffries, & Pace, 1994; Barns, Delwiche, Palmer, & Pace, 1996) ont décrit la diversité d'archées dans l'eau et les sédiments de deux sources chaudes du parc de Yellowstone. Différentes séquences de Crenarchaeota ont été décrites, branchant à proximité des Desulfurococcales et des Sulfolobales ou au sein des Thermoproteales. D'autres séquences branchaient avec le groupe I nouvellement décrit par DeLong et Furhmann, et d'autres encore branchaient au sein de l'arbre des archées avant même la divergence entre Crenarchaeota et Euryarchaeota. Cette position a amené les auteurs à proposer que ces organismes pouvaient être les représentants d'un nouveau phylum d'archées et ils proposèrent le nom de « Korarchaeota », du grec « *κορος* » pour « jeune homme », pour la divergence précoce de ce groupe dans l'histoire des archées (Barns et al. 1996).

Au cours des années 2000, un nouveau « type » d'archées a été découvert : de très petites archées, à commencer par *Nanoarchaeum equitans* en 2002 (H. Huber et al. 2002). Cette archée de seulement 400 nm de diamètre vit en symbiose avec une autre archée, *Ignicoccus hospitalis*, une crenarchée. Son génome ne fait que 0,5 megabases, et son séquençage fera apparaître de nombreux traits particuliers tels que l'absence d'un grand nombre de gènes essentiels chez les archées ou des gènes coupés en deux parties, notamment des gènes codants des ARNt. Une petite taille cellulaire et

un génome de taille très réduite sont des caractéristiques classiques rencontrées chez les bactéries endosymbiotiques ou parasites, mais n'avaient jamais été observées chez les archées ou chez des hyperthermophiles. Les analyses phylogénétiques de l'ANRr SSU ont placé *N. equitans* très basalement dans l'arbre des archées mais avec un faible soutien (H. Huber et al. 2002). Ceci a conduit les auteurs à proposer qu'elle serait la représentante d'un nouveau phylum d'archées, les « Nanoarchaeota » (« nano » pour sa petite taille cellulaire de l'ordre du nanomètre). Plus récemment, d'autres archées de petite taille ont été décrites, le groupe des ARMAN en 2006 par Baker et collaborateurs (Baker et al. 2006) et les Nanohaloarchaeota en 2011 par Narasingarao et collaborateurs (Narasingarao et al. 2012). Le groupe des ARMAN pour « archaeal Richmond Mine acidophilic nanoorganisms » a été trouvé dans les mines acides de Richmond, au niveau de biofilms principalement dominés par des archées de type thermoplasmatales (*Thermoplasma* et *Ferroplasma*) et des bactéries du genre *Leptospirillum*. En 2010, trois génomes ont été publiés, reconstruits à partir de données métagénomiques (Baker et al. 2010) : '*Candidatus* Micrarchaeum acidiphilum ARMAN-2', '*Candidatus* Parvarchaeum acidiphilum ARMAN-4' et '*Candidatus* Parvarchaeum acidophilus ARMAN-5'. Ces deux derniers génomes sont très proches en termes de similarité de séquences. Dans les trois cas, il s'agit de toutes petites archées comparables à *N. equitans* en termes de taille (environ 500 nm de diamètre), mais leurs génomes sont deux fois plus gros (environ 1 megabase) bien qu'ils aient des caractéristiques comparables telles que l'absence de nombreux gènes essentiels ou des gènes splittés chez '*Ca.* Micrarchaeum acidiphilum ARMAN-2'. Les absences de gènes demanderont à être confirmées car ces génomes ne sont pas totalement complets. Leurs caractéristiques cellulaires et génomiques portent à croire qu'elles ont un type de vie parasitique ou symbiotique mais n'étant pas cultivées, ce point reste sans réponse. Enfin, les Nanohaloarchaea sont de petites archées halophiles dont les génomes ont été séquencés d'après des données métagénomiques. C'est le cas pour '*Candidatus* Nanosalinarum sp.' et '*Candidatus* Nanosalina sp.' (Narasingarao et al. 2012) dont le séquençage a été suivi de près par celui de '*Candidatus* Haloredivivus' (Ghai et al. 2011). Ces trois génomes ont une taille d'environ 1,2 megabases et '*Ca.* Nanosalinarum sp.' et '*Ca.* Nanosalina sp.' ont des cellules de très petite taille, environ 600 nm. Les analyses phylogénétiques ont placé ces organismes au sein des Euryarchaeota, comme groupe frère des Halobacteria. C'est pourquoi la nouvelle classe des Nanohaloarchaea a été proposée par Narasingarao et collaborateurs (Narasingarao et al. 2012).

L'un des derniers phyla d'archée à avoir été proposé est celui des Aigarchaeota par Nunoura et collaborateurs en 2011 (Nunoura et al. 2011), lors de la publication du génome composite de '*Candidatus* Caldiarchaeum subterraneum'. Cet organisme appartient au groupe HWCGI (Hot Water Crenarchaeotic Group I), un groupe d'organismes thermophiles, non cultivés, et

phylogénétiquement proche des Thaumarchaeota et des Crenarchaeota tout en en étant distinct. Cette position basale dans l'arbre des archées et des caractéristiques génomiques particulières, telles que la présence de protéines typiquement eucaryotes intervenant dans le système ubiquitine-protéasome (système de marquage des protéines par l'ubiquitine en vue de leur dégradation par le protéasome) ont conduit les auteurs à proposer la création du nouveau phylum. Ils proposent le nom d'« Aigarchaeota », du grec « *αυγη* », « aigi », pour « aurore », ces organismes présentant des caractéristiques intermédiaires entre les hyperthermophiles et les mésophiles dans cette partie de l'arbre des archées.

Depuis, le phylum « Geoarchaeota » ou « NAG1 » (pour « novel archaeal group 1 ») a été proposé par Kozubal et collaborateurs en 2013 (Kozubal et al. 2013), d'après l'analyse de quatre génomes construits à partir de métagénomes issus de tapis microbiens du parc de Yellowstone. Cet environnement est acide, a une haute température (65-80°C), contient du fer ferreux et une faible concentration d'oxygène. Ces organismes se placeraient en groupe frère des Crenarchaeota et auraient des caractéristiques génomiques particulières par rapport aux autres archées, ce qui a conduit les auteurs à proposer la création d'un nouveau phylum. Néanmoins, leur position phylogénétique demanderait à être confirmée, particulièrement au regard de leur position basale supposée dans l'arbre des archées.

b. Les années 2000 : Séquençage de nombreux génomes complets

Depuis la découverte des archées en tant que domaine à part entière, la phylogénie moléculaire a toujours été un outil privilégié pour leur étude. Le séquençage de génomes complets a donc représenté une grande avancée, de façon générale en biologie mais particulièrement pour la connaissance de ce domaine. En effet, peu d'espèces d'archées sont cultivées par rapport aux eucaryotes et aux bactéries, et peu d'outils de génétique et de biologie moléculaire ont pu être développés du fait de leurs particularités cellulaires, de conditions de vie et même génétiques (bien que de plus en plus de modèles aient été développés ces dernières années (Leigh et al. 2011)). L'accès aux séquences complètes de leurs génomes a apporté une quantité de données sans précédent sur ces organismes encore mal connus. *Methanococcus jannaschii* fut la première archée dont le génome a été complètement séquencé en 1996 (Bult et al. 1996). Depuis, 227 génomes d'archées ont été complètement séquencés (Kyrpides 1999)(aout 2013). Bien que ne couvrant pas la totalité de la diversité des archées, au moins un génome représentant chaque phyla est disponible à l'heure actuelle (Euryarchaeota, Crenarchaeota, Korarchaeota, Thaumarchaeota, et Nanoarchaeota). Pour certaines espèces, plusieurs souches sont séquencées ce qui permet des études de génomique

comparée à petite échelle taxonomique. L'exemple le plus important est celui de l'espèce *Sulfolobus islandicus* dont 15 souches sont séquencées. Certaines ont été étudiées par génomique comparative par Reno et collaborateurs en 2009 (Reno et al. 2009) d'un point de vue biogéographique.

La publication de génomes complets d'archées et l'augmentation en nombre et en diversité est aussi très importante en ce qui concerne l'étude de leur histoire évolutive. Jusqu'ici, le marqueur phare était l'ARNr SSU, une molécule privilégiée pour la reconstruction phylogénétique de l'histoire des organismes, très conservé au niveau de sa séquence et de sa fonction, et peu transféré. Il reste néanmoins un marqueur unique et à ce titre soumis à certains biais tels qu'un nombre limité de positions, donc de caractères, pouvant de plus accumuler une saturation mutationnelle importante. L'accès aux séquences de gènes codant des protéines grâce aux génomes complets a donc permis d'utiliser de nouveaux marqueurs pour reconstruire la phylogénie des archées. Les protéines ribosomiques ont été parmi les premières séquences utilisées, puisque, intervenant dans le même système biologique que l'ARNr SSU, elles sont aussi très conservées et peu transférées, ce qui en fait des marqueurs phylogénétiques de choix. C'est d'ailleurs aussi le cas pour les sous-unités de l'ARN polymérase. Et plus que l'analyse de nouveaux marqueurs, l'analyse conjointe de plusieurs protéines a été rendue possible par l'arrivée des données génomiques. Ces études ont été commencées par Matte-Tailliez et collaborateurs en 2002 (Matte-Tailliez et al. 2002), et améliorées depuis par l'ajout de nouvelles protéines et l'augmentation de l'échantillonnage taxonomique (Simonetta Gribaldo and Brochier-Armanet 2006). Ces travaux ont permis de confirmer les grandes lignes de la phylogénie obtenue par Woese en 1980 et de raffiner la phylogénie interne des différents phyla d'archées.

Les génomes complets ont tout d'abord été obtenus à partir d'organismes cultivés pour des raisons évidentes de quantité de matériel génétique nécessaire au séquençage et de sécurité quant aux contaminations. Ainsi que je l'ai déjà évoqué, des génomes peuvent être séquencés pour des organismes non isolés en culture pure à partir de données métagénomiques, comme par exemple *Cenarchaeum symbiosum* (Hallam et al. 2006). De plus en plus de génomes sont reconstruits de cette manière, par exemple ceux des trois archées appartenant au groupe ARMAN (Baker et al. 2010), ou à partir d'une culture enrichie pour celui de '*Candidatus* Korarchaeum cryptofilum' (Elkins et al. 2008). Cette stratégie mène normalement à l'obtention de ce qu'on appelle un génome composite. Aujourd'hui, le séquençage à partir de cellules uniques est devenu possible et le premier génome ainsi séquencé, celui de la bactérie '*Candidatus* Sulcia muelleri DMIN', a été publié en 2010 (Woyke et al. 2010). La première archée séquencée de cette façon a été '*Candidatus* Nitrosoarchaeum limnia SFB1' en 2011 (Blainey et al. 2011), un membre des Thaumarchaeota (anciennement crenarchaeota du groupe I), dont les représentants sont très difficiles à cultiver

encore aujourd'hui. Cette nouvelle technique ouvre des portes quant à la connaissance de l'ensemble des groupes non cultivés d'archées, comme le montre la publication très récente de Rinke et collaborateurs, décrivant le séquençage de 201 archées et bactéries, principalement issues de groupes peu connus (Rinke et al. 2013). L'accès à leurs séquences génomiques pourrait permettre d'en savoir plus sur leur évolution, bien sûr, mais aussi sur leurs métabolismes potentiels et, peut être, sur les conditions nécessaires à leur développement, ce qui pourrait servir à mettre en place des milieux propices à leur culture pour une étude *in vivo* plus approfondie.

3. Conclusion

La découverte des Archaea en tant que domaine du vivant à part entière est étroitement liée à l'étude de l'évolution grâce à des données moléculaires, très vite remplacées par la phylogénie moléculaire. Les difficultés de culture d'une grande partie de leur diversité ont aussi renforcé l'importance des données génomiques dans l'étude des archées. Ces raisons historiques expliquent que l'étude des archées soit intimement liée à celle de leur histoire évolutive et de leurs génomes ; c'est pourquoi il m'a parut important de rappeler les grandes étapes historiques qui ont conduit à la connaissance actuelle de ce domaine du vivant, et l'importance de l'étude de leur évolution.

B. Phylogénie des Archaea

1. Diversité des Archaea : la phylogénie de l'ARNr SSU

Comme nous venons de le voir six phyla d'archées sont actuellement relativement acceptés : les Euryarchaeota et les Crenarchaeota, les deux phyla originellement proposé par Woese en 1980, les Korarchaeota, les Nanoarchaeota, les Thaumarchaeota et les Aigarchaeota. Ces six phyla regroupent des organismes très divers et j'aimerais ici présenter une vue d'ensemble non exhaustive de cette diversité. La Figure 7, ci-dessous, est un arbre publié par Schleper et collaborateurs en 2005 (Christa Schleper, Jurgens, and Jonuscheit 2005), inféré sur les séquences d'ARNr SSU d'archées cultivées et non cultivées. Bien que datant de 2005, cette phylogénie est très complète en ce qui concerne les groupes non cultivés et donne un bon aperçu de ce que l'on connaît de la diversité des archées actuellement.

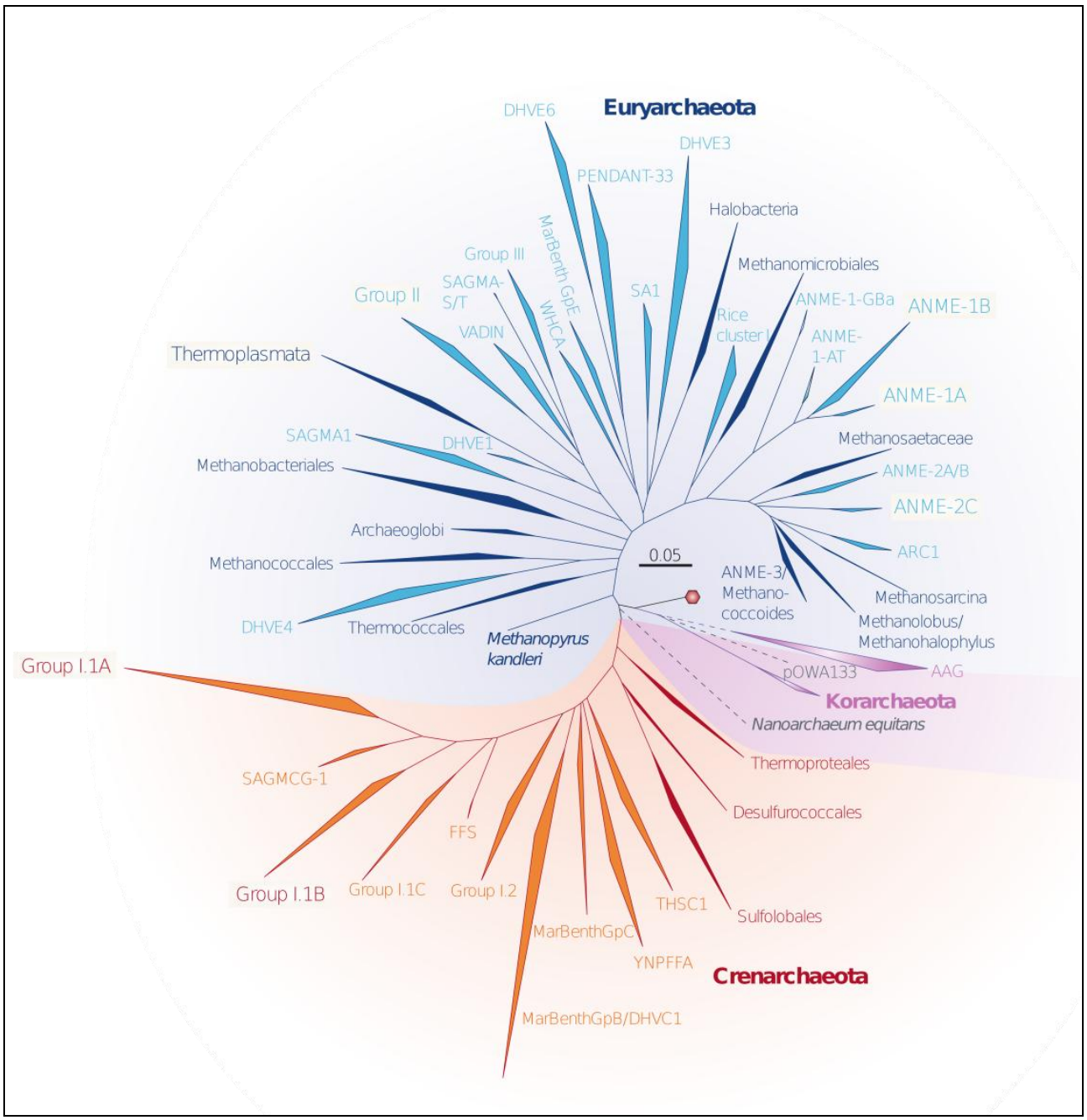


Figure 7. Phylogénie des archées cultivées et non cultivées. (Christa Schleper, Jurgens, and Jonuscheit 2005). Inféré à partir de l'ARNr SSU, cet arbre représente la diversité du domaine Archaea.

Dans cette phylogénie, le phylum Thaumarchaeota n'apparait pas en tant que tel mais regroupe le groupe I et d'autres branches comme les SAGMCG-1. Les séquences du phylum Aigarchaeota se placeront groupe frère de ces groupes et juste après les séquences de crenarchées lors de leur découverte (Nunoura et al. 2010). Par contre, on voit bien apparaitre les autres phyla, Euryarchaeota, Nanoarchaeota (placé sur une branche en pointillés car sa place n'est pas résolue), Korarchaeota et Crenarchaeota. Le groupe AAG « Ancient Archaeal Group » et le clone pOWA133 (Takai and Horikoshi 1999) représentent des séquences environnementales isolées dans une source

hydrothermale sous-marine profonde au Japon. Les analyses phylogénétiques basées sur l'ARNr SSU les placent en groupe frère d'un groupe contenant les Crenarchaeota, Korarchaeota et Thaumarchaeota (Takai and Horikoshi 1999). Des études supplémentaires seraient nécessaires pour évaluer l'importance de ce groupe, malheureusement très peu d'informations sont disponibles.

a. Les Euryarchaeota

Cette phylogénie d'ARNr SSU (Figure 7) montre neuf ordres d'Euryarchaeota (en bleu foncé) pour lesquels des génomes complets sont séquencés et des organismes cultivés : les *Methanopyrales* (représentés par *Methanopyrus kandleri*), les *Thermococcales*, les *Archaeoglobales*, les *Methanobacteriales*, les *Thermoplasmatales*, les *Halobacteriales*, les *Methanomicrobiales*, les *Methanosarcinales* et les *Methanococcales*. Depuis ont été ajoutés les ordres des *Methalocellales* (représentés ici par le groupe « Rice cluster I ») (Sakai et al. 2008) et des *Methanoplasmatales* (Mihajlovski, Alric, and Brugère 2008; Paul et al. 2012) et la classe des *Nanohaloarchaea* (Narasingarao et al. 2012). Le groupe des ARMAN a été défini sur la base de séquences environnementales et depuis, trois génomes ont été publiés (Baker et al. 2010), mais il n'est pas défini formellement comme un ordre taxonomique. Un organisme du groupe DHVE2 a été cultivé et séquencé, *Aciduliprofundum boonei* (Reysenbach et al. 2006) et un génome du groupe II d'archées marines a été reconstruit à partir de données métagénomiques (Iverson et al. 2012). À ces groupes se rajoute un nombre important de groupes d'organismes non cultivés, découverts lors d'études environnementales telles que celles décrites plus tôt. On trouve entre autre le groupe III des archées marines, découvertes par DeLong et Furhman et collaborateurs en 1992, les groupes ANME (pour Anaerobic Methanotrophic archaea) d'archées méthanotrophes, phylogénétiquement proches des archées méthanogènes, ou différents groupes DHVE, pour « Deep-sea Hydrothermal Vent Euryarchaeotic ».

Hyperthermophiles

En termes d'environnement, on trouve des Euryarchaeota dans quasiment tous les milieux. Puisque la réputation des archées est fondée sur leur extrêmophilie, commençons par les milieux hyperthermophiles (Figure 8). Différentes classes d'euryarchées vivent dans de tels milieux, i.e. à des températures supérieures ou égales à 80°C (Figure 9). Les *Thermococcales* sont tous hyperthermophiles, et ont été observés dans des sources hydrothermales terrestres ou sur des cheminées hydrothermales sous-marines, de surface ou profondes. Une espèce a même été isolée depuis un puits de pétrole, *Thermococcus sibiricus* (Mardanov et al. 2009). Leurs températures de

croissance optimale varient entre 80 et 100°C. Ce sont des organismes anaérobes, chimiorganotrophes ou chimiolitotrophes, utilisant le soufre comme accepteur final d'électrons. Les *Archaeoglobales* sont aussi des organismes hyperthermophiles anaérobes vivant dans des sédiments de sources chaudes ou des cheminées hydrothermales. Les espèces du genre *Archaeoglobus* sont sulfato-reductrices, propriété relativement rare dans le vivant, partagée seulement avec quelques bactéries comme les *Desulfurovibrio* (von Jan et al. 2010). L'organisme *Ferroglobus placidus*, archaeoglobale isolé dans un système hydrothermal sous-marin peu profond à Volcano en Italie, n'est pas sulfato-reducteur, mais est capable d'oxyder le fer à pH neutre alors que cette capacité n'avait été observée que dans des milieux soit à pH acide, soit à basse température (Hafenbradl et al. 1996). Des groupes d'organismes non cultivés tels que les groupes DHVE, ont été détectés sur des fumeurs sous-marins et un organisme du groupe DHVE2 a été isolé et séquencé, *Aciduliprofundum boonei*, phylogénétiquement proche des Thermoplasmatales et thermoacidophile (Reysenbach et al. 2006).

Un certain nombre de méthanogènes sont aussi hyperthermophiles, comme certains *Methanococcales*, *Methanobacteriales* et *Methanopyrus kandleri* (Kurr et al. 1991), seul représentant des *Methanopyrales* dont le génome ait été séquencé et pouvant vivre jusqu'à une température de 110°C.

Figure 8. Champ hydrothermal d'El Tatio au Chili. Photographies de Purificación López-García.

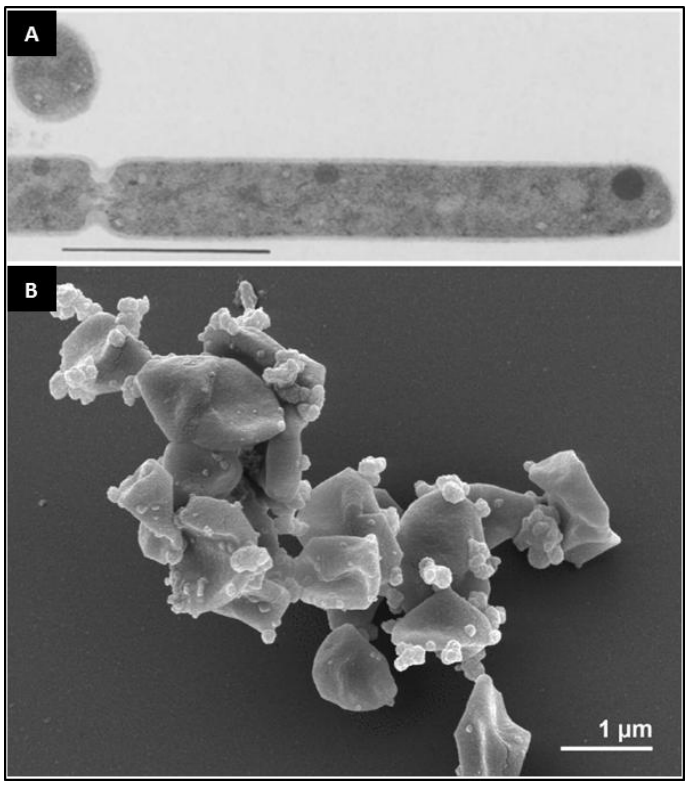


Figure 9. Euryarchées hyperthermophiles. A : Coupe fine de Methanopyrus kandleri (Kurr et al. 1991) Barre=1µm. B : Cellules d'Archaeoglobus profundus, image de microscopie électronique à balayage (von Jan et al. 2010).

Mésophiles et psychrophiles

A contrario, on trouve aussi des euryarchées dans des milieux extrêmement froids. Les groupes d'archées non cultivées II et III découverts dans les océans (DeLong 1992; Fuhrman, McCallum, and Davis 1992) ont été la première preuve que des archées sont présentes dans ces milieux océaniques froids, l'eau pouvant avoir une température de l'ordre de 4°C. En réalité, ces milieux froids ou tempérés (jusqu'à 50°C) sont plutôt majoritaires sur terre, et les organismes mésophiles (vivant entre 4 et 50°C) et psychrophiles (vivant à des températures inferieures à 4°C) sont très abondants, y compris parmi les archées (Cavicchioli 2006). L'espèce halophile *Halorubrum lacusprofundi* a été la première archée détectée dans ce genre de milieux extrêmement froid, en l'occurrence à *Deep Lake* en Antarctique (P.D. Franzmann et al. 1988). Des méthanogènes ont été isolés dans des eaux douces, par exemple à *Ace Lake* en Antarctique (P D Franzmann et al. 1997; P.D. Franzmann et al. 1992), ou dans de l'eau de mer, à *Skan Bay* en Alaska (Cavicchioli 2006). Mais la majeure partie de la diversité d'Euryarchaeota vivant à basse température est encore peu connue et non cultivée.

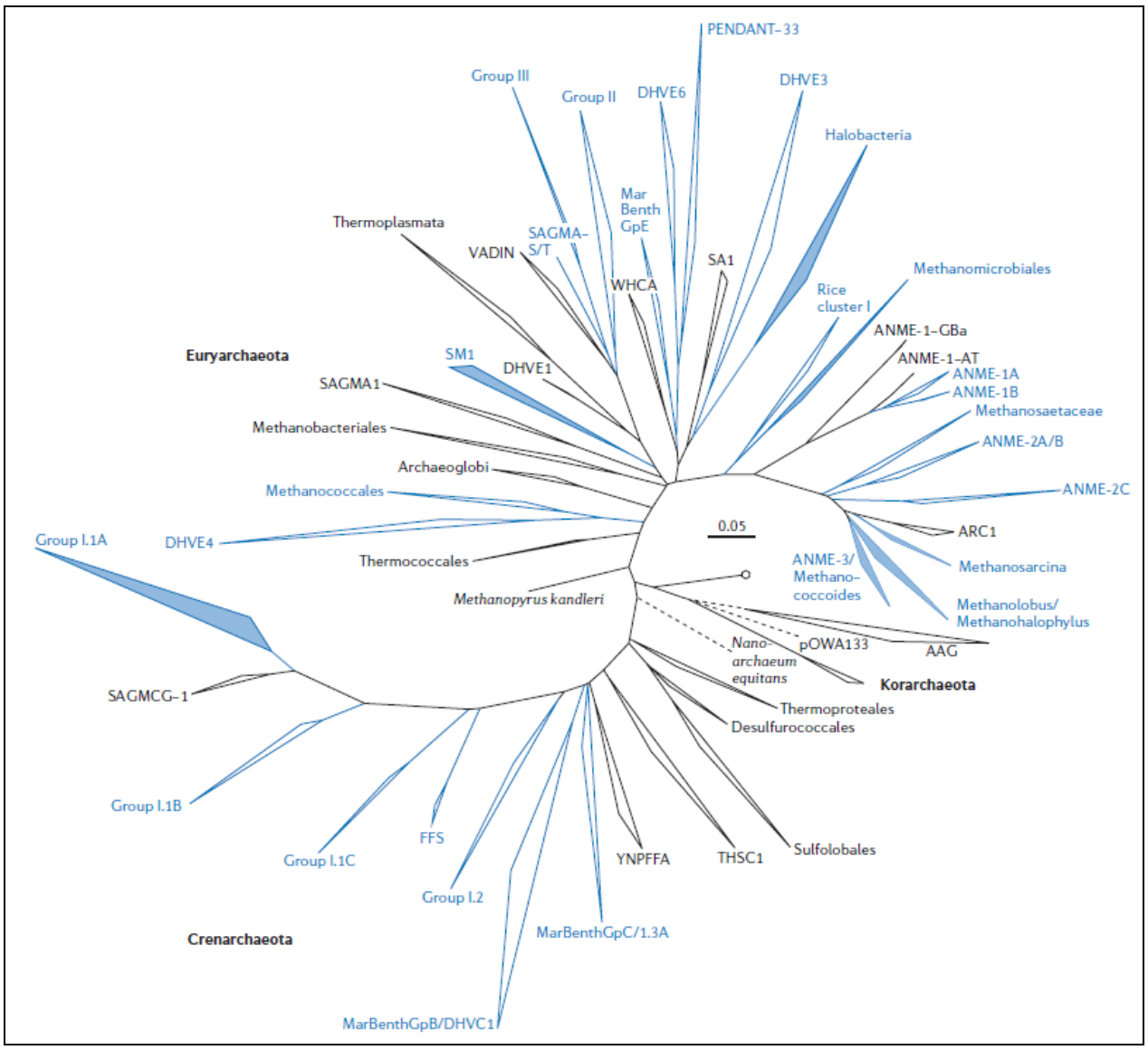


Figure 10. Diversité des archées mésophiles. Adapté de Schleper et collaborateurs 2005 (Christa Schleper, Jurgens, and Jonuscheit 2005) par Cavichioli en 2006 (Cavicchioli 2006), cet arbre montre en bleu les groupes d'archées connues pour vivre dans des environnements froids ou dont la description contient le mot « cold ».

La Figure 10 montre, en bleu, les groupes d'archées décrites dans des milieux froids. Parmi celles-ci on retrouve les archées du groupe II décrites dans des échantillons récoltés en Antarctique par DeLong en 1994 (DeLong et al. 1994), ou les archées du groupe SM1, identifiées dans des marais sulfureux à une température d'environ 10°C (Rudolph, Wanner, and Huber 2001). Sans avoir besoin de descendre à des températures aussi basses, de nombreuses archées sont observées dans des milieux mésophiles, tels que notre bouche ou des rizières pour certains méthanogènes (Knittel and Boetius 2009). La plupart des halophiles sont aussi mésophiles, et vivent par exemple dans des lacs salés ou sur la surface d'aliments soumis à salaison.

Halophiles

Un autre exemple d'extrêmophilie est l'halophilie, l'adaptation à des milieux hypersalins, tels que la Mer Morte ou des aliments salés. Deux ordres d'archées phylogénétiquement proches sont particulièrement adaptés à ces milieux, les *Halobacteriales* et les *Nanohaloarchaea*. Ces organismes sont capables de croître à des concentrations en sel entre 2,5 et 5,2 M de NaCl pour les plus extrêmes, et entre 0,5 et 2,5 M pour les halophiles modérés. Les environnements dans lesquels ils se développent sont très divers, dans la mesure où la concentration en sel est forte, par exemple, *Haloterrigena turkmenica* a été isolé dans des sols (Zvyagintseva and Tarasov 1988), *Halalkalicoccus jeotgali* sur des crustacés fermentés au sel (Roh et al. 2010) et *Haloferax volcanii* vit dans la Mer Morte (Mullakhanbhai and Larsen 1975).

Les *Halobacteriales* sont pour la plupart aérobies et contiennent des pigments de type caroténoïdes, donc rouges, ce qui donne à leurs colonies des couleurs vives et caractéristiques, comme l'on peut le voir sur la Figure 11. Certains sont anaérobes et photohétérotrophes grâce à la bactériorhodopsine, une pompe à proton spécifique à ce groupe, utilisant la lumière comme source d'énergie. Le gradient de protons généré permet la synthèse d'ATP (Hartmann, Sickinger, and Oesterhelt 1980). Une caractéristique génomique de ce groupe est la présence de plusieurs chromosomes ou de nombreux megaplasmides pouvant contenir jusqu'à plusieurs centaines de gènes, alors que le génome de la plupart des archées se trouve sur un seul chromosome et, potentiellement, quelques plasmides de petite taille. Enfin, on peut rappeler ici que la première archée décrite en 1880 par Farlow était une halophile, *Halococcus* (Farlow 1880; KOCUR and Hodgkiss 1973).

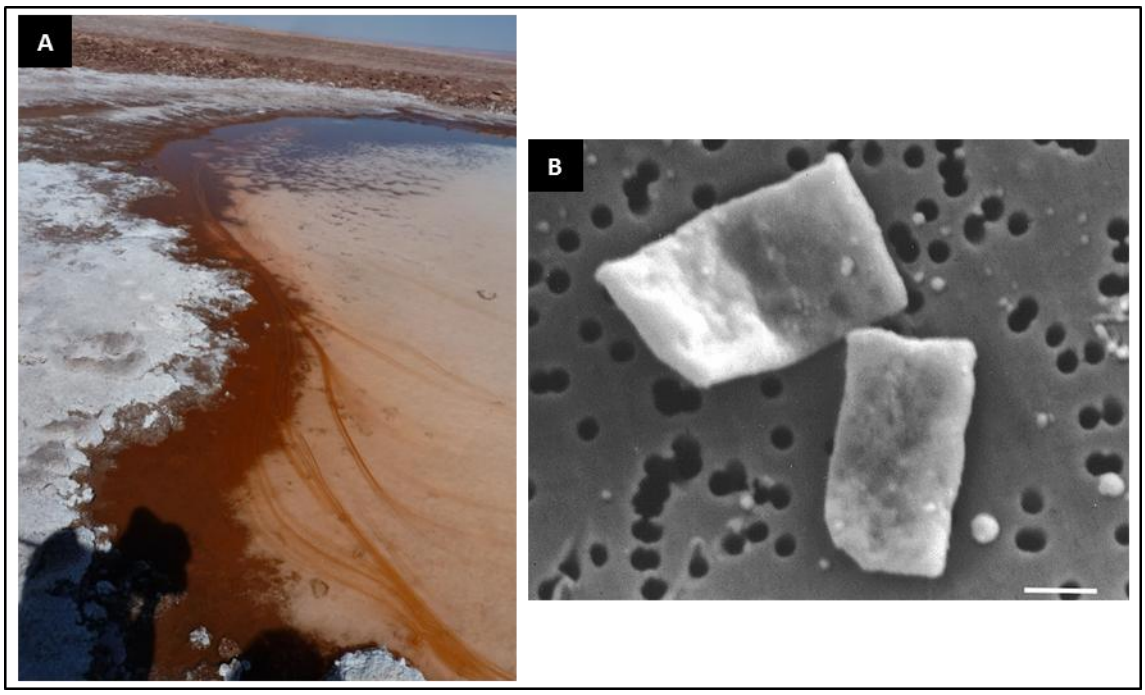


Figure 11. Archées halophiles. A : Salar du Lac d'Atacama au Chili. Les trainées rouges sont des colonies d'archées halophiles filamenteuses. Photographie de Purificación López-García. B : Haloquadratum walsbii, une halobacteriale à la forme caractéristique carrée et plate. Photographie de Francisco Rodríguez-Valera. Barre= 1µm

Acidophiles

Enfin, des archées sont adaptées aux milieux extrêmement acides, particulièrement les *Thermoplasmatales*, *Aciduliprofundum boonei*, archée du groupe DHVE2, phylogénétiquement proches des *Thermoplasmatales*, et les archées du groupe ARMAN. La plupart des *Thermoplasmatales* et les ARMAN ont été mises en évidence dans des mines acides ou dans des champs hydrothermaux volcaniques (Figure 12), à des pH compris entre 1 et 2 (Dopson et al. 2004; Darland et al. 1970). L'organisme *Picrophilus torridus*, isolé dans un sol volcanique sulfureux au Japon, est capable de croître à un pH de 0 (C Schleper et al. 1995). L'ordre *Thermoplasmatales* est aussi caractérisé par l'absence de paroi cellulaire (Darland et al. 1970; Dopson et al. 2004; C Schleper et al. 1995), leur donnant ainsi une morphologie proche des bactéries du groupe *Mycoplasmatales* avec lesquels ces organismes étaient classés jusqu'à ce que Woese analyse leurs ARNr SSU(C R Woese, Magrum, and Fox 1978). Ces organismes acidophiles peuvent aussi être thermophiles, vivant à des températures comprises entre 35 et 70°C. *Aciduliprofundum boonei* a été isolé depuis des échantillons de cheminées hydrothermales sous-marines. Cet organisme vit à un pH légèrement plus élevé (entre 3 et 4) mais à des températures plus hautes, autour de 70°C (Reysenbach et al. 2006).

Figure 12. Mine acide à Rio Tinto en Espagne. La couleur rouge-brune de l'eau sur les photos de gauche est due à la présence de métaux, et particulièrement de fer oxydé en solution. L'eau de la photo de droite est à pH très bas, autour de 1, et sa couleur verte est due à du fer ferreux, non oxydé. Photographie de Purificación López-García.

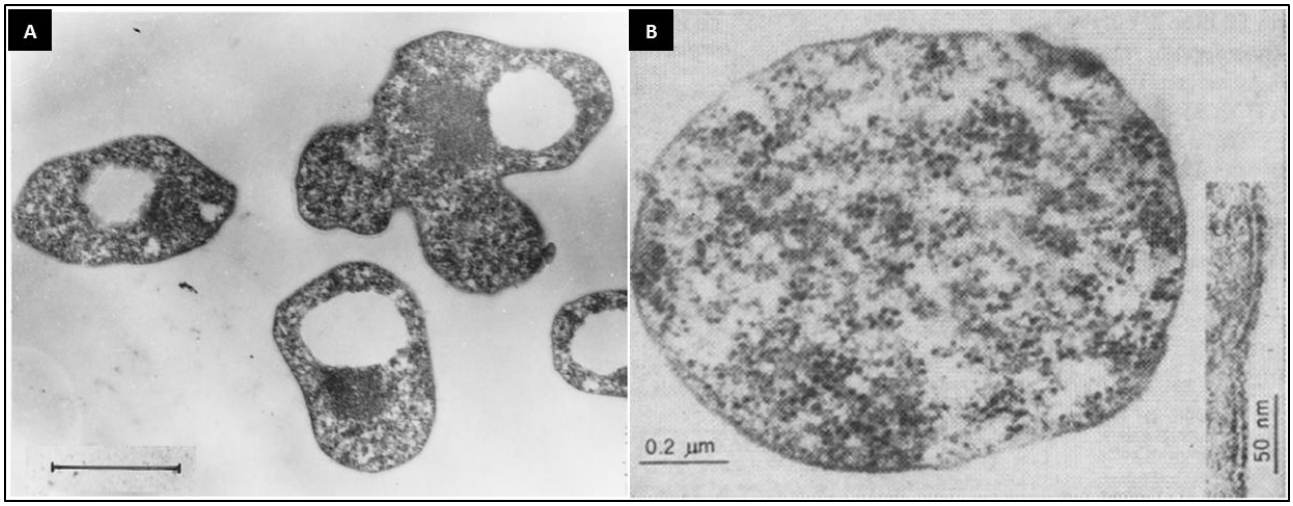


Figure 13. Archées acidophiles. A : Picrophilus oshimae en coupe fine. Cette archée a été isolée dans une source chaude sulfureuse au Japon, à une température de l'ordre de 55°C et un pH inferieur à 0,5 (C Schleper et al. 1995) Barre=1µm. B : Thermoplasma acidophilum en coupe fine à gauche. Elargissement sur la membrane à droite, montrant l'absence de paroi cellulaire, caractéristique des Thermoplasmatales (Darland et al. 1970).

Méthanogènes et méthanotrophes

Les méthanogènes sont des organismes capables de produire du méthane à partir de composés organiques ou inorganiques (CO_2, méthanol ou composés méthylés), le méthane étant le produit principal de ce métabolisme énergétique. Seules des archées possèdent les enzymes nécessaires à ce métabolisme complexe (Offre, Spang, and Schleper 2013; Y. Liu and Whitman 2008). Actuellement sept ordres de méthanogènes ont été décrits, tous des euryarchées, bien que ces ordres ne forment pas un ensemble monophylétique. Les *Methanopyrales*, les *Methanobacteriales* et les *Methanococcales* sont plus proches entre eux et ont été proposés comme « Methanogen Class I » par Bapteste et collaborateurs en 2005 (Eric Bapteste, Brochier, and Boucher 2005). Les *Methanosarcinales*, les *Methanomicrobiales* et les *Methanocellales* (anciennement Rice Cluster I (Sakai et al. 2008)) sont plus proches entre eux et ont été proposés comme « Methanogen Class II » (Eric Bapteste, Brochier, and Boucher 2005). Cependant, la monophylie de ces deux « classes » est toujours discutée (Brochier-Armanet, Forterre, and Gribaldo 2011). Le septième ordre de méthanogènes a été proposé très récemment par Paul et collaborateurs (Paul et al. 2012) suite à l'analyse de nombreuses séquences environnementales se plaçant à proximité des *Thermoplasmatales* (Mihajlovski, Alric, and Brugère 2008; Paul et al. 2012). Le gène mrcA qui code pour la sous-unité alpha du méthyl-coenzyme M, marqueur moléculaire classique des archées méthanogènes, a été amplifié à partir des mêmes échantillons que les ARNr SSU étudiés. Les phylogénies de ces deux marqueurs (mrcA et ARNr SSU) correspondent, renforçant l'idée que ces séquences environnementales représenteraient en réalité un nouvel ordre d'archées méthanogènes, et le nom de *Methanoplasmatales* a été proposé. Ces organismes ont été observés dans des milieux très divers, mais toujours anoxiques, tels que des sédiments marins, d'eau douce, des sols de marais, des rizières, le rumen de bovins, les tractus gastro-intestinaux de différents animaux (gros mammifères, insectes…), des tourbières, mais aussi dans des milieux considérés comme plus extrêmes, comme des sources chaudes, des hydrocarbures pétroliers ou des sédiments hypersalins (Y. Liu and Whitman 2008).

Les groupes ANME-1, -2 et -3 sont des groupes d'euryarchées non cultivées méthanotrophes, détectées dans les mêmes environnements que les méthanogènes (Knittel and Boetius 2009; Mills et al. 2005; Offre, Spang, and Schleper 2013). Ces organismes sont caractérisés par leur capacité à oxyder le méthane en anaérobiose. Il est donc logique qu'ils vivent dans les mêmes milieux que les méthanogènes. La plupart du temps des bactéries capables de réduire le sulfate se trouvent aussi dans ces mêmes milieux, ce qui permettrait une oxydation anaérobique du méthane (« AOM » pour « Anaerobic Oxydation of Methane ») dépendante du sulfate avec production d'énergie pour les archées ANME et les bactéries sulfato-réductrices (Offre, Spang, and

Schleper 2013; Knittel and Boetius 2009). D'un point de vue environnemental, ce processus AOM balancerait la production de méthane biologique et son émission dans l'atmosphère puisqu'il consommerait une grande part de cette production, particulièrement dans les environnements marins (Reeburgh 2007). L'impact écologique de ces deux populations d'archées, méthanogènes et méthanotrophes, est donc très important.

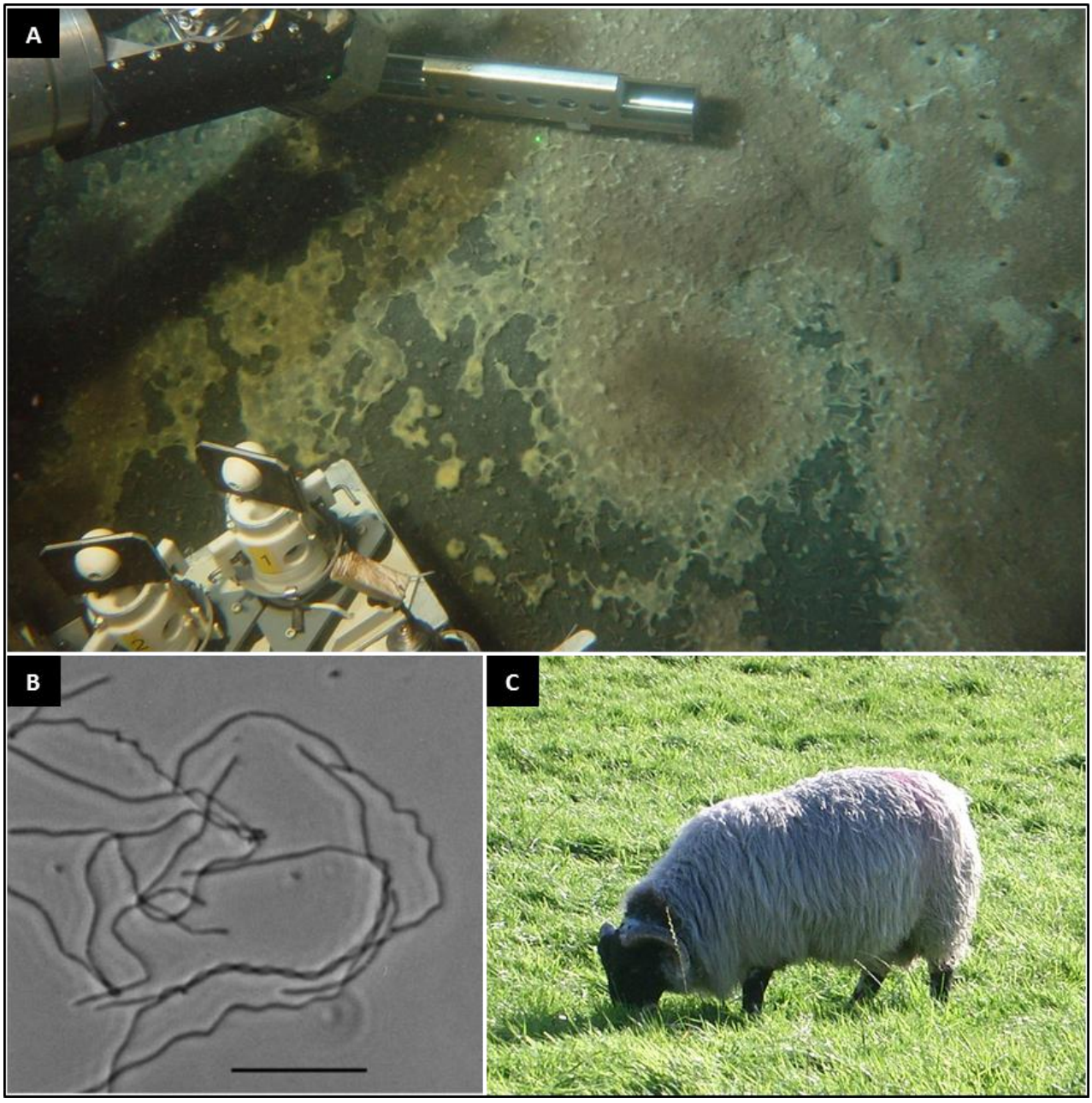


Figure 14. Archées méthanogènes et leurs habitats. A : Tapis microbien anoxique sous-marin, dans lequel sont présentes des archées méthanogènes et des méthanotrophes. Photographie de Purificación López-García. B : Methanobacterium oryzae, une archée méthanogène isolée dans un champ de riz aux Philippines. Barre= 10µm. (Joulian et al. 2000). C : Ovis arie. Certains méthanogènes ont été isolés depuis leur tractus digestif, comme Methanosarcina barkeri. Photographie de Celine Petitjean.

b. Les Nanoarchaeota

Actuellement *Nanoarchaeum equitans* est le seul représentant cultivé de ce phylum et dont le génome ait été séquencé (H. Huber et al. 2002). C'est une archée hyperthermophile, de toute petite taille (~ 400 nm), avec un génome très réduit et vivant justement en symbiose avec *Ignicoccus hospitalis*, une crenarchée avec laquelle elle a pu être cultivée (H. Huber et al. 2002) (Figure 13). Le séquençage de son génome a révélé des caractéristiques particulières typiques des organismes symbiotiques (Waters et al. 2003), à savoir un génome très réduit avec absence de beaucoup de gènes essentiels du métabolisme (biosynthèse des acides aminés, nucléotides, cofacteurs et lipides), des gènes intervenant dans les différentes voies d'assimilation du carbone ou de trois ARNt (glutamate, histidine et tryptophane). L'absence de ces gènes rend *N. equitans* complètement dépendant d'*I. hospitalis* et tend à confirmer son évolution par réduction de génome due à un mode de vie symbiotique. On trouve aussi 10 gènes coupés en deux dans ce génome alors qu'ils sont codés en une seule séquence chez la majorité des autres archées. Ce nombre est assez important comparativement aux autres archées, et certains des gènes concernés sont impliqués dans des processus cellulaires essentiels, comme l'ADN polymérase I, l'alanyl-ARNt synthétase ou la sous-unité β de l'ARN polymérase. Sa position phylogénétique et la justification de la création d'un phylum à part entière seront discutées plus tard.

Très récemment un deuxième génome de Nanoarchaeota a été séquencé par métagénomique : « Nst1 » (Podar et al. 2013). Cet organisme vivrait en symbiose avec une crenarchée de l'ordre des *sulfolobales*. Bien que son génome soit environ deux fois plus grand (de l'ordre de 1 mégabase) que celui de *N. equitans* (de l'ordre de 0,5 mégabase), on retrouve aussi des pertes massives de gènes du métabolisme et un certain nombre de gènes coupés en deux, certains communs avec *N. equitans*. Ce génome semble donc avoir subi une évolution par réduction mais moins drastique que celui de *N. equitans*.

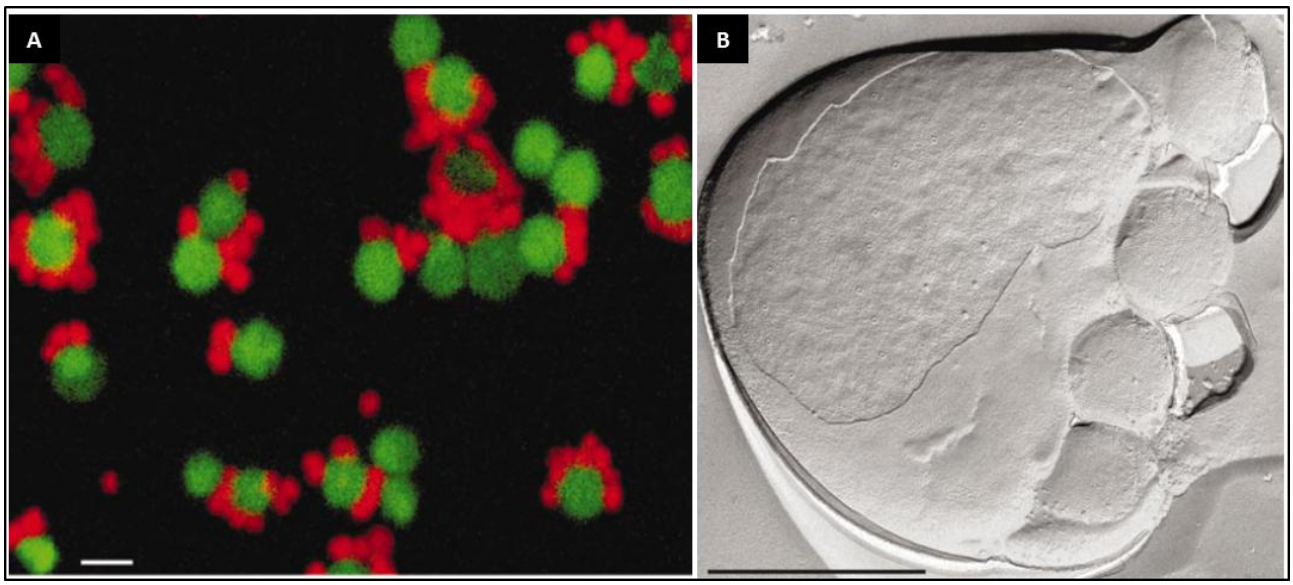


Figure 15. Nanoarchaeum equitans et Ignicoccus hospitalis, son hôte. (H. Huber et al. 2002). Barre= 1µm. A : Image de microscopie confocale laser après hybridation avec le marqueur rouge CY3 pour Nanoarchaeum equitans et le marqueur vert rhodamine-green pour I. hospitalis. B : Image de microscopie électronique d'I. hospitalis et de quatre cellules de N. equitans accrochées.

c. Les Crenarchaeota

Les Crenarchaeota sont des organismes hyperthermophiles ou thermophiles répartis en cinq ordres : les *Desulfurococcales*, les *Sulfolobales*, les *Thermoproteales*, ainsi que les *Acidilobales* et les *Fervidicoccales*, deux nouveaux ordres proposés en 2009 (Prokofeva et al. 2009) et en 2010 (Perevalova et al. 2010). La plupart des *Thermoproteales* et des *Desulfurococcales* sont anaérobes et hyperthermophiles, alors que les *Sulfolobales* ont plutôt tendance à être aérobes et que leurs températures de croissance sont plus basses que celles observées dans les deux autres groupes. Les environnements dans lesquels on les trouve peuvent être les mêmes que ceux dans lesquels sont présents les euryarchées hyperthermophiles (Figure 8), à savoir des sources chaudes terrestres ou des sources hydrothermales sous-marines, de surface ou profondes, par exemple les fumeurs noirs et blancs présents au niveau des dorsales océaniques (Figure 16).

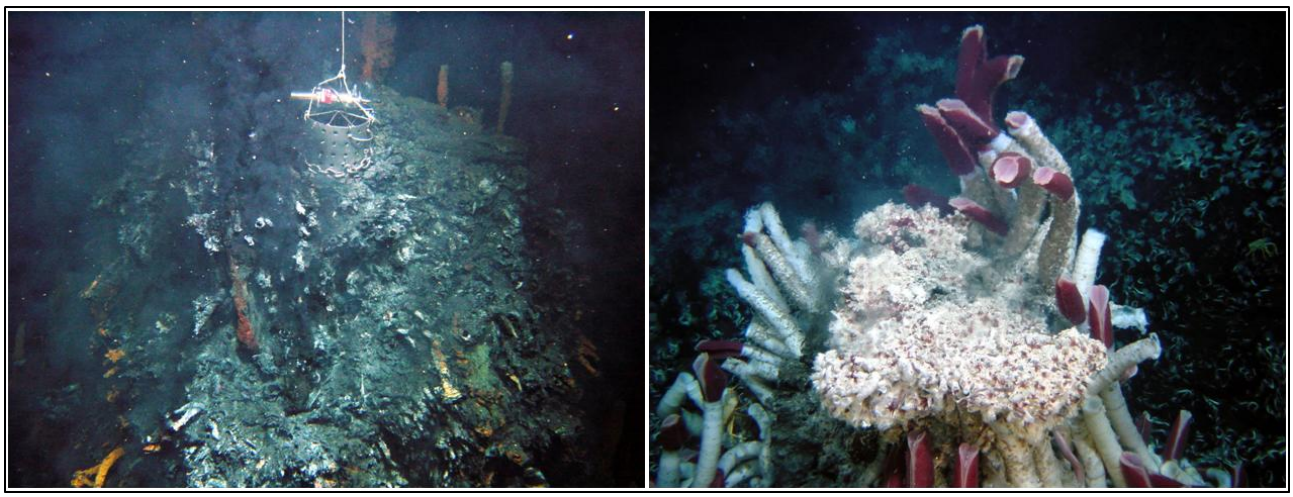


Figure 16. Fumeur noir de la dorsale océanique East Pacific Rise. Environnement hyperthermophile dans lequel se trouvent particulièrement des Desulfurococcales. Photographie de Purificación López-García.

Les *Desulfurococcales* sont les crenarchées les plus hyperthermophiles. Leurs températures de croissance varient autour de 90°C, et peuvent monter jusqu'à plus de 100°C. Par exemple, *Pyrolobus fumarii* a une température optimale de croissance de 106°C mais est capable de croître jusqu'à 113°C (Anderson, Göker, et al. 2011). Cette archée a été isolée dans des fumeurs hydrothermaux des fonds sous-marins, l'augmentation de la pression avec la profondeur permettant de telles températures. La majorité des *Desulfurococcales* sont anaérobes, stricts ou facultatifs, et leurs métabolismes sont assez variés. Par exemple, *Hyperthermus butylicus* (Brügger et al. 2007) ou *Ignicoccus hospitalis* (Podar et al. 2008) sont anaérobes et sulfato-réductrices et ont été isolées dans des sources hydrothermales sulfureuses. Certaines sont hétérotrophes, comme *Staphylothermus hellenicus* (Anderson, Wirth, et al. 2011), *Thermosphaera aggregans* (Spring et al. 2010), ou *Ignisphaera aggregans* (Göker et al. 2010). *Desulfurococcus fermentans*, isolée dans une source chaude au Kamchatka en Russie, est même capable de dégrader la cellulose (Susanti et al. 2012). *Ignicoccus hospitalis*, l'hôte de *Nanoarchaeum equitans*, est une désulfurococcale isolée à partir d'une source chaude sous-marine en Islande. Elle a une température de croissance de 95°C (Podar et al. 2008). Parmi les membres du genre *Ignicoccus*, seul *I. hospitalis* est l'hôte de *N. equitans* ou d'un symbiote tout court. Les *Ignicoccus* ont une cytologie particulière avec une double membrane cellulaire et un périplasme très important entre la membrane cellulaire interne et la membrane externe, dans lequel de nombreuses vésicules sont présentes (Rachel et al. 2002).

Les *Acidilobales* ont été proposés récemment par Prokofeva et collaborateurs en 2009 (Prokofeva et al. 2009), et incluraient les organismes des genres *Acidilobus* et *Caldisphaera*. Ces organismes sont thermophiles, anaérobes, organothrophes et acidophiles. Ils ont été isolés dans différentes sources chaudes acides, au Kamchatka en Russie, à Laguna aux Philippines ou dans le parc de Yellowstone aux Etats-Unis. Les analyses phylogénétiques de leur ARNr SSU les placent

groupe frère des *Desulfurococcales*, conduisant les auteurs à proposer un nouvel ordre de Crenarchaeota. Le génome complet d'*Acidilobus saccharovorans*, isolée au Kamchatka, a été séquencé en 2009 (Prokofeva et al. 2009). Il a révélé de nombreuses voies métaboliques laissant supposer qu'il est capable d'utiliser une très large gamme de composés organiques produits par les organismes lithoautotrophes vivant dans les mêmes environnements.

Peu de temps après, en 2010, un nouvel ordre de Crenarchaeota est proposé, les *Fervidicoccales* (Perevalova et al. 2010), après l'isolement de *Fervidicoccus fontis* dans les sources chaudes volcaniques du Kamchatka. C'est un organisme thermoacidophile (croissance optimale entre 65 et 70°C, à pH 5,5-6), anaérobe et organotrophe. L'analyse phylogénétique de l'ARNr SSU de cet organisme le place en groupe frère des *Acidilobales*, avec d'autres séquences environnementales trouvées dans des environnements de type sources hydrothermales à Yellowstone ou en Islande. Cette position phylogénétique a conduit les auteurs à proposer que ces séquences soient les représentantes d'un nouvel ordre.

Les *Thermoproteales* sont aussi des organismes hyperthermophiles et anaérobes pour la plupart. Ils vivent aussi dans ces mêmes milieux et principalement dans des sources chaudes terrestres sulfureuses, les solfatares, à des températures légèrement plus basses que les *Desulfurococcales*. Certains sont légèrement acidophiles et ont été isolés à des pH inferieurs à 7, comme *Thermoproteus tenax*, isolé à un pH de 5,6 dans un solfatare en Islande (Zillig et al. 1981; Siebers et al. 2011). On trouve aussi une variété importante de métabolismes : des organismes organotrophes comme *Caldivirga maquilingensis*, hétérotrophe et microaérophile, des organismes capables de croître de façon lithotrophe et organotrophe, comme *Pyrobaculum arsenaticum*, utilisant l'hydrogène comme donneur d'électrons pour la réduction de l'arsenate et de composés sulfurés pour une croissance chimiolithoautotrophe et qui utilise ces derniers composés comme accepteurs d'électrons pour une croissance organotrophe (R. Huber et al. 2000).

Enfin les *Sulfolobales* sont des organismes thermophiles ou hyperthermophiles plus modérés que les autres ordres, leurs températures de croissance sont plutôt comprises entre 60 et 85°C, et on les trouve principalement dans des sources chaudes terrestres, comme celles de Solfatare en Italie, de Yellowstone ou du Kamchatka. On trouve aussi plus d'organismes aérobies dans ce groupe, alors qu'ils sont minoritaires dans les autres ordres de Crenarchaeota, et d'acidophiles extrêmes, vivant à des pH entre 2 et 5. Par exemple *Sulfolobus acidocaldarius*, le premier hyperthermoacidophile caractérisé depuis un solfatare terrestre par Brock et collaborateurs en 1972 (Brock et al. 1972), croît entre 75 et 80°C, à un pH entre 2 et 3, et est strictement aérobie et organotrophe (Chen et al. 2005).

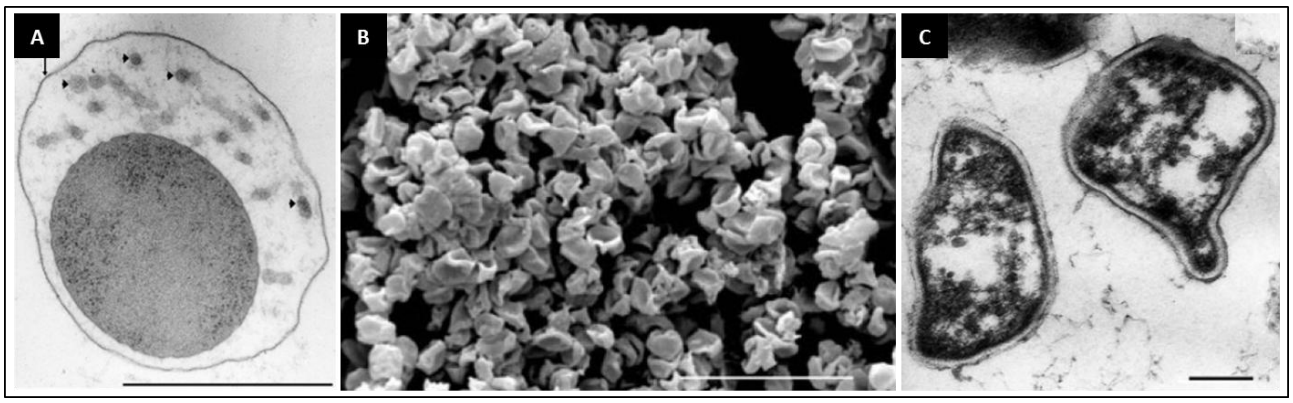


Figure 17. Crenarchées. A : Image de microscopie électronique d'une cellule d'Ignicoccus (Rachel et al. 2002). On observe le large périplasme entre la membrane interne et la membrane externe. Barre=1µm. B : Colonie de cellules de Metallosphaera cuprina, une sulfolobale. Barre= 5µm. (L.-J. Liu et al. 2011). C : Coupe fine de cellules d'Acidilobus saccharovorans, une acidilobale. Barre= 0,5µm. (Prokofeva et al. 2009).

d. Les Thaumarchaeota

Les Thaumarchaeota ont été découverts par DeLong et Furhmann en 1992 et nommés « groupe I ». Phylogénétiquement proches des Crenarchaeota elles étaient considérées comme des « Crenarchaeota mésophiles » (DeLong 1992; Fuhrman, McCallum, and Davis 1992). En 2008, après la publication du génome de *Cenarchaeum symbiosum*, Brochier-Armanet et collaborateurs ont réanalysé la position de ce groupe par des méthodes de phylogénomique et ont proposé que les membres du groupe I ne soient pas des Crenarchaeota mais un phylum à part entière. Ils ont proposé le nom de « *Thaumarchaeota* » (Brochier-Armanet et al. 2008). De nombreux organismes se placent parmi les Thaumarchaeota, mais très peu sont cultivés axéniquement, le premier a été *Nitrosopumilus maritimus* (Figure 18), 18 ans après la découverte de ces archées (Könneke et al. 2005).

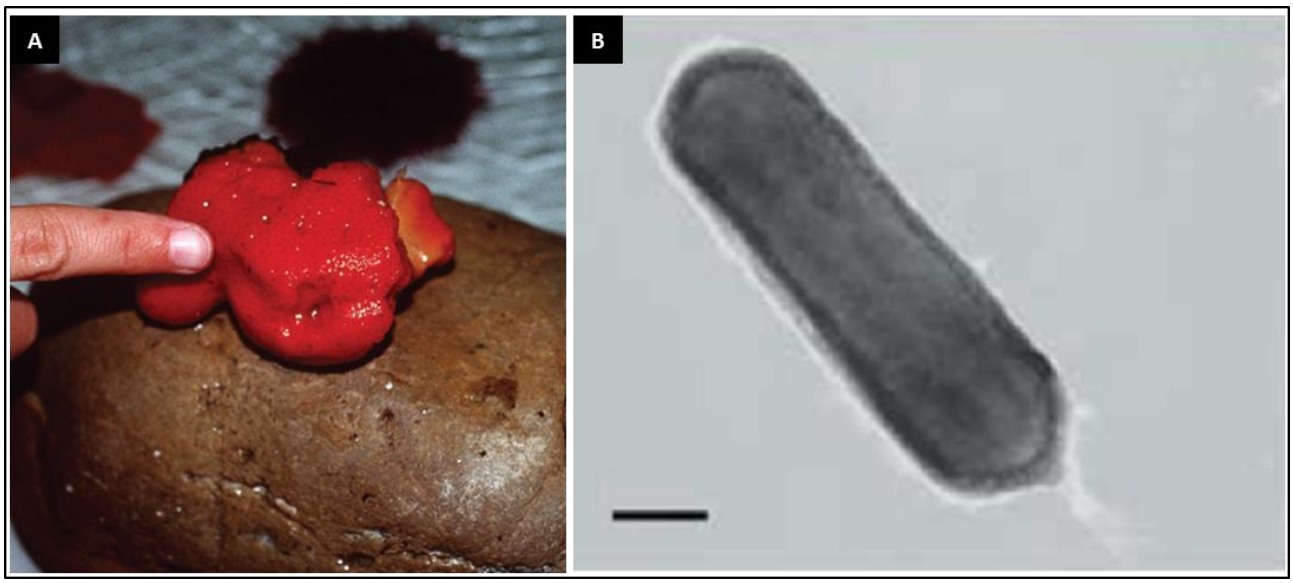


Figure 18. Thaumarchées. A : Axinella mexicana, une éponge, est l'hôte de Cenarchaeum symbiosum (Delong 2003). B : Nitrosopumilus maritimus, première thaumarchée isolée en 2005, vivant dans les milieux marins (Könneke et al. 2005) Barre=0,1µm.

Les Thaumarchaeota sont des organismes capables de réaliser la première étape de la nitrification, à savoir l'oxydation aérobie de l'ammoniac, NH_3, en nitrite, NO_2^- . Cette capacité est très rare dans le vivant, et jusqu'à la description des thaumarchées elle n'était connue que chez quelques beta- et gamma-protéobactéries (AOB, pour Ammonia-Oxydizing Bacteria). Cette réaction est catalysée par l'ammonia monooxygénase AMO, composée de trois sous-unités : AmoA, AmoB et AmoC chez les archées (Stahl and de la Torre 2012). Les thaumarchées dominent une partie de la biosphère des océans en eau profonde, comme l'ont montré Karner et collaborateur en 2001 (Karner, DeLong, and Karl 2001). Leur capacité d'oxydation de l'ammoniac en nitrite associée à l'importance en biomasse de ces organismes en fait des acteurs majeurs du cycle biogéochimique de l'azote.

Depuis, de nombreuses études environnementales ont démontré leur présence dans la plupart des environnements. En effet, bien qu'elles aient été découvertes dans des milieux mésophiles, des thaumarchées sont présentes dans des milieux extrêmement divers, et pas uniquement mésophiles. Dans l'océan profond, bien sûr, ou moins profond, par exemple *Cenarchaeum symbiosum* qui vit en symbiose avec l'éponge *Axinella mexicana* (Preston et al. 1996).On les trouve aussi dans les sols (*Nitrososphaera viennensis* (Tourna et al. 2011)), dans des sédiments marins ('*Candidatus* Nitrosopumilus salaria' (Mosier et al. 2012) ou '*Candidatus* Nitrosopumilus sediminis' (Park et al. 2012)). D'autres lignées sont présentes dans des milieux thermophiles tels que des tapis microbiens de sources chaudes (*Nitrososphaera gargensis* (Hatzenpichler et al. 2008)) ou dans des sources chaudes à Yellowstone ('*Candidatus* Nitrosocaldus yellowstonii' (de la Torre et al. 2008)) ; ces deux

derniers organismes sont donc thermophiles. Récemment de nouveaux groupes très divergents, HTC1 et HTC2, pour « Hot Thaumarchaeota-related Clade » 1 et 2, ont été décrits par Eme et collaborateurs en 2013 (Eme et al. 2013) à partir d'échantillons de sources chaudes du Kamchatka. La Figure 19 montre une phylogénie de l'ARNr SSU des Thaumarchaeota issue de cet article, représentative de leur diversité.

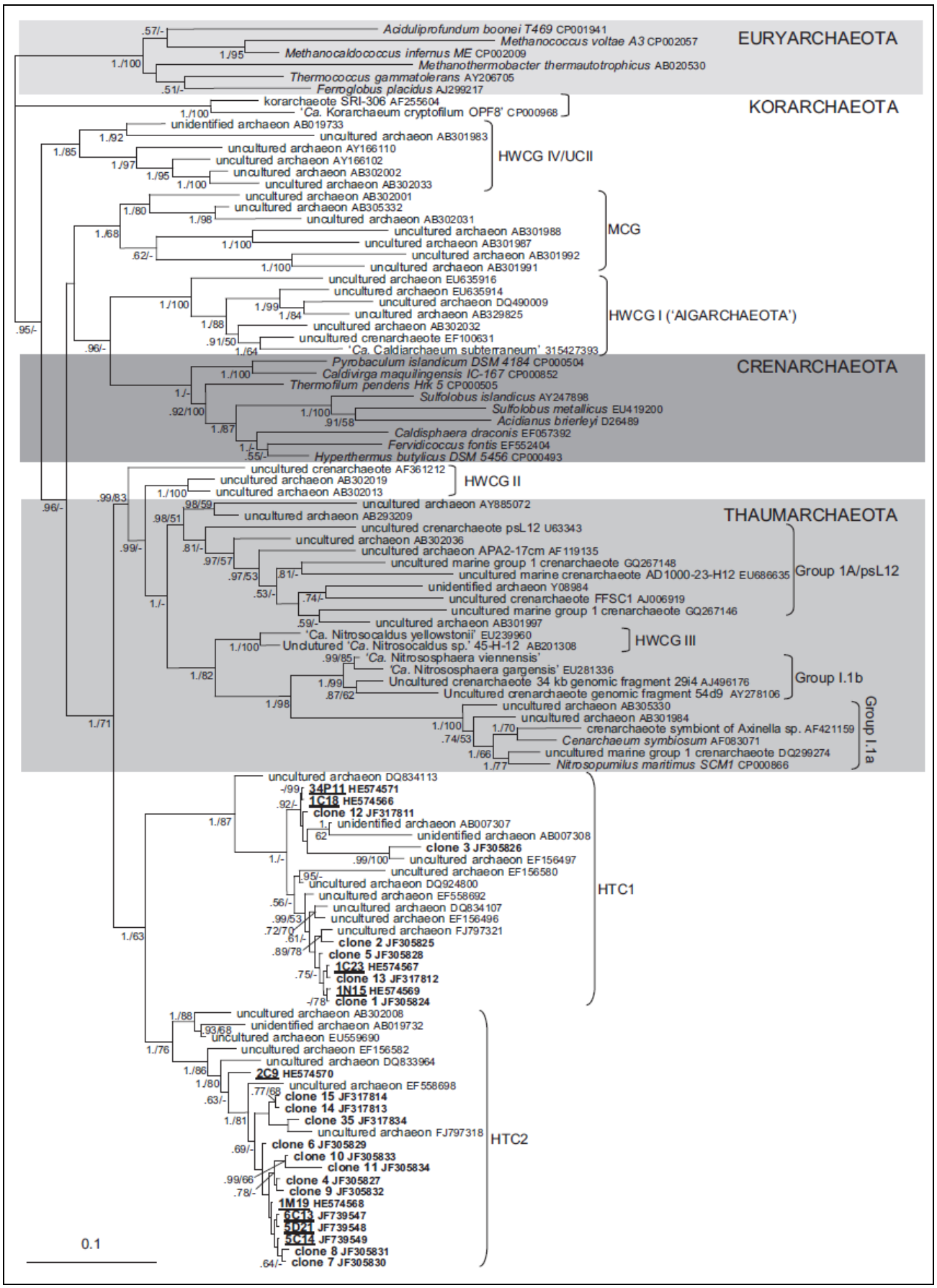


Figure 19. Phylogénie des Archées centrée sur les Thaumarchaeota non cultivées. Arbre non raciné, inféré sur 1064 positions de 105 séquences d'ARNr SSU (Eme et al. 2013).

Il est intéressant de noter que des organismes mésophiles et thermophiles sont présents chez les Thaumarchaeota mais que les organismes mésophiles sont polyphylétiques, plusieurs transitions ont donc dû se produire de la thermophilie à la mésophilie ou inversement. Plusieurs travaux de López-García et collaborateurs ont montré que les transferts horizontaux de gènes ont joué un rôle important dans l'évolution du génome des Thaumarchaeota marines (groupe I) et particulièrement dans l'adaptation à leur milieu océanique profond froid et pauvre en nutriments (López-garcía, Brochier, Moreira, & Rodríguez-valera, 2004 ; Brochier-Armanet et al., 2011) (cf. Chapitre 3).

La Figure 19 illustre également le fait que la majeure partie de la diversité des thaumarchées est non cultivée. Ainsi, très peu de génomes complets sont disponibles. En dehors de celui de *Cenarchaeum symbiosum* en 2006 (Hallam et al. 2006), de plus en plus de génomes sont séquencés pour ce phylum ces deux dernières années, mais la plupart appartiennent au groupe I.1a, proche de *Nitrosopumilus maritimus*, séquencé en 2010 (Walker et al. 2010).

e. Les Aigarchaeota

Ce phylum proposé par Nunoura et collaborateurs en 2011 (Nunoura et al. 2011) correspond au groupe HWCG I, visible sur la Figure 19, où il se place en groupe frère des Crenarchaeota. Ce phylum a été proposé en même temps que la publication du génome de '*Candidatus* Caldiarchaeum subterraneum' dont il a déjà été question plus tôt. Ces organismes sont présents dans des environnements thermophiles classiques : sources chaudes terrestres (Nunoura et al. 2005) et sous-marines, sources hydrothermales de mer profonde (Reykjavı et al. 2001). Aucun représentant n'est cultivé. Le génome de '*Ca.* Caldiarchaeum subterraneum' a été obtenu à partir de l'enrichissement d'un tapis microbien de surface dans de l'eau géothermale d'une mine d'or, dans lequel ce groupe HWCGI était dominant, et a été reconstruit à partir d'une banque de données métagénomiques (Nunoura et al. 2011). Dans cette publication, les analyses phylogénétiques placent '*Ca.* Caldiarchaeum subterraneum' en groupe frère des Thaumarchaeota.

f. Les Korarchaeota

Le phylum Korarchaeota a été proposé en 1996 par Barns et collaborateurs (Barns et al. 1996) sur la base de séquences environnementales produites à partir d'une source chaude de Yellowstone. En 2006 Auchtung et collaborateurs ont étudié la diversité des Korarchaeota en construisant des amorces ciblant spécifiquement l'ARNr SSU de ce groupe. Ils ont échantillonné un grand nombre d'environnements hydrothermaux, des sources chaudes à Yellowstone et des fumeurs

hydrothermaux de fonds sous-marins, mais aussi d'environnements non hydrothermaux tels que du compost, des végétaux, de la salive humaine, des sols, de l'eau, des sédiments de rivières, etc. Le résultat de cette étude est clair : les korarchées sont détectées uniquement dans des environnements thermophiles (température supérieure à 55°C). La plupart des nouvelles séquences, huit, ont été détectées dans les échantillons provenant des différentes sources chaudes de Yellowstone, et une autre séquence dans un échantillon de cheminée hydrothermale sous-marine. Les séquences provenant d'études précédentes sont aussi issues d'environnements thermophiles et l'ensemble forme un groupe monophylétique robuste, branchant avant la divergence entre Euryarchaeota et Crenarchaeota (Auchtung, Takacs-Vesbach, and Cavanaugh 2006). Cette étude confirme la présence de représentants de ce groupe phylogénétique dans divers environnements, mais toujours à haute température. Le seul génome de Korarchaeota disponible à ce jour a été publié en 2008. Il s'agit du génome de '*Ca.* Korarchaeum cryptofilum', reconstruit à partir d'enrichissement de cellules provenant de la source chaude « Obsidian Pool » à Yellowstone (Elkins et al. 2008). Ce génome se révèle assez particulier, avec l'absence de nombreux gènes permettant la synthèse *de novo* de molécules importantes, telles que les purines, le coenzyme A ou d'autres cofacteurs, présents chez la plupart des autres archées. Il partage avec les Euryarchaeota des fonctions liées à la maturation des ARNt, à la réplication et réparation de l'ADN et à la division cellulaire, mais la composition protéique de son ribosome et de son ARN Polymérase est plus proche de celles des Crenarchaeota. Sa position phylogénétique et ses relations avec ces deux groupes peuvent donc être considérées comme cohérentes et ces caractéristiques pourraient avoir été présentes dans leur ancêtre commun (Elkins et al. 2008). Ce groupe reste aujourd'hui encore très peu connu et il serait intéressant d'avoir plus de données, génomiques et de culture, sur ces organismes ayant une place si importante dans l'histoire évolutive des archées.

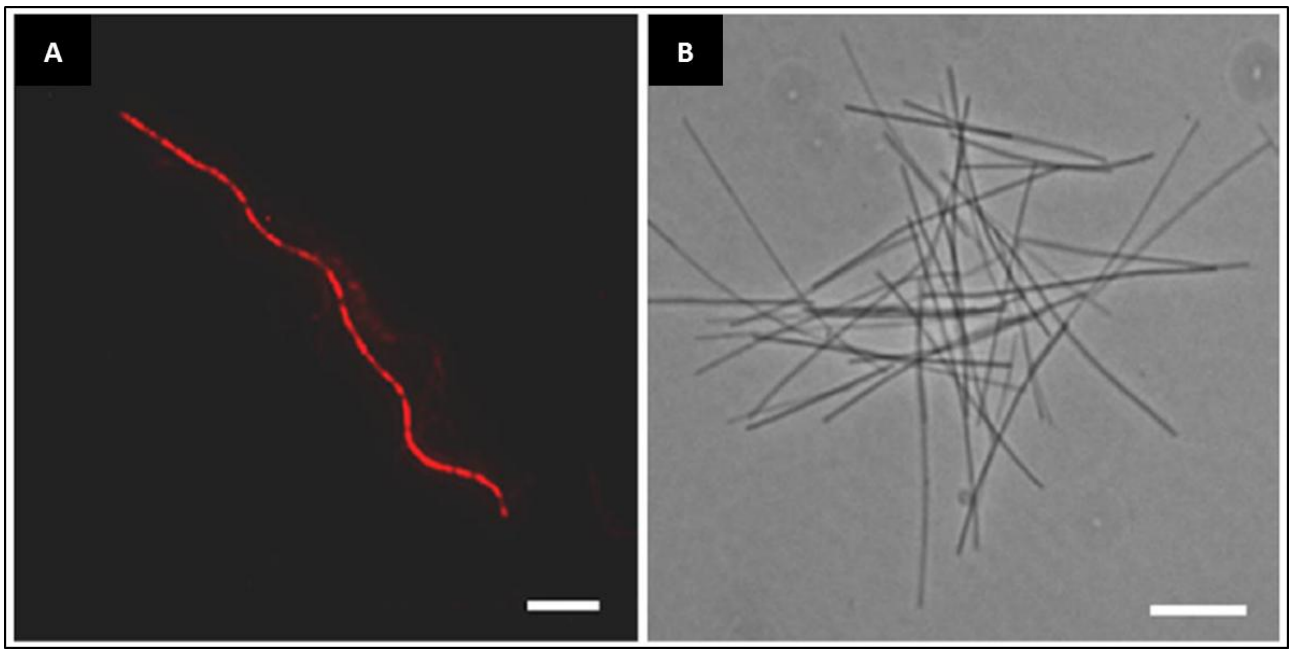


Figure 20. '*Candidatus* Korarchaeum cryptofilum'. Ces deux photographies de microscopie montrent bien la forme filamenteuse des cellules de '*Ca.* Korarchaeum cryptofilum'. Barre= 5µm. A : Analyse en FISH. B : Enrichissement de '*Ca.* Korarchaeum cryptofilum'. (Elkins et al. 2008).

g. Conclusion

La Figure 7, une phylogénie construite à partir de l'ARNr SSU, représente bien la diversité des archées, cultivées ou non, avec des génomes séquencés ou non. Les différents ordres d'Euryarchaeota et de Crenarchaeota apparaissent bien de façon monophylétique, ainsi que les différents groupes de Thaumarchaeota. Malgré tout, les relations entre ces groupes, ordres et phyla, ne sont pas résolues. Ceci montre que l'utilisation d'un seul marqueur, aussi puissant soit-il, n'est pas suffisante pour résoudre les relations anciennes entre autant d'organismes. Le séquençage massif de génomes a ouvert la porte à l'utilisation de nouveaux marqueurs, et surtout à l'utilisation conjointe de ces marqueurs dans des analyses phylogénomiques.

2. Phylogénies moléculaires inférées sur plusieurs marqueurs

a. La phylogénie des archées

Première phylogénie des archées

En 2002, Matte-Taillez et collaborateurs analysent pour la première fois des concaténations de protéines ribosomiques pour inférer la phylogénie des archées (Matte-Tailliez et al. 2002). Précédemment, cette phylogénie n'avait été inférée qu'à partir de l'ARNr SSU. Comme je l'ai déjà évoqué plus tôt, ce marqueur est sujet à certains biais comme la différence de vitesse d'évolution entre séquences, ce qui conduit à de possibles attractions de longues branches (Philippe and Laurent 1998). Le nombre limité de sites qui le composent peut être aussi responsable d'une mauvaise résolution, particulièrement au niveau des relations anciennes entre organismes, à cause du phénomène de saturation mutationnelle (Simonetta Gribaldo and Brochier-Armanet 2009). De plus, la phylogénie de l'ARNr SSU reste la phylogénie d'un seul gène, et pas des organismes dans leur globalité. Bien que l'histoire évolutive de l'ARNr SSU puisse refléter celle des organismes, la confirmation de cette hypothèse par l'utilisation de nouveaux marqueurs est importante. Plusieurs protéines avaient été utilisées dans ce but, mais toujours dans des phylogénies individuelles avec un seul marqueur utilisé à chaque fois ; par exemple, la sous-unité B de l'ARN polymérase (Klenk and Zillig 1994), différentes aminoacyl-ARNt synthétases (C R Woese et al. 2000), ou la chaperonne HSP70/DnaK (R. S. Gupta 2000).

Les protéines ribosomiques sont extrêmement conservées au niveau fonctionnel et sont universelles (présentes dans l'ensemble du vivant), deux qualités qui avaient fait de l'ARNr SSU un marqueur de choix pour la phylogénie des organismes. Treize génomes d'archées etaient disponibles en 2002 : trois Crenarchaeota et dix Euryarchaeota. Les auteurs ont donc analysé la phylogénie de 53 protéines ribosomiques clairement orthologues dans ces génomes, les ARNr SSU et LSU et la sous-unité B de l'ARN polymérase (Matte-Tailliez et al. 2002). A partir de ces 53 protéines, les auteurs ont réalisé différentes concaténations (Figure 21 A et B) d'une part, et d'autre part, fait une analyse de chaque phylogénie individuelle afin de détecter de potentiels transferts horizontaux de gènes. Enfin, ils ont ciblé dans une analyse plus fine les gènes qui apparaissaient comme ayant pu être transférés. Ceci aura permis d'identifier huit protéines porteuses d'un signal divergent. Cette méthode avait été mise en place et utilisée pour l'étude de la phylogénie des bactéries à partir de marqueurs multiples la même année (Brochier et al. 2002).

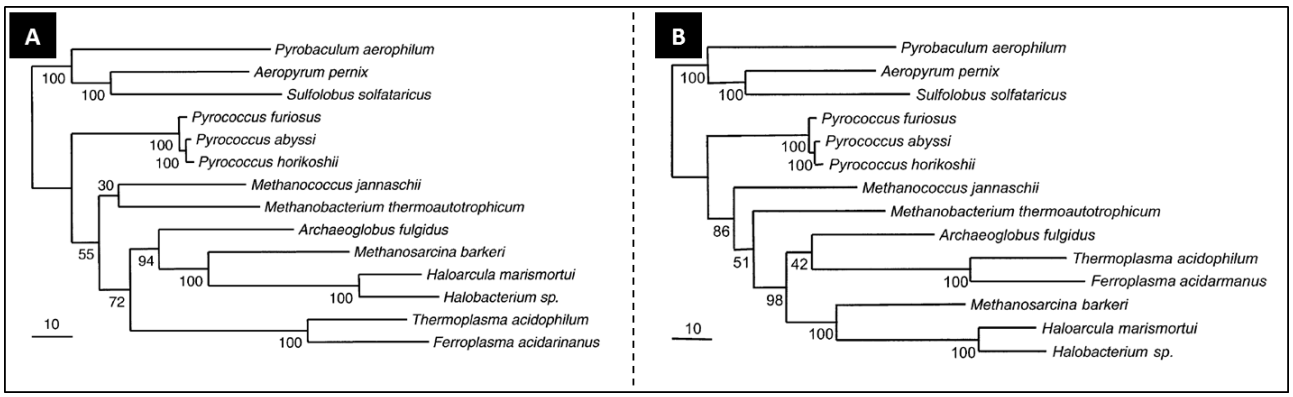


Figure 21. Phylogénies des archées publiées en 2002. Arbres inférés par maximum de vraisemblance. A : à partir de la concaténation de l'ARNr SSU et de l'ARNr LSU (3933 positions). B : à partir d'une concaténation de 45 protéines ribosomiques (6249 positions). (Matte-Tailliez et al. 2002).

La concaténation obtenue à partir de l'ensemble des 45 protéines ribosomiques ne montrant pas de transferts (Figure 21 B) a donné un résultat très similaire à celui obtenu par l'analyse de la concaténation des ARNr SSU et LSU (Figure 21 A) mais globalement avec un meilleur support des branches. Une dernière phylogénie a été inférée à partir de la sous-unité B de l'ARN polymérase. Elle a une topologie très proche de celles obtenues à partir des protéines ribosomiques. Le travail de Matte-Taillez et collaborateurs montre que les protéines ribosomiques portent dans l'ensemble un signal phylogénétique très proche de celui de l'ARNr SSU et de la sous-unité B de l'ARN polymérase. Les différents marqueurs semblent donc partager la même histoire évolutive et les implications de cette découverte sont importantes. Tout d'abord, la congruence de l'histoire évolutive de ces différents systèmes permet donc de poser l'hypothèse que cette histoire soit représentative de l'histoire évolutive des organismes. D'autre part, les travaux de Woese sur la phylogénie des archées basée sur l'ARNr SSU sont donc globalement confirmés par l'analyse de ces nouveaux marqueurs, et l'ARNr SSU semble donc un bon marqueur pour la phylogénie des organismes, même s'il n'apporte pas une résolution optimale. Enfin il apparaît que si certaines protéines ribosomiques semblent avoir subi des transferts au cours de leur histoire évolutive, l'analyse de leurs phylogénies individuelles comparées à une phylogénie de référence inférée à partir de plusieurs marqueurs permet de les écarter.

Les deux arbres montrés dans la Figure 21 A et B ont été racinés arbitrairement entre les Euryarchaeota et les Crenarchaeota, en accord avec les phylogénies universelles publiées précédemment. Seulement trois espèces de Crenarchaeota sont représentées pour dix d'Euryarchaeota. Parmi ces dernières, on peut observer 1) le groupe monophylétique formé par les trois *Pyrococcus*, première divergence au sein des Euryarchaeota ; 2) celui formé par deux *Halobacteriales* (*Haloarcula marismortui* et *Halobacterium* sp.) et *Methanosarcina barkeri*, seule représentante des *Methanosarcinales* ; et 3) celui formé par les *Thermoplasmatales* (*Thermoplasma*

acidophilum et *Ferroplasma acidarmanus*). Ces cinq espèces forment d'ailleurs un groupe monophylétique avec *Archaeoglobus fulgidus*, représentante des *Archaeoglobales*. Par contre, *Methanococcus jannaschii* (une méthanococcale) et *Methanobacterium thermoautotrophicum* (une méthanobacteriale) sont groupe frère dans l'arbre inféré sur les ARNr (Figure 21 A) et paraphylétiques dans l'arbre inféré sur les protéines ribosomiques (Figure 21 B). Ces deux espèces branchent néanmoins toujours juste après les *Thermococcales*.

Evolution de la méthanogenèse et implications pour la phylogénie des archées

Peu de temps après, le génome de *Methanopyrus kandleri* est séquencé (Slesarev et al. 2002). Alors que les premières analyses phylogénétiques de son ARNr SSU en 1991 (Burggraf et al. 1991) le plaçaient à la base de l'arbre des archées, les analyses phylogénétiques présentées dans l'article de Slesarev et collaborateurs, basées sur des concaténations de protéines ribosomiques ou sur le contenu en gènes, placent cet organisme au sein des Euryarchaeota, après la divergence des *Thermococcales* et plus précisément, groupe frère des *Methanococcales* et des *Methanobacteriales*. La Figure 22 illustre ces deux positions, bien qu'elle ne soit pas issue des articles en question. La position de *M. kandleri* au sein des archées est très importante pour comprendre l'évolution de la méthanogenèse. L'ensemble des méthanogènes connus partagent les mêmes enzymes homologues capables de réaliser la méthanogenèse (Eric Bapteste, Brochier, and Boucher 2005). Si *M. kandleri* a divergé avant les Euryarchaeota et les Crenarchaeota, il est possible que l'ancêtre des archées ait été méthanogène, et que cette voie métabolique ait été perdue dans les lignées non méthanogènes par la suite. L'hypothèse alternative est que les gènes impliqués dans la méthanogenèse auraient été transférés entre méthanogènes.

Deux articles, publiés en 2004 et 2005 par Brochier et collaborateurs (Brochier, Forterre, and Gribaldo 2004) et Bapteste et collaborateurs (Eric Bapteste, Brochier, and Boucher 2005) traitent de la position de *M. kandleri* chez les archées et de l'évolution de la méthanogenèse. Le premier (Brochier, Forterre, and Gribaldo 2004) est une analyse de la phylogénie des archées à partir d'une mise à jour des jeux de données de 53 protéines ribosomiques construits précédemment par Matte-Taillez et collaborateurs en 2002 et d'un nouveau jeu de données construit à partir de 15 protéines impliquées dans la transcription (12 sous-unités de l'ARN polymérase ; A', A'', B, D, E', E'', F, H, K, L, N, P ; et les facteurs de transcription NusA, NuG et TFS). La Figure 22 montre les arbres obtenus à partir de ces deux jeux de données.

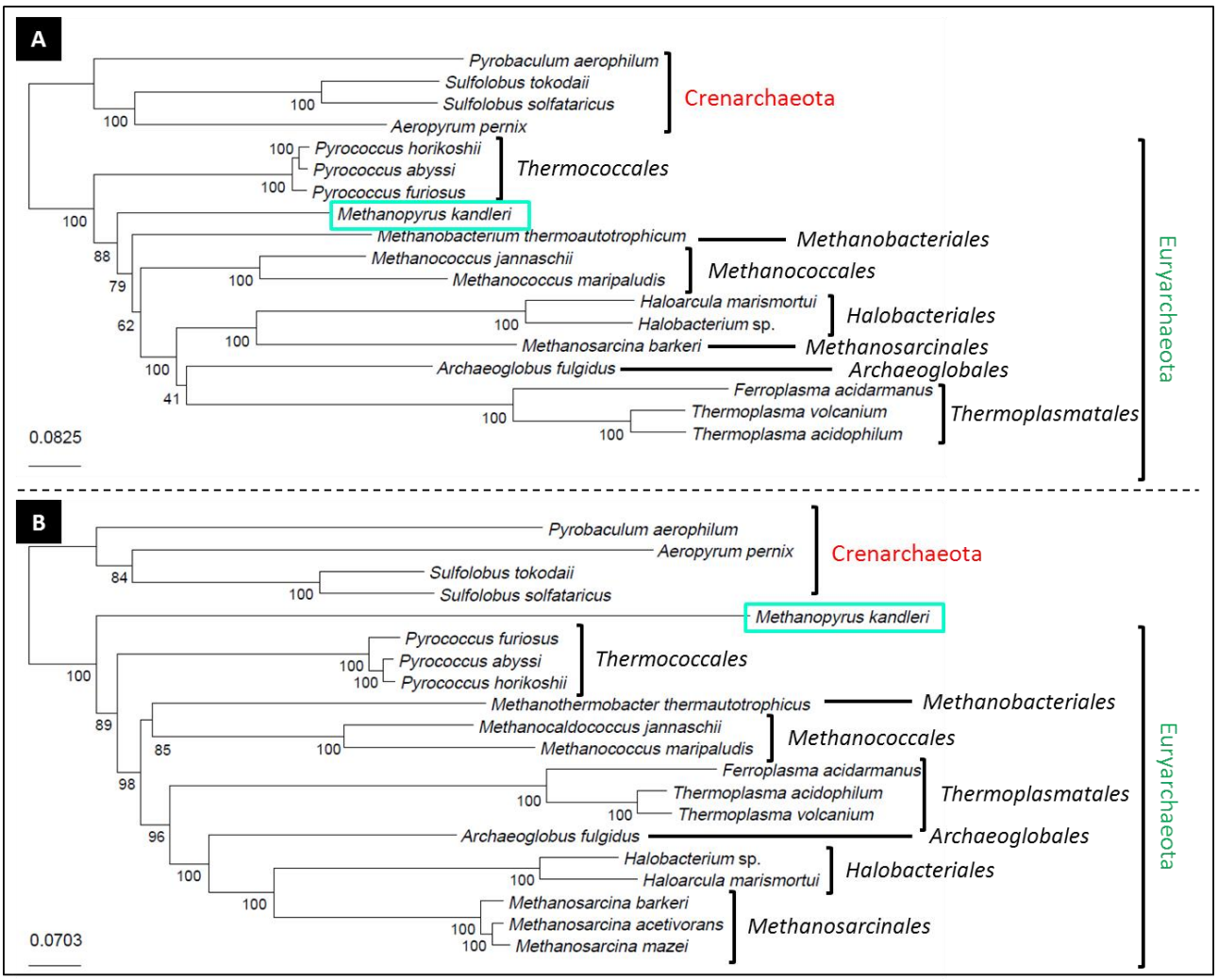


Figure 22. Position phylogénétique de M. kandleri. Phylogénies en maximum de vraisemblance adaptées de Brochier et collaborateurs 2004 (Brochier, Forterre, and Gribaldo 2004), racinées arbitrairement entre Euryarchaeota et Crenarchaeota. M. kandleri est entourée en bleu. A : Phylogénie inférée à partir de la concaténation de 53 protéines ribosomiques. M. kandleri branche après la divergence des Thermococcales au sein des Euryarchaota. B : Phylogénie inférée à partir de la concaténation des sous-unités de l'ARN polymérase et de trois facteurs de transcription. M. kandleri branche entre les Crenarchaeota et les Euryarchaeota.

La position de *M. kandleri* est différente selon le jeu de données utilisé, basale sur une très longue branche avec les protéines impliquées dans la transcription ou au sein des Euryarchaeota avec les protéines ribosomiques. L'analyse des phylogénies individuelles des protéines impliquées dans la transcription montre que certaines sous-unités de l'ARN polymérase, A' et A'' par exemple, ont un signal fort pour placer *M. kandleri* à la base des Euryarchaeota sur une très longue branche, alors que la sous-unité B la place avec un fort support avec les *Methanococcales* et les *Methanobacteriales*, après la divergence des *Thermococcales*. Cette différence de position serait probablement le résultat d'un biais d'attraction de longues branches (LBA) dû à un taux d'évolution rapide des sous-unités de l'ARN polymérase. Par ailleurs, l'analyse d'un caractère rare appuie

l'hypothèse d'une divergence tardive de *M. kandleri* au sein des Euryarchaeota ; la sous-unité B est coupée en deux dans le génome de certaines euryarchées : les méthanogènes, les *Halobacteriales* et *Archaeoglobus fulgidus*. La coupure entre les deux parties du gène codant cette protéine dans ces génomes est exactement au même endroit de la séquence, il est donc plus parcimonieux de supposer qu'une seule coupure s'est produite chez l'ancêtre des ces organismes. Si *M. kandleri* se place à la base des Euryarchaeota, la coupure aurait dû se faire chez l'ancêtre de l'ensemble des Euryarchaeota, et les deux parties du gène auraient dû être rassemblées deux fois, chez les *Thermococcales* et chez les *Thermoplasmatales*. Si *M. kandleri* se place après la divergence des *Thermococcales*, alors une seule fusion est nécessaire pour expliquer cette répartition, rendant cette hypothèse plus parcimonieuse.

Un an plus tard, Bapteste et collaborateurs (Eric Bapteste, Brochier, and Boucher 2005) analysent l'évolution de la méthanogenèse et montrent que les protéines impliquées dans ce processus sont orthologues. La voie principale de méthanogenèse est la voie hydrogénotrophique, permettant la production de méthane à partir de CO_2 ou de formate, dans laquelle au moins 25 protéines sont impliquées. Certains des gènes codant ces protéines sont spécifiques aux méthanogènes alors que d'autres sont présents chez d'autres organismes. Néanmoins la phylogénie des gènes est congruente avec la phylogénie des organismes inférée sur les protéines ribosomiques, et montre deux groupes de méthanogènes au sein des Euryarchaeota. Le premier groupe est composé des *Methanopyrales* (dont *M. kandleri*), des *Methanococcales* et des *Methanobacteriales*. Les auteurs proposent de les rassembler sous le nom de « Méthanogènes Classe I » et ce groupe émergerait au sein des Euryarchaeota après la divergence des *Thermococcales*. Le deuxième groupe, rassemblant les *Methanosarcinales* et les *Methanomicrobiales*, serait les « Méthanogènes Classe II » et émergerait plus tardivement. La monophylie de chacune de ces classes n'est par contre pas clairement définie ou rejetée par ces analyses. Par ailleurs, il est notable que certains gènes impliqués dans la méthanogenèse sont présents chez *Archaeoglobus fulgidus* et chez les *Halobacteriales*, appuyant ainsi fortement l'hypothèse d'une apparition unique de la méthanogenèse au sein des Euryarchaeota avec des pertes différentielles dans les groupes phylogénétiquement placés entre les deux classes de méthanogènes. L'apparition de la méthanogenèse serait donc relativement tardive dans l'histoire évolutive des archées. L'impact de la production biologique de méthane sur l'atmosphère terrestre est très significatif en tant que gaz avec un fort effet de serre (cf. Introduction B.1.a). Son apparition est donc importante en ce qui concerne l'évolution des conditions de vie sur Terre.

La position de Nanoarchaeum equitans : nouveau phylum ou artefact ?

Lors de sa découverte l'archée symbiote *N. equitans* avait été proposée comme étant représentante d'un nouveau phylum d'archées, les Nanoarchaeota (H. Huber et al. 2002). Le séquençage de son génome est publié en 2003 par Waters et collaborateurs (Waters et al. 2003) et l'analyse phylogénétique de la concaténation de 35 protéines ribosomiques en utilisant les eucaryotes comme groupe extérieur montre sa position basale chez les archées. Brochier et collaborateurs réanalysent la position de cette espèce en 2005 en utilisant différentes méthodes (Brochier, Gribaldo, Zivanovic, Confalonieri, & Forterre, 2005). D'une part, différentes concaténations de protéines ribosomiques sont construites : une première contenant toutes les protéines ribosomiques, la même moins 9 protéines dont les phylogénies individuelles ont montré de possibles transferts horizontaux depuis les Crenarchaeota vers *N. equitans* (sa symbiose avec la crenarchée *I. hospitalis* pourrait expliquer facilement ces transferts) et enfin deux concaténations composées respectivement des protéines de la petite et de la grande sous-unités du ribosome. La première concaténation et la concaténation des protéines de la grande sous-unité du ribosome ont la topologie attendue pour les archées et *N. equitans* émerge entre les Euryarchaeota et les Crenarchaeota. La concaténation des protéines de la petite sous-unité du ribosome place par contre *N. equitans* au sein des Euryarchaeota comme groupe frère des *Thermococcales*, le reste de la phylogénie est celle attendue. Le signal contradictoire entre les deux sous-unités du ribosome remet en question la pertinence de l'utilisation conjointe de toutes les protéines ribosomiques dans ce cas, et l'analyse des phylogénies individuelles de ces protéines permet d'en écarter neuf ; la concaténation résultante donne une phylogénie où *N. equitans* est groupe frère des *Thermococcales* au sein des Euryarchaeota avec un support aux branches assez fort (Figure 23).

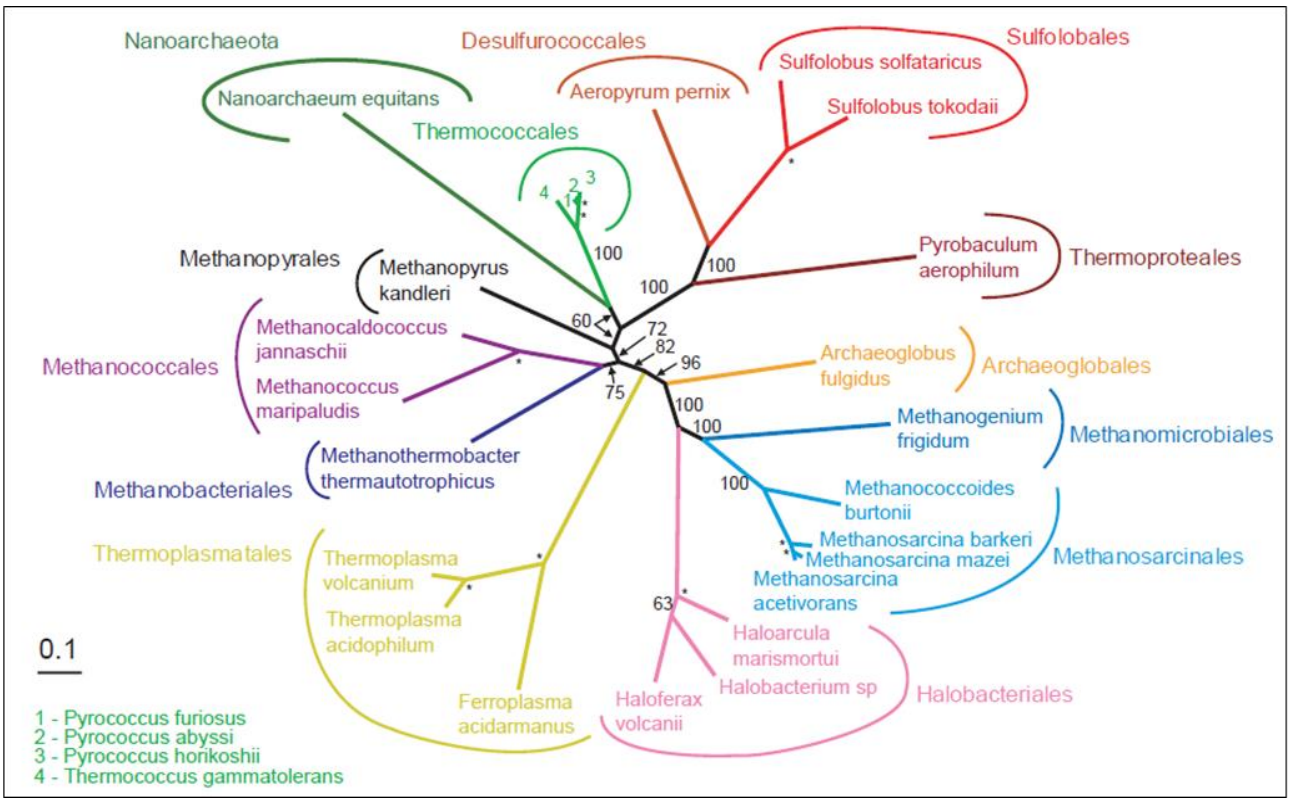


Figure 23. Nanoarchaeum equitans : une Euryarchaeota. Phylogénie de 41 protéines ribosomiques sélectionnées pour l'analyse de la position de N. equitans en maximum de vraisemblance par Brochier et collaborateurs 2005 (Brochier et al. 2005). On peut voir que N. equitans se place en groupe frère des Thermococcales avec un support de 60% et au sein des Euryarchaeota avec un support de 100%.

L'analyse de la phylogénie d'autres protéines, considérées comme des bons marqueurs de l'histoire évolutive des espèces, comme les facteurs d'élongation EF-1α et EF2, ou celle de la sous-unité A de la topoisomérase VI et de la reverse-gyrase placent également *N. equitans* comme groupe frère des *Thermococcales*. De même, le premier résultat obtenu par BLAST à partir de l'ensemble de ses gènes est d'origine thermococcale dans 25 % des cas environ. L'ensemble de ces résultats semble bien confirmer l'appartenance de *N. equitans* aux Euryarchaeota, sa position basale sur une longue branche étant très probablement le résultat d'une LBA avec le groupe extérieur eucaryote provoquée par une évolution très rapide classique des symbiotes (Philippe and Laurent 1998).

La position basale de *N. equitans* chez les archées avait naturellement conduit certains auteurs à proposer que *N. equitans* soit représentante d'une ancienne lignée d'archées possédant ses caractéristiques génomiques particulières (H. Huber et al. 2002). Néanmoins, ces caractéristiques peuvent être interprétées comme le résultat d'une évolution rapide de leur génome par réduction, déjà observées chez des bactéries symbiotiques (Waters et al. 2003; Boucher, Doolittle, and Raynes 2002). Malgré cela, DiGuilio a proposé que ces caractéristiques de *N. equitans* soient en fait des caractères ancestraux au moins aux archées (Di Giulio 2008a; Di Giulio 2008b) et peut être même à l'ensemble du vivant (Di Giulio 2007). Ces caractères que DiGiulio considère comme ancestraux

sont les gènes codant les ARNt coupés en deux et l'absence d'opérons. Ils feraient de *N. equitans* un « génome fossile » (Di Giulio 2006). Les opérons et l'unification des gènes codant les ARNt seraient des caractères complexes apparus dans un second temps. Cette hypothèse ne repose pas sur des analyses phylogénétiques. Les multiples systèmes ayant la même histoire évolutive analysés par Brochier et collaborateurs corroborent fortement l'hypothèse opposée selon laquelle *N. equitans* est une euryarchée probablement groupe frère des *Thermococcales* et que ses caractéristiques génomiques sont le résultat d'une évolution symbiotique. La récente analyse du génome de « Nanoarchaeote Nst1 » (Podar et al. 2013) montre une évolution par réduction de génome du même type que *N. equitans*, mais beaucoup moins avancée, ce qui apporte une preuve de plus de leur évolution atypique.

Nouveaux génomes et nouveaux marqueurs

En 2005, 25 génomes d'archées sont disponibles, 4 Crenarchaeota et 21 Euryarchaeota (incluant *N. equitans*), représentant 13 ordres. Les trois précédentes études portant sur la phylogénie des archées décrites plus haut portaient sur des points précis de cette histoire évolutive (position de *M. kandleri*, évolution de la méthanogenèse et position de *N. equitans*) ; le nombre de génomes d'archées disponibles depuis la dernière étude de leur phylogénie en général (Matte-Tailliez et al. 2002) a presque doublé à ce moment-là, et Brochier et collaborateurs font alors une mise à jour des jeux de données des protéines ribosomiques et des sous-unité de l'ARN polymérase pour tester la position des nouveaux génomes disponibles et vérifier la robustesse des topologies obtenues jusqu'ici (Brochier, Forterre, and Gribaldo 2005). Deux concaténations sont construites, la première nommée « traduction » à partir de 53 protéines ribosomiques, et la seconde « transcription » à partir des 15 protéines impliquées dans la transcription (12 sous-unité de l'ARN polymérase ; A', A'', B, D, E', E'', F, H, K, L, N, P ; et les facteurs de transcription NusA, NuG et TFS). Les deux phylogénies obtenues sont congruentes, excepté en ce qui concerne la position de *M. kandleri*, qui se place à la base des Euryarchaeota dans l'arbre de « transcription » et après les *Thermococcales* dans l'arbre de « traduction », comme attendu. La congruence de ces deux phylogénies tendrait à prouver qu'un ensemble de gènes, le « core génome », évolue principalement par héritage vertical chez les archées et enregistre l'histoire évolutive de ces organismes. Cette histoire serait donc reconstructible malgré les transferts horizontaux de gènes existants chez les archées. Néanmoins, ces deux systèmes, le ribosome et l'ARN polymérase sont impliqués dans des processus informationnels de la cellule (transcription et traduction) et qui, de plus, sont couplés (French et al. 2007). Il n'est pas exclu que l'histoire évolutive qu'ils reflètent soit l'histoire des systèmes informatifs mais que d'autres systèmes, métaboliques ou liés à d'autres processus cellulaires,

n'aient pas eu la même évolution.

L'année suivante, Gribaldo et Brochier-Armanet font une revue des travaux sur la phylogénie des archées (Simonetta Gribaldo and Brochier-Armanet 2006), et actualisent les jeux de données transcription et traduction avec quatre nouveaux génomes (Figure 24) et étudient la phylogénie d'autres systèmes, le complexe SRP (Signal Recognition Particule) et l'exosome.

Figure 24. Phylogénie des archées en 2006. Phylogénies en maximum de vraisemblance non racinées de Gribaldo et Brochier-Armanet en 2006 (Simonetta Gribaldo and Brochier-Armanet 2006). A : Phylogénie inférée à partir de la concaténation de 53 protéines ribosomiques. B : Phylogénie inférée à partir de la concaténation des sous-unités de l'ARN polymérase et de trois facteurs de transcription.

Les deux phylogénies de la Figure 24 sont congruentes sur leurs topologies globales et très proches de celles obtenues lors des études précédentes, mais mieux soutenues. On retrouve les Euryarchaeota et les Crenarchaeota, et *N. equitans* se place entre les deux. Au sein des Crenarchaeota, *Aeropyrum pernix*, représentant des *Desulfurococcales*, se place en groupe frère des *Sulfolobales*, et *Pyrobaculum aerophylum*, représentant les *Thermoproteales* est le premier à diverger. Ces relations de parenté ont déjà été observées dans les études précédentes. En ce qui concerne les Euryarchaeota, les ordres représentés par plusieurs organismes sont monophylétiques :

les *Thermococcales*, les *Methanococcales*, les *Thermoplasmatales*, les *Halobacteriales*, et les *Methanosarcinales*. L'ordre de divergence chez les euryarchées est aussi très ressemblant entre ces deux phylogénies. Les méthanogènes classe II (*Methanomicrobiales* et *Methanosarcinales*) forment un groupe monophylétique et sont groupe frère des *Halobacteriales*. Ce sous ensemble est à son tour groupe frère d'*Archaeoglobus fulgidus*, puis des *Thermoplasmatales*. On trouve par contre des différences en ce qui concerne les relations entre les méthanogènes classe I (*Methanopyrales*, *Methanococcales* et *Methanobacteriales*). Dans la phylogénie « traduction », *M. kandleri* et *Methanothermobacter thermoautotrophicus* (une méthanobacteriale) sont groupes frères et émergent juste après les *Thermococcales*, les *Methanococcales* se plaçant entre le groupe formé par *M. kandleri* et *M. thermoautotrophicus* et le reste des euryarchées. Dans la phylogénie « transcription », *M. kandleri* est groupe frère de *N. equitans*, et *M. thermoautotrophicus* se place entre les *Methanococcales* et le reste des euryarchées. La position de *M. kandleri* à la base des euryarchées est très probablement le résultat d'une LBA comme expliqué plus tôt dans l'arbre « transcription », mais de fait, les relations entre les méthanogènes classe I ne sont pas résolues. Il est aussi possible qu'un manque de signal phylogénétique provoque cette irrésolution comme le suggèrent les faibles soutiens de ces nœuds.

Le système SRP et l'exosome sont deux complexes protéiques impliqués dans d'autres mécanismes cellulaires que les processus informationnels. Ils sont présents chez toutes les archées et composés de plusieurs protéines ; dans chacun de ces deux complexes, les protéines partagent très probablement une même histoire évolutive puisque agissant en complexe dans la cellule. Il était donc intéressant de les comparer avec les systèmes informationnels de transcription et traduction. Le système SRP est composé de deux protéines paralogues, Srp54 et Srp19, d'un récepteur FtsY, et d'un ARN 7S. Srp54 et FtsY sont présents chez toutes les archées à l'exception de *N. equitans*, alors que Srp19 est absent dans plusieurs ordres (*Thermococcales*, *Thermoproteales*, et *N. equitans*) : ce dernier sera donc écarté de l'analyse. Les phylogénies individuelles de Srp54 et de FtsY ont des topologies proches de la phylogénie « informationnelle », avec une dichotomie Crenarchaeota/Euryarchaeota, la monophylie des ordres, et le même ordre de divergence chez les Crenarchaeota. Par contre, les relations entre euryarchées sont très mal résolues. Aucun transfert horizontal n'est clairement détecté, mais le manque de signal dû au peu de positions des alignements individuels est probablement la cause de cette irrésolution. La concaténation des deux protéines montre une phylogénie très proche de la phylogénie « informationelle », avec quelques points de divergences : les *Methanobacteriales* et *Methanococcales* sont groupes frères entre eux, et leur groupe est frère des *Thermococcales*, alors que *M. kandleri* branche à la base des euryarchées.

L'exosome est un complexe de dégradation des ARN dans la cellule. Chez les archées il est

formé de quatre protéines : Rrp41 et Rrp42, qui composent l'anneau central ayant une activité exonucléase, et Rrp4p et Cs14, deux protéines périphériques. Il est absent chez les *Halobacteriales* et chez les *Methanococcales*, et Cs14 est absent chez *N. equitans*. Les phylogénies individuelles ayant permis de détecter un transfert horizontal des crenarchées vers *N. equitans*, la séquence en question a donc été retirée. La phylogénie obtenue à partir de la concaténation des quatre protéines soutient les mêmes relations que pour les protéines du complexe SRP, avec un faible support et *N. equitans*, absente dans la phylogénie « SRP », branche à la base des archées. Cette faible résolution est probablement le résultat d'un manque de signal et des données manquantes pour Cs14 en ce qui concerne la position de *N. equitans*.

Les phylogénies de ces deux systèmes, bien que faiblement résolues et incomplètes pour l'exosome, montrent un signal très proche de celui obtenu à partir des protéines informationnelles. L'ajout de ces nouveaux marqueurs ne permet pas d'apporter plus d'éléments en ce qui concerne les relations problématiques entre certains groupes d'archées, mais, et c'est le plus important, confirme que l'histoire évolutive des archées est inscrite dans un certain nombre de protéines héritées verticalement.

Nouveaux phyla et phylogénie de référence actuelle

En 2008, le phylum des Thaumarchaeota est proposé par Brochier-Armanet et collaborateurs (Brochier-Armanet et al. 2008) et depuis plusieurs génomes de Thaumarchaeota ont été publiés. La même année, le génome de '*Ca.* Korarchaeum cryptophilum' est publié (Elkins et al. 2008). Enfin, en 2011, le génome de '*Ca.* Caldiarchaeum subterraneum' est publié (Nunoura et al. 2011) et le phylum Aigarchaeota est proposé par ses auteurs. Le séquençage de génomes représentant des lignées basales de la phylogénie des archées est donc très important pour comprendre les relations anciennes entre les différents phyla et mieux connaître la nature du dernier ancêtre commun aux archées.

Comme déjà expliqué précédemment, les Thaumarchaeota correspondent au groupe I d'archées mésophiles proches des crenarchées, décrites indépendamment par DeLong et Furhmann et collaborateurs en 1992. C'est un groupe d'organismes nombreux, en termes de biomasse et de diversité. Ils partagent la capacité d'oxydation anaérobie de l'ammonium, et leurs relations phylogénétiques avec les Crenarchaeota ne sont pas clairement résolues. La publication du génome de *Cenarchaeum symbiosum* (Hallam et al. 2006) permet enfin d'intégrer le groupe I à la phylogénie des archées en 2008 (Brochier-Armanet et al. 2008). L'arbre basé sur la concaténation des ARNr place le groupe I comme groupe frère des Crenarchaeota, et la racine de l'arbre est placée entre ce groupe et les Euryarchaeota. La phylogénie inférée à partir des protéines ribosomiques

(Figure 25) utilisant les eucaryotes comme groupe extérieur donne un résultat différent car *Cenarchaeum symbiosum* se place à la base des archées, avant l'émergence des Euryarchaeota et des Crenarchaeota.

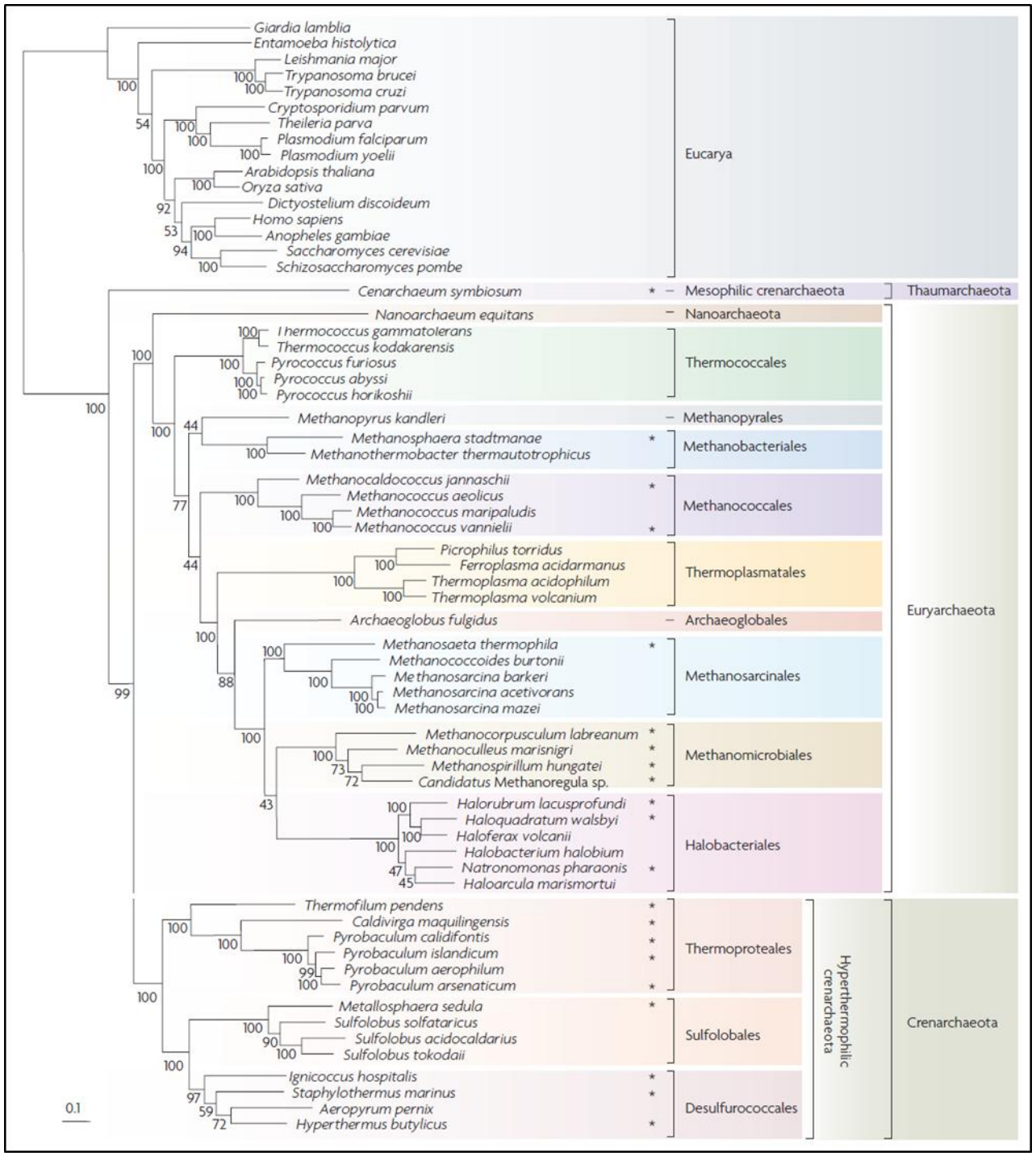


Figure 25. Les Thaumarchaeota : un nouveau phylum? Phylogénie en maximum de vraisemblance des protéines ribosomiques publiée par Brochier-Armanet et collaborateurs en 2008 (Brochier-Armanet et al. 2008). Cenarchaeum symbiosum, représentant les Thaumarchaeota, branche basalement au sein des archées.

Parallèlement à cette analyse phylogénétique, les auteurs ont fait une analyse du contenu en gènes dans le génome de *C. symbiosum*. L'analyse des COG, « Clusters of Orthologous Groups » (R L Tatusov et al. 2001) (qui permettent une classification des gènes orthologues selon leur fonction cellulaire) présents chez *C. symbiosum* montre qu'il partage un nombre de catégories fonctionnelles du même ordre avec les euryarchées (sept) et les crenarchées (un) que ces deux phyla entre eux (25), reflétant une grande différence biologique entre ces trois groupes. D'autres caractéristiques particulières placent ce génome à part des deux phyla d'archées : l'absence ou la présence de protéines ribosomiques particulières, l'absence de l'ADN topoisomérase IA (présente chez les autres archées, les bactéries et les eucaryotes) ou la présence de la ADN topoisomérase IB (absente chez les autres archées mais présente chez les eucaryotes). L'ensemble de ces caractéristiques, qui ne les rapprochent ni des Crenarchaeota ni des Euryarchaeota, et leurs particularités métaboliques sont donc des indices forts pour les considérer comme un groupe bien particulier. Ils ont conduit Brochier-Armanet et collaborateurs à proposer la création du nouveau phylum des Thaumarchaeota. La place de ce phylum au sein des archées n'est par contre pas résolue, puisque selon le groupe extérieur choisi, bactéries ou eucaryotes, la racine se place soit dans la branche menant aux Euryarchaeota, soit dans celle menant aux Thaumarchaeota, respectivement.

Du point de vue de la phylogénie des archées, la Figure 25 montre un arbre raciné avec les séquences eucaryotes. Cet arbre est très intéressant puisque qu'il montre les relations phylogénétiques entre 48 archées très diverses. La monophylie des Crenarchaeota et des Euryarchaeota est retrouvée, mais aussi celle des différents ordres d'archées. En ce qui concerne les points problématiques abordés plus tôt, il est intéressant de constater que *N. equitans* se place à la base des Euryarchaeota avec un support maximal. La phylogénie inférée à partir des ARNr et racinée par les séquences bactériennes place *N. equitans* au même endroit, si ce n'est qu'elle groupe avec *M. kandleri*. Par contre, dans la phylogénie inférée à partir des protéines ribosomiques, *M. kandleri* branche avec les *Methanobacteriales* et les méthanogènes classe I ne sont toujours pas monophylétiques. De même les méthanogènes classe II, qui semblaient jusque là être monophylétiques, ne le sont pas, et les *Methanomicrobiales* sont groupes frères avec les *Halobacteriales*.

J'ai déjà décrit précédemment comment avaient été proposés les phyla Korarchaeota (Barns et al. 1996) et Aigarchaeota (Nunoura et al. 2011). En 2008 et 2011 sont publiées les séquences des génomes des deux premiers représentants de ces phyla, '*Ca.* Korarchaeum cryptophylum' et '*Ca.* Caldiarchaeum subterraneum', respectivement. En 2011, Brochier-Armanet et collaborateurs (Brochier-Armanet, Forterre, & Gribaldo, 2011) intègrent ces nouveaux génomes à la phylogénie

des archées, ainsi que de nouvelles espèces découvertes entre temps, telles que les ARMAN, le nouvel ordre de méthanogènes, les *Methanocellales*, *Acidilobus saccharovorans* (représentant des *Acidilobales*), *Aciduliprofudum boonei* (représentant du groupe d'euryarchées DHVE2) et un fragment du génome de l'organisme « uncultured marine DeepAnt-JyKC7 » (représentant non cultivé du groupe II d'Euryarchaeota) ainsi que les nouveaux génomes séquencés appartenant aux ordres déjà représentés précédemment. Une phylogénie est inférée à partir de 57 protéines ribosomiques présentes dans au moins 89 des 99 génomes d'archées utilisés (Figure 26). Les six phyla d'archées (y compris le phylum Nanoarchaeota, bien qu'il ait été clairement remis en question) y sont représentés et le nombre de génomes utilisés par rapport à la phylogénie précédente a doublé. Néanmoins cette phylogénie n'est pas racinée, et il est donc impossible de définir les relations anciennes entre phyla.

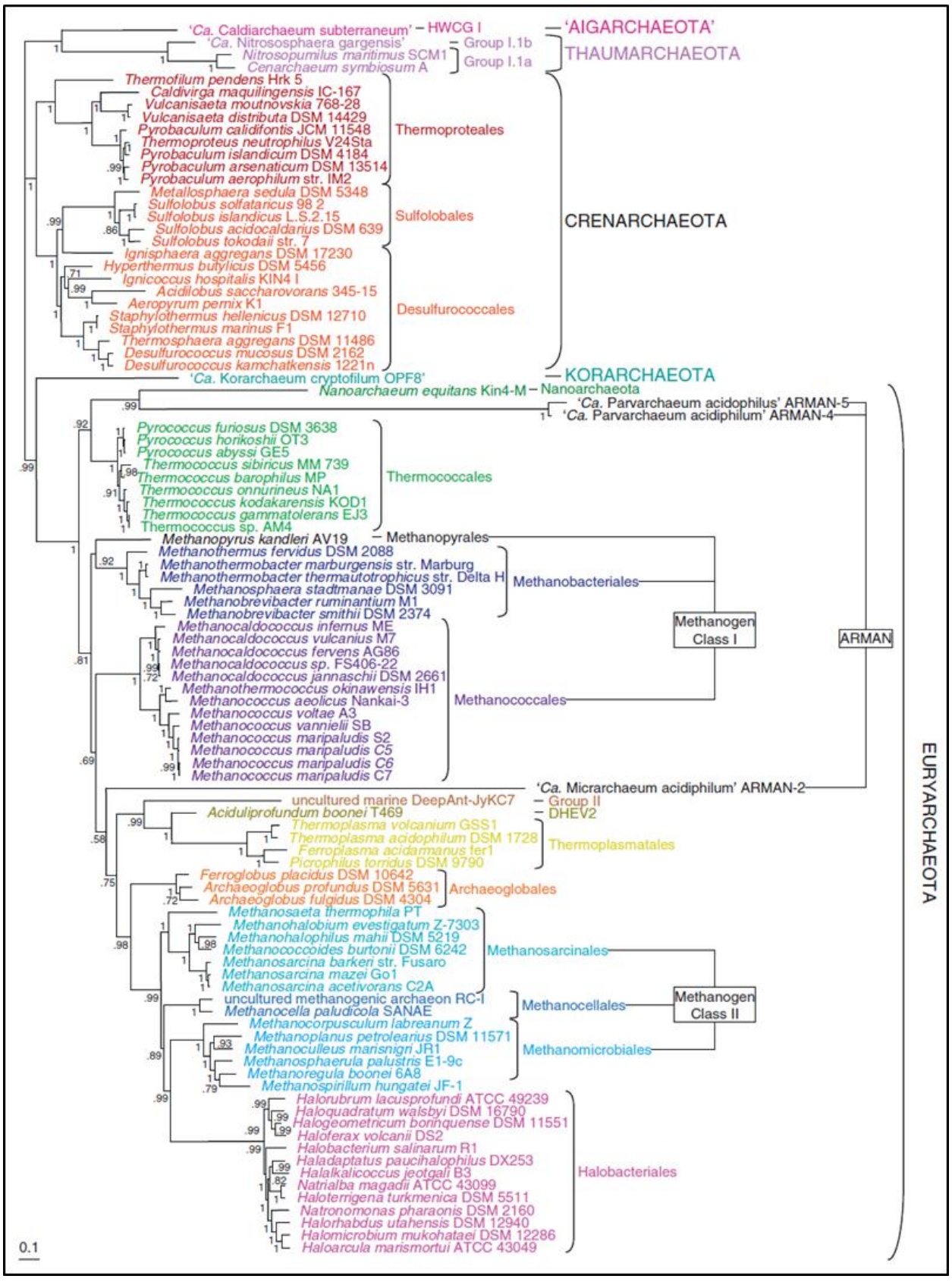


Figure 26. Phylogénie des archées en 2011 (référence actuelle). Phylogénie non racinée en inférence bayésienne publiée par Brochier-Armanet et collaborateurs en 2011 (Brochier-Armanet, Forterre, and Gribaldo 2011), inférée à partir d'un alignement de 57 protéines ribosomiques de 99 espèces (5838 positions).

La nouvelle phylogénie vient confirmer un certains nombre de points déjà observés dans les phylogénies précédentes : la monophylie des Euryarchaeota, des Crenarchaeota et des Thaumarchaeota ; la monophylie des différents ordres d'archées ; la position de *N. equitans* en groupe frère des *Thermococcales* et la position de *M. kandleri* au sein des Euryarchaeota après la divergence des *Thermococcales*. En ce qui concerne les relations entre phyla (excepté les Nanoarchaeota), on peut noter la relation entre les Thaumarchaeota et les Aigarchaeota ('*Ca.* Caldiarchaeum subterraneum'), avec un support maximal. Hormis cette relation privilégiée, les relations entre phyla ne peuvent pas être inférées clairement à partir de cet arbre étant donné qu'il n'est pas raciné. Il est donc impossible de savoir quel est l'ordre de divergence entre phyla, ou si un regroupement existe entre différents phyla.

Cependant, elle permet d'apporter de nouveaux éléments quant à la position de certaines espèces. Par exemple *Acidilobus saccaravorans*, une crenarchée appartenant à l'ordre récemment proposé *Acidilobales* (Prokofeva et al. 2009), se trouve en fait à l'intérieur des *Desulfurococcales* à proximité d'*Aeropyrum pernix*. Le groupe des *Acidilobales* semble donc faire partie intégrante des *Desulfurococcales* et ne pas être un ordre à part entière. L'ordre proposé récemment des *Fervidicoccales* (Perevalova et al. 2010) serait phylogénétiquement proche des *Acidilobales*, il est fort probable que sa situation soit donc la même. La position d'*Ignisphaera aggregans* est, par contre, problématique car elle se place à la base des *Sulfolobales* alors qu'elle est décrite comme appartenant aux *Desulfurococcales*, faisant de cet ordre un groupe paraphylétique contenant les *Sulfolobales*.

Chez les Euryarchaeota, en plus du placement de *N. equitans* avec les *Thermococcales*, cette phylogénie permet de placer le nouveau génome d'*Aciduliprofundum boonei*, du groupe DHVE2, en groupe frère des *Thermoplasmatales*, l'ensemble formant un groupe monophylétique avec le génome « uncultured marine DeepAnt-JyKC7 » (représentant non cultivé du groupe II d'Euryarchaeota). Etant donné la large diversité d'organismes non cultivés connus pour le groupe II et le groupe DHVE2 (Christa Schleper, Jurgens, and Jonuscheit 2005), il est attendu que cette région de l'arbre des Euryarchaeota accueille prochainement des nouveaux représentants. Par contre, la position des génomes des ARMAN, '*Ca.* Micrarchaeum acidiphilum ARMAN-2', '*Ca.* Parvarchaeum acidiphilum ARMAN-4' et '*Ca.* Parvarchaeum acidophilus ARMAN-5', n'est pas résolue. D'une part, les deux '*Ca.* Parvarchaeum' se placent ensemble, ce qui est attendu étant donné la forte similarité de leurs ARNr SSU, mais sur une très longue branche, à coté de *N. equitans*. Comme cela a déjà été abordé, ces organismes de petite taille ont un génome réduit (de l'ordre de 1 mégabase) et il est possible qu'ils aient eu une évolution rapide liée à un mode de vie

symbiotique ou parasitique, entrainant leur position sur une longue branche très instable. La position des '*Ca*. Parvarchaeum' à coté de *N. equitans*, elle-même sur une longue branche, semble donc artéfactuelle. En ce qui concerne la troisième ARMAN, '*Ca.* Micrarchaeum acidiphilum', elle se place après la divergence des *Methanococcales* et juste avant la divergence du groupe composé des *Thermoplasmatales*, du groupe DHVE2 et du groupe II, mais cette position n'est pas soutenue et devra être clarifiée. Enfin, ni les méthanogènes classe I ni les classe II ne sont monophylétiques, comme dans la dernière phylogénie des archées proposée en 2008 (Brochier-Armanet et al., 2008), mais contrairement à d'autres phylogénies plus anciennes (Simonetta Gribaldo and Brochier-Armanet 2006). Par ailleurs, le nouvel ordre d'archées méthanogènes *Methanocellales*, représenté par deux espèces, se place entre les *Methanosarcinales* et les *Methanomicrobiales*, elles mêmes groupe frère des *Halobacteriales*.

Actuellement, cette phylogénie des archées peut être considérée comme phylogénie de référence des espèces, dans la mesure où elle est centrée sur les archées, inférée à partir de plusieurs marqueurs fiables, assez bien soutenue et représentative de la diversité de ce domaine.

Le Tableau 1 récapitule les études de la phylogénie des archées que je viens de citer, avec le nombre de génomes utilisés par phyla d'archées, le nombre de protéines ribosomiques et d'autres protéines utilisées.

Article	Génomes	Protéines ribosomiques	Autres marqueurs protéiques
Matte-Taillez et collaborateurs 2002 (Matte-Tailliez et al. 2002)	3 crenarchées 10 euryarchées	53	ARN polymérase : Sous-unité RpoB
Brochier et collaborateurs 2004 (Brochier et al., 2004)	4 crenarchées 14 euryarchées	53	ARN polymérase : 12 sous-unités Facteurs de transcription : 3
Bapteste et collaborateurs 2005 (Eric Bapteste, Brochier, and Boucher 2005)	4 crenarchées 19 euryarchées (⊂ *N. equitans*)	53	Protéines impliquées dans la méthanogenèse
Brochier et collaborateurs 2005 (Brochier et al. 2005)	4 crenarchées 21 euryarchées (⊂ *N. equitans*)	50	Facteurs d'élongation : EF-1α et EF-2 Topoisomérase VI : Sous-unité A Reverse gyrase
Brochier et collaborateurs 2005 (Brochier, Forterre, and Gribaldo 2005)	4 crenarchées 21 euryarchées (⊂ *N. equitans*)	53	ARN polymérase : 12 sous-unités Facteurs de transcription : 3
Gribaldo & Brochier-Armanet 2006 (Simonetta Gribaldo and Brochier-Armanet 2006)	5 crenarchées 24 euryarchées (⊂ *N. equitans*)	53	ARN polymérase : 12 sous-unités Facteurs de transcription : 3 Système SRP : Srp 54 et FtsY Exosome : 4 sous-unités
Brochier-Armanet et collaborateurs 2008 (Brochier-Armanet et al. 2008)	14 crenarchées 1 thaumarchée 33 euryarchées (⊂ *N. equitans*)	53	
Brochier-Armanet et collaborateurs 2011 (Brochier-Armanet, Forterre, and Gribaldo 2011)	24 crenarchées 3 thaumarchées 1 aigarchée 1 korarchée 70 euryarchées (⊂ *N. equitans*)	57	

Tableau 1. Etudes de la phylogénie des archées : génomes et marqueurs utilisés.

b. La racine de l'arbre des archées et la relation avec les autres domaines

Comme nous venons de le voir, l'étude de la phylogénie des archées est très souvent passée par l'utilisation des protéines informationnelles, particulièrement des protéines ribosomiques et sans groupe extérieur. Les relations anciennes chez les archées, en l'occurrence entre les différentes grandes lignées d'archées et les premières divergences au sein de ces groupes, sont importantes pour comprendre, d'une part, la nature du dernier ancêtre commun des archées (LACA pour « Last Archaeal Common Ancestor ») et, d'autre part, les relations entre ce domaine du vivant et les deux autres domaines, les bactéries et les eucaryotes.

En ce qui concerne les relations entre les archées et les autres domaines du vivant, les archées partagent avec les eucaryotes un certain nombre de caractéristiques les rapprochant,

particulièrement les systèmes « informationnels », à savoir ceux impliqués dans la transmission et l'expression de l'information dans la cellule (réplication, transcription et traduction), qui sont très similaires en termes de composition protéique et de similarité de séquences entre archées et eucaryotes par rapport aux bactéries. Leur relation privilégiée a été établie par le positionnement de la racine de l'arbre du vivant dans la branche menant aux bactéries dès 1989, grâce aux couples de paralogues universels, EF-1α et EF-2, les sous-unités α et β de l'ATPase F_1 (Iwabe et al. 1989; Gogarten et al. 1989), et Srp54/Ffh et SRα/FtsY (S Gribaldo and Cammarano 1998). Même si la relation phylogénétique entre archées et eucaryotes semble acquise, la nature de cette relation n'est pas pour autant résolue. Deux écoles existent : la première propose que les archées et les eucaryotes soient groupes frères, comme l'illustre l'arbre du vivant de Woese en 1990 (Figure 4). Il y aurait ainsi trois domaines primaires : c'est l'hypothèse « 3D ». La seconde propose que les eucaryotes soient le résultat de l'association d'une archée, qui aurait donné la majeure partie des gènes informationnels eucaryotes, les plaçant ainsi phylogénétiquement proches des archées, et d'une bactérie. Il n'y aurait que deux domaines primaires : c'est l'hypothèse « 2D ». De nombreuses études phylogénomiques ont tenté de résoudre les relations entre les trois domaines du vivant comme l'illustre le Tableau 2. La plupart de ces études ont été faites à partir d'un nombre réduit de marqueurs, majoritairement des protéines ribosomiques, et en prenant en compte l'ensemble du vivant, archées, eucaryotes et bactéries. Les résultats varient selon les études, et le débat n'est pas tranché.

Article	Génomes	Marqueurs	Hypothèse soutenue
Harris et collaborateurs 2003 (Harris et al. 2003)	25 bactéries 1 crenarchée 7 euryarchées 3 eucaryotes	50 (⊂ 29 protéines ribosomiques)	3D
Rivera & Lake 2004 (Rivera and Lake 2004)	2 bactéries 1 crenarchée 2 euryarchées 2 eucaryotes	Génome complet d'*Archaeoglobus fulgidus*	2D (Eucaryotes groupe frère des Crenarchaeota)
Ciccarelli et collaborateurs 2006 (Ciccarelli et al. 2006)	150 bactéries 4 crenarchées 14 euryarchées 23 eucaryotes	31 (⊂ 23 protéines ribosomiques)	3D
Yutin et collaborateurs 2008 (Yutin et al. 2008)	Variable selon les marqueurs	136 (⊂ 31 protéines ribosomiques)	3D
Pisani et collaborateurs 2007 (Pisani, Cotton, and McInerney 2007)	97 bactéries 4 crenarchées 17 euryarchées 2 eucaryotes	Données non disponibles	2D (Eucaryotes groupe frère des Thermoplasmatales)
Cox et collaborateurs 2008 (Cox et al. 2008)	10 bactéries 3 crenarchées 11 euryarchées 16 eucaryotes	45 (⊂ 20 protéines ribosomiques, 2 ARN ribosomiques)	2D (Eucaryotes groupe frère des Crenarchaeota)
Foster et collaborateurs 2009 (Foster, Cox, and Embley 2009)	8 bactéries 8 crenarchées 2 thaumarchées 6 euryarchées 11 eucaryotes	41 (⊂ 20 protéines ribosomiques)	2D (Eucaryotes groupe frère du groupe Crenarchaeota+Thaumarchaeota)
Guy & Ettema 2010 (Guy and Ettema 2011)	10 bactéries 15 crenarchées 3 thaumarchées 1 aigarchée 1 korarchée 12 euryarchées (⊂ *N. equitans*) 7 eucaryotes	26 (⊂ 20 protéines ribosomiques)	2D (Eucaryotes groupe frère des « TACK »)
Williams et collaborateurs 2012 (Williams et al. 2012)	8 bactéries 8 crenarchées 3 thaumarchées 1 aigarchée 1 korarchée 6 euryarchées 11 eucaryotes	29 (⊂ 13 protéines ribosomiques)	2D (Eucaryotes groupe frère des « TACK »)
Lasek-Nesselquist & Gogarten 2013 (Lasek-Nesselquist and Gogarten 2013)	73 bactéries 17 crenarchées 3 thaumarchées 1 korarchée 30 euryarchées (⊂ *N. equitans*) 38 eucaryotes (variable selon les analyses)	85 protéines ribosomiques (variable selon les analyses)	Tendance 2D (variable selon les analyses)

Tableau 2. Analyses phylogénomiques sur l'arbre du vivant. Adapté de Gribaldo et collaborateurs 2010 (Simonetta Gribaldo et al. 2010).

Dans ce contexte, les relations anciennes chez les archées représentent un problème essentiel. En effet, il est important de connaître l'ordre de divergence des différents phyla d'archées pour pouvoir replacer les eucaryotes. Si nous sommes dans l'hypothèse « 2D », les eucaryotes brancheront au sein des archées, et il sera possible de savoir quelles sont les lignées d'archées les plus proches des eucaryotes ; dans une hypothèse « 3D », la position de la racine de l'arbre des archées permettra de savoir si les eucaryotes sont bien groupe frère de l'ensemble des archées ou s'ils émergent à l'intérieur de l'arbre des archées. Très peu d'études ont été faites ayant pour but de positionner la racine de l'arbre des archées. En 2008, Brochier-Armanet et collaborateurs (Brochier-Armanet et al. 2008) ont inféré la phylogénie des archées avec un groupe extérieur pour pouvoir comprendre les relations entre les Thaumarchaeota, les Crenarchaeota et les Euryarchaeota, mais un seul génome complet de Thaumarchaeota était disponible et le groupe extérieur choisi pour l'arbre inféré à partir des protéines ribosomiques était les eucaryotes, ce qui ne permet pas de résoudre les relations entre les deux domaines. Un arbre inféré à partir des ARNr avait pour groupe extérieur les bactéries, mais comme je l'ai déjà expliqué plus tôt, ces marqueurs sont insuffisants pour une bonne résolution de l'arbre des archées, et particulièrement des relations anciennes.

A contrario, des analyses à large échelle ont été faites pour essayer de résoudre les relations entre les trois domaines (Tableau 2). Les études les plus récentes prennent en compte les phyla d'archées pour lesquels des génomes ont été publiés durant ces trois dernières années. Par exemple, en 2011, Guy et Ettema (Guy and Ettema 2011) font une étude phylogénomique utilisant 26 protéines universelles (dont 20 protéines ribosomiques et deux sous-unités de l'ARN polymérase) (Tableau 2) sur 48 génomes (31 archées, sept eucaryotes et dix bactéries comme groupe extérieur). Ce jeu de données a été traité avec plusieurs méthodes, et le résultat place toujours les eucaryotes au sein des archées, soit en groupe frère du groupe monophylétique formé par les Thaumarchaeota et les Aigarchaeota, soit entre les Euryarchaeota (y compris *N. equitans*) et les autres phyla (Thaumarchaeota, Aigarchaeota, Crenarchaeota, Korarchaeota). Ils proposent ainsi de définir le groupe « TACK », acronyme des noms de ces quatre phyla, qui serait particulièrement relié aux eucaryotes. Néanmoins, ce travail pose un certain nombre de problèmes : il est basé sur un échantillonnage taxonomique très réduit, particulièrement en ce qui concerne les eucaryotes et les archées, principaux objets de l'étude ; *N. equitans* est intégré au jeu de donnés, alors que cette espèce est connue pour être soumise à l'artefact d'attraction de longues branches en raison de son évolution rapide ; les marqueurs utilisés sont très peu nombreux. Ce dernier point est particulièrement problématique. Les jeux de données utilisés ont été définis à partir des travaux antérieurs de Cox et collaborateurs (Cox et al. 2008) et de Ciccarelli et collaborateurs (Ciccarelli et al. 2006), qui avaient déjà été critiqués pour leur manque de fiabilité (Simonetta Gribaldo et al.

2010). De plus, l'utilisation de ces marqueurs sur l'ensemble du vivant réduit d'autant plus le nombre de positions utilisables pour l'inférence phylogénétique. Enfin, le parti pris des auteurs est clairement affiché dès le résumé de l'article et se place dans l'hypothèse « 2D ».

Dans l'article de Rinke et collaborateurs publié très récemment (Rinke et al. 2013), les auteurs présentent le séquençage en « single-cell » de 201 archées et bactéries non cultivées provenant de divers environnements. En termes d'apport de nouvelles données, cet article est extrêmement novateur car les organismes séquencés ont été choisis justement pour leur appartenance à des lignées peu connues, ce qu'ils appellent la « matière noire microbienne ». Les auteurs ont de plus utilisé ces nouveaux génomes complets pour étudier la phylogénie des archées et des bactéries. Une première phylogénie a été inférée à partir de l'ARNr SSU de l'ensemble des organismes séquencés dans ce travail et d'organismes de référence déjà disponibles. Elle permet aux auteurs de raciner des archées avec les bactéries et, réciproquement, les bactéries avec les archées. La racine de l'arbre des archées se place entre un groupe composé des phyla Thaumarchaeota, Aigarchaeota, Crenarchaeota et Korarchaeota (« TACK ») augmenté d'un certain nombre de lignées nouvelles, et un groupe composé des Euryarchaeota, comprenant aussi un certain nombre de nouvelles lignées. Ensuite, une phylogénie a été inférée pour chaque domaine à partir d'une concaténation de 38 protéines utilisées régulièrement pour ce type d'analyse (dont 27 protéines ribosomiques). En ce qui concerne les archées, 220 génomes ont été utilisés. Les auteurs retrouvent la monophylie du groupe « TACK » proposé par Guy et Ettema, qui se place en groupe frère des Euryarchaeota. Un autre groupe composé de plusieurs lignées anciennes et nouvelles semble avoir divergé avant les Euryarchaeota et les « TACK ». Parmi celle-ci se trouvent les Nanoarchaeota, les *Nanohaloarchaea*, les ARMAN, et deux nouvelles lignées, nommées 'Aenigmarchaeota' et 'Diapherotrites'. Les auteurs proposent d'élever ces quatre dernières lignées au rang de phyla, les *Nanohaloarchaea* prendraient alors le nom de 'Nanohaloarchaeota' et les ARMAN deviendraient 'Parvarchaeota'. Cet ensemble de 5 phyla est nommé d'après leur acronyme, « DPANN ». D'un point de vue phylogénétique, cette étude présente plusieurs lacunes. Tout d'abord, l'arbre des archées présenté n'est pas raciné, il est donc difficile de conclure quant à la monophylie des deux groupes proposés « TACK » + Euryarchaeota et « DPANN ». Ensuite, de nombreuses études précédentes ont montré que les trois lignées de petites archées Nanoarchaeota, *Nanohaloarchaea* et ARMAN avaient une évolution rapide et de type réduction de génome et, par conséquent, étaient en cela particulièrement sujettes à l'attraction de longues branches (Brochier et al. 2005; Baker et al. 2010; Narasingarao et al. 2012). Elles doivent donc être traitées avec beaucoup de précautions, particulièrement dans des études à large échelle où les artefacts peuvent être nombreux. Enfin, l'élévation au rang de phylum de plusieurs lignées doit aussi être soumise à la

plus grande rigueur étant donné les deux précédents arguments (Simonetta Gribaldo and Brochier-Armanet 2012).

3. Conclusion

Les relations anciennes entre les différentes lignées d'archées ne sont donc toujours pas complètement résolues. L'utilisation, à peu de chose près, du même set de protéines comme marqueurs de la phylogénie quelle que soit la question posée (résolution de la phylogénie des archées, position de la racine, relations avec les autres domaines…) et l'absence d'analyses clairement orientées dans ce but sont peut être des débuts de réponses quant à la non résolution de ces points essentiels dans la phylogénie et la connaissance des archées, mais aussi de l'ensemble du vivant. C'est dans ce contexte que se place mon travail de thèse, par la recherche de nouveaux marqueurs de la phylogénie des archées et de la racine de l'arbre des archées par une approche phylogénomique.

Objectifs

Comme je l'ai montré dans l'Introduction, la compréhension des relations de parenté au sein des archées et entre les archées et les deux autres domaines du vivant, bactéries et eucaryotes, est essentielle pour comprendre l'histoire évolutive de l'ensemble du vivant. L'augmentation du nombre de génomes d'archées séquencés et, surtout, la diversité qu'ils recouvrent, ont ouvert la porte à des études massives sur ces questions évolutives.

Mon travail de thèse s'inscrit dans ce contexte, avec pour objectif une amélioration de la connaissance de la phylogénie des archées, et particulièrement en ce qui concerne les relations anciennes dans ce domaine. Ce travail s'est articulé autour de deux axes:

Objectif 1 : La recherche de nouveaux marqueurs pour l'inférence de la phylogénie des archées.

Peut-on trouver de nouvelles protéines pour résoudre la phylogénie des archées ? L'histoire évolutive des protéines informationnelles, utilisées jusqu'à maintenant comme marqueurs de la phylogénie des archées, est-elle représentative de l'histoire évolutive des organismes ou donne-t-elle un signal particulier ? Peut-on améliorer la résolution de la phylogénie des archées ? De nouveaux marqueurs peuvent-ils apporter des éléments de réponse concernant la position phylogénétique de certaines lignées, non résolue ou en débat, comme, par exemple, les relations au sein des méthanogènes classe I ; entre les méthanogènes classe II et les *Halobacteriales* ; la position des ordres de crenarchées proposés récemment, *Acidilobales* et *Fervidicoccales* ; la position des lignées de d'archées de taille nanométrique, *Nanoarchaeota*, *Nanohaloarchaea* et le groupe ARMAN ; ou les relations entre les différents phyla d'archées ; etc. ?

Pour cela, nous avons cherché de nouveaux marqueurs pour la phylogénie des archées par des approches de phylogénomique, puis inféré une phylogénie d'une part à partir des marqueurs sélectionnés et d'autre part en les associant avec les marqueurs de type informationnel utilisés dans les études précédentes (protéines ribosomiques, sous-unités de l'ARN polymérase et facteurs de transcription s). Cette étude fait l'objet du Chapitre 1.

Objectif 2 : La recherche de la racine de l'arbre des archées grâce à des homologues bactériens.

Parmi les marqueurs utilisés pour inférer la phylogénie des archées, certains d'entre eux ont-ils des homologues bactériens pouvant être utilisés pour raciner cette phylogénie (i.e. des homologues bactériens issus de l'ancêtre commun de tous les organismes et non pas de transferts horizontaux depuis des archées) ? Peut-on placer la racine (i.e. la première divergence au sein des archées) grâce à ces homologues bactériens avec un support suffisant ? Si oui, où se place cette

racine? Quelles relations entre les phyla d'archées observe t-on ? Quelles sont les implications sur la taxonomie des archées ?

Pour répondre à ces questions, nous avons recherché parmi les marqueurs utilisés pour inférer la phylogénie des archées (nouveaux et anciens) ceux qui avaient des homologues bactériens suffisamment conservés. Après une analyse fine des phylogénies individuelles contenant les homologues bactériens, nous avons sélectionné un sous-ensemble de marqueurs dédiés à cette étude, et inféré les phylogénies afin de raciner l'arbre des archées. Cette étude fait l'objet du Chapitre 2.

L'analyse des génomes d'archées nécessaires pour attendre l'objectif 1 a produit une énorme masse d'information sur des protéines d'archées (leur répartition taxonomique, présence de transferts horizontaux, etc.). Cette information peut être exploitée pour comprendre l'évolution de ce domaine. Au cours de ma thèse, j'ai analysé l'histoire évolutive d'une de ces protéines, la protéine DnaJ-Fer, présente uniquement chez les thaumarchées et les plantes, et du système chaperonne DnaK-GrpE-DnaJ avec lequel la protéine DnaJ-Fer interagit. Ce travail est présenté dans le Chapitre 3.

Matériels et Méthodes

Les trois Chapitres contenus dans cette thèse sont le résultat d'une analyse globale des génomes de *Cenarchaeum symbiosum* et *Nitrosopumilus maritimus*, deux Thaumarchaeota, et de '*Candidatus* Caldiarchaeum subterraneum', seul représentant du phylum Aigarchaeota. Le but premier de ces analyses était de mettre en évidence des protéines portant un signal pouvant servir à la résolution de la phylogénie des Archaea, point qui sera traité dans le Chapitre 1. Le choix de deux génomes de thaumarchées et de l'aigarchée nous a semblé le plus intéressant pour rechercher de nouveaux marqueurs pour la phylogénie des archées. D'une part, ils partagent des gènes avec les crenarchées et/ou les euryarchées (Brochier-Armanet et al. 2008; Nunoura et al. 2011), ce qui nous permettra de trouver de nouvelles protéines présentes dans les différents phyla. D'autre part, la position phylogénétique de ces phyla récemment proposés n'est pas résolue, et l'utilisation de marqueurs présents dans leurs génomes pourrait nous aider à répondre à cette question. Le Chapitre 2 traite de la position de la racine de l'arbre des Archaea, donc des relations entre phyla. Le Chapitre 3 présente une histoire évolutive originale, celle de la protéine DnaJ-Fer et du système chaperonne DnaK-GrpE-DnaJ, mettant en jeux plusieurs transferts horizontaux entre les trois domaines du vivant. L'analyse des génomes de *N. maritimus*, *C. symbiosum* et '*Ca.* Caldiarchaeum subterraneum' a permis de souligner de nombreux autres cas de protéines avec des histoires évolutives potentiellement intéressantes, avec transferts horizontaux, duplications et/ou pertes de gènes, qui n'ont pu être traités en détails dans cette thèse.

J'ai d'abord étudié l'histoire évolutive de chaque protéine codée dans ces trois génomes par des analyses phylogénétiques préliminaires. J'ai ainsi pu sélectionner les protéines intéressantes pour le Chapitre 1, puis pour le Chapitre 2. Les méthodes utilisées pour l'analyse initiale et une partie des méthodes utilisées pour les Chapitres 1 et 2 sont donc communes ; je les expose ici.

A. Analyses phylogénétiques des protéines codées dans les génomes de *C. symbiosum*, *N. maritimus* et '*Ca*. Caldiarchaeum subterraneum'.

La première étape d'analyse a été la construction de la phylogénie de chacune des protéines codées dans les trois génomes de référence. Cette étape a été faite en collaboration avec Philippe Deschamps, qui avait mis en place précédemment une procédure qui génère des phylogénies de façon automatique pour les protéines choisies en utilisant les séquences similaires extraites d'une banque de données locale.

1. Construction d'une banque de données locale

La banque de données locale est composée des protéines codées par 92 génomes complets d'Archaea (correspondant à un représentant par espèce) disponibles en avril 2011; des protéines codées par les génomes complets de 297 bactéries et de 122 eucaryotes choisis parmi les différents phyla afin de représenter, dans la mesure du possible, la diversité actuellement connue et séquencée de ces deux domaines (Annexe 1, Supplementary Table S1). Ces séquences protéiques d'archées et de bactéries ainsi que la plupart de celles des eucaryotes ont été obtenues directement à partir de la banque de données GenBank (Benson et al. 2013), accessible sur le site du NCBI (http://www.ncbi.nlm.nih.gov/). Un certain nombre de génomes complets d'eucaryotes n'étant pas disponibles dans GenBank, ils ont pu être obtenus à partir d'autres banques, telles que le JGI (http://genome.jgi-psf.org/), et des banques spécialisées pour le génome de différents organismes (*Cyanidioschyzon merolae* : http://merolae.biol.s.u-tokyo.ac.jp/ ou *Galdieria sulphuraria* : http://genomics.msu.edu/galdieria/). La liste de ces génomes est donnée dans l'Annexe 1, Supplementary Table S1.

2. Génération des phylogénies préliminaires

Le nombre de protéines étudiées et conservées à chaque étape est indiqué dans le Tableau 3. Les génomes de *N. maritimus* et de *C. symbiosum* étant assez proches en termes de similarité de séquences (e.g. 94% d'identité entre leurs ARN de la petite sous unité du ribosome (SSU rRNA)) et de contenu en gènes, nous avons décidé de ne conserver que la protéine issue du génome de *N. maritimus* dans le cas où une protéine homologue proche existait chez *C. symbiosum*. Pour cela, nous avons fait une recherche de similarité par BLASTp (Altschul et al. 1990) entre chaque paire de protéines homologues dans ces deux génomes, et si deux protéines présentaient un pourcentage d'identité supérieur ou égal à 60%, nous n'avons conservé que la protéine de *N. maritimus* (Tableau

3). Aux protéines ainsi triées, nous avons rajouté celles de '*Ca.* Caldiarchaeum subterraneum'.

A partir de chaque protéine des trois génomes étudiés, nous avons réalisé une recherche de similarité sur la banque de données locale grâce au logiciel BLASTp (Altschul et al. 1990). Nous avons gardé uniquement les 200 premiers résultats trouvés, ayant une E-value inferieure à 10^{-5} ; au dessus de cette valeur, nous considérons que, même si les séquences peuvent être homologues, elles sont trop distantes pour être utilisées dans notre étude. En effet, nous cherchons seulement les orthologues les plus proches de nos séquences de référence et si ceux-ci sont trop divergents, le signal évolutif ne sera pas assez fiable pour les utiliser comme marqueurs de la phylogénie des archaea. Suite à cette sélection, nous avons éliminé tous les jeux de données contenant moins de 4 séquences, puisqu'aucune information phylogénétique utile ne pourrait en être extraite (Tableau 3).

L'ensemble de ces jeux de données ont été alignés par le logiciel MAFFT (Katoh et al. 2005), avec les options par défaut. De précédents tests de différents logiciels d'alignements (ClustalW (Larkin et al. 2007), Muscle (Edgar 2004), Probcons (Do et al. 2005), T-Coffee (Notredame, Higgins, and Heringa 2000), nous ont permis de comparer le rapport entre la qualité des résultats et le temps de traitement des données, et MAFFT nous semble être actuellement le programme le plus efficace dans la plupart des cas d'après ces deux critères. Certains cas complexes sont parfois mieux traités par d'autres logiciels (cf. Chapitre 3) mais dans une analyse massive de données, il était important d'utiliser la méthode qui donnait les meilleurs résultats de façon générale.

A partir de ces alignements, les positions dont l'homologie est relativement certaine ont été sélectionnées automatiquement grâce au logiciel BMGE (Criscuolo and Gribaldo 2010), avec les options par défaut. Ici encore, des tests effectués avec d'autres logiciels nous ont fait favoriser BMGE car il semblait être le plus efficace en termes de rapport qualité des résultats/temps de traitement.

Enfin, des phylogénies ont été inférées à partir de ces alignements grâce au logiciel FastTree version 2.1.4 (Price, Dehal, and Arkin 2010) implémentant une méthode de maximum de vraisemblance approximée, avec les options par défaut.

Génome Nombres de jeux de données	*N. maritimus*	*C. symbiosum*	'*Ca.* Caldiarchaeum subterraneum'
Total au départ (nombre de protéines codées dans le génome)	1796	2017	1704
Après BLAST réciproque entre *N. maritimus* et *C. symbiosum*	1796	674	1704
Après élimination des jeux de données contenant moins de 4 séquences	1381	626	1205
Total analysés phylogénétiquement	3212		

Tableau 3 : Nombre de protéines conservées à chaque étape de l'analyse des génomes complets de *N. maritimus*, *C. symbiosum* et '*Ca.* Caldiarchaeum subterraneum'.

3. Tri et sélection des phylogénies préliminaires

L'analyse des 3212 phylogénies inférées à été faite de façon manuelle avec l'aide du logiciel TreeReader développé dans notre équipe par Philippe Deschamps. Ce logiciel permet d'afficher des arbres phylogénétiques sur lesquels les noms d'espèces présents aux nœuds terminaux sont colorés automatiquement en fonction de leur taxonomie (d'après la classification utilisée par GenBank). L'autre avantage de cet outil est qu'il permet de charger un seul fichier contenant la totalité des arbres à étudier, au format Newick, et de passer d'un arbre à un autre d'un simple clic. Il est possible de raciner l'arbre observé sur le nœud choisi ou de faire permuter des branches. Cette facilité de lecture des arbres (à la fois en termes de manipulation et d'observation par la coloration automatique des noms d'espèces en fonction de leur taxonomie) a rendu plus efficace l'analyse de l'ensemble des phylogénies inférées par l'observation de chaque phylogénie.

Ainsi pour chaque arbre, j'ai noté dans un tableau (Annexe 2) une série d'observations. Tout d'abord, j'ai indiqué l'absence ou la présence de différents groupes taxonomiques, à commencer par les trois domaines Archaea, Bacteria et Eucarya, mais aussi à un niveau taxonomique plus détaillé, comme les différents phyla d'archées, ou les grands embranchements d'eucaryotes. Ensuite, j'ai inféré l'intérêt potentiel de la protéine comme marqueur de la phylogénie des Archaea. Pour cela, il fallait que la protéine soit présente dans un minimum de génomes d'Archaea (~10-15), que les séquences d'Archaea forment un groupe globalement monophylétique (afin de limiter les risques de transferts horizontaux et de paralogie ancienne) et que les groupes (particulièrement les ordres) dont la monophylie est décrite chez les Archaea le soient. J'ai aussi noté si de potentiels transferts

horizontaux étaient observables dans l'histoire évolutive de chaque protéine. Enfin, j'ai relevé un certain nombre d'observations, qui permettront une ré-analyse ciblée de certaines protéines pour aborder différentes problématiques. Les trois premiers points ont été notés de la façon la plus standardisée possible. Par exemple, l'absence d'un groupe est noté par un 0, sa présence par un 1 ; une protéine potentiellement marqueur de la phylogénie des Archaea est annotée « M ». Ce type d'annotation est précieux pour pouvoir faire différents tris automatiques sur ces données, et sortir la liste de protéines correspondante à une situation d'intérêt, même si elle n'avait pas été prévue au départ de l'analyse. Les analyses du Chapitre 2 en sont un parfait exemple, puisque ce projet n'avait pas été prévu lors du lancement de ce travail, mais a été possible grâce à un tri des familles protéiques sur le critère de la présence de bactéries dans les jeux de données.

B. Inférence de la phylogénie des Archaea : méthodes du Chapitre 1

1. Analyse des protéines d'intérêt pour l'étude de la phylogénie des Archaea

a. Sélection des protéines d'intérêt

A partir des informations collectées dans le tableau mentionné ci-dessus (Annexe 2) j'ai extrait une liste de protéines pouvant potentiellement être utilisées pour l'analyse de la phylogénie des Archaea. Comme précisé précédemment, il fallait que la protéine soit présente dans un minimum de génomes d'Archaea (~10-15), que les séquences d'archées soient globalement monophylétiques et que les groupes d'archées dont la monophylie est admise le soient. 209 protéines ont été sélectionnées à partir des génomes de *N. maritimus* et *C. symbiosum* (193 et 16 respectivement) et 55 à partir du génome de '*Ca.* Caldiarchaeum subterraneum' (Tableau 4) Ces chiffres excluent les protéines ribosomiques, les sous-unités de l'ARN polymérase et les facteurs de transcriptions y étant associés, bien connus et déjà utilisés dans de nombreuses études antérieures. Ces derniers ont fait l'objet d'une mise à jour au cours de mon travail à partir des données de Brochier-Armanet et collaborateurs en 2011 (Brochier-Armanet, Forterre, and Gribaldo 2011) et de Gribaldo et Brochier-Armanet en 2006 (Simonetta Gribaldo and Brochier-Armanet 2006), respectivement.

b. Construction d'une banque de données locale de génomes complets d'Archaea

L'objectif de cette analyse était de reconstruire la phylogénie globale des Archaea. De fait, nous avons décidé de ne travailler qu'à partir de séquences d'Archaea (un transfert horizontal vers un autre domaine ne pouvant modifier l'histoire évolutive au sein des Archaea) mais avec un échantillonnage le plus riche possible pour ce domaine. Pour cela, nous avons construit une banque de données locale nommée 'AllArchaea', contenant toutes les protéines prédites de 129 génomes complets d'Archaea (avec un représentant par espèce) disponibles au début de cette analyse, en avril 2012 (Annexe 1, Supplementary Table S3). Les protéines de ces 129 génomes ont été collectées sur GenBank (Benson et al. 2012).

c. Construction des jeux de données

Pour chaque protéine, un nouveau jeu de données a été construit contenant les homologues identifiés chez les différentes espèces d'archées. Afin de couvrir au mieux la diversité des archées

lors de la recherche d'homologues par similarité avec BLASTp, je suis partie systématiquement de plusieurs séquences graines. Pour cela, j'ai sélectionné dans les phylogénies de départ (contenant des homologues bactériens et eucaryotes) une séquence de thaumarchée, d'euryarchée et de crenarchée dans le groupe monophylétique d'archées m'intéressant. La plupart du temps, la séquence de départ de *N. maritimus*, *C. symbiosum* ou '*Ca.* Caldiarchaeum subterraneum' a été sélectionnée, mais dans certains cas rares, j'ai préféré me concentrer sur un groupe de séquences paralogues à ma séquence d'origine. Les critères pour le choix d'une séquence graine étaient : une position phylogénétique fiable (pas de transfert horizontal de gènes (THG) ou de cas de paralogie flagrant, et pas de séquence isolée loin des autres archées) et une longueur de branche pas trop importante par rapport aux autres branches de l'arbre. Dans la mesure du possible, j'ai sélectionnée les séquences de *Methanobrevibacter ruminantum* et *Sulfolobus tokodaii* comme graine euryarchée et crenarchée respectivement. Si l'une de ces espèces était absente, j'ai sélectionné la séquence d'une espèce proche.

A partir de chacune de ces séquences graines, j'ai fait une recherche de similarité en local grâce au logiciel BLASTp (Altschul et al. 1997) sur la banque de données 'AllArchaea', avec les options par défaut, sauf pour le nombre de résultats et le nombre d'alignements visibles qui ont été réglés à 500. Nous avons considéré qu'au-delà de 500 résultats, soit les séquences n'étaient pas des homologues à notre protéine de départ, soit elles étaient trop éloignées pour la problématique de notre analyse, qui nécessite l'utilisation d'orthologues et doit exclure absolument les paralogues. Les résultats de BLAST ont été analysés manuellement afin de s'assurer de l'homologie des séquences retenues, et si possible de l'orthologie, pour la suite de l'analyse. Les séquences sélectionnées ont été extraites de la banque de données locale 'AllArchaea' grâce à un script de traitement de données codé par Philippe Deschamps, permettant de retrouver à partir d'un fichier de sortie BLAST au format xml les premiers résultats de BLAST jusqu'à une limite choisie chiffrée, et d'aller chercher les séquences correspondantes dans la banque de données utilisée pour la recherche de similarité. Un fichier contenant ces séquences au format fasta est alors créé.

Pour chaque protéine d'intérêt, j'ai donc, à cette étape, trois fichiers contenant les séquences issues des recherches de similarités depuis les séquences graines de Thaumarchaeota, d'Euryarchaeota et de Crenarchaeota. Ces trois fichiers ont alors été fusionnés en un seul, qui a été trié automatiquement pour éliminer les doublons et ne conserver qu'une seule copie de chaque séquence. A ce moment, nous avons un fichier fasta pour chaque protéine sélectionnée au départ comme marqueur potentiel.

d. Analyse phylogénétique des jeux de données

Alignement

Chaque jeu de données a été aligné grâce au logiciel MAFFT (Katoh et al. 2005) avec les options par défaut. Les fichiers d'alignements ont ensuite été convertis dans un format auquel est associé un fichier d'information contenant la fiche d'information GenBank de chaque séquence du fichier (fichier.ali et fichier.inf respectivement), ces deux fichiers étant lisibles par le logiciel MUST (Philippe 1993).

Phylogénies préliminaires

Chaque alignement a été nettoyé avec le logiciel BMGE (options par défaut) (Criscuolo and Gribaldo 2010) pour la suppression des positions non homologues, et une phylogénie a été inférée par maximum de vraisemblance grâce au logiciel PhyML version 3.0 (Guindon et al. 2010), avec la matrice de substitution LG (Le and Gascuel 2008) et une loi gamma à quatre catégories pour tenir compte de l'hétérogénéité de la vitesse d'évolution des sites (LG+G4). A cette étape, nous avons choisi d'utiliser le logiciel PhyML plutôt que le logiciel FastTree (approximation de maximum de vraisemblance) car il s'agit d'une méthode plus robuste, même si le temps de calcul est beaucoup plus long (il n'était pas envisageable de l'utiliser pour traiter plus de 3000 jeux de données de l'analyse préliminaire).

A partir des phylogénies produites par PhyML il a été possible de supprimer les redondances entre les jeux de données. En effet, il est possible par exemple que deux paralogues présents chez les Thaumarchaeota aient été traités séparément, alors que chez le reste des archées il n'existe qu'une seule copie de cette protéine. Dans ce cas, nous avons décidé de ne garder qu'une seule copie du jeu de données et qu'une seule copie de la protéine chez les Thaumarchaeota. Après cette étape, restaient donc 181 jeux de données pour les génomes de *N. maritimus* et *C. symbiosum* (173 et 8 respectivement) et 55 pour le génome de '*Ca.* Caldiarchaeum subterraneum'.

Répartition taxonomique des séquences

Enfin, à partir de chaque fichier fasta, j'ai reporté automatiquement le nombre de copies de la protéine présent dans chaque espèce d'Archaea afin d'avoir un indicateur du nombre de paralogues ou de copies en double dans le jeu de données. Ce travail a été fait grâce au script « TaxRep.pl », que j'ai codé pour compter le nombre de séquences d'une espèce (à partir de son numéro TaxID) dans un fichier fasta. Le résultat obtenu est un tableau dans lequel chaque ligne correspond à une espèce et chaque colonne à un fichier fasta. Le croisement entre une ligne et une

colonne donne le nombre de séquences de cette espèce dans le fichier (Annexe 3).

Analyse de l'alignement

Chaque jeu de données aligné a été analysé grâce au logiciel MUST (Philippe 1993) qui permet de faire des aller-retour entre l'alignement et des phylogénies inférées par Neighbour Joining (Saitou and Nei 1987). Cette étape a permis de supprimer les séquences très divergentes qui auraient pu rester, de selectioner un groupe de paralogue dans les cas de duplication ancienne et de choisir entre deux copies paralogues pour une espèce ou pour un taxon si la duplication s'est faite après le dernier ancêtre commun des Archaea. Elle permet aussi de mettre en évidence d'éventuels transferts horizontaux de gènes entre Archaea, et donc d'écarter les séquences xénologues de l'analyse finale (dans ces cas particuliers, j'ai gardé une trace de l'alignement contenant les séquences issues de transferts horizontaux). Enfin, j'ai écarté un certain nombre de jeux de données dont la qualité n'était pas suffisante pour être utilisés pour l'analyse de la phylogénie des Archaea (irrésolution trop importante, doute sur des cas de THG ou de paralogie…) J'ai fait ce travail de nettoyage de l'alignement en tenant compte du tableau d'information de départ (Annexe 2), des phylogénies inférées par FastTree contenant les homologues bactériens et eucaryotes, des phylogénies PhyML ne contenant que les séquences d'archées, et du tableau de répartition taxonomique (Annexe 3). A la fin du nettoyage de l'alignement, il ne devait rester qu'une seule séquence par génome d'archée, donc au plus 129 séquences par jeu de données. Le nombre de jeux de données conservés est indiqué dans le Tableau 4. La liste des protéines retenues est indiquée dans l'Annexe 1, Supplementary Table S2.

Génome Nombres de jeux de données	*N. maritimus*	*C. symbiosum*	'*Ca.* Caldiarchaeum subterraneum'
Etudiés comme marqueurs potentiels	193	16	55
Après suppression des paralogues chez les Thaumarchaeota	173	8	55
Conservés comme marqueurs	153	4	43
Total des jeux de donnés retenus	200		

Tableau 4 : Nombre de protéines conservées à chaque étape de l'analyse des marqueurs potentiels depuis les génomes complets de *N. maritimus*, *C. symbiosum* et '*Ca.* Caldiarchaeum subterraneum'.

Les 200 jeux de données retenus ont été réalignés avec le logiciel MAFFT (Katoh et al.

2005) avec options par défaut afin d'éliminer le bruit qui avait pu être créé dans l'alignement à cause des séquences très divergentes ou non homologues. Les positions homologues conservées utilisées pour l'inférence des phylogénies ont ensuite été choisies de manière semi-automatique pour une meilleure expertise grâce au programme NET du package MUST (Philippe 1993).

Phylogénies individuelles définitives

Enfin, pour chaque jeu de données, une phylogénie a été inférée par maximum de vraisemblance grâce au logiciel Treefinder version octobre 2008 (Jobb, von Haeseler, and Strimmer 2004) avec le modèle d'évolution LG+G4.

e. Analyse fonctionnelle des 200 nouveaux marqueurs

Pour avoir une idée du rôle de ces protéines dans la cellule, nous avons utilisé les catégories fonctionnelles de la banque COG (R L Tatusov, Koonin, and Lipman 1997; Roman L Tatusov et al. 2003). L'assignation à ces catégories fonctionnelles a été faite par une recherche de similarité par BLASTp contre la banque COG (http://www.ncbi.nlm.nih.gov/COG/ (Roman L Tatusov et al. 2003)) à partir de chaque séquence protéique. Si le premier résultat de BLAST avait une E-value inferieure à 10-5, j'ai assigné sa catégorie fonctionnelle COG à la séquence testée (Annexe 1, Supplementary Table S2).

2. Mise à jour des jeux de données de protéines informationnelles

Les analyses précédentes de la phylogénie des archées par Brochier-Armanet et collaborateurs ont été faites principalement sur les protéines de type informationnel telles que les protéines ribosomiques (Brochier-Armanet, Forterre, and Gribaldo 2011) ou les protéines formant les sous-unités de l'ARN polymérase (Simonetta Gribaldo and Brochier-Armanet 2006), comme expliqué dans l'Introduction. Dans cette nouvelle analyse de la phylogénie des Archaea, nous avons voulu réutiliser cet ensemble de protéines en plus des nouvelles données apportées par notre analyse des génomes complets de *N. maritimus*, *C. symbiosum* et '*Ca.* Caldiarchaeum subterraneum'. La liste des protéines issues de ces deux précédentes études est indiquée dans l'Annexe 4. Les jeux de données correspondants ont été mis à jour pour qu'ils présentent le même échantillonnage taxonomique que les 200 marqueurs que j'ai identifiés. De la même façon, les positions homologues conservées utilisées pour l'inférence phylogénétique ont été choisies manuellement avec le logiciel MUST.

3. Inférence de la phylogénie globale des Archaea

a. Construction des supermatrices

Le but de ce travail était d'une part d'améliorer la résolution de la phylogénie des Archaea, et d'autre part d'étudier le signal phylogénétique contenu dans d'autres protéines que les protéines informationnelles utilisées classiquement. Pour cela, nous avons utilisé une approche de supermatrice de caractères en concaténant différents alignements afin d'avoir un maximum de signal pour l'inférence phylogénétique. Dans cette optique plusieurs concaténations ont été construites : une première avec l'ensemble des 200 jeux de données issus de l'analyse présentée ici, une deuxième avec l'ensemble des protéines ribosomiques et une troisième contenant les sous-unités de l'ARN polymérase et les facteurs de transcription. Enfin, deux supermatrices ont été construites à partir de l'ensemble des marqueurs : une première contenant la totalité des 273 jeux de données et, afin de tester l'impact des données manquantes, nous avons aussi construit une supermatrice en limitant le nombre d'espèces manquantes autorisé à 10 par jeux de données, appelée « Matrice 10 ». Si plus de 10 espèces étaient absentes, le jeu de données n'était utilisé pour la construction de la supermatrice. Les différentes matrices construites et analysées sont indiquées dans le Tableau 5.

	Nouveaux marqueurs	Protéines ribosomiques	ARN polymérase et facteurs de transcription	Matrice totale	Matrice 10
Nombre de positions	48 904	6228	2970	58 102	35 589
Nombre de jeux de données	200	57	16	273	179

Tableau 5 : Les différentes supermatrices construites pour les analyses phylogénétiques.

b. Désaturation

La saturation du signal phylogénétique dans certains marqueurs peut être responsable de l'irrésolution d'une phylogénie, ou du moins de l'irrésolution de la position d'espèces évoluant particulièrement vite (Philippe and Laurent 1998). Cette saturation est liée à des temps d'évolution très longs et/ou à des taux de mutation rapides, qui vont effacer la trace des mutations précédentes, donc du signal évolutif le plus ancien. Les protéines n'évoluant pas toutes à la même vitesse et, au sein de chaque protéine, chaque position pouvant être soumise à une pression de sélection

différente, une possibilité pour diminuer ce biais est de sélectionner les marqueurs et/ou les sites ayant une vitesse d'évolution plus lente. Les deux méthodes ont été employées dans cette étude afin d'essayer de résoudre certaines nœuds problématiques de la phylogénie des archées.

Désaturation par sélection de sites

La désaturation par sélection de sites a été faite par une méthode de « slow-fast » mise en place par Brinkmann et Philippe (Brinkmann and Philippe 1999). Le calcul des vitesses d'évolution par site et la construction des différentes matrices a été faite grâce au logiciel SlowFaster (Kostka et al. 2008). Cette méthode consiste à calculer pour chaque position de l'alignement la vitesse d'évolution du site (nombre de changements survenus par position) dans différents ensembles de séquences définies préalablement. Le taux d'évolution d'une position est estimé par la somme des changements survenus au sein des différents groupes de séquences sélectionnés. Les sites sont ensuite assemblés en différentes matrices : la première contenant les sites ayant le taux d'évolution le plus lent, auxquels sont ajoutés les sites ayant un taux d'évolution de plus en plus rapide. Ainsi, la matrice S_0 contient les sites invariants au sein de chaque ensemble de séquence, la matrice S_1 contient en plus les sites pour lesquels un seul changement a été mis en évidence sur l'ensemble des sous-groupes, et ainsi de suite.

Ici, nous avons choisi d'appliquer cette méthode sur la supermatrice contenant à la fois les nouveaux marqueurs et les protéines informationnelles contenant les jeux de données dans lesquels manquent au plus 10 espèces, ainsi que sur une sélection de 81 espèces (Annexe 1, Supplementary Table S3) correspondant à une liste réduite d'archées avec un choix équilibré de représentants de chaque groupe d'Archaea. 48 matrices ont été construites. Une deuxième série de matrices a été calculée à partir de la même supermatrice de départ mais sans les lignées d'archées nanométriques (*Nanoarchaeum equitans*, le groupe ARMAN (i.e. '*Candidatus* Micrarchaeum acidiphilum' ARMAN-2, '*Candidatus* Parvarchaeum acidiphilum' ARMAN-4, '*Candidatus* Parvarchaeum acidophilus' ARMAN-5) et les Nanohaloarchaea (i.e. '*Candidatus* Haloredivivus sp. G17', '*Candidatus* Nanosalina sp. J07AB43' et '*Candidatus* Nanosalinarum sp. J07AB56'). 47 matrices ont été construites. Pour les inférences phylogénétiques, nous avons choisi d'utiliser seulement les matrices séparées d'une taille d'au moins 500 positions dans un souci temps de calcul /gain de signal, ce qui représente 35 matrices. La liste des matrices utilisées et le nombre de positions qu'elles contiennent sont indiquées dans l'Annexe 5. Seuls les résultats sans les sept espèces placées sur des longues branches sont présentés dans le Chapitre 1.

Désaturation par sélection de gènes

Si la vitesse d'évolution est hétérogène entre les sites au sein d'une protéine, elle l'est également d'une protéine à l'autre. Nous avons voulu prendre en compte de cette hétérogénéité afin de désaturer la supermatrice totale et essayer de faire ressortir le signal porté par des protéines évoluant lentement. Pour cela, nous avons mis en place la méthodologie suivante : pour chaque protéine, un arbre individuel a été inféré par maximum de vraisemblance (avec Treefinder, cf. Matériels et Méthodes B.1.d), les longueurs de branches dans un arbre en maximum de vraisemblance étant le reflet de la vitesse d'évolution estimée de la protéine, nous avons utilisé ces longueurs de branches pour estimer une vitesse d'évolution moyenne de cette protéine. Ensuite, tout comme pour la désaturation par sélection de sites, nous avons construit différentes matrices, la première contenant les protéines évoluant le plus lentement, auxquelles sont ajoutés matrices après matrices, les protéines évoluant plus rapidement.

Concrètement, pour chaque protéine, j'ai extrait des matrices de distances entre chaque paire de séquences grâce au script Newick2Additive.jar développé par Alexis Criscuolo [Communication personnelle]. À partir de ces matrices de distances, j'ai calculé une « distance moyenne » pour l'arbre, en moyennant les distances séparant des couples d'espèces choisies préalablement. Pour cela, nous avons choisi de prendre en compte la distance entre deux groupes d'euryarchées (*Thermococcales* et *Methanomicrobiales*), entre deux groupes de crenarchées (*Sulfolobales* et *Thermoproteales*), entre chacun de ces groupes et les thaumarchées ou l'aigarchée, et entre chaque groupe d'euryarchées et chaque groupe de crenarchées. Pour chaque groupe, l'espèce présente dans le plus de jeux de données était utilisée comme référence, si elle était absente la suivante la plus représentée était utilisée. Au total, au maximum 10 distances ont donc été moyennées. Ce travail a été effectué automatiquement grâce a un script que j'ai mis en place « DistTree.pl », à partir des listes d'espèces utilisées et des matrices de distances extraites des arbres individuels. Enfin, nous avons construit, à partir des 273 jeux de données, 28 supermatrices, la première contenant les 10 jeux de données ayant les vitesses d'évolution les plus lentes, la seconde, les 20 jeux de donnés les plus lents et ainsi de suite (Annexe 6).

c. Inférences phylogénétiques

A partir de ces alignements, des phylogénies ont été inférées par différentes méthodes : en approximation de maximum de vraisemblance grâce au logiciel FastTree 2.1.4 (Price, Dehal, and Arkin 2010), en maximum de vraisemblance grâce au logiciel RaxML version 7.2.4 (Stamatakis 2006; Stamatakis, Hoover, and Rougemont 2008) avec des valeurs de soutien obtenues en Rapid

Bootstrapping et avec PhyML version 3.1 (Guindon et al. 2010), et par une méthode bayésienne avec Phylobayes version 3.3 (Lartillot, Lepage, and Blanquart 2009). Le calcul des phylogénies à partir de supermatrices aussi importantes peut prendre énormément de temps, et avec des différences très importantes selon les méthodes. Certains calculs sont donc toujours en cours, et seule une partie des résultats sont présentés dans le Chapitre 1.

Nous avons aussi fait différentes sélections d'espèces sur ces supermatrices afin d'essayer de résoudre des questions particulières de la phylogénie des Archaea. D'une part en enlevant les lignées correspondants aux archées nanométriques qui se positionnent sur des longues branches, qui peuvent être soumises à des artefacts tels que l'attraction de longues branches (LBA) (Philippe and Laurent 1998), et en les réinsérant une par une. D'autre part, en ne sélectionnant que les euryarchées d'une part et les crenarchées d'autre part permet d'obtenir un meilleur signal au sein de chaque phylum.

Le détail des méthodes et des paramètres utilisés est expliqué dans l'article du Chapitre 1.

C. Racinement de l'arbre des Archaea : méthodes du Chapitre 2

1. Génération des jeux de donnés pour le racinement de l'arbre des Archaea

Certaines des protéines que nous avons utilisées pour l'inférence de la phylogénie des Archaea ont des homologues bactériens. Les protéines ribosomiques, bien sûr, mais aussi certains des 200 nouveaux marqueurs mis en évidence dans cette étude. Nous avons donc utilisé ces homologues bactériens comme groupe extérieur pour essayer de raciner la phylogénie des Archaea.

a. Sélection des protéines d'intérêt

La présence d'homologues bactériens avait déjà été notée lors de l'analyse globale de l'ensemble des phylogénies des trois génomes de *N. maritimus*, *C. symbiosum* et '*Ca.* Caldiarchaeum subterraneum'. Il a suffit de sélectionner parmi les 200 marqueurs de la phylogénie des Archaea, les protéines ayant des homologues bactériens. Il y en avait 81.

b. Construction d'une banque de données locale de génomes complets de bactéries

L'objectif de cette analyse était de raciner la phylogénie des Archaea avec les homologues bactériens pour les marqueurs déjà analysés. Pour cela, nous avons construit une banque de données locale nommée 'BacteriaRacine', contenant les protéines prédites dans 117 génomes complets de bactéries, correspondant à environ cinq représentants par phylum bactérien. Nous avons choisi les représentants de chaque phylum en prenant les cinq génomes contenant le plus de protéines (un représentant par espèce et par genre, dans la mesure du possible) parmi les génomes complets disponibles au début de cette analyse en février 2013 (Annexe 7, Supplementary Table S1).

c. Construction des jeux de données pour l'analyse des marqueurs potentiels

Pour chacun des 81 marqueurs sélectionnés, nous avons fait une recherche de similarité par BLASTp (Altschul et al. 1997) sur la banque 'BacteriaRacine' à partir de la protéine de départ (en utilisant les homologues de *N. maritimus*, *C. symbiosum* ou '*Ca.* Caldiarchaeum subterraneum' comme graine dans la mesure du possible). Les recherches de similarité et l'analyse des résultats de BLAST ont été faites de la même façon que l'analyse des résultats de BLAST sur la banque 'AllArchaea' dans l'étape précédente de cette étude (cf. Matériels et Méthodes B.1). En cas de doute sur le fait que certains homologues bactériens n'aient pas été détectés par BLAST à partir de

la séquence d'archée, j'ai refait une recherche de similarité à partir de la première séquence bactérienne trouvée. Une fois que la sélection des séquences à garder pour l'analyse a été faite, la génération des jeux de données finaux a été réalisée en appliquant le même protocole que dans la première analyse (cf. Matériels et Méthodes B.1.c).

d. Analyse des jeux de données

Les jeux de donnés ont été analysés de la même façon que ceux lors de la première analyse, expliquée dans la partie Matériels et Méthodes B.1.d. A la fin du nettoyage de l'alignement, il ne devait rester qu'une seule séquence par espèce d'archée ou de bactérie, donc au plus 246 séquences par jeu de données. Le nombre de jeux de données conservés est indiqué dans le Tableau 6. La liste des 38 protéines conservées est indiquée dans l'Annexe 7, Supplementary Table S2 et dans l'Annexe 8. La répartition taxonomique de ces 38 protéines parmi les 246 génomes espèces est donnée dans l'Annexe 9.

Génome Nombre de jeux de données	*N. maritimus*	*C. symbiosum*	*'Ca.* Caldiarchaeum subterraneum'
Protéines ayant des homologues bactériens.	58	1	22
Jeux de donnés conservés pour le racinement de la phylogénie des Archaea	35	1	9
Total des jeux de donnés conservés pour l'analyse	38		

Tableau 6 : Nombre de jeux de données conservés à chaque étape de l'analyse visant à raciner l'arbre des Archaea avec les homologues bactériens.

2. Mise à jour des jeux de données de protéines informationnelles pour le racinement de l'arbre des Archaea

De la même façon que les jeux de données de protéines informationnelles ont été utilisés après mise à jour pour l'analyse de la phylogénie globale des Archaea, nous avons complété ces mêmes jeux de donnés en ajoutant les homologues bactériens des espèces que nous avions sélectionnées. Par contre, seules les protéines ribosomiques ont été utilisées ici, car peu de sous-unités de l'ARN polymérase archéenne ont des homologues bactériens, et pour les quelques sous unités pour lesquelles c'est le cas, ces homologues sont très divergents, donc très peu de signal serait réellement exploitable. 32 protéines ribosomiques d'archées ont des homologues bactériens

bien conservés et dont le signal peut être utilisé. La liste de ces protéines utilisées est indiquée dans l'Annexe 7, Supplementary Table S2.

3. Inférence phylogénétique et racinement de l'arbre des Archaea

a. Construction des supermatrices

De la même façon que les jeux de données utilisés pour la phylogénie globale des Archaea ont été concaténés en différentes supermatrices avec différents échantillonnages de jeux de données ou d'espèces, les jeux de données utilisés pour placer la racine de l'arbre des archaea l'ont été. La première supermatrice contient les 38 marqueurs issus de l'analyse précédente, une deuxième les 32 protéines ribosomiques, et une dernière contient l'ensemble de ces 70 jeux de donnés. Le but de cette analyse étant de placer la racine de l'arbre des archaea et non pas de résoudre la phylogénie intragroupe chez les archaea ou chez les bactéries, ces supermatrices ont été construites avec des listes de 81 archées et de 27 bactéries (108 espèces au totale) (Annexe 7, Supplementary Table S1) afin de limiter le temps de calcul, déjà très important pour des supermatrices de cette taille. Les différentes matrices construites et analysées sont indiquées dans le Tableau 7. À partir de ces supermatrices, d'autres sélections d'espèces ont été choisies pour analyser l'effet des espèces placées sur des longues branches par exemple, tout comme dans l'analyse précédente.

	Nouveaux marqueurs	Protéines ribosomiques	Concaténation totale
Nombre de positions	6890	2560	9450
Nombre de jeux de données	38	32	70

Tableau 7 : Différentes concaténations utilisées pour les analyses phylogénétiques de racinement de l'arbre des archées.

b. Désaturation

Le positionnement de la racine de l'arbre des archaea nécessitant d'extraire un signal encore plus ancien que celui pour résoudre les relations intra-archées, les analyses de désaturation étaient aussi indiquées dans ce cas là. La méthodologie appliquée est la même que dans la première analyse.

Désaturation par sélection de sites

La désaturation par sélection de sites pour les jeux de données construits pour la résolution du placement de la racine de l'arbre des archées a été faite de la même façon que dans l'analyse précédente. 10 matrices ont été construites, chacune ayant un gain d'environ 1000 positions en taille par rapport à la précédente. La liste des matrices utilisées et le nombre de positions qu'elles contiennent sont indiquées dans l'Annexe 10.

Désaturation par sélection de gènes

La désaturation par sélection de gènes a été faite de la même façon que dans l'analyse précédente, mais des distances entre trois espèces de bactéries et le reste des espèces d'archées déjà sélectionnées ont été utilisées en plus. Nous avons ainsi construit 10 matrices à partir des 70 jeux de données, la première contenant les 7 jeux de données ayant les vitesses d'évolution les plus lentes puis en ajoutant les jeux de données 7 par 7. (Annexe 11)

c. Inférences phylogénétiques

De même que dans l'analyse (cf. Matériels et Méthodes B.3), des phylogénies ont été inférées avec différentes méthodes, particulièrement avec les logiciel FastTree (Price, Dehal, and Arkin 2010) et RAxML (Stamatakis 2006) dans les mêmes versions que précédemment, et MrBayes v3.2.1 (Ronquist et al. 2012) pour les analyses en inférence bayésienne.

Le détail des méthodes et des paramètres utilisés est expliqué dans l'article du Chapitre 2.

Chapitre 1 : La Phylogénie des archées, au-delà des protéines informationnelles.

A. Introduction

Résoudre la phylogénie des archées est, comme je l'ai expliqué dans l'Introduction, important pour la compréhension de l'histoire évolutive de ce domaine mais aussi de l'ensemble du vivant étant données les interactions entre les archées, leurs milieux et les autres êtres vivants. Les protéines ribosomiques (et par extension les protéines informationnelles, avec les sous-unités de l'ARN polymérase) sont utilisées de plus en plus couramment depuis le séquençage des premiers génomes d'archées pour établir cette phylogénie. Récemment, de grandes avancées dans les techniques de séquençage ont permis l'augmentation exponentielle de la quantité de génomes disponibles et, notamment, pour des taxons couvrant une part de plus en plus grande de la diversité des archées. De nombreuses études ont été conduites récemment (cf. Introduction) pour essayer d'inférer la phylogénie des archées à partir de ces données, mais les marqueurs utilisés restent toujours, pour une grande partie, des protéines informationnelles. De ces analyses ressort le fait que malgré leur qualité en tant que marqueurs phylogénétiques, ces protéines ne permettent pas de résoudre tous les nœuds de l'arbre des archées. De plus, les protéines ribosomiques, les sous-unités de l'ARN polymérase ou les ARNr interviennent tous dans des processus interconnectés dans la cellule et sont peut être porteurs d'un même signal phylogénétique, représentatif non pas de la phylogénie des organismes mais de ces systèmes en particulier.

Afin de répondre à ces questions, nous avons cherché de nouveaux marqueurs pour la phylogénie des archées. Pour cela, nous avons inféré et analysé la phylogénie de chaque protéine codée dans trois génomes (deux thaumarchées, *N. maritimus* et *C. symbiosum* et de l'aigarchée '*Ca.* Caldiarchaeum subterraneum'). Des jeux de données ont été construits en collectant les 200 premières séquences trouvées par BLAST contre une banque locale contenant tous les génomes d'archées disponibles, ainsi que des génomes de bactéries et d'eucaryotes représentatifs de la diversité de ces domaines. Un certain nombre de protéines a été écarté très rapidement, car ne celles-ci ne présentaient pas assez d'homologues. L'analyse des 3212 jeux de données restants est passée par plusieurs étapes de tri, jusqu'à la sélection de 200 protéines porteuses d'un signal suffisamment fiable (sans transfert horizontal de gène ou paralogie). Ces 200 protéines interviennent dans de nombreux processus cellulaires, pas uniquement informationnels. En effet, l'analyse des catégories fonctionnelles COG auxquelles elles sont assignées montre que 53 interviennent dans des voies métaboliques, 24 dans des processus cellulaires et de signalisation, et

24 appartiennent à des familles encore peu caractérisées. Une concaténation de ces 200 nouveaux marqueurs a été construite pour 129 génomes d'archées. Bien que 149 génomes soient disponibles, nous avons décidé de ne conserver qu'un seul génome par espèce dans la mesure où notre travail ne porte pas sur la phylogénie à un niveau taxonomique aussi faible. Parmi ces 129 espèces, 34 n'étaient pas représentées dans la dernière analyse de la phylogénie des archées (Brochier-Armanet, Forterre, and Gribaldo 2011), dont sept nouveaux genres. La phylogénie inférée à partir de cette première supermatrice montre un signal très proche de celui des protéines informationnelles utilisées couramment. Nous avons aussi actualisé ces jeux de données (57 protéines ribosomiques (Brochier-Armanet, Forterre, and Gribaldo 2011) et 16 sous-unités de l'ARN polymérase et facteurs de transcription associés (Simonetta Gribaldo and Brochier-Armanet 2006)), afin d'avoir le même échantillonnage taxonomique que pour les nouveaux marqueurs. Deux supermatrices ont été construites, l'une avec les protéines ribosomiques, l'autre avec le système de transcription. Le signal phylogénétique de ces trois supermatrices étant congruent, nous avons pu construire deux nouvelles supermatrices à partir de l'ensemble de ces 273 protéines. Une première contenait 179 jeux de données pour lesquels moins de 10 espèces sont manquantes (afin d'estimer l'impact des données manquantes) et la seconde contenait l'ensemble des jeux de données, avec un total de 58102 positions. Des méthodes de désaturation site par site et gène par gène ont été appliquées à ces supermatrices afin de réduire la saturation mutationnelle des jeux de données et pour essayer de replacer des espèces évoluant rapidement et sujettes aux artefacts d'attraction de longues branches. Pour la même raison, différents échantillonnages d'espèces ont été réalisés sur les archées de taille nanométrique (*N. equitans*, '*Ca.* Micrarchaeum', '*Ca*. Parvarchaeum' et Nanohaloarchaea).

Ce travail est présenté dans l'article « *Extending the conserved phylogenetic core of Archaea disentangles the evolution of the third domain of Life.*» (Petitjean, Deschamps, López-García, Moreira, Brochier-Armanet, en préparation).

B. Manuscrit de l'article 1 : «Extending the conserved phylogenetic core of Archaea disentangles the evolution of the third domain of Life. »

Extending the conserved phylogenetic core of Archaea disentangles the evolution of the third domain of Life.

Céline Petitjean[1], Philippe Deschamps[1], Purificación López-García[1], David Moreira[1,*], Céline Brochier-Armanet[2,*].

[1]UMR CNRS 8079, Unité d'Ecologie, Systématique et Evolution Université Paris-Sud, 91405 Orsay, Cedex, France.

[2]Université de Lyon, Université Lyon 1, CNRS, UMR5558, Laboratoire de Biométrie et Biologie Evolutive, 43 boulevard du 11 novembre 1918, F-69622 Villeurbanne, France. Tel.: +33 4 26 23 44 76; fax: +33 4 72 43 13 88.

Corresponding authors:

David Moreira (david.moreira@u-psud.fr) and Céline Brochier-Armanet (celine.brochier-armanet@univ-lyon1.fr).

Abstract.

Seminal works aiming at studying the phylogeny of Archaea relied mainly on the analysis of the RNA component of the small subunit of the ribosome (SSU rRNA). The resulting phylogenies have provided interesting but partial information on the evolutionary history of the third domain of life because SSU rRNA sequences do not contain enough phylogenetic signal. Therefore many relationships, and especially the most ancient, remained elusive. Moreover, SSU rRNA phylogenies can be heavily biased by tree reconstruction artifacts. The sequencing of complete genomes allowed using protein markers as alternative to SSU rRNA and the ribosomal proteins are now used routinely to study ancient phylogenies. Taking the opportunity of the recent burst of archaeal complete genome sequences, we have carried out an in-depth phylogenomic analysis. We have identified 200 new protein families that form a conserved phylogenetic core of genes together with the ribosomal proteins and the subunits of the RNA polymerase. The accurate analysis of these markers sheds new light on the evolutionary history of this domain. We resolved a number of important relationships such as those among methanogens Class I. Furthermore the use of desaturation approaches revealed that several relationships recovered in recent analyses are the consequence of tree reconstruction artifacts and allowed replacing the three very fast evolving lineages of nanosized archaea.

Keywords. Phylogenomics, *Methanopyrus kandleri*, Methanohoma, ARMAN, Nanoarchaeota, Nanohaloarchaea, Horizontal gene transfer, mutational saturation, Slow-Fast method.

Running title.

The identification of 200 new conserved phylogenetic markers brings-up the phylogeny of Archaea.

Data deposition: All sequence alignments used in this work are available upon request to the corresponding authors.

Introduction.

The seminal work of Carl Woese and George Fox at the end of the 70's (Woese, Fox 1977) has contributed to establish the RNA component of the small subunit of the ribosome (SSU rRNA) as the gold standard to study the evolutionary relationships among living beings (and especially among microorganisms), and indeed this marker was subsequently proven to be a powerful tool for modern systematics and the exploration of microbial diversity. Among the most important discoveries relying on the analysis of SSU rRNA sequences was the awareness that the living world was divided into three domains (i.e. Archaea, Eucarya and Bacteria) (Woese, Fox 1977) and that most of the biological diversity was represented by uncultured microorganisms (for a recent review on the topic see (Lopez-Garcia, Moreira 2008)).

In the 90's, however, the question was asked whether the phylogenies based on SSU rRNA sequences actually reflect the evolutionary history of organisms or, in other words, whether the SSU rRNA is suitable to trace back the wealth of speciation events that have affected the cellular lineages, especially the most ancient ones (Stiller, Hall 1999; Philippe, Germot 2000). In fact, the phylogenetic signal carried by this molecular marker is too weak to resolve the deepest nodes of the archaeal phylogeny, leading to largely unresolved trees (Robertson et al. 2005; Cavicchioli 2011)but this is specific to neither Archaea nor SSU rRNA given that similar situations have been reported for Bacteria and Eucarya and for other molecular markers (Roger 1999; Philippe et al. 2000; Brochier, Philippe 2002). The lack of phylogenetic signal can result either from radiation, mutational saturation or from a combination of both (Gribaldo, Brochier 2009). Radiation is encountered when the diversification of the lineages under study occurred too rapidly to be recorded at the molecular level, meaning that too few substitutions were fixed between cladogenesis events. Conversely, mutational saturation results from the progressive erasure of the most ancient phylogenetic signal by the accumulation of more recent substitutions occurring at the same sites. As a consequence, in both cases the order of the speciation events is hardly traceable by the phylogenetic analysis of present-day homologues of the studied molecular marker. In addition, phylogenies based on SSU rRNA can be heavily affected by several tree reconstruction artifacts, such as the Long Branch Attraction (LBA), which is due to the heterogeneity of evolutionary rates among the studied sequences and leads to the grouping of the fastest- and the slowest-evolving sequences in different parts of the tree (Felsenstein 1978). This artifact has been particularly well documented in the case of Eucarya and Metazoa (see (Delsuc, Brinkmann, Philippe 2005) and references therein). Additional biases such as those linked to compositional heterogeneity of sequences can also affect phylogenies (Delsuc, Brinkmann, Philippe 2005). Prokaryotic SSU rRNA phylogenies are particularly sensitive to this bias because the base composition of structural RNAs (e.g. SSU and LSU rRNA, tRNA, etc.) is strongly correlated with the optimal growth temperature of the organisms (Woese et al. 1991; Galtier, Lobry 1997). Finally, cases of horizontal gene transfer affecting SSU rRNA genes have been reported (Yap, Zhang, Wang 1999; Bodilis et al. 2012; Kitahara, Miyazaki 2013).

Disentangle the deepest nodes of the Tree of Life, namely deciphering the relationships among the main lineages within each of the three domains, is however crucial because it provides the evolutionary frame indispensable to understand how the present-day diversity arose and how

biological features (e.g. metabolic processes, cellular structures, genomes, etc.) evolved all along the diversification of Life (Gribaldo, Brochier 2009). To overcome the limited phylogenetic signal carried by molecular markers, alternative approaches have been proposed and successfully applied(Delsuc, Brinkmann, Philippe 2005). These included the development of accurate evolutionary models overcoming some of the simplifying assumptions of the Markovian models currently used in molecular phylogenetics (Lartillot, Philippe 2004; Le, Lartillot, Gascuel 2008; Groussin, Boussau, Gouy 2013) and reducing the risk of tree reconstruction artifacts (Lartillot, Brinkmann, Philippe 2007). In parallel, the increase of computational power allowed generalizing the use of statistical methods for phylogenetic inference (e.g. Maximum Likelihood and Bayesian inference) that are less prone to tree reconstruction artifacts such as the LBA (Delsuc, Brinkmann, Philippe 2005).

Beside methodological aspects, the past five years have witnessed a burst of large scale genome sequencing projects covering an ever-growing part of the taxonomic diversity (including the uncultured one) within the three Domains (Wu et al. 2009; Rinke et al. 2013). This windfall of data provides valuable material to tackle complex evolutionary questions, allowing for instance the selection of accurate taxonomic samplings targeting the slowly-evolving sequences within each taxonomic group. Focusing on these sequences which are less susceptible to have undergone multiple substitutions can help to reduce the mutational saturation level of the datasets and thus limit the LBA (Delsuc, Brinkmann, Philippe 2005; Rodriguez-Ezpeleta et al. 2007). Last but not least, the availability of complete genomes has revolutionized phylogenetics, shifting progressively to phylogenomics and thus from single-gene analysis towards the analysis of hundreds of markers either through super-matrix or super-tree approaches (Delsuc, Brinkmann, Philippe 2005). This allows combining the weak phylogenetic signal carried by each individual molecular marker towards a stronger signal, and reducing the global level of noise contained in the data by diluting the noise carried by each individual marker, providing that the biases inherent to each marker are different. In return, phylogenomic approaches require a crucial preliminary and time-consuming step aiming at identifying and selecting the orthologous sequences of each studied marker.

Altogether these approaches have significantly improved our knowledge of ancient evolution. In the case of Archaea, we have shown that components of various biological systems, in particular transcription and translation, formed a conserved phylogenetic core that can be used to trace back the evolutionary history of this domain (Brochier, Forterre, Gribaldo 2005; Gribaldo, Brochier-Armanet 2006). From these analyses, a global picture of the evolutionary history of Archaea is emerging (see (Brochier-Armanet, Forterre, Gribaldo 2011) and references therein). These markers support the division of Archaea into four main phyla: Euryarchaeota, Crenarchaeota, Thaumarchaeota (including the candidate phylum ‘Aigarchaeota’) and Korarchaeota and has confirmed some relationships based on SSU rRNA analyses, among which the grouping of Desulfurococcales and Sulfolobales within Crenarchaeota or the divide of Euryarchaeota in three parts: a basal part containing Thermococcales and methanogens Class I (i.e. Methanopyrales, Methanobacteriales and Methanococcales), an intermediate region containing two groups: Thermoplasmatales and relatives (e.g. DHEV2, the uncultured marine group II, etc.), and Archaeoglobales, and an apical region gathering Halobacteriales and methanogens Class II (i.e. Methanocellales, Methanomicrobiales, Methanosarcinales, and

relatives). Despite these significant advances, the phylogeny of Archaea is far from being fully resolved and a number of important nodes require further investigations (Brochier-Armanet, Forterre, Gribaldo 2011). For instance, several lineages of uncultured nanosized archaea have been discovered in very hot (i.e. Nanoarchaeota (Huber et al. 2002), highly acidic (i.e. ARMAN, for Archaeal Richmond Mine Acidophilic Nano-organisms) (Baker et al. 2006)) and hypersaline (i.e. Nanohaloarchaea (Narasingarao et al. 2012)) environments. Among them, Nanohaloarchaea are widespread and could represent a significant part of the biodiversity of hypersaline ecosystems (Narasingarao et al. 2012). Genome sequencing has revealed that these lineages of tiny archaea are fast evolving and, accordingly, their phylogenetic position is highly debated. Nanoarchaeota were first proposed to represent the earliest branching lineage within Archaea, and therefore to represent a new phylum (Huber et al. 2002; Waters et al. 2003), whereas later analyses suggested that Nanoarchaeota represent in fact a fast evolving euryarchaeal lineage likely related to Thermococcales and that the basal branching observed in some phylogenetic analyses was the consequence of a LBA (Brochier et al. 2005). However, this point is still highly debated (see (Podar et al. 2013) and the comments of the reviewers therein). Regarding ARMAN, a relationship with Thermoplasmatales was initially proposed (Baker et al. 2006), but later analyses suggested that ARMAN could encompass distinct lineages, among which one could be related to Nanoarchaeota (Brochier-Armanet, Forterre, Gribaldo 2011). Finally, Nanohaloarchaea could represent a distant lineage related to Halobacteriales (Narasingarao et al. 2012). However, this should be confirmed by additional investigations. In addition to nanosized archaea, several other newly proposed lineages are debated, such as the Acidilobales within Crenarchaeota (Prokofeva et al. 2009; Mardanov et al. 2010) or the candidate phylum 'Aigarchaeota' (Nunoura et al. 2011), (see (Brochier-Armanet, Forterre, Gribaldo 2011) and references therein). Finally a number of important questions remain to be resolved, among which the relationships among the four archaeal phyla and the root of Archaea, the monophyly of some orders such as the Desulfurococcales, or the relationships among methanogens Class I, and among methanogen Class II and Halobacteriales.

Using exhaustive comparative genomics and phylogenomics approaches, we have identified 200 new conserved proteins widely distributed among the archaeal phyla. Most of them are involved in cellular activities other than transcription and translation. In combination with the classical markers previously used to study the phylogeny of Archaea (namely, 73 transcription and translation proteins), we have carried out phylogenetic analyses with different selections of markers and taxa. Our results confirm many previous observations and clarify several uncertain parts of the archaeal phylogeny, such as the monophyly of the methanogens Cass I, the position of Halobacteriales within the methanogens Class II or the refutation of the order Acidilobales.

Materials and methods.

Dataset assembly.

The proteomes of *Cenarchaeum symbiosum* (Hallam et al. 2006) and *Nitrosopumilus maritimus* (Walker et al. 2010), two species of Thaumarchaeota, and of '*Candidatus* Caldiarchaeum subterraneum', a composite genome assembled from a metagenomic library proposed to

represent a new candidate phylum tentatively called ‘Aigarchaeota’ (Nunoura et al. 2011), were used as starting point to identify protein families of potential interest to study the phylogeny of Archaea.

The 2017 proteins of *C. symbiosum* were compared to the 1796 proteins of *N. maritimus* using a best-reciprocal blast-hit approach. Pairs of proteins displaying more than 60% of identity were considered as members of the same family. This comparison led to the identification of 2470 protein families. Together with the 1704 proteins of ‘*Ca.* Caldiarchaeum subterraneum’, these represented 4174 protein families containing at least one representative of Thaumarchaeota or of the candidate phylum ‘Aigarchaeota’.

Then, these protein families were used to query a local sequence database with blastp (Altschul et al. 1997). This local database contained 92 archaeal, 297 bacterial and 122 eukaryotic proteomes publicly available at the NCBI (http://www.ncbi.nlm.nih.gov/) (Supplementary Table S1) that covered the taxonomic diversity of each domain of Life. The 200 first high-scoring segment pairs with e-values lower than 10^{-5} were retrieved and added to the corresponding protein family. At this step, 962 protein families containing less than four sequences were discarded. The remaining 3212 datasets were aligned with MAFFT version 7 (default parameters) (Katoh, Standley 2013). The resulting alignments were trimmed using BMGE (default parameters) (Criscuolo, Gribaldo 2010). Preliminary phylogenetic trees were inferred with FastTree version 2 (JTT+CAT model) (Price, Dehal, Arkin 2010). The resulting 3212 trees were visually inspected.

Among the 3212 datasets and beside ribosomal proteins and RNA polymerase subunits and transcription factors, 236 corresponded to distinct protein families which could represent good candidates to study the global phylogeny of Archaea because they produced phylogenetic trees supporting strongly the monophyly of this domain and most of its orders. Then these protein families were used to query a local database composed of all available archaeal complete genome sequences (representing 129 different species, see Supplementary Table S3) with blastp and at least one thaumarchaeotal, one crenarchaeotal and one euryarchaeotal sequence as seed. All archaeal homologous sequences within each protein family were aligned using MAFFT and the alignments were trimmed with BMGE. Maximum likelihood (ML) phylogenies were inferred upon the trimmed alignments with PhyML version 3.1 (with the accurate NNI+SPR and tlr options) (Guindon et al. 2010), with the Le and Gascuel (LG) model (Le, Gascuel 2008) and a gamma distribution (Γ) to model the heterogeneity of evolutionary rates across sites (four site categories). Branch robustness was estimated with the non-parametric bootstrap procedure implemented in PhyML (100 replicates of the original alignments). The resulting trees were visually inspected. 36 protein families of the original 236 were discarded at this step because they present complex patterns of gene duplications and losses, and possible cases of horizontal gene transfer among archaeal species that were not apparent when the taxonomic sampling was restricted to 91 archaea. As a result, 200 conserved proteins were kept for final analyses (Supplementary Table S2). When in-paralogous sequences were present in these 200 protein families, we kept the slowest-evolving copy for the phylogenetic analyses.

In parallel, the 57 ribosomal proteins and 16 protein families corresponding to the 14 subunits of the RNA polymerase and transcription factors used in previous studies on the phylogeny of

Archaea (Brochier, Forterre, Gribaldo 2004; Brochier-Armanet, Forterre, Gribaldo 2011) were updated according to the same procedure. These 73 protein families in combination to the 200 newly identified in this study represented 273 phylogenetic markers useful to investigate the evolutionary history of Archaea.

We made the choice to not include a bacterial outgroup in our analyses in order to limit long branch attraction (LBA) artifacts which could occur between the long branch leading to the outgroup and the fast evolving lineages within the ingroup, a situation which is frequently encountered when studying ancient evolution (Philippe, Laurent 1998; Gribaldo, Philippe 2002). Moreover, the inclusion of a bacterial outgroup would have severely limited the number of proteins that could be used to study the phylogeny of Archaea.

Supermatrix construction.

The trimmed alignments of the 273 proteins were combined to build five supermatrices. Supermatrices L2, L3 and L4 corresponded to the 200 new proteins (48,904 amino acid positions), the 57 ribosomal proteins (6,228 amino acid positions) and the 16 protein families corresponding to 14 subunits of the RNA polymerase and transcription factors (2,970 amino acid positions), respectively. Supermatrix L1 (35,589 amino acid positions) was built by concatenating the protein families present in more than 119 species (i.e. 106 newly identified proteins, 57 and 16 ribosomal and transcription proteins), which represented less than 10% of missing data per family, whereas the supermatrix XL1 gathered all the 273 proteins (58,102 amino acid positions).

Phylogenetic analysis of the supermatrices.

ML trees of the L1, XL1, L2, L3 and L4 amino acid supermatrices were inferred using PhyML version 3.1 (NNI+SPR and tlr options) with the LG+Γ8 evolutionary model. The robustness of the resulting trees was assessed using either the non-parametric bootstrap procedure implemented in PhyML (100 replicates of the original alignment) or the SH-like support. Bayesian inferences (BI) were performed using PhyloBayes 3.3b with the CAT+Γ8 model. In addition to the amino acid sequence supermatrices, two recoded versions of the L1 supermatrices were analyzed with phylobayes using dayhoff4 and dayhoff6 options (L1-REC4 and L1-REC6, respectively). The four and six Dayhoff's amino acid families corresponded to [(A,G,P,S,T) (D,E,N,Q) (H,K,R) (F,Y,W,I,L,M,V)] plus cysteins treated as missing data (C= ?) and to [(A,G,P,S,T) (D,E,N,Q) (H,K,R) (F,Y,W) (I,L,M,V) (C)]. Two chains were run for at least 10000 cycles, saving one tree in ten. The first 500 trees were discarded as "burn-in" and one on two of the remaining trees from each chain were sampled to test for convergence and to compute the 50% majority rule consensus tree.

Desaturation procedure.

We used two complementary approaches to reduce the mutational saturation level contained in our data.

First, a site-by-site desaturation of the L1 supermatrix was carried out with the Slow-Fast method (Brinkmann, Philippe 1999). This site-by-site desaturation strategy was applied only to the L1

supermatrix because the larger amount of missing data contained in the XL1 supermatrix could bias the estimation of the evolutionary rates. To do so, we subdivided the sequences of the L1 supermatrix into taxonomically balanced groups shown to be monophyletic in previous studies and in this analysis by selecting four to seven sequences per group. The considered groups were Thaumarchaeota + 'Aigarchaeota', Thermoproteales, Sulfolobales, Desulfurococcales, Thermococcales, Methanococcales, Methanobacteriales, Thermoplasmatales + uncultured marine group II + DHVE2, Archaeoglobales, Methanocellales, Methanosarcinales, Methanomicrobiales and Halobacteriales (Supplementary Table S4). Because they were crucial taxa, Korarchaeota and Methanopyrales were also included in the analysis, despite they were each represented by a single species. The evolutionary rate of each amino acid site of the L1 supermatrix was estimated with the program SlowFaster (Kostka et al. 2008). 35 alignments (S_0 to S_{34}) of increasing size were built by incorporating more and more rapidly-evolving sites.

Second, we applied a protein-by-protein desaturation strategy aiming at identifying and focusing the phylogenetic analyses on the slowly evolving proteins. To do so, we estimated the evolutionary rate of each protein by measuring the average evolutionary distances among sequences. Because the taxonomic sampling varied for the different proteins, those average distances among all pairs of sequences were not directly comparable. We thus calculated the average distances among five taxonomic groups widely represented among the 273 protein families: Methanobacteriales, Thermococcales, Sulfolobales, Thermoproteales and Thaumarchaeota + 'Aigarchaeota' to approximate the average evolutionary rate of each marker. We then constructed 28 supermatrices (P_0 to P_{27}) of increasing size by concatenating the protein datasets according to their average evolutionary rates, from the slowest- to the fastest-evolving ones.

Results.

Expending the conserved phylogenetic core of Archaea.

Most of the previous works on the phylogeny of Archaea have relied on the analysis of proteins involved in informational systems (in particular ribosomal proteins, subunits of the RNA polymerase, elongation factors, etc.). The in-depth survey of the proteomes of two thaumarchaeotal species *Cenarchaeum symbiosum* and *Nitrosopumilus maritimus* and the aigarchaeon '*Ca.* Caldiarchaeum subterraneum' led us to the identification of 200 new conserved proteins widely distributed among Archaea (Figure 1A and Supplementary Table S1). These were either specific to Archaea or strongly supported the monophyly of this domain and its orders, suggesting that they could have been present in the last common ancestor of archaea (LACA) and vertically inherited all along the diversification of this domain.

Together with the 57 ribosomal proteins and the 14 RNA polymerase subunits and transcription factors (encoded by 16 genes, because RpoA and RpoB are split in some species) retrieved from previous works (Matte-Tailliez et al. 2002; Brochier, Forterre, Gribaldo 2004), these represented 273 phylogenetic markers of potential interest to investigate the phylogeny of Archaea, most of them being totally new and never used before. Interestingly, the majority of the 200 proteins identified by analyzing the proteomes of the two thaumarchaeotes and of 'Ca. Caldiarchaeum subterraneum' did not correspond to genes involved in informational processes

(Figure 1B). More precisely, 51 were involved in metabolism, 23 in cellular processes and signaling, 2 in other functions, whereas 47 correspond to poorly characterized protein families. Importantly, to a few exceptions, these protein families were widely distributed in Archaea, indicating that genes involved in non-informational processes can be strongly conserved over large evolutionary times. Beside phylogenetic considerations, the likely presence of these genes in the last common ancestor of Archaea provided interesting biological information on this ancestor. Among the most interesting examples, we detected FtsZ, which, despite being absent in many crenarchaeotes (Bernander 2000), was probably involved in cell division in the archeal ancestor. We also detected enzymes involved in homoserine, thiamine, aspartate, and pseudouridine metabolism, as well as subunits of the exosome and proteasome complexes. The list also includes a large number of ATPase subunits, as well as enzymes involved in hydrogenase expression and maturation, all of them probably related to energy conversion. Given the wide distribution and phyletic patterns of these proteins in archaea, they were most likely present in the last archaeal common ancestor.

Archaeal homologues of the whole set of 273 protein families were retrieved from complete genome sequences of 129 species available at the NCBI (Figure 1). These included 34 new species compared to recent studies (Brochier-Armanet, Forterre, Gribaldo 2011), increasing significantly the taxonomic diversity of Archaea that can be considered in massive phylogenomic analyses. The new genomes included the three Nanohaloarchaeota , three additional thaumarchaeotes, six halobacteriales (including the new genera *Natrinema*, *Halopiger*, and *Natronobacterium*), six methanogen Class II and six methanogen Class I (including the new genera *Methanosalsum* and *Methanolinea*), one archaeoglobales, three thermococcales, two desulfurococcales (including the new genus *Pyrolobus*), three sulfolobales and four thermoproteales (including the new genus *Thermoproteus*).

The 273 proteins were checked to identify orthologues and remove paralogous and xenologous sequences. After this trimming step, the alignments were concatenated to generate three supermatrices: L2, L3 and L4 by gathering the 200 newly identified protein families (48,904 amino acid positions), the 57 ribosomal proteins (6,228 amino acid positions) and the 16 proteins involved in transcription (2,970 amino acid positions), respectively. The size of the ribosomal protein supermatrix L3 was in agreement with previous analyses (Matte-Tailliez et al. 2002; Gribaldo, Brochier-Armanet 2006) but smaller than supermatrices from recent reports (Yutin et al. 2012; Podar et al. 2013) despite the use of a similar set of proteins, most likely because of less stringent alignment-trimming criteria in these studies. At this step we removed the seven nanosized archaeal lineages (*Nanoarchaeum equitans*, three Nanohaloarchaeota and three ARMAN species) because of their fast evolutionary rates (Brochier-Armanet, Forterre, Gribaldo 2011; Narasingarao et al. 2012) and their lesser representation in the 273 protein families compared to other species (Figure 1D) in agreement with the relative small size of their genomes, which made them very prone to potential tree reconstruction artifacts (Roure, Baurain, Philippe 2013), meaning that only 122 species were represented in each supermatrix. The ML trees resulting from the three supermatrices showed similar topologies indicating that these supermatrices carried a consistent phylogenetic signal (Supplementary Figures S1-S3). More precisely, the monophyly of the archaeal phyla and orders represented by more than one sequence was recovered and well supported (all SH supports >0.95, except for

Desulfurococcales). Moreover, the relationships among the main archaeal orders within each phylum were consistent and in agreement with previous studies (Brochier-Armanet, Forterre, Gribaldo 2011). The strongest discrepancy concerned the position of *Methanopyrus kandleri*, which branched at the base of the Euryarchaeota in the L4 tree, whereas it emerged after the divergence of Thermococcales in the L2 and L3 trees. A similar branching pattern observed in previous analyses of the RNA polymerase and transcription factors was interpreted as the result of a LBA due to the fast evolutionary rate of some components of the *M. kandleri* transcription apparatus (Brochier, Forterre, Gribaldo 2004). The hypothesis of a LBA is strengthened by our new analyses because the basal branching of *M. kandleri* was not observed when we applied Bayesian inference (with the CAT+Γ8 model), less sensible to LBA, to analyze of the same RNA polymerase and transcription factors supermatrix (Supplementary Figures S4).

Because they carried a consistent global phylogenetic signal, the 273 protein families were then combined to build two very large supermatrices: L1 (35,589 amino acid positions), obtained by concatenating the 179 proteins present in more than 119 species, and XL1 (58,102 amino acid positions), which resulted from the combination of all the 273 proteins. Without surprise, the ML trees inferred with these two supermatrices (Figure 2) were consistent with the L2, L3 and L4 supermatrices (Supplementary Figures S1-S3) and with previous studies (Brochier-Armanet, Forterre, Gribaldo 2011).

Towards a well-resolved global phylogeny of Archaea.

The L1 and the XL1 supermatrices were the largest reported so far to study the phylogeny of Archaea and represented a unique opportunity to decipher in detail the evolutionary history of this domain. We were particularly interested in investigating the quantity and the quality of the phylogenetic signal contained in these protein families and the possible biases which could affect the phylogeny of Archaea. To do so, we applied a two-steps strategy. On the one hand, we applied a site-by-site desaturation strategy, the SlowFast method, to the L1 supermatrix to investigate the relationships among the archaeal phyla and the relative branching pattern of the orders within each phylum. This strategy aimed at reducing the mutational saturation of the data and thus limiting the risk of tree reconstruction artifacts. On the other hand, we used the recoded L1 supermatrices (L1-REC4 and L1-REC6) to reconstruct the BI phylogenies of Euryarchaeota and Crenarchaeota separately to study in detail the evolutionary history of these phyla. This allowed both reducing the mutational saturation and amino acid compositional biases.

For the site-by-site desaturation strategy we used the L1 supermatrix to generate 35 SF-supermatrices (S_0 to S_{34}) by including sites with increasing evolutionary rates (Figure 3A) and we inferred ML trees with them Supplementary Figures S5). First of all and without surprise, all the reconstructed trees strongly recovered the monophyly of Crenarchaeota and of Euryarchaeota (Figure 3B). We observed an increasing phylogenetic signal supporting the unrooted [(Thaumarchaeota+'Aigarchaeota',Korarchaeota),(Crenarchaeota,Euryarchaeota)] topology, with the maximal support being associated to the SF-ML-trees S_8 to S_{12} (Figure 3B), but the support in favor of this topology decreased progressively when faster evolving sites were considered. Because we did not include any outgroup sequence to avoid possible LBA artifacts and to maximize the number of conserved alignment positions (see methods), this topology could not be interpreted in terms of evolutionary relationships among the four main archaeal lineages.

However, this topology was in agreement with some studies (Elkins et al. 2008), but in contradiction with others (Elkins et al. 2008; Kelly, Wickstead, Gull 2010; Brochier-Armanet, Forterre, Gribaldo 2011; Guy, Ettema 2011; Nunoura et al. 2011; Eme et al. 2013). However, because it was supported by the slowly evolving positions in the site-by-site desaturation approach, this topology could reflect the true structure of the archaeal domain.

Furthermore, all the SF-ML trees supported the Thermococcales and Thermoproteales as the first diverging lineages within the Euryarchaeota and Crenarchaeota, respectively (Figure 3B), in agreement with earlier studies (Matte-Tailliez et al. 2002; Gribaldo, Brochier-Armanet 2006; Brochier-Armanet, Forterre, Gribaldo 2011). This provided additional support in favor of the hypothesis that methanogenesis was not the most ancient metabolism, neither in Archaea nor in Euryarchaeota and, therefore, not in Life history (Gribaldo, Brochier-Armanet 2006; Brochier-Armanet, Forterre, Gribaldo 2011), contrarily to what is sometime assumed (Xue et al. 2005; Wong et al. 2007; Yu et al. 2009).

The L1-REC6 BI tree of the Crenarchaeota and Euryarchaeota phyla were consistent (Figure 4) and in agreement with previous works (Brochier-Armanet, Forterre, Gribaldo 2011). The recoding of the L1 supermatrix with the Dayhoff4 scheme provided very similar trees (not shown). Regarding Crenarchaeota (Figure 4A), the monophyly of the order Thermoproteales and its genera was clearly supported (all Posterior Probabilities (PP) = 1.0). Similarly, the grouping of the genera *Pyrobaculum* and *Thermoproteus* on the one hand and of *Caldivirga* and *Vulcanisaeta* on the other hand was recovered (PP = 1.0 and PP = 0.82, respectively), whereas *Thermofilum* represented the first lineage diverging within this order (all PP = 1.0). Within Sulfolobales which appeared also as monophyletic (PP = 1.0), the monophyly of *Metallosphaera* was strongly supported as well as its grouping with *Acidianus* (PP = 1.0). In contrast, the monophyly of *Sulfolobus* was not recovered because *Sulfolobus islandicus* and *Sulfolobus solfataricus* branched together at the base of the *Acidianus* and *Metallosphaera* group, whereas *Sulfolobus tokodaii* and *Sulfolobus acidocaldarius* represented the first lineage diverging within Sulfolobales (PP = .95). This suggested that the genus *Sulfolobus* may be paraphyletic and encompasses in fact distinct taxonomic lineages.

While Thermoproteales and Sulfolobales represented monophyletic orders, the situation was less clear in the case of Desulfurococcales, which were monophyletic in the ML tree of the L1 supermatrix (Figure 2) but paraphyletic in the L1-REC6 BI tree, due to the grouping of Sulfolobales with either *Ignicoccus hospitalis* (the host of the nanoarchaeon *Nanoarchaeum equitans*) albeit with a weak support (PP = 0.8, Figure 4). The non monophyly of Desulfurococcales was reported and discussed earlier (Brochier-Armanet, Forterre, Gribaldo 2011). As for *Sulfolobus*, this could indicate that the order Desulfurococcales is actually non-monophyletic or, conversely, that the relationship between *I. hospitalis* and the Sulfolobales is artefactual. In that case, it could be the result of a few HGT events that were not detected by our trimming procedure or of a LBA artifact between the long stems leading to these two lineages.

Regarding the other Desulfurococcales lineages, the monophyly of *Desulfurococcus* and of *Staphylothermus* was strongly supported (PP = 1.0). *Desulfurococcus* grouped robustly with *Thermosphaera aggregans* (PP = 1.0), and represented the sister-lineage of *Staphylothermus*, whereas *Ignisphaera aggregans* emerged more deeply. *Aeropyrum pernix* and *Acidilobus*

saccharovorans (PP = 1.0) on the one hand and *Hyperthermus butylicus* and *Pyrolobus fumarii* (PP = 1.0) on the other hand, grouped together (PP = 0.93). This confirmed that *Acidilobus* is a true member of Desulfurococcales (Brochier-Armanet, Forterre, Gribaldo 2011) and does not represent a separate order of Crenarchaeota as recently proposed (Prokofeva et al. 2009; Mardanov et al. 2010).

As for the Crenarchaeota, the L1 REC6 and L1-REC4 BI euryarchaeal trees were consistent (Figure 4B and not shown) and mainly in agreement with previous work (Brochier-Armanet, Forterre, Gribaldo 2011). The monophyly of each order was recovered and well supported: Thermococcales, Methanobacteriales, Methanococcales, Thermoplasmatales, Halobacteriales, Methanosarcinales, Methanomicrobiales, Methanocellales and Archaeoglobales (all PP = 1.0, Figure 4B). Within Thermococcales, the monophyly of *Pyrococcus* was confirmed (PP = 1.0) but not of *Thermococcus*, due to the early divergence of *Thermococcus barophilus*, *Thermococcus sibiricus* and *Thermococcus litoralis* (PP = 1.0), whereas the four remaining *Thermococcus* species formed the sister-lineage of *Pyrococcus* (PP = 1.0). This suggested that the genus *Thermococcus* encompasses distinct lineages. The next diverging lineage gathered methanogens Class I: the Methanobacteriales, Methanopyrales and Methanococcales (PP = 1.0), the first two forming a monophyletic group (PP = 0.99), meaning that the monophyly of methanogens Class I was strongly recovered (Figure 4B). These groupings were strengthened by the site-by-site desaturation approach when the slowest evolving sites contained in the SF-supermatrix were considered (Figure 3C, red and blue lines). When faster and faster evolving sites were included (SF-supermatrices S_6 to S_{34}), the support for the monophyly of methanogens Class I decreased progressively and a grouping between Methanococcales and Methanobacteriales appeared. This indicated that the grouping of these two orders and the non monophyly of methanogens Class I observed in the ML L1 and XL1 trees (Figure 2) and in other studies (Brochier-Armanet et al. 2008; Wolf et al. 2012; Yutin et al. 2012; Podar et al. 2013) were likely artefactual. Worth noticing, the grouping of Methanopyrales and Methanobacteriales, which was supported by the slowly evolving sites, was also supported by phylogenies based on genome gene content (Makarova et al. 2007) and by the presence of pseudomurein in their cell wall, a unique biological feature shared by these two archaeal orders(see (Albers, Meyer 2011) and references therein). Pseudomurein could therefore represent a synapomorphy of this lineage.

Within Methanobacteriales (Figure4B), the monophyly of the genera *Methanothermobacter* and *Methanobrevibacter* was well recovered (PP =1.0), as well as the grouping of *Methanosphaera* and *Methanobacterium* sp. AL-21 (PP = 1.0) together with the genus *Methanobrevibacter* (PP = 1.0), whereas *Methanothermus* represented the first lineage branching within this order (PP = 1.0). In the case of the Methanococcales, the genera *Methanocaldococcus* and *Methanotorris* were monophyletic (PP = 1.0). In contrast, *Methanococcus* appeared paraphyletic due to the clustering of *Methanococcus aeolicus* with *Methanothermococcus okinawensis* (PP = 1.0). This was in agreement with a large scale phylogenomic analysis of Methanococcales based on the universally conserved proteins within this order (Céline Brochier-Armanet and Michel Lecocq, unpublished data) and suggested that *Methanococcus* does not form a monophyletic genus or, conversely, that *Methanothermococcus* should be reclassified within the genus *Methanococcus*.

Regarding the intermediate part of the euryarchaeotal tree, the L1-REC6 BI tree and the site-by-site desaturation analysis strongly supported the clustering of Thermoplasmatales, DHEV2 (represented by *Aciduliprofundum boonei*) and the uncultured marine group II lineage (PP = 1.0), as well as the divergence of Thermoplasmatales and relatives prior the divergence of Archaeoglobales (Figure 2E, and PP = 1.0, Figure 4B). Among these, Thermoplasmatales appeared more closely related to DHEV2 (PP = 1.0) than to the marine group II. However, the very long branch of the latter and its close relationship with Thermoplasmatales in L1-S_1 ML trees inferred with the slowest evolving sites (Figure 3E) deserved consideration: did it result of stochastic effects or could it indicate that basal branching of the marine group II resulted from a LBA? Due to the very restricted taxonomic sampling available for this clade (DHEV2 and marine group II were represented each by only one genome), a definitive answer regarding the phylogenetic position of marine group II would require additional data and analyses. Within Thermoplasmatales, the monophyly of *Thermoplasma* and the sister-relationship between *Ferroplasma* and *Picrophilus* was confirmed (PP = 1.0) (Figure4B). In contrast, the monophyly of the genus *Archaeoglobus* within the Archaeoglobales was not recovered, due to the basal branching of *Archaeoglobus profundus*, albeit with a weak support (PP = 0.83). The apical part of the euryarchaeotal tree gathered Halobacteriales and the Methanomicrobia, represented by three orders of methanogens: Methanomicrobiales, Methanocellales and Methanosarcinales. These orders were proposed to form a second class of methanogens, tentatively called methanogens Class II, representing the sister-lineage of the Halobacteriales (Bapteste, Brochier, Boucher 2005). While the monophyly of methanogens Class II was recovered in some phylogenetic analyses (Wolf et al. 2012; Yutin et al. 2012), it was not in others (Brochier-Armanet et al. 2008; Csuros, Miklos 2009; Kelly, Wickstead, Gull 2010; Groussin, Gouy 2011; Podar et al. 2013), and thus remained debated (Brochier-Armanet, Forterre, Gribaldo 2011). In our study, the monophyly of methanogens Class II was recovered neither in the L1-REC6 BI tree nor by the site-by-site desaturation strategy (Figure 4F and Figure 3B). The L1-REC6 BI tree strongly supported the emergence of Halobacteriales within Methanomicrobia as the sister-lineage of Methanomicrobiales (PP = 1.0), and the early divergence of Methanosarcinales (PP = 1.0). However, the picture provided by the analysis of the slowest evolving positions was slightly more complex and highlighted potential artifacts. First, the SF-ML trees based on the slowest evolving positions (S_1 to S_6, Figure 3F) favored a sister-grouping between Methanocellales and Halobacteriales, with Methanosarcinales as the first diverging lineage within methanogens Class II (S_1 to S_4, Figure 3D), suggesting that the branching pattern observed in ML and BI trees based on the complete L1 supermatrix could be artefactual (Figure 2B and Figure 4B). The relationships among methanogens Class II and Halobacteriales shifted dramatically following the addition of 1376 positions (transition from S_6 to S_7): Methanocellales moved to the base of methanogens Class II, whereas Halobacteriales grouped with Methanomicrobiales (Figures 2D and 2F). Interestingly, while the association between Halobacteriales and Methanomicrobiales remained stable in the ML trees built with SF-supermatrices with an index greater than 20 (Figure 2F), an association between Methanosarcinales and Methanocellales appeared (Figure 2D). However, this relation was highly suspicious because it occurred only when the fastest evolving positions of the L1 supermatrix were taken into account.

Origin of the nanosized archaeal lineages.

One of the most intriguing questions regarding the phylogeny of Archaea regarded the phylogenetic position of the newly discovered nanosized lineages, namely the two uncultured ARMAN lineages (Baker et al. 2006), the Nanoarchaeota (Huber et al. 2002) and the uncultured Nanohaloarchaea (Narasingarao et al. 2012). Due to fast evolutionary rates, their phylogenetic position has been particularly difficult to assess and is debated (Brochier-Armanet, Forterre, Gribaldo 2011; Podar et al. 2013). To address this question we applied a protein-by-protein desaturation strategy aiming at identifying the slowly evolving proteins to build 28 supermatrices (P_0 to P_{27}) by including faster and faster evolving proteins. The supermatrices gathering the slowest evolving proteins were expected to contain the most reliable phylogenetic signal. The ML trees based on the 10 slowly evolving genes (P_0, not shown) confirmed the sister-relationship between Nanohaloarchaea and Halobacteria (SH = 0.998) (Narasingarao et al. 2012). It also strongly supported the late divergence of '*Ca.* Micrarchaeum acidiphilum' within Euryarchaeota in agreement with SSU rRNA analyses (Baker et al. 2006; Baker et al. 2010) and a recent report (Brochier-Armanet, Forterre, Gribaldo 2011), but in contradiction analyses based on single protein (RadA, EF-2, EF-1A and SecY) (Baker et al. 2010). More precisely '*Ca.* Micrarchaeum acidiphilum' branched robustly at the base of Thermoplasmatales, DHVE2 and group II (SH = 0.977). Finally, '*Ca.* Parvarchaeum acidophilus' and *Nanoarchaeum equitans* grouped together (SH = 0.925) at the base of the Thermococcales (SH = 0.493) consistently with a recent report (Brochier-Armanet, Forterre, Gribaldo 2011). However, this relationship should be considered with caution because it could result from a LBA given the very long branches of these two nanosized lineages. Strengthening the hypothesis of a LBA, the two '*Ca.* Parvarchaeum' branched robustly at the base of the Thermoplasmatales, DHVE2 and group II (SH = 0.997) in the P2 ML tree, when the other fast evolving lineages were removed from the analysis (not shown).

Finally, we have showed that the L1 supermatrix contained a reliable phylogenetic signal, especially when the slowly evolving positions were considered. From our previous analysis, we established that the best ratio between phylogenetic signal and noise was found in the L1-S_2 to L1-S_5 supermatrices, and that a shift occurred between the L1-S_6 and L1-S_7 supermatrices. Accordingly, we used these supermatrices to test the phylogenetic position of *Nanoarchaeum equitans*. To do so, added the sequences of this tiny archaea in the S_1 to S_6 supermatrices. Their phylogenetic analysis provided trees supporting the grouping of Nanoarchaeota within Euryarchaeota and more precisely as the sister-lineage of Thermococcales, albeit with moderate supports, whereas a branching at the base of Euryarchaeota as observed for the S_7 and S_8 matrices (not shown).

Discussion

Taking the opportunity of the recent burst of complete genome sequences covering an ever-growing part of the archaeal diversity we have identified 200 proteins of phylogenetic interest which can be used in combination with the 73 proteins of the ribosome (Matte-Tailliez et al. 2002) and transcription apparatus (Brochier, Forterre, Gribaldo 2004) that have been employed in recent years to investigate the phylogeny of the third domain of Life. The discovery that the transcription and the translation apparatus carry a consistent phylogenetic signal was an

important step because it highlighted the existence of a conserved phylogenetic core of genes useful to study the ancient evolution of Archaea (Brochier, Forterre, Gribaldo 2005). However, one could argue that transcription and translation do not represent independent systems because they are coupled in archaea (French et al. 2007), as it is also the case in bacteria. Therefore, it can be speculated that they have coevolved (including gene losses and horizontal gene transfers), explaining why they carry a similar phylogenetic signal. Most of the 200 markers identified in this study are not functionally linked to translation or transcription. More important, the majority are not involved in informational processes. Therefore, the consistent phylogenetic signal carried by these markers and the transcription and translation apparatuses cannot be explained by functional links. This is a strong argument for the existence of a conserved core of genes carrying a phylogenetic signal that reflects the evolutionary relationships of organisms. This contrasts with recent studies claiming that horizontal gene transfers have overwhelmed the evolutionary history of prokaryotes, which cannot be represented by a tree and should be replaced by a network (or a rhizome, etc.) (Dagan, Martin 2006). However, networks of genes can neither capture the complexity of the evolutionary history of living beings (Gribaldo, Brochier 2009). In particular, they are not suitable to represent speciation events or gene duplications and losses that are important processes of biological evolution. On the contrary, deciphering the vertical relationships among organisms provides the indispensable framework to analyze the evolutionary history of the genomes (including gene transfers, duplications and losses). Altogether, the 273 genes composing the conserved phylogenetic core of Archaea represent 10-20% of the gene content in typical archaeal genomes. This is one order of magnitude more than the provocative picture of the tree of 1% (Dagan, Martin 2006). This conserved phylogenetic core of genes, which were inherited from the last common ancestor of Archaea, is not strictly conserved in all lineages, meaning that differential gene losses or punctual horizontal gene transfers could have occurred in one or the other lineage during the diversification of this domain. It corresponds to the “soft” conserved phylogenetic core of genes we defined in 2006 (Gribaldo, Brochier-Armanet 2006), which is much larger than the strictly conserved phylogenetic core of genes that is the set of universal archaeal genes. While the latter is fatted to disappear when a larger and larger part of the diversity will be taken into account in the analyses, the former is not. On the contrary, taking into consideration more genomes and more lineages allows distinguishing accessory genes composing the variable shell from those which were ancestrally present and punctually lost in each lineage, thus increasing the size of the soft conserved phylogenetic core of genes. For instance, taking into account ‘*Ca*. Caldiarchaeum subterraneum’ (the sole representative of the candidate phylum ‘Aigarchaeota’) allowed identifying 43 additional protein families that were not detected when we used as reference the genomes of the two thaumarchaeota *Cenarchaeum symbiosum* and *Nitrosopumilus maritimus*.

The accurate analysis of the 273 markers through supermatrix and desaturation approaches brought a new light on the evolutionary history of Archaea. First, we confirmed a number of nodes that were proposed before (e.g. the early emergence of Thermococcales within Euryarchaeota, the grouping of Thermoplasmatales with DHEV2 and the uncultured marine group II, or the branching of the candidate phylum ‘Aigarchaeota’ at the base of Thaumarchaeota). We also pointed some discrepancies between the phylogeny of Archaea and the current taxonomy of this domain (e.g. the refutation of the order Acidilobales, or the paraphyly of the genera *Methanococcus* and *Sulfolobus*) that deserves further investigation and

could lead to revise the current archaeal taxonomy. More importantly, we settled a number of important debated relationships such as the reliability of the methanogens Class I which appeared strongly supported by the slowest evolving sites and by Bayesian inferences. Therefore, it represents a *bona fide* class for which we propose the name Methanohoma (from the Greek ὸμάς meaning group), accordingly the classes Methanococci/Methanothermea, Methanobacteria and Methanopyri should be considered as sub-classes. The apparent non-monophyly of this class observed in some phylogenetic analyses was most likely due to tree reconstruction artifacts and/or biases. We also provided some support for the monophyly of Desulfurococcales. While significant advances emerged from our analyses, a few relationships remained elusive. The most important concerned the relationship among the methanogens Class II and Halobacteriales, which can reflect a lack of phylogenetic signal despite the use of very large supermatrices and accurate desaturation approaches and/or biases linked to HGT that would have escaped our trimming procedure. Even if a close relationship between Methanocellales and Halobacteriales seemed to emerge, additional data and analyses will be required to definitively close the question. Finally our site-by-site and protein-by-protein desaturation analyses clarified the origin of nanosized archaeal lineages. We showed that they are bona fide Euryarchaeota. More precisely, we confirmed the sister-ship between Nanohaloarchaea and Halobacteria. We refined the phylogenetic position of '*Ca.* Micrarchaeum acidiphilum' and supported their emergence of at the base of the large group comprising Thermoplasmatales, DHVE2 and group II. In contrast, additional studies are required to definitively conclude about the position of the two 'Ca. Parvarchaeum'. The analysis of the slowly evolving positions and proteins both confirmed the relationship between Thermococcales and Nanoarchaeota proposed nearly ten year ago (Brochier et al. 2005) and was in agreement with a recent analysis (Brochier-Armanet, Forterre, Gribaldo 2011). This suggested that their emergence at the base of Euryarchaeota recently reported (Podar et al. 2013) was the consequence of LBA. LBA could also explain a number of conflicting relationships observed. This was the case for instance of the grouping of Methanococcales and Methanobacteriales or of the Halobacteriales and Methanomicrobiales. The use of desaturation approaches, and accurate reconstruction methods and evolutionary models can overcome these artifacts. The use of rough phylogenetic reconstruction methods (e.g. distance or approximate maximum likelihood ones) or non-realistic evolutionary models should thus be avoided for phylogenomic studies aiming at reconstructing the very ancient relationships.

Figure legends.

Figure 1. Number of archaeal orthologues **(A)**,functional distribution **(B)** and **(D)** taxonomic distribution of the 200 newly identified protein markers. **(C)** Number of archaeal orthologues of transcription (grey) and ribosomal proteins (black).

Figure 2. Phylogeny of Archaea inferred with the L1 supermatrix (122 species, 35,589 amino acid positions). This provisional tree was inferred with FastTree (WAG+Γ8). The scale bars represent the average number of substitutions per site. Numbers at branches represent ML SH-like supports estimated with the L1 and XL1 supermatrices, respectively.

Figure 3. (A) Number of positions considered in each SF-matrix (S_0 to S_{34}). **(B-F)** Evolution of the SH-supports for major archaeal relationships according to the site-by-site desaturation strategy. X and Y axes correspond to the SF-matrices and to the SH-supports, respectively (see legends on the Figure).

Figure 4. BI trees of Crenarchaeota (33 species) **(A)** and Euryarchaeota (79 species) **(B)**, based on the L1-REC6 supermatrix (35,589 positions). The tree was inferred with Phylobayes (CAT+Γ8+Dayhoff6). The scale bars represent the average number of substitutions per site. Numbers at branches are posterior probabilities.

Supplementary material.

Supplementary Table S1. Table showing the taxonomic sampling used for the identification of new markers that can be potentially used to study the phylogeny of Archaea.

Supplementary Table S2. Table showing the 200 new proteins used for the phylogenetic analysis.

Supplementary Table S3. Table showing the taxonomic distribution of the 129 archaeal genomes used for the assembly of final datasets.

Supplementary Figure S1. Unrooted maximum likelihood phylogeny of Archaea based on the 200 newly identified proteins (L2 supermatrix, 48,904 amino acid positions, 122 species). The tree was inferred with PhyML (LG+Γ8). The scale bar represents the average number of substitutions per site. Numbers at branches are SH-like supports

Supplementary Figure S2. Unrooted maximum likelihood phylogeny of Archaea based on the 53 ribosomal proteins (L3 supermatrix, 6,228 amino acid positions, 122 organisms). The tree was inferred with PhyML (LG+Γ8). The scale bar represents the average number of substitutions per site. Numbers at branches are SH-like supports.

Supplementary Figure S3. Unrooted maximum likelihood phylogeny of Archaea based on the 15 subunits of the RNA polymerase and transcription factors (L4 supermatrix, 2,970 amino acid positions, 122 species). The tree was inferred with PhyML (LG+Γ8). The scale bar represents the average number of substitutions per site. Numbers at branches are SH-like supports.

Supplementary Figure S4. Unrooted Bayesian tree of Archaea based on the 15 subunits of the RNA polymerase and transcription factors (L4 supermatrix, 2970 amino acid positions, 122 species). The tree was inferred with PhyloBayes (CAT+Γ8). The scale bar represents the average number of substitutions per site. Numbers at branches are posterior probabilities.

Supplementary Figure S5. Unrooted maximum likelihood trees resulting from the site-by-site desaturation strategy. The S_0 to S_{34} supermatrices were extracted from the XL1 supermatrix (see material and methods). Trees were inferred with PhyML (CAT+Γ8). The scale bar represents the average number of substitutions per site. Numbers at branches are SH supports.

Acknowledgments.

This work was supported by the French National Agency for Research (EVOLDEEP project, contract number ANR-08-GENM-024-002) and by the Investissement d'Avenir grant (ANR-10-BINF-01-01). C.B-A is member of the Institut Universitaire de France. C. Petitjean was the recipient of a grant from the ANR EvolDeep.

Bibliography.

Albers, SV, BH Meyer. 2011. The archaeal cell envelope. Nat Rev Microbiol 9:414-426.

Altschul, SF, TL Madden, AA Schaffer, J Zhang, Z Zhang, W Miller, DJ Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389-3402.

Baker, BJ, LR Comolli, GJ Dick, LJ Hauser, D Hyatt, BD Dill, ML Land, NC Verberkmoes, RL Hettich, JF Banfield. 2010. Enigmatic, ultrasmall, uncultivated Archaea. Proc Natl Acad Sci U S A 107:8806-8811.

Baker, BJ, GW Tyson, RI Webb, J Flanagan, P Hugenholtz, EE Allen, JF Banfield. 2006. Lineages of acidophilic archaea revealed by community genomic analysis. Science 314:1933-1935.

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

Bernander, R. 2000. Chromosome replication, nucleoid segregation and cell division in archaea. Trends Microbiol 8:278-283.

Bodilis, J, S Nsigue-Meilo, L Besaury, L Quillet. 2012. Variable copy number, intra-genomic heterogeneities and lateral transfers of the 16S rRNA gene in Pseudomonas. PLoS One 7:e35647.

Brinkmann, H, H Philippe. 1999. Archaea sister group of Bacteria? Indications from tree reconstruction artifacts in ancient phylogenies. Mol Biol Evol 16:817-825.

Brochier-Armanet, C, B Boussau, S Gribaldo, P Forterre. 2008. Mesophilic Crenarchaeota: proposal for a third archaeal phylum, the Thaumarchaeota. Nat Rev Microbiol 6:245-252.

Brochier-Armanet, C, P Forterre, S Gribaldo. 2011. Phylogeny and evolution of the Archaea: one hundred genomes later. Curr Opin Microbiol 14:274-281.

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

Brochier, C, P Forterre, S Gribaldo. 2005. An emerging phylogenetic core of Archaea: phylogenies of transcription and translation machineries converge following addition of new genome sequences. BMC Evol Biol 5:36.

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

Brochier, C, H Philippe. 2002. A non-hyperthermophilic ancestor for bacteria. Nature 417:244.

Cavicchioli, R. 2011. Archaea--timeline of the third domain. Nat Rev Microbiol 9:51-61.

Criscuolo, A, S Gribaldo. 2010. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. BMC Evol Biol 10:210.

Csuros, M, I Miklos. 2009. Streamlining and large ancestral genomes in Archaea inferred with a phylogenetic birth-and-death model. Mol Biol Evol 26:2087-2095.

Dagan, T, W Martin. 2006. The tree of one percent. Genome Biol 7:118.

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

Elkins, JG, M Podar, DE Graham, et al. 2008. A korarchaeal genome reveals insights into the evolution of the Archaea. Proc Natl Acad Sci U S A 105:8102-8107.

Eme, L, LJ Reigstad, A Spang, A Lanzen, T Weinmaier, T Rattei, C Schleper, C Brochier-Armanet. 2013. Metagenomics of Kamchatkan hot spring filaments reveal two new major (hyper)thermophilic lineages related to Thaumarchaeota. Res Microbiol 164:425-438.

Felsenstein, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst Zool 27:401-410.

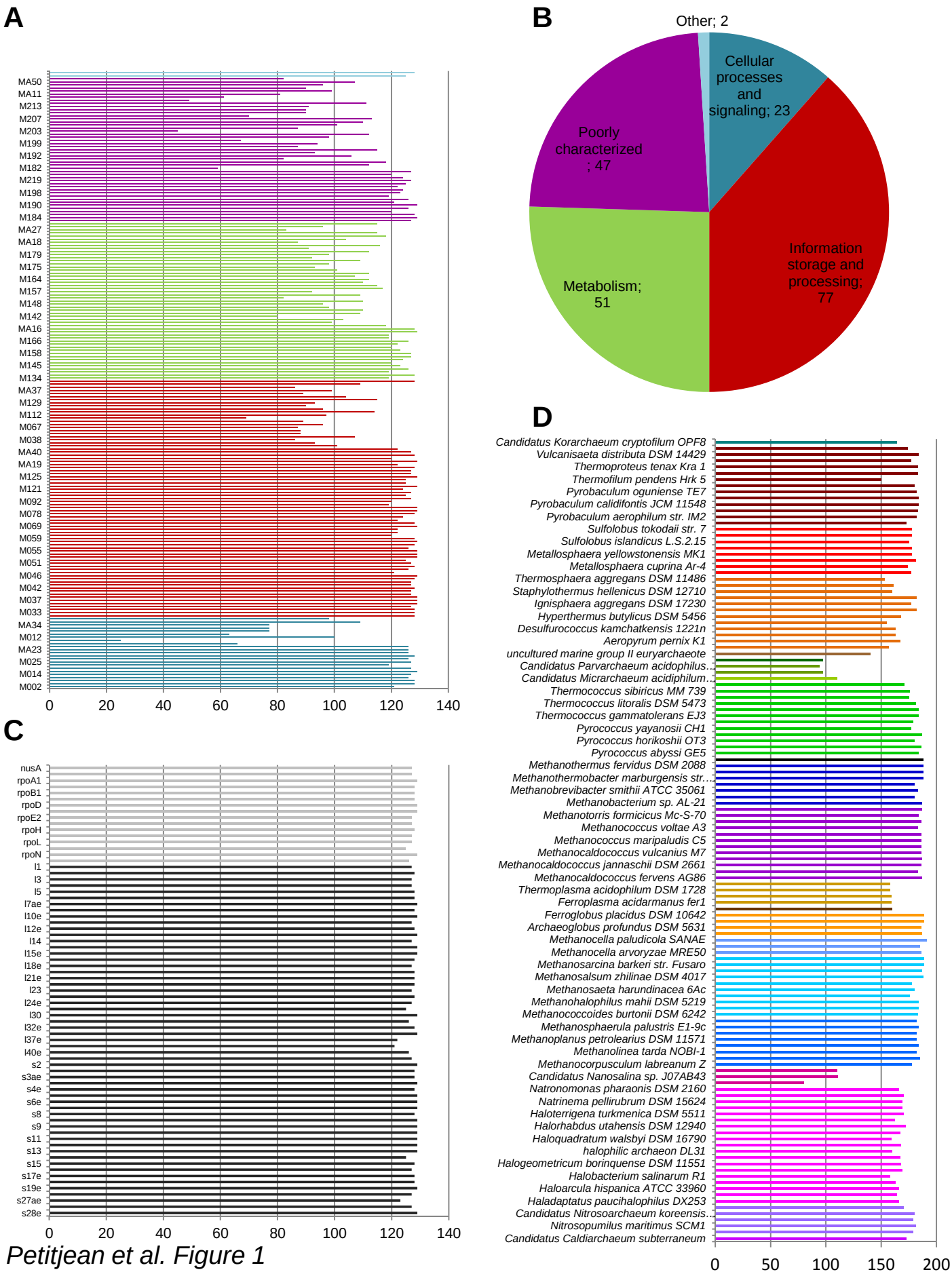
French, SL, TJ Santangelo, AL Beyer, JN Reeve. 2007. Transcription and translation are coupled in Archaea. Mol Biol Evol 24:893-895.

730 Galtier, N, JR Lobry. 1997. Relationships between genomic G+C content, RNA secondary structures, and
731 optimal growth temperature in prokaryotes. J Mol Evol 44:632-636.
732 Gribaldo, S, C Brochier-Armanet. 2006. The origin and evolution of Archaea: a state of the art. Philos
733 Trans R Soc Lond B Biol Sci 361:1007-1022.
734 Gribaldo, S, C Brochier. 2009. Phylogeny of prokaryotes: does it exist and why should we care? Res
735 Microbiol 160:513-521.
736 Gribaldo, S, H Philippe. 2002. Ancient phylogenetic relationships. Theor Popul Biol 61:391-408.
737 Groussin, M, B Boussau, M Gouy. 2013. A Branch-Heterogeneous Model of Protein Evolution for Efficient
738 Inference of Ancestral Sequences. Syst Biol.
739 Groussin, M, M Gouy. 2011. Adaptation to environmental temperature is a major determinant of molecular
740 evolutionary rates in archaea. Mol Biol Evol 28:2661-2674.
741 Guindon, S, JF Dufayard, V Lefort, M Anisimova, W Hordijk, O Gascuel. 2010. New algorithms and
742 methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0.
743 Syst Biol 59:307-321.
744 Guy, L, TJ Ettema. 2011. The archaeal 'TACK' superphylum and the origin of eukaryotes. Trends
745 Microbiol 19:580-587.
746 Hallam, SJ, KT Konstantinidis, N Putnam, C Schleper, Y Watanabe, J Sugahara, C Preston, J de la Torre,
747 PM Richardson, EF DeLong. 2006. Genomic analysis of the uncultivated marine crenarchaeote
748 Cenarchaeum symbiosum. Proc Natl Acad Sci U S A 103:18296-18301.
749 Huber, H, MJ Hohn, R Rachel, T Fuchs, VC Wimmer, KO Stetter. 2002. A new phylum of Archaea
750 represented by a nanosized hyperthermophilic symbiont. Nature 417:63-67.
751 Katoh, K, DM Standley. 2013. MAFFT multiple sequence alignment software version 7: improvements in
752 performance and usability. Mol Biol Evol 30:772-780.
753 Kelly, S, B Wickstead, K Gull. 2010. Archaeal phylogenomics provides evidence in support of a
754 methanogenic origin of the Archaea and a thaumarchaeal origin for the eukaryotes. Proc Biol Sci.
755 Kitahara, K, K Miyazaki. 2013. Revisiting bacterial phylogeny: Natural and experimental evidence for
756 horizontal gene transfer of 16S rRNA. Mob Genet Elements 3:e24210.
757 Kostka, M, M Uzlikova, I Cepicka, J Flegr. 2008. SlowFaster, a user-friendly program for slow-fast analysis
758 and its application on phylogeny of Blastocystis. BMC Bioinformatics 9:341.
759 Lartillot, N, H Brinkmann, H Philippe. 2007. Suppression of long-branch attraction artefacts in the animal
760 phylogeny using a site-heterogeneous model. BMC Evol Biol 7 Suppl 1:S4.
761 Lartillot, N, H Philippe. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid
762 replacement process. Mol Biol Evol 21:1095-1109.
763 Le, SQ, O Gascuel. 2008. An improved general amino acid replacement matrix. Mol Biol Evol 25:1307-
764 1320.
765 Le, SQ, N Lartillot, O Gascuel. 2008. Phylogenetic mixture models for proteins. Philos Trans R Soc Lond
766 B Biol Sci 363:3965-3976.
767 Lopez-Garcia, P, D Moreira. 2008. Tracking microbial biodiversity through molecular and genomic
768 ecology. Res Microbiol 159:67-73.
769 Makarova, KS, AV Sorokin, PS Novichkov, YI Wolf, EV Koonin. 2007. Clusters of orthologous genes for
770 41 archaeal genomes and implications for evolutionary genomics of archaea. Biol Direct 2:33.
771 Mardanov, AV, VA Svetlitchnyi, AV Beletsky, MI Prokofeva, EA Bonch-Osmolovskaya, NV Ravin, KG
772 Skryabin. 2010. The genome sequence of the crenarchaeon Acidilobus saccharovorans supports
773 a new order, Acidilobales, and suggests an important ecological role in terrestrial acidic hot
774 springs. Appl Environ Microbiol 76:5652-5657.
775 Matte-Tailliez, O, C Brochier, P Forterre, H Philippe. 2002. Archaeal phylogeny based on ribosomal
776 proteins. Mol Biol Evol 19:631-639.
777 Narasingarao, P, S Podell, JA Ugalde, C Brochier-Armanet, JB Emerson, JJ Brocks, KB Heidelberg, JF
778 Banfield, EE Allen. 2012. De novo metagenomic assembly reveals abundant novel major lineage
779 of Archaea in hypersaline microbial communities. ISME J 6:81-93.
780 Nunoura, T, Y Takaki, J Kakuta, et al. 2011. Insights into the evolution of Archaea and eukaryotic protein
781 modifier systems revealed by the genome of a novel archaeal group. Nucleic Acids Res 39:3204-
782 3223.
783 Philippe, H, A Germot. 2000. Phylogeny of eukaryotes based on ribosomal RNA: long-branch attraction
784 and models of sequence evolution. Mol Biol Evol 17:830-834.
785 Philippe, H, J Laurent. 1998. How good are deep phylogenetic trees? Curr Opin Genet Dev 8:616-623.

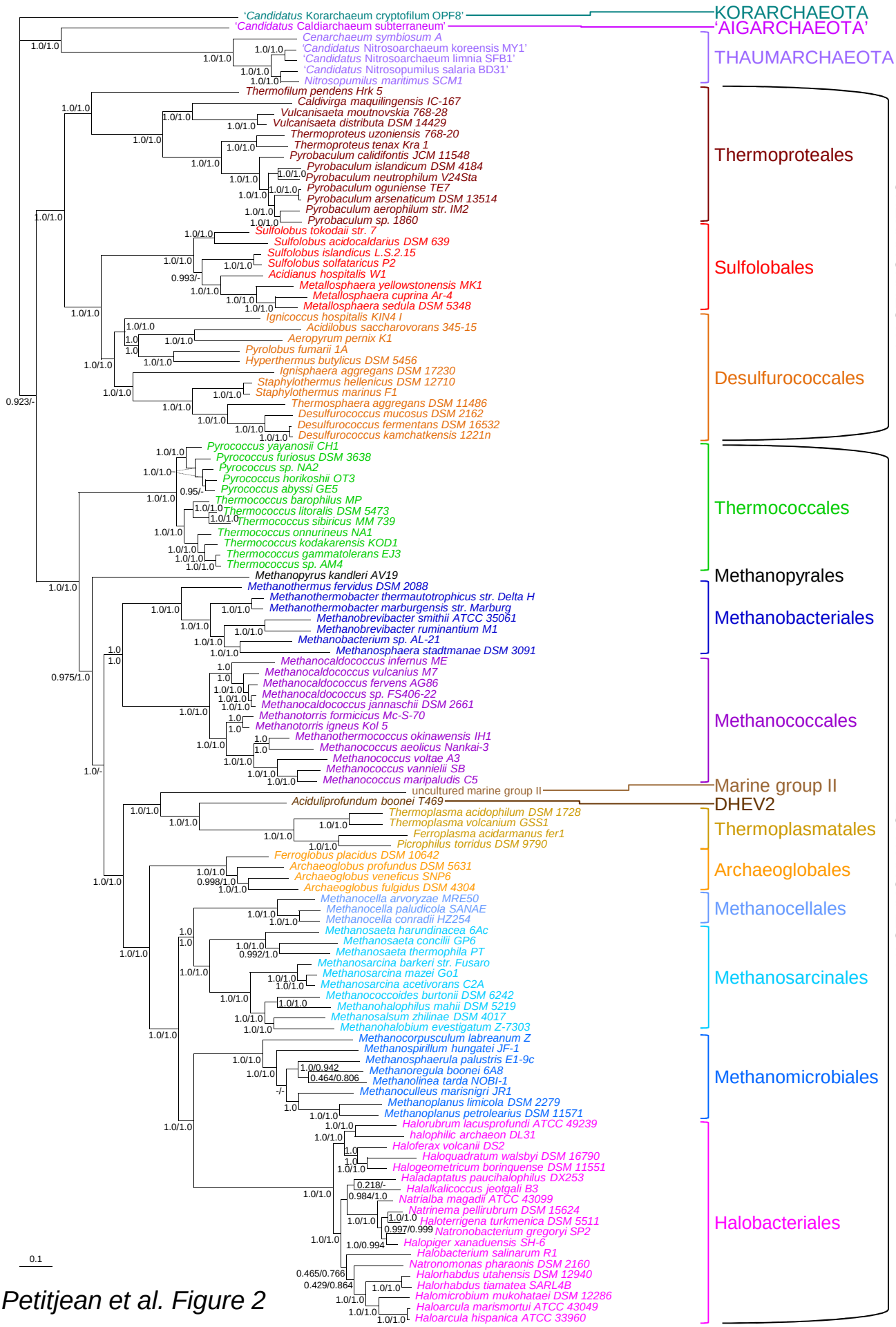
786 Philippe, H, P Lopez, H Brinkmann, K Budin, A Germot, J Laurent, D Moreira, M Muller, H Le Guyader.
787 2000. Early-branching or fast-evolving eukaryotes? An answer based on slowly evolving positions.
788 Proc Biol Sci 267:1213-1221.
789 Podar, M, KS Makarova, DE Graham, YI Wolf, EV Koonin, AL Reysenbach. 2013. Insights into archaeal
790 evolution and symbiosis from the genomes of a nanoarchaeon and its inferred crenarchaeal host
791 from Obsidian Pool, Yellowstone National Park. Biol Direct 8:9.
792 Price, MN, PS Dehal, AP Arkin. 2010. FastTree 2--approximately maximum-likelihood trees for large
793 alignments. PLoS One 5:e9490.
794 Prokofeva, MI, NA Kostrikina, TV Kolganova, TP Tourova, AM Lysenko, AV Lebedinsky, EA Bonch-
795 Osmolovskaya. 2009. Isolation of the anaerobic thermoacidophilic crenarchaeote Acidilobus
796 saccharovorans sp. nov. and proposal of Acidilobales ord. nov., including Acidilobaceae fam. nov.
797 and Caldisphaeraceae fam. nov. Int J Syst Evol Microbiol 59:3116-3122.
798 Rinke, C, P Schwientek, A Sczyrba, et al. 2013. Insights into the phylogeny and coding potential of
799 microbial dark matter. Nature.
800 Robertson, CE, JK Harris, JR Spear, NR Pace. 2005. Phylogenetic diversity and ecology of environmental
801 Archaea. Curr Opin Microbiol 8:638-642.
802 Rodriguez-Ezpeleta, N, H Brinkmann, G Burger, AJ Roger, MW Gray, H Philippe, BF Lang. 2007. Toward
803 resolving the eukaryotic tree: the phylogenetic positions of jakobids and cercozoans. Curr Biol
804 17:1420-1425.
805 Roger, AJ. 1999. Reconstructing early events in eukaryotic evolution. Am. Nat. 154:S146-S163.
806 Roure, B, D Baurain, H Philippe. 2013. Impact of missing data on phylogenies inferred from empirical
807 phylogenomic data sets. Mol Biol Evol 30:197-214.
808 Stiller, J, B Hall. 1999. Long-branch attraction and the rDNA model of early eukaryotic evolution. Mol Biol
809 Evol 16:1270-1279.
810 Walker, CB, JR de la Torre, MG Klotz, et al. 2010. Nitrosopumilus maritimus genome reveals unique
811 mechanisms for nitrification and autotrophy in globally distributed marine crenarchaea. Proc Natl
812 Acad Sci U S A 107:8818-8823.
813 Waters, E, MJ Hohn, I Ahel, et al. 2003. The genome of Nanoarchaeum equitans: insights into early
814 archaeal evolution and derived parasitism. Proc Natl Acad Sci U S A 100:12984-12988.
815 Woese, CR, L Achenbach, P Rouviere, L Mandelco. 1991. Archaeal phylogeny: reexamination of the
816 phylogenetic position of Archaeoglobus fulgidus in light of certain composition-induced artifacts.
817 Syst Appl Microbiol 14:364-371.
818 Woese, CR, GE Fox. 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc.
819 Natl. Acad. Sci. USA 74:5088-5090.
820 Wolf, YI, KS Makarova, N Yutin, EV Koonin. 2012. Updated clusters of orthologous genes for Archaea: a
821 complex ancestor of the Archaea and the byways of horizontal gene transfer. Biol Direct 7:46.
822 Wong, JT, J Chen, WK Mat, SK Ng, H Xue. 2007. Polyphasic evidence delineating the root of life and
823 roots of biological domains. Gene 403:39-52.
824 Wu, D, P Hugenholtz, K Mavromatis, et al. 2009. A phylogeny-driven genomic encyclopaedia of Bacteria
825 and Archaea. Nature 462:1056-1060.
826 Xue, H, SK Ng, KL Tong, JT Wong. 2005. Congruence of evidence for a Methanopyrus-proximal root of
827 life based on transfer RNA and aminoacyl-tRNA synthetase genes. Gene 360:120-130.
828 Yap, WH, Z Zhang, Y Wang. 1999. Distinct types of rRNA operons exist in the genome of the
829 actinomycete Thermomonospora chromogena and evidence for horizontal transfer of an entire
830 rRNA operon. J Bacteriol 181:5201-5209.
831 Yu, Z, K Takai, A Slesarev, H Xue, JT Wong. 2009. Search for primitive Methanopyrus based on genetic
832 distance between Val- and Ile-tRNA synthetases. J Mol Evol 69:386-394.
833 Yutin, N, P Puigbò, EV Koonin, YI Wolf. 2012. Phylogenomics of Prokaryotic Ribosomal Proteins. PLoS
834 One 7:e36972.

835

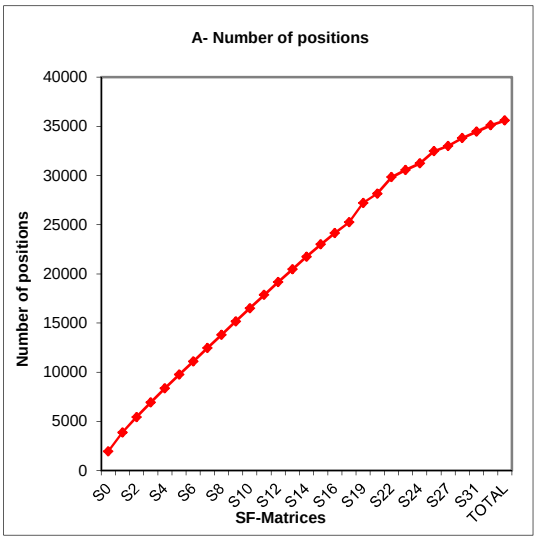
836

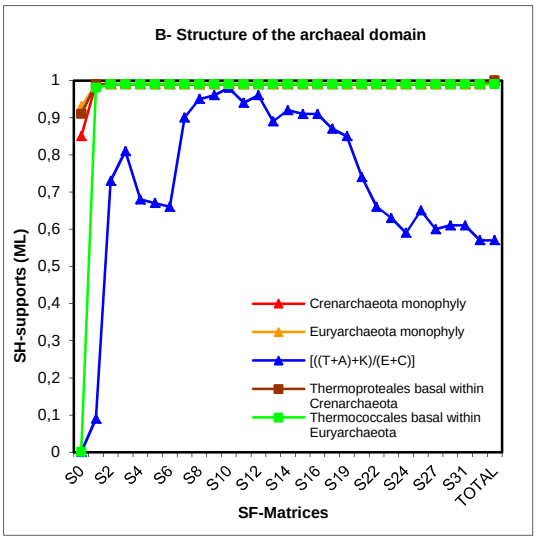


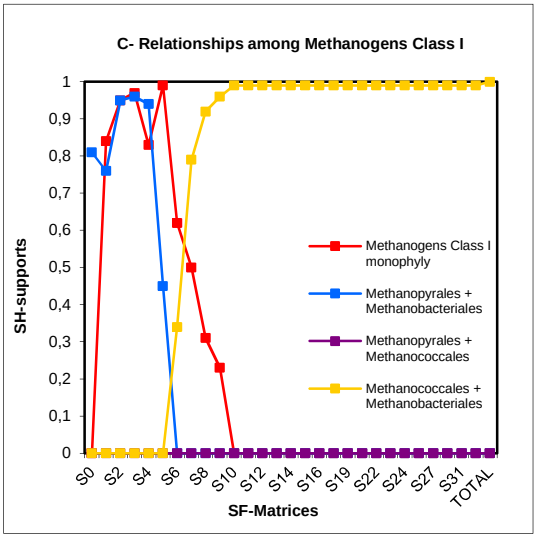
Petitjean et al. Figure 1

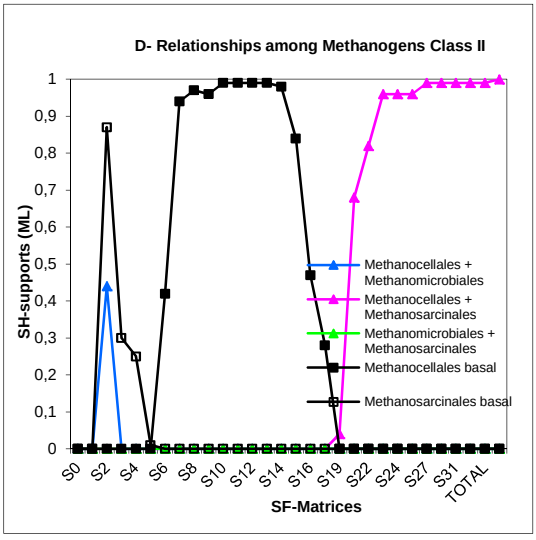


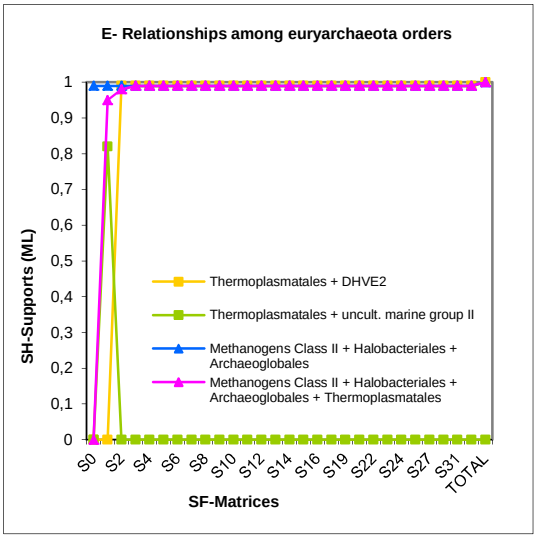
Petitjean et al. Figure 2

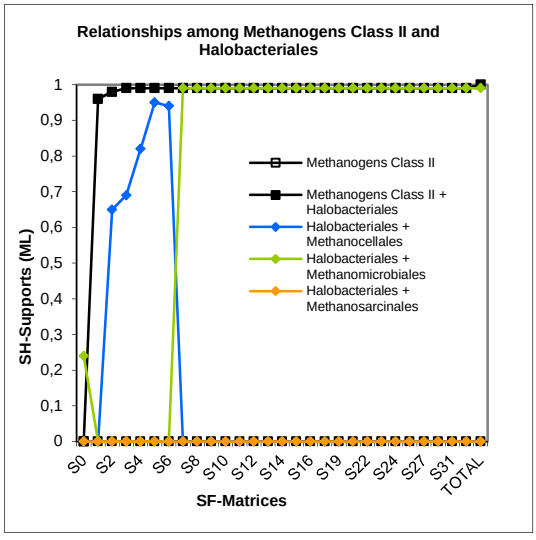




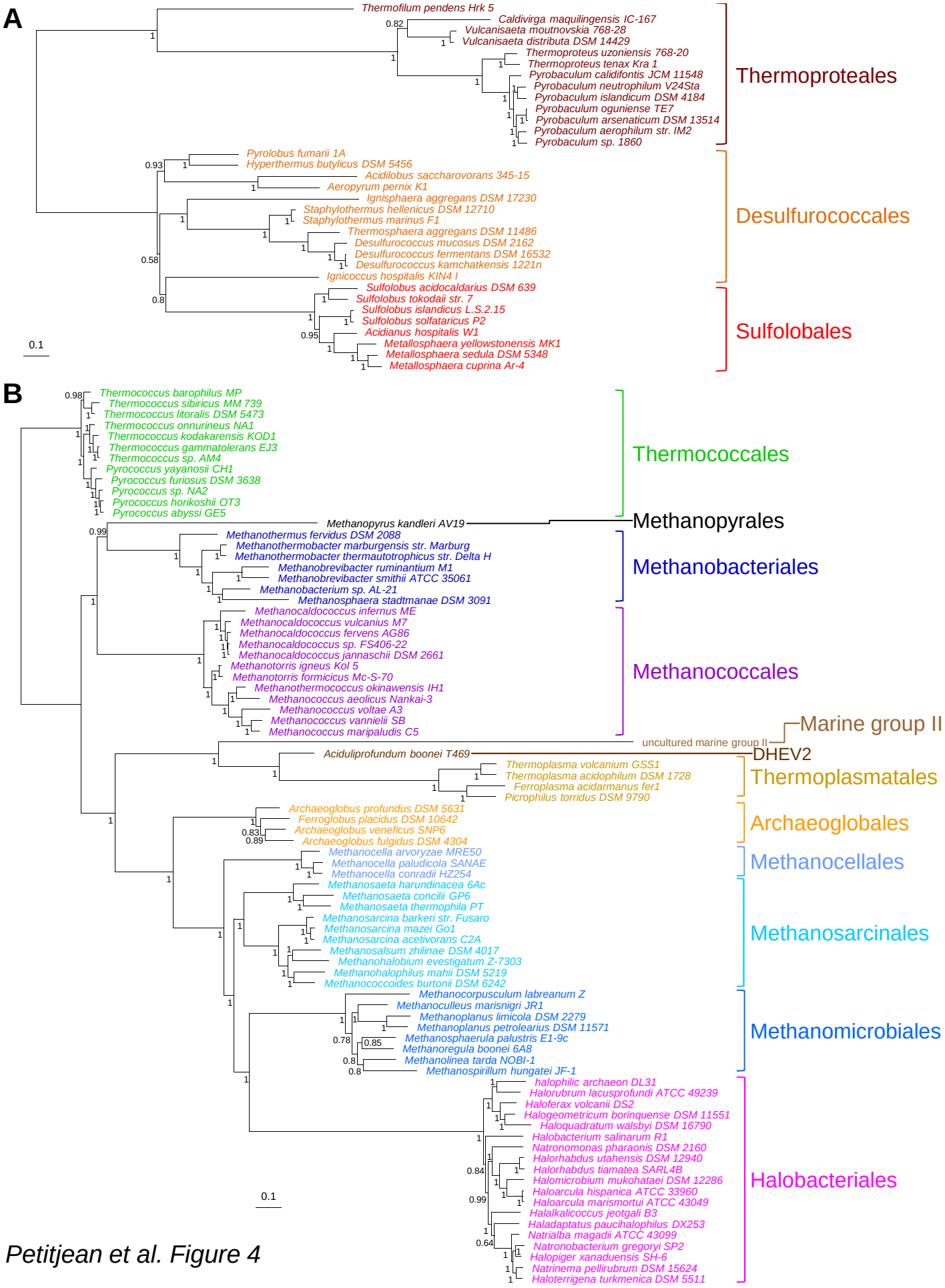








Petitjean et al. Figure 3



Petitjean et al. Figure 4

C. Synthèse et éléments de discussion

Les phylogénies obtenues en maximum de vraisemblance à partir des trois supermatrices constituées des protéines ribosomiques, du système de transcription et des 200 nouveaux marqueurs ont des topologies congruentes dans l'ensemble, ce qui est la preuve d'un signal phylogénétique consistant dans ces différents jeux de données. On retrouve la monophylie des différents grands groupes, comme les Euryarchaeota, Crenarchaeota et Thaumarchaeota, comme celle de la plupart des ordres de crenarchées et d'euryarchées. De plus, ces phylogénies montrent des supports aux nœuds assez élevés. Une différence subsiste en ce qui concerne la position de *M. kandleri*, à la base des euryarchées dans l'arbre construit à partir des protéines intervenant dans la transcription, alors qu'il se place au sein des euryarchées dans les autres phylogénies. Cette incongruence est attendue étant données les analyses précédentes qui montraient une attraction de longues branches dans le système de transcription pour cette espèce (Brochier, Forterre, and Gribaldo 2004). Ces résultats sont donc tout à fait en accord avec les résultats des études précédentes, et plus particulièrement avec la phylogénie de Brochier-Armanet et collaborateurs de 2011 (Brochier-Armanet, Forterre, and Gribaldo 2011). Le signal global de ces trois jeux de données n'étant pas contradictoire, nous avons construit deux supermatrices à partir de l'ensemble de ces 273 jeux de données, une avec la totalité des jeux de données et une avec seulement 179 jeux de données contenant moins de 10 espèces manquantes. Les topologies obtenues étaient, comme attendu, en accord avec les résultats précédents. Nous avons alors mené une analyse de ces jeux de données par désaturation par sélection de sites évoluant de plus en plus rapidement pour estimer l'impact de la vitesse d'évolution des sites sur les relations entre archées. Cette analyse a été conduite sur la supermatrice contenant les 179 jeux de données avec peu de données manquantes pour ne pas biaiser la répartition taxonomique dans les différentes supermatrices résultantes et sans les lignées d'archées de tailles nanométrique dont le cas a été traité par d'autres analyses. Les phylogénies obtenues montrent un signal très fort pour la monophylie des Euryarchaeota et des Crenarchaeota, et pour un regroupement des Thaumarchaeota, Aigarchaeota et Korarchaeota, bien que l'ordre de divergence entre ces phyla ne puisse pas être résolu en l'absence de groupe extérieur. Il est aussi intéressant de noter que les premiers ordres divergeant chez les Euryarchaeota et chez les Crenarchaeota sont, respectivement, les *Thermococcales* et les *Thermoproteales*.

Des supermatrices ont ensuite été construites contenant uniquement les Crenarchaeota et les Euryarchaeota afin d'analyser plus précisément les relations au sein de ces deux phyla. Les résultats obtenus correspondent à ce qui était attendu d'après les résultats des analyses sur la totalité des archées. En ce qui concerne les Crenarchaeota, la monophylie des *Thermoproteales* et des

Sulfolobales est retrouvée, mais la monophylie du genre *Sulfolobus* ne l'est pas. Dans ces analyses, la monophylie des *Desulfurococcales* n'est pas retrouvée car *I. hospitalis* (l'hôte de *N. equitans*) se place à la base des *Sulfolobales*. Cette position peut être le signe d'une paraphylie de ce groupe, indiquant que l'ordre des *Desulfurococcales* n'existerait pas, ou d'un artefact dû à une attraction de longues branches ou à des transferts horizontaux non détectés lors des analyses préliminaires. Les résultats préliminaires de la désaturation par sélection de gènes montrent la monophylie des *Desulfurococcales* dans les premières matrices, i.e. rassemblant les gènes évoluant le plus lentement. La paraphylie de ce groupe pourrait donc être plutôt le résultat d'un biais de type attraction de longues branches qu'une réalité biologique. L'espèce *A. saccharovorans*, à partir de laquelle a été proposé récemment l'ordre des *Acidilobales* (Prokofeva et al. 2009), se place en groupe frère d'*A. pernix* et au sein des *Desulfurococcales* avec des supports maximaux, ce qui tendrait à prouver définitivement que cette espèce ne représente pas un nouvel ordre de crenarchées. De même que pour les Crenarchaeota, on retrouve la monophylie des ordres d'Euryarchaeota. En ce qui concerne les points non résolus de la phylogénie de ce phylum, les méthanogènes classe I (*Methanobacteriales*, *Methanococcales* et *Methanopyrales*) sont monophylétiques et avec un support maximal, ce qui n'était pas établi jusqu'à présent. De plus, les *Methanobacteriales* et les *Methanopyrales* sont groupes frères avec un très fort support, et ces résultats sont confirmés par les analyses de désaturation par sélection de sites, sur les matrices contenant les sites évoluant le plus lentement. Nous proposons donc une redéfinition taxonomique de ce groupe en lui assignant le nom de « Methanohoma », du grec « ὁμάς » pour « groupe ». On retrouve également la monophylie du groupe Thermoplasmatales – DHVE2 - groupe II d'euryarchées, les deux premiers groupes semblant être plus poches entre eux. Néanmoins, un plus grand nombre de séquences génomiques des groupes DHVE2 et groupe II sera nécessaire pour confirmer ces relations. Un autre point non résolu concerne les relations entre les méthanogènes classe II (*Methanomicrobiales*, *Methanosarcinales* et *Methanocellales*) et les *Halobacteriales*. Nous observons ici que ces quatre ordres forment un groupe monophylétique et bien soutenu dans la plupart des analyses. Par contre, les relations au sein de ce groupe ne sont pas bien résolues. D'après la phylogénie obtenue sans désaturation, les *Halobacteriales* et les *Methanomicrobiales* formeraient un groupe monophylétique, alors que les analyses de désaturation par sélection de sites sur les sites évoluant lentement sont en faveur d'un regroupement entre *Halobacteriales* et *Methanocellales*.

La question de la position phylogénétique des archées de petite taille est un problème majeur. La plupart des phylogénies inférées ces dernières années les placent dans un groupe à la base de l'arbre des archées ou à la base des euryarchées, mais sur de très longues branches. Leur évolution rapide accompagnée d'une réduction de la taille du génome explique facilement pourquoi

une attraction de longues branches peut être la cause de cette position basale et groupée. Néanmoins, cet artefact s'est révélé être très fort et ce groupe reste très difficile à casser pour positionner ces espèces au sein des archées. Des analyses de désaturation par sélection de gènes avec différentes sélections d'espèces (*i.e.* introduction indépendante de chaque espèce à évolution rapide : *N. equitans*, les deux espèces du genre '*Ca.* Parvarchaeum', '*Ca.* Micrarchaeum acidiphilum', et les trois espèces de l'ordre Nanohaloarchaea) sont en cours. Cette approche avait déjà montré que *N. equitans* est probablement groupe frère des Thermococcales (Brochier et al. 2005) et nos résultats préliminaires confirment que les *Nanohaloarchaea* sont groupe frère des *Halobacteriales* et que les deux groupes d'ARMAN ('*Ca.* Parvarchaeum' et '*Ca.* Micrarchaeum') pourraient se placer au sein des euryarchées mais séparément. Ces analyses tendent à confirmer que le regroupement de ces sept espèces est artéfactuel dans certaines analyses phylogénétiques, même s'il reste très difficile de les repositionner avec certitude dans la phylogénie des archées. Le fait que ces archées nanométriques aux génomes particuliers ne soient très probablement pas des groupes basaux confirmerait aussi le fait que l'ancêtre des archées n'aurait pas ces caractéristiques génomiques, mais serait au contraire un organisme complexe.

En conclusion, cette analyse montre que le signal phylogénétique porté par les protéines ribosomiques et les sous-unités de l'ARN polymérase n'est pas le résultat d'une évolution particulière de ce système puisqu'on retrouve le même signal avec un grand nombre de marqueurs impliqués dans de nombreux processus cellulaires différents et non liés fonctionnellement. Néanmoins, l'ajout de 200 nouveaux marqueurs permet une amélioration sensible de la reconstruction de la phylogénie des archées en termes de soutien et permet d'apporter un nouvel éclairage sur certains points problématiques. Bien que ces nouveaux marqueurs ne soient pas présents dans la totalité des génomes d'archées, il est possible d'en extraire un signal phylogénétique fiable. Il semble donc qu'un génome « cœur » (« core genome » en anglais) existe et soit hérité majoritairement de manière verticale chez les archées, même s'il peut y avoir des pertes différentielles et HGT ponctuels de certaines protéines dans les différentes lignées. Cette idée avait été proposée par Gribaldo et Brochier-Armanet en 2006 (Simonetta Gribaldo and Brochier-Armanet 2006) et semble confirmée par cette étude. Il est important de souligner que nous avons pris pour point de départ des génomes de thaumarchées et de l'aigarchée, il est donc possible que d'autres marqueurs intéressants soient mis en évidence par l'analyse de génomes d'autres phyla d'archées, et en particulier des lignées non-cultivées à l'heure actuelle et qui constituent une grande part de la diversité du troisième domaine du vivant.

Chapitre 2 : Positionnement de la racine de l'arbre des Archaea.

A. Introduction

La position de la racine de l'arbre des archées est un sujet essentiel non seulement pour la phylogénie des archées mais pour celle du vivant dans son ensemble. En effet, résoudre les relations de parenté entre les phyla d'archées est important pour comprendre l'évolution de ce domaine, mais aussi ses relations avec les deux autres domaines du vivant, bactéries et eucaryotes. En 1990, la première divergence chez les archées a été placée entre Crenarchaeota et Euryarchaeota, définis comme deux phyla par Woese et collaborateurs sur la base des premières phylogénies d'ARNr SSU (C R Woese, Kandler, and Wheelis 1990) (Figure 4). La découverte successive de plusieurs lignées se plaçant près de la base de l'arbre des archées a peu à peu remis en cause cette dichotomie. Ces nouveaux phyla, les Korarchaeota (Barns et al. 1996), les Nanoarchaeota (H. Huber et al. 2002), les Thaumarchaeota (Brochier-Armanet et al. 2008) et les Aigarchaeota (Nunoura et al. 2011), ont multiplié les positions possibles de la racine de l'arbre. Par exemple, les premières phylogénies basées sur l'ARNr SSU et les protéines ribosomiques comprenant la nanoarchée *Nanoarchaeum equitans* plaçaient la racine entre cette espèce et le reste des archées (Waters et al. 2003). De la même manière, les analyses phylogénomiques du génome de la première thaumarchée séquencée, *Cenarchaeum symbiosum*, racinaient l'arbre des archées entre cette espèce et le reste, en utilisant les eucaryotes comme groupe extérieur (Brochier-Armanet et al. 2008). D'autres études phylogénomiques concernant les relations entre les trois domaines du vivant ont été conduites ces dernières années et la racine de l'arbre des archées a été proposée successivement comme étant à la base de chacun des phyla (cf. Introduction B.2.b). Ces analyses ont été conduites avec différentes méthodes et échantillonnages taxonomiques, et avec différents groupes extérieurs (bactéries et/ou eucaryotes), les rendant difficilement comparables. *A contrario*, les marqueurs utilisés sont souvent centrés sur les protéines ribosomiques et, par conséquent, les jeux de données sont assez proches en termes de composition protéique même s'ils ne sont pas analysés de la même façon. Enfin, pour la plupart ces études ont pour but de résoudre l'ensemble des relations entre les trois domaines du vivant, mais ne sont pas centrées sur la position de la racine de l'arbre des archées. D'autre part, une grande partie des phylogénies des archées ont été inférées sans groupe extérieur, ce qui est très important pour pouvoir résoudre la phylogénie interne des archées, mais ne permet pas de raciner l'arbre (Brochier-Armanet, Forterre, and Gribaldo 2011). Nous pensons que les relations anciennes entre les archées doivent être traitées pour elles mêmes, avec un set de marqueurs adapté ainsi qu'un échantillonnage taxonomique représentatif de leur diversité et un choix judicieux du groupe

extérieur.

Nous présentons ici une analyse de la racine de la phylogénie des archées basée sur 70 marqueurs dont 38 ne sont pas des protéines ribosomiques, utilisant des séquences homologues bactériennes comme groupe extérieur. A partir des 200 nouveaux marqueurs définis préalablement pour l'analyse de la phylogénie des archées présentée dans le Chapitre 1, nous avons trouvé 81 protéines ayant des homologues bactériens. Nous avons construit des jeux de données contenant les séquences d'archées et les séquences bactériennes homologues issues de 117 génomes représentatifs de la diversité bactérienne. La phylogénie de chaque protéine a été inférée et analysée afin d'exclure les jeux de données pour lesquels l'homologie des séquences bactériennes était le résultat de transferts horizontaux. 38 jeux de données ont été conservés. Le même travail a été fait à partir des protéines ribosomiques, et 32 ont été conservées. A partir de ces deux groupes de protéines, nous avons tout d'abord construit deux supermatrices différentes, afin d'analyser leur signal indépendamment, puis une troisième supermatrice contenant l'ensemble des 70 jeux de données. Une attention particulière a été donnée aux espèces d'archées de taille nanométrique des genres *Nanoarchaeum*, '*Ca.* Micrarchaeum', '*Ca.* Parvarchaeum' et les *Nanohaloarchaea*, connues pour être sujets aux LBA, ce qui est particulièrement problématique dans une analyse comprenant un groupe extérieur. Enfin, nous avons utilisés deux méthodes de désaturation de notre jeux de données, par gène et par site, afin d'analyser l'impact des différences de vitesse d'évolution sur la position de la racine des archées.

Ce travail est présenté dans l'article « *Phylogenomic Analysis Pinpoints the Root of the Domain Archaea and Supports the Foundation of the New Kingdom Proteoarchaeota* » (Petitjean, Deschamps, Brochier-Armanet, López-García, Moreira, en préparation).

B. Manuscrit de l'article 2 : « Phylogenomic Analysis Pinpoints the Root of the Domain Archaea and Supports the Foundation of the New Kingdom Proteoarchaeota »

Phylogenomic Analysis Pinpoints the Root of the Domain Archaea and Supports the Foundation of the New Kingdom Proteoarchaeota

Céline Petitjean[1], Philippe Deschamps[1], Céline Brochier-Armanet[2], Purificación López-García[1], and David Moreira[1,*]

[1]Unité d'Ecologie, Systématique et Evolution, CNRS UMR 8079, Université Paris-Sud, 91405 Orsay Cedex, France

[2]Université de Lyon, Université Lyon 1, CNRS, UMR5558, Laboratoire de Biométrie et Biologie Evolutive, 43 boulevard du 11 novembre 1918, F-69622 Villeurbanne, France. Tel.: +33 4 26 23 44 76; fax: +33 4 72 43 13 88

*Author for Correspondence: David Moreira, Unité d'Ecologie, Systématique et Evolution, CNRS UMR 8079, Université Paris-Sud, 91405 Orsay Cedex, France, +33 1 69157608, +33 1 69154697, david.moreira@u-psud.fr

Data deposition: All sequence alignments used in this work are available upon request to the corresponding author.

Abstract

The first phylogenetic trees of the domain Archaea, based on the analysis of 16S rRNA sequences, showed a deep division between two major groups, which were classified as the kingdoms Euryarchaeota and Crenarchaeota. This bipartite view has been challenged in recent years thanks to the discovery of several deeply-branching new archaeal lineages, in particular the Thaumarchaeota, Aigarchaeota, Nanoarchaeota, and Korarchaeota, which have also been given the same taxonomic status of kingdoms. However, the phylogenetic position of some of these lineages has been controversial. In this sense, one problem has been that very often the phylogenetic analyses of the Archaea has been carried out without outgroup, making very difficult to determine if these taxa actually represent independent archaeal lineages of the same level as the Euryarchaeota and Crenarchaeota. We have addressed this question by reconstructing archaeal phylogenetic trees rooted on bacterial sequences. These trees were based on conserved protein markers commonly used (32 ribosomal proteins) and on 38 new markers identified through phylogenomic analysis. We thus gathered a total of 70 markers that we analyzed as a concatenated dataset. Our trees consistently placed the root of the archaeal tree between the Euryarchaeota (including the Nanoarchaeota and other fast-evolving lineages) and the rest of archaeal species. We thus propose the erection of a new kingdom, the Proteoarchaeota, to contain these lineages. This has to be accompanied by the relegation to a lower taxonomic rank (class or order) of the lineages previously classified as kingdoms (Crenarchaeota, Thaumarchaeota, Aigarchaeota, and Korarchaeota) and now encompassed within the Proteoarchaeota.

Key words: Archaea, Euryarchaeota, Proteoarchaeota, root, phylogenomics.

Introduction

Despite the fact that several archaeal species had been isolated as soon as in the 30's (Barker, 1936), the recognition of their nature as members of an independent domain of life had to wait for 40 years, when Woese and Fox published the proposal for the division of life into three primary kingdoms: Bacteria, Archaebacteria and Urkaryotes, based on the analysis of small subunit rRNA sequences (SSU rRNA) (Woese and Fox 1977). Later on, to emphasize on the difference between the two major groups of prokaryotes (Bacteria and Archaebacteria), these three kingdoms were reclassified as the domains Bacteria, Archaea and Eucarya (Woese et al., 1990). Within the Archaea, the first SSU rRNA phylogenies supported the separation of two groups, one containing methanogenic species and another formed by thermoacidophilic ones (Fox et al., 1980). This was confirmed by subsequent analyses with a richer taxonomic sampling, leading to the division of the Archaea into two kingdoms: the Crenarchaeota, all hyperthermophilic, and the Euryarchaeota, containing species with a variety of phenotypes (hyperthermophilic, mesophilic, methanogenic, halophilic, etc.) (Woese et al., 1990). However, a number of discoveries have challenged this simple bipartite view in recent years.

First, environmental SSU rRNA sequence analyses revealed the existence of archaeal species related to the crenarchaeota thriving in mesophilic environments such as the open ocean (DeLong, 1992; Fuhrman et al., 1992), soils (Jurgens et al., 1997), lakes (Schleper et al., 1997) and both cold and hot terrestrial springs (Barns et al., 1996). Phylogenetic analysis of conserved genes involved in translation and some differences in gene content with the classical hyperthermophilic Crenarchaeota led to propose that these archaeal species may define a new archaeal phylum, the Thaumarchaeota, independent from the two phyla recognized until now, the Crenarchaeota and the Euryarchaeota (Brochier-Armanet et al., 2008). More recently, a distant relative of the Thaumarchaeota, the species *Candidatus* 'Caldiarchaeum subterraneum', was proposed to define an additional new phylum -Aigarchaeota- based on its distinct gene content (Nunoura et al., 2011). Similar arguments were used to suggest that the species *Candidatus* ' Korarchaeum cryptofilum' represented the first member of a new phylum Korarchaeota, very distantly related to the Crenarchaeota (Barns et al., 1996; Elkins et al., 2008). An even more divergent case was found with the discovery of the hyperthermophilic *Nanoarchaeum equitans*, a tiny symbiont of the crenarchaeote *Ingnicoccus hospitalis*. The phylogenetic analysis of concatenated ribosomal proteins suggested that this species diverged before the separation of the Crenarchaeota and the Euryarchaeota, thus defining a very ancient new archaeal phylum -the Nanoarchaeota- (Waters et al., 2003). Although being classified within the Euryarchaeota, a very deep branching has also been suggested for other ultrasmall archaea of the genera

Candidatus 'Micrarchaeum' and 'Parvarchaeum', using as argument their unusual gene content (Baker et al., 2010).

Thus, the last decade has seen a multiplication of deep-branching groups within the Archaea and, as a consequence, of possibilities to place the root of the archaeal tree, namely, the first divergence within this domain of Life. The first suggestion placed the root between the Crenarchaeota and the Euryarchaeota (Woese et al., 1990). However, the subsequent inclusion of the new deep-branching groups, in particular the Nanoarchaeota, has challenged this view. For example, the first phylogenetic trees incorporating this lineage, based on rRNA or ribosomal protein sequences, supported a rooting between *N. equitans* and the rest of archaea (Waters et al., 2003). However, other analyses rooted the archaeal tree on the branch leading to the Thaumarchaeota (Brochier et al., 2008). More recently, phylogenetic analyses using heterogeneous sequence evolution models aimed at determining the precise archaeal origin of the eukaryotic nucleocytoplasm have placed the root of the archaeal domain on the *N. equitans* branch or within the Euryarchaeota (Cox et al., 2008). These different analyses applied different tree reconstruction methods, different models of sequence evolution, different archaeal sequence samplings and different outgroup sequences (eukaryotic and/or bacterial), making them difficult to be compared. On the other hand, most archaeal phylogenetic analyses with a wide taxonomic archaeal sampling did not include outgroup sequences, so they only produced unrooted phylogenies (e.g., Brochier-Armanet et al., 2011). Therefore, there is a need for phylogenetic analyses specifically aimed at rooting the archaeal tree. For that, a wide representation of all archaeal phyla as well as a good sampling of outgroup sequences are necessary. In this work, we have used this approach to reconstruct archaeal phylogenetic trees rooted with a bacterial outgroup. These trees were based on the classical ribosomal protein dataset (32 proteins) and on a collection of 38 new conserved proteins identified by phylogenomic analysis. Our phylogenetic analyses supported the rooting between the Euryarchaeota (including the Nanoarchaeota and the genera *Micrarchaeum* and *Parvarchaeum*) and the rest of archaea (Crenarchaeota, Thaumarchaeota, Aigarchaeota and Korarchaeota). This deep division into two major groups prompted us to propose a re-classification of several major archaeal lineages from their current status of phyla or divisions into the rank of classes. This would make the whole archaeal taxonomy much more homogeneous.

Materials and Methods

Detection of Proteins Widespread in Archaea and Bacteria

We gathered the protein sequences coded by all publicly available complete genome sequences of archaeal species as well as those of 117 bacterial species into a local

database (for the list of organisms, see supplementary table S1, Supplementary Material online). We then carried out sequence similarity searches against the archaeal sequences in this database using BLAST (Altschul 1997) and all the protein sequences coded in the genomes of *Nitrosopumilus maritimus*, *Cenarchaeum symbiosum* and *Caldiarchaeum subterraneum* as queries. The BLAST results were filtered to retain only those containing representatives at least 4 archaeal sequences. The 4174 proteins thus identified were submitted to preliminary phylogenetic analysis (see below). The phylogenetic trees obtained were checked manually to retain only those were the different archaeal phyla were monophyletic and without evidence for horizontal gene transfer (HGT) events.

200 datasets were retained, for which we incorporated the respective homologous bacterial sequences and reconstructed new phylogenetic trees, which we used to finally retain 38 proteins for which the trees did not show evidence for HGT between archaea and bacteria (see the list in supplementary table S2, Supplementary Material online). To decrease the calculation time of the subsequent phylogenetic analyses, we selected 27 taxonomically diverse bacteria to be used as outgroup.

Phylogenetic Analyses

Sets of homologous protein sequences were aligned using MAFFT (Katoh 2013). Conserved positions in the alignments were detected using BMGE with default parameters (Criscuolo 2010) and verified by hand using the program NET of the MUST package (Philippe 1993). Maximum likelihood phylogenetic trees were reconstructed upon each individual protein or different concatenated datasets with RaxML v.7.2.4 (Stamatakis 2006) and the PROTGAMMALGF model. Tree robustness was estimated using the Rapid Bootstraping method implemented in RaxML. Bayesian inference analyses were carried out using MrBayes 3.2.1 (Ronquist et al., 2012) with a mixed model of amino acid sequence evolution and a Gamma distribution with six discrete categories to accommodate for among site rate variation.

Mutational Desaturation Analyses

We applied a site-by-site mutational desaturation approach using SlowFaster (Kostka 2008) to construct 10 multiple sequence alignments containing sites with increasing evolutionary rates. We also carried out a marker-by-marker desaturation analysis based on the average evolutionary rates of the different markers analyzed. For that, we selected 22 pairs of species (10 archaea-archaea, 9 archaea-bacteria, and 3 bacteria-bacteria, see supplementary table S3, Supplementary Material online) and we computed the species-to-species distance for each pair based on the ML phylogenetic tree for each individual marker. We then obtained the average value of those distances, which we attributed as the evolutionary rate of the

marker. As for the previous approach, we constructed 10 concatenations, beginning with one containing the slowest evolving markers and progressively adding more and more fast evolving ones. All the different concatenations were analyzed by maximum likelihood using with RaxML v.7.2.4 (Stamatakis 2006).

Results

Identification of New Markers Widely Distributed in Archaea and Bacteria

It has been proposed that the gene content in thaumarchaeotal genomes is to some extent intermediate between the euryarchaeotal and crenarchaeotal ones (Brochier et al., 2008, Spang et al., 2010, Brochier et al., 2012). By this reason we have used two thaumarchaeotal genomes (*Cenarchaeum symbiosum* and *Nitrosopumilus marinus*) and one from the closely related phylum Aigarchaeota (*Caldiarchaeum subterraneum*) as queries to look for conserved genes widely distributed in Archaea. Using BLAST searches and phylogenetic analysis, we detected 200 proteins with orthologs in most of the available archaeal complete genome sequences. Among them, 68 had homologues in bacterial species. We reconstructed maximum likelihood (ML) phylogenetic trees for all these proteins to verify that their presence in both Archaea and Bacteria was not due to horizontal gene transfer (HGT) events between these two domains. We also checked that these proteins were present in at least 26 bacterial phyla to have a sufficiently diverse set of outgroup sequences. This yielded 38 well conserved proteins widely distributed in Archaea and Bacteria (supplementary table S1, Supplementary Material online). We excluded eukaryotes from all our analyses because of the possibility that they might have inherited part of their genes from Archaea during their very early evolution (Cox et al., 2008; Embley and Martin 2006; Gribaldo et al., 2010; López-García and Moreira 1999). If this is actually the case, their use as an outgroup would lead to infer an erroneous rooting of the archaeal tree.

In contrast with the limited range of functions of the markers used until now to reconstruct rooted phylogenies of the Archaea (mostly ribosomal proteins), our newly detected proteins are involved in a variety of cellular processes, including the metabolism of amino acids, nucleotides and co-enzymes, post-translational protein modification and other functions (supplementary table S1, Supplementary Material online). This functional diversity is important because it can contribute to retrieve more homogeneous branch lengths among species because it is unlikely that this entire set of very different cellular processes might have an evolutionary rate in some lineages significantly more accelerated than in the rest. Such very unequal evolutionary rates among lineages, which represent a major problem for

phylogenetic reconstruction, are more likely to occur for small sets of proteins involved in a same process.

In addition to these new protein markers, we also included the classical set of translation proteins (32 ribosomal proteins) to construct concatenated sequence datasets for the subsequent phylogenetic analyses.

Rooted Phylogenetic Analyses of Translation Proteins and the Newly Detected Markers

Multi-marker phylogenetic analyses have revealed that several archaeal lineages have very high evolutionary rates, producing very long branches in phylogenetic trees which can lead to long branch attraction artefacts (LBA). This is notably the case for the ultrasmall archaea of the genera *Nanoarchaeum*, *Micrarchaeum*, *Parvarchaeum* and the Nanohaloarchaea (e.g., Brochier et al., 2011). Therefore, we carried out a first set of phylogenetic analyses excluding all these fast-evolving organisms. The first dataset that we analyzed contained the classical set of ribosomal proteins that have traditionally been used to reconstruct rooted and unrooted archaeal phylogenies (e.g., Cox et al., 2008, Elkins et al., 2008, Brochier et al., 2011, Guy and Ettema 2011). We reconstructed an updated phylogenetic analysis of those markers (32 markers and 2560 amino acid sites) incorporating representative species from all major archaeal lineages (with the exception of the fast-evolving ones, see above) to obtain a reference phylogeny. The resulting phylogeny was rooted within the Euryarchaeota (fig. 1). In fact, different euryarchaeotal lineages emerged as paraphyletic branches at the base of the tree, with strong statistical support in Bayesian trees (posterior probabilities PP=0.97-1) but moderate bootstrap support in maximum likelihood trees (BP<70%). This result was in agreement with a similar tree published by Cox et al., (2008). The apical part of our tree was occupied by a strongly supported group (PP=1, BP=100%) containing the Thaumarchaeota, Aigarchaeota, Crenarchaeota and Korarchaeota (a group tentatively defined as the "TACK" superphylum by Guy and Ettema (2011)).

We then carried out a phylogenetic analysis on a concatenation of our new markers (38 markers and 6890 amino acid sites). The Bayesian phylogenetic tree based on this new dataset (fig. 2) was rooted between the Euryarchaeota and the rest of archaeal species, namely the "TACK" group, in sharp contrast with the previous tree based on translation-related proteins. This rooting was robustly supported (PP=1, BP=99-100%). Moreover, the monophyly of most of the archaeal classes was also well supported (PP=1, BP>90%).

Finally, we analyzed a combined dataset contained all the previous markers (translation-related proteins plus our new markers, for a total of 70 markers and 9450 amino acid sites). The resulting tree (fig. 3) was very similar to the one obtained from the analysis of the concatenation of the new markers. In fact, the root was placed between the Euryarchaeota

and the rest of archaeal species with strong support, especially by the Bayesian analysis (PP=1 for all deep nodes in the tree) and moderately by the ML analysis (BP=79% for the monophyly of the Euryarchaeota and 100% for the "TACK" group). The internal phylogeny of the different archaeal lineages was also well supported in the Bayeasian tree, even for several difficult-to-resolve relationships. For example, we retrieved full support (PP=1, BP=100%) for the monophyly of the so-called "Methanogen Class I" containing the Methanopyrales, Methanobacteriales and Methanococcales, a relationship that was not supported by the ribosomal proteins, which placed the Methanococcales as sister group of the Thermococcales (fig. 1). Another highly supported (PP=1, BP=99%) interesting result was the position of the Halobacteria as a derived group within the class Methanomicrobia (also known as "Methanogen Class II").

The Position of Long-Branching Archaea

As mentioned above, we excluded from our initial phylogenetic analyses a number of taxa characterized by their very long branches. These included *Nanoarchaeum*, *Micrarchaeum*, *Parvarchaeum* and the Nanohaloarchaea. These species exhibited a very large amount of missing data in our concatenated alignments (>50%), in agreement with their very small genome sizes (especially for *N. equitans*, with only 540 protein-coding genes, and *Micrarchaeum* and *Parvarchaeum*, with ~1000 protein-coding genes). This, in addition to their long branches that suggest an accelerated evolutionary rate, made them very prone to potential LBA artefacts (Roure et al., 2013). Moreover, LBA can be exacerbated by the inclusion of distant outgroup sequences, leading to a basal emergence of the long-branching taxa attracted by the outgroup (Philippe and Laurent 1998). In fact, although poorly supported, the ML tree based on our complete set of markers and rooted using the bacterial sequences as outgroup showed a basal emergence of these long-branching taxa within a monophyletic group (BP=62%, supplementary fig. S1, Supplementary Material online). This would support that the root of the archaeal tree is located between these species and the rest of archaea. However, previous analyses have provided strong support for alternative placements of these species, in particular for *Nanoarchaeum* as sister group of the Thermococcales (Brochier et al., 2005) and the Nanohaloarchaea as sister group of the Halobacteria (Narasingarao et al., 2012). We thus tested different ways to reduce the potential LBA responsible of the basal emergence of the long-branching archaea. First, we carried out an unrooted phylogenetic analysis excluding the bacterial sequences. In the resulting ML tree all the ultrasmall archaea branched again as a strongly supported monophyletic group placed between the Crenarchaeota and the Euryarchaeota (supplementary fig. S2, Supplementary Material online). This suggested that the removal of the distant bacterial outgroup sequences was not enough to alleviate the possible LBA

artefact responsible of their basal emergence. We thus re-analyzed the same dataset using a mixed model of amino acid sequence evolution and Bayesian inference, which is known to be very often more robust against LBA than ML methods (Philippe et al., 2011). Indeed, the Bayesian tree showed a very different topology, with the four groups on long branches well nested at different locations within the Euryarchaeota (fig. 4). Nanoarchaeota branched as sister group of the Thermococcales, the genera *Micrarchaeum* and *Parvarchaeum* as sisters of the Thermoplasmata, and the Nanohaloarchaea as sisters of the Halobacteria. These results supported that these taxa can be considered as *bona fide* euryarchaeotal species and that, therefore, the root of the archaeal tree does not lie between them and the other archaea.

Desaturation Analyses

Since the root corresponds to the most ancient node in the archaeal tree, it can be one of the most affected by mutational saturation, namely the loss of phylogenetic signal due to the accumulation of successive mutations on variable sequence sites. To test if this was the case, we have analyzed our global dataset (translation related proteins plus our new markers) using two desaturation strategies based on site selection and marker selection, respectively. For the first one, we calculated the evolutionary rate at each position of the global alignment and constructed 10 concatenations of increasing size. The first contained only the slowest evolving sites and the following a progressively increasing number of more and more fast evolving sites. The ML phylogenetic tree based on the first dataset (with only 939 sites) placed the root within the Euryarchaeota, between the Thermococcales and the rest of archaeal species. However, all the other datasets recovered the same rooting pattern as the complete dataset, namely, the separation between the Euryarchaeota and the other archaea (see supplementary figs. S3-S12, Supplementary Material online). Whereas the monophyly of the "TACK" group was strongly retrieved by all datasets (BP=100% except for the first dataset, with BP=83%), the monophyly of the Euryarchaeota was less supported, especially by the small- and middle-size datasets. This probably reflected the influence of the ribosomal proteins, which tended to support a root within the Euryarchaeota (fig. 1).

The second desaturation strategy took into account the evolutionary rate of each marker to construct a series of 10 concatenations from the slowest evolving to the fastest evolving ones (see Materials and Methods). As in the site-by-site desaturation analysis, all datasets retrieved the monophyly of the "TACK" group with maximal support (BP=100%, supplementary fig. S13-S22, Supplementary Material online). However, the position of the root was unstable, with some of the smallest datasets supporting a root within the Euryarchaeota. This probably reflected the predominance of ribosomal proteins in these datasets. Interestingly, the last five datasets retrieved the root between the Euryarchaeota

and the other archaea, with highest support (BP=94-97%) in the intermediate datasets (49 and 56 proteins). They probably corresponded to the datasets with the optimum number of informative sites before an increase of noise provided by the fastest-evolving proteins, which led to a decrease of the statistical support in the final datasets.

In summary, the two desaturation strategies, site-by-site and gene-by-gene, produced similar results, with most trees supporting the rooting between the Euryarchaeota and the "TACK" group. The global stability of this result suggested that it is not due to any particular group of genes or site category.

Discussion

The Root of the Archaeal Tree

Since the initial proposal by Woese and co-workers that the deepest division in the archaeal domains was between the two kingdoms Crenarchaeota and Euryarchaeota (Woese 1990), many other alternatives have been advanced, most of them linked to the discovery of new deeply-branching archaeal lineages. As summarized in the Introduction, this has been the case especially for the hyperthermophilic *Nanoarchaeum equitans* (Waters et al., 2003, Cox et al., 2008), the Thaumarchaeota (Brochier et al., 2008), and the acidophilic genera *Candidatus* 'Micrarchaeum' and 'Parvarchaeum' (Baker et al., 2010). Di Giulio has repeatedly argued for the rooting on the Nanoarchaeota, to even propose that this group should be considered as a "living fossil" (Di Giulio 2006). This was based not only on phylogenetic analyses of rRNA sequences (Branciamore 2008) but also on the unusual discovery of some genes split in two fragments in the genome of *N. equitans* (Randau 2005), which was interpreted as an ancestral character according to the "introns early" hypothesis (Di Giulio 2008). However, several findings have undermined this proposal. Split genes have also been found in another small archaeon, *Micrarchaeum acidiphilum*, which is phylogenetically unrelated to *N. equitans* (Baket et al., 2010). In addition, phylogenetic analyses based on individual conserved proteins and on concatenated translation-related proteins strongly support that Nanoarchaeota are not basal but sisters to the Thermococcales (Brochier 2005; Brochier 2011). Likewise, a very deep-branching position for the genera *Micrarchaeum* and *Parvarchaeum* based on their gene content with both typical euryarchaeotal and crenarchaeotal genes (Baket et al., 2010) has not been validated by phylogenetic analysis of translation-related proteins, which placed them nested within the Euryarchaeota (Brochier 2011). Finally, the third group of long-branching archaea, the halophilic Nanohaloarchaea, has been shown to robustly branch close to the Halobacteria (Narasingarao et al., 2012). Our analyses with the complete set of markers are in agreement with these results. The unrooted Bayesian tree showed all the long-branching archaea well nested within the Euryarchaeota,

with the Thermococcales + Nanoarchaeota as the first group to diverge (fig. 4). Interestingly, when bacterial sequences were added to root this tree, all those long-branching archaea emerged together at the base of the archaeal tree, strongly suggesting a LBA artefact.

Thus, the atypical characteristics found in several fast-evolving ultrasmall archaea are most likely derived characters rather than ancestral features. One example concerns the split genes of *Nanoarchaeum* and *Micrarchaeum*, which have likely evolved by convergence in these two lineages. Massive genome size reduction has also occurred in parallel in these organisms as well as in *Parvarchaeum*, and, to a lesser extent, in the Nanohaloarchaea, all of them having gene numbers smaller than those of their respective closest relatives. This has been accompanied by a large acceleration of evolutionary rate, as attested by the long branches exhibited by all these species. It may be possible that anomalous characters, such as the split genes, are side products of that evolutionary acceleration and genome reduction. In fact, convergent acquisition of exceptional features has already been noticed in fast-evolving highly reduced genomes. One outstanding example is the migration of the rRNA genes to subtelomeric regions in the chromosomes of the highly reduced nucleomorph genomes of cryptophytes and chlorarachniophytes (Moore and Archibald 2009).

A Bipartite Division of the Domain Archaea and Proposal for the New Archaeal Kingdom Proteoarchaeota

Originally, the domain Archaea was divided into two major groups: the Euryarchaeota and the Crenarchaeota, based on the analysis of 16S rRNA sequences (Woese et al., 1990). These two groups were given the taxonomic rank of Kingdom as the one immediately below that of Domain and to insist on their clear phylogenetic distinctiveness (Woese et al., 1990). However, since the publication of this proposal, several new archaeal lineages have been discovered and some of them have been named using the kingdom suffix 'archaeota' to equal their ranks to that of the Euryarchaeota and Crenarchaeota. This is the case of the Nanoarchaeota (Waters et al., 2003), Thaumarchaeota (Brochier-Armanet et al., 2008), Aigarchaeota (Nunoura et al., 2011), and Korarchaeota (Elkins et al., 2008). The lack of clear criteria to establish a rank for those different lineages has fostered discussion on different aspects, such as the weight that molecular phylogeny has to have on the definition of new major taxa (Gribaldo and Brochier 2012; Garrity and Oren 2012).

Our phylogenetic analyses excluding the long-branching taxa (*Nanoarchaeum*, *Micrarchaeum*, *Parvarchaeum* and the Nanohaloarchaea) strongly supported that the root of the archaeal tree lies between the Euryarchaeota and the rest of archaeal lineages. This, together with the observation that the long-branching archaea can be considered as *bona fide* Euryarchaeota since they branch within this archaeal group when LBA problems are minimized, advocates for a major division of the domain Archaea into two major groups

between which the root is located: the Euryarchaeota and the so-called TACK supergroup. In addition to these phylogenetic considerations, these two groups have similar levels of ecological and evolutionary diversity. As their name evokes (*eurus* meaning wide), Euryarchaeota have for long been known to exhibit a variety of metabolic capacities and to occupy a broad range of habitats (Woese et al., 1990). This is also the case for the TACK supergroup (Guy and Ettema 2011), thriving in high-temperature environments (Crenarchaeota, Aigarchaeota and Korarchaeota) but also in mesophilic ones (Thaumarchaeota) and relying on a large variety of metabolisms. On the other hand, the evolutionary divergence among distant TACK members is very similar to that among distant Euryarchaeota. For example, the average 16S rRNA sequence identity between Thaumarchaeota and Desulfurococcales (Crenarchaeota) is of ~75%, identical to that between Thermococcales and Halobacteria within the Euryarchaeota. The phylogenetic depth, in terms of sequence divergence, for our set of protein markers was also similar for the Euryarchaeota and the TACK group (~65% amino acid sequence similarity for comparisons of the same taxa as above).

The separation of the Euryarchaeota and the TACK group represents the primary split among the known archaeal species and these two groups have comparable ecological and phylogenetic diversities. Thus, it would be logical to give the TACK group the same taxonomic level as the Euryarchaeota, namely a kingdom rank. This would require providing a formal name to the TACK group. We propose to call this new kingdom Proteoarchaeota, making reference to the Greek god of the sea Proteus, able to display many different forms. The same prefix Proteo- was used in the name Proteobacteria also to point to the vast diversity of this bacterial group (Stackebrandt et al., 1988). The erection of the kingdom Proteoarchaeota would entail the relegation in rank of several archaeal lineages that were given a kingdom (or superphylum) level. As mentioned above, this concerns the Nanoarchaeota, Thaumarchaeota, Aigarchaeota and Korarchaeota. They should be re-classified as classes as occurs for the different lineages that compose the Euryarchaeota. Therefore, we propose to apply them the new names Nanoarchaea, Thaumarchaea, Aigarchaea and Korarchaea, with their respective orders Nanoarchaeales, Thaumarchaeales, Aigarchaeales and Korarchaeales (table 1). Likewise, the former kingdom Crenarchaeota should be renamed as a class (Crenarchaea, which would be synonym of Thermoprotei). We consider that this amended scheme is the one that requires the smallest number of taxonomic changes and, at the same time, would be much more homogeneous for the whole archaeal domain than the current mix of different ranks (kingdoms, classes and orders) to refer to lineages that, actually, can be joined into only two major groups. Incidentally, our results reinforce that the Halobacteria branch with high support within the class Methanomicrobia, which turns out to be paraphyletic (fig. 3). Therefore, in order to

avoid the existence of a paraphyletic class, the Halobacteria should be retrograded to the status of order within the class Methanomicrobia. This is in agreement with the evidence supporting that the ancestor of the Halobacteria was a methanogen species that secondarily adapted to hypersaline environments (Nelson-Sathi et al., 2012).

Conclusions

In addition to the classical markers commonly used until now, most of them ribosomal proteins, our phylogenomic survey has allowed identifying 38 additional conserved proteins that can be used to reconstruct phylogenetic trees of the archaea rooted on bacterial homologues. The phylogenetic analyses of the complete set of all those markers (32 ribosomal and 38 new ones) converge to support a deep division of the domain Archaea in two major lineages. One corresponds to the kingdom Euryarchaeota, already defined 23 years ago (Woese et al., 1990) and the second to a miscellaneous collection of lineages that has been tentatively grouped under the informal denomination of TACK supergroup (Guy and Ettema 2011). This second group is clearly monophyletic in our analyses and has a level of phylogenetic diversity comparable to the one of the Euryarchaeota. In addition, the lineages composing it show a large panel of ecological adaptations. These points are not reflected by the taxonomic classification of the Archaea currently used, which gives similar ranks to groups as different in phylogenetic diversity and depth as the extremely wide Euryarchaeota and the much more reduced Nanoarchaeota, Korarchaeota or Aigarchaeota. We think that the best alternative would be to give the two major archaeal lineages the same taxonomic rank. For that, the most parsimonious solution is keeping the kingdom rank already given to the Euryarchaeota and to erect a new kingdom to contain the TACK lineages. We propose to call these new kingdom Proteoarchaeota to highlight its high ecological and phylogenetic diversity.

Supplementary Material

Supplementary figures S1–S22 and tables S1-S3 are available at http://www.xxx.xxx/).

Acknowledgments

This work was supported by the French National Agency for Research (EVOLDEEP project, contract number ANR-08-GENM-024-002) and by the Investissement d'Avenir grant (ANR-10-BINF-01-01). C.B-A is member of the Institut Universitaire de France.

Literature Cited

Altschul SF, Madden TL, Schäffer 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; 25:3389-402.

Barker, H. A. Studies upon the methane-producing bacteria. Arch. Mikrobiol. 1936; 7:420–438.

Baker BJ, Comolli LR, Dick GJ, Hauser LJ, Hyatt D, Dill BD, Land ML, Verberkmoes NC, Hettich RL, Banfield JF. Enigmatic, ultrasmall, uncultivated Archaea. Proc Natl Acad Sci U S A. 2010; 107:8806-8811.

Barns SM, Delwiche CF, Palmer JD, Pace NR. Perspectives on archaeal diversity, thermophily and monophyly from environmental rRNA sequences. Proc Natl Acad Sci USA. 1996; 93:9188–9193.

Branciamore S, Gallori E, Di Giulio M. The basal phylogenetic position of *Nanoarchaeum equitans* (Nanoarchaeota). Front Biosci. 2008; 13:6886-6892.

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.

Brochier-Armanet C, Boussau B, Gribaldo S, Forterre P. Mesophilic Crenarchaeota: proposal for a third archaeal phylum, the Thaumarchaeota. Nat Rev Microbiol. 2008; 6:245–252.

Brochier-Armanet C, Forterre P, Gribaldo S. Phylogeny and evolution of the Archaea: one hundred genomes later. Curr Opin Microbiol. 2011; 14:274-281.

Brochier-Armanet C, Gribaldo S, Forterre P. Spotlight on the Thaumarchaeota. ISME J. 2012; 6:227-230.

Cox CJ, Foster PG, Hirt RP, Harris SR, Embley TM. The archaebacterial origin of eukaryotes. Proc Natl Acad Sci U S A. 2008; 105:20356-20361.

Criscuolo A, Gribaldo S. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. BMC Evol Biol. 2010; 10:210.

DeLong EF. Archaea in coastal marine environments. Proc Natl Acad Sci USA. 1992; 89:5685–5689.

Di Giulio M. *Nanoarchaeum equitans* is a living fossil. J Theor Biol. 2006; 242:257-260.

Di Giulio M. The split genes of *Nanoarchaeum equitans* are an ancestral character. Gene. 2008; 421:20-26.

Elkins JG, Podar M, Graham DE, Makarova KS, Wolf Y, Randau L, Hedlund BP, Brochier-Armanet C, Kunin V, Anderson I, Lapidus A, Goltsman E, Barry K, Koonin EV, Hugenholtz P, Kyrpides N, Wanner G, Richardson P, Keller M, Stetter KO. A korarchaeal genome reveals insights into the evolution of the Archaea. Proc Natl Acad Sci U S A. 2008; 105:8102-8107.

Embley TM, Martin W. Eukaryotic evolution, changes and challenges. Nature. 2006; 440:623-630.

Fuhrman JA, McCallum K, Davis AA. Novel major archaebacterial group from marine plankton. Nature. 1992; 356:148–149.

Garrity GM, Oren A. Response to Gribaldo and Brochier-Armanet: time for order in microbial systematics. Trends Microbiol. 2012; 20:353-354.

Gribaldo S, Brochier-Armanet C. Time for order in microbial systematics. Trends Microbiol. 2012; 20:209-210.

Gribaldo S, Poole AM, Daubin V, Forterre P, Brochier-Armanet C. The origin of eukaryotes and their relationship with the Archaea: are we at a phylogenomic impasse? Nat Rev Microbiol. 2010; 8:743-752.

Guy L, Ettema TJ. The archaeal 'TACK' superphylum and the origin of eukaryotes. Trends Microbiol. 2011; 19:580-587.

Jurgens G, Lindstrom K, Saano A. Novel group within the kingdom Crenarchaeota from boreal forest soil. Appl Environ Microbiol. 1997; 63:803–805.

Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013; 30:772-780.

Kostka M, Uzlikova M, Cepicka I, Flegr J. SlowFaster, a user-friendly program for slow-fast analysis and its application on phylogeny of Blastocystis. BMC Bioinformatics. 2008; 9:341.

López-Garcia P, Moreira D. Metabolic symbiosis at the origin of eukaryotes. Trends Biochem Sci. 1999; 24:88-93.

Moore CE, Archibald JM. Nucleomorph genomes. Annu Rev Genet. 2009; 43:251-264.

Narasingarao P, Podell S, Ugalde JA, Brochier-Armanet C, Emerson JB, Brocks JJ, Heidelberg KB, Banfield JF, Allen EE. De novo metagenomic assembly reveals abundant novel major lineage of Archaea in hypersaline microbial communities. ISME J. 2012; 6:81-93.

Nelson-Sathi S, Dagan T, Landan G, Janssen A, Steel M, McInerney JO, Deppenmeier U, Martin WF. Acquisition of 1,000 eubacterial genes physiologically transformed a methanogen at the origin of Haloarchaea. Proc Natl Acad Sci U S A. 2012; 109:20537-20542.

Nunoura T, Takaki Y, Kakuta J, Nishi S, Sugahara J, Kazama H, Chee GJ, Hattori M, Kanai A, Atomi H, Takai K, Takami H. Insights into the evolution of Archaea and eukaryotic protein modifier systems revealed by the genome of a novel archaeal group. Nucleic Acids Res. 2011; 39:3204-3223.

Philippe H. MUST, a computer package of Management Utilities for Sequences and Trees. Nucleic Acids Res. 1993; 21:5264-5272.

Philippe H, Brinkmann H, Lavrov DV, Littlewood DT, Manuel M, Wörheide G, Baurain D. Resolving difficult phylogenetic questions: why more sequences are not enough. PLoS Biol. 2011; 9:e1000602.

Philippe H, Laurent J. How good are deep phylogenetic trees? Curr Opin Genet Dev. 1998; 8:616-623.

Randau L, Münch R, Hohn MJ, Jahn D, Söll D. *Nanoarchaeum equitans* creates functional tRNAs from separate genes for their 5'- and 3'-halves. Nature. 2005; 433:537-541.

Rodríguez-Ezpeleta N, Brinkmann H, Roure B, Lartillot N, Lang BF, Philippe H. Detecting and overcoming systematic errors in genome-scale phylogenies. Syst Biol. 2007; 56:389-399.

Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst Biol. 2012; 61:539-542.

Roure B, Baurain D, Philippe H. Impact of missing data on phylogenies inferred from empirical phylogenomic data sets. Mol Biol Evol. 2013; 30:197-214.

Schleper C, Holben W, Klenk HP. Recovery of crenarchaeotal ribosomal DNA sequences from freshwater-lake sediments. Appl Environ Microbiol. 1997; 63:321–323.

Spang A, Hatzenpichler R, Brochier-Armanet C, Rattei T, Tischler P, Spieck E, Streit W, Stahl DA, Wagner M, Schleper C. Distinct gene set in two different lineages of ammonia-oxidizing archaea supports the phylum Thaumarchaeota. Trends Microbiol. 2010; 18:331-340.

Stackebrandt E, Murray RGE, Trüper HG. *Proteobacteria* classis nov., a name for the phylogenetic taxon that includes the "purple bacteria and their relatives". Int J Syst Bacteriol. 1988; 38:321-325.

Stamatakis A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics. 2006; 22:2688-2690.

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 U S A. 2003; 100:12984-12988.

Table 1

Revised classification of Archaea into two major phyla and their respective classes and orders

Phylum	Class	Order
Euryarchaeota	Archaeoglobi	Archaeoglobales
	Methanobacteria	Methanobacteriales
	Methanococci	Methanococcales
	Methanomicrobia	Methanocellales
		Methanomicrobiales
		Methanosarcinales
		Halobacteriales
	Methanopyri	Methanopyrales
	Nanoarchaea	Nanoarchaeales
	Thermococci	Thermococcales
	Thermoplasmata	Thermoplasmatales
Proteoarchaeota	Aigarchaea	Aigarchaeales
	Crenarchaea*	Acidilobales
		Desulfurococcales
		Fervidicoccales
		Sulfolobales
		Thermoproteales
	Korarchaea	Korarchaeales
	Thaumarchaea	Cenarchaeales
		Nitrosopumilales
		Nitrososphaerales

* The current class Thermoprotei would be synonym of Crenarchaea.

Figure legends

FIG. 1.- Bayesian phylogenetic tree of Archaea rooted on bacterial sequences. The tree is based on the concatenation of 32 ribosomal proteins (2560 sites). Numbers at branches are posterior probabilities. The scale bar indicates the number of substitutions per position.

FIG. 2.- Bayesian phylogenetic tree of Archaea rooted on bacterial sequences. The tree is based on the concatenation of 38 diverse conserved proteins (6890 sites). Numbers at branches are posterior probabilities. The scale bar indicates the number of substitutions per position.

FIG. 3.- Bayesian phylogenetic tree of Archaea rooted on bacterial sequences. The tree is based on the concatenation of 32 ribosomal proteins and 38 diverse conserved proteins (9450 sites). Numbers at branches are posterior probabilities. The scale bar indicates the number of substitutions per position.

FIG. 4.- Unrooted Bayesian phylogenetic tree of Archaea including fast-evolving taxa. The tree is based on the concatenation of 32 ribosomal proteins and 38 diverse conserved proteins (9450 sites). Numbers at branches are posterior probabilities. The scale bar indicates the number of substitutions per position.

Supplementary material

Supplementary table S1. List of archaeal and bacterial complete genome sequences used in this work.

Supplementary table S2. List of all protein markers used in this work.

Supplementary table S3. List of species-pairs used for evolutionary estimations for gene-by-gene desaturation analysis.

Supplementary fig. S1. Maximum likelihood tree based on the complete set of markers (32 ribosomal proteins and 38 diverse conserved proteins, 9450 sites) and rooted using bacterial sequences as outgroup. Numbers at branches are bootstrap proportions. The scale bar indicates the number of substitutions per position.

Supplementary fig. S2. Unrooted maximum likelihood tree based on the complete set of markers (32 ribosomal proteins and 38 diverse conserved proteins, 9450 sites). Numbers at branches are bootstrap proportions. The scale bar indicates the number of substitutions per position.

Supplementary figs. S3-S12. Rooted maximum likelihood trees corresponding to the site-by-site desaturation analysis. Numbers at branches are bootstrap proportions. The scale bar indicates the number of substitutions per position.

Supplementary figs. S13-S22. Rooted maximum likelihood trees corresponding to the gene-by-gene desaturation analysis. Numbers at branches are bootstrap proportions. The scale bar indicates the number of substitutions per position.

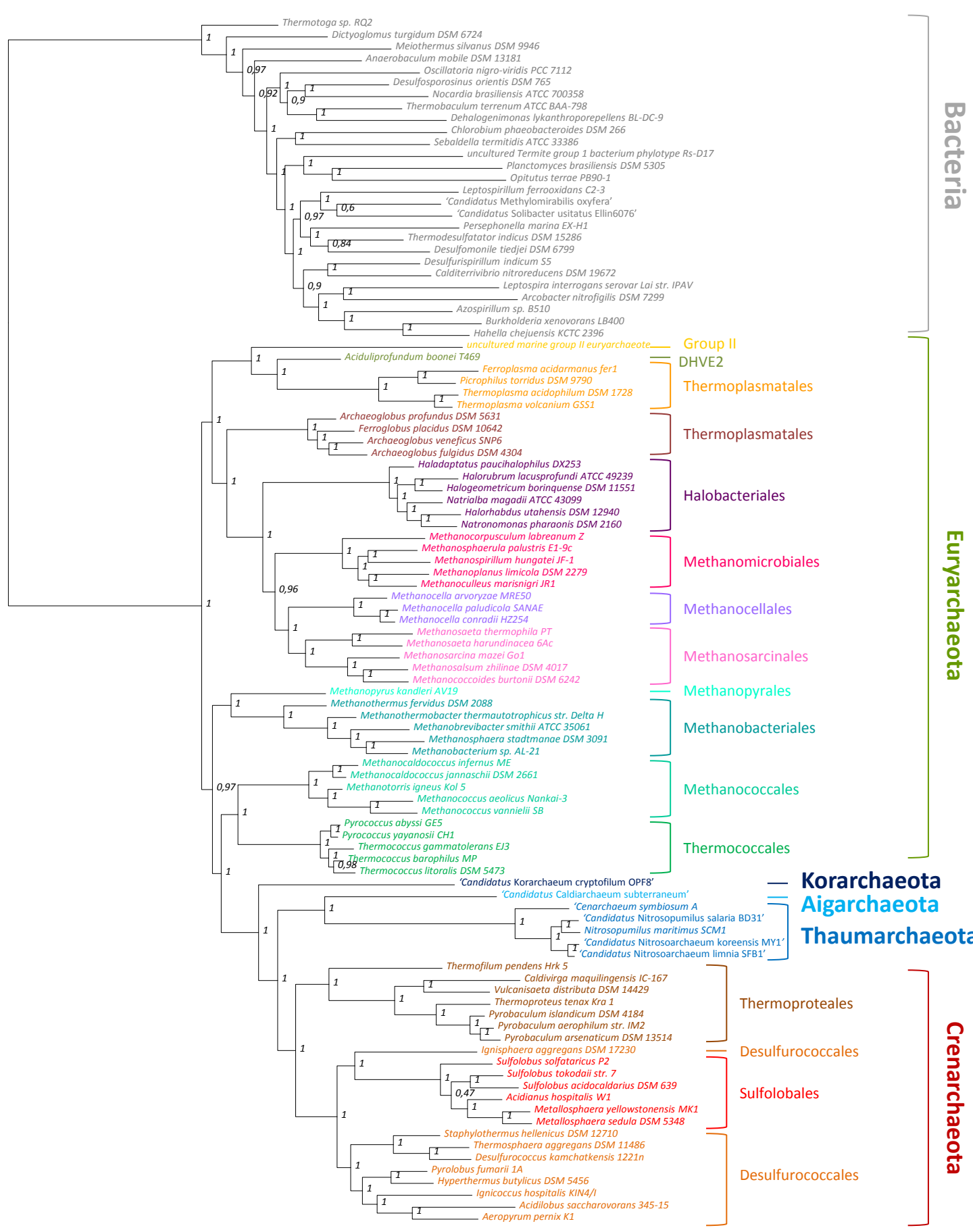


Figure 1

101 species - 2560 positions

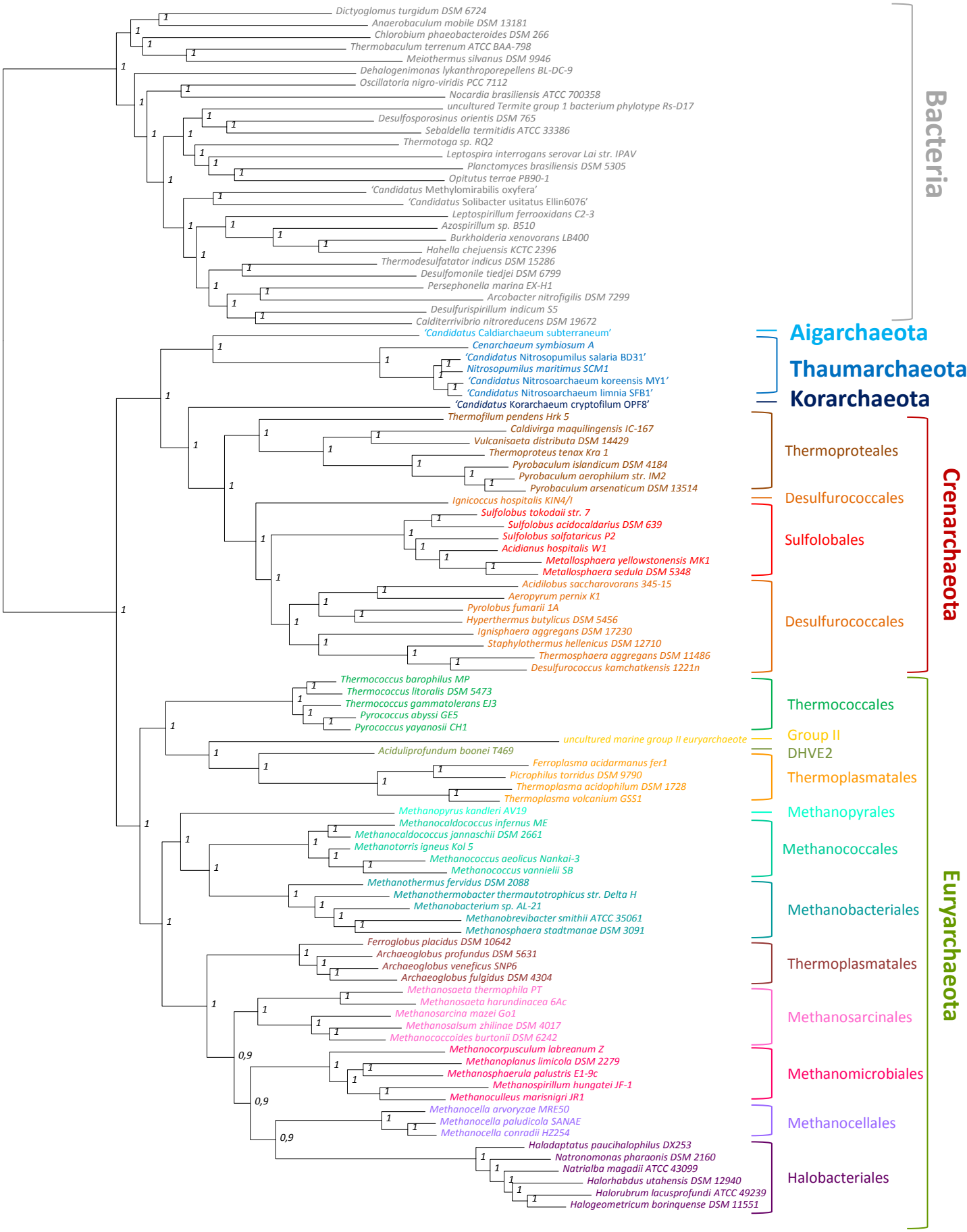


Figure 2

101 species - 9450 positions

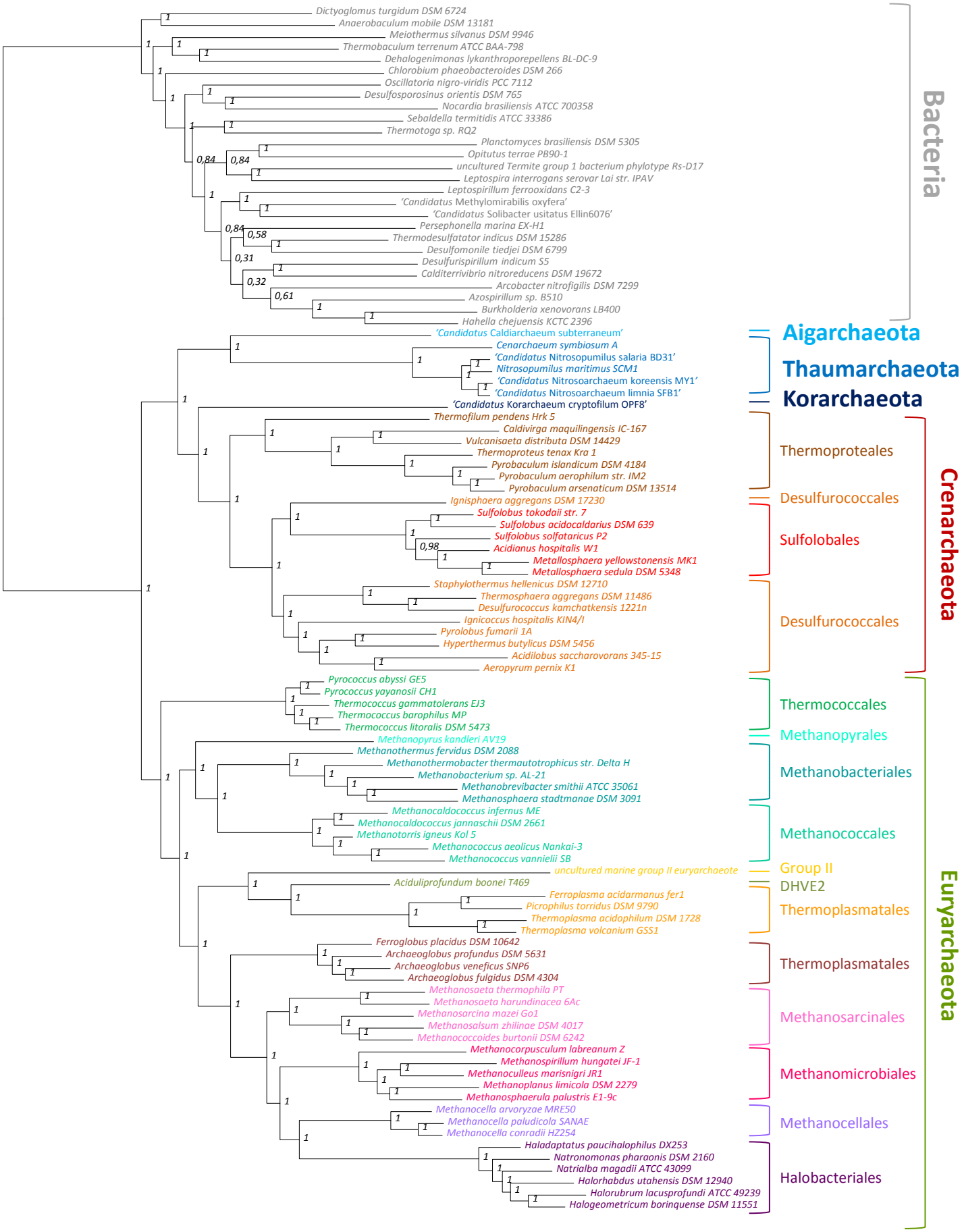


Figure 3

101 species - 9540 positions

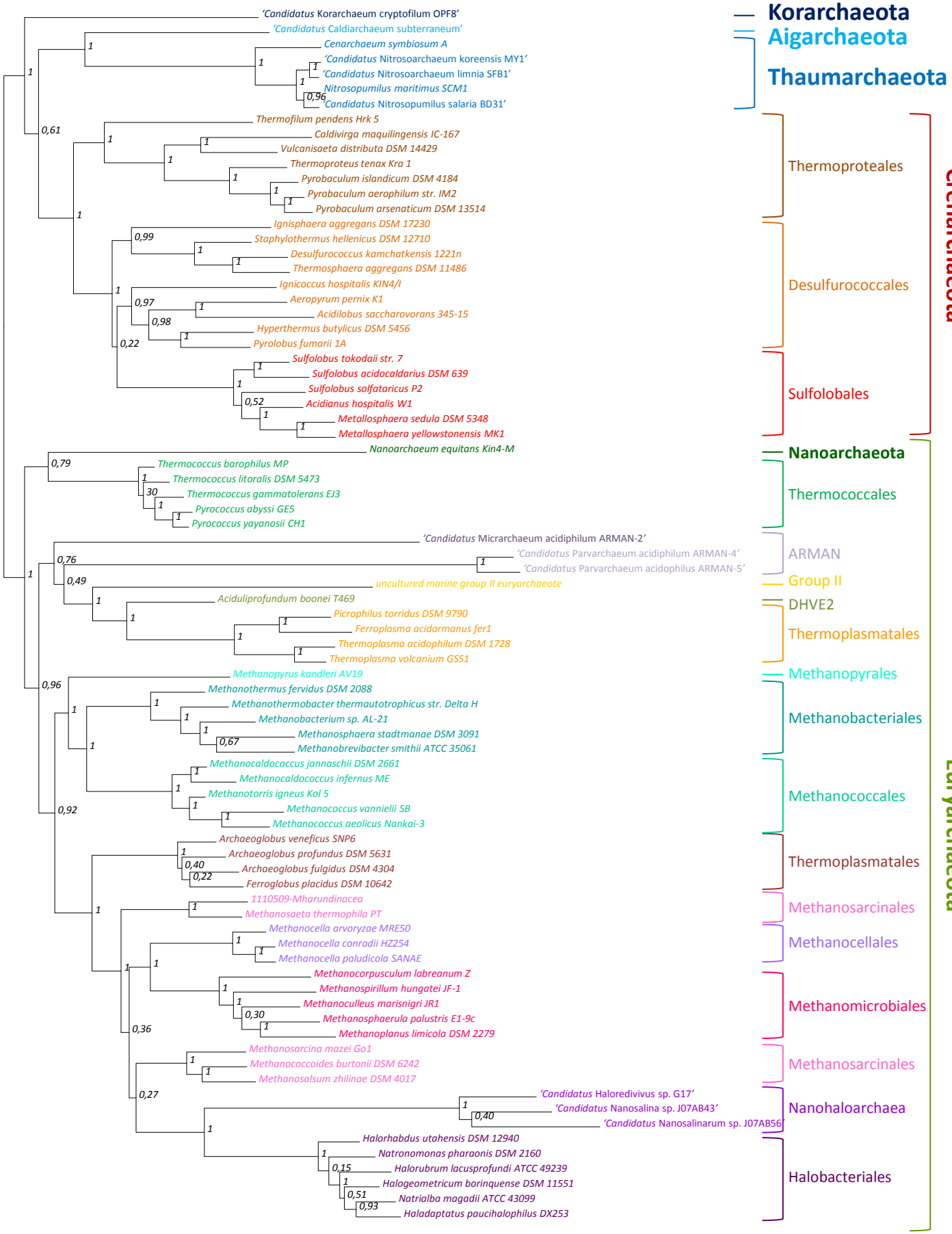


Figure 4

81 species - 9540 positions

C. Synthèse et éléments de discussion

Dans cet article, nous avons donc cherché à placer la racine de l'arbre des archées à partir de deux sets de protéines, 32 protéines ribosomiques et 38 nouveaux marqueurs impliqués dans différents processus cellulaires, avec leur homologues bactériens.

Nous avons tout d'abord écarté de nos premières analyses les lignées problématiques des genres *Nanoarchaeum*, '*Ca.* Micrarchaeum', '*Ca*. Parvarchaeum' et les Nanohaloarchaeota, afin de diminuer le risque de LBA. La première phylogénie inférée à partir des 32 protéines ribosomiques en maximum de vraisemblance et en bayésien place la racine au sein des Euryarchaeota. Différents groupes d'euryarchées émergent de façon paraphylétique et avec un très faible soutien, alors que les Thaumarchaeota, Aigarchaeota, Crenarchaeota et Korarchaeota (le superphylum « TACK » proposé par Guy et Ettema en 2011 (Guy and Ettema 2011)) groupent de façon monophylétique bien soutenue à l'intérieur des euryarchées. Les analyses du second jeu de données, contenant 38 nouveaux marqueurs, placent la racine entre les Euryarchaeota, monophylétiques, et le groupe « TACK » avec un support au nœud maximal. La concaténation de ces deux jeux de données permet aussi d'obtenir une phylogénie racinée entre les Euryarchaeota et le groupe « TACK » très bien soutenue. A partir de ce résultat, l'analyse en bayésien de la position des archées de petite taille dans cette phylogénie mais sans groupe extérieur (pour minimiser les problèmes de LBA) a montré qu'elles se replaçaient au sein des euryarchées et n'étaient pas groupées. La racine de l'arbre des archées serait donc entre les Euryarchaeota et le superphylum « TACK ».

Une diversité phylogénétique et physiologique d'une ampleur similaire à celle rencontrée au sein des Euryarchaeota (qui leur a valu leur nom, *eurus* voulant dire ample, spacieux) se retrouve aussi dans le groupe « TACK ». Cette observation, ainsi que la monophylie très robuste de ce groupe et la position de la racine le séparant des Euryarchaeota, nous ont conduits à proposer un changement majeur dans la taxonomie des archées, à savoir, la création d'un nouveau phylum 'Proteoarchaeota' afin de donner un nom formel au groupe « TACK ». Ceci doit être accompagné du déclassement des phyla composant le groupe « TACK » en ordres : Thaumarchaeales, Aigarchaeales, et Korarchaeales ; les Crenarchaeota deviendraient les Crenarchaea. Nous proposons aussi de supprimer le phylum Nanoarchaeota et de le considérer comme un ordre au sein des Euryarchaeota sous le nom de Nanoarchaeales.

La multiplication des phyla proposés ces dernières années est assez symptomatique de l'absence de règles claires en ce qui concerne la taxonomie de haut niveau chez les archées et les bactéries, créant un flou dans l'appellation des différents groupes, et par conséquent dans la compréhension et la perception de leur phylogénie. Certains de ces phyla ont été proposés sur

l'analyse de très peu de données en termes de génomes séquencés appartenant au groupe et du nombre de marqueurs utilisés. La publication très récente de Rinke et collaborateurs 2013 (Rinke et al. 2013), décrivant le séquençage de 201 nouveaux génomes microbiens et proposant plusieurs nouveaux phyla d'archées, est un exemple de plus de cette mouvance. Or, notre travail montre qu'une analyse fine permet d'obtenir une phylogénie des archées racinée et de bonne qualité. Il est important de remarquer ici que ce résultat a été obtenu grâce à l'analyse d'autres marqueurs que les protéines ribosomiques, montrant que malgré la puissance de ces marqueurs pour la résolution de la phylogénie des archées, il peut être nécessaire d'utiliser des protéines différentes pour répondre à des questions différentes. Une analyse fine avec un jeu de données adapté à la question posée, en l'occurrence, les relations phylogénétiques anciennes entre les lignées d'archées, semble donc nécessaire avant la proposition d'un nouveau groupe de haut niveau taxonomique tel qu'un phylum. La nouvelle classification que nous proposons rétablit un équilibre entre les ordres d'Euryarchaeota et les groupes composant les 'Proteoarchaeota', en termes de diversité et de distance phylogénétique.

Le positionnement de la racine de l'arbre des archées est aussi très important en ce qui concerne la nature du dernier ancêtre commun de ce domaine. D'après notre analyse, il semblerait avoir été hyperthermophile, les premiers ordres divergeant au sein des Euryarchaeota et des 'Proteoarchaeota' étant hyperthermophiles ou très thermophiles (les Thermococcales et Nanoarchaeales chez les euryarchées et les Korarchaeales et Aigarchaeales chez les 'proteoarchées'). Cet ancêtre aurait été aussi probablement très complexe, contenant au moins les caractères partagés entre les deux grands phyla, et non pas extrêmement réduit à l'image des nanoarchées comme *N. equitans* ou les ARMAN.

Enfin, en ce qui concerne la relation des eucaryotes avec les archées, il serait intéressant de construire un nouveau set de marqueurs sélectionnés pour répondre à cette question précise. Les homologues bactériens n'étant plus nécessaires, ces marqueurs pourront être choisis indépendamment de la présence d'homologues bactériens, de façon à s'affranchir de la nécessité d'avoir des protéines universelles, ainsi que du risque de biais liés à la différence de vitesse d'évolution entre les domaines, ou de la saturation. Cette méthode a déjà été proposé par Gribaldo et collaborateurs en 2010 (Simonetta Gribaldo et al. 2010).

Chapitre 3 : Répartition taxonomique et histoire évolutive des protéines d'archées : exemple du système chaperonne DnaK et de la protéine DnaJ-Fer.

A. Introduction

Notre recherche de nouveaux marqueurs pour l'inférence de la phylogénie des archées est passée par l'analyse de la phylogénie de chacune des protéines codées dans les trois génomes de *Cenarchaeum symbiosum*, *Nitrosopumilus maritimus* et '*Ca.* Caldiarchaeum subterraneum'. Cette analyse non automatique a permis de noter un certain nombre d'observations sur ces phylogénies, telles que l'absence ou la présence d'homologues bactériens et eucaryotes, les phyla d'archées représentés, des cas de transferts horizontaux potentiels, ou des cas de paralogie. Ces données n'ont pas été exploitées dans ma thèse, le but premier de cette étude étant la recherche de marqueurs pour l'étude de la phylogénie des archées, mais elles seraient très intéressantes à exploiter dans le futur. Deux exemples en sont la preuve. Tout d'abord, le travail présenté dans le Chapitre 2 a été entamé car il a été possible de vérifier très rapidement si des homologues bactériens étaient présents dans les jeux de données de départ des 200 marqueurs sélectionnés pour l'inférence de la phylogénie des archées. L'autre exemple de l'intérêt de ce travail est de mettre en évidence des histoires évolutives particulières comme celle de la protéine DnaJ-Fer présentée dans l'article ci-dessous.

Cette protéine a une répartition taxonomique très particulière, elle n'est présente que chez certaines Thaumarchaeota et les Viridiplantae (algues et plantes vertes), ce qui a attiré notre attention. L'étude de l'histoire évolutive de cette protéine nous a amenés à nous intéresser aussi à l'histoire évolutive du système chaperonne DnaK. En effet, les protéines contenant un domaine peptidique dnaJ, sont connues pour interagir avec DnaK dans les réponses aux chocs thermiques et d'autres types de stress. La protéine DnaK et ses co-chaperonnes GrpE et DnaJ sont présentes de façon ubiquitaire chez les bactéries et les eucaryotes, mais sont absentes chez certaines archées, et totalement absentes chez les archées hyperthermophiles. Il a été montré que la protéine DnaJ-Fer, composée d'un domaine dnaJ et d'un domaine ferredoxine, intervient en association avec DnaK dans le chloroplaste de l'algue verte *Chlamydomonas reinhardtii* (Dorn et al. 2010). Nous avons montré que chez les archées, DnaK, GrpE et DnaJ (une protéine différente de la protéine DnaJ-Fer, contenant aussi un domaine DnaJ et connue pour interagir avec DnaK) avaient été acquises par transfert horizontal de gène depuis des bactéries. Au moins deux évènements indépendants de transferts ont été observés, un vers les Halobacteriales et les Nanohaloarchaea et un autre vers d'autres euryarchées. Elles ont ensuite été transférées plusieurs fois au sein des archées. C'est le cas

chez les Thaumarchaeota, qui auraient acquis ce système depuis d'autres archées (euryarchées), même si le donneur exact n'a pas pu être identifié. Parallèlement, nous avons aussi mis en évidence que la protéine DnaJ-Fer était issue, chez les Thaumarchaeota, d'un transfert depuis les Viridiplantae, celles-ci ayant acquis au moins le domaine ferredoxine depuis des cyanobactéries lors de l'endosymbiose chloroplastique. Le domaine dnaJ ne semble pas avoir la même origine chez les Viridiplantae et chez les Thaumarchaeota mais l'origine commune du domaine ferredoxine nous incite à penser que la protéine a été transférée déjà fusionnée et qu'un remplacement homologue entre deux domaine dnaJ s'est produit chez les thaumarchées. La répartition taxonomique du système DnaK-GrpE-DnaJ et DnaJ-Fer et leur fonction cellulaire nous ont amenés à proposer que ces quatre protéines étaient peut-être impliquées dans l'adaptation à la mésophilie chez les archées et particulièrement chez les thaumarchées.

Cette étude a fait l'objet de la publication « *Horizontal gene transfer of a chloroplast DnaJ-Fer protein to Thaumarchaeota and the evolutionary history of the DnaK chaperone system in Archaea* » (Petitjean, Moreira, López-García, Brochier-Armanet) dans le journal *BMC Evolutionary Biology* (Petitjean et al. 2012).

B. Manuscrit de l'article 3 : « Horizontal gene transfer of a chloroplast DnaJ-Fer protein to Thaumarchaeota and the evolutionary history of the DnaK chaperone system in Archaea »

BMC Evolutionary Biology

RESEARCH ARTICLE Open Access

Horizontal gene transfer of a chloroplast DnaJ-Fer protein to Thaumarchaeota and the evolutionary history of the DnaK chaperone system in Archaea

Céline Petitjean[1,2], David Moreira[2], Purificación López-García[2] and Céline Brochier-Armanet[3*]

Abstract

Background: In 2004, we discovered an atypical protein in metagenomic data from marine thaumarchaeotal species. This protein, referred as DnaJ-Fer, is composed of a J domain fused to a Ferredoxin (Fer) domain. Surprisingly, the same protein was also found in Viridiplantae (green algae and land plants). Because J domain-containing proteins are known to interact with the major chaperone DnaK/Hsp70, this suggested that a DnaK protein was present in Thaumarchaeota. DnaK/Hsp70, its co-chaperone DnaJ and the nucleotide exchange factor GrpE are involved, among others, in heat shocks and heavy metal cellular stress responses.

Results: Using phylogenomic approaches we have investigated the evolutionary history of the DnaJ-Fer protein and of interacting proteins DnaK, DnaJ and GrpE in Thaumarchaeota. These proteins have very complex histories, involving several inter-domain horizontal gene transfers (HGTs) to explain the contemporary distribution of these proteins in archaea. These transfers include one from Cyanobacteria to Viridiplantae and one from Viridiplantae to Thaumarchaeota for the DnaJ-Fer protein, as well as independent HGTs from Bacteria to mesophilic archaea for the DnaK/DnaJ/GrpE system, followed by HGTs among mesophilic and thermophilic archaea.

Conclusions: We highlight the chimerical origin of the set of proteins DnaK, DnaJ, GrpE and DnaJ-Fer in Thaumarchaeota and suggest that the HGT of these proteins has played an important role in the adaptation of several archaeal groups to mesophilic and thermophilic environments from hyperthermophilic ancestors. Finally, the evolutionary history of DnaJ-Fer provides information useful for the relative dating of the diversification of Archaeplastida and Thaumarchaeota.

Keywords: DnaJ/Hsp40, DnaK/Hsp70, Hyperthermophily, Archaeplastida, Phylogeny, Archaea, Thaumarchaeota, Horizontal gene transfer, Mesophily

Background

The 70 kD heat shock proteins (called DnaK in bacteria and Hsp70 in eukaryotes) form a large family of molecular chaperones upregulated in cells suffering various stresses, including heat shocks and heavy metal exposure [1,2]. In addition, these proteins play a major role during protein synthesis by binding to the nascent peptides exiting the ribosome in order to prevent their aggregation and facilitating their folding in the optimal functional conformation [3]. During the interaction with the partially synthesized peptides, DnaK/Hsp70 increases its ATPase activity [3]. This chaperone has two main partners: the J-proteins [4,5] and the nucleotide exchange factor, called GrpE in bacteria (or Mge1 [6] in mitochondria and Cge1 [7] in chloroplasts) and Bag-1, a eukaryotic functional analogue of GrpE [8]. The nucleotide exchange factor promotes the exchange of ADP to fresh ATP in the nucleotide-binding region of DnaK/Hsp70, whereas the J-proteins stimulate the ATPase activity in order to stabilize the interaction of DnaK with unfolded proteins [5,9,10]. The J-proteins form a large family of proteins, which are structurally and functionally diverse but all have the capacity to interact with DnaK/Hsp70 through their J-domain [4,11]. Among them, DnaJ/Hsp40 proteins form the largest subfamily [12]. They control the flux of unfolded

* Correspondence: celine.brochier-armanet@univ-lyon1.fr
[3]CNRS, UMR5558, Laboratoire de Biométrie et Biologie Evolutive, Université de Lyon, Université Lyon 1, 43 boulevard du 11 novembre 1918, 69622, Villeurbanne, France
Full list of author information is available at the end of the article

polypeptides into and out of the substrate-binding domain of DnaK/Hsp70 [9,11].

DnaK proteins are widespread, being encoded by a single gene in most bacterial genomes, whereas most eukaryotic genomes harbor several Hsp70 genes that may have diverse evolutionary origins [1,13,14]. For example, in the green alga *Chlamydomonas reinhardtii*, five Hsp70 copies are present, all them encoded in the nuclear genome despite being targeted in diverse cellular compartments: three of them most likely originated by duplications from an ancestral eukaryotic gene (one expressed in the cytoplasm and two in the endoplasmic reticulum); one has a mitochondrial origin and is exported into the mitochondria, whereas the latter originated from the chloroplast endosymbiosis and is targeted into the chloroplast [15]. In contrast with DnaK, the J-proteins are encoded in multiple copies in bacterial genomes [9]. This is also the case in eukaryotes, where they work in the different cell compartments in association with the Hsp70 proteins cited above [9,11]. Finally, the nucleotide exchange factor GrpE is present in one copy in most of bacterial genomes, whereas the eukaryotic Mge1, Cge1 and Bag-1 are encoded in the nucleus but addressed to the mitochondria, chloroplasts, and to the nucleus and the cytoplasm, respectively [7,8].

The presence of DnaK, DnaJ and GrpE has been reported in several archaeal genomes [16], more precisely in several euryarchaeota but never in crenarchaeotal species. The best studied case concerns DnaK. A phylogenetic analysis by Gribaldo and coworkers suggested that this protein was acquired by several archaea by horizontal gene transfer (HGT) from different bacterial donors [17]. These authors observed three different groups of archaeal DnaK sequences branching specifically with certain bacterial homologues. More precisely, *Methanosarcina mazei* (Methanosarcinales) was related to the *Clostridium* group of Firmicutes (low G+C Gram positive bacteria), *Halobacterium cutirubrum* and *Halobacterium marismortui* (Halobacteriales) to the Actinobacteria (high G+C Gram positive bacteria), whereas *Methanobacterium thermautotrophicum* (Methanobacteriales) and *Thermoplasma acidophilum* (Thermoplasmatales) branched with *Thermotoga maritima* (Thermotogales) [17]. More recently, Macario *et al.* (2006) studied in various bacteria and archaea the taxonomic distribution and the phylogeny not only of DnaK but also of GrpE and DnaJ. They showed that the genes coding for these three proteins were clustered in most of the genomes examined [16]. They also confirmed the results of Gribaldo *et al.* (1999), i.e. the likely existence of three HGT events from bacteria to archaea. However, they proposed a more complex scenario where the DnaK/DnaJ/GrpE cluster was first acquired from a bacterial donor by the ancestor of the Euryarchaeota, then lost in Methanococcales and in the common ancestor of Archaeoglobales, Halobacteriales and Methanosarcinales, and finally reacquired independently by Halobacteriales and Methanosarcinales from Actinobacteria and from Firmicutes, respectively [16]. Worth noting, in these two studies, none of the three proteins was detected in hyperthermophilic archaea.

In addition to these relatively well-characterized chaperones and co-chaperones, the study of a genomic fragment of an uncultured deep marine archaeon from an environmental DNA fosmid library revealed a very unusual J-protein, referred as DnaJ-Fer, composed of a J-domain fused with a Ferredoxin (Fer) domain [18]. The phylogenetic analysis of a 16S rRNA gene also found in this genomic fragment showed that it belonged to a member of the Thaumarchaeota, more precisely in the I.1a subgroup. These archaea, formerly classified as Group I, a sublineage of Crenarchaeota [19,20], have been recently proposed to represent a third phylum of Archaea together with the Euryarchaeota and Crenarchaeota [21]. Thaumarchaeota are widespread in many environments, including marine and freshwater, soil and sediment [22,23]. Surprisingly, the presence of DnaJ-Fer proteins has also been reported in Viridiplantae (including green algae and plants), with three homologues (CDJ3, 4 and 5) in *C. reinhardtii* [24]. These proteins are localized in the chloroplast of this green alga where they interact with the chloroplast Hsp70B and Cge1 proteins. However, the precise function of these DnaJ-Fer proteins in *C. reinhardtii* remains to be elucidated. According to the location and the nature of its partners, it would be tempting to hypothesize a cyanobacterial origin of the DnaJ-Fer protein. However, no homologue has been detected in Cyanobacteria [24].

Two hypotheses can explain the unexpected taxonomic distribution of the DnaJ-Fer protein in Thaumarchaeota and Viridiplantae: either two independent and convergent fusions of the two protein domains occurred in these two distantly related lineages, or a single fusion occurred in one of them followed by a HGT to the other lineage [24]. In this work, we have taken advantage of the recent burst of available archaeal complete genome sequences [25], including representatives of new major lineages such as the Thaumarchaeota, ARMAN or Nanohaloarchaeales, to decipher the evolutionary history of DnaK and its co-chaperones in Archaea, with especial attention on the intriguing DnaJ-Fer protein. Our results support a complex scenario in which HGT appears to have played an important role. In addition to other cases of HGT, Thaumarchaeota appear to have most likely acquired their DnaK, co-chaperones and DnaJ-Fer proteins by independent HGTs from multiple donors, including other archaea and plants.

Results

DnaJ-Fer proteins are widespread in viridiplantae and thaumarchaeota

We carried out an intensive survey of public sequence databases to find that DnaJ-Fer homologues are present in all Viridiplantae (green algae and land plants) for which complete genome sequences were available. In contrast, we did not detect them in Rhodophyta and Glaucophyta, the two other lineages composing the Plantae or Archaeplastida eukaryotic supergroup [26]. However, due to the scarcity of sequence data from these two lineages, we can not exclude the future discovery of DnaJ-Fer in some species belonging to them. In addition to green algae and land plants, DnaJ-Fer homologues were detected in the four available complete genomes of Thaumarchaeota (Additional file 1): *Cenarchaeum symbiosum* (a sponge symbiont) [27], the planktonic *Nitrosopumilus maritimus* (the first isolated thaumarchaeote) [28] and its two close relatives '*Candidatus* (*Ca*.) Nitrosoarchaeum limnia SFB1' [29] and in '*Ca*. Nitrosoarchaeum koreensis MY1' [30] which live in low salinity sediments and in the soil rhizosphere, respectively, as well as in several environmental fosmid sequences, all likely members of the mesophilic group I.1a. The protein was also present in *Nitrososphaera viennensis* (Schleper and Spang, personal communication) and '*Ca*. Nitrososphaera gargensis' [31], two moderate thermophilic representatives of the group I.1b. In contrast, it was absent in the thermophilic species '*Ca*. Nitrosocaldus yellowstonii', a representative of the more distant Hot Water Crenarchaeotic Group (HWCG) III (de la Torre, personal communication), and in '*Ca*. Caldiarchaeum subterraneum' [32], a representative of the 'Aigarchaeota' (formerly group HWCG I) which seems to be either the sister group of Thaumarchaeota or a deeply branching thaumarchaeotal lineage [22].

The J and Fer domains are two small domains (less than 100 amino acids) well conserved in the plant and thaumarchaeotal DnaJ-Fer sequences. The Fer domain was characterized by an amino acid motif CXXCXXC observed in all those sequences except in '*Ca*. N. gargensis' and *N. viennensis*, where the motif was CXXFXXC. Contrasting with the conservation of these two domains, we observed different sequence organizations of the DnaJ-Fer proteins in the Viridiplantae and in the Thaumarchaeota (Figure 1). In Viridiplantae, an N-terminal chloroplast signal region preceded the J and the Fer domains, and the protein ended with a long C-terminal region (up to 150 amino acids) of unknown function. In Thaumarchaeota, these N- and C-terminal regions were absent, but an inter-domain region (ranging between 54 and 92 amino acids) was present between the J and the Fer domains. This region was well conserved in *N. maritimus*, *C. symbiosum*, '*Ca*. Nitrosoarchaeum limnia SFB1', '*Ca*. Nitrosoarchaeum koreensis MY1' and the fosmids found in the environmental database (all belonging to the group I.1a), but was divergent and shorter (54 amino acids) in the sequences of '*N. gargensis*' and *N. viennensis*, the two representatives of group I.1b. The presence of this variable central region suggested that its role is probably structural and not functional in Thaumarchaeota. By contrast, much shorter or no central regions were present between the two domains in the plant sequences.

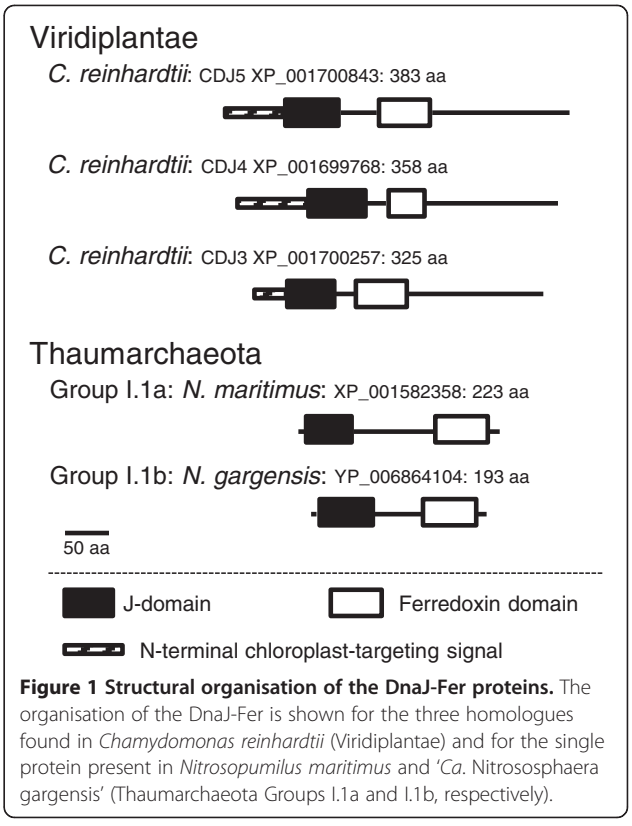


Figure 1 Structural organisation of the DnaJ-Fer proteins. The organisation of the DnaJ-Fer is shown for the three homologues found in *Chamydomonas reinhardtii* (Viridiplantae) and for the single protein present in *Nitrosopumilus maritimus* and '*Ca*. Nitrososphaera gargensis' (Thaumarchaeota Groups I.1a and I.1b, respectively).

The taxonomic distribution of the DnaJ-Fer protein results from an ancient HGT

Maximum likelihood (ML) analyses and Bayesian inference (BI) of the DnaJ-Fer alignment revealed three monophyletic groups (Figure 2). Two corresponded to Viridiplantae (ML bootstrap values (BV) = 71% and 52%, and BI posterior probabilities (PP) = 0.98 and 0.78, respectively) whereas the third gathered the thaumarchaeotal sequences (BV = 100% and PP = 1.00). Interestingly, the relationships among sequences within each of these groups were in agreement with the accepted species phylogeny and relatively well supported despite the small number positions (127 amino acids) kept for the phylogenetic analysis. More precisely, the dichotomy between group I.1a and group I.1b Thaumarchaeota was well supported (BV = 96% and PP = 1.00). The relationships among the green algae and land plant

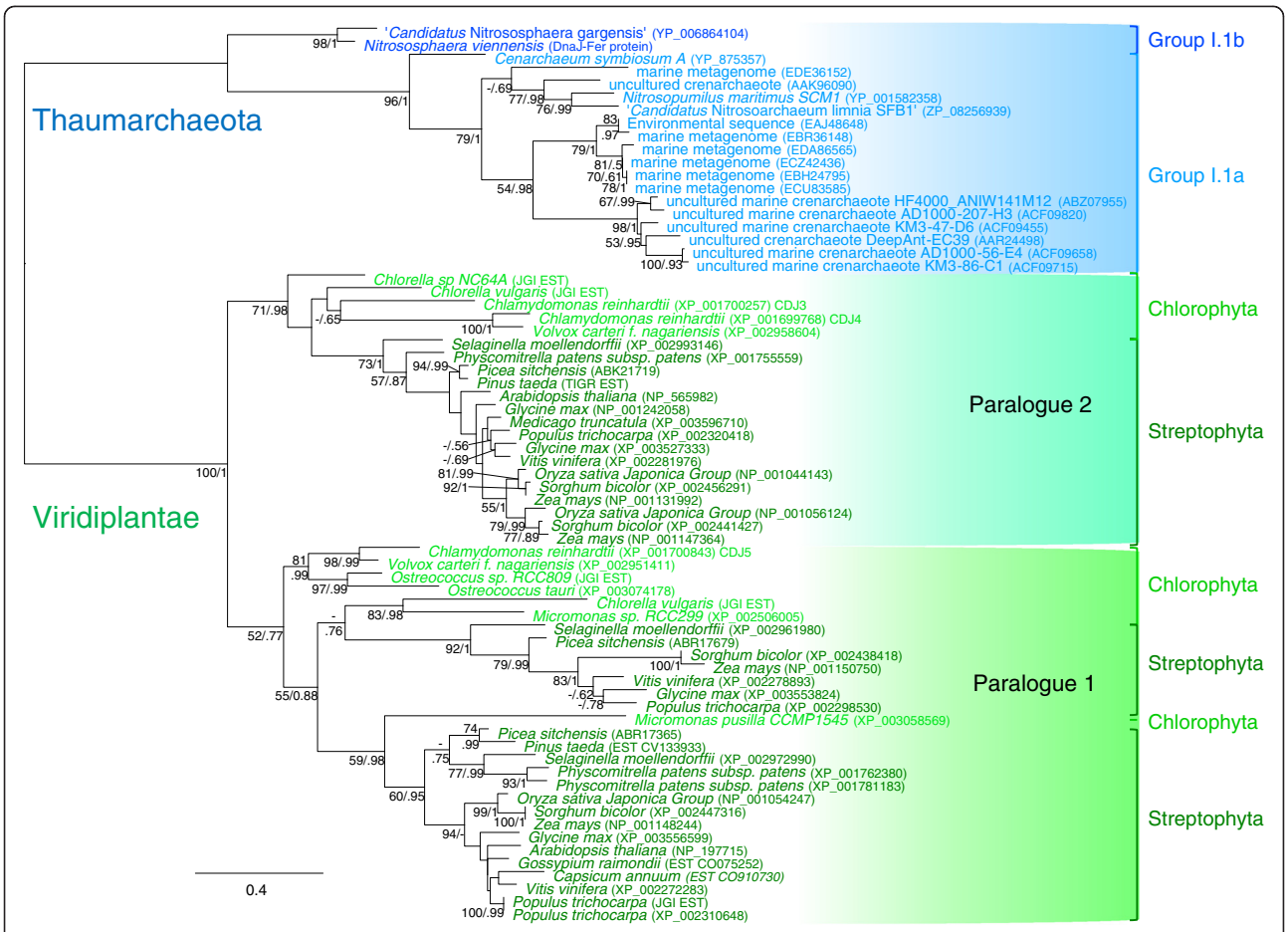


Figure 2 Unrooted ML phylogenetic tree of the DnaJ-Fer protein. The tree was reconstructed with 69 sequences and 127 positions with TreeFinder and the LG model + Γ4. Numbers at nodes represent bootstrap values and Bayesian posterior probabilities computed by TreeFinder and PhyloBayes, respectively (only values >50% and 0.5 are shown, dashes indicate that the corresponding support is inferior to the threshold, whereas when both supports are inferior to the thresholds no support values are indicated). The scale bar represents the average number of substitutions per site. The Viridiplantae sequences are shown in green and those of Thauamrchaeota in blue.

sequences were more complex since there were several copies of this protein in these species, most likely resulting from duplication events. A first duplication occurred almost certainly in the ancestor of Viridiplantae, leading to the two paralogues present in green algae and land plants. This event was followed by additional duplication events at the origin of the multiple copies of paralogues 1 and 2 observed in the viridiplantae lineages (Figure 2).

Phylogeny results indicated that the ancestor of thaumarchaeotal groups I.1a and I.1b already harboured the DnaJ-Fer gene and that the ancestor of Viridiplantae had two copies. If the unusual taxonomic distribution of DnaJ-Fer proteins was indicative of an HGT between Thaumarchaeota and Viridiplantae, the inferred phylogenies suggested that this HGT took place before the diversification of these two major lineages and was therefore relatively ancient (event 3 on Figure 3A). However, due to the lack of any suitable outgroup (no other lineage contained the DnaJ-Fer protein) it was not possible to determine the precise evolutionary origin of the DnaJ-Fer gene and the direction of the HGT between Thaumarchaeota and Viridiplantae. To tackle this issue we carried out phylogenetic analyses of the J and Fer domains separately. Indeed, although the association between these two domains is specific of Thaumarchaeota and Viridiplantae, each domain is widely distributed in present day organisms, opening the possibility to reconstruct rooted phylogenies for each of them.

The J and Fer domains have two different evolutionary origins

As expected because of the small number of conserved sequence positions, the ML phylogeny of the Fer domain was largely unresolved (data not shown). Nevertheless, the Fer domain of the DnaJ-Fer proteins of Viridiplantae and Thaumarchaeota branched within a single cluster, which also contained various bacterial and archaeal sequences. To improve the resolution of the phylogenetic

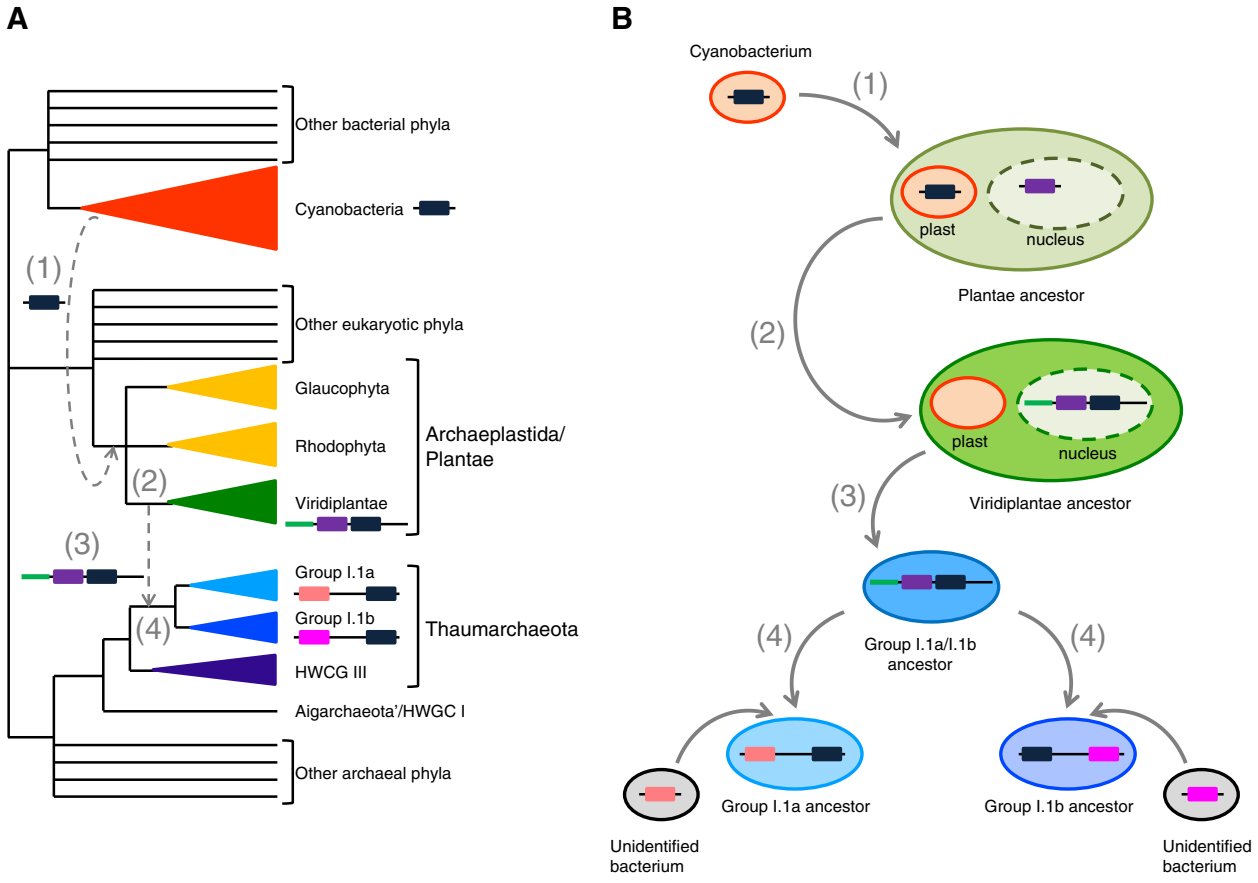


Figure 3 Origin and evolution of the DnaJ-Fer protein. (**A**) Schematic representation of the tree of life with the three Domains (Archaea, Bacteria and Eucarya) showing the time of the evolutionary events that have affected the DnaJ-Fer protein. (**B**) Evolutionary scenario for the origin and evolution of the DnaJ-Fer protein: (1) Acquisition of a cyanobacterial Fer domain-containing protein by the ancestor of Archaeplastida/Plantae; (2) translocation of the corresponding gene in the nucleus, fusion with a J domain coding gene and addition of a chloroplast signal peptide; (3) horizontal gene transfer of the DnaJ-Fer coding gene to the ancestor of thaumarchaeota groups I.1a and I.1b and (4) independent replacement of the J domain by J domains of bacterial origin in thaumarchaeotal groups I.1a and I.1b.

relationships between these sequences, we carried out an analysis of the sequences composing this cluster and close relatives using several more distantly related sequences as outgroup. The resulting ML tree supported the grouping of thaumarchaeotal and viridiplantae sequences (BV = 77% and PP = 0.99, Figure 4A), indicating that the Fer domain of the DnaJ-Fer proteins had a single origin and, most likely, that a HGT event occurred between these two distant lineages. Interestingly, Fer domains from cyanobacterial and stramenopile species branched in the same cluster (Figure 4A). Stramenopiles are eukaryotes that acquired a chloroplast secondarily from Rhodophyta [33]. Therefore, the grouping of viridiplantae, stramenopile and cyanobacterial sequences strongly suggested a cyanobacterial origin of the Fer domain in these two eukaryotic photosynthetic lineages, even if the sequences of the photosynthetic eukaryotes did not appear nested within the cyanobacterial sequences. In fact, this was likely due to a poor resolution of the phylogenetic tree, which is frequent in similar studies of proteins of cyanobacterial origin, where most often only a sister-grouping of cyanobacteria and plant sequences is observed in phylogenetic trees [34]. The hypothesis of an HGT from plants to cyanobacteria can be discarded because the protein is present in *Gloeobacter*, which is a deeply branching cyanobacterial lineage that has diverged before the chloroplastic endosymbiosis and, consequently, before the origin of plants [35]. The HGT of the Fer domain from cyanobacteria to plants is also strongly supported by the functional data showing that the DnaJ-Fer protein is targeted to the chloroplast in the green alga *Chlamydomonas* [24]. It is important to notice that, in contrast with the two-domain DnaJ-Fer proteins of Viridiplantae and Thaumarchaeota, the stramenopile and cyanobacterial proteins were composed uniquely of the Fer domain. Thus, the association between the J and the Fer domains probably occurred in the Viridiplantae lineage after the divergence of the present-day three main Archaeplastida phyla (i.e., Viridiplantae, Rhodophyta and Glaucophyta) but prior to the

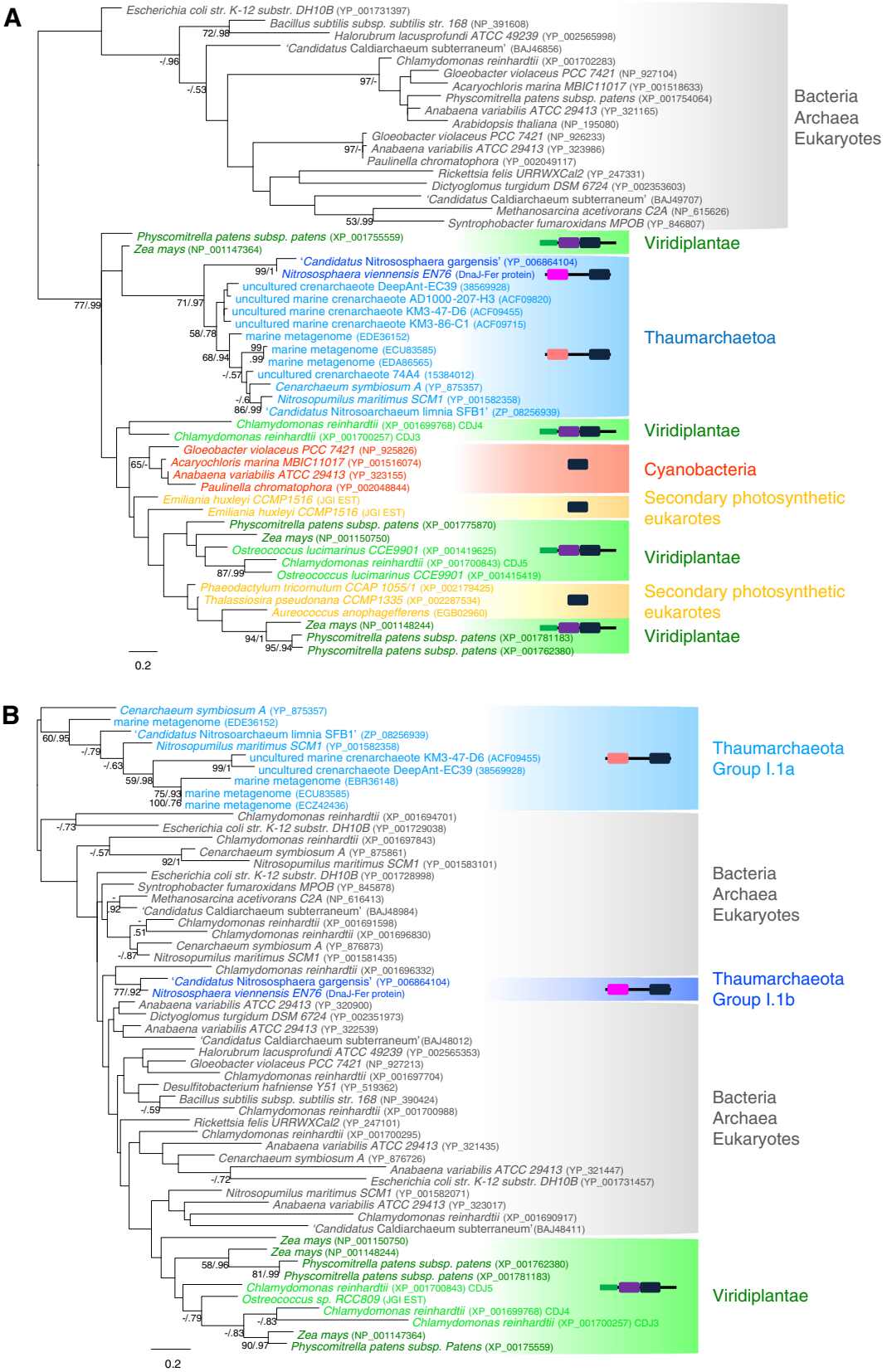


Figure 4 (See legend on next page.)

(See figure on previous page.)
Figure 4 Unrooted ML trees of the Fer and J domains. The ML tree of the Fer domain (**A**) was inferred with 52 sequences and 41 positions, whereas 55 sequences and 40 positions were kept to reconstruct the J domain tree (**B**). The two trees were inferred with TreeFinder (LG model). Numbers at nodes represent bootstrap values and Bayesian posterior probabilities computed with TreeFinder and PhyloBayes, respectively (only values >50% and 0.5 are shown, dashes indicate that the corresponding support is inferior to the threshold, whereas when both supports are inferior to the thresholds no support values are indicated). The scale bars represent the average number of substitutions per site. For clarity, the sequences relevant for the understanding of the history of DnaJ-Fer proteins have been coloured according to their taxonomy.

internal diversification of Viridiplantae (Figure 3A). This phylogeny also supported that the ancestor of the thaumarchaeotal groups I.1a and I.1b acquired secondarily the DnaJ-Fer protein from an ancestor of present-day Viridiplantae (Figure 3A). Another possibility would be that Viridiplantae and Cyanobacteria acquired their Fer domain from Thaumarchaeota. This would imply two HGT events, one from Thaumarchaeota to Cyanobacteria and a second one from Cyanobacteria to photosynthetic eukaryotes through the chloroplast endosymbiosis. In addition, that hypothesis would also imply the dissociation of the Fer and J domains in Cyanobacteria and their reassociation in the Viridiplantae lineage. Therefore, this scenario would require two HGTs as well as two independent associations and one split between the J and Fer domains, what is less parsimonious than the previous one that only requires one association and two HGT events.

Although poorly resolved as in the case of the phylogeny of the Fer domain, the phylogeny of the entire data set of J domain sequences yielded a very different picture. In fact, Viridiplantae and Thaumarchaeota did not cluster together, which was confirmed by a second analysis based on a more restricted sequence sampling. The J domains from the DnaJ-Fer proteins formed three distinct groups (indicated by colours in Figure 4B) scattered among J domain sequences of very different origins (bacterial, eukaryotic and archaeal) and being part of very diverse multidomain proteins. One group contained the J domains from Viridiplantae DnaJ-Fer proteins, another contained those from the group I.1b Thaumarchaeota (i.e., '*Ca*. N. gargensis' and *N. viennensis*), whereas group I.1a Thaumarchaeota emerged in another part of the tree (Figure 4B). This separation in three groups suggested that the J domains of the DnaJ-Fer proteins have different origins. However, this could be due just to the overall poor resolution of the trees. Thus, to discriminate between these two hypotheses (i.e. different origins or lack of phylogenetic signal) we compared the topology of the ML tree with AU tests against four constrained topologies reflecting alternative scenarios for the origin of the DnaJ domain contained in the DnaJ-Fer proteins: 1) the grouping of the J domains of the DnaJ-Fer proteins of the two groups of Thaumarchaeota I.1a and I.1b (Topology 2); 2) the monophyly of these sequences plus the J domains of the DnaJ-Fer proteins of the Viridiplantae (Topology 3); 3) the monophyly of group I.1a Thaumarchaeota and Viridiplantae DnaJ-Fer J domains (Topology 4); and 4) the monophyly of group I.1b Thaumarchaeota and Viridiplantae DnaJ-Fer J domains (Topology 5) (Table 1), the other nodes remaining unchanged. The five topologies were used for the AU test with the alignment of J domain sequences used for the inference of the initial topology (Topology 1). All the four alternative topologies were significantly rejected ($p<0.05$, Table 1), which indicated that the J domains found in the DnaJ-Fer proteins probably have three independent evolutionary origins.

To reconcile this observation with those from the Fer domain (see above), the most parsimonious hypothesis would be that homologous replacements of the J domain occurred twice in Thaumarchaeota after their acquisition of the DnaJ-Fer protein from Viridiplantae (Figure 4B). Such independent homologous replacements could also explain the structural differences observed between the sequences of Viridiplantae and Thaumarchaeota, namely the presence of different central regions separating the J and the Fer domains (large in group I.1a Thaumarchaeota, short in group I.1b Thaumarchaeota, and its absence in Viridiplantae, see above).

The complex evolutionary history of the DnaK/DnaJ/GrpE system in Archaea

Two of the three DnaJ-Fer proteins of *C. reinhardtii* (CDJ3 and CDJ4) have been shown to interact with the chloroplast Hsp70B proteins [24]. These proteins together with their partners, the co-chaperone DnaJ and the nucleotide exchange factor GrpE, are widely distributed in eukaryotes and also in bacteria (where they are encoded in a gene cluster). In contrast, in Archaea, they were initially reported only in lineages of mesophilic and thermophilic euryarchaeota [16,17]. However, at that time the available complete genome sequences were far from covering the whole diversity of the archaeal phyla [25] and many major lineages were not represented. Our survey of about a hundred archaeal genomes now available allowed us confirming the presence of the three genes in all members of the lineages where they were initially reported (Methanobacteriales, Thermoplasmatales, Halobacteriales and Methanosarcinales, Additional file 1) [16,17]. In addition, we also found them in many other major lineages, such as Thaumarchaeota, 'Aigarchaeota', ARMAN group, DHEV2 group, Nanohaloarchaeales, Methanomicrobiales, and Methanocellales (Additional file 1:

Table S1). This increased considerably the diversity of archaea harboring the DnaK system. Worth noting, all these archaea were either mesophilic or thermophilic organisms, underlying the absence of the DnaK/DnaJ/GrpE system in hyperthermophilic archaea. In fact, to the noticeable exception of mesophilic Methanococcales for which we identified DnaK and GrpE homologues only in *Methanococcus vannielii* SB, which were not included in our phylogenetic trees because of their extreme sequence divergence, all mesophilic and thermophilic archaea encoded these three genes and, in most of these archaeal genomes, the three genes were clustered together as occurs in Bacteria (Additional file 1).

In agreement with previous studies [13,14,16,17] our phylogenetic analysis of a subset of 136 sequences representative of the genetic diversity of bacterial, archaeal and eukaryotic DnaK/Hsp70 sequences from complete genomes supported a clear separation between eukaryotic and prokaryotic sequences (BV = 100% and PP = 1.00), and the grouping of mitochondrial and chloroplast sequences with Alphaproteobacteria (BV < 50% and PP < 0.50) and Cyanobacteria (BV = 60% and PP = 1.00), respectively (Additional file 2). By contrast, bacterial and archaeal sequences did not form two separated monophyletic groups but appeared intricately mixed, suggesting that HGT occurred between these two domains of life. To increase the resolution of the evolutionary relationships among prokaryotic DnaK proteins, we reanalysed this dataset after removing the eukaryotic sequences (Figure 5). The monophyly of most bacterial phyla was recovered, often with strong statistical support: Aquificae (BV = 100%; PP = 1.00); Cyanobacteria (BV = 81%; PP = 1.00, a second copy that groups with Deinococcus/Thermus exists in some Cyanobacteria); Actinobacteria (BV = 100%; PP = 1.00); Thermotogae (BV = 100%; PP = 1.00); Dictyoglomi (BV = 100%; PP = 1.00); Deinococcus/Thermus (BV = 98%; PP = 0.82); Spirochaetes (BV = 58%; PP = 0.68); Chlamydiae and Verrucomicrobia (BV < 50%; PP = 0.78); Alpha-, Beta-, Gamma-, Deltaproteobacteria (BV < 50%; PP = 1.00). Similarly, the monophyly of most archaeal orders and classes harbouring a DnaK gene was recovered with high support: Halobacteria and Nanohaloarchaea (BV = 100% ; PP = 1.00); Methanosarcinales (except *Methanosaeta thermophila*, BV = 100%; PP = 1.00); Methanomicrobiales (BV = 100%; PP = 1.00); Methanobacteriales (BV = 100%; PP = 1.00); Methanocellales (BV = 100%; PP = 1.00); ARMAN (BV = 52%; PP = 1.00); Thermoplasmatales (BV = 100%; PP = 1.00) together with *Aciduliprofundum boonei* (BV = 88%; PP = 1.00) and Thaumarchaeota (BV = 95%; PP = 1.00). This indicated the ancestral presence of DnaK in these groups (i.e. prior to their diversification) and that very few HGTs among them occurred after their diversification.

As in most molecular phylogenies, the relationships among bacterial phyla remained mostly unresolved (BV < 50% and PP < 0.95, Figure 5). However, the ancestral presence of DnaK in most of them suggested that this protein was present in the last common ancestor of bacteria. By contrast, the relationships among archaeal orders and classes were well resolved but in strong contradiction with the reference species phylogeny of this domain [25]. To assess the robustness of this contradiction, we tested if the reference archaeal phylogeny was significantly rejected by the archaeal DnaK dataset. For that, the archaeal relationships observed in the DnaK ML tree were compared to those of the archaeal species reference phylogeny [25]. The AU test indicated that the DnaK dataset strongly rejected the reference phylogeny (p = 0.0). In agreement with previous proposals [16,17], this supported the hypothesis that DnaK was not present in the ancestor of Archaea and that it was acquired secondarily by some members of this domain by HGT from bacteria. A careful examination of the DnaK trees suggested that at least two independent inter-domain HGT events occurred: one to the ancestor of Halobacteria and Nanohaloarchaea and another to the ancestor of Class II methanogens (i.e., Methanosarcinales, Methanomicrobiales and Methanocellales [25]). However, the phylogeny of DnaK was not resolved enough to determine without ambiguities the bacterial donors at the origin of these HGT even if, as previously proposed [17], Firmicutes may be a possible donor in the case of the HGT to methanogenic archaea. The initial acquisitions were probably followed by secondary HGTs to and among Methanobacteriales, ARMAN, Thermoplasmatales + DHEV2, '*Ca*. Caldiarchaeum subterraneum' and Thaumarchaeota. Worth noting, the two latter lineages did not form a monophyletic group (Figure 5), in contradiction with the expected species phylogeny [25]. This suggests two independent acquisitions of the gene coding for DnaK from two different euryarchaeotal donors, but the statistical support for the corresponding branches are too low to reach a definitive conclusion. Interestingly, if inter-domain HGTs from Bacteria to Archaea were clearly supported by our analyses, at least one HGT occurred in the opposite direction. This concerned the DnaK of *Elusimicrobium minutum* (Class Elusimicrobia, previously referred as Termite Group 1), which was nested among archaeal sequences with strong support (Figure 5). This ultramicrobacterium was isolated from humivorous beetle larvae and some close relatives have been detected in gut or faeces of termites, cockroaches, and mammals such as chimpanzee, horses or cows [36], environments that are also inhabited by diverse methanogenic archaea. Accordingly, HGT among those microorganisms is not unexpected.

The phylogenies of GrpE and DnaJ were less resolved than that of DnaK, in particular for the deepest nodes (Additional file 3 and Additional file 4), as expected from the smaller number of conserved positions that

Table 1 AU tests of scenarios for the origin of the J domain contained in DnaJ-Fer proteins

Topology	Scenario	*p*-value
1	Viridiplantae, Group I.1a and Group I.1b not monophyletic	0.991
2	Group I.1a and Group I.1b monophyletic	0
3	Viridiplantae, Group I.1a, and Group I.1b monophyletic	0
4	Viridiplantae and Group I.1a monophyletic	0
5	Viridiplantae and Group I.1b monophyletic	0.022

could be kept for phylogenetic analyses (105 and 227 positions, respectively). However, despite the weaker signal, most of the monophyletic groups observed in the DnaK trees were also recovered in the GrpE and DnaJ phylogenies, which suggested that the three proteins have undergone similar evolutionary histories. This agreed with the fact that the corresponding genes are clustered in most prokaryotic genomes (Additional file 1).

Discussion

The unexpected discovery eight years ago of an atypical protein composed of Ferredoxin domain associated to a J domain in the marine archaeal fosmid EC-39 led to the proposal that Thaumarchaeota should have a DnaK protein [18]. This prediction was confirmed a few years later after the identification of genes coding for the DnaK/DnaJ/GrpE system in complete genome sequences of representatives of this phylum [37]. Our phylogenomic analysis showed that the DnaJ-Fer and the DnaK/DnaJ/GrpE proteins have two different origins in this archaeal phylum. The thaumarchaeotal DnaJ-Fer protein resulted from a complex history involving at least two inter-domain HGTs: from Cyanobacteria to Viridiplantae and then from Viridiplantae to Thaumarchaeota, in addition to two independent replacements of the original viridiplantae J domain by J domains of unknown bacterial origin during the diversification of Thaumarchaeota (Figure 3B). By contrast, the phylogenetic analysis of the DnaK, DnaJ and GrpE proteins suggested that these proteins were acquired by the ancestor of Thaumarchaeota by HGT from an unidentified euryarchaeotal donor. The thaumarchaeotal DnaK/DnaJ/GrpE and DnaJ-Fer form therefore a complex chimera, mixing components from bacterial, euryarchaeotal and eukaryotic origin. Identifying the precise functional role of these proteins in Thaumarchaeota will require further investigation.

Due to its large distribution in present day organisms, DnaK was initially proposed as a good phylogenetic marker to infer ancient phylogenies and, more precisely, to decipher the relationships among the three domains of life and their main phyla [38-40]. In particular, the presence of a 23 amino acids deletion shared by Actinobacteria and Firmicutes (referred as Monoderma) and Archaea, but not by Gram negative bacteria (referred as Diderma) and Eucarya was interpreted as the evidence that Archaea derived from Gram positive bacteria [38], dismissing the hypothesis of inter-domain HGT. However, the phylogenetic analysis of DnaK (and of its two partners DnaJ and GrpE) showed later that the evolutionary history of these proteins has been largely affected by HGT. In particular, the strong discrepancies observed between the phylogeny of archaeal DnaK and the species tree indicates that multiple HGTs are responsible of the taxonomic distribution of DnaK in Archaea [17,41]. Thus, the deletion detected by Gupta (which now has been shown to be present also in Thermotogae, Dictyoglomi and Fusobacteria) should not be interpreted as relevant for species phylogeny but just as a strong signature for the gene transfers mentioned above. Therefore, DnaK appears not to be a reliable marker to infer ancient evolutionary relationships, as already suggested in previous works [41].

Phylogenetic and molecular analyses have shown that adaptation to mesophily occurred several times independently during the diversification of Archaea [22,42,43]. Because, DnaK is an important heat shock chaperone involved, among others, in thermal stresses [1], we have postulated previously that its acquisition could have help for the adaptation of archaeal hyperthermophiles to mesophilic environments [18]. Strengthening this hypothesis, DnaK, DnaJ and GrpE are found only in thermophilic and mesophilic archaea which have acquired these genes several times independently, either from bacteria through inter-domain HGTs or from archaea already adapted to mesophilic environments. However, the acquisition of DnaK cannot be considered as obligatory for that adaptation. For example, in the case of Methanococcales, we identified an atypical DnaK gene in *Methanococcus vannielii* SB, but we did not detect it in other mesophilic members of this archaeal order. This illustrated the fact that additional contributing factors, such as protein amino acid composition, are surely important to determine the optimal growth temperature of microorganisms. For instance, always among the Methanococcales, the mesophilic *Methanococcus maripaludis* has been shown to harbour amino acid signatures typical of thermophilic or hyperthermophilic organisms [44], making tempting to speculate that the proteins of this archaeon are intrinsically resistant to heat shocks and, thus, that heat shock chaperones are dispensable in this organism.

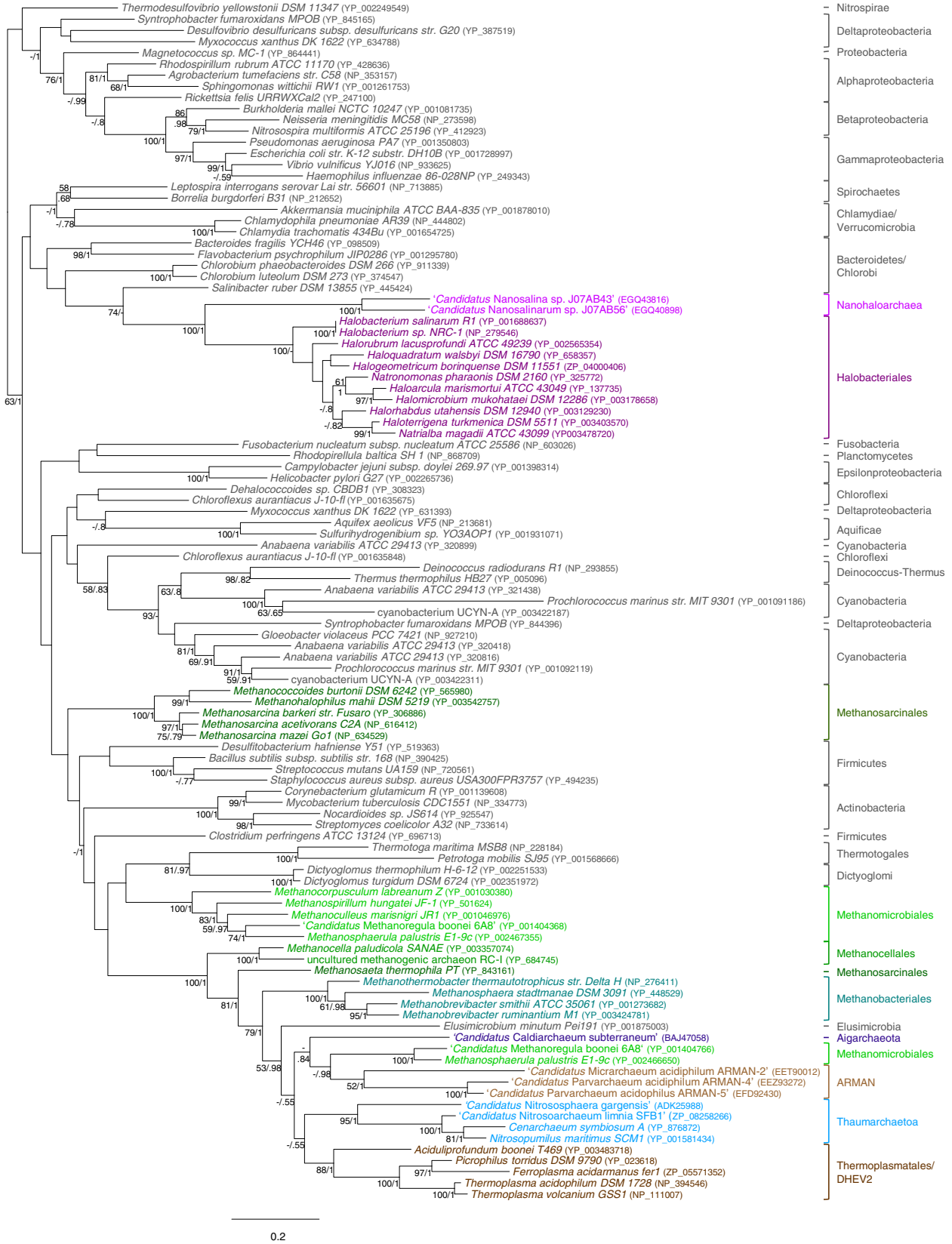


Figure 5 (See legend on next page.)

(See figure on previous page.)
Figure 5 Unrooted ML tree of the DnaK protein. The tree was inferred with TreeFinder with the LG + Γ4 model (107 sequences and 444 positions). Numbers at nodes represent bootstrap values (100 replicates of the original alignment) and Bayesian posterior probabilities computed with TreeFinder and MrBayes, respectively (only values >50% and 0.5 are shown, dashes indicate that the corresponding support is inferior to the threshold, whereas when both supports are inferior to the thresholds no support values are indicated). The scale bar represents the average number of substitutions per site.

Beside those aspects, the evolutionary history of the DnaJ-Fer protein provided an interesting temporal landmark between the Eucarya and Archaea domains. Indeed, the association between the Fer and the J domains composing this protein very likely occurred in an ancestor of the Viridiplantae, before their diversification but after their divergence from the two other Archaeplastida lineages (i.e. the Glaucophyta and the Rhodophyta, which do not have this fused protein) (Figure 3A). Then, the resulting gene was transferred to the ancestor of Thaumarchaeota groups I.1a and I.1b (Figure 3B), more precisely before the divergence of these two lineages but likely after their separation from the HWCG III group (Figure 3A). This indicated that the divergence of the groups I.1a and I.1b is more recent than the divergence between Viridiplantae and the two other Archaeplastida lineages but more ancient than the diversification of Viridiplantae. This illustrates how HGTs can be useful to date evolutionary events relatively against each other [45]. According to fossil record and molecular dating estimates, the divergence of Viridiplantae from the two other Archaeplastida lineages occurred ~950 million years ago whereas the diversification of Viridiplantae started ~750 million years ago [46]. The HGT from Viridiplantae to Thaumarchaeota occurred most likely during this time window, so the divergence of the groups I.1a and I.1b Thaumarchaeota and their diversification could be less than ~950 million years old.

Conclusions

Phylogenomic analysis supports that the proteins DnaK, DnaJ, GrpE and DnaJ-Fer have a chimerical origin in Thaumarchaeota, which acquired them by HGT from different donors, including bacterial and eukaryotic species. Similar HGT events have occurred independently in other archaeal groups. This suggests that the acquisition of these proteins has probably played an important role in the convergent adaptation of these archaea to mesophilic and thermophilic lifestyles from their hyperthermophilic ancestors. In addition, these HGT events can be used as markers for the relative dating of the diversification of donor and acceptor groups as, for example, the Thaumarchaeota, which have received their DnaJ-Fer protein from Archaeplastida.

Methods

Dataset assembly

The DnaJ-Fer protein homologues were retrieved from the non-redundant (nr) and the environmental databases at the NCBI (http://www.ncbi.nlm.nih.gov) with the BlastP program (default parameters) [47] using as seeds the DnaJ-Fer protein from the uncultured archaeon DeepAnt-EC39 fosmid (AY316120.1), and the sequences of *Chlamydomonas reinhardtii* CDJ3, 4 and 5 (XP_001700257.1, XP_001699768.1 and XP_001700843.1 respectively). The DnaJ-Fer sequence from *Nitrososphaera viennensis* was kindly provided by C. Schleper and A. Spang. To ensure the exhaustive retrieval of eukaryote sequences we queried EST and ongoing genome project databases: the JGI (http://genome.jgi-psf.org/) for *Ostreococcus* sp. RCC809, *Emiliania huxleyi*, *Chlorella vulgaris* and *Chlorella* sp. NC64A; the TIGR (http://plantta.jcvi.org/) for *Pinus taeda*; the *Cyanidioschyzon merolae* Genome Project (http://merolae.biol.s.u-tokyo.ac.jp/), and the *Galdieria sulphuraria* Genome Project (http://genomics.msu.edu/galdieria/about.html). The absence of DnaJ-Fer homologues in any archaeal or eukaryotic complete genome was verified by tBlastN searches against the corresponding nucleic acid sequence. The presence of the J and the Fer domains in the retrieved homologues was systematically verified using Pfam (Pfam profiles PF00226 and PF13459, respectively). The J and Fer domains were then analysed separately using the same strategy as previously to retrieve proteins containing these domains.

DnaK, GrpE and DnaJ homologues were retrieved from 92 archaeal complete genome sequences available at NCBI with BlastP (default parameters) using the sequences from *N. maritimus* as seeds (YP_001581434, YP_001581433.1 and YP_001581435, respectively). The absence of homologues in any genome was systematically verified by tBlastN searches against the corresponding nucleic acid sequence. Eukaryotic and bacterial homologues were retrieved from a subset of four and 86 complete genomes representative of the taxonomic diversity of these two domains using BlastP (default parameters). In the case of DnaJ, we checked the domain composition of the retrieved homologues with PFAM in order to distinguish *bona fide* DnaJ proteins (harbouring a J-domain (PF00226), the cysteine rich central domain (PF00684) and the C-terminal domain (PF01556)) from other J-proteins.

We thus obtained six different sequence datasets, and we tested various programs to align them, including (Mafft v6.833b [48], Probcons v1.12 [49], and Muscle v3.6 [50]). The quality of the resulting alignments was visually inspected in order to keep those for which the residues of the conserved domains were correctly aligned.

Probcons provided better results for the DnaJ-Fer, the J domain and the Fer domain datasets, whereas Mafft provided better results in the case of the DnaK, GrpE and DnaJ datasets. The selected alignments were edited and manually refined with the program ED of the MUST package [51]. The regions where the alignment was ambiguous were removed using the NET program from the MUST package.

Phylogenetic reconstruction

The DnaJ-Fer, DnaK, GrpE and DnaJ alignments were analysed by maximum likelihood (ML) and Bayesian approaches (BI). For each dataset, the LG model was proposed as the best suited evolutionary model according to the "propose model tool" of TreeFinder v2011 [52] with the AICc criterion. Alternative models (e.g. WAG, JTT, etc.) were also tested. The resulting trees were consistent with those inferred with the LG model (not shown).

ML tree reconstructions were performed using PhyML v3.0 [53] and TreeFinder v2011 [52]. The robustness of the resulting trees was estimated by the non-parametric procedure implemented in PhyML and TreeFinder (100 replicates of the original dataset). The resulting trees were very similar, so we decided to show only the ML trees inferred with TreeFinder.

BI of DnaJ-Fer, DnaK, DnaJ and GrpE proteins was carried out with PhyloBayes v 3.3 with the LG model and a gamma distribution of substitution rates with four categories [54]. Phylobayes was run with four independent chains for at least 10,000 cycles, saving one tree in ten. The first 300 trees were discarded as "burn-in", and the remaining trees from each chain were used to test for convergence and compute the 50% majority rule consensus tree. In the case of DnaK, the chains did not converge even after 10,000 cycles. Therefore, BI trees for this marker were computed with MrBayes v.3.0b4 [55] with a mixed substitution model and a Gamma distribution of substitution rates with 4 categories. Searches were run with 4 chains of 1,000,000 generations for which the first 1,000 generations were discarded as “burn in”, trees being sampled every 100 generations.

The analyses of the J and Fer domains were divided in two steps. First, all the homologous sequences were analysed by neighbor-joining (NJ) using the MUST package [51]. Based on this preliminary phylogenetic tree, we selected the closest homologues of the DeepAnt-EC39 fosmid and *C. reinhardtii* sequences and a subset of more distantly related homologues representative of the genetic diversity of these domains. These sequences were used to carry out ML and BI analysis with TreeFinder, PhyML and PhyloBayes as previously described.

The comparison of different tree topologies was done by applying the Approximately Unbiased test [56] implemented in TreeFinder with the same evolutionary models and parameters as for ML phylogenetic inference.

Additional files

Additional file 1: Table showing the taxonomic distribution of DnaJ-Fer, DnaK, DnaJ and GrpE proteins in Archaea. Numbers correspond to accession numbers in the NCBI Genpep database in the 92 complete genomes available in July 2011 and that of, '*Ca.* Nitrosoarchaeum koreensis', a thaumarchaeotal genome available more recently. The two divergent sequences of DnaK and GrpE found in *Methanococcus vannielii* SB are underlined.

Additional file 2: Unrooted ML tree of the DnaK protein (136 sequences and 444 positions) inferred with TreeFinder and the LG + Γ4 model. Numbers at nodes represent bootstrap values and Bayesian posterior probabilities computed with TreeFinder and MrBayes, respectively (only values >50% and 0.5 are shown, dashes indicate the corresponding support is inferior to the threshold, whereas when both supports are inferior to the thresholds no support values are indicated). Archaeal sequences are shown with colours according to their taxonomic classification. The scale bar represents the average number of substitutions per site.

Additional file 3: Unrooted ML tree of the DnaJ protein (102 sequences and 227 positions) inferred with TreeFinder and the LG + Γ4. Numbers at nodes represent bootstrap values and Bayesian posterior probabilities computed with TreeFinder and MrBayes, respectively (only values >50% and 0.5 are shown, dashes indicate that the corresponding support is inferior to the threshold, whereas when both supports are inferior to the thresholds no support values are indicated). Archaeal sequences are shown with colours according to their taxonomic classification. The scale bar represents the average number of substitutions per site.

Additional file 4: Unrooted ML tree of the GrpE protein (101 sequences and 105 positions) inferred with TreeFinder and the LG + Γ4. Numbers at nodes represent bootstrap values and Bayesian posterior probabilities computed with TreeFinder and MrBayes, respectively (only values >50% and 0.5 are shown, dashes indicate the corresponding value is inferior to the threshold, whereas when both supports are inferior to the thresholds no support values are indicated). Archaeal sequences are shown with colours according to their taxonomic classification. The scale bar represents the average number of substitutions per site.

Competing interests

The authors have declared that no competing interests exist.

Authors' contributions

CB-A and DM conceived this study. CP, CB-A and DM designed and carried out the phylogenetic analyses. CP, CB-A, PL-G and DM wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements

This work was supported by the Agence Nationale de la Recherche (ANR EvolDeep; contract number ANR-08-GENM-024-002). C. Petitjean was the recipient of a grant from the ANR EvolDeep. C. Brochier-Armanet was member of the Institut Universitaire de France, and was funded by an ATIP from the Centre National de la Recherche Scientifique (CNRS) and by the Ancestrome project (ANR-10-BINF-01-01) and by the ANR-10-BINF-01-01 "Ancestrome" grant. We acknowledge C. Schleper, A. Spang and J. de la Torre for sharing unpublished data.

Author details

[1]UPR CNRS 9043, Laboratoire de Chimie Bactérienne, Université d'Aix-Marseille (AMU), 13402 Marseille, Cedex 20, France. [2]UMR CNRS 8079, Unité d'Ecologie, Systématique et Evolution Université Paris-Sud, 91405 Orsay, Cedex, France. [3]CNRS, UMR5558, Laboratoire de Biométrie et Biologie Evolutive, Université de Lyon, Université Lyon 1, 43 boulevard du 11 novembre 1918, 69622, Villeurbanne, France.

Received: 19 March 2012 Accepted: 25 October 2012
Published: 26 November 2012

References

1. Mayer MP, Bukau B: **Hsp70 chaperones: cellular functions and molecular mechanism.** *Cell Mol Life Sci* 2005, **62**(6):670–684.
2. Young JC: **Mechanisms of the Hsp70 chaperone system.** *Biochem Cell Biol* 2010, **88**(2):291–300.
3. Morano KA: **New tricks for an old dog: the evolving world of Hsp70.** *Ann N Y Acad Sci* 2007, **1113:**1–14.
4. Kampinga HH, Craig EA: **The HSP70 chaperone machinery: J proteins as drivers of functional specificity.** *Nat Rev Mol Cell Biol* 2010, **11**(8):579–592.
5. Harrison C: **GrpE, a nucleotide exchange factor for DnaK.** *Cell Stress Chaperones* 2003, **8**(3):218–224.
6. Laloraya S, Gambill BD, Craig EA: **A role for a eukaryotic GrpE-related protein, Mge1p, in protein translocation.** *Proc Natl Acad Sci USA* 1994, **91**(14):6481–6485.
7. Schroda M, Vallon O, Whitelegge JP, Beck CF, Wollman FA: **The chloroplastic GrpE homolog of Chlamydomonas: two isoforms generated by differential splicing.** *Plant Cell* 2001, **13**(12):2823–2839.
8. Alberti S, Esser C, Hohfeld J: **BAG-1–a nucleotide exchange factor of Hsc70 with multiple cellular functions.** *Cell Stress Chaperones* 2003, **8**(3):225–231.
9. Craig EA, Huang P, Aron R, Andrew A: **The diverse roles of J-proteins, the obligate Hsp70 co-chaperone.** *Rev Physiol Biochem Pharmacol* 2006, **156:**1–21.
10. Liberek K, Marszalek J, Ang D, Georgopoulos C, Zylicz M: **Escherichia coli DnaJ and GrpE heat shock proteins jointly stimulate ATPase activity of DnaK.** *Proc Natl Acad Sci USA* 1991, **88**(7):2874–2878.
11. Walsh P, Bursac D, Law YC, Cyr D, Lithgow T: **The J-protein family: modulating protein assembly, disassembly and translocation.** *EMBO Rep* 2004, **5**(6):567–571.
12. Qiu XB, Shao YM, Miao S, Wang L: **The diversity of the DnaJ/Hsp40 family, the crucial partners for Hsp70 chaperones.** *Cell Mol Life Sci* 2006, **63**(22):2560–2570.
13. Boorstein WR, Ziegelhoffer T, Craig EA: **Molecular evolution of the HSP70 multigene family.** *J Mol Evol* 1994, **38**(1):1–17.
14. Renner T, Waters ER: **Comparative genomic analysis of the Hsp70s from five diverse photosynthetic eukaryotes.** *Cell Stress Chaperones* 2007, **12**(2):172–185.
15. Nordhues A, Miller SM, Muhlhaus T, Schroda M: **New insights into the roles of molecular chaperones in Chlamydomonas and Volvox.** *Int Rev Cell Mol Biol* 2010, **285:**75–113.
16. Macario AJ, Brocchieri L, Shenoy AR, Conway de Macario E: **Evolution of a protein-folding machine: genomic and evolutionary analyses reveal three lineages of the archaeal hsp70(dnaK) gene.** *J Mol Evol* 2006, **63**(1):74–86.
17. Gribaldo S, Lumia V, Creti R, de Macario EC, Sanangelantoni A, Cammarano P: **Discontinuous occurrence of the hsp70 (dnaK) gene among Archaea and sequence features of HSP70 suggest a novel outlook on phylogenies inferred from this protein.** *J Bacteriol* 1999, **181**(2):434–443.
18. Lopez-Garcia P, Brochier C, Moreira D, Rodriguez-Valera F: **Comparative analysis of a genome fragment of an uncultivated mesopelagic crenarchaeote reveals multiple horizontal gene transfers.** *Environ Microbiol* 2004, **6**(1):19–34.
19. DeLong EF: **Archaea in coastal marine environments.** *Proc Natl Acad Sci USA* 1992, **89**(12):5685–5689.
20. Fuhrman JA, McCallum K, Davis AA: **Novel major archaebacterial group from marine plankton.** *Nature* 1992, **356**(6365):148–149.
21. Brochier-Armanet C, Boussau B, Gribaldo S, Forterre P: **Mesophilic Crenarchaeota: proposal for a third archaeal phylum, the Thaumarchaeota.** *Nat Rev Microbiol* 2008, **6**(3):245–252.
22. Brochier-Armanet C, Gribaldo S, Forterre P: **Spotlight on the Thaumarchaeota.** *ISME J* 2012, **6**(2):227–230.
23. Pester M, Schleper C, Wagner M: **The Thaumarchaeota: an emerging view of their phylogeny and ecophysiology.** *Curr Opin Microbiol* 2011, **14**(3):300–306.
24. Dorn KV, Willmund F, Schwarz C, Henselmann C, Pohl T, Hess B, Veyel D, Usadel B, Friedrich T, Nickelsen J, *et al*: **Chloroplast DnaJ-like proteins 3 and 4 (CDJ3/4) from Chlamydomonas reinhardtii contain redox-active Fe-S clusters and interact with stromal HSP70B.** *Biochem J* 2010, **427**(2):205–215.
25. Brochier-Armanet C, Forterre P, Gribaldo S: **Phylogeny and evolution of the Archaea: one hundred genomes later.** *Curr Opin Microbiol* 2011, **14**(3):274–281.
26. Adl SM, Simpson AG, Farmer MA, Andersen RA, Anderson OR, Barta JR, Bowser SS, Brugerolle G, Fensome RA, Fredericq S, *et al*: **The new higher level classification of eukaryotes with emphasis on the taxonomy of protists.** *J Eukaryot Microbiol* 2005, **52**(5):399–451.
27. Hallam SJ, Konstantinidis KT, Putnam N, Schleper C, Watanabe Y, Sugahara J, Preston C, de la Torre J, Richardson PM, DeLong EF: **Genomic analysis of the uncultivated marine crenarchaeote Cenarchaeum symbiosum.** *Proc Natl Acad Sci USA* 2006, **103**(48):18296–18301.
28. Walker CB, de la Torre JR, Klotz MG, Urakawa H, Pinel N, Arp DJ, Brochier-Armanet C, Chain PS, Chan PP, Gollabgir A, *et al*: **Nitrosopumilus maritimus genome reveals unique mechanisms for nitrification and autotrophy in globally distributed marine crenarchaea.** *Proc Natl Acad Sci USA* 2010, **107**(19):8818–8823.
29. Blainey PC, Mosier AC, Potanina A, Francis CA, Quake SR: **Genome of a Low-Salinity Ammonia-Oxidizing Archaeon Determined by Single-Cell and Metagenomic Analysis.** *PLoS One* 2011, **6**(2):e16626.
30. Kim BK, Jung MY, Yu DS, Park SJ, Oh TK, Rhee SK, Kim JF: **Genome sequence of an ammonia-oxidizing soil archaeon, "Candidatus Nitrosoarchaeum koreensis" MY1.** *J Bacteriol* 2011, **193**(19):5539–5540.
31. Spang A, Poehlein A, Offre P, Zumbragel S, Haider S, Rychlik N, Nowka B, Schmeisser C, Lebedeva EV, Rattei T *et al*: **The genome of the ammonia-oxidizing Candidatus Nitrososphaera gargensis: insights into metabolic versatility and environmental adaptations.** *Environmental microbiology* 2012.
32. Nunoura T, Takaki Y, Kakuta J, Nishi S, Sugahara J, Kazama H, Chee GJ, Hattori M, Kanai A, Atomi H, *et al*: **Insights into the evolution of Archaea and eukaryotic protein modifier systems revealed by the genome of a novel archaeal group.** *Nucleic Acids Res* 2011, **39**(8):3204–3223.
33. Keeling PJ: **The endosymbiotic origin, diversification and fate of plastids.** *Philos Trans R Soc Lond B Biol Sci* 2010, **365**(1541):729–748.
34. Deschamps P, Moreira D: **Signal conflicts in the phylogeny of the primary photosynthetic eukaryotes.** *Mol Biol Evol* 2009, **26**(12):2745–2753.
35. Criscuolo A, Gribaldo S: **Large-scale phylogenomic analyses indicate a deep origin of primary plastids within cyanobacteria.** *Mol Biol Evol* 2011, **28**(11):3019–3032.
36. Geissinger O, Herlemann DP, Morschel E, Maier UG, Brune A: **The ultramicrobacterium "Elusimicrobium minutum" gen. nov., sp. nov., the first cultivated representative of the termite group 1 phylum.** *Appl Environ Microbiol* 2009, **75**(9):2831–2840.
37. Spang A, Hatzenpichler R, Brochier-Armanet C, Rattei T, Tischler P, Spieck E, Streit W, Stahl DA, Wagner M, Schleper C: **Distinct gene set in two different lineages of ammonia-oxidizing archaea supports the phylum Thaumarchaeota.** *Trends Microbiol* 2010, **18**(8):331–340.
38. Gupta RS: **What are archaebacteria: life's third domain or monoderm prokaryotes related to Gram-positive bacteria? A new proposal for the classification of prokaryotic organisms.** *Mol Microbiol* 1998, **229**(3):695–708.
39. Griffiths E, Gupta RS: **The use of signature sequences in different proteins to determine the relative branching order of bacterial divisions: evidence that Fibrobacter diverged at a similar time to Chlamydia and the Cytophaga-Flavobacterium-Bacteroides division.** *Microbiology* 2001, **147**(Pt 9):2611–2622.
40. Gupta RS: **Origin of diderm (Gram-negative) bacteria: antibiotic selection pressure rather than endosymbiosis likely led to the evolution of bacterial cells with two membranes.** *Antonie Van Leeuwenhoek* 2011, **100**(2):171–182.
41. Philippe H, Budin K, Moreira D: **Horizontal transfers confuse the prokaryotic phylogeny based on the HSP70 protein family.** *Mol Microbiol* 1999, **31**(3):1007–1009.
42. Gribaldo S, Brochier-Armanet C: **The origin and evolution of Archaea: a state of the art.** *Philos Trans R Soc Lond B Biol Sci* 2006, **361**(1470):1007–1022.
43. Groussin M, Gouy M: **Adaptation to environmental temperature is a major determinant of molecular evolutionary rates in archaea.** *Mol Biol Evol* 2011, **28**(9):2661–2674.
44. Puigbo P, Pasamontes A, Garcia-Vallve S: **Gaining and losing the thermophilic adaptation in prokaryotes.** *Trends in genetics: TIG* 2008, **24**(1):10–14.
45. Huang J, Gogarten JP: **Ancient horizontal gene transfer can benefit phylogenetic reconstruction.** *Trends in genetics: TIG* 2006, **22**(7):361–366.
46. Douzery EJ, Snell EA, Bapteste E, Delsuc F, Philippe H: **The timing of eukaryotic evolution: does a relaxed molecular clock reconcile proteins and fossils?** *Proc Natl Acad Sci USA* 2004, **101**(43):15386–15391.
47. 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**(17):3389–3402.
48. Katoh K, Misawa K, Kuma K, Miyata T: **MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform.** *Nucleic Acids Res* 2002, **30**(14):3059–3066.

49. Do CB, Mahabhashyam MS, Brudno M, Batzoglou S: **ProbCons: Probabilistic consistency-based multiple sequence alignment.** *Genome Res* 2005, **15**(2):330–340.
50. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput.** *Nucleic Acids Res* 2004, **32**(5):1792–1797.
51. Philippe H: **MUST, a computer package of Management Utilities for Sequences and Trees.** *Nucleic Acids Res* 1993, **21**(22):5264–5272.
52. Jobb G, von Haeseler A, Strimmer K: **TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics.** *BMC Evol Biol* 2004, **4**:18.
53. Guindon S, Gascuel O: **A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood.** *Syst Biol* 2003, **52**(5):696–704.
54. Lartillot N, Lepage T, Blanquart S: **PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating.** *Bioinformatics* 2009, **25**(17):2286–2288.
55. Ronquist F, Huelsenbeck JP: **MrBayes 3: Bayesian phylogenetic inference under mixed models.** *Bioinformatics* 2003, **19**(12):1572–1574.
56. Shimodaira H: **An approximately unbiased test of phylogenetic tree selection.** *Syst Biol* 2002, **51**(3):492–508.

doi:10.1186/1471-2148-12-226
Cite this article as: Petitjean *et al.*: Horizontal gene transfer of a chloroplast DnaJ-Fer protein to Thaumarchaeota and the evolutionary history of the DnaK chaperone system in Archaea. *BMC Evolutionary Biology* 2012 **12**:226.

C. Eléments de discussion

La masse de données générées par notre analyse de l'ensemble des protéines codées dans les génomes de *C. symbiosum*, *N. maritimus* et '*Ca.* Caldiarchaeum subterraneum' n'a pas encore été étudiée et, de par son ampleur, pourrait faire l'objet d'un travail à part entière. Le cas de la protéine DnaJ-Fer n'est qu'un premier essai d'exploitation de cette information. Parmi les 3085 protéines dont la phylogénie a pu être inférée et observée, j'ai extrait quelques observations qui pourraient être exploitées. Néanmoins, je voudrais insister sur le fait que, d'une part, les jeux de données construits pour chaque protéine contiennent au plus les 200 séquences les plus proches retrouvées par BLAST. Dans certains cas, la présence de certains homologues plus lointains (particulièrement bactériens ou eucaryotes) n'a donc pu être détectée. D'autre part, l'analyse de ces données avait pour but premier la recherche de nouveaux marqueurs pour la phylogénie des archées et non une analyse rigoureuse de génomique comparative d'archées. Tout élément ayant retenu mon attention sur une phylogénie a été noté, et ce de la façon la plus systématique et standardisée possible. Enfin, un premier tri a été effectué entre les génomes de *C. symbiosum* et *N. maritimus* pour éviter une trop grande redondance, mais, par précaution et pour ne rater aucune protéine intéressante, nous avons préféré garder le génome de '*Ca.* Caldiarchaeum subterraneum' dans son ensemble quitte à analyser la phylogénie de la même famille protéique plusieurs fois, les jeux de données redondants ayant été supprimés dans la deuxième étape de sélection des marqueurs. Il existe donc une redondance entre certains jeux de données analysés (mais qui a été éliminée pour les analyses visant à étudier la phylogénie des archées et la place de la racine). Les chiffres donnés sont donc à considérer avec précaution et doivent être pris comme un ordre de grandeur et non pas comme un chiffre définitif.

On trouve ainsi un peu moins de 800 protéines pour lesquelles aucun homologue bactérien ou eucaryote n'était présent, dont 135 spécifiques aux thaumarchées et 16 spécifiques à '*Ca.* Caldiarchaeum subterraneum'. Ces protéines spécifiques aux archées au premier abord pourraient faire l'objet d'une étude et apporter de nombreux éléments quant à l'évolution du domaine Archaea et aux innovations évolutives apparues spécifiquement dans cette lignée. Parallèlement, 650 protéines sont présentes chez au moins un représentant de chaque phylum archée (Euryarchaeota, Crenarchaeota, Aigarchaeota, Thaumarchaeota, et Korarchaeota). Nous avons aussi observé la présence d'euryarchées et de crenarchées dans environ deux tiers des jeux de données, la présence de '*Ca.* Korarchaeum cryptophilum', d'une part, et des thaumarchées et de l'aigarchée, d'autre part, dans environ un tiers de ces jeux de données. Ces protéines partagées entre archées peuvent être la cible d'une étude du contenu en gènes et de la nature du dernier ancêtre commun aux archées (« LACA » pour « Last Archaeal Common Ancestor »). En effet, la présence d'une protéine dans au

moins deux phyla, dont notamment chez les euryarchées et au moins un des autres phyla (cf. Chapitre 2), pourrait être due à sa présence chez LACA.

De la même manière on retrouve environ un tiers des jeux de données dans lesquels des homologues bactériens et eucaryotes sont présents. Ces protéines pourraient être des cibles privilégiées pour une étude sur le dernier ancêtre commun de tous les organismes et sur les caractères conservés depuis cet ancêtre chez les archées. En ce qui concerne particulièrement les relations entre les archées et les deux autres domaines du vivant, environ 300 jeux de données contiennent des eucaryotes mais pas de bactéries ; en parallèle, environ un autre tiers des jeux de données contiennent des homologues bactériens sans eucaryotes. L'analyse des phylogénies de ces protéines pourrait être un point de départ pour analyser les relations entre les archées et les deux autres domaines du vivant, ensemble ou indépendamment.

Il est important de souligner que la présence d'un homologue bactérien, eucaryote ou d'un autre phylum d'archées n'est pas forcement le signe de la présence de cette protéine dans l'ancêtre commun de ces deux lignées, mais peut être aussi le résultat d'un transfert horizontal. Lors de ma première analyse, j'ai noté tous les cas potentiels de transfert mais il serait nécessaire de mener une nouvelle étude de ces jeux de données pour les valider. Seul un petit nombre de transferts entre différentes lignées d'archées a été considéré dans cette première analyse car il est souvent très difficile de faire la différence entre un transfert horizontal entre archées, un cas de paralogie cachée ou un manque de résolution dans une phylogénie générée automatiquement et sans analyse poussée. C'est pourquoi je préconiserais, pour étudier ces transferts, de réanalyser les jeux de données dans lesquels plusieurs phyla d'archées sont présents mais qui n'ont pas été conservés pour l'analyse de la phylogénie des archées. Une nouvelle analyse, avec un échantillonnage taxonomique différent centré sur les archées et avec toutes les informations apportées dans ma thèse concernant la phylogénie des archées et la position de la racine de ce domaine, pourrait permettre de faire un tri plus efficace et de détecter réellement les transferts entre archées. Un problème similaire se pose en ce qui concerne les transferts horizontaux entre eucaryotes et archées. Le débat existant autour de la relation entre archées et eucaryotes et les hypothèses proposant que les eucaryotes soient en réalité issus de lignées d'archées incitent à prendre avec beaucoup de précaution les observations de transferts horizontaux potentiels entre ces deux domaines. La présence d'homologues bactériens dans la plupart des phylogénies avec des homologues eucaryotes (environ 1000 sur 1300) peut rendre plus difficile l'inférence de la position des eucaryotes de par leur importante distante évolutive (i.e. générant de longues branches) et en diminuant probablement la résolution de ces phylogénies. De même que pour les relations entre phyla d'archées, une analyse centrée sur les relations entre eucaryotes et archées avec un échantillonnage taxonomique adapté serait nécessaire

pour conclure quant aux cas de transferts entre archées et eucaryotes. Les mêmes risques existent en ce qui concerne les transferts horizontaux entre archées et bactéries et les mêmes recommandations sont donc à préconiser pour leur analyse, bien que la polémique en ce qui concerne les relations entre bactéries et archées soit moins importante actuellement. Dans mon analyse initiale, j'ai noté plus de 350 cas de transferts potentiels entre archées et bactéries (le sens de ces transferts devant être analysé en détails) dont un grand nombre impliquant les thaumarchées. Il a été proposé que l'adaptation des thaumarchées mésophiles et marines à leur environnement ait impliqué de nombreux transferts depuis des bactéries vivant dans ces même milieux (López-garcía et al. 2004; Brochier-Armanet et al. 2011), l'analyse de ces phylogénies pourrait apporter de nouvelles informations quant à ces transferts et leur rôle dans cette adaptation.

La liste de l'ensemble des protéines de départ ainsi que quelques informations sur la répartition taxonomique des homologues présents dans les jeux de données résultants et sur l'analyse des phylogénies sont disponibles dans l'Annexe 2.

Discussion

Mon travail de thèse porte sur la phylogénie des archées avec des approches de phylogénomique. Il a permis de mettre en évidence un nombre relativement important de nouveaux marqueurs utilisables pour reconstruire la phylogénie des archées et améliorer sa résolution. Dans cette discussion, j'aimerais aborder certains points qui me semblent importants dans le travail autour de la phylogénomique, particulièrement lorsque celle-ci est appliquée au monde microbien.

De l'importance de la phylogénie en biologie.

La très célèbre phrase de Dobzhansky « *Nothing in biology makes sense except in the light of evolution* » (Dobzhansky 1973)[1] est une introduction parfaite à l'importance de la phylogénie en biologie. J'ai déjà parlé de l'importance de l'étude de la phylogénie des archées pour la compréhension de leur histoire évolutive ainsi que des relations qui les lient aux deux autres domaines du vivant. D'un point de vue géologique aussi, la connaissance de l'histoire évolutive des organismes est importante, par exemple l'apparition de la méthanogenèse chez les archées qui a pu modifier les conditions atmosphériques. La reconstruction de la phylogénie des organismes est aussi extrêmement importante pour comprendre les mécanismes évolutifs sous-tendant leur évolution. En effet, l'utilisation d'une phylogénie des organismes comme référence est essentielle pour comprendre l'histoire de leurs caractères, gènes ou systèmes. Par exemple, le Chapitre 3 de ma thèse présente l'histoire évolutive de quatre protéines qui ont subi de nombreux transferts horizontaux entre les trois domaines du vivant. Les protéines DnaJ et DnaJ-Fer font partie d'une famille protéique extrêmement large composée de protéines contenant un domaine protéique dnaJ. Cette famille est le résultat de multiples événements de duplications et de pertes de gènes, mais aussi de fusion avec d'autres gènes, comme dans le cas de la protéine DnaJ-Fer. Ces différents événements évolutifs particuliers, et rares pour certains, n'ont pu être mis en évidence que par la comparaison de la phylogénie de ces protéines individuelles avec une phylogénie de référence, inférée à partir de multiples marqueurs, telle que celle que nous avons construite dans le Chapitre 1. L'on pourra objecter que cette phylogénie des organismes dite « de référence », n'est pas totalement

[1] Si cet article de Dobzhansky est très important, pour l'éloquence de son titre mais aussi pour son explication de ce qu'est l'évolution biologique et de son message clairement opposé à l'antiévolutionnisme, je pense que sa lecture attentive est importante. Il me semble empreint d'une certaine forme de gradisme, mettant l'homme « à l'apex de l'évolution », et d'une volonté de réconcilier science et religion alors que ces deux domaines n'ont ni à être opposés ni à être réconciliés, puisque ne traitant pas de la même chose. Je tenais à souligner ce point à propos d'un article cité si souvent en biologie (y compris ici).

résolue et l'existence même d'une telle phylogénie est remise en question très souvent (E Bapteste et al. 2005; Dagan and Martin 2006). La découverte de nouvelles espèces et le séquençage de nouveaux génomes viennent forcément modifier la phylogénie, et de nouvelles méthodes permettent de corriger certains biais ; en revanche nous avons montré dans le Chapitre 1 que si certains points sont toujours discutés, le signal global de la phylogénie des archées reste relativement stable depuis de nombreuses années. La phylogénie globale proposée pourra donc être considérée comme une phylogénie de référence des organismes, jusqu'à la prochaine amélioration de celle-ci.

De la sélection des données.

Mes travaux de recherche montrent qu'un grand nombre de nouveaux marqueurs phylogénétiques ont pu être identifiés pour inférer la phylogénie des archées. Ainsi nous avons presque triplé le nombre de marqueurs utilisables pour aborder cette question. Même si les 200 protéines utilisées ici ne sont pas présentes chez toutes les archées, elles contiennent un signal phylogénétique suffisant pour reconstruire l'histoire évolutive de ce domaine. Les supermatrices construites à partir de ces 200 protéines, et avec la totalité des 273 marqueurs en incluant les protéines ribosomiques, l'ARN polymérase et les facteurs de transcription conservés à l'échelle des archées, permettent de clarifier certaines relations en débat et d'apporter un meilleur soutien général à la phylogénie de ce domaine. Ces nouvelles protéines peuvent avoir des répartitions taxonomiques très différentes et elles ne sont pas forcément présentes dans la totalité des archées. Elles ont été sélectionnées sur la base de leur présence dans au moins dix espèces d'archées représentant deux phyla différents ainsi que sur la base de la monophylie des séquences d'archées et des différents ordres d'archées. La présence de ces protéines chez des bactéries ou chez des eucaryotes n'a eu aucun impact sur leur choix comme marqueurs pour la phylogénie des archées. En effet, cet ensemble de protéines a été sélectionné exclusivement pour aborder la question de la phylogénie des archées et leur analyse phylogénétique a conduit à des résultats très concluants.

Les données collectées m'ont ensuite permis d'aborder la question de la position de la racine de l'arbre des archées. Compte-tenu des possibles liens qui unissent eucaryotes et archées, je me suis orientée vers l'utilisation d'un groupe extérieur bactérien. Comme dans le cas précédent, nous avons cherché le plus grand nombre possible de protéines pouvant répondre spécifiquement à cette question. L'analyse minutieuse des phylogénies individuelles des 200 marqueurs identifiés précédemment m'a permis de mettre en évidence que 38 d'entre eux présentaient les qualités requises pour étudier la position de la racine chez les archées auxquels il a été possible de rajouter 32 protéines ribosomiques. Ici encore, une sélection de marqueurs adaptés a été nécessaire pour aborder une nouvelle problématique précise et obtenir des résultats robustes.

Pour ces deux études, nous avons sélectionné des ensembles de marqueurs différents pour répondre à des questions différentes. De plus, cette sélection s'est faite uniquement sur la base de critères phylogénétiques, sans prendre en compte a priori la fonction des gènes. En ce sens ma démarche est très différente des études précédentes, qui étaient centrées sur l'analyse d'un petit nombre de protéines informationnelles, principalement les protéines ribosomiques. Certains travaux présentés dans l'Introduction, dont le but était d'inférer l'arbre du vivant et de résoudre les relations entre les trois domaines du vivant, sont passés par des recherches automatiques de protéines présentes dans un maximum de génomes. Les gènes ainsi sélectionnés correspondaient aux fonctions les plus conservées et convergeaient en général vers une liste très réduite et centrée autour des protéines ribosomiques (Tableau 2) (Ciccarelli et al. 2006; Yutin et al. 2008). Cependant, nous avons confirmé qu'une répartition taxonomique exhaustive pour chaque protéine n'est pas nécessaire dans une approche phylogénomique car les données manquantes dans certains jeux de données peuvent être compensées par la concaténation de l'ensemble dans une supermatrice. Ainsi, l'accès à de grandes quantités de données génomiques nous permet aujourd'hui d'aller chercher d'autres protéines portant un signal intéressant, et surtout de définir un nouvel ensemble de marqueurs adaptés à chaque problématique traitée.

Par exemple, les relations entre eucaryotes et archées sont très débattues actuellement (Simonetta Gribaldo et al. 2010). Le placement de la racine de l'arbre des archées en utilisant les bactéries comme groupe extérieur nous a permis d'obtenir une phylogénie racinée des archées robuste. Considérant cette phylogénie racinée comme point de départ, il serait maintenant envisageable de chercher un nouvel ensemble de marqueurs adaptés à cette question précise. Les protéines sélectionnées n'auraient pas besoin d'être présentes chez les bactéries et pourraient être choisies avec la même méthodologie que celle utilisée dans le Chapitre 2. Cette idée a déjà été proposée par Gribaldo et collaborateurs en 2010 (Simonetta Gribaldo et al. 2010), et pourrait être mise en ouvre.

De l'augmentation de la quantité de données génomiques.

Le nombre de génomes d'archées publiés est en augmentation exponentielle depuis le début des années 2000 et les dernières avancées en termes de séquençage de cellules uniques commencent à faire exploser les quantités de données disponibles (Rinke et al. 2013). Il m'est ainsi arrivé deux fois d'entendre au cours d'une conversation, de façon plus ou moins sérieuse, que la quantité de données (génomiques ou scientifiques de façon générale) était suffisante actuellement pour répondre à toutes les questions et qu'il n'était pas forcément nécessaire d'en produire plus. Je trouve plutôt dommage d'avoir ce point de vue. Certes, l'augmentation du nombre de génomes disponibles

est complexe à gérer pour de nombreuses raisons, mais c'est aussi une mine d'information exploitable du plus grand intérêt. Il est possible que seule une petite part de ces données soit utilisée un jour, mais la génération de la totalité est très utile pour pouvoir trouver des informations intéressantes pour répondre à de nouvelles questions. Il ne faut pas oublier que de nombreux champs de la biologie sont à même d'utiliser ces séquences pour des études tout à fait différentes. Par exemple, une véritable génomique des populations commence à devenir possible pour les microorganismes, mais aussi des études très détaillées sur des caractères métaboliques, écologiques, fonctionnels ou évolutifs. Dans mon cas, la sélection des 200 nouveaux marqueurs pour la phylogénie des archées, puis des 38 pour la position de la racine, n'a pu être faite que parce qu'un nombre suffisant de génomes complets était disponible pour les archées mais aussi pour les bactéries et les eucaryotes. Avoir une répartition taxonomique représentative de la diversité de chaque domaine était nécessaire pour révéler les cas de transferts horizontaux ou de paralogies. Et dans la mesure où la recherche de ces marqueurs était par définition sans a priori, il aurait été impossible de les cibler pour un séquençage individuel. L'accès aux séquences génomiques complètes était donc nécessaire.

Parallèlement, lors de mes travaux de recherche j'ai été confrontée de nombreuses fois à l'artefact d'attraction de longues branches dû à l'évolution rapide de certains génomes. L'augmentation du nombre de génomes disponibles permet un choix plus intéressant et plus adapté en termes d'échantillonnage taxonomique. Si dans un groupe nous ne disposons que d'un seul génome fortement sujet à certains biais, nous devrons quand même l'utiliser afin de représenter ce groupe dans les analyses. Par contre, si de nouveaux génomes sont disponibles, il sera possible d'une part de vérifier si ces biais sont caractéristiques de l'espèce ou de son groupe entier et de choisir un nouvel échantillonnage taxonomique plus adapté à la question scientifique posée. Par exemple, les Nanoarchaeota, groupe taxonomique d'intérêt lorsqu'on s'intéresse à la phylogénie des archées, étaient représentés jusqu'à récemment, par un seul organisme (*N. equitans*) au génome très réduit et évoluant vite. Le séquençage d'un nouveau représentant (Podar et al. 2013), ayant un génome moins réduit et évoluant moins vite, devrait permettre de réduire les biais et artefacts autrefois induits par l'utilisation de *N. equitans*.

De l'intérêt des analyses automatiques.

L'énorme masse de données générée par toutes ces séquences génomiques est bien sûr extrêmement complexe à exploiter et, par conséquent, très chronophage. Ainsi que je l'ai déjà exposé plus tôt, la plupart des analyses phylogénétiques massives n'ont, d'une part, pas réussi à dégager de nouveaux marqueurs malgré le développement de nouvelles approches pour

l'identification des marqueurs et, d'autre part, elles ont probablement été biaisées par un certain nombre d'artefacts tels que des attractions de longues branches ou la présence de paralogies cachées (Cox, Foster, Hirt, Harris, & Embley, 2008; Rinke et al., 2013…). L'analyse que j'ai menée sur l'ensemble des protéines de trois génomes complets m'a permis de mettre en évidence 200 nouveaux marqueurs, exploités immédiatement pour l'inférence de la phylogénie des archées, mais aussi un grand nombre de protéines ayant des profils évolutifs notables, avec des cas de transferts horizontaux inattendus ou des répartitions taxonomiques particulières (cf. Chapitre 3). Ces événements évolutifs n'étaient pas l'information recherchée au départ, et c'est l'observation humaine de chaque phylogénie qui a permis de les détecter et qui permettra potentiellement de les exploiter par la suite. Parallèlement, certaines protéines ont été sélectionnées comme marqueurs potentiels dans cette première étude alors qu'elles avaient des profils inattendus qui n'auraient pas été retenus par une analyse automatique dont les critères sont forcément préétablis et fixes. Par exemple, l'absence de certains ordres complets d'archées, la présence d'homologues bactériens ou eucaryotes au sein des archées issus de transferts horizontaux ponctuels et récents mais pouvant être supprimés par la suite, ou simplement l'absence de suffisamment de signal pour avoir une phylogénie individuelle bien résolue, auraient empêché la sélection automatique de ces protéines. Bien que l'ensemble des critères utilisés pour remplir les tableaux présentés en Annexe 2, soit, à mon avis, parfaitement programmables un par un par un bon informaticien, il me semble beaucoup plus difficile, pour ne pas dire impossible, d'imaginer à l'avance tous les cas de figure auxquels on peut être confronté lors d'une analyse de cette ampleur. Un autre exemple provient des paralogies spécifiques à certains groupes d'archées : j'ai dû sélectionner certaines copies de paralogues, mais elles n'avaient pas forcement exactement la même répartition taxonomique ou la même longueur de branches. Les mêmes questions se sont posées sur l'analyse des jeux de données contenant les séquences bactériennes (cf. Chapitre 2). Malgré toutes ces remarques, je tiens à préciser que les critères de sélection des jeux de données, puis des séquences retenues auraient été difficilement programmables a priori et qu'ils peuvent même être considérés comme subjectifs, il n'en reste pas moins qu'ils sont reproductibles a posteriori.

Mon analyse a d'ailleurs nécessité l'automatisation de nombreuses procédures, en particulier pour réaliser les premières étapes de construction des jeux de données et des phylogénies individuelles. Cependant, la phase d'expertise des phylogénies résultantes a été cruciale. Il est important de préciser que les biais inhérents à ce protocole automatique (la limitation du nombre de séquences à 200 par jeu de données et la limite d'E-value à $e10^{-5}$ dans les recherches par BLAST, la non vérification de l'homologie des séquences dans les jeux de données, la sélection automatique des positions conservées pour la reconstruction phylogénétique et enfin, l'inférence des arbres

préliminaires en approximation de maximum de vraisemblance (méthode particulièrement sujette à l'attraction de longues branches)) ont été pris en compte dans la phase d'observation et les annotations que j'ai faites sur ces phylogénies ont été considérées avec de grandes précautions. Ceci n'aurait pas pu être le cas dans une analyse totalement automatique. Il me semble donc que l'automatisation de certaines taches précises est nécessaire pour pouvoir traiter ces masses de données, mais qu'elle ne doit pas se faire aux dépens d'une analyse humaine à certaines étapes clés du protocole.

Le temps et les compétences humaines nécessaires à ce genre d'analyse ne sont évidement pas compatibles avec l'analyse à flux tendu de l'ensemble de l'information génomique publiée. Un décalage se crée donc entre la disponibilité d'une information et son utilisation. De mon point de vue, il est important de choisir une question scientifique précise et de passer le temps nécessaire à l'analyse des données pour prétendre y répondre de façon fiable, plutôt que d'analyser de façon massive des données en perpétuelle expansion sans prendre en compte toutes les spécificités de la question à laquelle on veut répondre, au risque d'obtenir des résultats biaisés. La totalité des données disponibles ne sera peut être jamais traitée, mais certaines questions scientifiques auront des réponses de plus en plus fiables.

De la multiplication des phyla archéens.

Ces dernières années (fin des années 2000 et début des années 2010), de nombreux nouveaux phyla d'archées ont été proposés. D'abord les Thaumarchaeota en 2008, definis sur l'analyse phylogénétique et génomique de *C. symbiosum* et l'analyse de diverses séquences environnementales (Brochier-Armanet et al. 2008). En 2011, Nunoura et collaborateurs proposent le phylum Aigarchaeota (Nunoura et al. 2011) à partir de l'analyse de '*Ca*. Caldiarchaeum subterraneum'. Au cours de cette dernière année, pas moins de cinq phyla ont été proposés : Geoarchaeota (Kozubal et al. 2013), suivit de Aenigmarchaeota, Diapherotrites, Nanohaloarchaeota et Parvarchaeota (Rinke et al. 2013). Les phyla Thaumarchaeota et Aigarchaeota reposent sur une analyse assez poussée des génomes disponibles et, surtout, sur le fait que ces groupes sont connus depuis longtemps par les études environnementales ; de plus, en ce qui concerne les Thaumarchaeota, des représentants cultivés étaient disponibles en 2008. Par contre, les cinq derniers nouveaux phyla reposent sur des métagénomes ou des génomes séquencés à partir de cellules uniques. Je ne remets pas en cause ici la validité de ces phyla, mais il me semble intéressant de noter le fait qu'aucune règle ne régit la taxonomie procaryote, et a fortiori des archées, en ce qui concerne les rangs taxonomiques de haut niveau (supérieurs à la classe). Certains des phyla proposés (Nanohaloarchaeota et Parvarchaeota) sont particulièrement sujets aux LBA et aucune

précaution particulière n'a été prise pour minimiser cet artefact dans l'article où Rinke et collaborateurs les ont proposés (Rinke et al. 2013). En ce qui concerne les Geoarchaeota (Kozubal et al. 2013), Aenigmarchaeota et Diapherotrites (Rinke et al. 2013), ces phyla ont été proposés dans l'article décrivant leurs génomes, mais la biologie de ces organismes reste en très grande partie inconnue. Il serait donc prudent d'attendre d'en savoir plus à leur propos et de préciser leurs positions phylogénétiques avec plus de séquences et des analyses phylogénétiques fiables avant de proposer la création de nouveaux phyla.

Du risque de découplage entre génome et organisme.

La multiplication des études de séquençage de génomes à partir de cellules uniques ou de métagénomes est extrêmement intéressante de par la quantité de nouvelles informations génomiques qu'elles apportent. Mais ces études peuvent être critiquées dans le sens où le génome n'est pas l'organisme. La séquence de tous les gènes d'un organisme ne suffit pas pour comprendre son fonctionnement cellulaire dans sa totalité. Les protéines sont prédites à partir d'ORF (« Open Reading Frame » pour « cadre ouvert de lecture ») et il n'est jamais certain que ces gènes soient réellement exprimés dans la cellule. Pour beaucoup d'ORF aucune fonction n'est assignable, et même si une fonction peut être assignée à un ORF, c'est par comparaison avec son homologue le plus proche, et dans la plupart des cas cette fonction n'est pas vérifiée de façon expérimentale. En revanche, ces génomes et métagénomes sont de nouvelles sources d'information très riches sur le métabolisme potentiel de ces organismes. Elles pourraient permettre de tenter de créer un environnement favorable à leur culture, pure ou en association avec d'autres organismes, cultures qui seraient utilisables pour des confirmations expérimentales. Dans le cas des archées, cela fut notamment le cas pour la thaumarchée *Nitrosopumilus maritimus*, la première à être isolée pour ce groupe, car des données métagénomiques avaient prédit quelques années plus tôt l'importance du métabolisme de l'ammonium chez ces organismes (Treusch et al. 2005). Il est aussi intéressant de pouvoir coupler l'analyse de génomes à d'autres techniques ne nécessitant pas d'utiliser l'organisme lui-même, telle que la transcriptomique par exemple, pour savoir si les gènes prédits sont exprimés, une idée qui est à la base des techniques de métatranscriptomiques et de métaprotéomiques de plus en plus utilisées (Shi, Tyson, and DeLong 2009; Bertin et al. 2011). L'application de ces méthodes pour comprendre le fonctionnement et l'évolution des communautés microbiennes, souvent très complexes, peut paraître utopique, mais dans la mesure où de plus en plus de génomes sont accessibles et où les organismes sont mal connus, cette voie, explorée depuis plusieurs années déjà, me semble tout à fait justifiée.

De la phylogénie moléculaire comme révolution scientifique ?

La première partie de l'Introduction de ma thèse portait principalement sur les travaux de Woese et de ses collaborateurs dans les années 1970 à 1990. Ici, je voudrais évoquer une question qui m'interpelle et l'introduire comme idée de réflexion. La notion de révolution scientifique est définie par Kuhn en 1962 dans son essai *The structure of scientific revolutions* (Kuhn 1962) comme le résultat d'un changement de paradigme dans un domaine de la science. Le paradigme scientifique selon Kuhn correspond à un modèle de pensée reposant sur une théorie, un socle d'observations, de faits et d'expériences tendant à confirmer la théorie en cours, et une série d'hypothèses et de prédictions qui viendraient confirmer la théorie. Il vient cadrer la pensée scientifique en lui donnant un socle de connaissances acquises et une méthode. Cette théorie de philosophie des sciences a été proposée d'après l'analyse de l'histoire de la physique et de la chimie principalement. En biologie, un élément fondateur est la publication de *On the Origin of Species* de Darwin en 1859 (Darwin 1859). Il a permis de donner un sens aux observations des naturalistes jusqu'à ce moment là et de donner un cadre de pensée pour l'étude des organismes vivants et des relations qui les lient. La phylogénie moléculaire, et particulièrement son application par Woese sur l'ARNr SSU à partir des années 1970, a permis de passer d'une phylogénie de l'ensemble du vivant théorique proposée par Darwin, à une phylogénie réelle du vivant. De plus, Woese a donné un nouveau cadre de pensée à la classification du vivant et à l'évolution microbienne. Le travail de Woese et de ses collaborateurs ne remet pas en question la théorie de Darwin sur l'évolution des espèces mais, au contraire, vient même la confirmer. Par contre, il apporte un nouveau cadre de pensée à un domaine de la biologie, la microbiologie, et une série d'expériences et de résultats en concordance avec ce cadre de pensée (comme la prédiction de la découverte d'une grande diversité d'archées, qui a été réalisée). Étant donné l'avancée fondamentale du travail de Woese dans les années 1970 et 1980, peut-on considérer que ce travail et cette redéfinition de la classification du vivant par la phylogénie moléculaire basée sur l'ARNr SSU soit une révolution scientifique ? Je ne prétends pas avoir de réponse à cette question et je pense qu'elle pourrait faire le sujet d'un travail à part entière.

Conclusion

Le sujet de cette thèse a été l'étude de la phylogénie des archées. Mon travail s'est articulé autour de deux axes présentés dans la partie Objectifs : la recherche de nouveaux marqueurs pour l'inférence de la phylogénie des archées et l'élucidation de la position de la racine de l'arbre des archées grâce à des homologues bactériens. Ces deux objectifs ont été atteints et les résultats obtenus font l'objet de deux articles en préparation. Grâce à l'utilisation de marqueurs et de méthodes qui n'avaient jamais été appliquées à l'étude des archées, mes travaux apportent un nouvel éclairage sur la phylogénie de ces organismes et, d'une manière plus générale, sur le travail en phylogénomique.

De façon plus précise, le Chapitre 1 a pour résultats majeurs :

- La mise en évidence de 200 nouveaux marqueurs pour la reconstruction de la phylogénie des archées, avec une grande diversité fonctionnelle par rapport aux protéines informationnelles utilisées préalablement.
- L'inférence d'une phylogénie des archées à partir de 273 protéines (200 nouvelles + 57 ribosomiques + 16 sous-unités de l'ARN polymérase et facteurs de transcription) avec un meilleur support que les phylogénies précédentes.
- La résolution de certaines relations phylogénétiques problématiques : la monophylie des méthanogènes classe I et la proposition d'une nouvelle classe les regroupant ; la non-monophylie des méthanogènes classe II et le groupement de ces méthanogènes avec les *Halobacteriales* ; la monophylie des *Thermoproteales* ; la position d'*A. saccharavorans* groupe frère d'*A. pernix* au sein des *Desulfurococcales* ; la divergence des *Thermococcales* et des *Thermoproteales* en premier au sein des euryarchées et des crenarchées respectivement ; la relation fortement soutenue entre les Thaumarchaeota et l'Aigarchaeota.

En ce qui concerne le Chapitre 2 les résultats majeurs sont :

- L'identification de 38 protéines fonctionnellement très diverses et avec des homologues bactériens utilisables pour placer la racine de l'arbre des archées, indépendamment de la présence d'homologues eucaryotes.
- L'apport d'un signal phylogénétique important par rapport aux 32 protéines ribosomiques partagées entre archées et bactéries.
- Le placement de la racine avec un support très fort entre les Euryarchaeota et les autres phyla d'archées considérés dans cette étude (Thaumarchaeota, Aigarchaeota, Crenarchaeota,

Korarchaeota).

- La proposition d'une révision de la taxonomie des archées, qui consiste principalement à déclasser quatre phyla (Thaumarchaeota, Aigarchaeota, Crenarchaeota, Korarchaeota) au rang de classe et la proposition d'un nouveau phylum englobant ces quatre groupes sous le nom de « Proteoarchaeota »

Enfin, le Chapitre 3, par l'analyse de l'histoire évolutive du « système chaperonne » DnaK-GrpE-DnaJ chez les archées, montre que des histoires évolutives complexes impliquant des événements de transfert, de duplication et de fusion de gènes ont marqué l'évolution de ce système.

Bibliographie

Altschul, S F, W Gish, W Miller, E W Myers, and D J Lipman. 1990. “Basic Local Alignment Search Tool.” *Journal of Molecular Biology* 215 (3) (October 5): 403–10. doi:10.1016/S0022-2836(05)80360-2. http://www.ncbi.nlm.nih.gov/pubmed/2231712.

Altschul, S F, T L Madden, a a Schäffer, J Zhang, Z Zhang, W Miller, and D J Lipman. 1997. “Gapped BLAST and PSI-BLAST: a New Generation of Protein Database Search Programs.” *Nucleic Acids Research* 25 (17) (September 1): 3389–402. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=146917&tool=pmcentrez&rendertype=abstract.

Anderson, Iain, Markus Göker, Matt Nolan, Susan Lucas, Nancy Hammon, Shweta Deshpande, Jan-Fang Cheng, et al. 2011. “Complete Genome Sequence of the Hyperthermophilic Chemolithoautotroph Pyrolobus Fumarii Type Strain (1A).” *Standards in Genomic Sciences* 4 (3) (July 1): 381–92. doi:10.4056/sigs.2014648. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3156397&tool=pmcentrez&rendertype=abstract.

Anderson, Iain, Reinhard Wirth, Susan Lucas, Alex Copeland, Alla Lapidus, Jan-Fang Cheng, Lynne Goodwin, et al. 2011. “Complete Genome Sequence of Staphylothermus Hellenicus P8.” *Standards in Genomic Sciences* 5 (1) (October 15): 12–20. doi:10.4056/sigs.2054696. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3236042&tool=pmcentrez&rendertype=abstract.

Atoji, M, N Bjerrum, L Pauling, Hydrogen Bonding, A Geiger, G E Fox, E Stackebrandt, et al. 1980. “The Phylogeny of Prokaryotes.” *Science (New York, N.Y.)* 209 (4455) (July 25): 457–63. http://www.ncbi.nlm.nih.gov/pubmed/6771870.

Auchtung, Thomas a, Cristina D Takacs-Vesbach, and Colleen M Cavanaugh. 2006. “16S rRNA Phylogenetic Investigation of the Candidate Division ‘Korarchaeota’.” *Applied and Environmental Microbiology* 72 (7) (July): 5077–82. doi:10.1128/AEM.00052-06. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1489347&tool=pmcentrez&rendertype=abstract.

Baker, Brett J, Luis R Comolli, Gregory J Dick, Loren J Hauser, Doug Hyatt, Brian D Dill, Miriam L Land, Nathan C Verberkmoes, Robert L Hettich, and Jillian F Banfield. 2010. “Enigmatic, Ultrasmall, Uncultivated Archaea.” *Proceedings of the National Academy of Sciences of the United States of America* 107 (19) (May 11): 8806–11. doi:10.1073/pnas.0914470107. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2889320&tool=pmcentrez&rendertype=abstract.

Baker, Brett J, Gene W Tyson, Richard I Webb, Judith Flanagan, Philip Hugenholtz, Eric E Allen, and Jillian F Banfield. 2006. “Lineages of Acidophilic Archaea Revealed by Community Genomic Analysis.” *Science (New York, N.Y.)* 314 (5807) (December 22): 1933–5. doi:10.1126/science.1132690. http://www.ncbi.nlm.nih.gov/pubmed/17185602.

Balch, W E, L J Magrum, G E Fox, R S Wolfe, and C R Woese. 1977. “An Ancient Divergence Among the Bacteria.” *Journal of Molecular Evolution* 9 (4) (August 5): 305–11. http://www.ncbi.nlm.nih.gov/pubmed/408502.

Bapteste, E, E Susko, J Leigh, D MacLeod, R L Charlebois, and W F Doolittle. 2005. "Do Orthologous Gene Phylogenies Really Support Tree-thinking?" *BMC Evolutionary Biology* 5 (January): 33. doi:10.1186/1471-2148-5-33. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1156881&tool=pmcentrez&rendertype=abstract.

Bapteste, Eric, Céline Brochier, and Yan Boucher. 2005. "Higher-level Classification of the Archaea: Evolution of Methanogenesis and Methanogens." *Archaea (Vancouver, B.C.)* 1 (5) (May): 353–63. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2685549&tool=pmcentrez&rendertype=abstract.

Barns, S M, C F Delwiche, J D Palmer, and N R Pace. 1996. "Perspectives on Archaeal Diversity, Thermophily and Monophyly from Environmental rRNA Sequences." *Proceedings of the National Academy of Sciences of the United States of America* 93 (17) (August 20): 9188–93. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=38617&tool=pmcentrez&rendertype=abstract.

Barns, S M, R E Fundyga, M W Jeffries, and N R Pace. 1994. "Remarkable Archaeal Diversity Detected in a Yellowstone National Park Hot Spring Environment." *Proceedings of the National Academy of Sciences of the United States of America* 91 (5) (March 1): 1609–13. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=43212&tool=pmcentrez&rendertype=abstract.

Benson, Dennis A, Mark Cavanaugh, Karen Clark, Ilene Karsch-Mizrachi, David J Lipman, James Ostell, and Eric W Sayers. 2013. "GenBank." *Nucleic Acids Research* 41 (Database issue) (January): D36–42. doi:10.1093/nar/gks1195. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3531190&tool=pmcentrez&rendertype=abstract.

Benson, Dennis A, Ilene Karsch-Mizrachi, Karen Clark, David J Lipman, James Ostell, and Eric W Sayers. 2012. "GenBank." *Nucleic Acids Research* 40 (Database issue) (January): D48–53. doi:10.1093/nar/gkr1202. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3245039&tool=pmcentrez&rendertype=abstract.

Bertin, Philippe N, Audrey Heinrich-Salmeron, Eric Pelletier, Florence Goulhen-Chollet, Florence Arsène-Ploetze, Sébastien Gallien, Béatrice Lauga, et al. 2011. "Metabolic Diversity Among Main Microorganisms Inside an Arsenic-rich Ecosystem Revealed by Meta- and Proteo-genomics." *The ISME Journal* 5 (11) (November): 1735–47. doi:10.1038/ismej.2011.51. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3197163&tool=pmcentrez&rendertype=abstract.

Blainey, Paul C, Annika C Mosier, Anastasia Potanina, Christopher a Francis, and Stephen R Quake. 2011. "Genome of a Low-salinity Ammonia-oxidizing Archaeon Determined by Single-cell and Metagenomic Analysis." *PloS One* 6 (2) (January): e16626. doi:10.1371/journal.pone.0016626. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3043068&tool=pmcentrez&rendertype=abstract.

Boucher, Yan, W Ford Doolittle, and Peter Raynes. 2002. "Something New Under the Sea Liquid Crystal Painting" 417 (May).

Brinkmann, H, and H Philippe. 1999. “Archaea Sister Group of Bacteria? Indications from Tree Reconstruction Artifacts in Ancient Phylogenies.” *Molecular Biology and Evolution* 16 (6) (June): 817–25. http://www.ncbi.nlm.nih.gov/pubmed/10368959.

Brochier, Céline, Eric Bapteste, David Moreira, and Hervé Philippe. 2002. “Eubacterial Phylogeny Based on Translational Apparatus Proteins.” *Trends in Genetics : TIG* 18 (1) (January): 1–5. http://www.ncbi.nlm.nih.gov/pubmed/11750686.

Brochier, Céline, Patrick Forterre, and Simonetta Gribaldo. 2004. “Archaeal Phylogeny Based on Proteins of the Transcription and Translation Machineries: Tackling the Methanopyrus Kandleri Paradox.” *Genome Biology* 5 (3) (January): R17. doi:10.1186/gb-2004-5-3-r17. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=395767&tool=pmcentrez&rendertype=abstract.

———. 2005. “An Emerging Phylogenetic Core of Archaea: Phylogenies of Transcription and Translation Machineries Converge Following Addition of New Genome Sequences.” *BMC Evolutionary Biology* 5 (January): 36. doi:10.1186/1471-2148-5-36. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1177939&tool=pmcentrez&rendertype=abstract.

Brochier, Céline, Simonetta Gribaldo, Yvan Zivanovic, Fabrice Confalonieri, and Patrick Forterre. 2005. “Nanoarchaea: Representatives of a Novel Archaeal Phylum or a Fast-evolving Euryarchaeal Lineage Related to Thermococcales?” *Genome Biology* 6 (5) (January): R42. doi:10.1186/gb-2005-6-5-r42. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1175954&tool=pmcentrez&rendertype=abstract.

Brochier-Armanet, Céline, Bastien Boussau, Simonetta Gribaldo, and Patrick Forterre. 2008. “Mesophilic Crenarchaeota: Proposal for a Third Archaeal Phylum, the Thaumarchaeota.” *Nature Reviews. Microbiology* 6 (3) (March): 245–52. doi:10.1038/nrmicro1852. http://www.ncbi.nlm.nih.gov/pubmed/18274537.

Brochier-Armanet, Céline, Philippe Deschamps, Purificación López-García, Yvan Zivanovic, Francisco Rodríguez-Valera, and David Moreira. 2011. “Complete-fosmid and Fosmid-end Sequences Reveal Frequent Horizontal Gene Transfers in Marine Uncultured Planktonic Archaea.” *The ISME Journal* 5 (8) (August): 1291–302. doi:10.1038/ismej.2011.16. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3146271&tool=pmcentrez&rendertype=abstract.

Brochier-Armanet, Céline, Patrick Forterre, and Simonetta Gribaldo. 2011. “Phylogeny and Evolution of the Archaea: One Hundred Genomes Later.” *Current Opinion in Microbiology* 14 (3) (June): 274–81. doi:10.1016/j.mib.2011.04.015. http://www.ncbi.nlm.nih.gov/pubmed/21632276.

Brock, T D, K M Brock, R T Belly, and R L Weiss. 1972. “Sulfolobus: a New Genus of Sulfur-oxidizing Bacteria Living at Low pH and High Temperature.” *Archiv Für Mikrobiologie* 84 (1) (January): 54–68. http://www.ncbi.nlm.nih.gov/pubmed/4559703.

Brügger, Kim, Lanming Chen, Markus Stark, Arne Zibat, Peter Redder, Andreas Ruepp, Mariana Awayez, Qunxin She, Roger a Garrett, and Hans-Peter Klenk. 2007. “The Genome of Hyperthermus Butylicus: a Sulfur-reducing, Peptide Fermenting, Neutrophilic Crenarchaeote Growing up to 108 Degrees C.” *Archaea (Vancouver, B.C.)* 2 (2) (May): 127–35.

http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2686385&tool=pmcentrez&rendertype=abstract.

Bult, Carol J, Owen White, Gary J Olsen, Lixin Zhou, Robert D Fleischmann, Granger G Sutton, Judith A Blake, et al. 1996. “Complete Genome Sequence of the Methanogenic Arc Haeon , Me Thanococcus Jannaschii The Genome of Methanococcus Jannaschii” 273 (August).

Burggraf, S, K O Stetter, P Rouviere, and C R Woese. 1991. “Methanopyrus Kandleri: An Archaeal Methanogen Unrelated to All Other Known Methanogens.” *Systematic and Applied Microbiology* 14 (January): 346–51. http://www.ncbi.nlm.nih.gov/pubmed/11540073.

Cavicchioli, Ricardo. 2006. “Cold-adapted Archaea.” *Nature Reviews. Microbiology* 4 (5) (May): 331–43. doi:10.1038/nrmicro1390. http://www.ncbi.nlm.nih.gov/pubmed/16715049.

Chen, Lanming, Kim Bru, Marie Skovgaard, Peter Redder, Qunxin She, Elfar Torarinsson, Bo Greve, et al. 2005. “The Genome of Sulfolobus Acidocaldarius , a Model Organism of the Crenarchaeota †” 187 (14): 4992–4999. doi:10.1128/JB.187.14.4992.

Ciccarelli, Francesca D, Tobias Doerks, Christian Von Mering, Christopher J Creevey, Christian von Mering, Berend Snel, and Peer Bork. 2006. “Toward Automatic Reconstruction of a Highly Resolved Tree of Life.” *Science (New York, N.Y.)* 311 (5765) (March 3): 1283–7. doi:10.1126/science.1123061. http://www.ncbi.nlm.nih.gov/pubmed/16513982.

Cox, Cymon J, Peter G Foster, Robert P Hirt, Simon R Harris, and T Martin Embley. 2008. “The Archaebacterial Origin of Eukaryotes.” *Proceedings of the National Academy of Sciences of the United States of America* 105 (51) (December 23): 20356–61. doi:10.1073/pnas.0810647105. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2629343&tool=pmcentrez&rendertype=abstract.

Criscuolo, Alexis, and Simonetta Gribaldo. 2010. “BMGE (Block Mapping and Gathering with Entropy): a New Software for Selection of Phylogenetic Informative Regions from Multiple Sequence Alignments.” *BMC Evolutionary Biology* 10 (January): 210. doi:10.1186/1471-2148-10-210. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3017758&tool=pmcentrez&rendertype=abstract.

Dagan, Tal, and William Martin. 2006. “The Tree of One Percent.” *Genome Biology* 7 (10) (January): 118. doi:10.1186/gb-2006-7-10-118. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1794558&tool=pmcentrez&rendertype=abstract.

Darland, G, T D Brock, W Samsonoff, and S F Conti. 1970. “A Thermophilic, Acidophilic Mycoplasma Isolated from a Coal Refuse Pile.” *Science (New York, N.Y.)* 170 (3965) (December 25): 1416–8. http://www.ncbi.nlm.nih.gov/pubmed/5481857.

Darwin, Charles. 1859. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*.

De la Torre, José R, Christopher B Walker, Anitra E Ingalls, Martin Könneke, and David a Stahl. 2008. “Cultivation of a Thermophilic Ammonia Oxidizing Archaeon Synthesizing Crenarchaeol.” *Environmental Microbiology* 10 (3) (March): 810–8. doi:10.1111/j.1462-2920.2007.01506.x. http://www.ncbi.nlm.nih.gov/pubmed/18205821.

DeLong, E F. 1992. “Archaea in Coastal Marine Environments.” *Proceedings of the National Academy of Sciences of the United States of America* 89 (12) (June 15): 5685–9. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=49357&tool=pmcentrez&rendertype=abstract.

DeLong, E F, K Y Wu, B B Prézelin, and R V Jovine. 1994. “High Abundance of Archaea in Antarctic Marine Picoplankton.” *Nature* 371 (6499) (October 20): 695–7. doi:10.1038/371695a0. http://www.ncbi.nlm.nih.gov/pubmed/7935813.

Delong, Edward F. 2003. “Oceans of Archaea.” *ASM News* 69 (10): 503–511.

Di Giulio, Massimo. 2006. “The Non-monophyletic Origin of the tRNA Molecule and the Origin of Genes Only after the Evolutionary Stage of the Last Universal Common Ancestor (LUCA).” *Journal of Theoretical Biology* 240 (3) (June 7): 343–52. doi:10.1016/j.jtbi.2005.09.023. http://www.ncbi.nlm.nih.gov/pubmed/16289209.

———. 2007. “The Tree of Life Might Be Rooted in the Branch Leading to Nanoarchaeota.” *Gene* 401 (1-2) (October 15): 108–13. doi:10.1016/j.gene.2007.07.004. http://www.ncbi.nlm.nih.gov/pubmed/17689206.

———. 2008a. “Transfer RNA Genes in Pieces Are an Ancestral Character.” *EMBO Reports* 9 (9) (September): 820; author reply 820–1. doi:10.1038/embor.2008.153. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2529356&tool=pmcentrez&rendertype=abstract.

———. 2008b. “The Split Genes of Nanoarchaeum Equitans Are an Ancestral Character.” *Gene* 421 (1-2) (September 15): 20–6. doi:10.1016/j.gene.2008.06.010. http://www.ncbi.nlm.nih.gov/pubmed/18590807.

Do, Chuong B, Mahathi S P Mahabhashyam, Michael Brudno, and Serafim Batzoglou. 2005. “ProbCons: Probabilistic Consistency-based Multiple Sequence Alignment.” *Genome Research* 15 (2) (February 1): 330–40. doi:10.1101/gr.2821705. http://genome.cshlp.org/content/15/2/330.abstract.

Dobzhansky, Theodosuis. 1973. “Nothing Makes Sense in Biology Except in the Light of Evolution.” *The American Biology Teacher*.

Dopson, Mark, Craig Baker-austin, Andrew Hind, John P Bowman, Philip L Bond, and A P P L E Nviron M Icrobiol. 2004. “Characterization of Ferroplasma Isolates and Ferroplasma Acidarmanus Sp . Nov ., Extreme Acidophiles from Acid Mine Drainage and Industrial Bioleaching Environments” 70 (4): 2079–2088. doi:10.1128/AEM.70.4.2079.

Dorn, Karolin V, Felix Willmund, Christian Schwarz, Christine Henselmann, Thomas Pohl, Barbara Hess, Daniel Veyel, et al. 2010. “Chloroplast DnaJ-like Proteins 3 and 4 (CDJ3/4) from Chlamydomonas Reinhardtii Contain Redox-active Fe-S Clusters and Interact with Stromal HSP70B.” *The Biochemical Journal* 427 (2) (April 15): 205–15. doi:10.1042/BJ20091412. http://www.ncbi.nlm.nih.gov/pubmed/20113313.

Edgar, Robert C. 2004. “MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput.” *Nucleic Acids Research* 32 (5) (January): 1792–7. doi:10.1093/nar/gkh340. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=390337&tool=pmcentrez&rendertype=abstract.

Elkins, James G, Mircea Podar, David E Graham, Kira S Makarova, Yuri Wolf, Lennart Randau, Brian P Hedlund, 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 (23) (June 10): 8102–7. doi:10.1073/pnas.0801980105. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2430366&tool=pmcentrez&rendertype=abstract.

Eme, Laura, Laila J Reigstad, Anja Spang, Anders Lanzén, Thomas Weinmaier, Thomas Rattei, Christa Schleper, and Céline Brochier-Armanet. 2013. "Metagenomics of Kamchatkan Hot Spring Filaments Reveal Two New Major (hyper)thermophilic Lineages Related to Thaumarchaeota." *Research in Microbiology* 164 (5) (June): 425–38. doi:10.1016/j.resmic.2013.02.006. http://www.ncbi.nlm.nih.gov/pubmed/23470515.

Farlow, W. G. 1880. "On the Nature of the Peculiar Reddening of Salted Codfish During the Summer Season." *Fish Comm.* 3: 969–97.

Fitch, W M, and E Margoliash. 1967. "Construction of Phylogenetic Trees." *Science (New York, N.Y.)* 155 (3760) (January 20): 279–84. http://www.ncbi.nlm.nih.gov/pubmed/5334057.

Foster, Peter G, Cymon J Cox, and T Martin Embley. 2009. "The Primary Divisions of Life: a Phylogenomic Approach Employing Composition-heterogeneous Methods." *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 364 (1527) (August 12): 2197–207. doi:10.1098/rstb.2009.0034. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2873002&tool=pmcentrez&rendertype=abstract.

Fox, G. E., K. R. Pechman, and C. R. Woese. 1977. "Comparative Cataloging of 16S Ribosomal Ribonucleic Acid: Molecular Approach to Procaryotic Systematics." *International Journal of Systematic Bacteriology* 27 (1) (January 1): 44–57. doi:10.1099/00207713-27-1-44. http://ijs.sgmjournals.org/cgi/doi/10.1099/00207713-27-1-44.

Fox, George E, Linda J Magrum, W E Balch, R S Wolfe, Carl R Woese, William E Balcht, and Ralph S Wolfef. 1977. "Classification of Methanogenic Bacteria by 16S Ribosomal RNA Characterization." *Proceedings of the National Academy of Sciences of the United States of America* 74 (10) (October): 4537–41. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=431980&tool=pmcentrez&rendertype=abstract.

Franzmann, P D, Y Liu, D L Balkwill, H C Aldrich, E Conway de Macario, and D R Boone. 1997. "Methanogenium Frigidum Sp. Nov., a Psychrophilic, H2-using Methanogen from Ace Lake, Antarctica." *International Journal of Systematic Bacteriology* 47 (4) (October): 1068–72. http://www.ncbi.nlm.nih.gov/pubmed/9336907.

Franzmann, P.D., N. Springer, W. Ludwig, E. Conway De Macario, and M. Rohde. 1992. "A Methanogenic Archaeon from Ace Lake, Antarctica: Methanococcoides Burtonii Sp. Nov." *Systematic and Applied Microbiology* 15 (4) (December): 573–581. doi:10.1016/S0723-2020(11)80117-7. http://dx.doi.org/10.1016/S0723-2020(11)80117-7.

Franzmann, P.D., E. Stackebrandt, K. Sanderson, J.K. Volkman, D.E. Cameron, P.L. Stevenson, T.A. Mcmeekin, and H.R. Burton. 1988. "Halobacterium Lacusprofundi Sp. Nov., a Halophilic Bacterium Isolated from Deep Lake, Antarctica." *Systematic and Applied Microbiology* 11 (1) (November): 20–27. doi:10.1016/S0723-2020(88)80044-4. http://dx.doi.org/10.1016/S0723-2020(88)80044-4.

French, Sarah L, Thomas J Santangelo, Ann L Beyer, and John N Reeve. 2007. “Transcription and Translation Are Coupled in Archaea.” *Molecular Biology and Evolution* 24 (4) (April): 893–5. doi:10.1093/molbev/msm007. http://www.ncbi.nlm.nih.gov/pubmed/17237472.

Fuhrman, J a, K McCallum, and a a Davis. 1992. “Novel Major Archaebacterial Group from Marine Plankton.” *Nature* 356 (6365) (March 12): 148–9. doi:10.1038/356148a0. http://www.ncbi.nlm.nih.gov/pubmed/1545865.

Ghai, Rohit, Lejla Pašić, Ana Beatriz Fernández, Ana-Belen Martin-Cuadrado, Carolina Megumi Mizuno, Katherine D McMahon, R Thane Papke, et al. 2011. “New Abundant Microbial Groups in Aquatic Hypersaline Environments.” *Scientific Reports* 1 (January): 135. doi:10.1038/srep00135. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3216616&tool=pmcentrez&rendertype=abstract.

Gogarten, J P, H Kibak, P Dittrich, L Taiz, E J Bowman, B J Bowman, M F Manolson, R J Poole, T Date, and T Oshima. 1989. “Evolution of the Vacuolar H+-ATPase: Implications for the Origin of Eukaryotes.” *Proceedings of the National Academy of Sciences of the United States of America* 86 (17) (September): 6661–5. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=297905&tool=pmcentrez&rendertype=abstract.

Göker, Markus, Brittany Held, Alla Lapidus, Matt Nolan, Stefan Spring, Montri Yasawong, Susan Lucas, et al. 2010. “Complete Genome Sequence of Ignisphaera Aggregans Type Strain (AQ1.S1).” *Standards in Genomic Sciences* 3 (1) (January): 66–75. doi:10.4056/sigs.1072907. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3035270&tool=pmcentrez&rendertype=abstract.

Gribaldo, S, and P Cammarano. 1998. “The Root of the Universal Tree of Life Inferred from Anciently Duplicated Genes Encoding Components of the Protein-targeting Machinery.” *Journal of Molecular Evolution* 47 (5) (November): 508–16. http://www.ncbi.nlm.nih.gov/pubmed/9797401.

Gribaldo, Simonetta, and Celine Brochier. 2009. “Phylogeny of Prokaryotes: Does It Exist and Why Should We Care?” *Research in Microbiology* 160 (7) (September): 513–21. doi:10.1016/j.resmic.2009.07.006. http://www.ncbi.nlm.nih.gov/pubmed/19631737.

Gribaldo, Simonetta, and Celine Brochier-Armanet. 2006. “The Origin and Evolution of Archaea: a State of the Art.” *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 361 (1470) (June 29): 1007–22. doi:10.1098/rstb.2006.1841. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1578729&tool=pmcentrez&rendertype=abstract.

Gribaldo, Simonetta, and Céline Brochier-Armanet. 2012. “Time for Order in Microbial Systematics.” *Trends in Microbiology* 20 (5) (May): 209–10. doi:10.1016/j.tim.2012.02.006. http://www.ncbi.nlm.nih.gov/pubmed/22440793.

Gribaldo, Simonetta, Anthony M Poole, Vincent Daubin, Patrick Forterre, and Céline Brochier-Armanet. 2010. “The Origin of Eukaryotes and Their Relationship with the Archaea: Are We at a Phylogenomic Impasse?” *Nature Reviews. Microbiology* 8 (10) (October): 743–52. doi:10.1038/nrmicro2426. http://www.ncbi.nlm.nih.gov/pubmed/20844558.

Guindon, Stéphane, Jean-François Dufayard, Vincent Lefort, Maria Anisimova, Wim Hordijk, and Olivier Gascuel. 2010. “New Algorithms and Methods to Estimate Maximum-likelihood

Phylogenies: Assessing the Performance of PhyML 3.0." *Systematic Biology* 59 (3) (May): 307–21. doi:10.1093/sysbio/syq010. http://www.ncbi.nlm.nih.gov/pubmed/20525638.

Gupta, R, J M Lanter, and C R Woese. 1983. "Sequence of the 16S Ribosomal RNA from Halobacterium Volcanii, an Archaebacterium." *Science (New York, N.Y.)* 221 (4611) (August 12): 656–9. doi:10.1126/science.221.4611.656. http://www.ncbi.nlm.nih.gov/pubmed/17787735.

Gupta, R S. 2000. "The Natural Evolutionary Relationships Among Prokaryotes." *Critical Reviews in Microbiology* 26 (2) (January): 111–31. doi:10.1080/10408410091154219. http://www.ncbi.nlm.nih.gov/pubmed/10890353.

Guy, Lionel, and Thijs J G Ettema. 2011. "The Archaeal 'TACK' Superphylum and the Origin of Eukaryotes." *Trends in Microbiology* 19 (12) (December): 580–7. doi:10.1016/j.tim.2011.09.002. http://www.ncbi.nlm.nih.gov/pubmed/22018741.

Hafenbradl, Doris, Martin Keller, Reinhard Dirmeier, Reinhard Rachel, Petra Roßnagel, Siegfried Burggraf, Harald Huber, and Karl O Stetter. 1996. "a Novel Hyperthermophilic Archaeum That Oxidizes Fe 2 + at Neutral pH Under Anoxic Conditions" 2: 308–314.

Hallam, Steven J, Konstantinos T Konstantinidis, Nik Putnam, Christa Schleper, Yoh-ichi Watanabe, Junichi Sugahara, Christina Preston, José de la Torre, Paul M Richardson, and Edward F DeLong. 2006. "Genomic Analysis of the Uncultivated Marine Crenarchaeote Cenarchaeum Symbiosum." *Proceedings of the National Academy of Sciences of the United States of America* 103 (48) (November 28): 18296–301. doi:10.1073/pnas.0608549103. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1643844&tool=pmcentrez&rendertype=abstract.

Harris, J Kirk, Scott T Kelley, George B Spiegelman, and Norman R Pace. 2003. "The Genetic Core of the Universal Ancestor" (February): 407–412. doi:10.1101/gr.652803.

Hartmann, R, H D Sickinger, and D Oesterhelt. 1980. "Anaerobic Growth of Halobacteria." *Proceedings of the National Academy of Sciences of the United States of America* 77 (7) (July): 3821–5. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=349718&tool=pmcentrez&rendertype=abstract.

Hatzenpichler, Roland, Elena V Lebedeva, Eva Spieck, Kilian Stoecker, Andreas Richter, Holger Daims, and Michael Wagner. 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 (6) (February 12): 2134–9. doi:10.1073/pnas.0708857105. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2538889&tool=pmcentrez&rendertype=abstract.

Huber, Harald, Michael J Hohn, Reinhard Rachel, Tanja Fuchs, Verena C Wimmer, and Karl O Stetter. 2002. "A New Phylum of Archaea Represented by a Nanosized Hyperthermophilic Symbiont." *Nature* 417 (6884) (May 2): 63–7. doi:10.1038/417063a. http://www.ncbi.nlm.nih.gov/pubmed/11986665.

Huber, R, M Sacher, a Vollmann, H Huber, and D Rose. 2000. "Respiration of Arsenate and Selenate by Hyperthermophilic Archaea." *Systematic and Applied Microbiology* 23 (3) (October): 305–14. doi:10.1016/S0723-2020(00)80058-2. http://www.ncbi.nlm.nih.gov/pubmed/11108007.

Iverson, Vaughn, Robert M Morris, Christian D Frazar, Chris T Berthiaume, Rhonda L Morales, and E Virginia Armbrust. 2012. “Untangling Genomes from Metagenomes: Revealing an Uncultured Class of Marine Euryarchaeota.” *Science (New York, N.Y.)* 335 (6068) (February 3): 587–90. doi:10.1126/science.1212665. http://www.ncbi.nlm.nih.gov/pubmed/22301318.

Iwabe, N, K Kuma, M Hasegawa, S Osawa, and T Miyata. 1989. “Evolutionary Relationship of Archaebacteria, Eubacteria, and Eukaryotes Inferred from Phylogenetic Trees of Duplicated Genes.” *Proceedings of the National Academy of Sciences of the United States of America* 86 (23) (December): 9355–9. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=298494&tool=pmcentrez&rendertype=abstract.

Jobb, Gangolf, Arndt von Haeseler, and Korbinian Strimmer. 2004. “TREEFINDER: a Powerful Graphical Analysis Environment for Molecular Phylogenetics.” *BMC Evolutionary Biology* 4 (June 28): 18. doi:10.1186/1471-2148-4-18. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=459214&tool=pmcentrez&rendertype=abstract.

Joulian, C, B K Patel, B Ollivier, J L Garcia, and P a Roger. 2000. “Methanobacterium Oryzae Sp. Nov., a Novel Methanogenic Rod Isolated from a Philippines Ricefield.” *International Journal of Systematic and Evolutionary Microbiology* 50 Pt 2 (March): 525–8. http://www.ncbi.nlm.nih.gov/pubmed/10758856.

Karner, M B, E F DeLong, and D M Karl. 2001. “Archaeal Dominance in the Mesopelagic Zone of the Pacific Ocean.” *Nature* 409 (6819) (January 25): 507–10. doi:10.1038/35054051. http://www.ncbi.nlm.nih.gov/pubmed/11206545.

Katoh, Kazutaka, Kei-ichi Kuma, Hiroyuki Toh, and Takashi Miyata. 2005. “MAFFT Version 5: Improvement in Accuracy of Multiple Sequence Alignment.” *Nucleic Acids Research* 33 (2) (January): 511–8. doi:10.1093/nar/gki198. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=548345&tool=pmcentrez&rendertype=abstract.

Klenk, H P, and W Zillig. 1994. “DNA-dependent RNA Polymerase Subunit B as a Tool for Phylogenetic Reconstructions: Branching Topology of the Archaeal Domain.” *Journal of Molecular Evolution* 38 (4) (April): 420–32. http://www.ncbi.nlm.nih.gov/pubmed/8007009.

Knittel, Katrin, and Antje Boetius. 2009. “Anaerobic Oxidation of Methane: Progress with an Unknown Process.” *Annual Review of Microbiology* 63 (January): 311–34. doi:10.1146/annurev.micro.61.080706.093130. http://www.ncbi.nlm.nih.gov/pubmed/19575572.

KOCUR, M., and W. Hodgkiss. 1973. “Taxonomic Status of the Genus Halococcus Schoop.” *International Journal of Systematic Bacteriology* 23 (2) (April 1): 151–156. doi:10.1099/00207713-23-2-151. http://ijs.sgmjournals.org/content/23/2/151.abstract.

Könneke, Martin, Anne E Bernhard, José R de la Torre, Christopher B Walker, John B Waterbury, and David a Stahl. 2005. “Isolation of an Autotrophic Ammonia-oxidizing Marine Archaeon.” *Nature* 437 (7058) (September 22): 543–6. doi:10.1038/nature03911. http://www.ncbi.nlm.nih.gov/pubmed/16177789.

Kostka, Martin, Magdalena Uzlikova, Ivan Cepicka, and Jaroslav Flegr. 2008. “SlowFaster, a User-friendly Program for Slow-fast Analysis and Its Application on Phylogeny of Blastocystis.” *BMC Bioinformatics* 9 (January): 341. doi:10.1186/1471-2105-9-341.

http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2529323&tool=pmcentrez&rendertype=abstract.

Kozubal, Mark a, Margaret Romine, Ryan deM Jennings, Zack J Jay, Susannah G Tringe, Doug B Rusch, Jacob P Beam, Lee Ann McCue, and William P Inskeep. 2013. "Geoarchaeota: a New Candidate Phylum in the Archaea from High-temperature Acidic Iron Mats in Yellowstone National Park." *The ISME Journal* 7 (3) (March): 622–34. doi:10.1038/ismej.2012.132. http://www.ncbi.nlm.nih.gov/pubmed/23151644.

Kuhn, Thomas S. 1962. *The Structure of Scientific Revolutions*.

Kurr, Margit, Robert Huber, Helmut K�nig, Holger W. Jannasch, Hans Fricke, Antonio Trincone, Jakob K. Kristjansson, and Karl O. Stetter. 1991. "Methanopyrus Kandleri, Gen. and Sp. Nov. Represents a Novel Group of Hyperthermophilic Methanogens, Growing at 110�C." *Archives of Microbiology* 156 (4) (September): 239–247. doi:10.1007/BF00262992. http://link.springer.com/10.1007/BF00262992.

Kyrpides, N C. 1999. "Genomes OnLine Database (GOLD 1.0): a Monitor of Complete and Ongoing Genome Projects World-wide." *Bioinformatics (Oxford, England)* 15 (9) (September): 773–4. http://www.ncbi.nlm.nih.gov/pubmed/10498782.

Larkin, M A, G Blackshields, N P Brown, R Chenna, P A McGettigan, H McWilliam, F Valentin, et al. 2007. "Clustal W and Clustal X Version 2.0." *Bioinformatics (Oxford, England)* 23 (21) (November 1): 2947–8. doi:10.1093/bioinformatics/btm404. http://www.ncbi.nlm.nih.gov/pubmed/17846036.

Lartillot, Nicolas, Thomas Lepage, and Samuel Blanquart. 2009. "PhyloBayes 3: a Bayesian Software Package for Phylogenetic Reconstruction and Molecular Dating." *Bioinformatics (Oxford, England)* 25 (17) (September 1): 2286–8. doi:10.1093/bioinformatics/btp368. http://www.ncbi.nlm.nih.gov/pubmed/19535536.

Lasek-Nesselquist, Erica, and Johann Peter Gogarten. 2013. "The Effects of Model Choice and Mitigating Bias on the Ribosomal Tree of Life." *Molecular Phylogenetics and Evolution* 69 (1) (May 21): 17–38. doi:10.1016/j.ympev.2013.05.006. http://www.ncbi.nlm.nih.gov/pubmed/23707703.

Le, Si Quang, and Olivier Gascuel. 2008. "An Improved General Amino Acid Replacement Matrix." *Molecular Biology and Evolution* 25 (7) (July): 1307–20. doi:10.1093/molbev/msn067. http://www.ncbi.nlm.nih.gov/pubmed/18367465.

Leigh, John A, Sonja-Verena Albers, Haruyuki Atomi, and Thorsten Allers. 2011. "Model Organisms for Genetics in the Domain Archaea: Methanogens, Halophiles, Thermococcales and Sulfolobales." *FEMS Microbiology Reviews* 35 (4) (July): 577–608. doi:10.1111/j.1574-6976.2011.00265.x. http://www.ncbi.nlm.nih.gov/pubmed/21265868.

Liu, Li-Jun, Xiao-Yan You, Xu Guo, Shuang-Jiang Liu, and Cheng-Ying Jiang. 2011. "Metallosphaera Cuprina Sp. Nov., an Acidothermophilic, Metal-mobilizing Archaeon." *International Journal of Systematic and Evolutionary Microbiology* 61 (Pt 10) (October): 2395–400. doi:10.1099/ijs.0.026591-0. http://www.ncbi.nlm.nih.gov/pubmed/21057050.

Liu, Yuchen, and William B Whitman. 2008. "Metabolic, Phylogenetic, and Ecological Diversity of the Methanogenic Archaea." *Annals of the New York Academy of Sciences* 1125 (March): 171–89. doi:10.1196/annals.1419.019. http://www.ncbi.nlm.nih.gov/pubmed/18378594.

López-garcía, Purificación, Céline Brochier, David Moreira, and Francisco Rodríguez-valera. 2004. "Comparative Analysis of a Genome Fragment of an Uncultivated Mesopelagic Crenarchaeote Reveals Multiple Horizontal Gene Transfers" 6: 19–34. doi:10.1046/j.1462-2920.2003.00533.x.

Magrum, L J, K R Luehrsen, and C R Woese. 1978. "Are Extreme Halophiles Actually 'Bacteria'?" *Journal of Molecular Evolution* 11 (1) (May 12): 1–8. http://www.ncbi.nlm.nih.gov/pubmed/660662.

Mardanov, Andrey V, Nikolai V Ravin, Vitali a Svetlitchnyi, Alexey V Beletsky, Margarita L Miroshnichenko, Elizaveta a Bonch-Osmolovskaya, and Konstantin G Skryabin. 2009. "Metabolic Versatility and Indigenous Origin of the Archaeon Thermococcus Sibiricus, Isolated from a Siberian Oil Reservoir, as Revealed by Genome Analysis." *Applied and Environmental Microbiology* 75 (13) (July): 4580–8. doi:10.1128/AEM.00718-09. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2704819&tool=pmcentrez&rendertype=abstract.

Margulis, L, and R Guerrero. 1991. "Kingdoms in Turmoil." *New Scientist (1971)* 1761 (March 23): 46–50. http://www.ncbi.nlm.nih.gov/pubmed/11538109.

Matte-Tailliez, Oriane, Céline Brochier, Patrick Forterre, and Hervé Philippe. 2002. "Archaeal Phylogeny Based on Ribosomal Proteins." *Molecular Biology and Evolution* 19 (5) (May): 631–9. http://www.ncbi.nlm.nih.gov/pubmed/11961097.

Mayr, Ernst. 1990. "A Natural System of Organisms." *Nature* 348 (6301) (December 6): 491–491. doi:10.1038/348491a0. http://dx.doi.org/10.1038/348491a0.

———. 1991. "More Natural Classification." *Nature* 353 (6340) (September 12): 122–122. doi:10.1038/353122a0. http://www.nature.com/nature/journal/v353/n6340/pdf/353122a0.pdf.

Mihajlovski, Agnès, Monique Alric, and Jean-François Brugère. 2008. "A Putative New Order of Methanogenic Archaea Inhabiting the Human Gut, as Revealed by Molecular Analyses of the mcrA Gene." *Research in Microbiology* 159 (7-8): 516–21. doi:10.1016/j.resmic.2008.06.007. http://www.ncbi.nlm.nih.gov/pubmed/18644435.

Mills, Heath J, Robert J Martinez, Sandra Story, and Patricia A Sobecky. 2005. "Characterization of Microbial Community Structure in Gulf of Mexico Gas Hydrates : Comparative Analysis of DNA- and RNA-Derived Clone Libraries" 71 (6): 3235–3247. doi:10.1128/AEM.71.6.3235.

Moore, R L, and B J McCarthy. 1967. "Comparative Study of Ribosomal Ribonucleic Acid Cistrons in Enterobacteria and Myxobacteria." *Journal of Bacteriology* 94 (4) (October): 1066–74. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=276777&tool=pmcentrez&rendertype=abstract.

Mosier, Annika C, Eric E Allen, Maria Kim, Steven Ferriera, and Christopher a Francis. 2012. "Genome Sequence of 'Candidatus Nitrosopumilus Salaria' BD31, an Ammonia-oxidizing Archaeon from the San Francisco Bay Estuary." *Journal of Bacteriology* 194 (8) (April): 2121–2. doi:10.1128/JB.00013-12. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3318490&tool=pmcentrez&rendertype=abstract.

Mullakhanbhai, M F, and H Larsen. 1975. “Halobacterium Volcanii Spec. Nov., a Dead Sea Halobacterium with a Moderate Salt Requirement.” *Archives of Microbiology* 104 (3) (August 28): 207–14. http://www.ncbi.nlm.nih.gov/pubmed/1190944.

Narasingarao, Priya, Sheila Podell, Juan a Ugalde, Céline Brochier-Armanet, Joanne B Emerson, Jochen J Brocks, Karla B Heidelberg, Jillian F Banfield, and Eric E Allen. 2012. “De Novo Metagenomic Assembly Reveals Abundant Novel Major Lineage of Archaea in Hypersaline Microbial Communities.” *The ISME Journal* 6 (1) (January): 81–93. doi:10.1038/ismej.2011.78. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3246234&tool=pmcentrez&rendertype=abstract.

Notredame, C, D G Higgins, and J Heringa. 2000. “T-Coffee: A Novel Method for Fast and Accurate Multiple Sequence Alignment.” *Journal of Molecular Biology* 302 (1) (September 8): 205–17. doi:10.1006/jmbi.2000.4042. http://www.ncbi.nlm.nih.gov/pubmed/10964570.

Nunoura, Takuro, Hisako Hirayama, Hideto Takami, Hanako Oida, Shinro Nishi, Shigeru Shimamura, Yohey Suzuki, et al. 2005. “Genetic and Functional Properties of Uncultivated Thermophilic Crenarchaeotes from a Subsurface Gold Mine as Revealed by Analysis of Genome Fragments.” *Environmental Microbiology* 7 (12) (December): 1967–84. doi:10.1111/j.1462-2920.2005.00881.x. http://www.ncbi.nlm.nih.gov/pubmed/16309394.

Nunoura, Takuro, Hanako Oida, Miwako Nakaseama, Ayako Kosaka, Satoru B Ohkubo, Toru Kikuchi, Hiromi Kazama, et al. 2010. “Archaeal Diversity and Distribution Along Thermal and Geochemical Gradients in Hydrothermal Sediments at the Yonaguni Knoll IV Hydrothermal Field in the Southern Okinawa Trough.” *Applied and Environmental Microbiology* 76 (4) (February): 1198–211. doi:10.1128/AEM.00924-09. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2820965&tool=pmcentrez&rendertype=abstract.

Nunoura, Takuro, Yoshihiro Takaki, Jungo Kakuta, Shinro Nishi, Junichi Sugahara, Hiromi Kazama, Gab-Joo Chee, et al. 2011. “Insights into the Evolution of Archaea and Eukaryotic Protein Modifier Systems Revealed by the Genome of a Novel Archaeal Group.” *Nucleic Acids Research* 39 (8) (April): 3204–23. doi:10.1093/nar/gkq1228. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3082918&tool=pmcentrez&rendertype=abstract.

Offre, Pierre, Anja Spang, and Christa Schleper. 2013. “Archaea in Biogeochemical Cycles.” *Annual Review of Microbiology* 67 (1) (November): 130628184403000. doi:10.1146/annurev-micro-092412-155614. http://www.annualreviews.org/doi/abs/10.1146/annurev-micro-092412-155614.

Olsen, G J, D J Lane, S J Giovannoni, N R Pace, and D a Stahl. 1986. “Microbial Ecology and Evolution: a Ribosomal RNA Approach.” *Annual Review of Microbiology* 40 (January): 337–65. doi:10.1146/annurev.mi.40.100186.002005. http://www.ncbi.nlm.nih.gov/pubmed/2430518.

Olsen, G J, N R Pace, M Nuell, B P Kaine, R Gupta, and C R Woese. 1985. “Sequence of the 16S rRNA Gene from the Thermoacidophilic Archaebacterium Sulfolobus Solfataricus and Its Evolutionary Implications.” *Journal of Molecular Evolution* 22 (4) (January): 301–7. http://www.ncbi.nlm.nih.gov/pubmed/3936935.

Olsen, G J, and C R Woese. 1993. “Ribosomal RNA: a Key to Phylogeny.” *FASEB Journal : Official Publication of the Federation of American Societies for Experimental Biology* 7 (1) (January): 113–23. http://www.ncbi.nlm.nih.gov/pubmed/8422957.

Pace, B, and L L Campbell. 1971. “Homology of Ribosomal Ribonucleic Acid Diverse Bacterial Species with Escherichia Coli and Bacillus Stearothermophilus.” *Journal of Bacteriology* 107 (2) (August): 543–7. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=246958&tool=pmcentrez&rendertype=abstract.

Pace, N R, G J Olsen, and C R Woese. 1986. “Ribosomal RNA Phylogeny and the Primary Lines of Evolutionary Descent.” *Cell* 45 (3) (May 9): 325–6. http://www.ncbi.nlm.nih.gov/pubmed/3084106.

Park, Soo-Je, Jong-Geol Kim, Man-Young Jung, So-Jeong Kim, In-Tae Cha, Rohit Ghai, Ana-Belén Martín-Cuadrado, Francisco Rodríguez-Valera, and Sung-Keun Rhee. 2012. “Draft Genome Sequence of an Ammonia-oxidizing Archaeon, ‘Candidatus Nitrosopumilus Sediminis’ AR2, from Svalbard in the Arctic Circle.” *Journal of Bacteriology* 194 (24) (December): 6948–9. doi:10.1128/JB.01869-12. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3510607&tool=pmcentrez&rendertype=abstract.

Paul, Kristina, James O Nonoh, Lena Mikulski, and Andreas Brune. 2012. “‘Methanoplasmatales,’ Thermoplasmatales-related Archaea in Termite Guts and Other Environments, Are the Seventh Order of Methanogens.” *Applied and Environmental Microbiology* 78 (23) (December): 8245–53. doi:10.1128/AEM.02193-12. http://www.ncbi.nlm.nih.gov/pubmed/23001661.

Perevalova, Anna a, Salima Kh Bidzhieva, Ilya V Kublanov, Kai-Uwe Hinrichs, Xiaolei L Liu, Andrey V Mardanov, Alexander V Lebedinsky, and Elizaveta a Bonch-Osmolovskaya. 2010. “Fervidicoccus Fontis Gen. Nov., Sp. Nov., an Anaerobic, Thermophilic Crenarchaeote from Terrestrial Hot Springs, and Proposal of Fervidicoccaceae Fam. Nov. and Fervidicoccales Ord. Nov.” *International Journal of Systematic and Evolutionary Microbiology* 60 (Pt 9) (September): 2082–8. doi:10.1099/ijs.0.019042-0. http://www.ncbi.nlm.nih.gov/pubmed/19837732.

Petitjean, Céline, David Moreira, Purificación López-García, and Céline Brochier-Armanet. 2012. “Horizontal Gene Transfer of a Chloroplast DnaJ-Fer Protein to Thaumarchaeota and the Evolutionary History of the DnaK Chaperone System in Archaea.” *BMC Evolutionary Biology* 12 (January): 226. doi:10.1186/1471-2148-12-226. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3564930&tool=pmcentrez&rendertype=abstract.

Philippe, H. 1993. “MUST, a Computer Package of Management Utilities for Sequences and Trees.” *Nucleic Acids Research* 21 (22) (November 11): 5264–72. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=310646&tool=pmcentrez&rendertype=abstract.

Philippe, H, and J Laurent. 1998. “How Good Are Deep Phylogenetic Trees?” *Current Opinion in Genetics & Development* 8 (6) (December): 616–23. http://www.ncbi.nlm.nih.gov/pubmed/9914208.

Pisani, Davide, James a Cotton, and James O McInerney. 2007. “Supertrees Disentangle the Chimerical Origin of Eukaryotic Genomes.” *Molecular Biology and Evolution* 24 (8) (August): 1752–60. doi:10.1093/molbev/msm095. http://www.ncbi.nlm.nih.gov/pubmed/17504772.

Podar, Mircea, Iain Anderson, Kira S Makarova, James G Elkins, Natalia Ivanova, Mark a Wall, Athanasios Lykidis, et al. 2008. “A Genomic Analysis of the Archaeal System Ignicoccus hospitalis-Nanoarchaeum Equitans.” *Genome Biology* 9 (11) (January): R158. doi:10.1186/gb-2008-9-11-r158. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2614490&tool=pmcentrez&rendertype=abstract.

Podar, Mircea, Kira S Makarova, David E Graham, Yuri I Wolf, Eugene V Koonin, and Anna-Louise Reysenbach. 2013. “Insights into Archaeal Evolution and Symbiosis from the Genomes of a Nanoarchaeon and Its Inferred Crenarchaeal Host from Obsidian Pool, Yellowstone National Park.” *Biology Direct* 8 (January): 9. doi:10.1186/1745-6150-8-9. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3655853&tool=pmcentrez&rendertype=abstract.

Preston, Christina M, K E Ying Wu, Tadeusz F Molinskit, and Edward F Delong. 1996. “A Psychrophilic Crenarchaeon Inhabits a Marine Sponge :” 93 (June): 6241–6246.

Price, Morgan N, Paramvir S Dehal, and Adam P Arkin. 2010. “FastTree 2--approximately Maximum-likelihood Trees for Large Alignments.” Edited by Art F. Y. Poon. *PloS One* 5 (3) (January): e9490. doi:10.1371/journal.pone.0009490. http://dx.plos.org/10.1371/journal.pone.0009490.

Prokofeva, M I, N a Kostrikina, T V Kolganova, T P Tourova, a M Lysenko, a V Lebedinsky, and E a Bonch-Osmolovskaya. 2009. “Isolation of the Anaerobic Thermoacidophilic Crenarchaeote Acidilobus Saccharovorans Sp. Nov. and Proposal of Acidilobales Ord. Nov., Including Acidilobaceae Fam. Nov. and Caldisphaeraceae Fam. Nov.” *International Journal of Systematic and Evolutionary Microbiology* 59 (Pt 12) (December): 3116–22. doi:10.1099/ijs.0.010355-0. http://www.ncbi.nlm.nih.gov/pubmed/19643887.

Rachel, Reinhard, Irith Wyschkony, Sabine Riehl, and Harald Huber. 2002. “The Ultrastructure of Ignicoccus: Evidence for a Novel Outer Membrane and for Intracellular Vesicle Budding in an Archaeon.” *Archaea (Vancouver, B.C.)* 1 (1) (March): 9–18. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2685547&tool=pmcentrez&rendertype=abstract.

Reeburgh, William S. 2007. “Oceanic Methane Biogeochemistry.” *Chemical Reviews* 107 (2) (February): 486–513. doi:10.1021/cr050362v. http://www.ncbi.nlm.nih.gov/pubmed/17261072.

Reno, Michael L, Nicole L Held, Christopher J Fields, Patricia V Burke, and Rachel J Whitaker. 2009. “Biogeography of the Sulfolobus Islandicus Pan-genome.” *Proceedings of the National Academy of Sciences of the United States of America* 106 (21) (May 26): 8605–10. doi:10.1073/pnas.0808945106. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2689034&tool=pmcentrez&rendertype=abstract.

Reykjavı, Is-, V T Marteinsson, S Hauksdóttir, C F Hobel, H Kristmannsdóttir, G O Hreggvidsson, and J K Kristjánsson. 2001. “Phylogenetic Diversity Analysis of Subterranean Hot Springs in Iceland.” *Applied and Environmental Microbiology* 67 (9) (September): 4242–8.

doi:10.1128/AEM.67.9.4242. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=93153&tool=pmcentrez&rendertype=abstract.

Reysenbach, Anna-Louise, Yitai Liu, Amy B Banta, Terry J Beveridge, Julie D Kirshtein, Stefan Schouten, Margaret K Tivey, Karen L Von Damm, and Mary a Voytek. 2006. "A Ubiquitous Thermoacidophilic Archaeon from Deep-sea Hydrothermal Vents." *Nature* 442 (7101) (July 27): 444–7. doi:10.1038/nature04921. http://www.ncbi.nlm.nih.gov/pubmed/16871216.

Rinke, Christian, Patrick Schwientek, Alexander Sczyrba, Natalia N. Ivanova, Iain J. Anderson, Jan-Fang Cheng, Aaron Darling, et al. 2013. "Insights into the Phylogeny and Coding Potential of Microbial Dark Matter." *Nature* advance on (July 14): 1–7. doi:10.1038/nature12352. http://dx.doi.org/10.1038/nature12352.

Rivera, Maria C, and James a Lake. 2004. "The Ring of Life Provides Evidence for a Genome Fusion Origin of Eukaryotes." *Nature* 431 (7005) (September 9): 152–5. doi:10.1038/nature02848. http://www.ncbi.nlm.nih.gov/pubmed/15356622.

Roh, Seong Woon, Young-Do Nam, Seong-Hyeuk Nam, Sang-Haeng Choi, Hong-Seog Park, and Jin-Woo Bae. 2010. "Complete Genome Sequence of Halalkalicoccus Jeotgali B3(T), an Extremely Halophilic Archaeon." *Journal of Bacteriology* 192 (17) (September): 4528–9. doi:10.1128/JB.00663-10. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2937367&tool=pmcentrez&rendertype=abstract.

Ronquist, Fredrik, Maxim Teslenko, Paul van der Mark, Daniel L Ayres, Aaron Darling, Sebastian Höhna, Bret Larget, Liang Liu, Marc A Suchard, and John P Huelsenbeck. 2012. "MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space." *Systematic Biology* 61 (3) (May): 539–42. doi:10.1093/sysbio/sys029. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3329765&tool=pmcentrez&rendertype=abstract.

Rudolph, Christian, Gerhard Wanner, and Robert Huber. 2001. "Natural Communities of Novel Archaea and Bacteria Growing in Cold Sulfurous Springs with a String-of-Pearls-Like Morphology" 67 (5): 2336–2344. doi:10.1128/AEM.67.5.2336.

Saitou, N, and M Nei. 1987. "The Neighbor-joining Method: a New Method for Reconstructing Phylogenetic Trees." *Molecular Biology and Evolution* 4 (4) (July): 406–25. http://www.ncbi.nlm.nih.gov/pubmed/3447015.

Sakai, Sanae, Hiroyuki Imachi, Satoshi Hanada, Akiyoshi Ohashi, Hideki Harada, and Yoichi Kamagata. 2008. "Methanocella Paludicola Gen. Nov., Sp. Nov., a Methane-producing Archaeon, the First Isolate of the Lineage 'Rice Cluster I', and Proposal of the New Archaeal Order Methanocellales Ord. Nov." *International Journal of Systematic and Evolutionary Microbiology* 58 (Pt 4) (April): 929–36. doi:10.1099/ijs.0.65571-0. http://www.ncbi.nlm.nih.gov/pubmed/18398197.

Schleper, C, G Puehler, I Holz, a Gambacorta, D Janekovic, U Santarius, H P Klenk, and W Zillig. 1995. "Picrophilus Gen. Nov., Fam. Nov.: a Novel Aerobic, Heterotrophic, Thermoacidophilic Genus and Family Comprising Archaea Capable of Growth Around pH 0." *Journal of Bacteriology* 177 (24) (December): 7050–9. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=177581&tool=pmcentrez&rendertype=abstract.

Schleper, Christa, German Jurgens, and Melanie Jonuscheit. 2005. “Genomic Studies of Uncultivated Archaea.” *Nature Reviews. Microbiology* 3 (6) (June): 479–88. doi:10.1038/nrmicro1159. http://www.ncbi.nlm.nih.gov/pubmed/15931166.

Shi, Yanmei, Gene W Tyson, and Edward F DeLong. 2009. “Metatranscriptomics Reveals Unique Microbial Small RNAs in the Ocean’s Water Column.” *Nature* 459 (7244) (May 14): 266–9. doi:10.1038/nature08055. http://www.ncbi.nlm.nih.gov/pubmed/19444216.

Siebers, Bettina, Melanie Zaparty, Guenter Raddatz, Britta Tjaden, Sonja-Verena Albers, Steve D Bell, Fabian Blombach, et al. 2011. “The Complete Genome Sequence of Thermoproteus Tenax: a Physiologically Versatile Member of the Crenarchaeota.” *PloS One* 6 (10) (January): e24222. doi:10.1371/journal.pone.0024222. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3189178&tool=pmcentrez&rendertype=abstract.

Slesarev, Alexei I, Katja V Mezhevaya, Kira S Makarova, Nikolai N Polushin, Olga V Shcherbinina, Vera V Shakhova, Galina I Belova, 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 (7) (April 2): 4644–9. doi:10.1073/pnas.032671499. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=123701&tool=pmcentrez&rendertype=abstract.

Sogin, S J, M L Sogin, and C R Woese. 1972. “Phylogenetic Measurement in Procaryotes by Primary Structural Characterization.” *Journal of Molecular Evolution* 1 (1) (January): 173–84. http://www.ncbi.nlm.nih.gov/pubmed/5006250.

Spring, Stefan, Reinhard Rachel, Alla Lapidus, Karen Davenport, Hope Tice, Alex Copeland, Jan-Fang Cheng, et al. 2010. “Complete Genome Sequence of Thermosphaera Aggregans Type Strain (M11TL).” *Standards in Genomic Sciences* 2 (3) (January): 245–59. doi:10.4056/sigs.821804. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3035292&tool=pmcentrez&rendertype=abstract.

Stahl, David a, and José R de la Torre. 2012. “Physiology and Diversity of Ammonia-oxidizing Archaea.” *Annual Review of Microbiology* 66 (January): 83–101. doi:10.1146/annurev-micro-092611-150128. http://www.ncbi.nlm.nih.gov/pubmed/22994489.

Stamatakis, Alexandros. 2006. “RAxML-VI-HPC: Maximum Likelihood-based Phylogenetic Analyses with Thousands of Taxa and Mixed Models.” *Bioinformatics (Oxford, England)* 22 (21) (November 1): 2688–90. doi:10.1093/bioinformatics/btl446. http://www.ncbi.nlm.nih.gov/pubmed/16928733.

Stamatakis, Alexandros, Paul Hoover, and Jacques Rougemont. 2008. “A Rapid Bootstrap Algorithm for the RAxML Web Servers.” *Systematic Biology* 57 (5) (October 1): 758–71. doi:10.1080/10635150802429642. http://sysbio.oxfordjournals.org/content/57/5/758.short.

Stanier, R. Y., Doudoroff, M. & Adelberg, E. A. 1963. *The Microbial World 2nd Edn*. Prentice-Hall, Englewood Cliffs.

Stein, J L, T L Marsh, K Y Wu, H Shizuya, and E F DeLong. 1996. “Characterization of Uncultivated Prokaryotes: Isolation and Analysis of a 40-kilobase-pair Genome Fragment from a Planktonic Marine Archaeon.” *Journal of Bacteriology* 178 (3) (February): 591–9.

http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=177699&tool=pmcentrez&rendertype=abstract.

Susanti, Dwi, Eric F Johnson, Jason R Rodriguez, Iain Anderson, Anna a Perevalova, Nikos Kyrpides, Susan Lucas, et al. 2012. "Complete Genome Sequence of Desulfurococcus Fermentans, a Hyperthermophilic Cellulolytic Crenarchaeon Isolated from a Freshwater Hot Spring in Kamchatka, Russia." *Journal of Bacteriology* 194 (20) (October): 5703–4. doi:10.1128/JB.01314-12. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3458677&tool=pmcentrez&rendertype=abstract.

Takai, K, and K Horikoshi. 1999. "Genetic Diversity of Archaea in Deep-sea Hydrothermal Vent Environments." *Genetics* 152 (4) (August): 1285–97. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1460697&tool=pmcentrez&rendertype=abstract.

Tatusov, R L, E V Koonin, and D J Lipman. 1997. "A Genomic Perspective on Protein Families." *Science (New York, N.Y.)* 278 (5338) (October 24): 631–7. http://www.ncbi.nlm.nih.gov/pubmed/9381173.

Tatusov, R L, D a Natale, I V Garkavtsev, T a Tatusova, U T Shankavaram, B S Rao, B Kiryutin, M Y Galperin, N D Fedorova, and E V Koonin. 2001. "The COG Database: New Developments in Phylogenetic Classification of Proteins from Complete Genomes." *Nucleic Acids Research* 29 (1) (January 1): 22–8. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=29819&tool=pmcentrez&rendertype=abstract.

Tatusov, Roman L, Natalie D Fedorova, John D Jackson, Aviva R Jacobs, Boris Kiryutin, Eugene V Koonin, Dmitri M Krylov, et al. 2003. "The COG Database: An Updated Version Includes Eukaryotes." *BMC Bioinformatics* 4 (September 11): 41. doi:10.1186/1471-2105-4-41. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=222959&tool=pmcentrez&rendertype=abstract.

Tourna, Maria, Michaela Stieglmeier, Anja Spang, Martin Könneke, Arno Schintlmeister, Tim Urich, Marion Engel, et al. 2011. "Nitrososphaera Viennensis, an Ammonia Oxidizing Archaeon from Soil." *Proceedings of the National Academy of Sciences of the United States of America* 108 (20) (May 17): 8420–5. doi:10.1073/pnas.1013488108. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3100973&tool=pmcentrez&rendertype=abstract.

Treusch, Alexander H, Sven Leininger, Arnulf Kletzin, Stephan C Schuster, Hans-Peter Klenk, and Christa Schleper. 2005. "Novel Genes for Nitrite Reductase and Amo-related Proteins Indicate a Role of Uncultivated Mesophilic Crenarchaeota in Nitrogen Cycling." *Environmental Microbiology* 7 (12) (December): 1985–95. doi:10.1111/j.1462-2920.2005.00906.x. http://www.ncbi.nlm.nih.gov/pubmed/16309395.

Von Jan, Mathias, Alla Lapidus, Tijana Glavina Del Rio, Alex Copeland, Hope Tice, Jan-Fang Cheng, Susan Lucas, et al. 2010. "Complete Genome Sequence of Archaeoglobus Profundus Type Strain (AV18)." *Standards in Genomic Sciences* 2 (3) (January): 327–46. doi:10.4056/sigs.942153. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3035285&tool=pmcentrez&rendertype=abstract.

Walker, C B, J R de la Torre, M G Klotz, H Urakawa, N Pinel, D J Arp, C Brochier-Armanet, et al. 2010. "Nitrosopumilus Maritimus Genome Reveals Unique Mechanisms for Nitrification and Autotrophy in Globally Distributed Marine Crenarchaea." *Proceedings of the National Academy of Sciences of the United States of America* 107 (19) (May 11): 8818–23. doi:10.1073/pnas.0913533107. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2889351&tool=pmcentrez&rendertype=abstract.

Waters, Elizabeth, Michael J Hohn, Ivan Ahel, David E Graham, Mark D Adams, Mary Barnstead, Karen Y Beeson, 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 (22) (October 28): 12984–8. doi:10.1073/pnas.1735403100. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=240731&tool=pmcentrez&rendertype=abstract.

Wheelis, M L, O Kandler, and C R Woese. 1992. "On the Nature of Global Classification." *Proceedings of the National Academy of Sciences of the United States of America* 89 (April): 2930–4. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=48777&tool=pmcentrez&rendertype=abstract.

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 (3863) (January 10): 150–60. http://www.ncbi.nlm.nih.gov/pubmed/5762760.

Williams, Tom a, Peter G Foster, Tom M W Nye, Cymon J Cox, and T Martin Embley. 2012. "A Congruent Phylogenomic Signal Places Eukaryotes Within the Archaea." *Proceedings. Biological Sciences / The Royal Society* 279 (1749) (December 22): 4870–9. doi:10.1098/rspb.2012.1795. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3497233&tool=pmcentrez&rendertype=abstract.

Woese, C R, and G E Fox. 1977. "Phylogenetic Structure of the Prokaryotic Domain: The Primary Kingdoms." *Proceedings of the National Academy of Sciences of the United States of America* 74 (11) (November): 5088–90. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=432104&tool=pmcentrez&rendertype=abstract.

Woese, C R, O Kandler, and M L Wheelis. 1990. "Towards a Natural System of Organisms: Proposal for the Domains Archaea, Bacteria, and Eucarya." *Proceedings of the National Academy of Sciences of the United States of America* 87 (12) (June): 4576–9. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=54159&tool=pmcentrez&rendertype=abstract.

Woese, C R, L J Magrum, and G E Fox. 1978. "Archaebacteria." *Journal of Molecular Evolution* 11 (3) (August 2): 245–51. http://www.ncbi.nlm.nih.gov/pubmed/691075.

Woese, C R, J Maniloff, and L B Zablen. 1980. "Phylogenetic Analysis of the Mycoplasmas." *Proceedings of the National Academy of Sciences of the United States of America* 77 (1) (January): 494–8.

http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=348298&tool=pmcentrez&rendertype=abstract.

Woese, C R, G J Olsen, M Ibba, and D Söll. 2000. “Aminoacyl-tRNA Synthetases, the Genetic Code, and the Evolutionary Process.” *Microbiology and Molecular Biology Reviews : MMBR* 64 (1) (March): 202–36. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=98992&tool=pmcentrez&rendertype=abstract.

Woese, Carl R, Kenneth R Luehrsen, Cheryl D Pribula, and George E Fox. 1976. “Molecular Evolution Sequence Characterization of 5S Ribosomal RNA from Eight Gram Positive Procaryotes” 8: 143–153.

Woyke, Tanja, Damon Tighe, Konstantinos Mavromatis, Alicia Clum, Alex Copeland, Wendy Schackwitz, Alla Lapidus, et al. 2010. “One Bacterial Cell, One Complete Genome.” *PloS One* 5 (4) (January): e10314. doi:10.1371/journal.pone.0010314. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2859065&tool=pmcentrez&rendertype=abstract.

Yutin, Natalya, Kira S Makarova, Sergey L Mekhedov, Yuri I Wolf, and Eugene V Koonin. 2008. “The Deep Archaeal Roots of Eukaryotes.” *Molecular Biology and Evolution* 25 (8) (August): 1619–30. doi:10.1093/molbev/msn108. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2464739&tool=pmcentrez&rendertype=abstract.

Zillig, W., K.O. Stetter, W. Schäfer, D. Janekovic, S. Wunderl, I. Holz, and P. Palm. 1981. “Thermoproteales: A Novel Type of Extremely Thermoacidophilic Anaerobic Archaebacteria Isolated from Icelandic Solfataras.” *Zentralblatt Für Bakteriologie Mikrobiologie Und Hygiene: I. Abt. Originale C: Allgemeine, Angewandte Und Ökologische Mikrobiologie* 2 (3): 227–205. doi:10.1016/S0721-9571(81)80001-4. http://dx.doi.org/10.1016/S0721-9571(81)80001-4.

Zuckerkandl, E, and L Pauling. 1965. “Molecules as Documents of Evolutionary History.” *Journal of Theoretical Biology* 8 (2) (March): 357–66. http://www.ncbi.nlm.nih.gov/pubmed/5876245.

Zvyagintseva, I.S., and A.L. Tarasov. 1988. “Extreme Halophilic Bacteria from Saline Soils.” *Microbiology* v. 56(5) p (March 1). http://agris.fao.org/agris-search/search/display.do?f=1989/US/US89321.xml;US8858510.

Annexes

Table des Annexes

Annexe 1 : Matériels supplémentaires de l'Article 1 (Chapitre 1).

Supplementary material.

Supplementary Table S1. Table showing the taxonomic sampling used for the identification of new markers that can be potentially used to study the phylogeny of Archaea.

Supplementary Table S2. Table showing the 200 new proteins used for the phylogenetic analysis.

Supplementary Table S3. Table showing the taxonomic distribution of the 129 archaeal genomes used for the assembly of final datasets.

Supplementary Figure S1. Unrooted maximum likelihood phylogeny of Archaea based on the 200 newly identified proteins (L2 supermatrix, 48,904 amino acid positions, 122 species). The tree was inferred with PhyML (LG+Γ8). The scale bar represents the average number of substitutions per site. Numbers at branches are SH-like supports

Supplementary Figure S2. Unrooted maximum likelihood phylogeny of Archaea based on the 53 ribosomal proteins (L3 supermatrix, 6,228 amino acid positions, 122 organisms). The tree was inferred with PhyML (LG+Γ8). The scale bar represents the average number of substitutions per site. Numbers at branches are SH-like supports.

Supplementary Figure S3. Unrooted maximum likelihood phylogeny of Archaea based on the 15 subunits of the RNA polymerase and transcription factors (L4 supermatrix, 2,970 amino acid positions, 122 species). The tree was inferred with PhyML (LG+Γ8). The scale bar represents the average number of substitutions per site. Numbers at branches are SH-like supports.

Supplementary Figure S4. Unrooted Bayesian tree of Archaea based on the 15 subunits of the RNA polymerase and transcription factors (L4 supermatrix, 2970 amino acid positions, 122 species). The tree was inferred with PhyloBayes (CAT+Γ8). The scale bar represents the average number of substitutions per site. Numbers at branches are posterior probabilities.

Supplementary Figure S5. Unrooted maximum likelihood trees resulting from the site-by-site desaturation strategy. The S_0 to S_{34} supermatrices were extracted from the XL1 supermatrix (see material and methods). Trees were inferred with PhyML (CAT+Γ8). The scale bar represents the average number of substitutions per site. Numbers at branches are SH supports.

Eucarya

Allomyces macrogynus
Arabidopsis thaliana
Arabidopsis thaliana Plastid
Aspergillus niger
Aureococcus anophagefferens
Aureococcus anophagefferens Plastid
Batrachochytrium dendrobatidis JAM81
Bigelowiella natans Nucleomorph
Bigelowiella natans Plastid
Blastocystis hominis
Bodo saltans
Botrytis cinerea
Brachypodium distachyon
Brachypodium distachyon Plastid
Caenorhabditis elegans
Calliarthron tuberculosum
Capsaspora owczarzaki
Chaetomium globosum
Chlamydomonas reinhardtii
Chlamydomonas reinhardtii Plastid
Chlorella sp. NC64A
Chlorella vulgaris C-169
Chlorella vulgaris Plastid
Chondrus crispus EST
Chromera velia Plastid
Coprinopsis cinerea
Cryptococcus neoformans JEC21
Cryptosporidium parvum
Cyanidioschyzon merolae
Cyanidioschyzon merolae Plastid
Cyanophora paradoxa
Cyanophora paradoxa EST
Cyanophora paradoxa Plastid
Cyanophora paradoxa T
Daphnia pulex
Debaryomyces hansenii CBS767
Dictyostelium discoideum
Dictyostelium purpureum QSDP1
Diplonema papillatum
Ectocarpus siliculosus
Ectocarpus siliculosus Plastid
Emiliania huxleyi CCMP1516
Emiliania huxleyi Plastid
Entamoeba histolytica
Eucheuma denticulatum EST
Euglena gracilis
Fragilariopsis cylindrus
Furcellaria lumbricalis EST
Fusarium graminearum
Galdieria sulphuraria
Giardia lamblia
Gracilaria changii EST
Griffithsia okiensis EST
Guillardia theta Nucleomorph
Guillardia theta Plastid
Helobdella robusta

Histoplasma capsulatum
Homo sapiens
Laccaria bicolor
Leishmania infantum
Leishmania major strain Friedlin
Melampsora larici-populina
Micromonas pusilla CCMP1545
Micromonas pusilla CCMP1545 Plastid
Micromonas sp. RCC299 Plastid
Micromonas strain RCC299
Mimulus guttatus
Monosiga brevicollis
Mucor circinelloides
Mycosphaerella graminicola
Naegleria gruberi
Nematostella vectensis
Neurospora crassa OR74A
Oryza sativa
Oryza sativa Plastid
Ostreococcus lucimarinus
Ostreococcus tauri
Ostreococcus tauri Plastid
Paramecium tetraurelia
Perkinsus marinus
Phaeodactylum tricornutum
Phaeodactylum tricornutum Plastid
Phanerochaete chrisosporium
Phycomyces blakesleeanus
Physcomitrella patens
Physcomitrella patens Plastid
Phytophthora capsici
Phytophthora sojae
Plasmodium falciparum 3D7
Populus trichocarpa
Populus trichocarpa Plastid
Porphyra sp EST
Porphyra yezoensis Plastid
Porphyridium cruentum EST
Postia placenta
Procryptobia sorokini
Puccinia graminis
Rhizopus oryzae
Rhynchomonas nasuta
Saccharomyces cerevisiae RM11-1a
Salpingoeca rosetta
Schizophylum commune
Schizosaccharomyces pombe
Selaginella moellendorffii
Selaginella moellendorffii Plastid
Spizellomyces punctatus
Sporobolomyces roseus
Stagonospora nodorum
Tetrahymena thermophila
Thalassiosira pseudonana
Thalassiosira pseudonana Plastid
Thecamonas trahens
Toxoplasma gondii GT1
Toxoplasma gondii Plastid

Trichoderma reesei
Trichomonas vaginalis G3
Trichoplax adhaerens Grell-BS-1999
Trypanosoma brucei TREU927
Trypanosoma cruzi strain CL Brener
Ustilago maydis
Volvox carteri
Yarrowia lipolytica

Bacteria

Acaryochloris marina MBIC11017
Acholeplasma laidlawii PG_8A
Acidiphilium cryptum JF_5
Acidithiobacillus ferrooxidans ATCC 23270
Acidobacteria bacterium Ellin345
Acidothermus cellulolyticus 11B
Acinetobacter baumannii AB0057
Actinobacillus pleuropneumoniae L20
Aeromonas hydrophila ATCC_7966
Agrobacterium tumefaciens C58
Akkermansia muciniphila ATCC BAA-835
Alcanivorax borkumensis SK2
Aliivibrio salmonicida LFI1238
Alkalilimnicola ehrlichei MLHE_1
Alkaliphilus metalliredigens QYMF
Alteromonas macleodii Deep ecotype
Amoebophilus asiaticus 5a2
Anabaena variabilis
Anaerocellum thermophilum DSM 6725
Anaeromyxobacter sp. Fw109_5
Anaplasma marginale StMaries
Anoxybacillus flavithermus WK1
Aquifex aeolicus VF5
Arcobacter butzleri RM4018
Aromatoleum aromaticum EbN1
Arthrobacter aurescens TC1
Aster yellows witches-broom phytoplasma AYWB
Azoarcus sp. BH72
Azorhizobium caulinodans ORS_571
Bacillus amyloliquefaciens FZB42
Bacteroides fragilis NCTC_9343
Bacteroides fragilis YCH46
Bartonella bacilliformis KC583
Baumannia cicadellinicola Hc
Bdellovibrio bacteriovorus HD100
Beijerinckia indica subsp. indica ATCC 9039
Bifidobacterium adolescentis ATCC_15703
Blochmannia floridanus
Bordetella bronchiseptica RB50
Borrelia afzelii PKo
Brachyspira hyodysenteriae WA1
Bradyrhizobium sp. BTAi1
Brucella abortus 9_941
Buchnera aphidicola str. APS
Burkholderia sp. 383
Caldicellulosiruptor saccharolyticus DSM_8903

Campylobacter concisus 13826
Carboxydothermus hydrogenoformans Z_2901
Carsonella ruddii PV
Caulobacter crescentus CB15
Cellvibrio japonicus Ueda107
Chlamydia muridarum Nigg
Chlamydia trachomatis 70
Chlamydophila abortus S26_3
Chlamydophila pneumoniae CWL029
Chlorobaculum parvum NCIB 8327
Chlorobium chlorochromatii CaD3
Chloroflexus aurantiacus J_10_fl
Chloroherpeton thalassium ATCC 35110
Chromobacterium violaceum ATCC_12472
Chromohalobacter salexigens DSM 3043
Citrobacter koseri ATCC_BAA895
Clavibacter michiganensis NCPPB_382
Clostridium acetobutylicum ATCC_824
Colwellia psychrerythraea 34H
Coprothermobacter proteolyticus DSM 5265
Corynebacterium diphtheriae NCTC_13129
Coxiella burnetii RSA_493
Crocosphaera watsonii WH8501
Cronobacter sakazakii ATCC BAA-894
Cupriavidus taiwanensis
Cyanobacterium Yellowstone A
Cyanobacterium Yellowstone B
Cyanothece sp. ATCC51142
Cyanothece sp. CCY0110
Cytophaga hutchinsonii ATCC_33406
Dechloromonas aromatica RCB
Dehalococcoides sp. BAV1
Deinococcus geothermalis DSM_11300
Delftia acidovorans SPH_1
Desulfitobacterium hafniense Y51
Desulfococcus oleovorans Hxd3
Desulfotalea psychrophila LSv54
Desulfotomaculum reducens MI_1
Desulfovibrio desulfuricans G20
Desulfovibrio vulgaris DP4
Desulfovibrio vulgaris Hildenborough
Diaphorobacter sp. TPSY
Dichelobacter nodosus VCS1703A
Dictyoglomus thermophilum H-6-12
Dinoroseobacter shibae DFL_12
Ehrlichia canis Jake
Elusimicrobium minutum Pei191
Enterobacter sp. 638
Enterococcus faecalis V583
Erythrobacter litoralis HTCC2594
Escherichia coli 536
Exiguobacterium sibiricum 255-15
Fervidobacterium nodosum Rt17_B1
Finegoldia magna ATCC_29328
Flavobacterium johnsoniae UW101
Flavobacterium psychrophilum JIP02_86
Francisella novicida U112
Frankia alni ACN14a

Fusobacterium nucleatum ATCC_25586
Geobacillus kaustophilus HTA426
Geobacter metallireducens GS_15
Geobacter sulfurreducens PCA
Geobacter uraniumreducens Rf4
Gloeobacter violaceus PCC7421
Gluconacetobacter diazotrophicus PAl_5
Gramella forsetii KT0803
Granulibacter bethesdensis CGDNIH1
Haemophilus ducreyi 35000HP
Hahella chejuensis KCTC_2396
Haliangium ochraceum DSM 14365
Halorhodospira halophila SL1
Halothermothrix orenii H 168
Helicobacter acinonychis Sheeba
Heliobacterium modesticaldum Ice1
Herminiimonas arsenicoxydans
Herpetosiphon aurantiacus ATCC_23779
Hydrogenobaculum sp. Y04AAS1
Hyphomonas neptunium ATCC_15444
Idiomarina loihiensis L2TR
Jannaschia sp. CCS1
Janthinobacterium sp. Marseille
Kineococcus radiotolerans SRS30216
Klebsiella pneumoniae MGH_78578
Kocuria rhizophila DC2201
Lactobacillus acidophilus NCFM
Lactococcus lactis Il1403
Lawsonia intracellularis PHE_MN1_00
Legionella pneumophila Corby
Leifsonia xyli CTCB07
Leptospira borgpetersenii JB197
Leptothrix cholodnii SP-6
Leuconostoc citreum KM20
Listeria innocua Clip11262
Lyngbya sp PCC8106
Lysinibacillus sphaericus C3_41
Macrococcus caseolyticus JCSC5402
Magnetococcus sp. MC_1
Magnetospirillum magneticum AMB_1
Mannheimia succiniciproducens MBEL55E
Maricaulis maris MCS10
Marinobacter aquaeolei VT8
Marinomonas sp. MWYL1
Mesoplasma florum L1
Mesorhizobium sp. BNC1
Methylibium petroleiphilum PM1
Methylobacillus flagellatus KT
Methylobacterium extorquens PA1
Methylocella silvestris BL2
Methylococcus capsulatus Bath
Microcystis aeruginosa NIES843
Moorella thermoacetica ATCC_39073
Mycobacterium abscessus
Mycoplasma agalactiae PG2
Myxococcus xanthus DK_1622
Natranaerobius thermophilus JWNM-WN-LF
Nautilia profundicola AmH

Neisseria gonorrhoeae FA_1090
Neorickettsia sennetsu Miyayama
Nitratiruptor sp. SB155_2
Nitrobacter hamburgensis X14
Nitrosococcus oceani ATCC_19707
Nitrosomonas europaea ATCC_19718
Nitrosospira multiformis ATCC_25196
Nocardia farcinica IFM_10152
Nocardioides sp. JS614
Nodularia spumigena CCY9414
Nostoc punctiforme
Nostoc sp. PCC7120
Novosphingobium aromaticivorans DSM_12444
Oceanobacillus iheyensis HTE831
Ochrobactrum anthropi ATCC_49188
Oenococcus oeni PSU_1
Oligotropha carboxidovorans OM5
Onion yellows phytoplasma OY-M
Opitutus terrae PB90-1
Parabacteroides distasonis ATCC_8503
Parachlamydia acanthamoebae
Paracoccus denitrificans PD1222
Parvibaculum lavamentivorans DS1
Pasteurella multocida Pm70
Pectobacterium atrosepticum SCRI1043
Pediococcus pentosaceus ATCC_25745
Pelagibacter ubique HTCC1062
Pelobacter carbinolicus DSM_2380
Pelobacter propionicus DSM_2379
Pelodictyon luteolum DSM_273
Pelotomaculum thermopropionicum SI
Petrotoga mobilis SJ95
Phenylobacterium zucineum HLK1
Photobacterium profundum SS9
Photorhabdus luminescens TTO1
Phytoplasma australiense
Plesiocystis pacifica SIR-1
Polaromonas sp. JS666
Polynucleobacter sp. QLW_P1DMWA_1
Porphyromonas gingivalis W83
Prochlorococcus marinus AS9601
Prochlorococcus marinus CCMP1375
Prochlorococcus marinus CCMP1986
Prochlorococcus marinus MIT9215
Prochlorococcus marinus MIT9301
Prochlorococcus marinus MIT9303
Prochlorococcus marinus MIT9312
Prochlorococcus marinus MIT9313
Prochlorococcus marinus MIT9515
Prochlorococcus marinus NATL1A
Prochlorococcus marinus NATL2A
Propionibacterium acnes KPA171202
Prosthecochloris vibrioformis DSM_265
Proteus mirabilis HI4320
Protochlamydia amoebophila UWE25
Pseudoalteromonas atlantica T6c
Pseudomonas aeruginosa PAO1
Psychrobacter arcticus 273_4

Psychromonas ingrahamii 37
Ralstonia eutropha H16
Renibacterium salmoninarum ATCC_33209
Rhizobium etli CFN_42
Rhodobacter sphaeroides 2_4_1
Rhodococcus sp. RHA1
Rhodoferax ferrireducens T118
Rhodopirellula baltica SH_1
Rhodopseudomonas palustris BisA53
Rhodospirillum rubrum ATCC_11170
Rickettsia akari Hartford
Roseiflexus castenholzii DSM_13941
Roseobacter denitrificans OCh_114
Rubrobacter xylanophilus DSM_9941
Ruegeria sp. TM1040
Saccharophagus degradans 2_40
Saccharopolyspora erythraea NRRL_2338
Salinibacter ruber DSM_13855
Salinispora arenicola CNS_205
Salmonella enterica arizonae
Serratia proteamaculans 568
Shewanella amazonensis SB2B
Shigella boydii Sb227
Silicibacter pomeroyi DSS_3
Sinorhizobium medicae WSM419
Sodalis glossinidius
Solibacter usitatus Ellin6076
Sorangium cellulosum
Sphingomonas wittichii RW1
Sphingopyxis alaskensis RB2256
Staphylococcus aureus MRSA252
Stenotrophomonas maltophilia K279a
Streptococcus agalactiae 2603V_R
Streptomyces avermitilis MA_4680
Sulcia muelleri GWSS
Sulfurihydrogenibium sp. YO3AOP1
Sulfurimonas denitrificans DSM 1251
Sulfurovum sp. NBC37_1
Symbiobacterium thermophilum IAM_14863
Synechococcus elongatus PCC6301
Synechococcus elongatus PCC7942
Synechococcus sp WH5701
Synechococcus sp_CC9311
Synechococcus sp_CC9605
Synechococcus sp_CC9902
Synechococcus sp_RCC307
Synechococcus sp_WH7803
Synechococcus sp_WH8102
Synechocystis sp_PCC6803
Syntrophobacter fumaroxidans MPOB
Syntrophomonas wolfei subsp. wolfei str. Goettingen
Syntrophus aciditrophicus SB
Thermoanaerobacter pseudethanolicus ATCC_33223
Thermobifida fusca YX
Thermodesulfovibrio yellowstonii DSM 11347
Thermomicrobium roseum DSM 5159
Thermosipho melanesiensis BI429
Thermosynechococcus elongatus BP1

Thermotoga lettingae TMO
Thermus thermophilus HB8
Thioalkalivibrio sp. HL-EbGR7
Thiobacillus denitrificans ATCC_25259
Thiomicrospira crunogena XCL_2
Treponema denticola ATCC_35405
Trichodesmium erythraeum
Tropheryma whipplei TW08_27
Ureaplasma parvum ATCC_700970
Verminephrobacter eiseniae EF01_2
Vibrio cholerae N16961
Wigglesworthia glossinidia
Wolbachia endosymbiont TRS
Wolinella succinogenes DSM_1740
Xanthobacter autotrophicus Py2
Xanthomonas campestris ATCC_33913
Xylella fastidiosa 9a5c
Yersinia enterocolitica 8081
Zymomonas mobilis ZM4

Archaea

Acidilobus saccharovorans 345-15
Aciduliprofundum boonei T469
Aeropyrum pernix K1
Archaeoglobus fulgidus DSM_4304
Archaeoglobus profundus DSM 5631
Caldivirga maquilingensis IC_167
Candidatus Caldiarchaeum subterraneum
Candidatus Korarchaeum cryptofilum OPF8
Candidatus Methanoregula boonei 6A8
Candidatus Micrarchaeum acidiphilum ARMAN-2
Candidatus Parvarchaeum acidiphilum ARMAN-4
Candidatus Parvarchaeum acidophilus ARMAN-5
Cenarchaeum symbiosum A
Desulfurococcus kamchatkensis
Desulfurococcus mucosus DSM2162
Ferroglobus placidus DSM 10642
Ferroplasma acidarmanus fer1
Halalkalicoccus jeotgali B3
Haloarcula marismortui ATCC_43049
Halobacterium salinarum R1
Haloferax volcanii DS2
Halogeometricum borinquense DSM 11551
Halomicrobium mukohataei DSM 12286
Haloquadratum walsbyi DSM_16790
Halorhabdus utahensis DSM 12940
Halorubrum lacusprofundi ATCC 49239
Haloterrigena turkmenica DSM 5511
Hyperthermus butylicus DSM_5456
Ignicoccus hospitalis KIN4_I
Ignisphaera aggregans DSM17230
Metallosphaera sedula DSM_5348
Methanobrevibacter ruminantium M1
Methanobrevibacter smithii ATCC_35061
Methanocaldococcus fervens AG86
Methanocaldococcus infernus ME

Methanocaldococcus jannaschii DSM 2661
Methanocaldococcus sp. FS406-22
Methanocaldococcus vulcanius M7
Methanocella paludicola SANAE
Methanococcoides burtonii DSM_6242
Methanococcus aeolicus Nankai_3
Methanococcus maripaludis C5
Methanococcus vannielii
Methanococcus voltae A3
Methanocorpusculum labreanum Z
Methanoculleus marisnigri JR1
Methanohalobium evestigatum Z-7303
Methanohalophilus mahii DSM 5219
Methanoplanus petrolearius DSM11571
Methanopyrus kandleri AV19
Methanosaeta thermophila PT
Methanosarcina acetivorans C2A
Methanosarcina barkeri str. Fusaro
Methanosarcina mazei Go1
Methanosphaera stadtmanae DSM_3091
Methanosphaerula palustris E1-9c
Methanospirillum hungatei JF_1
Methanothermobacter marburgensis str. Marburg
Methanothermobacter thermautotrophicus
Methanothermococcus okinawensis IH1
Methanothermus fervidus
Nanoarchaeum equitans Kin4_M
Natrialba magadii ATCC 43099
Natronomonas pharaonis DSM_2160
Nitrosopumilus maritimus SCM1
Picrophilus torridus DSM_9790
Pyrobaculum aerophilum IM2
Pyrobaculum arsenaticum DSM 13514
Pyrobaculum calidifontis JCM 11548
Pyrobaculum islandicum DSM 4184
Pyrobaculum neutrophilum V24Sta
Pyrococcus abyssi GE5
Pyrococcus furiosus DSM 3638
Pyrococcus horikoshii OT3
Staphylothermus hellenicus DSM 12710
Staphylothermus marinus F1
Sulfolobus acidocaldarius DSM_639
Sulfolobus islandicus L.S.2.15
Sulfolobus solfataricus P2
Sulfolobus tokodaii str. 7
Thermococcus barophilus MP
Thermococcus gammatolerans EJ3
Thermococcus kodakarensis KOD1
Thermococcus onnurineus NA1
Thermococcus sibiricus MM 739
Thermococcus sp. AM4
Thermofilum pendens Hrk_5
Thermoplasma acidophilum DSM_1728
Thermoplasma volcanium GSS1
Thermosphaera aggregans DSM 11486
uncultured methanogenic archaeon RC-I
Vulcanisaeta distributa DSM14429

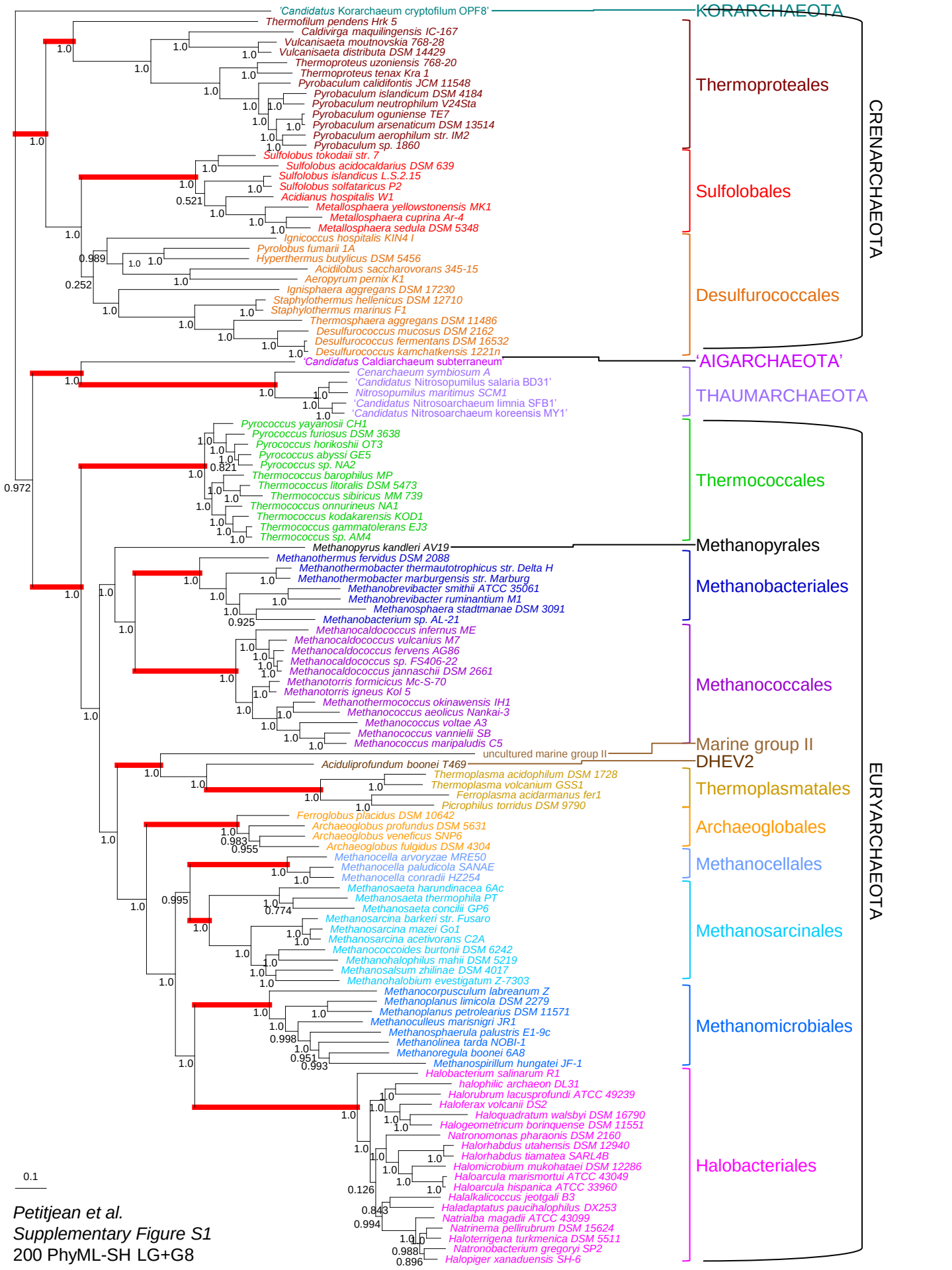
	Annotation	Reference gi	Reference accession	study annotation	COG number	COG class	COG definition	COG Classification	Supermatrice	Sequences
1	PP-loop domain-containing protein	161527617	YP_001581443.1	M002	COG0037	D	Cell cycle control, cell division, chromosome partitioning	CELLULAR PROCESSES AND SIGNALING	10	121
2	cell division protein FtsZ	315425227	BAJ46896.1	MA10	COG0206	D	Cell cycle control, cell division, chromosome partitioning	CELLULAR PROCESSES AND SIGNALING	Totale	63
3	hypothetical protein Nmar_0029	161527541	YP_001581367.1	M005	COG5491	N	Cell motility	CELLULAR PROCESSES AND SIGNALING	Totale	25
4	peptidase M50	161527619	YP_001581445.1	M003	COG0750	M	Cell wall/membrane/envelope biogenesis	CELLULAR PROCESSES AND SIGNALING	10	128
5	peptidase M50	161527799	YP_001581625.1	M004	COG0750	M	Cell wall/membrane/envelope biogenesis	CELLULAR PROCESSES AND SIGNALING	Totale	66
6	glucosamine--fructose-6-phosphate aminotransferase isomerizing	315426952	BAJ48571.1	MA48	COG0449	M	Cell wall/membrane/envelope biogenesis	CELLULAR PROCESSES AND SIGNALING	Totale	109
7	preprotein translocase SecY subunit	161527910	YP_001581736.1	M027	COG0201	U	Intracellular trafficking, secretion, and vesicular transport	CELLULAR PROCESSES AND SIGNALING	10	128
8	GTP-binding signal recognition particle	161528039	YP_001581865.1	M028	COG0541	U	Intracellular trafficking, secretion, and vesicular transport	CELLULAR PROCESSES AND SIGNALING	10	126
9	signal recognition particle receptor	315425933	BAJ47583.1	MA23	COG0552	U	Intracellular trafficking, secretion, and vesicular transport	CELLULAR PROCESSES AND SIGNALING	10	126
10	signal recognition particle subunit SRP54	315425555	BAJ47216.1	MA56	COG0541	U	Intracellular trafficking, secretion, and vesicular transport	CELLULAR PROCESSES AND SIGNALING	10	126
11	metalloendopeptidase glycoprotease family	161529041	YP_001582867.1	M010	COG0533	O	Post-translational modification, protein turnover, and chaperones	CELLULAR PROCESSES AND SIGNALING	10	128
12	peptidylprolyl isomerase FKBP-type	161529238	YP_001583064.1	M011	COG1047	O	Post-translational modification, protein turnover, and chaperones	CELLULAR PROCESSES AND SIGNALING	10	126
13	proteasome subunit alpha	161528319	YP_001582145.1	M014	COG0638	O	Post-translational modification, protein turnover, and chaperones	CELLULAR PROCESSES AND SIGNALING	10	127
14	proteasome endopeptidase complex	161529236	YP_001583062.1	M018	COG0638	O	Post-translational modification, protein turnover, and chaperones	CELLULAR PROCESSES AND SIGNALING	10	129
15	prefoldin chaperonin cofactor	118575219	YP_874962.1	M021	COG1382	O	Post-translational modification, protein turnover, and chaperones	CELLULAR PROCESSES AND SIGNALING	10	127
16	PUA domain-containing protein	161529111	YP_001582937.1	M022	COG1370	O	Post-translational modification, protein turnover, and chaperones	CELLULAR PROCESSES AND SIGNALING	10	119
17	proteasome endopeptidase complex	161528202	YP_001582028.1	M012	COG0638	O	Post-translational modification, protein turnover, and chaperones	CELLULAR PROCESSES AND SIGNALING	Totale	100
18	hydrogenase maturation protein HypF	315426449	BAJ48087.1	MA32	COG0068	O	Post-translational modification, protein turnover, and chaperones	CELLULAR PROCESSES AND SIGNALING	Totale	77
19	hydrogenase expressionformation protein HypD	315426462	BAJ48095.1	MA33	COG0409	O	Post-translational modification, protein turnover, and chaperones	CELLULAR PROCESSES AND SIGNALING	Totale	77
20	hydrogenase expressionformation protein HypE	315426464	BAJ48097.1	MA34	COG0309	O	Post-translational modification, protein turnover, and chaperones	CELLULAR PROCESSES AND SIGNALING	Totale	77
21	pyruvate formate-lyase activating enzyme	315427296	BAJ48908.1	MA55	COG1180	O	Post-translational modification, protein turnover, and chaperones	CELLULAR PROCESSES AND SIGNALING	Totale	98
22	Mn2+-dependent serine/threonine protein kinase	161529039	YP_001582865.1	M025	COG3642	T	Signal transduction mechanisms	CELLULAR PROCESSES AND SIGNALING	10	127
23	non-specific serine/threonine protein kinase	161528918	YP_001582744.1	M026	COG1718	TD	Signal transduction mechanisms / Cell cycle control, cell division, chromosome partitioning	CELLULAR PROCESSES AND SIGNALING	10	126
24	proliferating cell nuclear antigen PCNA	315426658	BAJ48284.1	MA39	COG0592	L	Replication, recombination and repair	INFORMATION STORAGE AND PROCESSING	10	128
25	ATPase central domain-containing protein	161527778	YP_001581604.1	M116	COG0470	L	Replication, recombination and repair	INFORMATION STORAGE AND PROCESSING	10	127
26	DNA ligase I ATP-dependent Dnl1	161528545	YP_001582371.1	M117	COG1793	L	Replication, recombination and repair	INFORMATION STORAGE AND PROCESSING	10	125
27	DNA primase	161528810	YP_001582636.1	M119	COG0358	L	Replication, recombination and repair	INFORMATION STORAGE AND PROCESSING	10	127
28	DNA primase small subunit	161527578	YP_001581404.1	M121	COG1467	L	Replication, recombination and repair	INFORMATION STORAGE AND PROCESSING	10	124
29	DNA repair and recombination protein RadA	161528894	YP_001582720.1	M122	COG0468	L	Replication, recombination and repair	INFORMATION STORAGE AND PROCESSING	10	129
30	DNA topoisomerase VI B subunit	161528916	YP_001582742.1	M123	COG1389	L	Replication, recombination and repair	INFORMATION STORAGE AND PROCESSING	10	125
31	DNA topoisomerase VI subunit A	161528915	YP_001582741.1	M124	COG1697	L	Replication, recombination and repair	INFORMATION STORAGE AND PROCESSING	10	125
32	MCM family protein	161527750	YP_001581576.1	M125	COG1241	L	Replication, recombination and repair	INFORMATION STORAGE AND PROCESSING	10	129
33	replication factor C	161527748	YP_001581574.1	M126	COG0470	L	Replication, recombination and repair	INFORMATION STORAGE AND PROCESSING	10	127
34	flap endonuclease-1	161528211	YP_001582037.1	M128	COG0258	L	Replication, recombination and repair	INFORMATION STORAGE AND PROCESSING	10	127
35	methyltransferase	315426518	BAJ48149.1	MA35	COG1041	L	Replication, recombination and repair	INFORMATION STORAGE AND PROCESSING	10	120
36	DNA polymerase II large subunit	161529302	YP_001583128.1	M118	COG1933	L	Replication, recombination and repair	INFORMATION STORAGE AND PROCESSING	Totale	96
37	RecA/RadA recombinase-like protein	161527599	YP_001581425.1	M127	COG0468	L	Replication, recombination and repair	INFORMATION STORAGE AND PROCESSING	Totale	90
38	DEAD/DEAH box helicase domain-containing protein	161528452	YP_001582278.1	M129	COG1111	L	Replication, recombination and repair	INFORMATION STORAGE AND PROCESSING	Totale	93
39	ERCC4 domain-containing protein	161527945	YP_001581771.1	M130	COG1948	L	Replication, recombination and repair	INFORMATION STORAGE AND PROCESSING	Totale	115
40	DNA repair protein RadA	315427207	BAJ48820.1	MA54	COG0468	L	Replication, recombination and repair	INFORMATION STORAGE AND PROCESSING	Totale	128
41	RNA 3p-terminal phosphate cyclase	118576735	YP_876478.1	M030	COG0430	A	RNA processing and modification	INFORMATION STORAGE AND PROCESSING	Totale	101
42	2-alkenal reductase	161528106	YP_001581932.1	M080	COG2101	K	Transcription	INFORMATION STORAGE AND PROCESSING	10	129
43	hypothetical protein Nmar_0055	161527567	YP_001581393.1	M085	COG1293	K	Transcription	INFORMATION STORAGE AND PROCESSING	10	129
44	like-Sm ribonucleoprotein core	161527729	YP_001581555.1	M086	COG1958	K	Transcription	INFORMATION STORAGE AND PROCESSING	10	119
45	transcription factor TFIIE alpha subunit	161529115	YP_001582941.1	M092	COG1675	K	Transcription	INFORMATION STORAGE AND PROCESSING	10	120
46	XRE family transcriptional regulator	161528460	YP_001582286.1	M093	COG3620	K	Transcription	INFORMATION STORAGE AND PROCESSING	Totale	69
47	Alba DNA/RNA-binding protein	161527763	YP_001581589.1	M112	COG1581	K	Transcription	INFORMATION STORAGE AND PROCESSING	Totale	97
48	hypothetical conserved protein	68445573	BAE03281.1	MA01	COG1709	K	Transcription	INFORMATION STORAGE AND PROCESSING	Totale	104
49	archaea-specific DNA-binding protein AlbA	315426570	BAJ48200.1	MA37	COG1581	K	Transcription	INFORMATION STORAGE AND PROCESSING	Totale	99
50	ELP3 family histone acetyltransferase	161528087	YP_001581913.1	M115	COG1243	KB	Transcription / Chromatin structure and dynamics	INFORMATION STORAGE AND PROCESSING	Totale	114
51	alanyl-tRNA synthetase	161527876	YP_001581702.1	M031	COG0013	J	Translation, ribosomal structure and biogenesis	INFORMATION STORAGE AND PROCESSING	10	128
52	cytidylate kinase	161527913	YP_001581739.1	M033	COG2163	J	Translation, ribosomal structure and biogenesis	INFORMATION STORAGE AND PROCESSING	10	128
53	diphthamide biosynthesis protein	161529071	YP_001582897.1	M034	COG1736	J	Translation, ribosomal structure and biogenesis	INFORMATION STORAGE AND PROCESSING	10	128
54	diphthine synthase	161528805	YP_001582631.1	M035	COG1798	J	Translation, ribosomal structure and biogenesis	INFORMATION STORAGE AND PROCESSING	10	127
55	elongation factor 1-alpha	161528542	YP_001582368.1	M036	COG5256	J	Translation, ribosomal structure and biogenesis	INFORMATION STORAGE AND PROCESSING	10	129
56	elongation factor EF-2	161528256	YP_001582082.1	M037	COG0480	J	Translation, ribosomal structure and biogenesis	INFORMATION STORAGE AND PROCESSING	10	129
57	fibrillarin	161528900	YP_001582726.1	M039	COG1889	J	Translation, ribosomal structure and biogenesis	INFORMATION STORAGE AND PROCESSING	10	129
58	glutamyl-tRNA synthetase	161527819	YP_001581645.1	M040	COG0008	J	Translation, ribosomal structure and biogenesis	INFORMATION STORAGE AND PROCESSING	10	127
59	glutamyl-tRNA(Gln) amidotransferase subunit E	161527611	YP_001581437.1	M041	COG2511	J	Translation, ribosomal structure and biogenesis	INFORMATION STORAGE AND PROCESSING	10	127
60	glycyl-tRNA synthetase	161527575	YP_001581401.1	M042	COG0423	J	Translation, ribosomal structure and biogenesis	INFORMATION STORAGE AND PROCESSING	10	128
61	hypothetical protein Nmar_1382	161528890	YP_001582716.1	M043	COG1491	J	Translation, ribosomal structure and biogenesis	INFORMATION STORAGE AND PROCESSING	10	127
62	isoleucyl-tRNA synthetase	161528628	YP_001582454.1	M044	COG0060	J	Translation, ribosomal structure and biogenesis	INFORMATION STORAGE AND PROCESSING	10	127
63	peptide chain release factor 1	161528390	YP_001582216.1	M045	COG1503	J	Translation, ribosomal structure and biogenesis	INFORMATION STORAGE AND PROCESSING	10	128
64	Pre-mRNA processing ribonucleoprotein	161528899	YP_001582725.1	M046	COG1498	J	Translation, ribosomal structure and biogenesis	INFORMATION STORAGE AND PROCESSING	10	129
65	PUA domain-containing protein	161528203	YP_001582029.1	M047	COG2016	J	Translation, ribosomal structure and biogenesis	INFORMATION STORAGE AND PROCESSING	10	121
66	putative RNA-associated protein	161527938	YP_001581764.1	M048	COG1500	J	Translation, ribosomal structure and biogenesis	INFORMATION STORAGE AND PROCESSING	10	126
67	RNA-binding S1 domain-containing protein	161527589	YP_001581415.1	M049	COG1093	J	Translation, ribosomal structure and biogenesis	INFORMATION STORAGE AND PROCESSING	10	128
68	translation elongation factor aEF-1 beta	161528730	YP_001582556.1	M051	COG2888	J	Translation, ribosomal structure and biogenesis	INFORMATION STORAGE AND PROCESSING	10	127
69	translation initiation factor eIF-1A	161528034	YP_001581860.1	M052	COG0361	J	Translation, ribosomal structure and biogenesis	INFORMATION STORAGE AND PROCESSING	10	125
70	translation initiation factor eIF-5A	161528037	YP_001581863.1	M053	COG0231	J	Translation, ribosomal structure and biogenesis	INFORMATION STORAGE AND PROCESSING	10	129
71	translation initiation factor IF-2	161527737	YP_001581563.1	M054	COG0532	J	Translation, ribosomal structure and biogenesis	INFORMATION STORAGE AND PROCESSING	10	129
72	translation initiation factor IF-2 subunit gamma	161527584	YP_001581410.1	M055	COG5257	J	Translation, ribosomal structure and biogenesis	INFORMATION STORAGE AND PROCESSING	10	129
73	translation initiation factor IF-6	161527745	YP_001581571.1	M056	COG1976	J	Translation, ribosomal structure and biogenesis	INFORMATION STORAGE AND PROCESSING	10	126
74	translation initiation factor IF2/IF5	161528920	YP_001582746.1	M057	COG1601	J	Translation, ribosomal structure and biogenesis	INFORMATION STORAGE AND PROCESSING	10	128
75	tRNA (guanine-N(2)-)-methyltransferase	161528902	YP_001582728.1	M058	COG1867	J	Translation, ribosomal structure and biogenesis	INFORMATION STORAGE AND PROCESSING	10	129
76	tRNA cytidylyltransferase	161529290	YP_001583116.1	M059	COG1746	J	Translation, ribosomal structure and biogenesis	INFORMATION STORAGE AND PROCESSING	10	128

77	tRNA methyltransferase complex GCD14 subunit	161528198	YP_001582024.1	M060	COG2519	J	Translation, ribosomal structure and biogenesis	INFORMATION STORAGE AND PROCESSING	10	120
78	tRNA-guanine transglycosylase	161529044	YP_001582870.1	M061	COG0343	J	Translation, ribosomal structure and biogenesis	INFORMATION STORAGE AND PROCESSING	10	122
79	pseudouridylate synthase-like protein	161528040	YP_001581866.1	M068	COG1258	J	Translation, ribosomal structure and biogenesis	INFORMATION STORAGE AND PROCESSING	10	122
80	methionine aminopeptidase	161528109	YP_001581935.1	M069	COG0024	J	Translation, ribosomal structure and biogenesis	INFORMATION STORAGE AND PROCESSING	10	129
81	translation initiation factor SUI1	161528130	YP_001581956.1	M070	COG0023	J	Translation, ribosomal structure and biogenesis	INFORMATION STORAGE AND PROCESSING	10	128
82	MiaB-like tRNA modifying enzyme	161528769	YP_001582595.1	M071	COG0621	J	Translation, ribosomal structure and biogenesis	INFORMATION STORAGE AND PROCESSING	10	122
83	2p-5p RNA ligase	161529289	YP_001583115.1	M076	COG1514	J	Translation, ribosomal structure and biogenesis	INFORMATION STORAGE AND PROCESSING	10	124
84	phenylalanyl-tRNA synthetase alpha subunit	161528997	YP_001582823.1	M078	COG0016	J	Translation, ribosomal structure and biogenesis	INFORMATION STORAGE AND PROCESSING	10	128
85	phenylalanyl-tRNA synthetase beta chain	315425004	BAJ46679.1	MA03	COG0072	J	Translation, ribosomal structure and biogenesis	INFORMATION STORAGE AND PROCESSING	10	128
86	histidyl-tRNA synthetase	315425704	BAJ47360.1	MA19	COG0124	J	Translation, ribosomal structure and biogenesis	INFORMATION STORAGE AND PROCESSING	10	122
87	methionyl-tRNA synthetase	315426202	BAJ47845.1	MA30	COG0143	J	Translation, ribosomal structure and biogenesis	INFORMATION STORAGE AND PROCESSING	10	129
88	arginyl-tRNA synthetase	315426742	BAJ48366.1	MA40	COG0018	J	Translation, ribosomal structure and biogenesis	INFORMATION STORAGE AND PROCESSING	10	127
89	tRNA-intron endonuclease	315426913	BAJ48533.1	MA45	COG1676	J	Translation, ribosomal structure and biogenesis	INFORMATION STORAGE AND PROCESSING	10	122
90	beta-lactamase domain-containing protein	161529047	YP_001582873.1	M032	COG1236	J	Translation, ribosomal structure and biogenesis	INFORMATION STORAGE AND PROCESSING	Totale	93
91	exosome complex RNA-binding protein Rrp4	161527939	YP_001581765.1	M038	COG1097	J	Translation, ribosomal structure and biogenesis	INFORMATION STORAGE AND PROCESSING	Totale	86
92	transcription factor TFIIB cyclin-related	161527525	YP_001581351.1	M050	COG3277	J	Translation, ribosomal structure and biogenesis	INFORMATION STORAGE AND PROCESSING	Totale	107
93	glutamyl-tRNA(Gln) amidotransferase B subunit	161528376	YP_001582202.1	M064	COG0064	J	Translation, ribosomal structure and biogenesis	INFORMATION STORAGE AND PROCESSING	Totale	88
94	exosome complex exonuclease 1	161527940	YP_001581766.1	M066	COG0689	J	Translation, ribosomal structure and biogenesis	INFORMATION STORAGE AND PROCESSING	Totale	88
95	exosome complex RNA-binding protein Rrp42	161527941	YP_001581767.1	M067	COG2123	J	Translation, ribosomal structure and biogenesis	INFORMATION STORAGE AND PROCESSING	Totale	87
96	orotidine 5p-phosphate decarboxylase	161528926	YP_001582752.1	M075	COG1185	J	Translation, ribosomal structure and biogenesis	INFORMATION STORAGE AND PROCESSING	Totale	96
97	threonyl-tRNA synthetase	161528434	YP_001582260.1	M077	COG0441	J	Translation, ribosomal structure and biogenesis	INFORMATION STORAGE AND PROCESSING	Totale	89
98	prolyl-tRNA synthetase	315425642	BAJ47301.1	MA17	COG0442	J	Translation, ribosomal structure and biogenesis	INFORMATION STORAGE AND PROCESSING	Totale	89
99	5-nucleotidase SurE	315426747	BAJ48371.1	MA42	COG0522	J	Translation, ribosomal structure and biogenesis	INFORMATION STORAGE AND PROCESSING	Totale	86
100	chorismate synthase	315426893	BAJ48513.1	MA44	COG0594	J	Translation, ribosomal structure and biogenesis	INFORMATION STORAGE AND PROCESSING	Totale	109
101	serine hydroxymethyltransferase	161529301	YP_001583127.1	M140	COG0112	E	Amino acid transport and metabolism	METABOLISM	10	119
102	aspartyl-tRNA synthetase	161528574	YP_001582400.1	M144	COG0436	E	Amino acid transport and metabolism	METABOLISM	10	126
103	putative agmatinase	161528433	YP_001582259.1	M145	COG0010	E	Amino acid transport and metabolism	METABOLISM	10	123
104	argininosuccinate synthase	161528794	YP_001582620.1	M138	COG0137	E	Amino acid transport and metabolism	METABOLISM	Totale	103
105	homoserine kinase	161527553	YP_001581379.1	M142	COG0083	E	Amino acid transport and metabolism	METABOLISM	Totale	80
106	homoserine dehydrogenase	161528204	YP_001582030.1	M143	COG0460	E	Amino acid transport and metabolism	METABOLISM	Totale	109
107	aspartate kinase	161529262	YP_001583088.1	M146	COG0527	E	Amino acid transport and metabolism	METABOLISM	Totale	110
108	threonine synthase	315425153	BAJ46824.1	MA05	COG0498	E	Amino acid transport and metabolism	METABOLISM	Totale	91
109	S-adenosylmethionine synthetase	315425167	BAJ46837.1	MA08	COG1812	E	Amino acid transport and metabolism	METABOLISM	Totale	116
110	hypothetical protein HGMM_F27H06C04	315425678	BAJ47335.1	MA18	COG0367	E	Amino acid transport and metabolism	METABOLISM	Totale	87
111	ketol-acid reductoisomerase	161528467	YP_001582293.1	M147	COG0059	EH	Amino acid transport and metabolism / Coenzyme transport and metabolism	METABOLISM	Totale	98
112	acetolactate synthase large subunit biosynthetic type	161528580	YP_001582406.1	M148	COG0028	EH	Amino acid transport and metabolism / Coenzyme transport and metabolism	METABOLISM	Totale	96
113	glyceraldehyde-3-phosphate dehydrogenase	161528339	YP_001582165.1	M163	COG0057	G	Carbohydrate transport and metabolism	METABOLISM	10	120
114	phosphoglycerate kinase	161528008	YP_001581834.1	M165	COG0126	G	Carbohydrate transport and metabolism	METABOLISM	10	122
115	triosephosphate isomerase	161528458	YP_001582284.1	M166	COG0149	G	Carbohydrate transport and metabolism	METABOLISM	10	126
116	leucyl-tRNA synthetase	161527875	YP_001581701.1	M164	COG0800	G	Carbohydrate transport and metabolism	METABOLISM	Totale	112
117	phosphopantothenoylcysteine decarboxylase/phosphopantothenate--cysteine ligase	161529229	YP_001583055.1	M170	COG0452	H	Coenzyme transport and metabolism	METABOLISM	10	119
118	thiamine-monophosphate kinase	161528956	YP_001582782.1	M174	COG0611	H	Coenzyme transport and metabolism	METABOLISM	10	119
119	3 4-dihydroxy-2-butanone 4-phosphate synthase	161529124	YP_001582950.1	M168	COG0108	H	Coenzyme transport and metabolism	METABOLISM	Totale	107
120	pyridoxine biosynthesis protein	161528987	YP_001582813.1	M171	COG0214	H	Coenzyme transport and metabolism	METABOLISM	Totale	112
121	hypothetical protein Nmar_0322	161527830	YP_001581656.1	M173	COG2038	H	Coenzyme transport and metabolism	METABOLISM	Totale	101
122	UbiD family decarboxylase	161529105	YP_001582931.1	M175	COG0043	H	Coenzyme transport and metabolism	METABOLISM	Totale	93
123	porphobilinogen deaminase	118576540	YP_876283.1	M176	COG0181	H	Coenzyme transport and metabolism	METABOLISM	Totale	98
124	6 7-dimethyl-8-ribityllumazine synthase	161529123	YP_001582949.1	M177	COG0054	H	Coenzyme transport and metabolism	METABOLISM	Totale	109
125	thiamine biosynthesis protein ThiC	161528051	YP_001581877.1	M178	COG0422	H	Coenzyme transport and metabolism	METABOLISM	Totale	92
126	glutamate-1-semialdehyde-2 1-aminomutase	161527998	YP_001581824.1	M179	COG0001	H	Coenzyme transport and metabolism	METABOLISM	Totale	98
127	SNO glutamine amidotransferase	161528986	YP_001582812.1	M180	COG0311	H	Coenzyme transport and metabolism	METABOLISM	Totale	112
128	adenosylhomocysteinase	315425904	BAJ47555.1	MA22	COG0499	H	Coenzyme transport and metabolism	METABOLISM	Totale	115
129	molybdenum cofactor biosynthesis protein C	315427188	BAJ48802.1	MA53	COG0315	H	Coenzyme transport and metabolism	METABOLISM	Totale	115
130	3-dehydroquinate synthase	161529237	YP_001583063.1	M134	COG0371	C	Energy production and conversion	METABOLISM	10	119
131	V-type ATP synthase subunit B	161529198	YP_001583024.1	M136	COG1156	C	Energy production and conversion	METABOLISM	10	128
132	tRNA-modifying enzyme	161529037	YP_001582863.1	M135	COG0731	C	Energy production and conversion	METABOLISM	Totale	118
133	V-type ATP synthase alpha chain	68445582	BAE03290.1	MA02	COG1155	C	Energy production and conversion	METABOLISM	10	129
134	V-type H+-transporting ATPase subunit D	315425639	BAJ47298.1	MA16	COG1394	C	Energy production and conversion	METABOLISM	10	128
135	isopentenyl-diphosphate delta-isomerase	315425794	BAJ47448.1	MA20	COG1304	C	Energy production and conversion	METABOLISM	Totale	104
136	pyruvate ferredoxin oxidoreductase beta subunit	315426090	BAJ47736.1	MA27	COG1013	C	Energy production and conversion	METABOLISM	Totale	83
137	isopropylmalate/isohomocitrate dehydrogenase	161528577	YP_001582403.1	M137	COG0473	CE	Energy production and conversion / Amino acid transport and metabolism	METABOLISM	Totale	99
138	conserved hypothetical protein	315425902	BAJ47553.1	MA21	COG0575	I	Lipid transport and metabolism	METABOLISM	Totale	118
139	conserved hypothetical protein	315426923	BAJ48542.1	MA47	COG3425	I	Lipid transport and metabolism	METABOLISM	Totale	96
140	aspartate carbamoyltransferase	161529194	YP_001583020.1	M149	COG0540	F	Nucleotide transport and metabolism	METABOLISM	10	120
141	hypothetical protein Nmar_1535	161529043	YP_001582869.1	M150	COG1936	F	Nucleotide transport and metabolism	METABOLISM	10	124
142	uridylate kinase putative	161529213	YP_001583039.1	M153	COG0528	F	Nucleotide transport and metabolism	METABOLISM	10	127
143	RdgB/HAM1 family non-canonical purine NTP pyrophosphatase	161529038	YP_001582864.1	M158	COG0127	F	Nucleotide transport and metabolism	METABOLISM	10	127
144	phosphoribosylformylglycinamidine cyclo-ligase	161528090	YP_001581916.1	M151	COG0150	F	Nucleotide transport and metabolism	METABOLISM	Totale	110
145	adenylate kinase	161527911	YP_001581737.1	M154	COG2019	F	Nucleotide transport and metabolism	METABOLISM	Totale	82
146	phosphoribosylformylglycinamidine synthase II	161528288	YP_001582114.1	M156	COG0046	F	Nucleotide transport and metabolism	METABOLISM	Totale	109
147	adenylosuccinate lyase	161528963	YP_001582789.1	M157	COG0015	F	Nucleotide transport and metabolism	METABOLISM	Totale	92
148	aspartate carbamoyltransferase regulatory subunit	161529195	YP_001583021.1	M159	COG1781	F	Nucleotide transport and metabolism	METABOLISM	Totale	117
149	adenylosuccinate synthetase	161528928	YP_001582754.1	M160	COG0104	F	Nucleotide transport and metabolism	METABOLISM	Totale	115
150	glutamine amidotransferase class-II	161528289	YP_001582115.1	M161	COG0034	F	Nucleotide transport and metabolism	METABOLISM	Totale	110
151	ribose-phosphate pyrophosphokinase	161529282	YP_001583108.1	M162	COG0462	FE	Nucleotide transport and metabolism / Amino acid transport and metabolism	METABOLISM	10	123
152	glutamyl-tRNA(Gln) amidotransferase subunit D	161527612	YP_001581438.1	M132	COG0252	EJ	Amino acid transport and metabolism / Translation, ribosomal structure and biogenesis	OTHER	10	128
153	cytidyltransferase-like protein	161528806	YP_001582632.1	M029	COG0615	MI	Cell wall/membrane/envelope biogenesis / Lipid transport and metabolism	OTHER	10	125
154	pseudouridylate synthase subunit TruB	161527914	YP_001581740.1	M212	COG1598	S	Function unknown	POORLY CHARACTERIZED	10	124

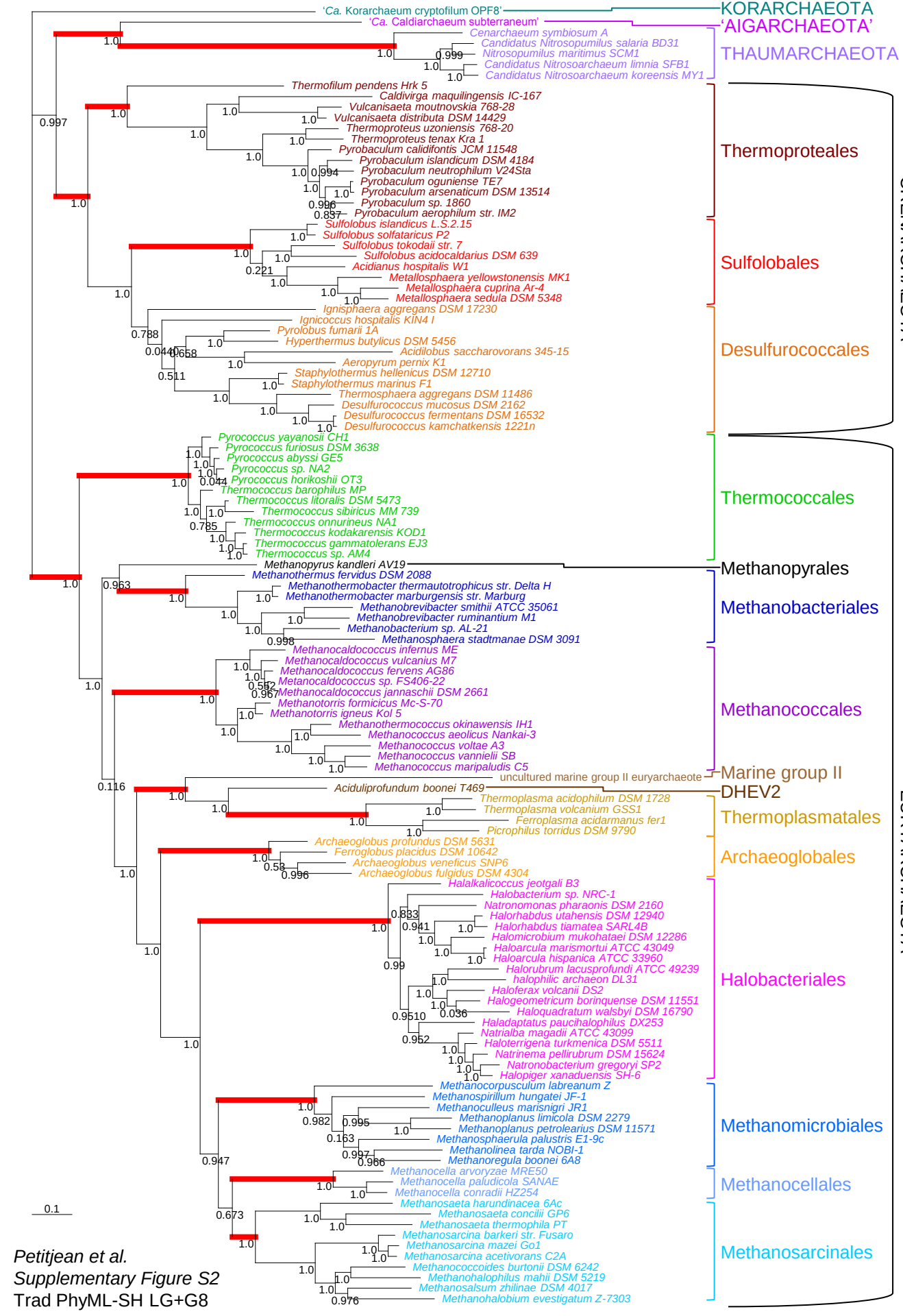
155	hypothetical protein Nmar_1233	161528741	YP_001582567.1	M217	COG1909	S	Function unknown	POORLY CHARACTERIZED	10	122
156	tRNA pseudouridine synthase D TruD	161528706	YP_001582532.1	M218	COG0585	S	Function unknown	POORLY CHARACTERIZED	10	125
157	peptidyl-tRNA hydrolase	161528708	YP_001582534.1	M219	COG1990	S	Function unknown	POORLY CHARACTERIZED	10	127
158	conserved hypothetical protein	315425963	BAJ47612.1	MA25	COG1371	S	Function unknown	POORLY CHARACTERIZED	10	124
159	DNA polymerase II small subunit	161527514	YP_001581340.1	M199	COG4304	S	Function unknown	POORLY CHARACTERIZED	Totale	94
160	GTP cyclohydrolase IIA	161529122	YP_001582948.1	M200	COG2429	S	Function unknown	POORLY CHARACTERIZED	Totale	67
161	hypothetical protein Nmar_0307	161527815	YP_001581641.1	M201	COG1698	S	Function unknown	POORLY CHARACTERIZED	Totale	98
162	hypothetical protein Nmar_0387	161527895	YP_001581721.1	M202	COG1650	S	Function unknown	POORLY CHARACTERIZED	Totale	112
163	hypothetical protein Nmar_0599	161528107	YP_001581933.1	M203	COG4046	S	Function unknown	POORLY CHARACTERIZED	Totale	45
164	hypothetical protein Nmar_0810	161528318	YP_001582144.1	M204	COG2106	S	Function unknown	POORLY CHARACTERIZED	Totale	87
165	hypothetical protein Nmar_0812	161528320	YP_001582146.1	M205	COG2090	S	Function unknown	POORLY CHARACTERIZED	Totale	101
166	hypothetical protein Nmar_1415	161528923	YP_001582749.1	M206	COG2237	S	Function unknown	POORLY CHARACTERIZED	Totale	110
167	hypothetical protein Nmar_1719	161529227	YP_001583053.1	M207	COG1701	S	Function unknown	POORLY CHARACTERIZED	Totale	113
168	LPPG:FO 2-phospho-L-lactate transferase	161528134	YP_001581960.1	M208	COG0391	S	Function unknown	POORLY CHARACTERIZED	Totale	70
169	hypothetical protein Nmar_0053	161527565	YP_001581391.1	M209	COG1679	S	Function unknown	POORLY CHARACTERIZED	Totale	90
170	hypothetical protein Nmar_0054	161527566	YP_001581392.1	M210	COG1786	S	Function unknown	POORLY CHARACTERIZED	Totale	90
171	hypothetical protein Nmar_0617	161528125	YP_001581951.1	M213	COG2042	S	Function unknown	POORLY CHARACTERIZED	Totale	91
172	hypothetical protein Nmar_1606	161529114	YP_001582940.1	M215	COG1303	S	Function unknown	POORLY CHARACTERIZED	Totale	111
173	hypothetical conserved protein	315425224	BAJ46893.1	MA09	COG1839	S	Function unknown	POORLY CHARACTERIZED	Totale	61
174	ATP phosphoribosyltransferase	315425361	BAJ47027.1	MA11	COG4077	S	Function unknown	POORLY CHARACTERIZED	Totale	81
175	conserved hypothetical protein	315426248	BAJ47890.1	MA31	COG1469	S	Function unknown	POORLY CHARACTERIZED	Totale	90
176	ATPase of the PP-loop superfamily	118575616	YP_875359.1	M181	COG2102	R	General function prediction only	POORLY CHARACTERIZED	10	127
177	DNA-binding TFAR19-related protein	161529131	YP_001582957.1	M184	COG2118	R	General function prediction only	POORLY CHARACTERIZED	10	129
178	eRF1 domain-containing 1 protein	161527621	YP_001581447.1	M185	COG1537	R	General function prediction only	POORLY CHARACTERIZED	10	128
179	hypothetical protein Nmar_1562	161529070	YP_001582896.1	M186	COG1571	R	General function prediction only	POORLY CHARACTERIZED	10	120
180	putative RNA-processing protein	161528917	YP_001582743.1	M188	COG1094	R	General function prediction only	POORLY CHARACTERIZED	10	126
181	beta-lactamase domain-containing protein	161529235	YP_001583061.1	M190	COG1782	R	General function prediction only	POORLY CHARACTERIZED	10	129
182	phosphopyruvate hydratase	161527825	YP_001581651.1	M191	COG2102	R	General function prediction only	POORLY CHARACTERIZED	10	121
183	ATPase	161528562	YP_001582388.1	M193	COG1855	R	General function prediction only	POORLY CHARACTERIZED	10	126
184	hypothetical protein Nmar_1571	161529079	YP_001582905.1	M194	COG1634	R	General function prediction only	POORLY CHARACTERIZED	10	119
185	putative ATPase RIL	161527803	YP_001581629.1	M198	COG1245	R	General function prediction only	POORLY CHARACTERIZED	10	123
186	ribonuclease Z	315426519	BAJ48150.1	MA36	COG1234	R	General function prediction only	POORLY CHARACTERIZED	10	120
187	GTP-binding protein	315426744	BAJ48368.1	MA41	COG1163	R	General function prediction only	POORLY CHARACTERIZED	10	127
188	5-formaminoimidazole-4-carboxamide-1-(beta)-D-ribofuranosyl 5p-monophosphate synthe	161528278	YP_001582104.1	M182	COG1759	R	General function prediction only	POORLY CHARACTERIZED	Totale	59
189	DEAD/DEAH box helicase domain-containing protein	161527751	YP_001581577.1	M183	COG1204	R	General function prediction only	POORLY CHARACTERIZED	Totale	112
190	nucleotide binding protein PINc	161527947	YP_001581773.1	M187	COG1439	R	General function prediction only	POORLY CHARACTERIZED	Totale	118
191	5-formaminoimidazole-4-carboxamide-1-(beta)-D-ribofuranosyl 5p-monophosphate synthe	161528380	YP_001582206.1	M189	COG1759	R	General function prediction only	POORLY CHARACTERIZED	Totale	82
192	SPP-like hydrolase	161528193	YP_001582019.1	M192	COG0561	R	General function prediction only	POORLY CHARACTERIZED	Totale	106
193	GTP1/OBG protein	161529113	YP_001582939.1	M195	COG2262	R	General function prediction only	POORLY CHARACTERIZED	Totale	93
194	GHMP kinase	161529226	YP_001583052.1	M196	COG1829	R	General function prediction only	POORLY CHARACTERIZED	Totale	115
195	geranylgeranylglyceryl phosphate synthase-like protein	161529303	YP_001583129.1	M197	COG1646	R	General function prediction only	POORLY CHARACTERIZED	Totale	87
196	acetylacyl transferase related protein	315425161	BAJ46831.1	MA06	COG0110	R	General function prediction only	POORLY CHARACTERIZED	Totale	49
197	phosphopantetheine adenylyltransferase	315426113	BAJ47758.1	MA28	COG1019	R	General function prediction only	POORLY CHARACTERIZED	Totale	99
198	nucleic-acid-binding protein	315426921	BAJ48540.1	MA46	COG1545	R	General function prediction only	POORLY CHARACTERIZED	Totale	96
199	conserved hypothetical protein	315427069	BAJ48685.1	MA50	COG1938	R	General function prediction only	POORLY CHARACTERIZED	Totale	107
200	conserved hypothetical protein	315427074	BAJ48690.1	MA52	COG2129	R	General function prediction only	POORLY CHARACTERIZED	Totale	82

	TaxID	Species		Phylum	Class/Order
		* : species retained for the construction of the final datasets (108 species)			
1	311458	*Candidatus* Caldiarchaeum subterraneum	*	Aigarchaeota	Unclassified
2	414004	*Cenarchaeum symbiosum A*	*	Thaumarchaeota	Cenarchaeales
3	436308	*Nitrosopumilus maritimus SCM1*	*	Thaumarchaeota	Nitrosopumilales
4	886738	*Candidatus* Nitrosoarchaeum limnia SFB1	*	Thaumarchaeota	Nitrosopumilales
5	1001994	*Candidatus* Nitrosoarchaeum koreensis MY1	*	Thaumarchaeota	Nitrosopumilales
6	859350	*Candidatus* Nitrosopumilus salaria BD31	*	Thaumarchaeota	Nitrosopumilales
7	797209	*Haladaptatus paucihalophilus DX253*	*	Euryarchaeota	Halobacteriales
8	795797	*Halalkalicoccus jeotgali B3*		Euryarchaeota	Halobacteriales
9	634497	*Haloarcula hispanica ATCC 33960*		Euryarchaeota	Halobacteriales
10	272569	*Haloarcula marismortui ATCC 43049*		Euryarchaeota	Halobacteriales
11	478009	*Halobacterium salinarum R1*		Euryarchaeota	Halobacteriales
12	309800	*Haloferax volcanii DS2*		Euryarchaeota	Halobacteriales
13	469382	*Halogeometricum borinquense DSM 11551*	*	Euryarchaeota	Halobacteriales
14	485914	*Halomicrobium mukohataei DSM 12286*		Euryarchaeota	Halobacteriales
15	756883	*halophilic archaeon DL31*		Euryarchaeota	Halobacteriales
16	797210	*Halopiger xanaduensis SH-6*		Euryarchaeota	Halobacteriales
17	362976	*Haloquadratum walsbyi DSM 16790*		Euryarchaeota	Halobacteriales
18	1033806	*Halorhabdus tiamatea SARL4B*		Euryarchaeota	Halobacteriales
19	519442	*Halorhabdus utahensis DSM 12940*	*	Euryarchaeota	Halobacteriales
20	416348	*Halorubrum lacusprofundi ATCC 49239*	*	Euryarchaeota	Halobacteriales
21	543526	*Haloterrigena turkmenica DSM 5511*		Euryarchaeota	Halobacteriales
22	547559	*Natrialba magadii ATCC 43099*	*	Euryarchaeota	Halobacteriales
23	797303	*Natrinema pellirubrum DSM 15624*		Euryarchaeota	Halobacteriales
24	797304	*Natronobacterium gregoryi SP2*		Euryarchaeota	Halobacteriales
25	348780	*Natronomonas pharaonis DSM 2160*	*	Euryarchaeota	Halobacteriales
26	1072681	*Candidatus* Haloredivivus sp. G17	*	Euryarchaeota	Nanohaloarchaea
27	889948	*Candidatus* Nanosalina sp. J07AB43	*	Euryarchaeota	Nanohaloarchaea
28	889962	*Candidatus* Nanosalinarum sp. J07AB56	*	Euryarchaeota	Nanohaloarchaea
29	410358	*Methanocorpusculum labreanum Z*	*	Euryarchaeota	Methanomicrobiales
30	368407	*Methanoculleus marisnigri JR1*	*	Euryarchaeota	Methanomicrobiales
31	882090	*Methanolinea tarda NOBI-1*		Euryarchaeota	Methanomicrobiales
32	937775	*Methanoplanus limicola DSM 2279*	*	Euryarchaeota	Methanomicrobiales
33	679926	*Methanoplanus petrolearius DSM 11571*		Euryarchaeota	Methanomicrobiales
34	456442	*Methanoregula boonei 6A8*		Euryarchaeota	Methanomicrobiales
35	521011	*Methanosphaerula palustris E1-9c*	*	Euryarchaeota	Methanomicrobiales
36	323259	*Methanospirillum hungatei JF-1*	*	Euryarchaeota	Methanomicrobiales
37	259564	*Methanococcoides burtonii DSM 6242*	*	Euryarchaeota	Methanosarcinales
38	644295	*Methanohalobium evestigatum Z-7303*		Euryarchaeota	Methanosarcinales
39	547558	*Methanohalophilus mahii DSM 5219*		Euryarchaeota	Methanosarcinales
40	990316	*Methanosaeta concilii GP6*		Euryarchaeota	Methanosarcinales
41	1110509	*Methanosaeta harundinacea 6Ac*	*	Euryarchaeota	Methanosarcinales
42	349307	*Methanosaeta thermophila PT*	*	Euryarchaeota	Methanosarcinales
43	679901	*Methanosalsum zhilinae DSM 4017*	*	Euryarchaeota	Methanosarcinales
44	188937	*Methanosarcina acetivorans C2A*		Euryarchaeota	Methanosarcinales
45	269797	*Methanosarcina barkeri str. Fusaro*		Euryarchaeota	Methanosarcinales
46	192952	*Methanosarcina mazei Go1*	*	Euryarchaeota	Methanosarcinales
47	351160	*Methanocella arvoryzae MRE50*	*	Euryarchaeota	Methanocellales
48	1041930	*Methanocella conradii HZ254*	*	Euryarchaeota	Methanocellales
49	304371	*Methanocella paludicola SANAE*	*	Euryarchaeota	Methanocellales
50	224325	*Archaeoglobus fulgidus DSM 4304*	*	Euryarchaeota	Archaeoglobales
51	572546	*Archaeoglobus profundus DSM 5631*	*	Euryarchaeota	Archaeoglobales
52	693661	*Archaeoglobus veneficus SNP6*	*	Euryarchaeota	Archaeoglobales
53	589924	*Ferroglobus placidus DSM 10642*	*	Euryarchaeota	Archaeoglobales
54	439481	*Aciduliprofundum boonei T469*	*	Euryarchaeota	Thermoplasmatales
55	333146	*Ferroplasma acidarmanus fer1*	*	Euryarchaeota	Thermoplasmatales
56	263820	*Picrophilus torridus DSM 9790*	*	Euryarchaeota	Thermoplasmatales
57	273075	*Thermoplasma acidophilum DSM 1728*	*	Euryarchaeota	Thermoplasmatales
58	273116	*Thermoplasma volcanium GSS1*	*	Euryarchaeota	Thermoplasmatales
59	573064	*Methanocaldococcus fervens AG86*		Euryarchaeota	Methanococcales
60	573063	*Methanocaldococcus infernus ME*	*	Euryarchaeota	Methanococcales
61	243232	*Methanocaldococcus jannaschii DSM 2661*	*	Euryarchaeota	Methanococcales
62	644281	*Methanocaldococcus sp. FS406-22*		Euryarchaeota	Methanococcales
63	579137	*Methanocaldococcus vulcanius M7*		Euryarchaeota	Methanococcales
64	419665	*Methanococcus aeolicus Nankai-3*	*	Euryarchaeota	Methanococcales
65	402880	*Methanococcus maripaludis C5*		Euryarchaeota	Methanococcales
66	406327	*Methanococcus vannielii SB*	*	Euryarchaeota	Methanococcales

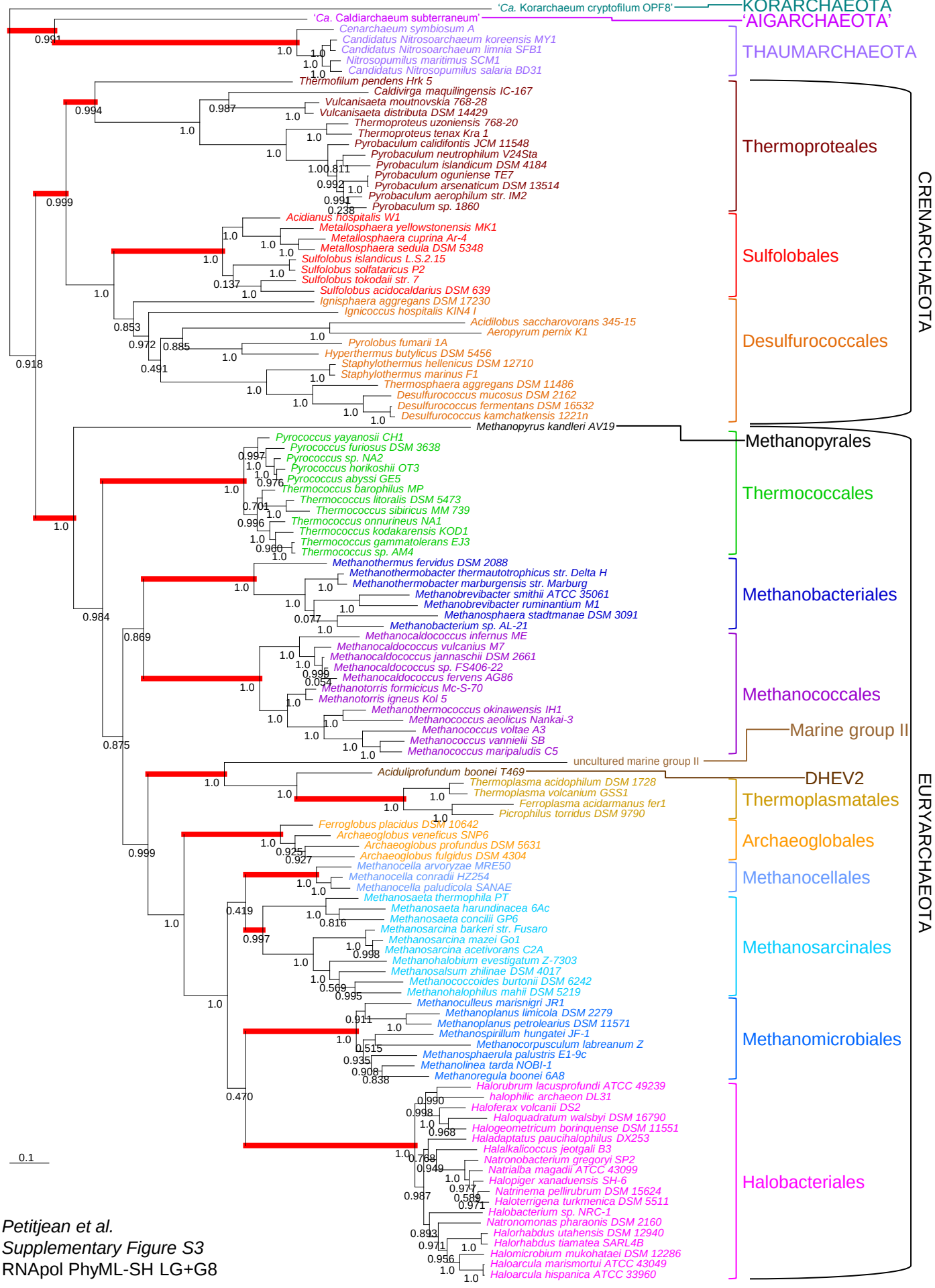
67	456320	*Methanococcus voltae A3*		Euryarchaeota	Methanococcales
68	647113	*Methanothermococcus okinawensis IH1*		Euryarchaeota	Methanococcales
69	647171	*Methanotorris formicicus Mc-S-70*		Euryarchaeota	Methanococcales
70	880724	*Methanotorris igneus Kol 5*	*	Euryarchaeota	Methanococcales
71	868132	*Methanobacterium sp. AL-21*	*	Euryarchaeota	Methanobacteriales
72	634498	*Methanobrevibacter ruminantium M1*		Euryarchaeota	Methanobacteriales
73	420247	*Methanobrevibacter smithii ATCC 35061*	*	Euryarchaeota	Methanobacteriales
74	339860	*Methanosphaera stadtmanae DSM 3091*	*	Euryarchaeota	Methanobacteriales
75	79929	*Methanothermobacter marburgensis str. Marburg*		Euryarchaeota	Methanobacteriales
76	187420	*Methanothermobacter thermautotrophicus str. Delta H*	*	Euryarchaeota	Methanobacteriales
77	523846	*Methanothermus fervidus DSM 2088*	*	Euryarchaeota	Methanobacteriales
78	190192	*Methanopyrus kandleri AV19*	*	Euryarchaeota	Methanopyrales
79	272844	*Pyrococcus abyssi GE5*	*	Euryarchaeota	Thermococcales
80	186497	*Pyrococcus furiosus DSM 3638*		Euryarchaeota	Thermococcales
81	70601	*Pyrococcus horikoshii OT3*		Euryarchaeota	Thermococcales
82	342949	*Pyrococcus sp. NA2*		Euryarchaeota	Thermococcales
83	529709	*Pyrococcus yayanosii CH1*	*	Euryarchaeota	Thermococcales
84	391623	*Thermococcus barophilus MP*	*	Euryarchaeota	Thermococcales
85	593117	*Thermococcus gammatolerans EJ3*	*	Euryarchaeota	Thermococcales
86	69014	*Thermococcus kodakarensis KOD1*		Euryarchaeota	Thermococcales
87	523849	*Thermococcus litoralis DSM 5473*	*	Euryarchaeota	Thermococcales
88	523850	*Thermococcus onnurineus NA1*		Euryarchaeota	Thermococcales
89	604354	*Thermococcus sibiricus MM 739*		Euryarchaeota	Thermococcales
90	246969	*Thermococcus sp. AM4*		Euryarchaeota	Thermococcales
91	425595	*Candidatus* Micrarchaeum acidiphilum ARMAN-2	*	Euryarchaeota	ARMAN
92	662760	*Candidatus* Parvarchaeum acidiphilum ARMAN-4	*	Euryarchaeota	ARMAN
93	662762	*Candidatus* Parvarchaeum acidophilus ARMAN-5	*	Euryarchaeota	ARMAN
94	228908	*Nanoarchaeum equitans Kin4-M*	*	Euryarchaeota	Nanoarchaeota
95	274854	uncultured marine group II euryarchaeote	*	Euryarchaeota	Unclassified
96	666510	*Acidilobus saccharovorans 345-15*	*	Crenarchaeota	Desulfurococcales
97	272557	*Aeropyrum pernix K1*	*	Crenarchaeota	Desulfurococcales
98	768672	*Desulfurococcus fermentans DSM 16532*		Crenarchaeota	Desulfurococcales
99	490899	*Desulfurococcus kamchatkensis 1221n*	*	Crenarchaeota	Desulfurococcales
100	765177	*Desulfurococcus mucosus DSM 2162*		Crenarchaeota	Desulfurococcales
101	415426	*Hyperthermus butylicus DSM 5456*	*	Crenarchaeota	Desulfurococcales
102	453591	*Ignicoccus hospitalis KIN4/I*	*	Crenarchaeota	Desulfurococcales
103	583356	*Ignisphaera aggregans DSM 17230*	*	Crenarchaeota	Desulfurococcales
104	694429	*Pyrolobus fumarii 1A*	*	Crenarchaeota	Desulfurococcales
105	591019	*Staphylothermus hellenicus DSM 12710*	*	Crenarchaeota	Desulfurococcales
106	399550	*Staphylothermus marinus F1*		Crenarchaeota	Desulfurococcales
107	633148	*Thermosphaera aggregans DSM 11486*	*	Crenarchaeota	Desulfurococcales
108	933801	*Acidianus hospitalis W1*	*	Crenarchaeota	Sulfolobales
109	1006006	*Metallosphaera cuprina Ar-4*		Crenarchaeota	Sulfolobales
110	399549	*Metallosphaera sedula DSM 5348*	*	Crenarchaeota	Sulfolobales
111	671065	*Metallosphaera yellowstonensis MK1*	*	Crenarchaeota	Sulfolobales
112	330779	*Sulfolobus acidocaldarius DSM 639*	*	Crenarchaeota	Sulfolobales
113	429572	*Sulfolobus islandicus L.S.2.15*		Crenarchaeota	Sulfolobales
114	273057	*Sulfolobus solfataricus P2*	*	Crenarchaeota	Sulfolobales
115	273063	*Sulfolobus tokodaii str. 7*	*	Crenarchaeota	Sulfolobales
116	397948	*Caldivirga maquilingensis IC-167*	*	Crenarchaeota	Thermoproteales
117	178306	*Pyrobaculum aerophilum str. IM2*	*	Crenarchaeota	Thermoproteales
118	340102	*Pyrobaculum arsenaticum DSM 13514*	*	Crenarchaeota	Thermoproteales
119	410359	*Pyrobaculum calidifontis JCM 11548*		Crenarchaeota	Thermoproteales
120	384616	*Pyrobaculum islandicum DSM 4184*	*	Crenarchaeota	Thermoproteales
121	698757	*Pyrobaculum neutrophilum V24Sta*		Crenarchaeota	Thermoproteales
122	1104324	*Pyrobaculum oguniense TE7*		Crenarchaeota	Thermoproteales
123	368408	*Pyrobaculum sp. 1860*		Crenarchaeota	Thermoproteales
124	444157	*Thermofilum pendens Hrk 5*	*	Crenarchaeota	Thermoproteales
125	768679	*Thermoproteus tenax Kra 1*	*	Crenarchaeota	Thermoproteales
126	999630	*Thermoproteus uzoniensis 768-20*		Crenarchaeota	Thermoproteales
127	572478	*Vulcanisaeta distributa DSM 14429*	*	Crenarchaeota	Thermoproteales
128	985053	*Vulcanisaeta moutnovskia 768-28*		Crenarchaeota	Thermoproteales
129	374847	*Candidatus* Korarchaeum cryptofilum OPF8	*	Korachaeota	



Petitjean et al.
Supplementary Figure S1
200 PhyML-SH LG+G8



Petitjean et al.
Supplementary Figure S2
Trad PhyML-SH LG+G8

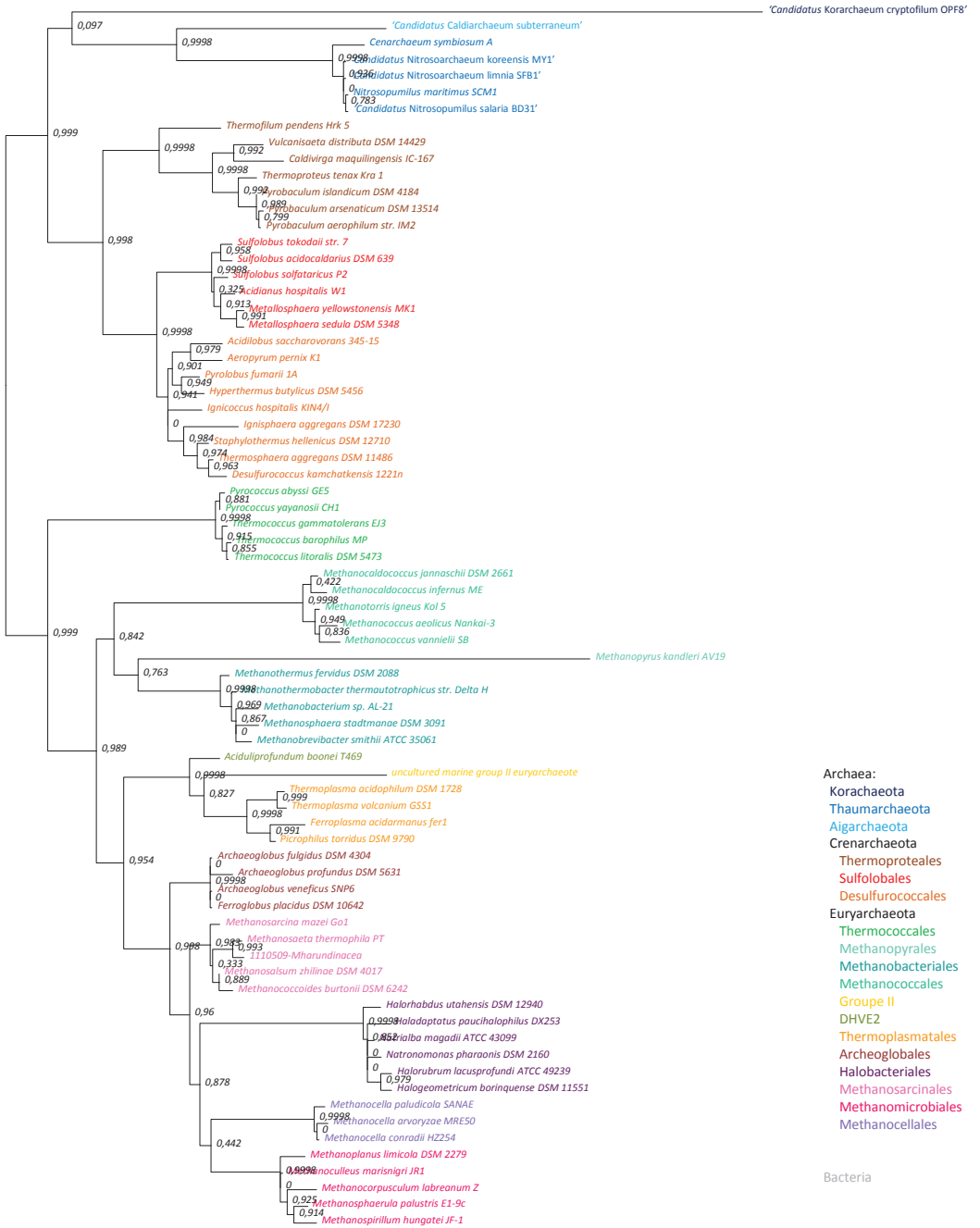


Petitjean et al.
Supplementary Figure S3
RNApol PhyML-SH LG+G8

Petitjean et al.
Supplementary Figure S4
RNApol Phylobayes CAT+G8

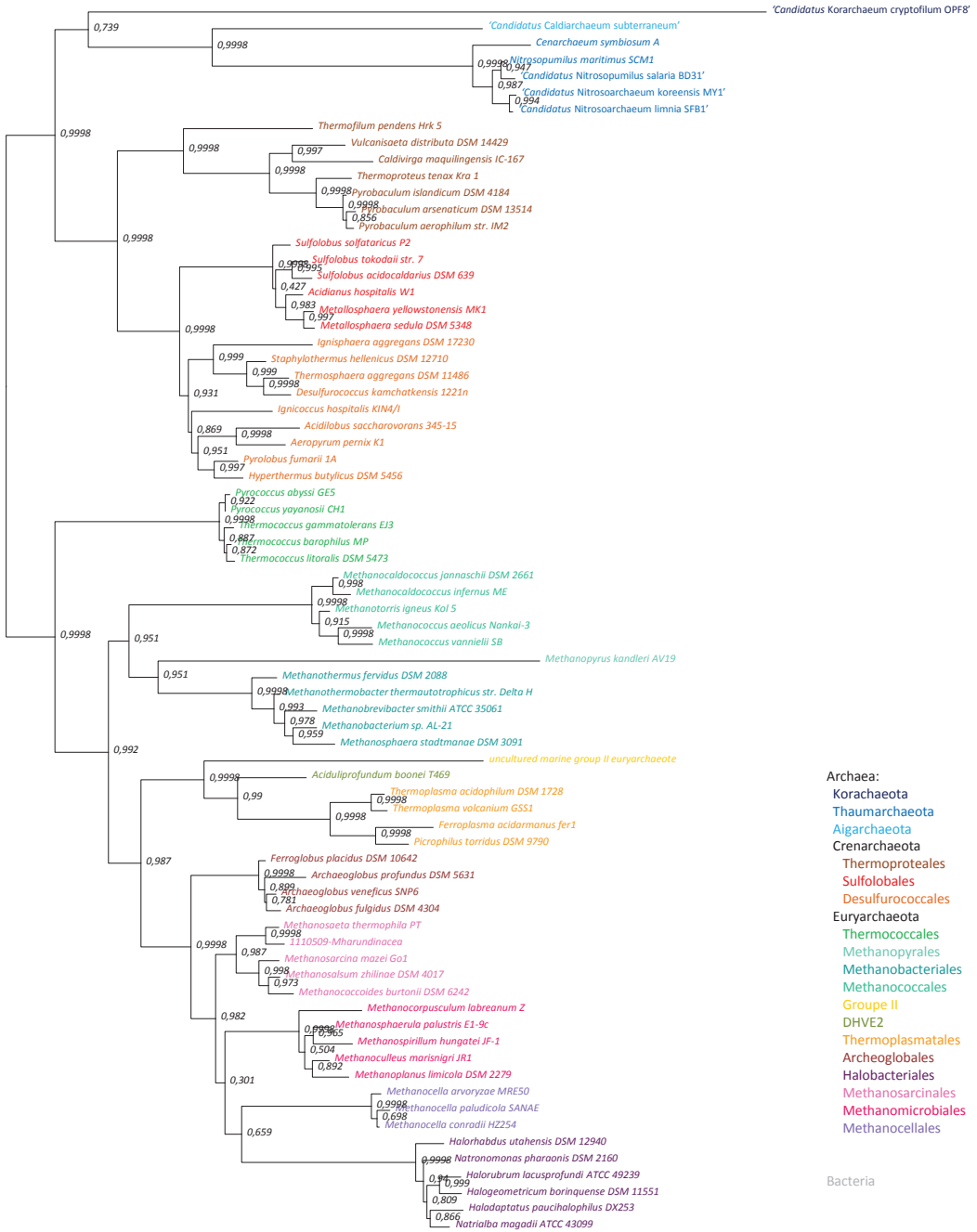
Petitjean et al.
Supplementary Figure S5
SF_0 PhyML-SH LG+G8

74 species - 1958 positions



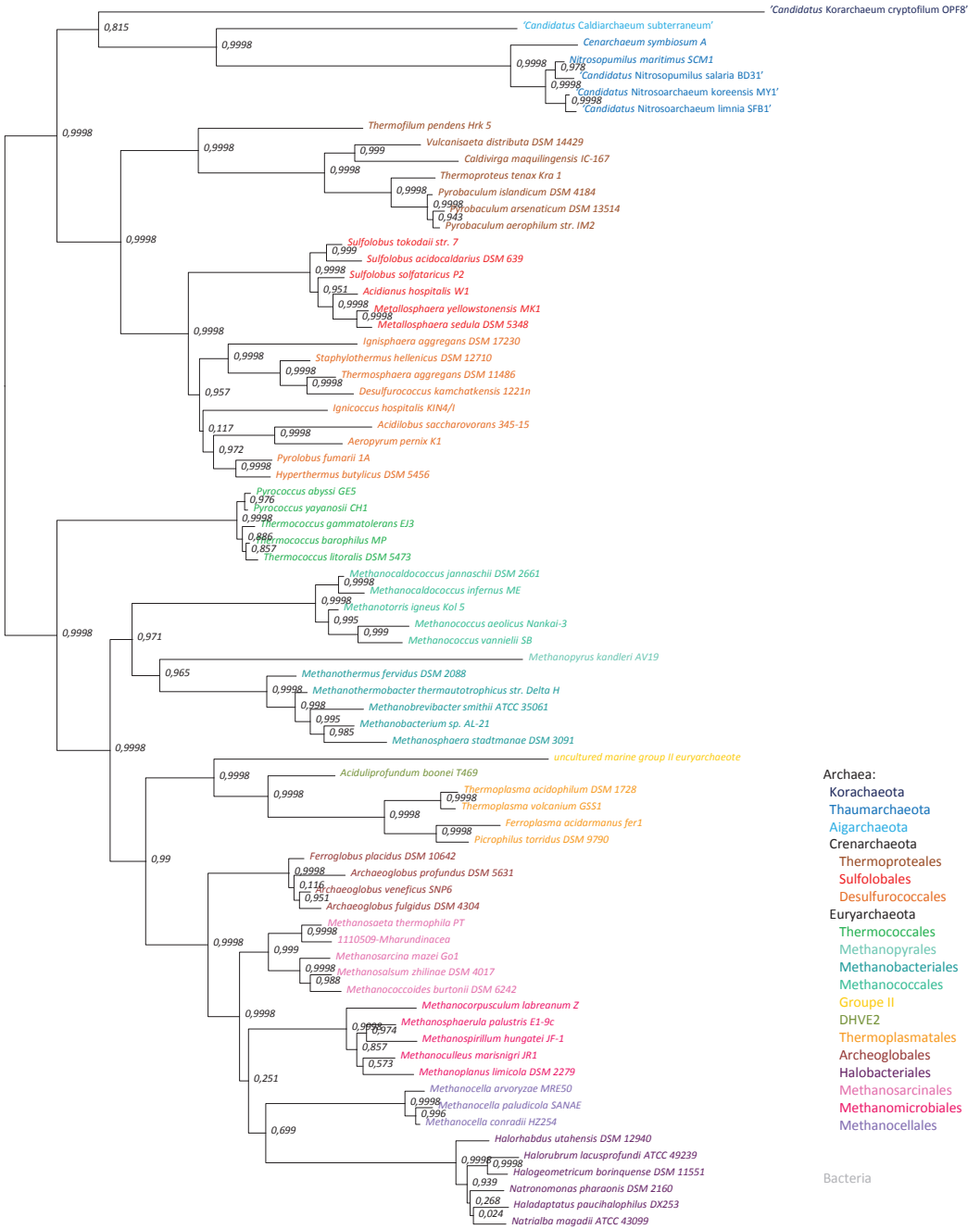
Petitjean et al.
Supplementary Figure S5
SF_1 PhyML-SH LG+G8

74 species - 3884 positions



Petitjean et al.
Supplementary Figure S5
SF_2 PhyML-SH LG+G8

74 species - 5439 positions



Petitjean et al.
Supplementary Figure S5
SF_3 PhyML-SH LG+G8

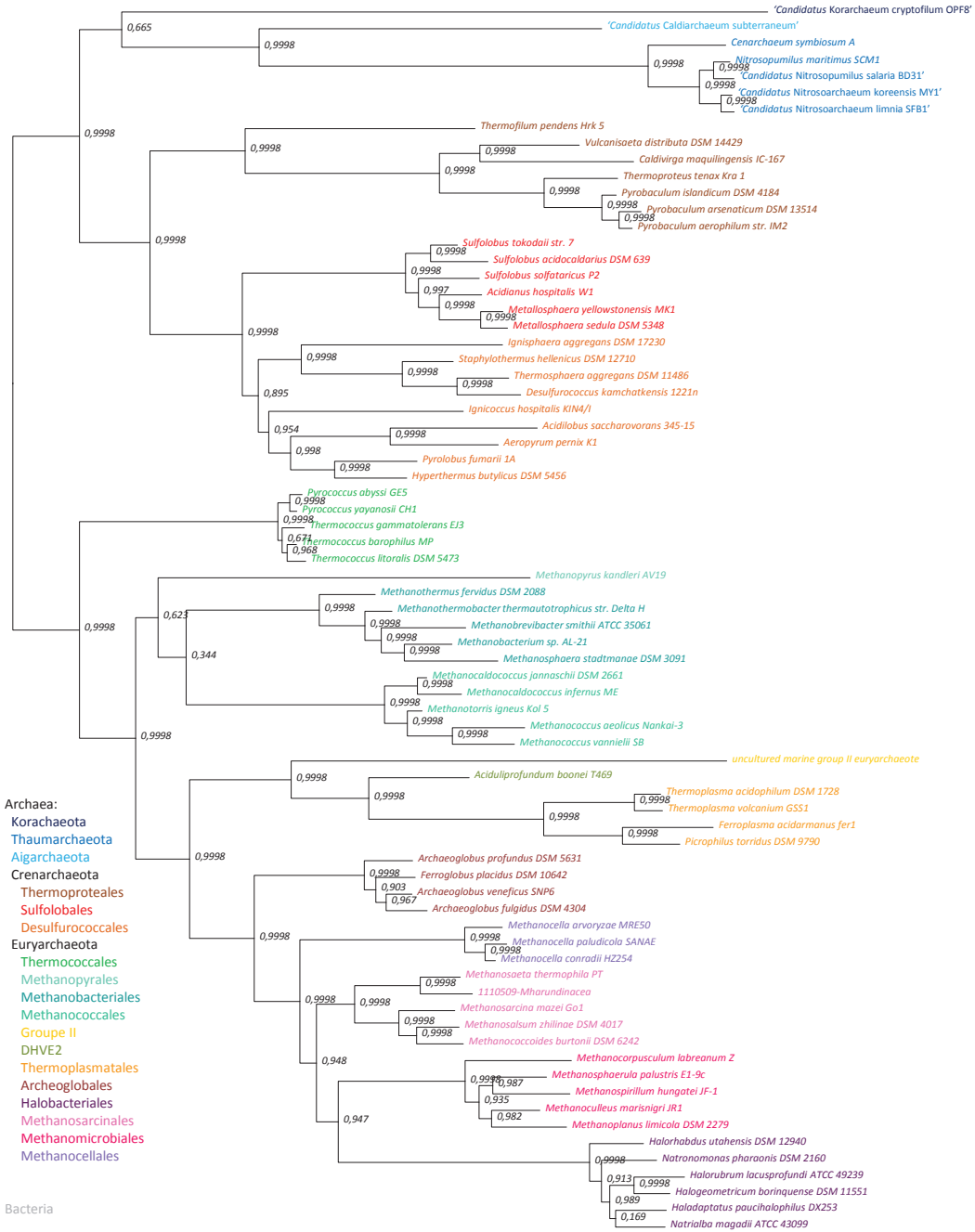
74 species - 6929 positions

Petitjean et al.
Supplementary Figure S5
SF_4 PhyML-SH LG+G8

74 species - 8373 positions

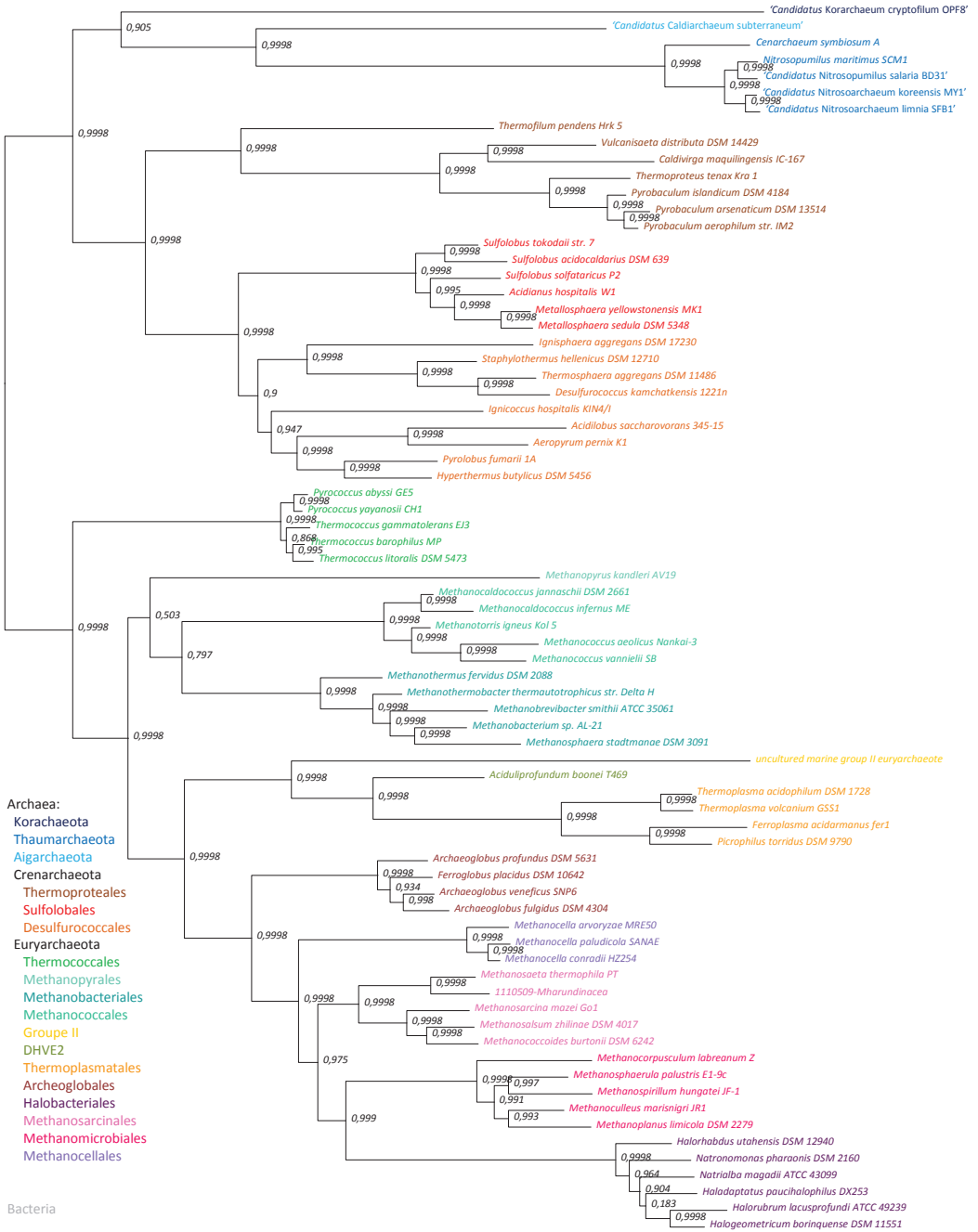
Petitjean et al.
Supplementary Figure S5
SF_5 PhyML-SH LG+G8

74 species - 9742 positions



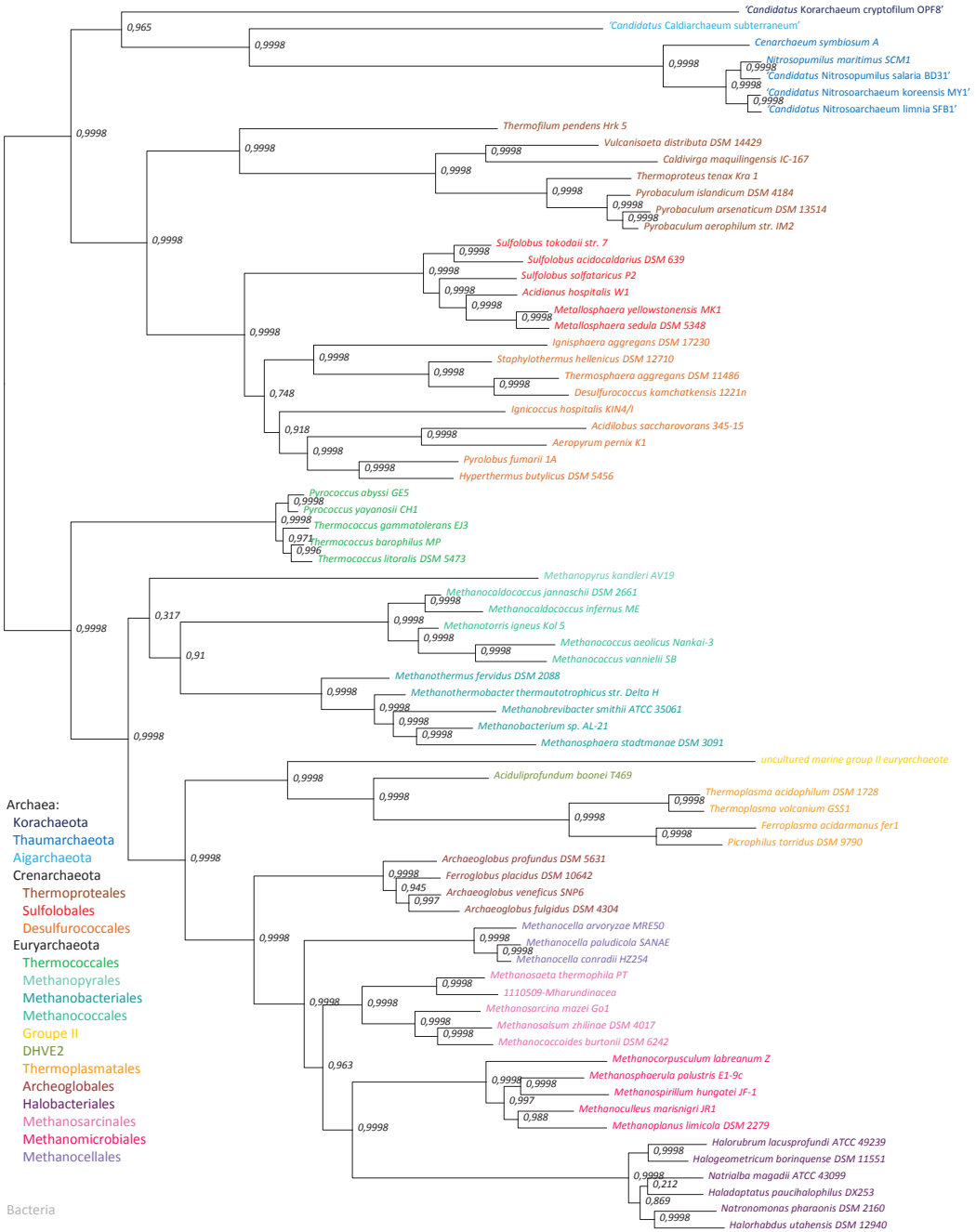
Petitjean et al.
Supplementary Figure S5
SF_6 PhyML-SH LG+G8

74 species – 11 089 positions



Petitjean et al.
Supplementary Figure S5
SF_7 PhyML-SH LG+G8

74 species – 12 465 positions



Petitjean et al.
Supplementary Figure S5
SF_8 PhyML-SH LG+G8

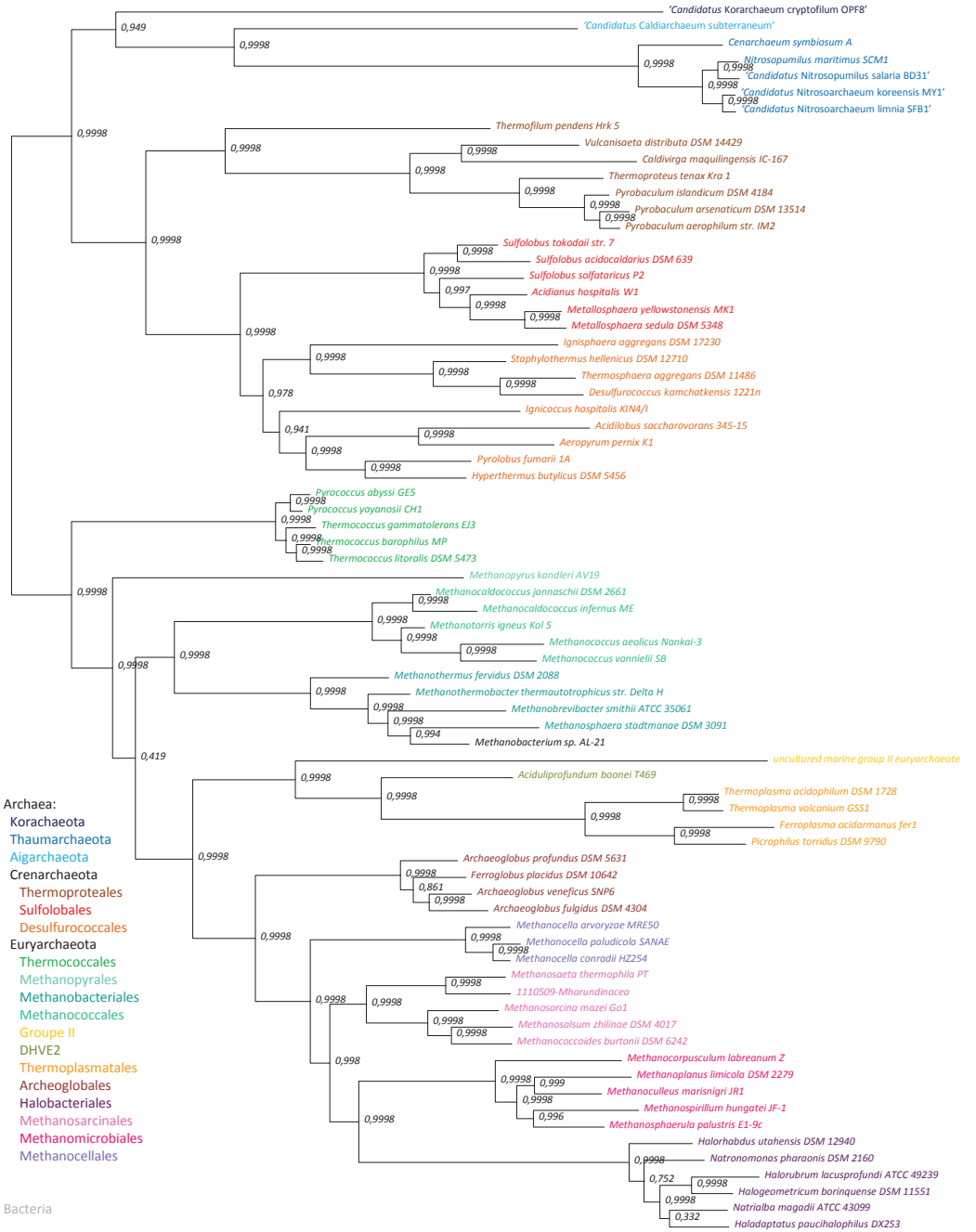
74 species – 13 806 positions

Petitjean et al.
Supplementary Figure S5
SF_9 PhyML-SH LG+G8

74 species – 15 176 positions

Petitjean et al.
Supplementary Figure S5
SF_{10} PhyML-SH LG+G8

74 species – 16 508 positions

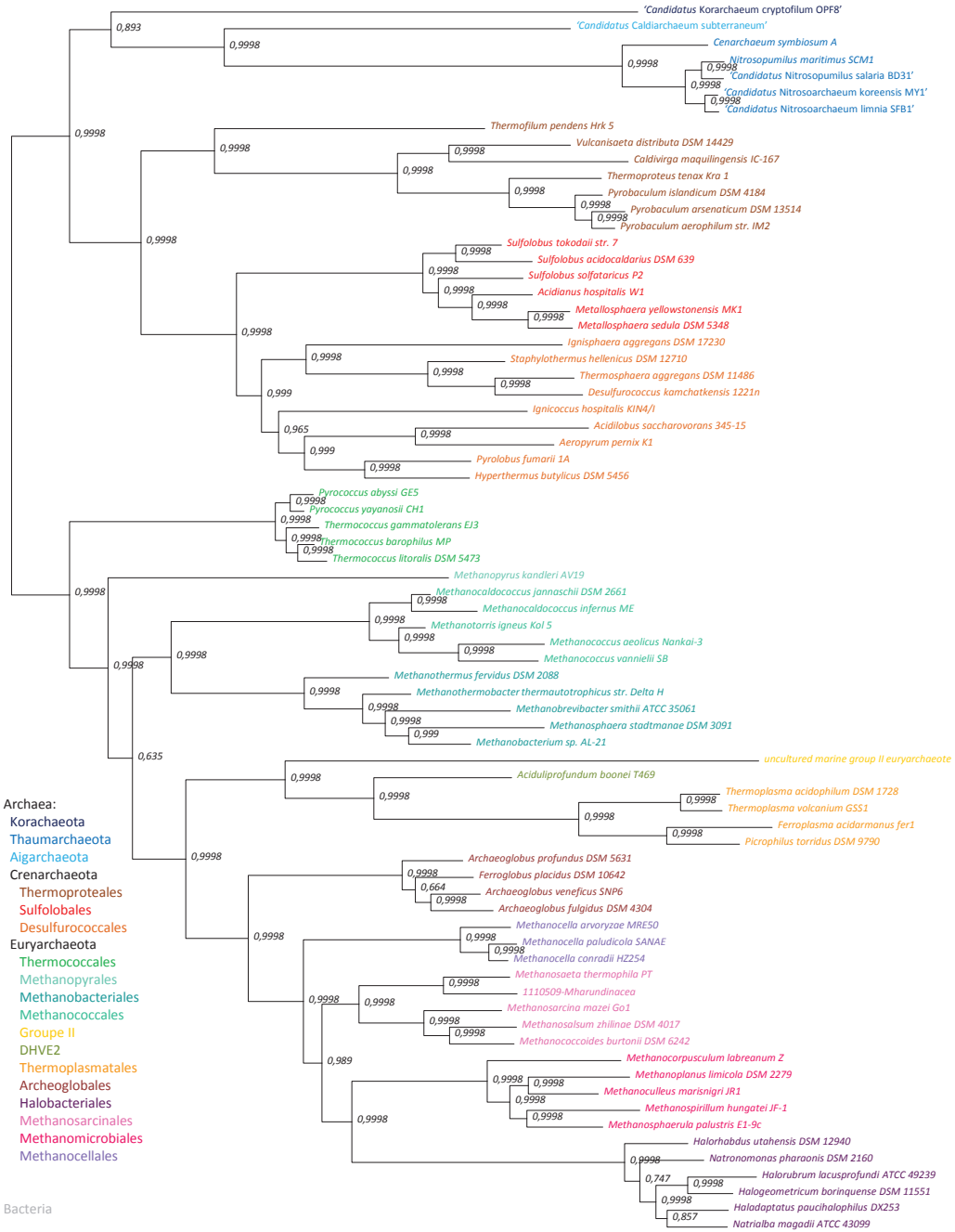


Petitjean et al.
Supplementary Figure S5
SF_{11} PhyML-SH LG+G8

74 species – 17 845 positions

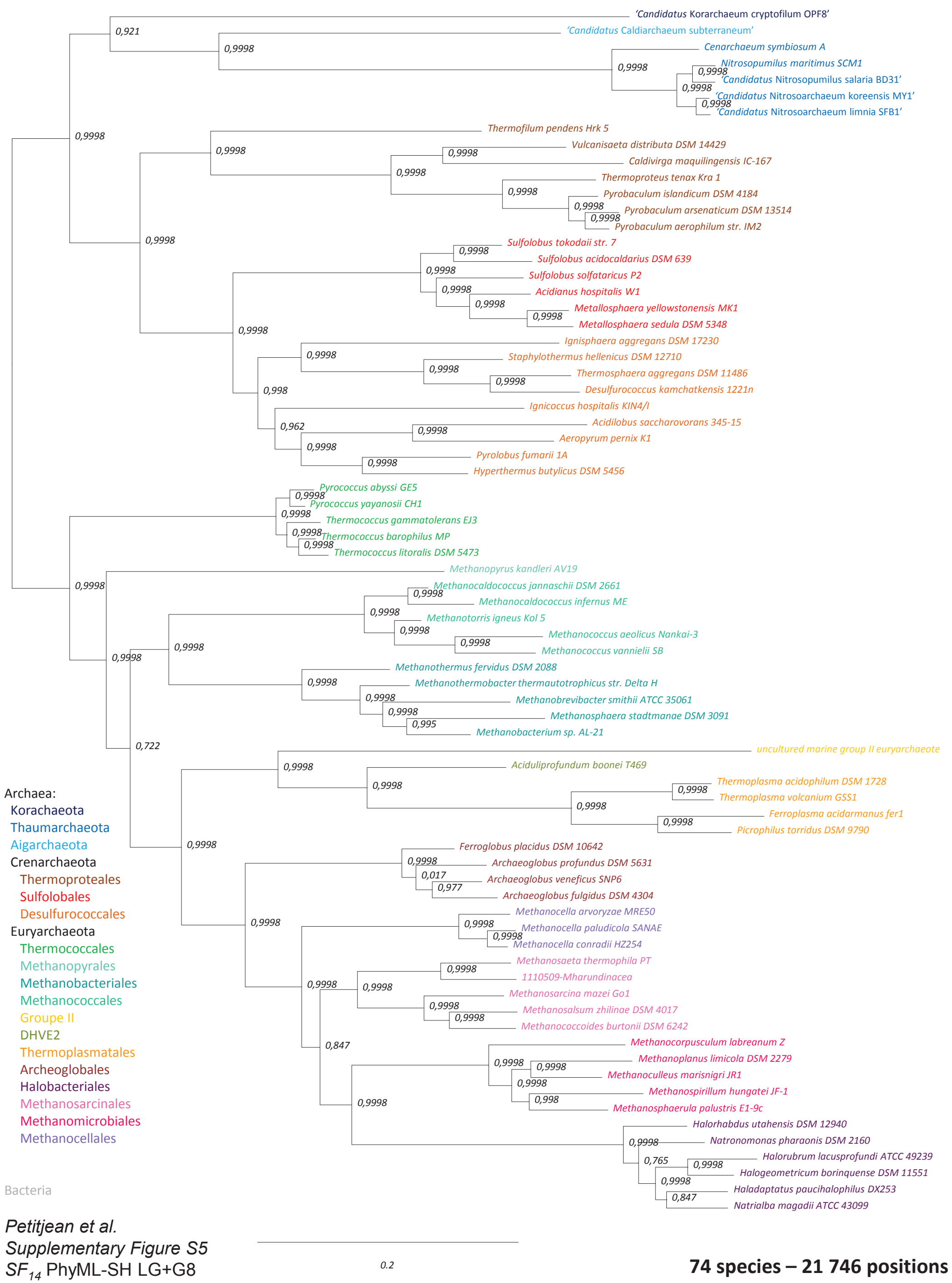
Petitjean et al.
Supplementary Figure S5
SF_{12} PhyML-SH LG+G8

74 species – 19 167 positions



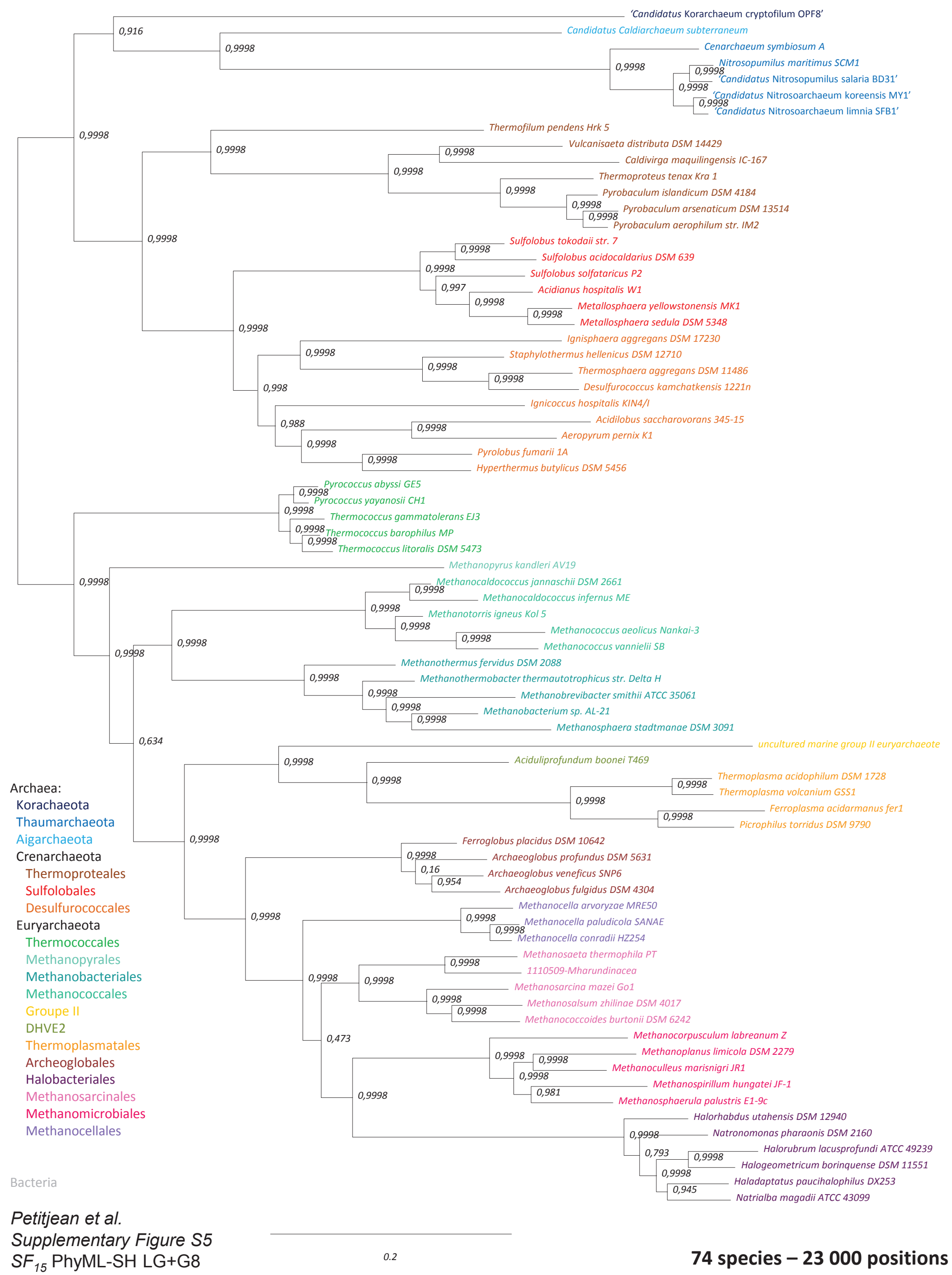
Petitjean et al.
Supplementary Figure S5
SF_{12} PhyML-SH LG+G8

74 species – 20 458 positions



Petitjean et al.
Supplementary Figure S5
SF_{14} PhyML-SH LG+G8

74 species – 21 746 positions



Petitjean et al.
Supplementary Figure S5
SF_{15} PhyML-SH LG+G8

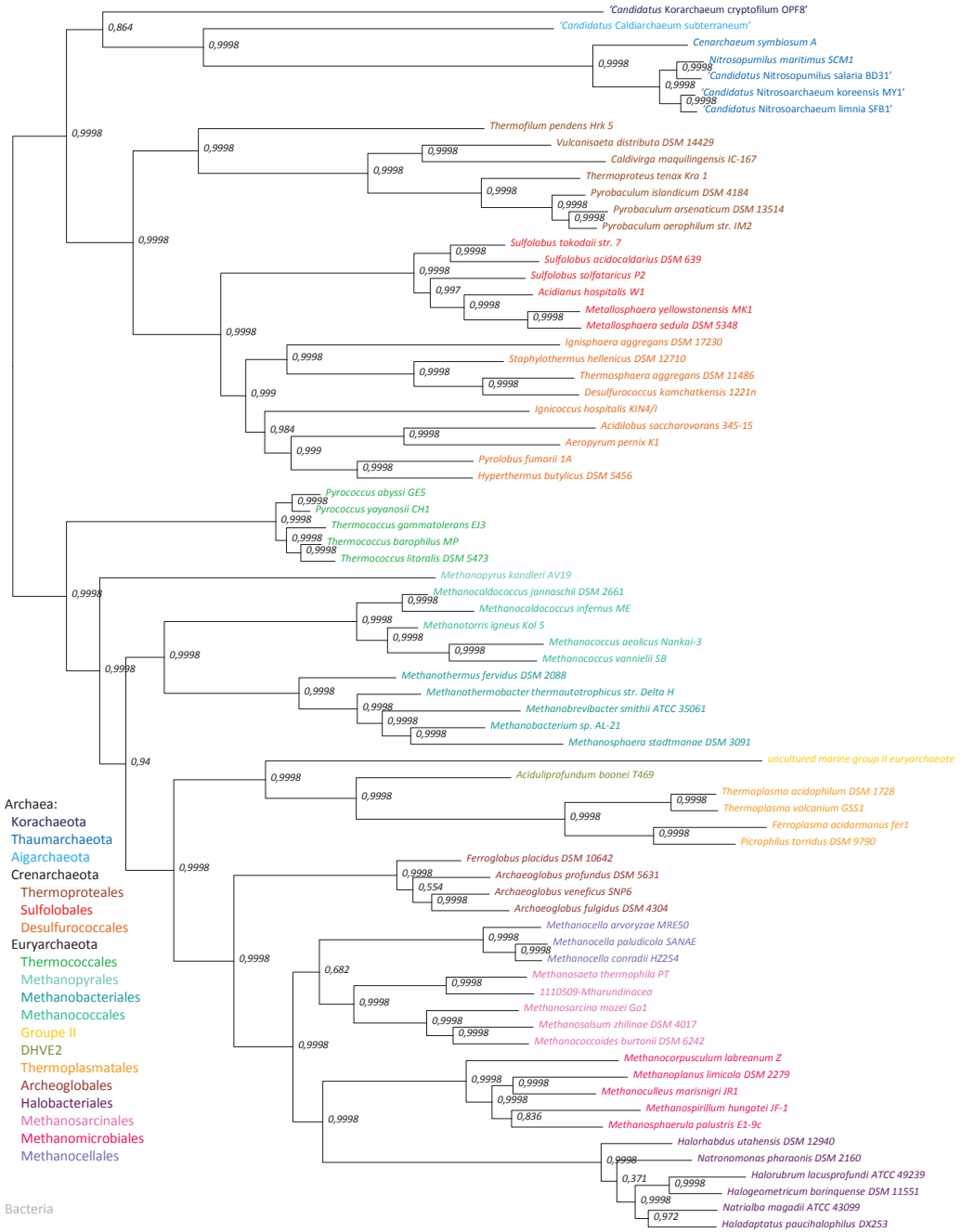
74 species – 23 000 positions

Petitjean et al.
Supplementary Figure S5
SF$_{16}$ PhyML-SH LG+G8

74 species – 24 143 positions

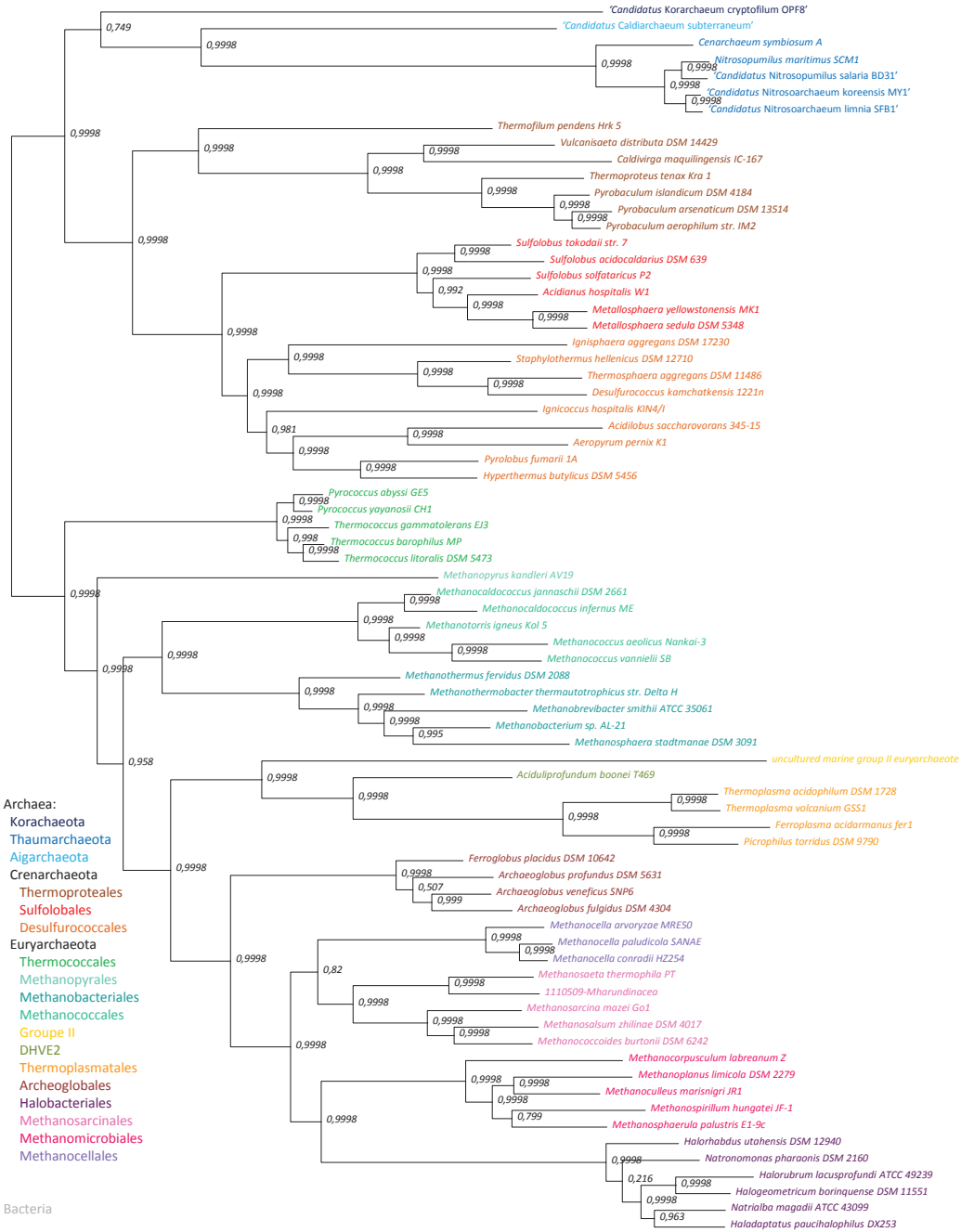
Petitjean et al.
Supplementary Figure S5
SF$_{17}$ PhyML-SH LG+G8

74 species – 25 237 positions



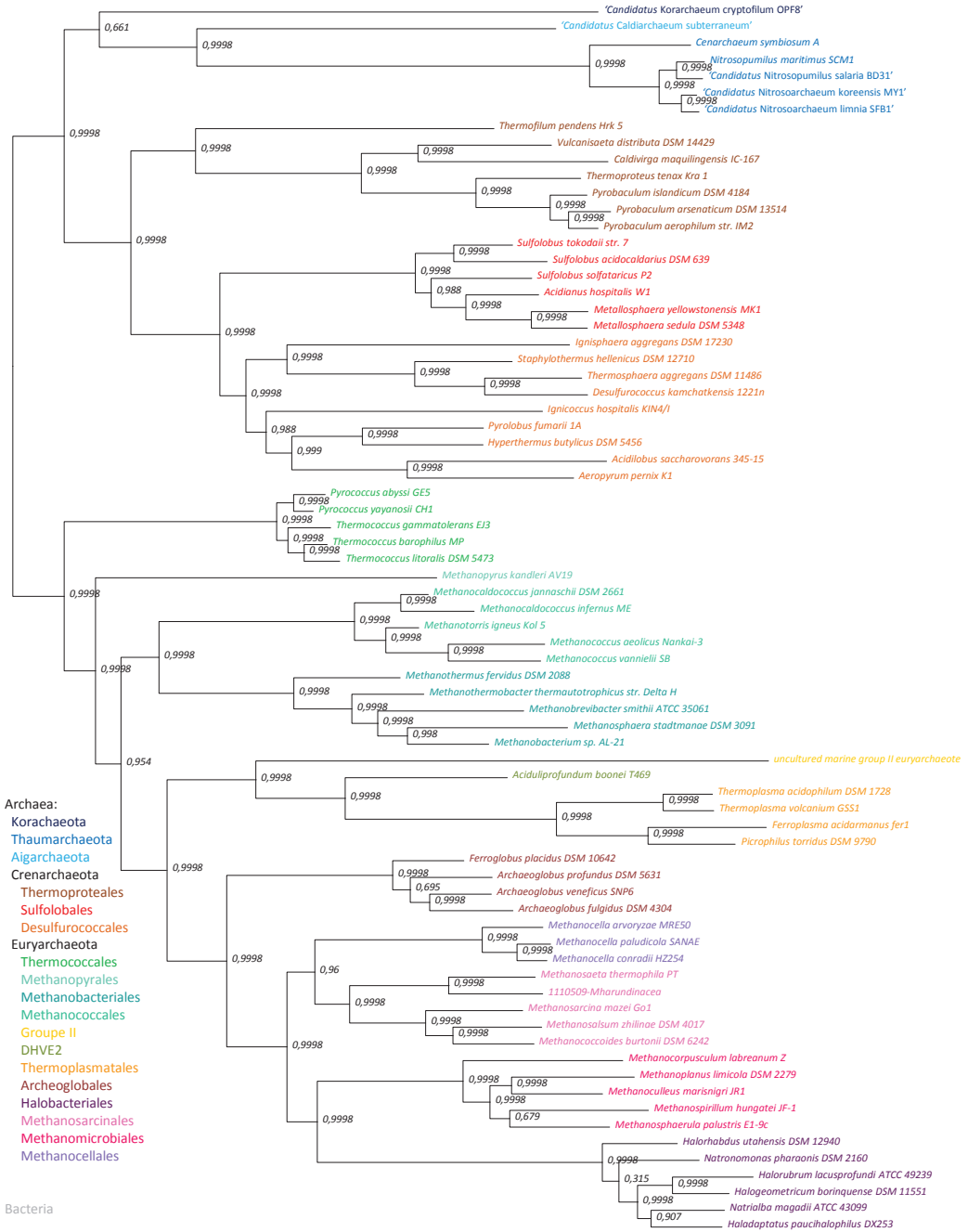
Petitjean et al.
Supplementary Figure S5
SF$_{19}$ PhyML-SH LG+G8

74 species – 27 217 positions



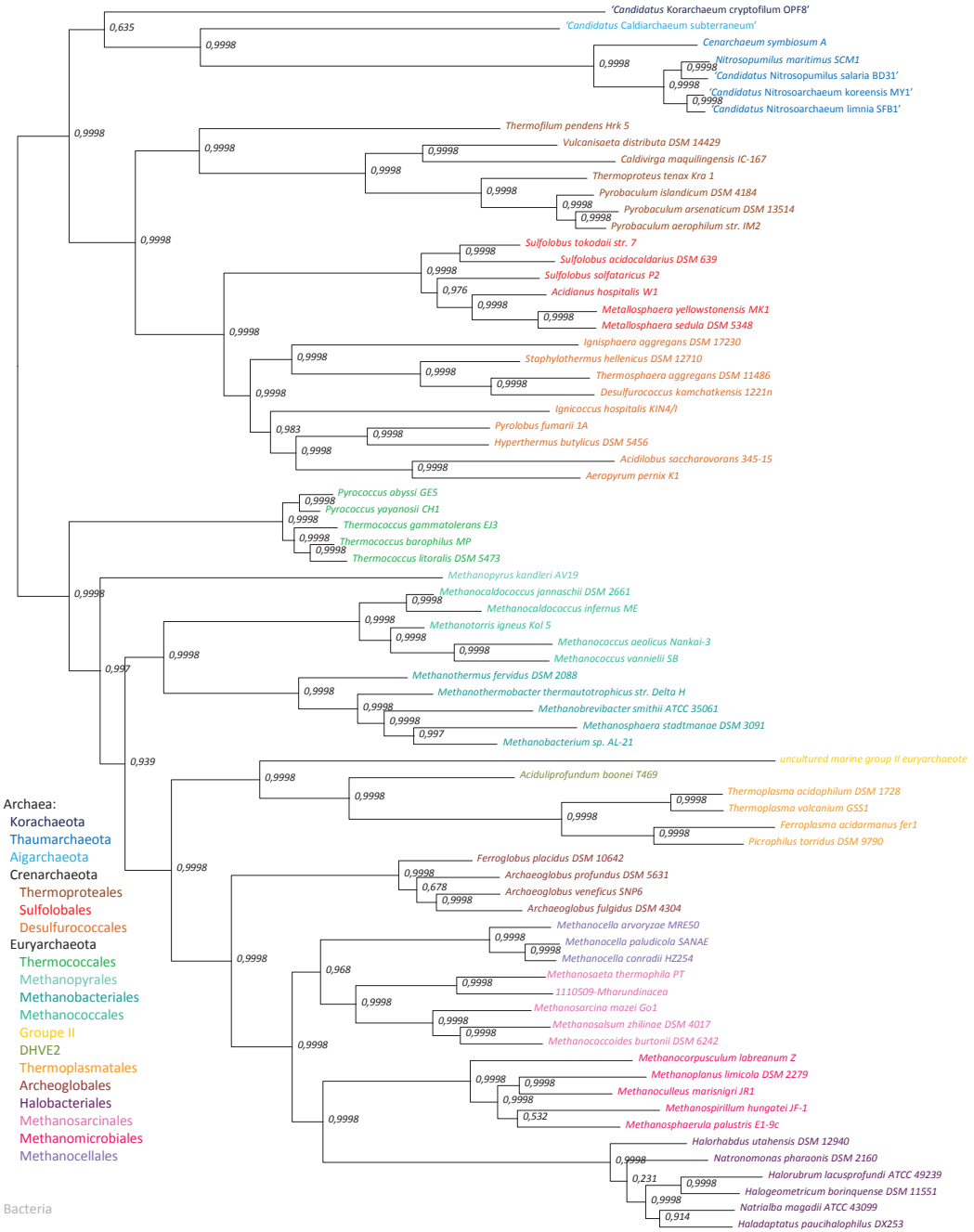
Petitjean et al.
Supplementary Figure S5
SF$_{20}$ PhyML-SH LG+G8

74 species – 28 146 positions



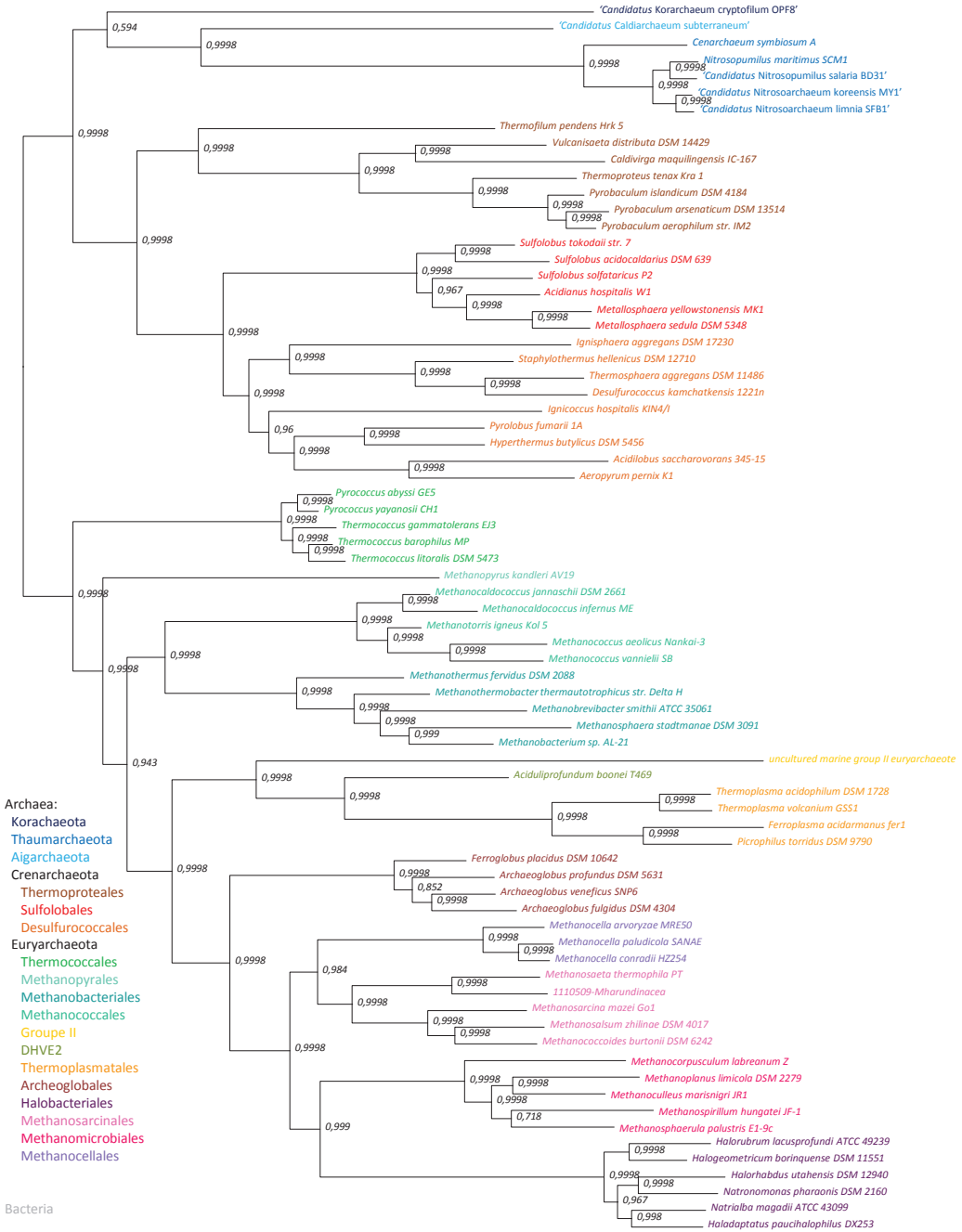
Petitjean et al.
Supplementary Figure S5
SF_{22} PhyML-SH LG+G8

74 species – 29 825 positions



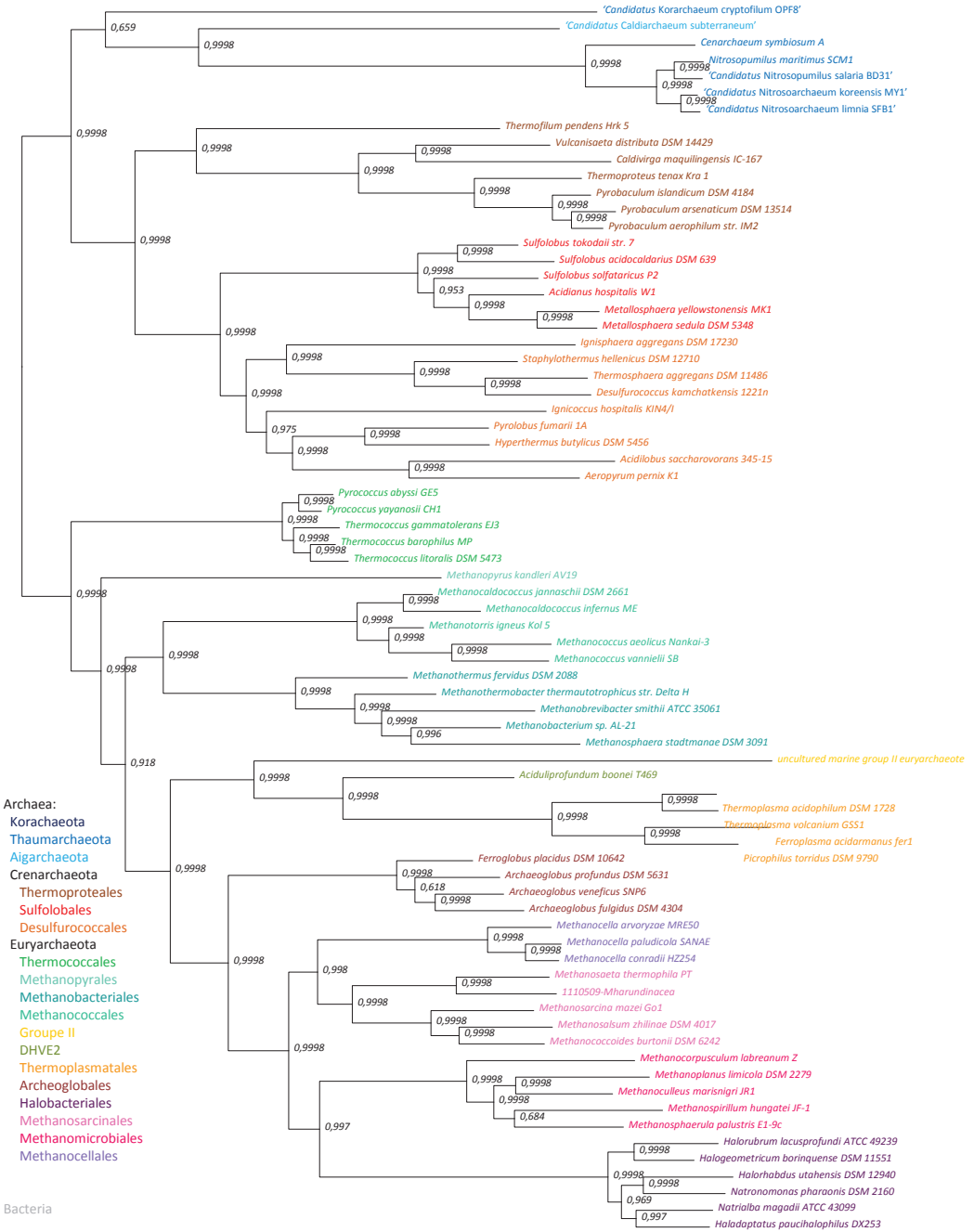
Petitjean et al.
Supplementary Figure S5
SF_{23} PhyML-SH LG+G8

74 species – 30 561 positions



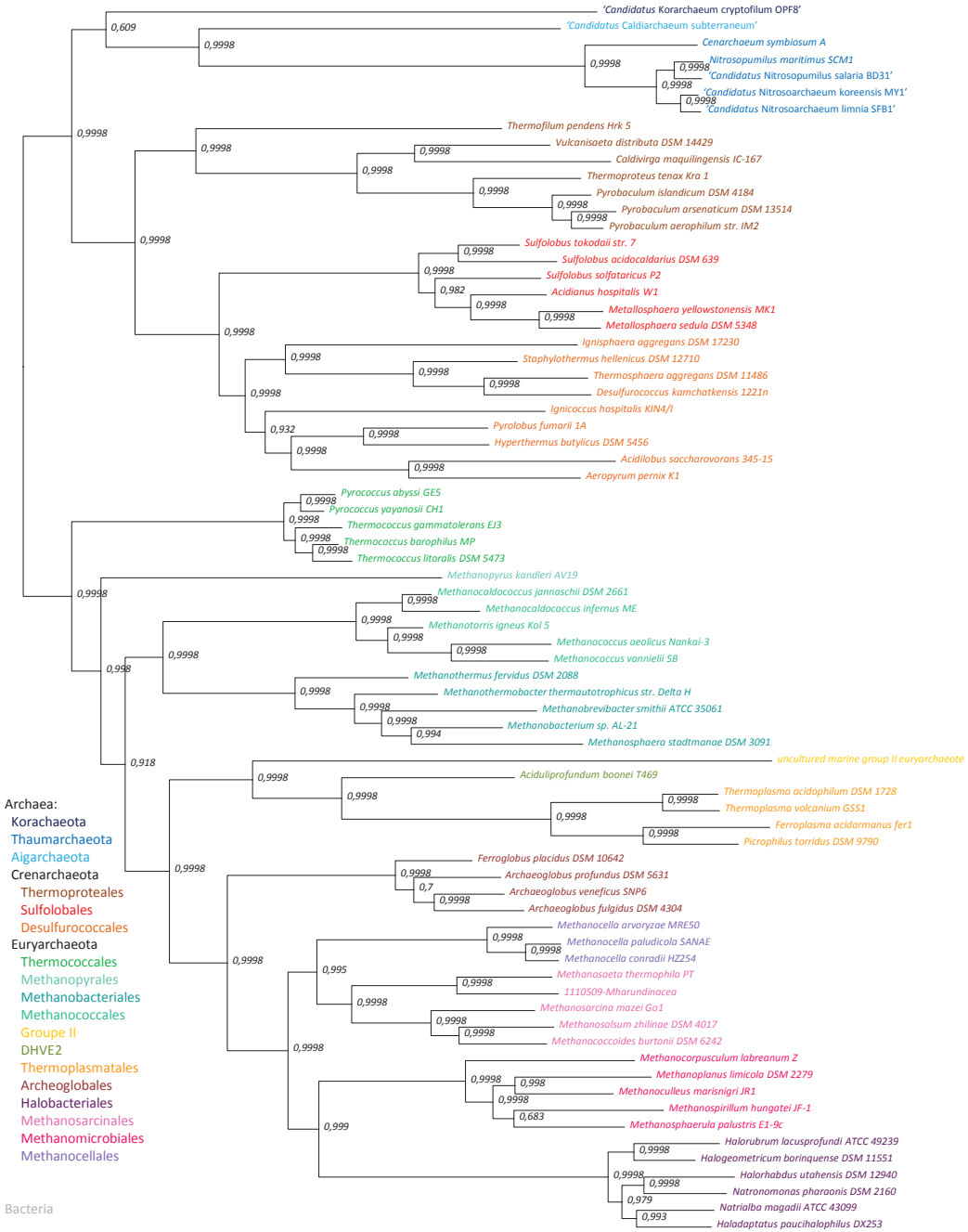
Petitjean et al.
Supplementary Figure S5
SF_{24} PhyML-SH LG+G8

74 species – 31 225 positions



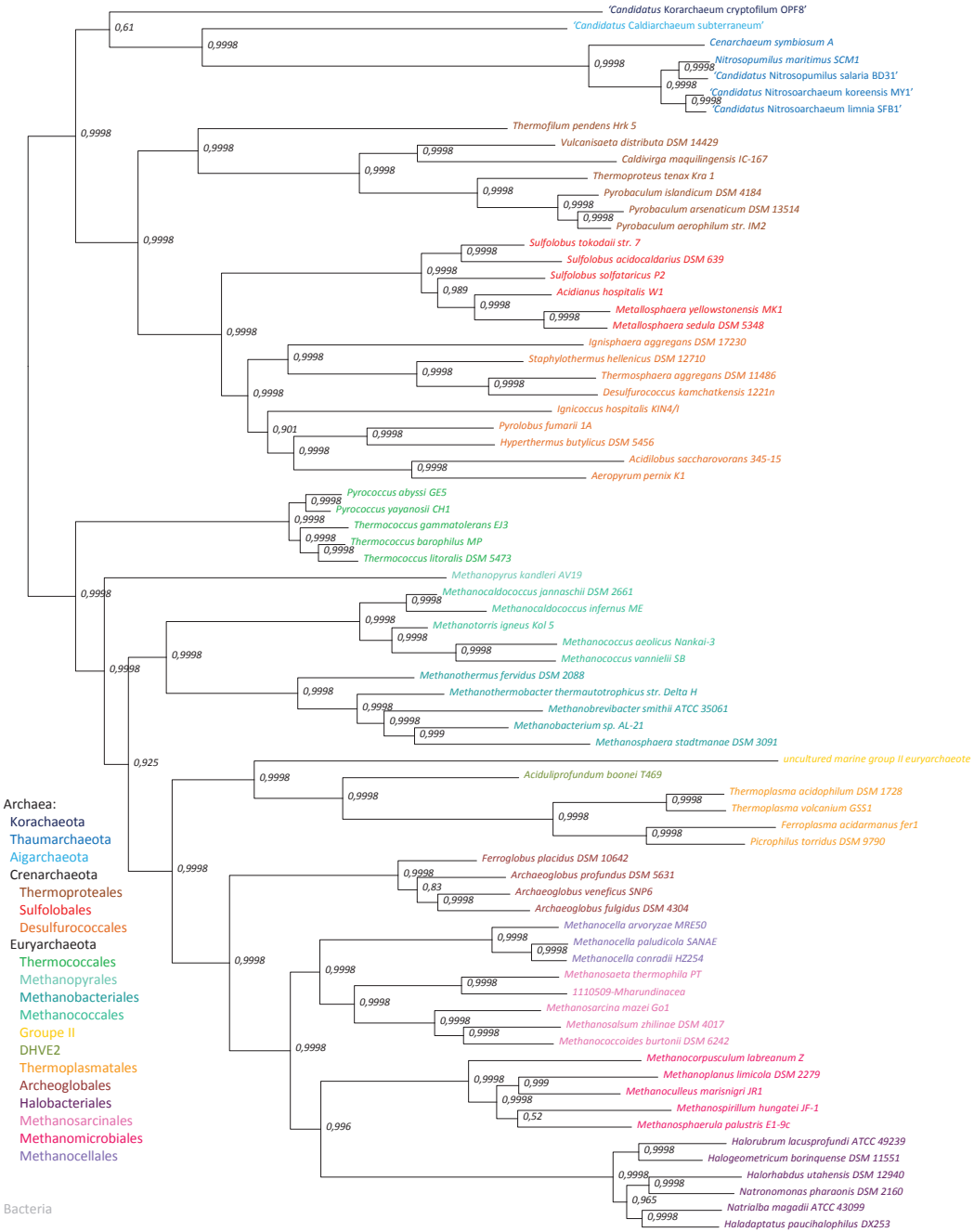
Petitjean et al.
Supplementary Figure S5
SF_{26} PhyML-SH LG+G8

74 species – 22 472 positions



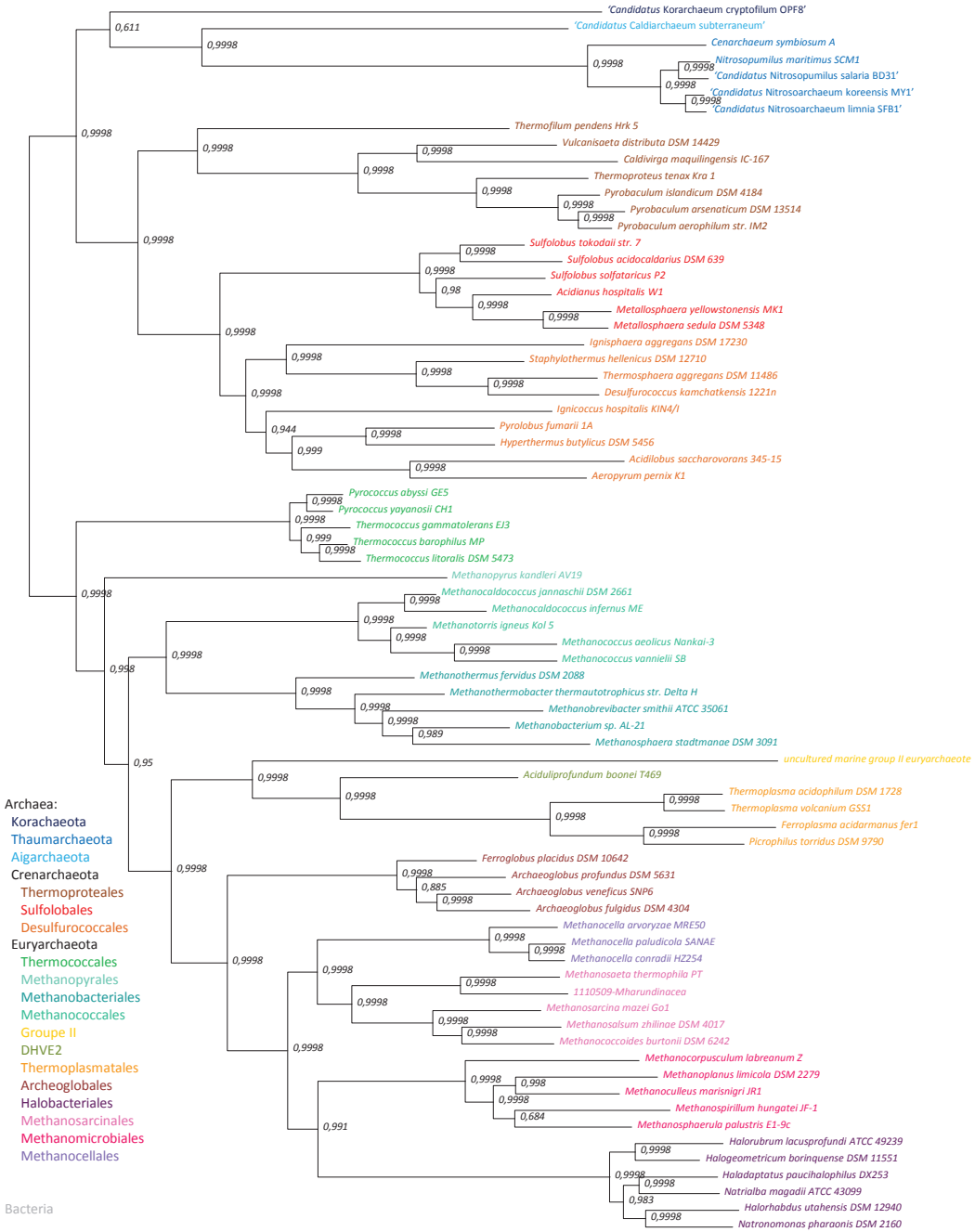
Petitjean et al.
Supplementary Figure S5
SF$_{27}$ PhyML-SH LG+G8

74 species – 32 999 positions



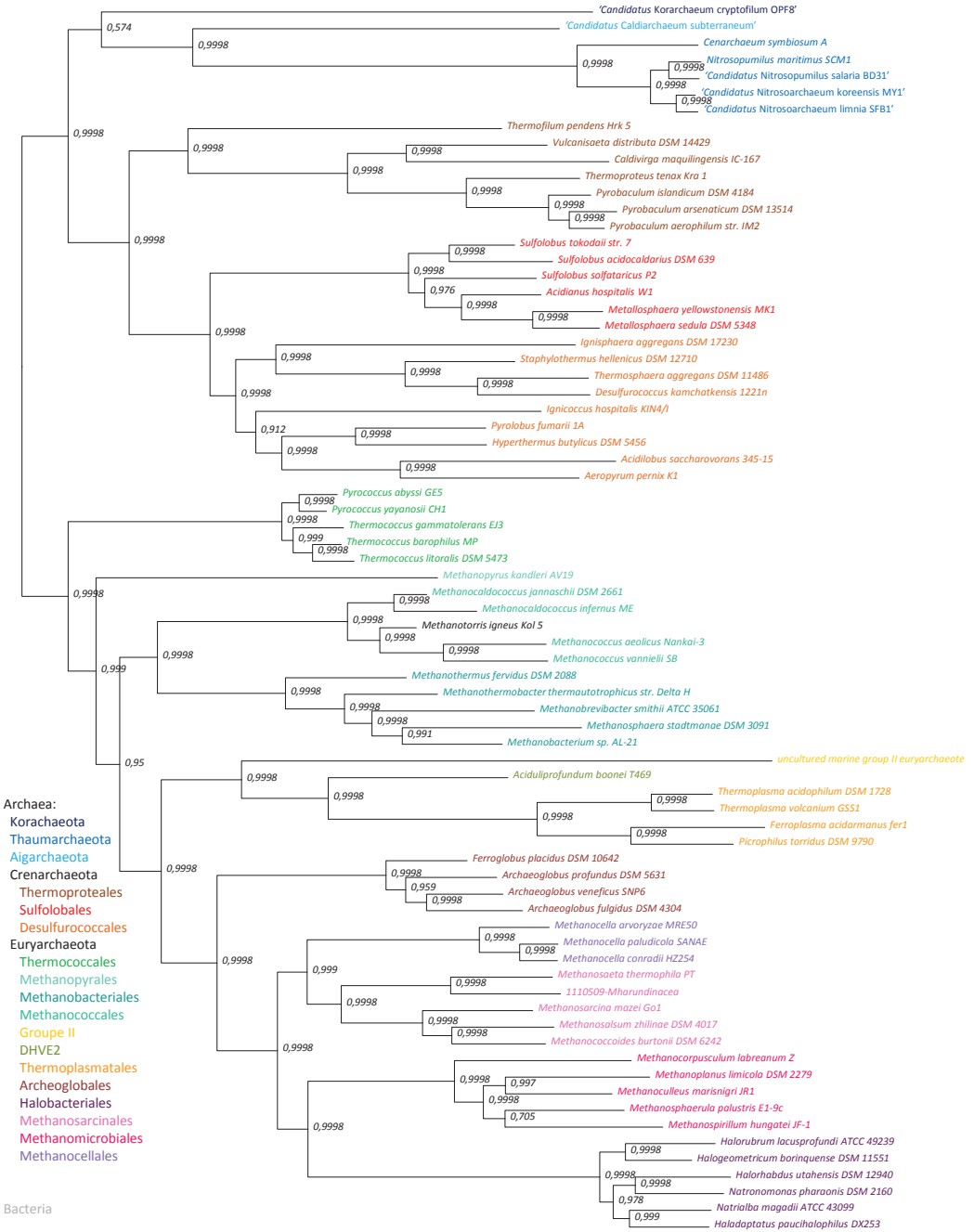
Petitjean et al.
Supplementary Figure S5
SF$_{29}$ PhyML-SH LG+G8

74 species – 33 799 positions



Petitjean et al.
Supplementary Figure S5
SF$_{31}$ PhyML-SH LG+G8

74 species – 34 449 positions



Petitjean et al.
Supplementary Figure S5
SF$_{34}$ PhyML-SH LG+G8

74 species – 35 589 positions

Annexe 2 : Listes et annotation des protéines des trois génomes étudiés.

a. Liste des protéines de *Nitrosopumilus maritimus*

b. Liste des protéines de *Cenarchaeum symbiosum*

c. Liste des protéines de '*Candidatus* Caldiarchaeum subterraneum'

La colonne « Taille » correspond à la taille de la protéine en acides aminés.

Les colonnes « Alignement » correspondent aux nombre de positions et nombre de séquences des alignements générés automatiquement au départ de l'analyse.

Dans la colonne « Spéciale archées », une protéine annotée « 1 » correspond à l'absence d'homologue chez les bactéries et les eucaryotes, une protéine annotée « 0 » correspond à la présence d'au moins un homologue dans un de ces domaines.

Dans les autres colonnes « Archaea », « Eucarya » et « Bacteria », une protéine annotée « 1 » correspond à la présence d'au moins un homologue dans le groupe considéré, une protéine annotée « 0 » correspond à l'absence d'homologue dans ce groupe.

Dans la colonne « Marqueur », une protéine annotée « M », « M2 » ou « M3 » est considérée comme un marqueur potentiel de premier choix, deuxième choix ou troisième choix respectivement.

Dans la colonne « THG potentiel », sont notés les transferts horizontaux potentiels :

Les lettres correspondent à différents groupes taxonomiques :

A : Archées

B : Bactéries

E : Eucaryotes

T : Thaumarchées

EA : Euryarchées

CA : Crenarchées

K : Korarchées

AA : Aigarchées

Les flèches marquent le sens du transfert potentiel :

B=>A : Transfert potentiel de bactérie vers une ou plusieurs archées.

B<=>A : Transferts potentiel entre bactéries et archées, mais dont le sens reste à déterminer.

a.	Sequence de départ				Alignement		Présence (1) /Absence (0) d'homologues dans les phylogénies							Observations	
							Archaea								
	Référence GI	Référence RefSeq	annotation	taille (aa)	Nombre de positions	Nombre de sequences	spéciale archée	Eury-archées	Cren-archées	Aig-archée	Kor-archée	Eucarya	Bacteria	Marqueur potentiel	THG potentiel
1	161527513	YP_001581339.1	cell division control protein 6 family protein	400	174	200	1	1	1	1	1	0	0	M	
2	161527514	YP_001581340.1	DNA polymerase II small subunit	481	218	88	0	1	0	1	1	1	0	M	
3	161527515	YP_001581341.1	hypothetical protein Nmar_0003	168	163	5	1	0	0	0	0	0	0		
4	161527517	YP_001581343.1	precorrin-6y C5 15-methyltransferase (decarboxylating) CbiE subunit	238	112	69	0	1	1	0	0	0	1		
5	161527518	YP_001581344.1	hypothetical protein Nmar_0006	163	161	4	1	0	0	0	0	0	0		
6	161527519	YP_001581345.1	peptidylprolyl isomerase	509	75	200	0	1	0	0	0	1	1		B<=>T
7	161527520	YP_001581346.1	FAD-dependent pyridine nucleotide-disulphide oxidoreductase	451	197	200	0	0	0	0	0	1	1		
8	161527521	YP_001581347.1	hypothetical protein Nmar_0009	216	209	6	1	0	1	0	0	0	0		
9	161527522	YP_001581348.1	heat shock protein HSP20	163	105	55	1	1	1	1	1	0	1		
10	161527523	YP_001581349.1	AAA family ATPase CDC48 subfamily protein	713	574	200	0	1	1	1	1	1	1	M2	T=>B
11	161527525	YP_001581351.1	transcription factor TFIIB cyclin-related	306	244	200	0	1	1	1	1	1	0	M2	
12	161527527	YP_001581353.1	camphor resistance CrcB protein	126	81	56	0	1	1	0	0	1	1		
13	161527532	YP_001581358.1	transcription factor TFIIB cyclin-related	301	255	200	1	1	1	1	1	0	0	M2	
14	161527533	YP_001581359.1	sodium/hydrogen exchanger	484	248	39	0	1	1	1	0	0	1		
15	161527534	YP_001581360.1	hypothetical protein Nmar_0022	180	177	7	1	1	1	0	0	0	0		
16	161527535	YP_001581361.1	adenylylsulfate kinase	179	132	200	0	1	1	0	0	1	1		
17	161527536	YP_001581362.1	inositol monophosphatase	271	105	200	0	1	1	0	0	1	1		
18	161527538	YP_001581364.1	short-chain dehydrogenase/reductase SDR	269	131	200	0	1	1	0	1	1	1		
19	161527539	YP_001581365.1	hypothetical protein Nmar_0027	142	138	4	1	0	0	0	0	0	0		
20	161527541	YP_001581367.1	hypothetical protein Nmar_0029	215	130	42	1	0	1	1	0	0	0	M2	
21	161527542	YP_001581368.1	hypothetical protein Nmar_0030	334	83	37	1	0	0	0	0	0	0		
22	161527545	YP_001581371.1	hypothetical protein Nmar_0033	86	92	6	1	0	0	0	0	0	0		
23	161527546	YP_001581372.1	hypothetical protein Nmar_0034	335	176	18	1	1	1	0	0	0	0		
24	161527549	YP_001581375.1	radical SAM domain-containing protein	237	65	143	0	1	1	0	0	0	1		EA=>T
25	161527550	YP_001581376.1	GTP cyclohydrolase I	185	143	200	0	0	1	0	0	1	1		B=>A
26	161527552	YP_001581378.1	ExsB family protein	224	137	200	0	1	1	0	0	1	1		
27	161527553	YP_001581379.1	homoserine kinase	307	124	200	0	1	1	0	1	1	1	M3	
28	161527554	YP_001581380.1	GTPase	252	112	200	0	1	1	1	0	1	0		
29	161527556	YP_001581382.1	short-chain dehydrogenase/reductase SDR	264	98	200	0	1	1	0	0	1	1		B=>T
30	161527558	YP_001581384.1	hypothetical protein Nmar_0046	103	79	4	1	0	0	0	0	0	0		
31	161527559	YP_001581385.1	hypothetical protein Nmar_0047	411	345	8	1	1	0	0	0	0	0		
32	161527560	YP_001581386.1	thiamine biosynthesis ATP pyrophosphatase-like protein	387	368	5	1	1	0	0	0	0	0		
33	161527561	YP_001581387.1	adenylylsulfate reductase thioredoxin dependent	241	114	200	0	1	1	0	0	1	1		B=>T
34	161527562	YP_001581388.1	sulfate adenylyltransferase	380	213	200	0	1	1	0	0	1	1		
35	161527563	YP_001581389.1	hypothetical protein Nmar_0051	221	216	6	1	0	0	0	0	0	0		
36	161527564	YP_001581390.1	DNA-(apurinic or apyrimidinic site) lyase	218	125	200	0	1	0	0	0	1	1		T<=>A
37	161527565	YP_001581391.1	hypothetical protein Nmar_0053	384	223	96	0	1	1	0	1	1	1	M2	A=>B
38	161527566	YP_001581392.1	hypothetical protein Nmar_0054	128	88	78	0	1	1	0	0	1	1	M2	
39	161527567	YP_001581393.1	hypothetical protein Nmar_0055	652	208	168	0	1	1	1	1	1	0	M	
40	161527568	YP_001581394.1	hypothetical protein Nmar_0056	158	143	4	1	0	0	0	0	0	0		
41	161527569	YP_001581395.1	precorrin-6Y C5 15-methyltransferase (decarboxylating) CbiT subunit	196	83	162	0	1	1	0	0	1	1		A=>B
42	161527570	YP_001581396.1	precorrin-2 C20-methyltransferase	240	74	200	0	1	1	0	0	0	1		B=>T
43	161527571	YP_001581397.1	precorrin-4 C11-methyltransferase	256	174	200	0	1	1	0	0	0	1		
44	161527572	YP_001581398.1	glutamate racemase	251	99	200	0	0	0	0	0	1	1		B=>T
45	161527573	YP_001581399.1	hypothetical protein Nmar_0061	216	120	43	1	0	1	0	0	0	0	M	
46	161527574	YP_001581400.1	glycyl transferase family protein	386	126	200	0	1	1	1	1	1	1		
47	161527575	YP_001581401.1	glycyl-tRNA synthetase	485	284	200	0	1	1	1	1	1	1	M	A=>E
48	161527576	YP_001581402.1	apurinic endonuclease Apn1	279	123	178	0	1	1	1	0	1	1		B=>A
49	161527577	YP_001581403.1	DNA primase large subunit	341	133	70	1	1	1	1	1	0	0	M	
50	161527578	YP_001581404.1	DNA primase small subunit	373	138	69	1	1	1	1	1	0	0	M	
51	161527579	YP_001581405.1	hypothetical protein Nmar_0067	953	493	62	0	1	0	1	0	0	1		B=>T // B=>EA
52	161527580	YP_001581406.1	glutamine amidotransferase class-I	180	108	200	0	1	1	0	1	1	1	M2	
53	161527581	YP_001581407.1	TPR repeat-containing protein	320	95	28	0	1	0	0	0	1	1		E=>T
54	161527583	YP_001581409.1	ribosomal protein S6e	137	101	97	1	1	1	1	1	0	0	M	
55	161527584	YP_001581410.1	translation initiation factor IF-2 subunit gamma	420	347	200	0	1	1	1	1	1	0	M	
56	161527586	YP_001581412.1	PEBP family protein	149	80	200	0	1	1	0	1	1	1		
57	161527587	YP_001581413.1	dolichyl-diphosphooligosaccharide--protein glycotransferase	712	357	167	0	1	1	1	1	1	0	M2	
58	161527588	YP_001581414.1	RNA-binding protein Nop10p	51	50	38	1	1	1	0	0	1	0	M2	A=>E
59	161527589	YP_001581415.1	RNA-binding S1 domain-containing protein	265	151	173	0	1	1	1	1	1	0	M	
60	161527590	YP_001581416.1	cob(I)alamin adenosyltransferase	176	104	200	0	1	1	0	1	1	1		B=>T // B=>EA
61	161527591	YP_001581417.1	cobyrinic acid a c-diamide synthase	450	237	200	0	1	1	1	1	1	1		
62	161527592	YP_001581418.1	precorrin-8X methylmutase CbiC/CobH	213	105	200	0	1	1	0	0	0	1		B=>T/B=>C+EA
63	161527593	YP_001581419.1	cobalamin (vitamin B12) biosynthesis CbiX protein	250	70	84	0	1	0	0	0	1	1		T<=>E
64	161527594	YP_001581420.1	cobalamin (vitamin B12) biosynthesis CbiG protein	350	114	143	0	1	1	0	0	0	1		B=>T // B=>A
65	161527595	YP_001581421.1	cobalamin biosynthesis protein CbiD	361	150	148	0	1	1	0	0	0	1		B=>T // B=>A
66	161527596	YP_001581422.1	NAD-dependent epimerase/dehydratase	289	78	162	0	1	0	0	0	1	1		B=>T // B=>EA
67	161527597	YP_001581423.1	methionine adenosyltransferase	388	334	200	0	1	0	0	0	1	1		B=>T // B=>EA
68	161527598	YP_001581424.1	like-Sm ribonucleoprotein core	81	57	181	0	1	1	1	1	1	0	M	
69	161527599	YP_001581425.1	RecA/RadA recombinase-like protein	217	229	200	0	1	1	1	1	1	0	M2	E<=>A
70	161527600	YP_001581426.1	DEAD/DEAH box helicase domain-containing protein	913	336	200	0	1	1	1	0	0	1	M2	
71	161527601	YP_001581427.1	nucleic acid binding OB-fold tRNA/helicase-type	466	167	33	1	1	0	0	0	0	0		T<=>EA
72	161527602	YP_001581428.1	ABC transporter related	312	159	200	0	1	1	0	0	0	1		
73	161527603	YP_001581429.1	ABC-2 type transporter	253	167	77	0	1	1	0	1	0	1		A=>B
74	161527604	YP_001581430.1	UDP-glucose/GDP-mannose dehydrogenase	246	245	5	0	0	1	0	0	1	0		
75	161527605	YP_001581431.1	hypothetical protein Nmar_0093	350	179	50	0	1	1	1	1	0	1	M2	A=>B
76	161527607	YP_001581433.1	GrpE protein	187	74	200	0	1	0	0	0	1	1		B=>A
77	161527608	YP_001581434.1	chaperone protein DnaK	636	531	200	0	1	0	1	0	1	1		B=>A
78	161527609	YP_001581435.1	chaperone protein DnaJ	361	291	200	0	1	0	0	0	1	1		B=>A
79	161527610	YP_001581436.1	hypothetical protein Nmar_0098	177	177	6	1	1	0	1	0	0	0		
80	161527611	YP_001581437.1	glutamyl-tRNA(Gln) amidotransferase subunit E	634	240	200	0	1	1	1	1	1	1	M	
81	161527612	YP_001581438.1	glutamyl-tRNA(Gln) amidotransferase subunit D	429	158	200	0	1	1	1	1	1	1	M2	
82	161527613	YP_001581439.1	AAA family ATPase CDC48 subfamily protein	722	567	200	0	1	1	1	1	1	1	M2	T=>B
83	161527614	YP_001581440.1	50S ribosomal protein L2P	244	209	200	0	1	1	1	1	1	0	M	
84	161527615	YP_001581441.1	acylphosphatase	92	47	170	0	1	1	0	1	1	1		B=>A B=>E
85	161527617	YP_001581443.1	PP-loop domain-containing protein	304	159	200	0	1	1	0	1	1	1	M2	A=>B
86	161527618	YP_001581444.1	methyltransferase type 12	306	48	39	0	1	0	0	0	0	1		B=>T // B=>EA
87	161527619	YP_001581445.1	peptidase M50	398	132	67	0	1	1	1	0	0	1	M2	A=>B
88	161527620	YP_001581446.1	CutA1 divalent ion tolerance protein	103	50	200	0	1	1	1	1	1	1		
89	161527621	YP_001581447.1	eRF1 domain-containing 1 protein	349	137	144	0	1	1	1	1	1	0	M	
90	161527622	YP_001581448.1	hypothetical protein Nmar_0110	137	132	4	1	0	0	1	0	0	0		
91	161527623	YP_001581449.1	hypothetical protein Nmar_0111	150	91	200	0	1	1	1	1	1	1		T<=>B // A<=>B
92	161527625	YP_001581451.1	DtxR family iron dependent repressor	150	54	138	0	1	1	0	0	0	1		A<=>B
93	161527626	YP_001581452.1	transcriptional regulator TrmB	255	45	162	0	1	1	1	0	0	1		
94	161527627	YP_001581453.1	radical SAM domain-containing protein	559	122	185	0	1	1	1	1	1	1		
95	161527628	YP_001581454.1	glycosyl transferase group 1	341	55	200	0	1	1	0	1	1	1		A<=>B
96	161527629	YP_001581455.1	polysaccharide biosynthesis protein CapD	335	186	200	0	1	1	0	0	0	1		B=>T // B=>EA
97	161527632	YP_001581458.1	glycosyl transferase group 1	393	63	200	0	1	1	0	0	1	1		A<=>B
98	161527633	YP_001581459.1	NAD-dependent epimerase/dehydratase	295	137	200	0	1	1	1	1	1	1		A<=>B
99	161527634	YP_001581460.1	NAD-dependent epimerase/dehydratase	314	137	200	0	1	1	1	1	1	1		A<=>B
100	161527635	YP_001581461.1	glycosyl transferase family protein	346	64	200	0	1	1	0	0	0	1		A<=>B
101	161527636	YP_001581462.1	glucose-1-phosphate thymidyltransferase	351	153	200	0	1	1	0	0	1	1		A<=>B
102	161527637	YP_001581463.1	dTDP-glucose 4 6-dehydratase	329	256	200	0	1	1	0	0	0	1		A<=>B
103	161527639	YP_001581465.1	carbamoyltransferase	567	228	115	0	1	1	0	0	1	1		B=>T // B=>EA
104	161527643	YP_001581469.1	N-acyneuraminate-9-phosphate synthase	351	135	139	0	1	1	0	0	1	1		B=>T/B=>EA
105	161527644	YP_001581470.1	glutamine--scyllo-inositol transaminase	362	200	200	0	1	1	1	1	1	1		B=>T/B=>EA
106	161527646	YP_001581472.1	methionyl-tRNA formyltransferase-like protein	197	217	8	0	0	0		0	0	1		B=>T
107	161527647	YP_001581473.1	acylneuraminate cytidylyltransferase	252	80	64	0	1	0	0	0	0	1		B=>T // B=>EA
108	161527648	YP_001581474.1	oxidoreductase domain-containing protein	348	58	176	0	1	1	1	0	1	1		
109	161527649	YP_001581475.1	aminopeptidase domain-containing protein	412	281	18	0	1	0	0	0	0	1		B=T
110	161527650	YP_001581476.1	4p-phosphopantetheinyl transferase	115	68	73	0	1	0	0	0	1	1		B=>T
111	161527652	YP_001581478.1	pyruvate carboxyltransferase	352	155	200	0	1	1	1	1	0	1		B=>T
112	161527653	YP_001581479.1	AMP-dependent synthetase and ligase	485	207	200	0	1	1	1	0	1	1		B=>T // B=>EA
113	161527654	YP_001581480.1	cyclase family protein	214	51	200	0	1	1	0	0	1	1		
114	161527656	YP_001581482.1	aminoglycoside N3p-acetyltransferase	261	196	9	0	0	0	0	0	0	0		B=>T
115	161527658	YP_001581484.1	polysaccharide biosynthesis protein CapD	329	194	200	0	1	0	0	0	1	1		B=>T // B=>EA
116	161527660	YP_001581486.1	aminotransferase class-III	442	312	200	0	1	1	1	0	1	1		T<=>B
117	161527661	YP_001581487.1	acylneuraminate cytidylyltransferase	247	205	9	0	1	0	0	0	0	1		
118	161527662	YP_001581488.1	NAD-dependent epimerase/dehydratase	327	284	200	0	1	1	0	0	1	1		B=>A
119	161527663	YP_001581489.1	oxidoreductase domain-containing protein	335	125	29	0	1	1	1	0	0	1		
120	161527664	YP_001581490.1	sulfatase	454	329	7	0	1	0	0	0	1	1		
121	161527666	YP_001581492.1	sulfatase	374	162	52	0	1	0	0	0	1	1		B=>T
122	161527667	YP_001581493.1	sugar nucleotidyltransferase-like protein	247	150	16	0	1	0	0	0	0	1		
123	161527668	YP_001581494.1	putative phosphoenolpyruvate phosphomutase	299	155	200	0	1	1	0	0	1	1		B=>A
124	161527669	YP_001581495.1	iron-containing alcohol dehydrogenase	311	141	200	0	1	1	0	0	1	1		B=>T // B=>A
125	161527670	YP_001581496.1	Rieske (2Fe-2S) domain-containing protein	98	79	18	0	0	1	1	0	0	1		B=>T
126	161527671	YP_001581497.1	hypothetical protein Nmar_0163	440	116	18	0	1	0	0	0	1	1		B=>T
127	161527672	YP_001581498.1	signal transduction protein	274	151	33	1	1	1	0	0	0	0		
128	161527673	YP_001581499.1	glycosyl transferase family protein	321	79	200	0	1	1	0	1	0	1		
129	161527674	YP_001581500.1	sulfotransferase	295	48	127	0	1	0	0	0	1	1		B=>T
130	161527675	YP_001581501.1	RpiR family transcriptional regulator	323	159	31	0	1	1	1	0	0	1		
131	161527676	YP_001581502.1	NAD-dependent epimerase/dehydratase	308	151	200	0	1	1	1	1	1	1		
132	161527677	YP_001581503.1	aIF-2B1 family translation initiation factor	345	212	200	0	1	1	1	1	1	1	M3	B<=>A
133	161527680	YP_001581506.1	hypothetical protein Nmar_0172	212	56	200	0	1	1	1	1	0	1		B<=>A
134	161527681	YP_001581507.1	short-chain dehydrogenase/reductase SDR	215	123	200	0	1	1	0	1	0	1		B=>T // CA=>EA
135	161527683	YP_001581509.1	hypothetical protein Nmar_0175	197	184	4	1	0	0	0	0	0	0		
136	161527685	YP_001581511.1	alanine aminotransferase	391	250	200	0	1	1	0	1	0	1		
137	161527686	YP_001581512.1	pyridoxamine 5p-phosphate oxidase-related FMN-binding	133	128	8	0	0	0	1	0	0	1		
138	161527688	YP_001581514.1	NADP oxidoreductase coenzyme F420-dependent	223	148	75	0	1	0	1	0	0	1	M2	EA=>B // EA=>T
139	161527689	YP_001581515.1	phosphoglycerate mutase	207	47	200	0	1	1	0	0	1	1		
140	161527691	YP_001581517.1	cytochrome c oxidase subunit II	143	70	190	0	1	1	1	0	1	1		B=>T // B=>A
141	161527692	YP_001581518.1	cytochrome c oxidase subunit I	508	449	200	0	1	0	1	0	1	1		
142	161527693	YP_001581519.1	blue (type1) copper domain-containing protein	315	47	49	0	1	0	1	0	0	1		
143	161527694	YP_001581520.1	hypothetical protein Nmar_0186	137	166	44	0	1	0	0	0	0	1		A<=>B
144	161527696	YP_001581522.1	hypothetical protein Nmar_0188	198	113	4	0	0	0	0	0	0	1		
145	161527697	YP_001581523.1	histidine kinase	598	77	200	0	1	0	0	0	0	1		B=>T
146	161527698	YP_001581524.1	methyltransferase type 11	184	59	22	0	1	0	0	0	0	1		
147	161527700	YP_001581526.1	branched-chain amino acid aminotransferase	304	213	200	0	1	1	1	0	0	1		B<=>A
148	161527701	YP_001581527.1	aspartate/glutamate/uridylate kinase	471	187	200	0	1	1	1	1	1	1		
149	161527702	YP_001581528.1	glycosyltransferase family 28 protein	340	202	27	0	1	0	0	0	0	1		B=>A
150	161527704	YP_001581530.1	hypothetical protein Nmar_0196	115	109	7	1	0	0	0	0	0	0		
151	161527705	YP_001581531.1	hypothetical protein Nmar_0197	121	104	11	1	0	0	0	0	0	0		
152	161527709	YP_001581535.1	radical SAM domain-containing protein	293	135	57	0	1	0	0	0	1	1		B=>T // B=>EA
153	161527710	YP_001581536.1	ArsR family transcriptional regulator	100	71	39	1	1	1	1	0	0	0	M2	
154	161527712	YP_001581538.1	hypothetical protein Nmar_0204	147	139	5	1	0	0	0	0	0	0		
155	161527713	YP_001581539.1	ribonuclease H	118	118	7	0	1	0	0	0	0	1		
156	161527714	YP_001581540.1	CoA-binding domain-containing protein	698	360	200	0	1	1	0	0	1	0		
157	161527715	YP_001581541.1	vinylacetyl-CoA Delta-isomerase	508	214	95	0	1	1	0	0	1	1		
158	161527716	YP_001581542.1	peptidase S26B signal peptidase	236	132	53	0	1	1	0	1	1	0		
159	161527717	YP_001581543.1	hypothetical protein Nmar_0209	256	103	7	1	0	0	0	0	0	0		
160	161527718	YP_001581544.1	thioredoxin	108	86	200	0	0	0	0	0	1	1		B=>T
161	161527719	YP_001581545.1	hypothetical protein Nmar_0211	119	118	4	1	0	0	0	0	0	0		
162	161527720	YP_001581546.1	hypothetical protein Nmar_0212	179	169	4	1	0	0	0	0	0	0		
163	161527722	YP_001581548.1	dihydropyrimidinase	464	183	200	0	1	1	1	0	1	1		B<=>A
164	161527723	YP_001581549.1	hypothetical protein Nmar_0215	275	100	200	0	1	1	1	1	1	1	M3	A=>B
165	161527724	YP_001581550.1	AMMECR1 domain-containing protein	205	73	200	0	1	1	1	1	1	1	M3	B=>EA

166	161527725	YP_001581551.1	radical SAM domain-containing protein	356	202	200	0	1	1	1	1	1	1		
167	161527726	YP_001581552.1	ribose 5-phosphate isomerase	224	136	200	0	1	1	0	0	1	1		B=>T
168	161527727	YP_001581553.1	hypothetical protein Nmar_0219	319	178	4	1	0	0	0	0	0	0		
169	161527728	YP_001581554.1	50S ribosomal protein L37e	55	57	55	0	1	1	1	0	1	0	M	
170	161527729	YP_001581555.1	like-Sm ribonucleoprotein core	78	60	84	1	1	1	1	1	1	0	M	
171	161527730	YP_001581556.1	creatininase	240	64	200	0	1	1	0	1	1	1		
172	161527731	YP_001581557.1	formyl transferase domain-containing protein	289	196	200	0	1	0	0	0	1	1		B=>A
173	161527732	YP_001581558.1	phosphoglucosamine mutase	449	212	200	0	1	1	1	1	0	1	M2	
174	161527733	YP_001581559.1	50S ribosomal protein L7Ae	128	87	200	0	1	1	1	1	1	0	M	
175	161527734	YP_001581560.1	30S ribosomal protein S28e	70	59	200	0	1	1	1	1	1	0	M	
176	161527735	YP_001581561.1	ribosomal protein L24E	66	50	115	0	1	1	1	1	1	0	M	
177	161527736	YP_001581562.1	nucleoside-diphosphate kinase	133	126	200	0	1	1	1	0	1	1		
178	161527737	YP_001581563.1	translation initiation factor IF-2	593	402	200	0	1	1	1	1	1	0	M	
179	161527738	YP_001581564.1	thioredoxin	133	84	200	0	1	1	1	0	1	1		B=>T // B=>A
180	161527739	YP_001581565.1	lg family protein	556	104	38	0	1	0	0	1	0	1		
181	161527740	YP_001581566.1	hypothetical protein Nmar_0232	138	84	94	0	1	1	1	1	1	1	M2	A=>B
182	161527741	YP_001581567.1	glyoxalase	122	87	21	0	1	0	0	0	0	1		
183	161527742	YP_001581568.1	SNARE associated Golgi protein	178	91	200	0	1	1	0	0	1	1		B=>A
184	161527744	YP_001581570.1	histidyl-tRNA synthetase	427	165	200	0	1	1	0	1	1	1	M2	
185	161527745	YP_001581571.1	translation initiation factor IF-6	221	164	200	0	1	1	1	1	1	0	M	
186	161527746	YP_001581572.1	4Fe-4S ferredoxin iron-sulfur binding domain-containing protein	163	98	42	1	1	1	1	0	0	0		
187	161527747	YP_001581573.1	4Fe-4S ferredoxin iron-sulfur binding domain-containing protein	178	98	42	1	1	1	1	0	0	0		
188	161527748	YP_001581574.1	replication factor C	316	193	200	0	1	1	1	1	1	0	M	
189	161527749	YP_001581575.1	hypothetical protein Nmar_0241	172	169	5	1	0	0	0	0	0	0		
190	161527750	YP_001581576.1	MCM family protein	695	303	200	0	1	1	1	1	1	0	M2	
191	161527751	YP_001581577.1	DEAD/DEAH box helicase domain-containing protein	709	294	200	0	1	1	1	1	1	1	M2	A=>B
192	161527752	YP_001581578.1	glycosyl transferase family protein	330	166	58	0	1	1	1	1	0	1	M3	A<=>B
193	161527753	YP_001581579.1	acetyltransferase	158	72	200	0	1	1	0	1	1	1		B=>T // B=>A
194	161527754	YP_001581580.1	hypothetical protein Nmar_0246	89	83	32	1	1	1	1	0	0	0	M2	
195	161527755	YP_001581581.1	oxidoreductase domain-containing protein	311	107	200	0	1	1	1	1	1	1		
196	161527756	YP_001581582.1	branched-chain amino acid aminotransferase	305	214	200	0	1	1	1	0	1	1		
197	161527757	YP_001581583.1	histidine kinase	459	79	200	0	1	0	0	0	1	1		B=>T
198	161527758	YP_001581584.1	ABC-type phosphate/phosphonate transport system periplasmic component-like protein	309	194	19	0	1	0	0	0	0	1		B=>T
199	161527759	YP_001581585.1	response regulator receiver protein	126	56	200	0	1	0	0	0	1	1		B=>T
200	161527760	YP_001581586.1	hypothetical protein Nmar_0252	201	188	5	1	0	0	0	0	0	0		
201	161527761	YP_001581587.1	oxidoreductase FAD/NAD(P)-binding subunit	281	89	200	0	1	0	0	0	1	1		B=>T
202	161527763	YP_001581589.1	Alba_DNA/RNA-binding protein	95	75	79	1	1	1	1	1	0	0	M2	
203	161527764	YP_001581590.1	putative methylisocitrate lyase	288	167	200	0	1	1	0	0	1	1		
204	161527765	YP_001581591.1	PadR-like family transcriptional regulator	169	59	41	0	1	1	0	0	0	1		B=>A
205	161527766	YP_001581592.1	succinate dehydrogenase or fumarate reductase flavoprotein subunit	570	416	200	0	1	1	1	1	1	1		A<=>B
206	161527767	YP_001581593.1	hypothetical protein Nmar_0259	143	139	8	1	0	1	1	0	0	1		
207	161527769	YP_001581595.1	succinate dehydrogenase and fumarate reductase iron-sulfur protein	249	153	200	0	1	1	1	1	1	1		B=>T+A
208	161527770	YP_001581596.1	hypothetical protein Nmar_0262	118	99	4	1	1	0	0	0	0	0		
209	161527771	YP_001581597.1	arginninosuccinate lyase	485	321	200	0	1	0	0	0	1	1		B=>T+EA
210	161527772	YP_001581598.1	hypothetical protein Nmar_0264	243	237	5	1	0	0	0	0	0	0		
211	161527773	YP_001581599.1	glutamyl-tRNA reductase	358	196	200	0	1	1	0	0	1	1	M2	
212	161527775	YP_001581601.1	oxidoreductase FAD/NAD(P)-binding subunit	280	115	113	0	1	0	1	0	0	1		B=>T // B=>EA
213	161527776	YP_001581602.1	inositol-phosphate phosphatase	266	92	200	0	1	1	0	1	1	1	M3	
214	161527777	YP_001581603.1	hypothetical protein Nmar_0269	373	357	5	1	0	0	0	0	0	0		
215	161527778	YP_001581604.1	ATPase central domain-containing protein	385	114	200	0	1	1	1	1	1	0	M	A=>EK
216	161527780	YP_001581606.1	carboxyl transferase	515	449	200	0	1	1	0	0	1	1		B=>A
217	161527781	YP_001581607.1	carbamoyl-phosphate synthase L chain ATP-binding	495	388	200	0	1	0	0	0	1	1		B=>T // B=>EA
218	161527782	YP_001581608.1	biotin/lipoyl attachment domain-containing protein	170	56	96	0	1	0	0	0	1	1		B=>T // B=>EA
219	161527783	YP_001581609.1	alkyl hydroperoxide reductase/ Thiol specific antioxidant/ Mal allergen	154	84	200	0	1	1	1	0	1	1		B=>A
220	161527784	YP_001581610.1	NADH-ubiquinone/plastoquinone oxidoreductase chain 3	110	106	46	0	1	0	0	0	1	1		B=>T // B=>EA
221	161527785	YP_001581611.1	NADH-quinone oxidoreductase_B subunit	174	140	200	0	1	1	1	1	1	1		
222	161527786	YP_001581612.1	NADH dehydrogenase (ubiquinone) 30 kDa subunit	200	60	200	0	1	1	1	1	1	1		B<=>A
223	161527787	YP_001581613.1	NADH dehydrogenase (quinone)	378	288	200	0	1	1	1	1	1	1		B<=>A
224	161527788	YP_001581614.1	NADH dehydrogenase (quinone)	432	265	200	0	1	1	1	1	1	1		B=>A
225	161527789	YP_001581615.1	4Fe-4S ferredoxin iron-sulfur binding domain-containing protein	165	63	200	0	1	1	1	1	1	1		B<=>A
226	161527790	YP_001581616.1	NADH-ubiquinone/plastoquinone oxidoreductase chain 6	170	119	60	0	1	0	0	0	0	1		B=>T // B=>EA
227	161527791	YP_001581617.1	NADH-ubiquinone oxidoreductase chain 4L	101	83	171	0	1	0	0	1	0	1		B=>A
228	161527792	YP_001581618.1	proton-translocating NADH-quinone oxidoreductase_chain M	517	313	200	0	1	1	1	0	1	1		A=>B
229	161527793	YP_001581619.1	proton-translocating NADH-quinone oxidoreductase_chain L	692	404	200	0	1	1	1	1	1	1		B=>EA
230	161527794	YP_001581620.1	proton-translocating NADH-quinone oxidoreductase_chain N	494	280	200	0	1	0	1	1	1	1		
231	161527795	YP_001581621.1	polyprenyl synthetase	283	143	200	0	1	1	0	1	1	1		EA=>B
232	161527796	YP_001581622.1	TENA/THI-4 domain-containing protein	246	194	16	0	0	0	0	0	0	1		B=>T
233	161527797	YP_001581623.1	orotate phosphoribosyltransferase	201	100	200	0	1	1	0	1	1	1		
234	161527798	YP_001581624.1	hypothetical protein Nmar_0290	360	352	8	1	1	0	0	0	0	0		T<=>EA
235	161527799	YP_001581625.1	peptidase M50	368	173	127	0	1	1	1	0	1	1	M2	
236	161527800	YP_001581626.1	ribosomal-protein-alanine acetyltransferase	171	68	200	0	1	1	1	1	1	1	M3	
237	161527801	YP_001581627.1	transcription regulator	72	71	5	1	0	1	0	0	0	0		
238	161527802	YP_001581628.1	protein of unknown function Mrt10	275	120	200	0	1	1	1	1	1	1	M2	A=>B
239	161527803	YP_001581629.1	putative ATPase RIL	595	456	200	0	1	1	1	1	1	1	M	A=>B
240	161527804	YP_001581630.1	leucyl aminopeptidase	488	248	200	0	0	1	0	0	0	1		B=>T // B=>C
241	161527806	YP_001581632.1	hypothetical protein Nmar_0298	102	48	32	0	1	0	0	0	0	1		EA<=>T // A=>B
242	161527807	YP_001581633.1	ribulose-phosphate 3-epimerase	555	98	200	0	1	1	0	0	1	1		B=>A
243	161527808	YP_001581634.1	transketolase central region	324	180	200	0	1	1	0	0	1	1		B=>A
244	161527809	YP_001581635.1	putative transaldolase	222	139	200	0	1	1	0	0	1	1		B=>A
245	161527811	YP_001581637.1	Mg2 transporter protein CorA family protein	340	119	194	0	1	0	0	0	1	1		B=>T // B=>A
246	161527812	YP_001581638.1	hypothetical protein Nmar_0304	129	105	10	1	0	0	0	0	0	0		
247	161527813	YP_001581639.1	hypothetical protein Nmar_0305	115	112	4	1	0	0	0	0	0	0		
248	161527814	YP_001581640.1	hypothetical protein Nmar_0306	253	137	28	0	1	0	0	0	1	1		
249	161527815	YP_001581641.1	hypothetical protein Nmar_0307	89	62	73	1	1	1	1	1	0	0	M	
250	161527816	YP_001581642.1	Rieske (2Fe-2S) domain-containing protein	102	35	121	0	1	0	1	0	1	1		B=>T // B=>A
251	161527817	YP_001581643.1	GCN5-related N-acetyltransferase	258	214	5	1	1	1	0	0	0	0		
252	161527818	YP_001581644.1	5-carboxymethyl-2-hydroxymuconate Delta-isomerase	289	145	200	0	1	1	1	1	1	1		A<=>B
253	161527819	YP_001581645.1	glutamyl-tRNA synthetase	569	246	200	0	1	1	1	1	1	1	M	
254	161527820	YP_001581646.1	polyprenyl synthetase	326	146	200	0	1	1	0	0	1	1		A<=>B
255	161527821	YP_001581647.1	isopentenyl-diphosphate delta-isomerase_type 1	216	85	150	0	1	1	0	0	1	1		B=>A
256	161527822	YP_001581648.1	aspartate/glutamate/uridylate kinase	247	83	120	0	1	1	1	1	1	1		
257	161527823	YP_001581649.1	mevalonate kinase	313	113	131	0	1	1	1	1	1	1		
258	161527824	YP_001581650.1	30S ribosomal protein S2	207	164	200	0	1	1	1	1	1	0	M	
259	161527825	YP_001581651.1	phosphopyruvate hydratase	412	341	200	0	1	1	1	1	0	1	M2	
260	161527826	YP_001581652.1	DNA-directed RNA polymerase subunit N	90	54	184	0	1	1	1	1	1	0	M	
261	161527827	YP_001581653.1	translin family protein	204	97	62	0	1	1	0	1	0	1		A=>B
262	161527828	YP_001581654.1	carbohydrate kinase_YjeF related protein	215	190	200	0	1	1	1	1	1	1	M2	A=>B/EK
263	161527829	YP_001581655.1	beta-lactamase domain-containing protein	202	86	200	0	1	1	1	0	1	1		B=T // B=>A
264	161527830	YP_001581656.1	hypothetical protein Nmar_0322	353	183	114	0	1	1	0	0	0	1	M2	A=>B
265	161527831	YP_001581657.1	glucosamine--fructose-6-phosphate aminotransferase_isomerizing	586	404	200	0	1	1	1	0	0	1		EA=>T // B=>EA
266	161527832	YP_001581658.1	30S ribosomal protein S4	201	149	200	0	1	1	1	1	1	0		
267	161527833	YP_001581659.1	30S ribosomal protein S13P	194	134	200	0	1	1	1	0	1			
268	161527834	YP_001581660.1	hypothetical protein Nmar_0326	102	99	4	1	0	0	0	0	0	0		
269	161527836	YP_001581662.1	periplasmic solute binding protein	323	115	200	0	0	0	0	0	0	1		B=>T
270	161527837	YP_001581663.1	ABC transporter related	256	122	200	0	1	0	0	0	0	1		B=>T // B=>EA
271	161527838	YP_001581664.1	ABC-3 protein	314	180	200	0	1	0	0	0	1	1		B=>T // B=>EA
272	161527839	YP_001581665.1	ABC-3 protein	368	192	200	0	1	0	0	0	1	1		B=>T // B=>EA
273	161527840	YP_001581666.1	cation diffusion facilitator family transporter	450	118	200	0	1	1	0	0	1	1		B=>A
274	161527841	YP_001581667.1	hypothetical protein Nmar_0333	1042	72	56	1	0	0	0	0	0	0		
275	161527842	YP_001581668.1	hypothetical protein Nmar_0334	343	116	28	0	0	0	0	1	0	1		T/K/B
276	161527843	YP_001581669.1	hypothetical protein Nmar_0335	107	107	4	1	0	0	0	0	0	0		
277	161527844	YP_001581670.1	radical SAM domain-containing protein	526	294	73	1	1	1	1	1	0	1	M3	
278	161527845	YP_001581671.1	1-(5-phosphoribosyl)-5-amino-4-imidazole-carboxylate (AIR) carboxylase	258	146	146	0	1	1	0	1	1	1		A=>B
279	161527846	YP_001581672.1	malate dehydrogenase	304	209	200	0	1	1	1	1	1	1		B=>A
280	161527847	YP_001581673.1	hypothetical protein Nmar_0339	406	170	153	0	1	1	0	1	1	1	M2	
281	161527848	YP_001581674.1	ExsB family protein	265	134	149	0	1	1	0	1	1	1		B<=>A // B=>EA
282	161527849	YP_001581675.1	cysteine desulfurase	388	300	200	0	1	0	0	0	1	1		B=>A
283	161527850	YP_001581676.1	hypothetical protein Nmar_0342	263	216	17	1	0	0	0	0	0	0		
284	161527851	YP_001581677.1	hypothetical protein Nmar_0343	294	185	20	1	0	0	0	0	0	0		
285	161527852	YP_001581678.1	hypothetical protein Nmar_0344	285	265	6	1	0	0	0	0	0	0		
286	161527853	YP_001581679.1	hypothetical protein Nmar_0345	279	154	25	1	0	0	0	0	0	0		
287	161527854	YP_001581680.1	RNA polymerase Rpb5	83	68	164	0	1	1	1	1	1	0	M	
288	161527855	YP_001581681.1	DNA-directed RNA polymerase subunit beta	1115	730	200	0	1	1	1	1	1	0	M	
289	161527856	YP_001581682.1	DNA-directed RNA polymerase subunit Ap	1263	583	200	0	1	1	1	1	1	0	M	
290	161527857	YP_001581683.1	like-Sm ribonucleoprotein core	80	61	98	1	1	1	0	1	0	0	M2	
291	161527858	YP_001581684.1	hypothetical protein Nmar_0350	392	336	18	1	0	0	0	0	0	0		
292	161527860	YP_001581686.1	ribosomal protein L7Ae/L30e/S12e/Gadd45	107	93	43	0	1	1	0	0	1	0	M	
293	161527861	YP_001581687.1	NusA family protein	154	102	104	1	1	1	1	1	0	0	M	
294	161527862	YP_001581688.1	30S ribosomal protein S12P	145	139	200	0	1	1	1	1	1	0	M	
295	161527863	YP_001581689.1	30S ribosomal protein S7P	199	181	200	0	1	1	1	1	1	0	M	
296	161527864	YP_001581690.1	hypothetical protein Nmar_0356	109	106	5	1	0	0	0	0	0	0		
297	161527865	YP_001581691.1	hypothetical protein Nmar_0357	234	234	6	1	0	0	0	0	0	0		
298	161527866	YP_001581692.1	GCN5-related N-acetyltransferase	149	65	33	0	1	0	0	0	1	1		B=>T
299	161527867	YP_001581693.1	dehydrogenase (flavoprotein)-like protein	332	252	16	1	1	1	1	0	0	0		
300	161527868	YP_001581694.1	nitroreductase	200	67	171	0	1	0	0	0	0	1		B=>T // B=>EA
301	161527874	YP_001581700.1	uracil-DNA glycosylase superfamily protein	217	77	141	0	1	1	1	0	0	1		
302	161527875	YP_001581701.1	leucyl-tRNA synthetase	955	456	200	0	1	1	1	1	1	0	M	
303	161527876	YP_001581702.1	alanyl-tRNA synthetase	889	413	200	0	1	1	1	1	1	1	M	
304	161527877	YP_001581703.1	alcohol dehydrogenase	321	125	200	0	1	1	0	0	1	1		B=>T // B=>A
305	161527878	YP_001581704.1	hypothetical protein Nmar_0370	130	104	14	1	0	0	0	0	0	0		
306	161527879	YP_001581705.1	glucose sorbosone dehydrogenase	410	115	200	0	1	1	1	0	1	1		B=>T // B=>A
307	161527880	YP_001581706.1	histidine kinase	505	68	200	0	1	0	0	0	1	1		B=>T
308	161527881	YP_001581707.1	hypothetical protein Nmar_0373	169	170	5	1	0	0	0	0	0	0		
309	161527882	YP_001581708.1	50S ribosomal protein L12P	105	96	88	1	1	1	1	1	0	0	M	
310	161527883	YP_001581709.1	hypothetical protein Nmar_0375	210	102	200	0	1	1	0	0	1	1		B=>T
311	161527884	YP_001581710.1	response regulator receiver protein	125	61	200	0	1	0	0	0	0	1		B=>T // B=>EA
312	161527888	YP_001581714.1	hypothetical protein Nmar_0380	357	184	4	1	0	0	0	0	0	0		
313	161527889	YP_001581715.1	acidic ribosomal protein P0	288	173	200	0	1	1	1	1	1	0	M	
314	161527890	YP_001581716.1	ribosomal protein L1	220	145	200	0	1	1	1	1	1	1	M	
315	161527891	YP_001581717.1	AsnC family transcriptional regulator	145	123	39	0	1	1	1	0	0	1		A=>B
316	161527893	YP_001581719.1	ribosomal protein L11	159	114	200	0	1	1	1	1	1	1		
317	161527894	YP_001581720.1	NusG antitermination factor	152	115	98	1	1	1	1	0	0	0	M	
318	161527895	YP_001581721.1	hypothetical protein Nmar_0387	258	100	97	0	1	1	1	1	1	0	M2	
319	161527896	YP_001581722.1	methylated-DNA--protein-cysteine methyltransferase	99	51	200	0	1	1	1	1	1	1		B<=>A
320	161527897	YP_001581723.1	glyoxalase/bleomycin resistance protein/dioxygenase	131	118	5	0	0	0	0	0	0	1		
321	161527898	YP_001581724.1	50S ribosomal protein L19e	151	127	150	0	1	1	1	1	1	0	M	
322	161527899	YP_001581725.1	ribosomal protein L32e	134	96	200	0	1	1	1	0	1	0	M	
323	161527900	YP_001581726.1	phosphoenolpyruvate carboxykinase	499	382	200	0	0	0	0	0	1	1		B=>T
324	161527901	YP_001581727.1	peptidase S8 and S53 subtilisin kexin sedolisin	660	172	200	0	1	0	0	0	1	1		B=>T // B=>EA
325	161527902	YP_001581728.1	superoxide dismutase	206	147	200	0	1	0	0	0	1	1		B=>T
326	161527903	YP_001581729.1	prefoldin_alpha subunit	145	131	14	1	1	0	0	0	0	0		
327	161527904	YP_001581730.1	signal recognition particle-docking protein FtsY	513	188	200	0	1	1	0	0	1	1	M3	A<=>B
328	161527905	YP_001581731.1	ornithine carbamoyltransferase	304	229	200	0	1	1	1	1	1	1		
329	161527906	YP_001581732.1	50S ribosomal protein L19P	160	137	192	0	1	1	1	1	1	0	M	
330	161527907	YP_001581733.1	30S ribosomal protein S5P	238	180	200	0	1	1	1	1	1	0	M	
331	161527908	YP_001581734.1	50S ribosomal protein L30P	155	135	96	1	1	1	1	1	0	0	M	
332	161527909	YP_001581735.1	ribosomal protein L15	154	112	53	0	1	1	1	1	1	0	M	
333	161527910	YP_001581736.1	preprotein translocase_SecY subunit	476	351	200	0	1	1	1	1	1	0	M	
334	161527911	YP_001581737.1	adenylate kinase	192	135	66	1	1	1	1	0	0	0	M2	

335	161527912	YP_001581738.1	membrane protein-like protein	207	208	7	1	0	0	0	0	0	0			
336	161527913	YP_001581739.1	cytidylate kinase	186	124	98	0	1	1	1	1	0	1	M		A=>B
337	161527914	YP_001581740.1	pseudouridylate synthase subunit TruB	332	247	200	0	1	1	0	1	1	0	M2		
338	161527915	YP_001581741.1	hypothetical protein Nmar_0407	148	147	7	1	0	0	0	0	0	0			
339	161527916	YP_001581742.1	secreted protein	160	81	19	0	0	0	0	0	0	1			T<=>B
340	161527917	YP_001581743.1	FAD dependent oxidoreductase	437	164	137	0	1	0	0	0	1	1			B=>T // B=>EA
341	161527918	YP_001581744.1	hypothetical protein Nmar_0410	93	86	7	1	0	0	0	0	0	0			
342	161527919	YP_001581745.1	hypothetical protein Nmar_0411	126	123	4	1	0	0	0	0	0	0			
343	161527920	YP_001581746.1	glyoxylate reductase	322	190	200	0	1	1	0	0	1	1			
344	161527921	YP_001581747.1	pyruvate flavodoxin/ferredoxin oxidoreductase domain-containing protein	636	191	200	1	1	1	1	0	0	1			
345	161527922	YP_001581748.1	pyruvate ferredoxin/flavodoxin oxidoreductase_beta subunit	320	162	200	1	1	1	1	0	0	1			
346	161527923	YP_001581749.1	hypothetical protein Nmar_0415	146	101	6	1	0	0	0	0	0	0			
347	161527924	YP_001581750.1	hypothetical protein Nmar_0416	177	172	6	1	0	0	0	0	0	0			
348	161527926	YP_001581752.1	hypothetical protein Nmar_0418	189	162	12	1	1	1	0	0	0	0			
349	161527927	YP_001581753.1	hypothetical protein Nmar_0419	259	167	103	0	1	0	1	1	0	1			A=>B
350	161527928	YP_001581754.1	3-isopropylmalate dehydratase	422	321	200	0	1	1	0	1	1	1			
351	161527929	YP_001581755.1	3-isopropylmalate dehydratase_small subunit	160	89	200	0	1	1	0	1	1	1			
352	161527930	YP_001581756.1	3-isopropylmalate dehydratase_large subunit	471	422	200	0	0	0	0	0	1	1			B=>T
353	161527931	YP_001581757.1	3-isopropylmalate dehydratase_small subunit	194	140	200	0	0	0	0	0	1	1			B=>T
354	161527932	YP_001581758.1	nucleotidyl transferase	239	108	200	0	1	1	1	0	0	1			
355	161527933	YP_001581759.1	ribosomal protein S9	149	130	200	0	1	1	0	1	1	0	M		
356	161527934	YP_001581760.1	ribosomal protein L13	154	109	200	0	1	1	1	0	1	0	M		
357	161527935	YP_001581761.1	ribosomal protein L15	115	107	92	1	1	1	1	1	0	0	M		
358	161527936	YP_001581762.1	RNA polymerase insert	221	100	200	0	1	1	1	1	1	0	M		
359	161527937	YP_001581763.1	hypothetical protein Nmar_0429	117	76	19	1	0	0	0	0	0	0			
360	161527938	YP_001581764.1	putative RNA-associated protein	228	117	200	0	1	1	1	1	1	0	M		
361	161527939	YP_001581765.1	exosome complex RNA-binding protein Rrp4	226	133	98	0	1	1	1	1	1	0	M		
362	161527940	YP_001581766.1	exosome complex exonuclease 1	244	131	200	0	1	1	1	1	1	1	M2		
363	161527941	YP_001581767.1	exosome complex RNA-binding protein Rrp42	272	122	200	0	1	1	1	1	1	0	M		
364	161527942	YP_001581768.1	hypothetical protein Nmar_0434	70	70	24	0	1	0	1	0	1	0			
365	161527943	YP_001581769.1	hypothetical protein Nmar_0435	81	81	4	1	0	0	0	0	0	0			
366	161527944	YP_001581770.1	prefoldin_beta subunit	129	120	8	1	0	1	1	0	0	0			
367	161527945	YP_001581771.1	ERCC4 domain-containing protein	228	73	178	0	1	1	0	1	1	1	M2		A=>B
368	161527946	YP_001581772.1	TPR repeat-containing protein	164	93	62	0	1	0	0	0	1	1			
369	161527947	YP_001581773.1	nucleotide binding protein_PINc	159	86	81	1	1	1	1	0	1	0	M2		
370	161527948	YP_001581774.1	ATP-NAD/AcoX kinase	278	97	200	0	1	1	1	0	0	1			B=>A
371	161527949	YP_001581775.1	rhodanese domain-containing protein	281	101	200	0	1	1	1	0	1	1			B=>A
372	161527951	YP_001581777.1	4Fe-4S ferredoxin iron-sulfur binding domain-containing protein	315	97	43	1	1	1	1	0	0	0			
373	161527952	YP_001581778.1	glycosyl transferase family protein	688	136	200	0	1	1	1	0	1	1			B=>A
374	161527953	YP_001581779.1	hypothetical protein Nmar_0445	509	185	4	1	0	0	0	0	0	0			
375	161527954	YP_001581780.1	sodium/hydrogen exchanger	441	193	200	0	1	0	0	0	1	1			B=>A
376	161527957	YP_001581783.1	50S ribosomal protein L10e	169	143	200	0	1	1	1	0	1	0	M		
377	161527958	YP_001581784.1	tRNA intron endonuclease	199	104	59	1	1	1	1	1	0	0	M2		
378	161527959	YP_001581785.1	DNA repair helicase	218	360	20	0	1	0	0	0	0	1			B<=>A
379	161527960	YP_001581786.1	response regulator receiver protein	130	77	37	0	1	0	0	0	0	1			B<=>A
380	161527963	YP_001581789.1	hypothetical protein Nmar_0455	240	109	29	1	0	0	0	0	0	0			
381	161527964	YP_001581790.1	hypothetical protein Nmar_0456	2764	147	12	0	0	0	0	0	1	1			
382	161527967	YP_001581793.1	hypothetical protein Nmar_0459	211	117	25	1	0	0	0	0	0	0			
383	161527968	YP_001581794.1	histidine kinase	600	70	200	0	1	0	0	0	0	1			B=>T
384	161527970	YP_001581796.1	hypothetical protein Nmar_0462	174	159	5	1	0	0	1	0	0	0			
385	161527971	YP_001581797.1	hypothetical protein Nmar_0463	333	232	27	0	1	0	0	0	0	1			B<=>A
386	161527972	YP_001581798.1	hypothetical protein Nmar_0464	199	123	4	0	1	0	0	0	0	0			
387	161527973	YP_001581799.1	response regulator receiver protein	128	58	176	0	1	0	0	0	0	1			B=>A
388	161527975	YP_001581801.1	hypothetical protein Nmar_0467	473	72	46	1	0	0	0	0	0	0			
389	161527979	YP_001581805.1	copper amine oxidase-like protein	741	145	63	0	1	1	0	0	1	1			B<=>A
390	161527980	YP_001581806.1	ArsR family transcriptional regulator	239	104	31	1	1	0	1	0	0	0			
391	161527983	YP_001581809.1	daunorubicin resistance ABC transporter ATPase subunit	333	188	200	0	1	1	1	0	0	1			
392	161527984	YP_001581810.1	ABC-2 type transporter	261	124	200	0	1	1	1	1	1	1			B<=>A
393	161527985	YP_001581811.1	NHL repeat-containing protein	525	33	200	0	1	0	0	0	1	1			B=>A
394	161527987	YP_001581813.1	phosphate ABC transporter_periplasmic phosphate-binding protein	417	167	200	0	1	1	0	1	1	1			B=>T // B=>A
395	161527989	YP_001581815.1	phosphate ABC transporter_inner membrane subunit PstC	325	183	200	0	1	1	0	1	1	1			B=>T // B=>A
396	161527990	YP_001581816.1	phosphate ABC transporter_inner membrane subunit PstA	310	175	200	0	1	1	0	1	1	1			B=>T // B=>A
397	161527991	YP_001581817.1	phosphate ABC transporter_ATPase subunit	277	210	200	0	1	1	0	1	1	1			B=>T // B=>A
398	161527992	YP_001581818.1	phosphate uptake regulator_PhoU	219	84	34	0	1	1	1	1	0	1			B<=>A
399	161527993	YP_001581819.1	hypothetical protein Nmar_0485	359	278	66	0	0	0	0	0	1	1			B=>T
400	161527994	YP_001581820.1	hypothetical protein Nmar_0486	100	97	9	0	0	0	0	0	0	1			B=>T
401	161527995	YP_001581821.1	ArsR family transcriptional regulator	99	52	93	0	1	1	0	0	0	1			B=>A
402	161527996	YP_001581822.1	protein tyrosine phosphatase	132	61	163	0	1	0	0	0	1	1			B=>A
403	161527997	YP_001581823.1	major intrinsic protein	227	148	200	0	1	0	0	0	1	1			
404	161527998	YP_001581824.1	glutamate-1-semialdehyde-2 1-aminomutase	424	346	200	0	1	1	0	0	1	1	M2		
405	161527999	YP_001581825.1	porphobilinogen deaminase	311	193	200	0	1	1	0	0	1	1			
406	161528000	YP_001581826.1	uroporphyrin-III C-methyltransferase	246	172	200	0	1	1	0	0	1	1			B=>A
407	161528001	YP_001581827.1	uroporphyrinogen III synthase HEM4	267	105	80	0	1	0	0	0	0	1			
408	161528003	YP_001581829.1	FeS assembly protein SufB	466	401	200	0	1	1	1	0	1	1			B=>T // B<=>A
409	161528004	YP_001581830.1	FeS assembly protein SufD	466	313	200	0	1	1	1	0	1	1			A<=>B
410	161528005	YP_001581831.1	SufS subfamily cysteine desulfurase	414	330	200	0	1	1	1	0	1	1			B=>T // B=>A
411	161528006	YP_001581832.1	NifU family SUF system FeS assembly protein	147	87	84	0	1	0	1	1	0	1			
412	161528008	YP_001581834.1	phosphoglycerate kinase	408	260	200	0	1	1	1	1	1	1	M		
413	161528009	YP_001581835.1	phosphoesterase PA-phosphatase related	222	59	33	0	1	1	0	0	0	1			B=>T
414	161528013	YP_001581839.1	hypothetical protein Nmar_0505	104	97	9	1	1	1	1	0	0	0			
415	161528015	YP_001581841.1	metal dependent phosphohydrolase	206	78	135	0	1	1	0	0	1	1			B=>T // B=>EA // EA=>CA
416	161528016	YP_001581842.1	hypothetical protein Nmar_0508	103	84	4	1	0	0	0	0	0	0			
417	161528017	YP_001581843.1	delta-aminolevulinic acid dehydratase	326	266	200	0	1	1	0	0	1	1			B=>T
418	161528018	YP_001581844.1	glutamyl-tRNA reductase	421	205	200	0	1	1	0	0	1	1	M3		
419	161528019	YP_001581845.1	siroheme synthase	217	102	36	0	1	1	0	0	0	1			B=>A
420	161528020	YP_001581846.1	AsnC family transcriptional regulator	341	53	166	0	1	1	0	0	1	1			
421	161528022	YP_001581848.1	aldo/keto reductase	332	123	200	0	1	1	0	0	1	1			B<=>A
422	161528023	YP_001581849.1	chlorite dismutase	225	100	50	0	1	1	0	0	0	1			B=>A
423	161528024	YP_001581850.1	FeS assembly ATPase SufC	256	210	200	0	1	1	1	0	1	1			
424	161528025	YP_001581851.1	zinc finger TFIIB-type domain-containing protein	318	244	200	1	1	1	1	1	0	0	M2		
425	161528028	YP_001581854.1	30S ribosomal protein S8e	127	106	81	1	1	1	0	1	0	0	M		
426	161528029	YP_001581855.1	chromosome segregation and condensation protein ScpB	192	85	107	0	1	0	1	1	0	1			A<=>B
427	161528031	YP_001581857.1	alcohol dehydrogenase	350	159	200	0	1	1	1	1	1	1			
428	161528032	YP_001581858.1	hypothetical protein Nmar_0524	253	115	49	0	1	1	0	0	0	1			B=>A
429	161528034	YP_001581860.1	translation initiation factor eIF-1A	103	74	151	0	1	1	1	1	1	0	M		
430	161528035	YP_001581861.1	cold-shock DNA-binding domain-containing protein	63	59	200	0	0	0	0	0	0	1			B=>T
431	161528037	YP_001581863.1	translation initiation factor eIF-5A	136	95	200	0	1	1	1	1	1	0	M		
432	161528038	YP_001581864.1	putative ATP binding protein	229	89	192	0	1	1	1	0	1	1			A<=>B
433	161528039	YP_001581865.1	GTP-binding signal recognition particle	442	284	200	0	1	1	1	1	1	1	M2		
434	161528040	YP_001581866.1	pseudouridylate synthase-like protein	381	234	79	1	1	1	1	1	0	0	M2		
435	161528042	YP_001581868.1	ribosomal protein S27a	57	46	38	1	1	0	0	0	1	0			
436	161528043	YP_001581869.1	hypothetical protein Nmar_0535	147	131	4	0	0	0	0	0	1	0			
437	161528044	YP_001581870.1	hypothetical protein Nmar_0536	186	147	13	1	1	1	1	0	0	1			
438	161528045	YP_001581871.1	phosphoglycerate mutase	428	201	176	0	1	1	1	0	1	1			
439	161528046	YP_001581872.1	galactose-1-phosphate uridylyltransferase	338	142	150	0	1	0	0	0	1	1			B/T/EA
440	161528048	YP_001581874.1	GCN5-related N-acetyltransferase	145	74	121	0	1	0	0	0	1	1			B<=>A
441	161528049	YP_001581875.1	nucleotidyl transferase	238	145	200	0	1	1	1	0	0	1			
442	161528050	YP_001581876.1	phosphomethylpyrimidine kinase	432	139	200	0	1	1	1	1	1	1			
443	161528051	YP_001581877.1	thiamine biosynthesis protein ThiC	442	353	200	0	1	1	0	0	1	1	M2		B=>EA
444	161528052	YP_001581878.1	NUDIX hydrolase	134	108	15	0	1	0	0	0	1	1			
445	161528053	YP_001581879.1	prephenate dehydrogenase	283	99	101	0	1	1	1	1	0	1			
446	161528054	YP_001581880.1	aminotransferase class I and II	456	209	200	0	1	1	1	1	1	1			
447	161528055	YP_001581881.1	chorismate synthase	365	270	200	0	1	1	1	1	1	1			B<=>A
448	161528056	YP_001581882.1	3-phosphoshikimate 1-carboxyvinyltransferase	422	231	200	0	1	1	1	1	1	1			
449	161528057	YP_001581883.1	shikimate kinase	284	139	76	1	1	1	1	1	0	0	M2		
450	161528058	YP_001581884.1	shikimate 5-dehydrogenase	273	129	200	0	1	1	0	0	1	1			B<=>A
451	161528059	YP_001581885.1	3-dehydroquinate dehydratase_type I	223	76	156	0	1	1	1	1	1	1			
452	161528060	YP_001581886.1	3-dehydroquinate synthase	353	193	97	0	1	0	1	1	1	1	M2		A=>B
453	161528061	YP_001581887.1	phospho-2-dehydro-3-deoxyheptonate aldolase	260	140	200	0	1	1	1	1	0	1			
454	161528062	YP_001581888.1	radical SAM domain-containing protein	372	225	200	0	1	1	0	0	1	1			B<=>A
455	161528063	YP_001581889.1	UbiD family decarboxylase	480	326	200	0	1	1	0	1	0	1			
456	161528064	YP_001581890.1	histidine kinase	826	38	104	0	1	1	1	1	1	1			B=>EA
457	161528066	YP_001581892.1	hypothetical protein Nmar_0558	243	225	7	0	0	0	0	0	1	0			
458	161528067	YP_001581893.1	trans-sialidase	331	130	10	0	0	0	0	0	1	1			
459	161528068	YP_001581894.1	alkyl hydroperoxide reductase/ Thiol specific antioxidant/ Mal allergen	154	92	200	0	1	1	0	0	0	1			B=>A
460	161528069	YP_001581895.1	major intrinsic protein	237	151	200	0	1	0	0	0	1	1			
461	161528072	YP_001581898.1	acyl-CoA hydrolase	137	119	18	0	0	0	0	0	0	1			B=>T
462	161528073	YP_001581899.1	methyltransferase type 12	248	255	5	0	0	0	0	0	1	0			
463	161528076	YP_001581902.1	signal-transduction protein	133	38	200	0	1	1	1	1	1	1			
464	161528077	YP_001581903.1	hypothetical protein Nmar_0569	141	139	5	1	0	0	0	0	0	0			
465	161528078	YP_001581904.1	DNA-3-methyladenine glycosylase I	183	131	200	0	1	0	0	0	1	1			B=>T
466	161528080	YP_001581906.1	hypothetical protein Nmar_0572	421	298	37	0	1	0	0	0	0	1			B=>T // B=>EA
467	161528081	YP_001581907.1	Ferritin Dps family protein	171	98	173	0	1	1	0	0	1	1			B=>T
468	161528084	YP_001581910.1	pyridoxamine 5p-phosphate oxidase-related FMN-binding	148	93	20	0	1	0	0	0	0	1			
469	161528086	YP_001581912.1	lysyl-tRNA synthetase	529	218	135	0	1	1	1	1	0	1	M2		A=>B
470	161528087	YP_001581913.1	ELP3 family histone acetyltransferase	527	384	200	0	1	1	1	1	1	1	M2		
471	161528089	YP_001581915.1	sodium/hydrogen exchanger	416	208	99	0	1	1	0	1	1	1			
472	161528090	YP_001581916.1	phosphoribosylformylglycinamidine cyclo-ligase	340	213	200	0	1	1	0	0	1	1	M2		A=>E
473	161528091	YP_001581917.1	hypothetical protein Nmar_0583	241	85	131	1	1	1	1	1	0	0	M3		
474	161528092	YP_001581918.1	hypothetical protein Nmar_0584	551	509	12	1	0	0	0	0	0	0			
475	161528095	YP_001581921.1	nitrogen regulatory protein P-II	110	112	8	0	0	0	0	0	1	1			T<=>B
476	161528096	YP_001581922.1	ammonium transporter	435	296	200	0	1	1	0	0	1	1			
477	161528097	YP_001581923.1	hypothetical protein Nmar_0589	122	108	9	1	0	0	0	0	0	0			
478	161528098	YP_001581924.1	CaCA family Na(+)/Ca(+) antiporter	296	177	200	0	1	0	0	0	1	1			
479	161528099	YP_001581925.1	peptidase U62 modulator of DNA gyrase	452	104	200	0	1	1	1	1	0	1			
480	161528100	YP_001581926.1	radical SAM domain-containing protein	419	213	200	0	1	1	0	0	1	1			
481	161528101	YP_001581927.1	radical SAM domain-containing protein	390	126	186	0	1	1	0	0	1	1			
482	161528102	YP_001581928.1	2-alkenal reductase	379	171	200	0	1	0	1	0	1	1			B=>A
483	161528103	YP_001581929.1	glycosyl transferase family protein	402	42	200	0	1	1	1	1	0	1			
484	161528104	YP_001581930.1	cyclase/dehydrase	138	110	13	1	1	1	0	0	0	0			
485	161528105	YP_001581931.1	tRNA-binding domain-containing protein	136	93	200	0	1	1	0	1	1	1	M2		A=>B
486	161528106	YP_001581932.1	2-alkenal reductase	186	135	200	0	1	1	1	1	1	0	M		
487	161528107	YP_001581933.1	hypothetical protein Nmar_0599	373	256	33	1	1	1	1	1	0	0	M2		CA=>EA
488	161528109	YP_001581935.1	methionine aminopeptidase	297	198	200	0	1	1	1	1	1	0	M		
489	161528113	YP_001581939.1	restriction modification system DNA specificity subunit	453	82	38	0	1	1	0	0	0	1			
490	161528114	YP_001581940.1	type I restriction-modification system_M subunit	523	198	200	0	1	0	0	0	0	1			
491	161528115	YP_001581941.1	restriction modification system DNA specificity subunit	438	28	114	0	1	1	0	0	0	1			
492	161528116	YP_001581942.1	HsdR family type I site-specific deoxyribonuclease	1006	197	200	0	1	1	0	0	0	1			
493	161528117	YP_001581943.1	hypothetical protein Nmar_0609	179	46	187	0	1	1	0	0	1	1			
494	161528118	YP_001581944.1	restriction modification system DNA specificity subunit	730	159	66	0	1	0	0	0	1	1			
495	161528124	YP_001581950.1	hypothetical protein Nmar_0616	1096	140	7	0	1	0	0	0	0	1			
496	161528125	YP_001581951.1	hypothetical protein Nmar_0617	163	99	161	0	1	1	0	1	1	0	M2		
497	161528126	YP_001581952.1	solute binding protein-like protein	820	328	45	0	1	1	1	1	0	1			A<=>B
498	161528127	YP_001581953.1	hypothetical protein Nmar_0619	217	135	14	1	0	1	0	0	0	0			
499	161528128	YP_001581954.1	membrane protein-like protein	581	186	50	1	1	1	1	0	0	0	M2		
500	161528129	YP_001581955.1	hypothetical protein Nmar_0621	109	97	5	1	0	1	0	0	0	0			
501	161528130	YP_001581956.1	translation initiation factor SUI1	102	81	114	0	1	1	1	1	0	1	M		T<=>EA
502	161528131	YP_001581957.1	abortive infection protein	329	282	4	1	0	0	0	0	0	0			
503	161528132	YP_001581958.1	transcription factor TFIIB cyclin-related	302	252	200	1	1	1	1	0	0	0	M		

504	161528133	YP_001581959.1	hypothetical protein Nmar_0625	212	122	25	0	1	0	0	0	0	1			
505	161528134	YP_001581960.1	LPPG:FO 2-phospho-L-lactate transferase	319	188	84	0	1	0	1	0	0	1	M2		
506	161528135	YP_001581961.1	hypothetical protein Nmar_0627	458	395	5	1	0	0	0	0	0	0			
507	161528137	YP_001581963.1	methyltransferase type 11	207	133	25	0	1	0	0	0	1	1			
508	161528139	YP_001581965.1	geranylgeranyl reductase	402	101	200	0	1	1	1	1	1	1			
509	161528140	YP_001581966.1	hypothetical protein Nmar_0632	69	68	4	1	0	0	0	0	0	0			
510	161528141	YP_001581967.1	hypothetical protein Nmar_0633	193	112	65	0	1	1	1	0	0	1			EA=>B
511	161528142	YP_001581968.1	short-chain dehydrogenase/reductase SDR	253	129	200	0	1	1	0	1	1	1			
512	161528143	YP_001581969.1	2-hydroxy-3-oxopropionate reductase	292	152	200	0	1	1	1	0	1	1			
513	161528144	YP_001581970.1	hypothetical protein Nmar_0636	128	128	4	1	0	0	0	0	0	0			
514	161528145	YP_001581971.1	zinc finger C2H2-type domain-containing protein	122	122	4	1	0	0	0	0	0	0			
515	161528146	YP_001581972.1	hypothetical protein Nmar_0638	100	135	6	1	0	0	0	0	0	0			
516	161528147	YP_001581973.1	DSBA oxidoreductase	220	42	177	0	1	0	0	0	1	1			B=>T
517	161528148	YP_001581974.1	hypothetical protein Nmar_0640	159	134	6	1	0	0	0	0	0	0			
518	161528149	YP_001581975.1	hypothetical protein Nmar_0641	313	309	6	1	0	0	0	0	0	0			
519	161528150	YP_001581976.1	DNA-directed DNA polymerase	364	165	200	0	1	1	1	0	1	1			
520	161528151	YP_001581977.1	hypothetical protein Nmar_0643	83	80	5	1	0	0	0	0	0	0			
521	161528152	YP_001581978.1	DNA-binding protein	77	83	26	1	1	1	0	0	0	0			
522	161528154	YP_001581980.1	lysine exporter protein LysE/YggA	206	123	56	0	1	1	0	1	0	1			A<=>B
523	161528155	YP_001581981.1	hypothetical protein Nmar_0647	237	127	74	0	1	1	1	1	1	0	M3		
524	161528156	YP_001581982.1	putative deoxyhypusine synthase	313	183	200	0	1	1	1	1	1	1	M2		A=>B
525	161528157	YP_001581983.1	myo-inositol-1-phosphate synthase	364	273	107	0	1	1	1	1	0	1			
526	161528158	YP_001581984.1	Nitrilase/cyanide hydratase and apolipoprotein N-acyltransferase	245	114	200	0	1	1	1	0	1	1			
527	161528159	YP_001581985.1	SMC domain-containing protein	1174	249	200	0	1	0	1	0	1	1			A=>B
528	161528160	YP_001581986.1	oligopeptide/dipeptide ABC transporter ATPase subunit	322	231	200	0	1	1	1	1	0	1			
529	161528161	YP_001581987.1	hypothetical protein Nmar_0653	339	145	141	0	1	0	0	0	1	1			B=>T
530	161528162	YP_001581988.1	NADPH-dependent FMN reductase	188	52	200	0	1	1	0	0	1	1			
531	161528163	YP_001581989.1	DSBA oxidoreductase	219	40	198	0	1	1	0	0	1	1			B=>T
532	161528164	YP_001581990.1	histidine kinase	417	79	200	0	0	0	0	0	0	1			B=>T
533	161528165	YP_001581991.1	hypothetical protein Nmar_0657	436	76	200	0	0	0	0	0	1	1			B=>T
534	161528166	YP_001581992.1	prolyl-tRNA synthetase	480	312	200	0	1	1	0	0	1	1			B<=>A
535	161528167	YP_001581993.1	hypothetical protein Nmar_0659	94	92	5	1	0	1	0	0	0	0			
536	161528168	YP_001581994.1	hypothetical protein Nmar_0660	815	595	8	0	1	0	0	0	0	1			
537	161528169	YP_001581995.1	peptidase C26	212	58	171	0	1	1	0	0	1	1			
538	161528170	YP_001581996.1	hypothetical protein Nmar_0662	263	66	164	0	1	1	1	0	1	1			
539	161528171	YP_001581997.1	sulfatase	371	197	27	0	0	0	0	0	1	1			B=>T
540	161528174	YP_001582000.1	phosphoserine phosphatase SerB	238	138	200	0	1	0	0	0	1	1			EA<=>B
541	161528175	YP_001582001.1	F420-dependent oxidoreductase putative	241	166	109	0	1	0	0	1	0	1			B<=>A
542	161528176	YP_001582002.1	phosphotransferase domain-containing protein	304	79	101	0	1	1	0	0	0	1	M3		A=>B
543	161528177	YP_001582003.1	SirA family protein	78	49	91	0	0	1	1	0	0	1			
544	161528178	YP_001582004.1	pyridoxal-5p-phosphate-dependent protein beta subunit	330	229	200	0	1	0	1	0	1	1			B<=>A
545	161528180	YP_001582006.1	thioredoxin reductase	334	252	200	0	1	0	1	0	1	1			B=>A
546	161528181	YP_001582007.1	signal-transduction protein	165	38	200	0	1	1	1	1	1	1			
547	161528182	YP_001582008.1	UspA domain-containing protein	143	63	49	0	1	1	0	0	1	1			B<=>A
548	161528184	YP_001582010.1	lactate/malate dehydrogenase	303	207	200	0	1	1	1	0	1	1			
549	161528185	YP_001582011.1	hypothetical protein Nmar_0677	256	215	9	0	1	0	0	0	0	0			T<=>EA
550	161528186	YP_001582012.1	ribulose-1 5-biphosphate synthetase	272	213	159	0	1	1	0	0	1	1	M2		A=>B
551	161528187	YP_001582013.1	sulfite reductase (ferredoxin)	602	225	200	0	1	1	0	0	1	1			B=>T // B=>Halo
552	161528188	YP_001582014.1	rhodanese domain-containing protein	265	97	200	0	1	1	1	0	1	1			
553	161528189	YP_001582015.1	TPR repeat-containing protein	375	56	200	0	1	1	0	0	1	0			
554	161528190	YP_001582016.1	triple helix repeat-containing collagen	542	407	6	0	0	0	0	0	1	0			
555	161528192	YP_001582018.1	small GTP-binding protein	369	309	200	0	1	0	0	0	1	0			
556	161528193	YP_001582019.1	SPP-like hydrolase	231	67	124	0	1	1	0	0	1	1	M2		A=>B
557	161528194	YP_001582020.1	arginyl-tRNA synthetase	628	204	200	0	1	1	1	1	1	1	M3		A<=>B
558	161528195	YP_001582021.1	peptidyl-arginine deiminase	432	160	25	0	1	0	0	0	1	1			
559	161528197	YP_001582023.1	hypothetical protein Nmar_0689	113	103	7	1	0	0	0	0	0	0			
560	161528198	YP_001582024.1	tRNA methyltransferase complex GCD14 subunit	293	100	200	0	1	1	1	1	1	1	M2		B=>EA
561	161528199	YP_001582025.1	carbohydrate kinase YjeF related protein	287	131	200	0	1	1	1	1	1	1			A=>E
562	161528200	YP_001582026.1	peptidase M20	450	175	200	0	1	1	1	0	1	1			B=>T
563	161528201	YP_001582027.1	DoxX family protein	138	116	13	0	0	0	0	0	0	1			B=>T
564	161528202	YP_001582028.1	proteasome endopeptidase complex	198	133	200	0	1	1	1	1	1	0	M3		
565	161528203	YP_001582029.1	PUA domain-containing protein	158	77	118	0	1	1	1	1	1	0	M		
566	161528204	YP_001582030.1	homoserine dehydrogenase	336	190	200	0	1	1	1	1	1	1	M2		A=>B
567	161528205	YP_001582031.1	methionine-R-sulfoxide reductase	135	87	200	0	1	0	0	0	1	1			B=>T // B=>EA
568	161528206	YP_001582032.1	thymidylate synthase complementing protein ThyX	545	329	22	0	1	0	0	0	0	1			B<=>A
569	161528208	YP_001582034.1	hypothetical protein Nmar_0700	292	199	16	1	0	1	0	0	0	0			
570	161528209	YP_001582035.1	acetate--CoA ligase	638	511	200	0	1	1	1	0	1	1			A<=>B
571	161528210	YP_001582036.1	putative PAS/PAC sensor protein	174	155	4	0	0	0	0	0		1			
572	161528211	YP_001582037.1	flap endonuclease-1	340	221	200	0	1	1	1	1	1	0	M		
573	161528213	YP_001582039.1	transcriptional regulator TrmB	283	76	36	1	1	1	1	0	0	0			
574	161528215	YP_001582041.1	response regulator receiver protein	216	56	200	0	1	0	0	0	1	1			B=>T // B=>EA
575	161528217	YP_001582043.1	nitroreductase	215	81	77	0	1	1	1	1	1	1			B<=>A
576	161528218	YP_001582044.1	AsnC/Lrp family regulatory protein	77	77	24	1	0	1	1	0	0	0			
577	161528220	YP_001582046.1	hypothetical protein Nmar_0712	113	113	9	1	0	0	0	0	0	0			
578	161528222	YP_001582048.1	hypothetical protein Nmar_0714	426	427	9	1	0	0	0	0	0	0			
579	161528223	YP_001582049.1	Mov34/MPN/PAD-1 family protein	133	93	41	1	1	0	1	0	0	0			
580	161528225	YP_001582051.1	hypothetical protein Nmar_0717	103	103	8	1	0	0	0	0	0	0			
581	161528226	YP_001582052.1	binding-protein-dependent transport systems inner membrane component	455	180	200	0	1	1	1	1		1			B<=>A
582	161528227	YP_001582053.1	binding-protein-dependent transport systems inner membrane component	349	175	200	0	1	1	1	1	1	1			B<=>A
583	161528230	YP_001582056.1	arsenate reductase and related	90	104	29	0	0	0	0	0	0	1			B=>T
584	161528232	YP_001582058.1	hypothetical protein Nmar_0724	219	207	8	1	0	1	0	1	0	0			
585	161528233	YP_001582059.1	hypothetical protein Nmar_0725	226	201	40	0	1	0	0	0	0	1			EA=>T // EA=>B
586	161528234	YP_001582060.1	SNF2-related protein	574	215	200	0	1	0	0	0	1	1			
587	161528237	YP_001582063.1	superoxide dismutase	207	148	200	0	1	0	0	0	1	1			B=>EA
588	161528238	YP_001582064.1	hypothetical protein Nmar_0730	432	458	5	1	0	0	0	0	0	0			
589	161528241	YP_001582067.1	hypothetical protein Nmar_0733	47	45	6	0	0	0	0	0	0	1			
590	161528242	YP_001582068.1	FAD linked oxidase domain-containing protein	481	200	200	0	1	1	1	0	1	1			
591	161528245	YP_001582071.1	heat shock protein DnaJ domain-containing protein	182	114	200	0	0	0	0	0	1	1			
592	161528246	YP_001582072.1	AsnC family transcriptional regulator	147	69	200	0	1	1	0	1	0	1			A=>B
593	161528247	YP_001582073.1	TPR repeat-containing protein	75	88	42	0	1	0	0	0	1	1			
594	161528248	YP_001582074.1	TPR repeat-containing protein	158	45	200	0	1	0	0	0	1	1			
595	161528249	YP_001582075.1	hypothetical protein Nmar_0741	57	48	5	1	0	0	0	0	0	0			
596	161528250	YP_001582076.1	radical SAM domain-containing protein	365	226	200	0	1	1	0	0	1	1			
597	161528251	YP_001582077.1	hypothetical protein Nmar_0743	101	76	56	0	1	1	0	1	1	1			
598	161528254	YP_001582080.1	hypothetical protein Nmar_0746	275	160	70	0	1	1	0	0	1	1			B<=>A
599	161528255	YP_001582081.1	hypothetical protein Nmar_0747	317	262	9	1	0	0	0	0	0	0			
600	161528256	YP_001582082.1	elongation factor EF-2	730	592	200	0	1	1	1	1	1	0	M		
601	161528263	YP_001582089.1	O-methyltransferase family protein	205	88	200	0	1	0	0	0	1	1			B=>T+EA
602	161528264	YP_001582090.1	hypothetical protein Nmar_0756	125	123	8	1	0	0	0	0	0	0			
603	161528271	YP_001582097.1	hypothetical protein Nmar_0763	270	272	5	1	0	0	0	0	0	0			
604	161528272	YP_001582098.1	CBS domain-containing protein	605	21	138	0	1	1	1	1	1	1			
605	161528273	YP_001582099.1	hypothetical protein Nmar_0765	147	139	6	1	0	0	0	0	0	0			
606	161528275	YP_001582101.1	hypothetical protein Nmar_0767	130	344	16	1	0	0	0	0	0	0			
607	161528278	YP_001582104.1	5-formaminoimidazole-4-carboxamide-1-(beta)-D-ribofuranosyl 5p-monophosphate synthetase-like protein	363	261	110	1	1	1	1	1	0	0	M2		
608	161528279	YP_001582105.1	3-octaprenyl-4-hydroxybenzoate carboxy-lyase	185	142	200	0	1	1	0	0	1	1			B=>T // B=>E // B=>EA
609	161528282	YP_001582108.1	Sec-independent protein translocase TatC subunit	263	141	200	0	1	1	1	0	1	1			
610	161528283	YP_001582109.1	hypothetical protein Nmar_0775	154	154	5	1	0	0	0	0	0	0			
611	161528285	YP_001582111.1	phosphoribosylformylglycinamidine synthase purS	86	75	13	1	1	1	0	0	0	0			
612	161528286	YP_001582112.1	signal transduction protein	140	36	200	0	1	1	1	0	1	1			
613	161528287	YP_001582113.1	phosphoribosylformylglycinamidine synthase I	228	149	200	0	1	1	0	0	0	1	M2		
614	161528288	YP_001582114.1	phosphoribosylformylglycinamidine synthase II	721	483	200	0	1	1	0	0	0	1	M2		
615	161528289	YP_001582115.1	glutamine amidotransferase class-II	489	314	200	0	1	1	0	0	1	1	M3		EA<=>EA // EA=>B
616	161528290	YP_001582116.1	hypothetical protein Nmar_0782	92	61	113	0	1	0	0	0	1	1			B=>T
617	161528292	YP_001582118.1	phosphoribosylaminoimidazolesuccinocarboxamide synthase	274	157	200	0	1	1	0	0	1	1			B=>A // B=>E
618	161528293	YP_001582119.1	ArsR family transcriptional regulator	167	164	8	1	1	1	0	0	0	0			
619	161528295	YP_001582121.1	pyridoxamine 5p-phosphate oxidase-related FMN-binding	134	98	25	0	1	0	1	0	0	1			
620	161528302	YP_001582128.1	ribosomal protein L6	185	108	200	0	1	1	1	0	1	0	M		
621	161528303	YP_001582129.1	30S ribosomal protein S8P	130	123	200	0	1	1	1	1	1	0	M		
622	161528304	YP_001582130.1	ribosomal protein S14	59	43	80	0	1	1	1	0	1	0	M		
623	161528305	YP_001582131.1	50S ribosomal protein L5P	173	163	200	0	1	1	1	1	1	0	M		
624	161528306	YP_001582132.1	30S ribosomal protein S4e	238	195	200	0	1	1	1	1	1	0	M		
625	161528307	YP_001582133.1	KOW domain-containing protein	168	111	200	0	1	1	1	1	1	0	M		
626	161528308	YP_001582134.1	ribosomal protein L14b/L23e	140	132	200	0	1	1	1	1	1	0	M		
627	161528309	YP_001582135.1	ribosomal protein S17	107	103	200	0	1	1	1	1	1	0	M		
628	161528310	YP_001582136.1	ribonuclease P Pop29	88	80	25	0	1	0	0	0	0	1			A=>B
629	161528312	YP_001582138.1	ribosomal protein S3	251	149	200	0	1	1	1	1	1	0	M		
630	161528313	YP_001582139.1	ribosomal protein L22	152	137	200	0	1	1	1	0	1	0	M		
631	161528314	YP_001582140.1	30S ribosomal protein S19P	131	118	200	0	1	1	1	1	1	0	M		
632	161528315	YP_001582141.1	ribosomal protein L25/L23	88	78	78	0	1	1	1	1	1	0	M		
633	161528316	YP_001582142.1	50S ribosomal protein L4P	271	191	200	0	1	1	1	1	1	0	M		
634	161528317	YP_001582143.1	50S ribosomal protein L3P	331	255	200	0	1	1	1	1	1	0	M		
635	161528318	YP_001582144.1	hypothetical protein Nmar_0810	270	138	108	0	1	1	1	0	1	0	M		
636	161528319	YP_001582145.1	proteasome subunit alpha	245	178	200	0	1	1	1	1	1	0	M		
637	161528320	YP_001582146.1	hypothetical protein Nmar_0812	138	90	71	1	1	1	0	1	0	0	M		
638	161528323	YP_001582149.1	blue (type1) copper domain-containing protein	539	42	50	0	1	0	1	0	0	1			A<=>B
639	161528324	YP_001582150.1	SNF7	216	123	43	1	0	1	1	0	0	0	M		
640	161528325	YP_001582151.1	redoxin domain-containing protein	369	66	94	0	1	0	0	0	1	1			
641	161528326	YP_001582152.1	cytochrome c biogenesis protein transmembrane region	244	131	111	0	1	0	0	0	0	1			B=>T // B=>EA
642	161528327	YP_001582153.1	8-amino-7-oxononanoate synthase	374	241	200	0	1	1	0	0	1	1			B=>T // B=>EA
643	161528329	YP_001582155.1	biotin synthase	326	222	200	0	1	0	0	0	1	1			B=>T // B=>EA
644	161528330	YP_001582156.1	aminotransferase class-III	426	220	200	0	1	0	0	0	1	1			B=>T // B=>EA
645	161528331	YP_001582157.1	dethiobiotin synthase	228	90	200	0	1	0	0	0	1	1			B=>T // B=>EA
646	161528332	YP_001582158.1	hypothetical protein Nmar_0824	75	74	7	1	0	0	0	0	0	0			
647	161528333	YP_001582159.1	band 7 protein	287	223	200	0	0	0	0	0	1	1			
648	161528336	YP_001582162.1	hydroxymethylglutaryl-CoA synthase	464	175	192	0	1	1	0	0	1	0	M2		
649	161528337	YP_001582163.1	DSBA oxidoreductase	236	43	191	0	1	0	0	0	1	1			B=>T // B=>EA
650	161528339	YP_001582165.1	glyceraldehyde-3-phosphate dehydrogenase	350	295	94	0	1	1	1	0	0	1	M		A=>B
651	161528340	YP_001582166.1	hypothetical protein Nmar_0832	417	153	200	0	1	0	0	0	1	1			B=>T // B=>EA
652	161528341	YP_001582167.1	threonine synthase	445	263	200	0	1	1	1	1	1	1	M2		A<=>B
653	161528349	YP_001582175.1	Propanoyl-CoA C-acyltransferase	371	229	200	0	1	1	1	0	1	1	M3		
654	161528350	YP_001582176.1	luciferase family protein	321	66	200	0	1	0	1	0	0	1			A<=>B
655	161528351	YP_001582177.1	nucleotidyl transferase	222	144	200	0	1	1	0	0	1	1			
656	161528354	YP_001582180.1	phosphoribosylaminoimidazole carboxylase ATPase subunit	380	187	200	0	1	1	0	0	1	1			
657	161528355	YP_001582181.1	phosphoribosylaminoimidazole carboxylase catalytic subunit	191	123	200	0	1	1	0	0	1	1			
658	161528356	YP_001582182.1	threonine dehydratase	402	214	200	0	1	1	0	0	1	1			B=>T
659	161528358	YP_001582184.1	adenylate/guanylate cyclase	278	87	53	0	0	0	0	0	1	1			B=>T
660	161528361	YP_001582187.1	glyoxalase/bleomycin resistance protein/dioxygenase	136	98	12	0	0	0	0	0	1	1			
661	161528366	YP_001582192.1	ATP--cobalamin adenosyltransferase	178	111	200	0	1	1	0	0	1	1			B=>T // B=>EA
662	161528367	YP_001582193.1	metal dependent phosphohydrolase	177	70	159	0	1	1	1	0	1	1			A<=>B
663	161528370	YP_001582196.1	hypothetical protein Nmar_0862	115	106	5	0	0	0	1	0	0	0			
664	161528371	YP_001582197.1	LOR/SDH bifunctional protein	407	355	69	0	1	0	0	0	0	1			EA=>T
665	161528372	YP_001582198.1	hypothetical protein Nmar_0864	170	83	7	1	0	0	0	0	0	0			
666	161528373	YP_001582199.1	aminotransferase class I and II	452	237	200	0	1	1	0	1	1	1			A<=>B
667	161528374	YP_001582200.1	AMP-dependent synthetase and ligase	641	480	200	0	1	1	1	0	1	1			A<=>B
668	161528376	YP_001582202.1	glutamyl-tRNA(Gln) amidotransferase B subunit	467	335	200	0	1	1	0	0	1	1	M2		
669	161528377	YP_001582203.1	glutamyl-tRNA(Gln) amidotransferase A subunit	481	366	200	0	1	1	0	0	1	1	M3		
670	161528378	YP_001582204.1	hypothetical protein Nmar_0870	88	72	4	1	0	0	0	0	0	0			
671	161528379	YP_001582205.1	aspartyl-tRNA synthetase	436	271	200	0	1	1	1	1	1	1	M3		
672	161528380	YP_001582206.1	5-formaminoimidazole-4-carboxamide-1-(beta)-D-ribofuranosyl 5p-monophosphate synthetase	341	261	110	1	1	1	1	1	0	0	M2		

673	161528381	YP_001582207.1	phosphonate ABC transporter inner membrane subunit	265	131	200	0	1	0	1	0	1	1		B=>T // B=>EA
674	161528382	YP_001582208.1	ABC transporter related	276	111	200	0	1	0	1	0	1	1		B=>T // B=>EA
675	161528383	YP_001582209.1	phosphonate ABC transporter periplasmic phosphonate-binding protein	362	37	89	0	0	1	1	0	0	1		B=>T
676	161528384	YP_001582210.1	hypothetical protein Nmar_0876	116	106	7	1	0	0	0	0	0	0		
677	161528385	YP_001582211.1	putative transcriptional regulator	110	82	9	1	0	0	0	0	0	0		
678	161528386	YP_001582212.1	hypothetical protein Nmar_0878	242	191	20	0	0	0	0	0	0	1		B<=>T
679	161528388	YP_001582214.1	hypothetical protein Nmar_0880	209	175	13	1	1	1	1	0	0	0		
680	161528389	YP_001582215.1	DSBA oxidoreductase	214	49	200	0	1	0	0	0	1	1		B=>T // B=>EA
681	161528390	YP_001582216.1	peptide chain release factor 1	426	322	200	0	1	1	1	1	1	0	M	
682	161528391	YP_001582217.1	histidine triad (HIT) protein	134	49	200	0	1	1	1	0	1	1		
683	161528392	YP_001582218.1	hypothetical protein Nmar_0884	199	159	4	1	0	0	0	0	0	0		
684	161528394	YP_001582220.1	dihydroxy-acid dehydratase	558	476	200	0	1	1	1	1	1	1	M3	A<=>B
685	161528396	YP_001582222.1	helicase c2	466	359	28	0	1	0	0	0	0	1		A<=>B
686	161528411	YP_001582237.1	hypothetical protein Nmar_0903	138	134	4	1	0	0	0	0	0	0		
687	161528414	YP_001582240.1	hypothetical protein Nmar_0906	223	190	8	0	0	0	0	0	1	1		
688	161528419	YP_001582245.1	anthranilate synthase	449	278	200	0	1	1	0	0	1	1		A=>B
689	161528420	YP_001582246.1	glutamine amidotransferase of anthranilate synthase	198	149	200	0	1	1	0	0	1	1		EA<=>T
690	161528421	YP_001582247.1	anthranilate phosphoribosyltransferase	348	211	200	0	1	1	0	0	1	1		
691	161528422	YP_001582248.1	indole-3-glycerol-phosphate synthase	258	163	200	0	1	1	0	0	1	1		
692	161528423	YP_001582249.1	tryptophan synthase subunit beta	395	353	200	0	1	0	0	0	1	1		EA=>T
693	161528424	YP_001582250.1	tryptophan synthase alpha subunit	268	149	200	0	1	0	0	0	1	1		B=>T
694	161528426	YP_001582252.1	hypothetical protein Nmar_0918	280	75	32	0	1	0	0	0	0	1		A=>B
695	161528427	YP_001582253.1	CTP synthetase	533	427	200	0	1	1	1	0	1	1	M2	A=>E
696	161528428	YP_001582254.1	histidine kinase	505	99	200	0	1	0	0	0	0	1		B<=>A
697	161528429	YP_001582255.1	NAD(+) kinase	332	99	200	0	1	1	1	1	1	1		
698	161528430	YP_001582256.1	ribokinase-like domain-containing protein	289	156	19	1	1	1	1	1	0	0		
699	161528431	YP_001582257.1	30S ribosomal protein S26e	95	97	182	0	0	1	1	1	1	0	M2	
700	161528432	YP_001582258.1	CDP-alcohol phosphatidyltransferase	190	109	66	0	1	1	1	1	0	1		A=>B
701	161528433	YP_001582259.1	putative agmatinase	291	124	200	0	1	1	1	0	0	1	M2	A=>B
702	161528434	YP_001582260.1	threonyl-tRNA synthetase	621	310	200	0	1	1	1	0	1	1	M2	EA=>T
703	161528436	YP_001582262.1	deoxyhypusine synthase-like protein	354	148	200	0	1	1	1	1	1	0	M2	
704	161528440	YP_001582266.1	RNA polymerase Rpb6	182	59	155	0	1	1	1	0	1	1	M2	
705	161528441	YP_001582267.1	Alba DNA/RNA-binding protein	104	87	5	1	1	0	0	0	0	0		
706	161528442	YP_001582268.1	hypothetical protein Nmar_0934	138	74	20	1	1	1	1	1	0	0		
707	161528443	YP_001582269.1	asparagine synthase	339	242	200	0	1	1	0	0	1	1		
708	161528444	YP_001582270.1	hypothetical protein Nmar_0936	121	75	20	1	0	0	0	0	0	0		
709	161528445	YP_001582271.1	aldo/keto reductase	366	167	43	0	1	0	0	0	1	1		
710	161528446	YP_001582272.1	UbiA prenyltransferase	296	148	117	0	1	1	1	0	1	1		
711	161528447	YP_001582273.1	hypothetical protein Nmar_0939	133	70	17	0	0	1	0	0	0	1		
712	161528450	YP_001582276.1	PpiC-type peptidyl-prolyl cis-trans isomerase	92	64	200	0	1	0	0	0	1	1		B=>T
713	161528452	YP_001582278.1	DEAD/DEAH box helicase domain-containing protein	502	136	200	0	1	1	1	0	1	0	M3	
714	161528455	YP_001582281.1	CMP/dCMP deaminase zinc-binding	200	78	196	0	1	0	0	0	1	1		
715	161528456	YP_001582282.1	DNA polymerase I	852	326	200	0	1	1	1	1	1	1	M2	A=>B
716	161528457	YP_001582283.1	hypothetical protein Nmar_0949	93	82	6	0	1	0	0	0	0	1		T/EA/B
717	161528458	YP_001582284.1	triosephosphate isomerase	218	164	99	0	1	1	1	1	0	1	M2	A=>B
718	161528459	YP_001582285.1	pyruvate phosphate dikinase	887	628	200	0	1	1	1	1	1	1		
719	161528460	YP_001582286.1	XRE family transcriptional regulator	185	139	51	1	1	1	0	1	0	0	M2	EA=>CA
720	161528461	YP_001582287.1	glyoxalase/bleomycin resistance protein/dioxygenase	131	84	163	0	1	1	0	0	1	1		
721	161528462	YP_001582288.1	methylmalonyl-CoA mutase large subunit	532	466	200	0	1	1	0	0	1	0		
722	161528463	YP_001582289.1	LAO/AO transport system ATPase	305	144	200	0	1	1	0	0	1	1		
723	161528464	YP_001582290.1	hypothetical protein Nmar_0956	135	129	4	1	0	0	0	0	0	0		
724	161528465	YP_001582291.1	peptidase M1 membrane alanine aminopeptidase	830	295	200	0	1	1	0	0	1	1		
725	161528466	YP_001582292.1	cobalamin B12-binding domain-containing protein	140	101	200	0	1	1	0	0	1	1		EA=>T
726	161528467	YP_001582293.1	ketol-acid reductoisomerase	332	300	200	0	1	1	1	0	0	1	M3	
727	161528468	YP_001582294.1	cytidyltransferase-like protein	176	116	120	0	1	1	1	1	0	1	M3	
728	161528469	YP_001582295.1	elongation factor Tu domain-containing protein	306	141	123	0	1	1	0	0	1	1		
729	161528470	YP_001582296.1	hypothetical protein Nmar_0962	250	240	5	1	1	0	0	0	0	0		
730	161528471	YP_001582297.1	methionyl-tRNA synthetase	551	296	200	0	1	1	1	1	1	1		A<=>B // A<=>E
731	161528472	YP_001582298.1	adenosylhomocysteinase	413	337	200	0	1	1	1	1	1	1	M3	A=>B
732	161528473	YP_001582299.1	Rieske (2Fe-2S) domain-containing protein	100	41	68	0	1	0	0	0	0	1		B=>A
733	161528474	YP_001582300.1	NAD+ synthetase	237	135	200	0	1	1	1	1	0	1	M3	A=>B
734	161528475	YP_001582301.1	cysteinyl-tRNA synthetase	462	330	200	0	1	1	1	1	1	1	M3	A=>B
735	161528477	YP_001582303.1	hypothetical protein Nmar_0969	109	99	8	0	1	0	0	0	0	1		
736	161528478	YP_001582304.1	hypothetical protein Nmar_0970	93	34	4	1	0	0	0	0	0	0		
737	161528479	YP_001582305.1	MscS mechanosensitive ion channel	272	105	200	0	1	1	0	0	0	1		
738	161528480	YP_001582306.1	hypothetical protein Nmar_0972	258	179	14	1	1	0	0	0	0	0		EA=>T
739	161528481	YP_001582307.1	hypothetical protein Nmar_0973	299	73	48	1	0	0	0	0	0	0		
740	161528482	YP_001582308.1	hypothetical protein Nmar_0974	175	203	53	0	1	0	0	0	0	1		B=>A
741	161528483	YP_001582309.1	hypothetical protein Nmar_0975	385	206	118	1	1	1	1	1	0	0	M2	
742	161528484	YP_001582310.1	periplasmic binding protein	308	95	110	0	1	1	0	0	1	1		
743	161528485	YP_001582311.1	NAD-dependent epimerase/dehydratase	459	63	199	0	1	0	0	0	1	1		
744	161528486	YP_001582312.1	response regulator receiver protein	280	74	33	0	1	0	0	0	0	1		B=>A
745	161528487	YP_001582313.1	transcription factor TFIIB cyclin-related	303	242	200	0	1	1	1	1	1	0	M2	
746	161528488	YP_001582314.1	translation-associated GTPase	403	181	200	0	1	1	1	1	1	0	M2	
747	161528489	YP_001582315.1	hypothetical protein Nmar_0981	121	79	9	1	0	1	0	0	0	0		
748	161528490	YP_001582316.1	CoA-binding domain-containing protein	136	69	182	0	1	1	1	1	1	1		
749	161528491	YP_001582317.1	hypothetical protein Nmar_0983	173	165	5	1	0	0		0	0	0		
750	161528492	YP_001582318.1	biotin--acetyl-CoA-carboxylase ligase	335	79	200	0	1	1	0	0	0	1		B<=>A
751	161528494	YP_001582320.1	TPR repeat-containing protein	479	17	86	0	1	0	0	0	1	1		B<=>A
752	161528495	YP_001582321.1	transcription factor TFIIB cyclin-related	315	249	200	0	1	1	1	1	1	0	M2	
753	161528496	YP_001582322.1	phosphoribosyltransferase-like protein	135	110	10	0	0	1	0	0	0	1		
754	161528497	YP_001582323.1	hypothetical protein Nmar_0989	130	92	14	1	0	0	0	0	0	0		
755	161528499	YP_001582325.1	poly(R)-hydroxyalkanoic acid synthase class III PhaC subunit	354	177	176	0	1	0	0	0	1	1		B=>A
756	161528500	YP_001582326.1	hypothetical protein Nmar_0992	142	134	4	1	0	0	0	0	0	0		
757	161528501	YP_001582327.1	AbrB family transcriptional regulator	141	84	12	1	1	0	0	0	0	0		
758	161528502	YP_001582328.1	alpha/beta hydrolase fold	281	48	200	0	1	1	1	0	1	1		B=>A
759	161528503	YP_001582329.1	peptide methionine sulfoxide reductase	150	107	200	0	1	1	0	0	1	1		
760	161528504	YP_001582330.1	hypothetical protein Nmar_0996	132	114	10	1	0	0	0	0	0	0		
761	161528505	YP_001582331.1	methyltransferase type 11	282	106	23	0	0	0	0	0	1	1		B<=>T
762	161528509	YP_001582335.1	AsnC family transcriptional regulator	83	75	13	1	1	1	0	0	0	0		
763	161528510	YP_001582336.1	hypothetical protein Nmar_1002	130	107	9	1	0	0	0	0	0	0		
764	161528511	YP_001582337.1	N-formylglutamate amidohydrolase	259	139	77	1	1	0	0	0	1	1		B=>A
765	161528512	YP_001582338.1	gamma-glutamyltransferase	541	264	200	0	1	0	0	0	1	1		B=>A
766	161528513	YP_001582339.1	Ion transport protein	267	114	200	0	1	0	0	0	1	1		B=>A
767	161528514	YP_001582340.1	magnesium and cobalt transport protein CorA	547	117	200	0	1	0	0	0	1	1		B=>A
768	161528515	YP_001582341.1	hypothetical protein Nmar_1007	422	175	200	0	1	0	0	0	1	1		B=>A
769	161528516	YP_001582342.1	D-alanine--D-alanine ligase	743	129	200	0	1	0	0	0	1	1		B=>A
770	161528519	YP_001582345.1	nitroreductase	329	68	164	0	1	0	0	1	1	1		B=>A
771	161528522	YP_001582348.1	transcriptional regulator protein-like protein	266	248	7	1	1	0	0	0	0	0		
772	161528523	YP_001582349.1	proteinase inhibitor I1 Kazal	343	66	16	0	0	0	0	0	0	1		B=>A
773	161528525	YP_001582351.1	aminotransferase class-III	435	310	200	0	1	1	1	0	1	1	M3	
774	161528527	YP_001582353.1	TPR repeat-containing protein	273	58	200	0	1	0	0	0	1	1		EA=>T
775	161528528	YP_001582354.1	hypothetical protein Nmar_1020	234	231	13	1	0	0	0	0	0	0		
776	161528530	YP_001582356.1	cupin 2 domain-containing protein	151	109	20	0	1	0	0	0	0	1		B=>T
777	161528531	YP_001582357.1	cupin 2 domain-containing protein	120	81	33	0	1	1	0	0	0	1		A<=>B
778	161528532	YP_001582358.1	heat shock protein DnaJ domain-containing protein	223	56	24	0	0	0	1	0	1	1		
779	161528533	YP_001582359.1	histidine triad (HIT) protein	135	58	200	0	1	1	0	0	1	1		
780	161528534	YP_001582360.1	hypothetical protein Nmar_1026	101	92	8	1	0	0	0	0	0	0		
781	161528535	YP_001582361.1	hypothetical protein Nmar_1027	55	53	6	1	0	0	0	0	0	0		
782	161528536	YP_001582362.1	3-hydroxybutyryl-CoA dehydrogenase	379	184	200	0	1	1	1	1	1	1		B=>A
783	161528538	YP_001582364.1	Rossmann fold nucleotide-binding protein	169	87	80	0	1	1	1	0	1	1		A<=>B
784	161528539	YP_001582365.1	hypothetical protein Nmar_1031	136	136	7	1	0	0	0	0	0	0		
785	161528540	YP_001582366.1	RNA polymerase Rpb10	67	69	5	1	0	0	0	0	0	0		
786	161528541	YP_001582367.1	ribosomal protein S10	102	98	200	0	1	1	1	0	1	0	M	
787	161528542	YP_001582368.1	elongation factor 1-alpha	432	401	200	0	1	1	1	1	1	0	M	
788	161528543	YP_001582369.1	hypothetical protein Nmar_1035	381	346	85	0	1	1	1	1	0	1	M2	A=>B
789	161528544	YP_001582370.1	hypothetical protein Nmar_1036	135	135	6	1	0	0	0	0	0	0		
790	161528545	YP_001582371.1	DNA ligase I ATP-dependent Dnl1	588	292	200	0	1	1	1	1	1	1	M2	A=>B
791	161528546	YP_001582372.1	hypothetical protein Nmar_1038	217	213	10	1	0	1	1	0	0	0		
792	161528547	YP_001582373.1	tRNA intron endonuclease	178	68	93	0	1	1	1	1	1	0	M2	
793	161528548	YP_001582374.1	fibronectin type III domain-containing protein	771	0	1									
794	161528549	YP_001582375.1	hypothetical protein Nmar_1041	125	123	6	1	0	0	0	0	0	0		
795	161528550	YP_001582376.1	hypothetical protein Nmar_1042	78	78	6	1	0	0	0	0	0	0		
796	161528551	YP_001582377.1	short-chain dehydrogenase/reductase SDR	588	92	200	0	1	1	0	0	1	1		
797	161528553	YP_001582379.1	response regulator receiver protein	125	62	33	0	0	0	0	0	1	1		B=>T
798	161528554	YP_001582380.1	putative PAS/PAC sensor protein	530	44	200	0	1	0	0	0	1	1		B=>T
799	161528555	YP_001582381.1	fibronectin type III domain-containing protein	903	0	1									
800	161528556	YP_001582382.1	hypothetical protein Nmar_1048	257	247	8	0	1	0	0	0	0	1		
801	161528558	YP_001582384.1	hypothetical protein Nmar_1050	163	63	81	0	1	0	1	0	1	1		B=>A
802	161528559	YP_001582385.1	hypothetical protein Nmar_1051	247	159	4	1	0	0	0	0	0	0		
803	161528560	YP_001582386.1	hypothetical protein Nmar_1052	118	116	5	1	0	0	0	0	0	0		
804	161528561	YP_001582387.1	DtxR family iron dependent repressor	154	59	170	0	1	1	1	1	0	1		
805	161528562	YP_001582388.1	ATPase	605	374	101	0	1	1	1	1	1	1	M2	A=>B
806	161528563	YP_001582389.1	peptidase U62 modulator of DNA gyrase	473	195	200	0	1	1	1	1	0	1		
807	161528564	YP_001582390.1	phage SPO1 DNA polymerase-related protein	196	112	200	0	1	1	1	1	0	1		
808	161528565	YP_001582391.1	DNA-3-methyladenine glycosylase	187	91	150	0	1	1	1	1	1	1		
809	161528566	YP_001582392.1	SNARE associated Golgi protein	210	160	16	1	1	1	1	0	0	0		
810	161528569	YP_001582395.1	hypothetical protein Nmar_1061	276	163	5	1	0	0	0	0	0	0		
811	161528572	YP_001582398.1	oligopeptide/dipeptide ABC transporter ATPase subunit	311	230	200	0	1	1	1	1	0	1		B=>A
812	161528573	YP_001582399.1	transcription elongation factor NusA	172	142	24	1	1	1	1	0	0	0	M3	
813	161528574	YP_001582400.1	aspartyl-tRNA synthetase	428	277	200	0	1	1	1	1	1	1	M	A=>B
814	161528577	YP_001582403.1	isopropylmalate/isohomocitrate dehydrogenase	337	218	200	0	1	1	1	1	1	1		A=>B
815	161528578	YP_001582404.1	isopropylmalate/citramalate/homocitrate synthase	505	356	200	0	1	1	1	1	1	1	M2	
816	161528579	YP_001582405.1	acetolactate synthase small subunit	161	136	200	0	1	0	0	0	1	1		B=A
817	161528580	YP_001582406.1	acetolactate synthase large subunit biosynthetic type	566	408	200	0	1	1	1	0	1	1	M3	
818	161528581	YP_001582407.1	fibronectin type III domain-containing protein	9585	3	200	0	1	1	0	0	1	1		
819	161528583	YP_001582409.1	chromosome segregation ATPase-like protein	1206	6	34	0	0	0	0	0	1	0		
820	161528584	YP_001582410.1	sensor protein	630	182	30	0	1	0	0	0	0	1		B=>T
821	161528585	YP_001582411.1	hypothetical protein Nmar_1077	134	131	5	1	0	0	1	0	0	0		
822	161528586	YP_001582412.1	fructose-bisphosphate aldolase	331	148	180	0	1	1	1	1	0	1		A=>B
823	161528587	YP_001582413.1	alcohol dehydrogenase	342	148	200	0	1	1	1	0	1	1		B=>A
824	161528588	YP_001582414.1	vacuolar H+transporting two-sector ATPase F subunit	97	95	4	1	0	1	1	0	0	0		
825	161528589	YP_001582415.1	hypothetical protein Nmar_1081	100	98	9	1	0	0	0	0	0	0		
826	161528591	YP_001582417.1	hypothetical protein Nmar_1083	100	96	6	1	0	0	1	0	0	0		
827	161528592	YP_001582418.1	metallophosphoesterase	249	97	70	1	1	1	1	1	0	0	M2	
828	161528593	YP_001582419.1	carbamoyl-phosphate synthase large subunit	1080	780	200	0	1	1	0	0	1	1		A<=>B
829	161528594	YP_001582420.1	carbamoyl-phosphate synthase small subunit	384	232	200	0	1	1	0	0	1	1		A<=>B
830	161528595	YP_001582421.1	hypothetical protein Nmar_1087	518	75	48	1	0	0	0	0	0	0		
831	161528596	YP_001582422.1	ATPase central domain-containing protein	397	201	200	0	1	1	1	0	1	0	M3	
832	161528599	YP_001582425.1	transcriptional regulator TrmB	386	60	73	0	1	1	1	0	0	1		A=>B
833	161528600	YP_001582426.1	endoribonuclease L-PSP	151	102	168	0	1	0	0	0	1	1		B=>A
834	161528601	YP_001582427.1	sodium:dicarboxylate symporter	425	249	200	0	1	1	0	0	1	1		B=>A
835	161528602	YP_001582428.1	pyridoxamine 5p-phosphate oxidase family protein	130	95	16	0	0	0	0	0	0	1		B=>T
836	161528604	YP_001582430.1	hypothetical protein Nmar_1096	375	148	25	0	1	0	0	0	0	0		
837	161528605	YP_001582431.1	DoxX family protein	148	136	12	0	0	0	0	0	0	1		B=>T
838	161528606	YP_001582432.1	phosphoesterase DHHA1	380	263	25	0	1	1	0	0	1	0	M3	
839	161528607	YP_001582433.1	2-methylcitrate synthase/citrate synthase II	387	236	200	0	1	1	1	0	1	1		
840	161528608	YP_001582434.1	hypothetical protein Nmar_1100	68	67	4	1	0	0	0	0	0	0		
841	161528610	YP_001582436.1	blue (type1) copper domain-containing protein	318	40	61	0	1	1	1	0	0	1		A<=>B

842	161528611	YP_001582437.1	excinuclease ABC subunit C	528	312	200	0	1	0	0	0	0	1		B=>A
843	161528612	YP_001582438.1	excinuclease ABC_A subunit	939	808	200	0	1	0	0	0	0	1		B=>A
844	161528613	YP_001582439.1	excinuclease ABC subunit B	650	554	200	0	1	0	0	0	0	1		B=>A
845	161528614	YP_001582440.1	pyridoxamine 5p-phosphate oxidase-related FMN-binding	216	63	52	0	1	0	0	0	0	1		B<=>A
846	161528617	YP_001582443.1	luciferase family protein	368	74	200	0	1	1	1	0	0	1	M3	
847	161528618	YP_001582444.1	iron-containing alcohol dehydrogenase	314	155	200	0	1	1	0	0	1	1		B=>T
848	161528619	YP_001582445.1	hydroxypyruvate reductase	425	225	171	0	1	1	1	1	1	1		
849	161528620	YP_001582446.1	hypothetical protein Nmar_1112	78	77	6	1	0	0	0	0	0	0		
850	161528621	YP_001582447.1	hypothetical protein Nmar_1113	321	142	95	0	1	1	0	1	0	1		
851	161528622	YP_001582448.1	phosphoribosylglycinamide formyltransferase	191	138	200	0	1	0	0	0	1	1		B=>T
852	161528623	YP_001582449.1	cupin 2 domain-containing protein	106	105	5	0	0	0	0	0	0	1		B<=>T
853	161528624	YP_001582450.1	antibiotic biosynthesis monooxygenase	94	72	4	0	0	0	0	0	0	1		B<=>T
854	161528625	YP_001582451.1	methylenetetrahydrofolate dehydrogenase (NADP(+))	287	217	200	0	1	0	0	0	1	1		B=>T
855	161528626	YP_001582452.1	5-formyltetrahydrofolate cyclo-ligase	187	54	200	0	1	0	0	0	1	1		B=>T
856	161528627	YP_001582453.1	phosphoribosylamine--glycine ligase	421	321	200	0	0	0	0	0	1	1		B=>T
857	161528628	YP_001582454.1	isoleucyl-tRNA synthetase	1065	538	200	0	1	1	1	1	1	1	M	A=>B
858	161528629	YP_001582455.1	DNA protection protein DPS	195	132	45	0	1	1	1	0	0	1		
859	161528630	YP_001582456.1	hypothetical protein Nmar_1122	109	109	5	1	0	0	0	0	0	0		
860	161528631	YP_001582457.1	succinate--CoA ligase (ADP-forming)	368	283	200	0	1	1	1	0	1	1		A<=>B
861	161528632	YP_001582458.1	CoA-binding domain-containing protein	305	260	200	0	1	1	1	1	1	1		A<=>B
862	161528633	YP_001582459.1	50S ribosomal protein L40e	55	53	13	1	0	1	1	1	0	0		
863	161528636	YP_001582462.1	periplasmic solute binding protein	296	110	200	0	1	0	0	0	0	1		B=>T // B=>EA
864	161528637	YP_001582463.1	hypothetical protein Nmar_1129	454	26	60	0	1	0	0	1	0	1		A=>B
865	161528638	YP_001582464.1	putative metal cation transporter	392	184	21	0	1	1	0	0	0	1		
866	161528639	YP_001582465.1	multicopper oxidase type 3	351	77	200	0	1	0	1	0	1	1		B=>A
867	161528640	YP_001582466.1	DtxR family iron dependent repressor	160	67	177	0	1	1	0	0	0	1		B=>T
868	161528641	YP_001582467.1	kelch repeat-containing protein	341	45	200	0	0	1	1	0	1	1		
869	161528643	YP_001582469.1	spermine synthase	558	68	46	0	1	1	1	0	1	1		B=>A
870	161528644	YP_001582470.1	multicopper oxidase type 3	535	127	200	0	1	0	0	0	1	1		B=>A
871	161528646	YP_001582472.1	hypothetical protein Nmar_1138	232	183	4	0	1	0	0	0	0	1		
872	161528647	YP_001582473.1	hypothetical protein Nmar_1139	69	68	5	1	0	0	0	0	0	0		
873	161528648	YP_001582474.1	protein-disulfide isomerase-like protein	226	100	40	0	1	0	0	0	0	1		B<=>A // T=>B
874	161528650	YP_001582476.1	blue (type1) copper domain-containing protein	153	102	14	1	0	0	1	0	0	0		
875	161528651	YP_001582477.1	DSBA oxidoreductase	221	44	200	0	1	0	0	0	1	1		
876	161528653	YP_001582479.1	glutaredoxin	100	74	5	1	1	0	0	0	0	0		
877	161528654	YP_001582480.1	2-alkenal reductase	374	165	200	0	1	0	1	0	1	1		
878	161528655	YP_001582481.1	AsnC family transcriptional regulator	147	133	12	1	1	1	0	0	0	0		
879	161528656	YP_001582482.1	DSBA oxidoreductase	265	46	200	0	1	0	0	0	1	1		B=>T
880	161528657	YP_001582483.1	hypothetical protein Nmar_1149	229	72	4	1	0	0	0	0	0	0		
881	161528658	YP_001582484.1	DSBA oxidoreductase	247	49	200	0	1	0	0	0	1	1		B=>T
882	161528660	YP_001582486.1	hypothetical protein Nmar_1152	89	108	5	1	0	0	0	0	0	0		
883	161528661	YP_001582487.1	high-affinity nickel-transporter	233	155	16	0	0	0	0	0	1	1		B=>T
884	161528663	YP_001582489.1	ArsR family transcriptional regulator	122	43	199	0	1	1	1	0	0	1		B=>A
885	161528664	YP_001582490.1	cation efflux protein	205	159	88	0	0	0	0	0	0	1		B=>T
886	161528666	YP_001582492.1	hypothetical protein Nmar_1158	392	157	22	0	0	0	0	0	0	1		B=>T
887	161528667	YP_001582493.1	hypothetical protein Nmar_1159	314	156	4	1	1	1	0	0	0	0		
888	161528668	YP_001582494.1	Gp37Gp68 family protein	279	149	56	0	1	0	0	0	0	1		B=>T // B=>EA
889	161528669	YP_001582495.1	hypothetical protein Nmar_1161	456	62	29	1	0	0	0	0	0	0		
890	161528673	YP_001582499.1	MT-A70 family protein	305	123	96	0	0	0	0	0	1	1		B=>T
891	161528677	YP_001582503.1	integrase family protein	400	328	4	1	0	0	0	0	0	0		
892	161528678	YP_001582504.1	NAD(P)H dehydrogenase (quinone)	181	85	86	0	1	0	0	0	1	1		B=>T
893	161528679	YP_001582505.1	hypothetical protein Nmar_1171	126	123	4	1	0	0	0	0	0	0		
894	161528680	YP_001582506.1	vitamin K-dependent gamma-carboxylase	415	303	21	0	0	0	0	0	1	1		
895	161528683	YP_001582509.1	inorganic diphosphatase	179	132	200	0	1	1	0	0	1	1		B<=>A
896	161528684	YP_001582510.1	zinc finger UBP-type protein	85	61	27	0	1	0	0	0	0	1		B=>T // B=>EA
897	161528685	YP_001582511.1	pepF/M3 family oligoendopeptidase	590	136	183	0	1	0	0	0	1	1		B=>T // B=>EA
898	161528686	YP_001582512.1	hypothetical protein Nmar_1178	190	182	4	1	0	0	0	0	0	0		
899	161528687	YP_001582513.1	hypothetical protein Nmar_1179	381	323	4	1	0	0	0	0	0	0		
900	161528688	YP_001582514.1	arginine decarboxylase_pyruvoyl-dependent	183	127	62	0	1	1	0	0	0	1		A=>B
901	161528689	YP_001582515.1	DSBA oxidoreductase	265	46	200	0	1	0	0	0	1	1		
902	161528690	YP_001582516.1	DEAD/DEAH box helicase domain-containing protein	440	277	200	0	1	0	0	0	0	1		B<=>A
903	161528691	YP_001582517.1	AsnC/Lrp family regulatory protein	77	76	31	1	0	1	1	0	0	0	M2	
904	161528693	YP_001582519.1	hypothetical protein Nmar_1185	231	159	5	1	0	0	0	0	0	0		
905	161528694	YP_001582520.1	hypothetical protein Nmar_1186	252	169	4	1	0	0	0	0	0	0		
906	161528698	YP_001582524.1	hypothetical protein Nmar_1190	86	86	4	1	0	0	0	0	0	0		
907	161528699	YP_001582525.1	hypothetical protein Nmar_1191	675	266	200	0	1	1	1	0	1	1		
908	161528701	YP_001582527.1	alcohol dehydrogenase	344	197	200	0	1	1	0	0	1	1		B=>A
909	161528705	YP_001582531.1	AsnC family transcriptional regulator	76	74	17	1	0	1	1	0	0	0		
910	161528706	YP_001582532.1	tRNA pseudouridine synthase D TruD	398	126	183	0	1	1	0	1	1	1	M2	A=>B
911	161528707	YP_001582533.1	hypothetical protein Nmar_1199	533	407	10	1	0	0	0	0	0	0		
912	161528708	YP_001582534.1	peptidyl-tRNA hydrolase	124	85	200	0	1	1	1	1	1	0	M2	EA<=>T
913	161528709	YP_001582535.1	hypothetical protein Nmar_1201	1708	93	24	1	0	1	0	0	0	0		
914	161528710	YP_001582536.1	response regulator receiver protein	124	58	200	0	1	1	0	0	1	1		B=>T // B=>EA // EA<=>T
915	161528712	YP_001582538.1	hypothetical protein Nmar_1204	304	217	4	1	0	0	0	0	0	0		
916	161528714	YP_001582540.1	response regulator receiver protein	145	62	200	0	1	0	0	0	1	1		B=>T // B=>EA
917	161528715	YP_001582541.1	histidine kinase	577	73	200	0	1	0	0	0	1	1		B=>T // B=>EA
918	161528718	YP_001582544.1	metal dependent phosphohydrolase	434	109	200	0	1	1	1	1	1	1		
919	161528719	YP_001582545.1	hypothetical protein Nmar_1211	184	112	18	0	1	0	0	0	0	1		T<=>EA
920	161528720	YP_001582546.1	metallophosphoesterase	415	75	92	0	1	1	1	0	1	1		
921	161528721	YP_001582547.1	SMC domain-containing protein	806	137	66	0	1	1	0	1	0	1		A=>T // A=>B
922	161528722	YP_001582548.1	cobyrinic acid ac-diamide synthase	330	92	99	0	1	1	0	0	1	1		B=>T
923	161528727	YP_001582553.1	integrase family protein	351	328	4	1	0	0	0	0	0	0		
924	161528728	YP_001582554.1	dehydrogenase (flavoprotein)-like protein	378	118	165	0	1	1	1	1	1	1	M3	
925	161528730	YP_001582556.1	translation elongation factor aEF-1 beta	93	73	53	1	1	1	1	0	0	0	M	
926	161528731	YP_001582557.1	AAA family ATPase_CDC48 subfamily protein	721	571	200	0	1	1	1	1	1	1	M2	A=>B
927	161528733	YP_001582559.1	protoheme IX farnesyltransferase	289	201	200	0	1	1	1	1	0	1		B=>A // T=>EA
928	161528734	YP_001582560.1	blue (type1) copper domain-containing protein	358	41	65	0	1	0	1	0	0	1		A<=>B
929	161528735	YP_001582561.1	Mov34/MPN/PAD-1 family protein	130	62	42	0	1	1	0	0	0	1		A<=>B
930	161528736	YP_001582562.1	hypothetical protein Nmar_1228	127	99	18	0	1	1	0	0	0	1		A=>B
931	161528737	YP_001582563.1	hypothetical protein Nmar_1229	85	84	9	0	0	0	0	0	0	1		B<=>T
932	161528739	YP_001582565.1	membrane-bound proton-translocating pyrophosphatase	679	522	200	0	1	1	1	1	1	1		
933	161528741	YP_001582567.1	hypothetical protein Nmar_1233	154	86	81	1	1	1	1	1	0	0	M3	
934	161528742	YP_001582568.1	hypothetical protein Nmar_1234	370	209	200	0	1	0	1	0	0	1		B=>T // B=>EA
935	161528743	YP_001582569.1	DNA-directed RNA polymerase subunit Ep	198	114	118	0	1	1	1	1	1	0	M	
936	161528744	YP_001582570.1	DNA-directed RNA polymerase subunit E_RpoE2	62	50	75	1	1	1	1	1	0	0	M	
937	161528745	YP_001582571.1	nicotinate-nucleotide pyrophosphorylase	272	156	200	0	1	1	0	0	0	1		
938	161528747	YP_001582573.1	quinolinate synthetase complex_A subunit	318	212	200	0	1	1	0	0	1	1		B=>T
939	161528748	YP_001582574.1	aspartate dehydrogenase	272	146	79	0	1	1	0	0	1	1		A=>B
940	161528749	YP_001582575.1	hypothetical protein Nmar_1241	317	231	8	1	0	0	0	0	0	0		
941	161528750	YP_001582576.1	hypothetical protein Nmar_1242	278	273	5	1	1	1	0	0	0	0		
942	161528751	YP_001582577.1	hypothetical protein Nmar_1243	510	112	172	0	1	1	1	1	0	1		
943	161528752	YP_001582578.1	metallophosphoesterase	380	71	110	0	1	1	1	1	0	1		A=>B
944	161528753	YP_001582579.1	SMC domain-containing protein	693	74	141	0	1	1	1	1	1	1		A=>B EA<=>T
945	161528755	YP_001582581.1	histidine kinase	546	84	200	0	1	0	0	0	1	1		B=>T
946	161528756	YP_001582582.1	response regulator receiver protein	102	52	169	0	1	0	0	0	1	1		EA<=>T
947	161528758	YP_001582584.1	protease inhibitor Kazal-type	239	52	39	0	1	0	1	0	0	1		T=>B
948	161528759	YP_001582585.1	HEAT repeat-containing PBS lyase	162	169	30	0	1	0	1	0	1	1		
949	161528760	YP_001582586.1	alkyl hydroperoxide reductase/ Thiol specific antioxidant/ Mal allergen	221	143	9	0	0	0	0	0	0	1		T<=>B
950	161528761	YP_001582587.1	hypothetical protein Nmar_1253	257	125	29	0	0	0	0	1	0	1		T<=>B
951	161528762	YP_001582588.1	50S ribosomal protein L15e	154	173	200	0	1	1	1	1	1	0	M	
952	161528763	YP_001582589.1	hypothetical protein Nmar_1255	219	214	7	1	0	0	0	0	0	0		
953	161528765	YP_001582591.1	hypothetical protein Nmar_1257	404	147	4	1	0	0	0	0	0	0		
954	161528766	YP_001582592.1	D-isomer specific 2-hydroxyacid dehydrogenase NAD-binding	310	255	200	0	1	1	1	0	1	1		
955	161528767	YP_001582593.1	hypothetical protein Nmar_1259	470	134	85	0	1	0	1	0	1	1		B<=>T
956	161528768	YP_001582594.1	fumarate lyase	453	364	200	0	1	0	0	0	1	1		B=>A
957	161528769	YP_001582595.1	MiaB-like tRNA modifying enzyme	422	183	200	0	1	1	1	1	1	1	M2	
958	161528770	YP_001582596.1	tubulin/FtsZ GTPase	316	288	24	1	1	0	0	0	0	0		EA<=>T
959	161528771	YP_001582597.1	ribosomal protein L31e	187	70	81	0	1	1	1	0	1	0	M3	
960	161528772	YP_001582598.1	ribosomal protein L39e	52	49	4	1	0	1	0	0	0	0		
961	161528773	YP_001582599.1	NUDIX hydrolase	183	70	200	0	1	1	1	0	1	1		
962	161528775	YP_001582601.1	methionine synthase	832	845	200	0	0	0	0	0	1	1		B=>T
963	161528776	YP_001582602.1	homocysteine S-methyltransferase	320	610	200	0	0	0	0	0	1	1		B=>T
964	161528777	YP_001582603.1	major facilitator transporter	382	308	15	0	1	0	0		0	1		
965	161528779	YP_001582605.1	signal transduction protein	285	82	69	0	1	1	0	1	1	1		T<=>EA
966	161528780	YP_001582606.1	UspA domain-containing protein	138	49	164	0	1	1	0	1	1	1		
967	161528781	YP_001582607.1	blue (type1) copper domain-containing protein	287	75	35	0	1	0	1	0	0	1		T=>B
968	161528782	YP_001582608.1	UspA domain-containing protein	145	50	200	0	1	1	0	1	0	1		
969	161528783	YP_001582609.1	rhodanese domain-containing protein	257	37	140	0	1	0	0	0	1	1		
970	161528784	YP_001582610.1	hypothetical protein Nmar_1276	185	105	23	1	1	0	0	0	0	0		T<=>EA
971	161528789	YP_001582615.1	ArsR family transcriptional regulator	154	75	33	1	1	0	0	0	0	0		T<=>EA
972	161528790	YP_001582616.1	hypothetical protein Nmar_1282	127	88	13	1	0	0	0	0	0	0		
973	161528791	YP_001582617.1	binding-protein-dependent transport systems inner membrane component	252	134	200	0	1	0	1	0	0	1		B=>T
974	161528792	YP_001582618.1	ABC transporter related	254	163	200	0	1	1	0	1	0	1		B=>T
975	161528793	YP_001582619.1	aliphatic sulfonate ABC transporter periplasmic ligand-binding protein	341	96	76	0	0	0	0	0	0	1		B=>T
976	161528794	YP_001582620.1	argininosuccinate synthase	399	322	200	0	1	1	1	1	1	1	M	
977	161528795	YP_001582621.1	hypothetical protein Nmar_1287	55	53	11	0	1	1	1	1	0	0		
978	161528796	YP_001582622.1	lysine biosynthesis enzyme LysX	285	126	200	0	1	1	1	1	0	1		A=>B
979	161528797	YP_001582623.1	N-acetyl-gamma-glutamyl-phosphate reductase	348	217	200	0	1	1	1	1	1	1	M2	A=>B
980	161528798	YP_001582624.1	acetylglutamate kinase	267	161	200	0	1	1	1	1	1	1	M3	A=>B // EA<=>T
981	161528799	YP_001582625.1	acetylornithine and succinylornithine aminotransferase	393	227	200	0	1	1	0	1	1	1	M3	A<=>B
982	161528800	YP_001582626.1	AsnC family transcriptional regulator	139	67	107	0	1	1	1	1	0	1		A<=>B
983	161528801	YP_001582627.1	pyruvate carboxyltransferase	393	262	200	0	1	1	1	1	1	1	M3	
984	161528802	YP_001582628.1	hypothetical protein Nmar_1294	55	54	8	1	1	1	0	0	0	0		
985	161528803	YP_001582629.1	lysine biosynthesis enzyme LysX	280	121	200	0	1	1	1	1	1	1	M3	
986	161528804	YP_001582630.1	N-acetyl-ornithine/N-acetyl-lysine deacetylase	375	82	200	0	1	1	1	1	1	1	M3	
987	161528805	YP_001582631.1	diphthine synthase	345	180	188	0	1	1	1	0	1	0	M	
988	161528806	YP_001582632.1	cytidyltransferase-like protein	206	78	200	0	1	1	1	1	1	1	M	EA<=>T
989	161528807	YP_001582633.1	PAS/PAC sensor signal transduction histidine kinase	618	52	200	0	1	0	0	0	1	1		B=>T // B=>EA
990	161528809	YP_001582635.1	hypothetical protein Nmar_1301	227	84	90	0	1	1	1	1	0	1	M2	A=>B
991	161528810	YP_001582636.1	DNA primase	380	258	102	0	1	1	1	1	0	0	M	
992	161528811	YP_001582637.1	iron-sulfur cluster assembly accessory protein	116	76	200	0	1	0	0	0	1	1		B=>T // B=>EA
993	161528812	YP_001582638.1	signal-transduction protein	148	37	200	0	1	1	1	1	1	1	M3	A<=>B // T<=>EA
994	161528814	YP_001582640.1	hypothetical protein Nmar_1306	116	75	13	1	1	0	0	0	0	0		
995	161528815	YP_001582641.1	blue (type1) copper domain-containing protein	174	62	39	0	1	0	1	0	0	1		T<=>EA // A=>B
996	161528816	YP_001582642.1	enoyl-CoA hydratase/isomerase	253	172	200	0	1	1	0	1	1	1		
997	161528817	YP_001582643.1	CoA-binding domain-containing protein	705	279	200	0	1	1	0	1	1	1		
998	161528818	YP_001582644.1	glutamine amidotransferase class-I	227	65	200	0	1	1	0	0	1	1		
999	161528819	YP_001582645.1	TOPRIM domain-containing protein	124	123	13	0	1	0	0	0	0	0		A=>B
1000	161528820	YP_001582646.1	Glu/Leu/Phe/Val dehydrogenase	424	290	200	0	1	1	1	1	1	1		
1001	161528821	YP_001582647.1	ArsR family transcriptional regulator	127	71	39	1	1	1	1	0	0	0	M2	EA=>T
1002	161528822	YP_001582648.1	proteasome subunit alpha	240	173	200	0	1	1	1	1	1	0	M	
1003	161528823	YP_001582649.1	methyltransferase type 11	230	92	200	0	1	1	1	1	1	1	M3	CA=>EA
1004	161528824	YP_001582650.1	major facilitator transporter	384	328	10	0	1	0	1	0	0	1		
1005	161528825	YP_001582651.1	nitrogen regulatory protein P-II	104	105	200	0	1	0	0	0	1	1		B=>T // B=>EA
1006	161528827	YP_001582653.1	DNA methylase N-4/N-6 domain-containing protein	258	63	183	0	1	1	0	0	0	1		B=>A
1007	161528828	YP_001582654.1	DEAD/DEAH box helicase domain-containing protein	815	214	200	0	1	1	0	0	1	1	M3	
1008	161528829	YP_001582655.1	short-chain dehydrogenase/reductase SDR	276	105	200	0	1	0	0	0	1	1		B=>T // B=>EA
1009	161528830	YP_001582656.1	2-alkenal reductase	551	326	120	0	0	0	1	0	1	1		
1010	161528832	YP_001582658.1	integrase family protein	423	140	39	0	1	0	0	0	0	1		A=>B

1011	161528834	YP_001582660.1	transcriptional regulator TrmB	171	46	98	1	1	0	0	0	0	0		
1012	161528836	YP_001582662.1	hypothetical protein Nmar_1328	486	428	7	1	0	1	0	0	0	0		
1013	161528839	YP_001582665.1	parB-like partition protein	269	134	152	0	0	0	0		1	1		B=>T
1014	161528843	YP_001582669.1	MscS mechanosensitive ion channel	275	113	200	0	1	1	0	0	0	1	M3	
1015	161528844	YP_001582670.1	hypothetical protein Nmar_1336	247	126	21	1	1	1	0	0	0	0	M3	
1016	161528846	YP_001582672.1	transcriptional regulator protein-like protein	265	243	8	1	1	0	0	0	0	0		
1017	161528848	YP_001582674.1	transcription factor TFIIB cyclin-related	306	254	200	0	1	1	1	1	1	0	M	
1018	161528849	YP_001582675.1	transcription factor TFIIB cyclin-related	309	247	200	0	1	1	1	1	1	0	M	
1019	161528851	YP_001582677.1	ectoine hydroxylase	304	123	43	0	0	0	0	0	1	1		B=>T
1020	161528852	YP_001582678.1	L-ectoine synthase	128	114	38	0	0	0	0	0	1	1		B=>T
1021	161528853	YP_001582679.1	diaminobutyrate--2-oxoglutarate aminotransferase	431	199	200	0	1	1	1	1	1	1		B=>T
1022	161528854	YP_001582680.1	L-2,4-diaminobutyric acid acetyltransferase	167	124	30	0	0	0	0	0	1	1		B=>T
1023	161528856	YP_001582682.1	translation initiation factor eIF-1A	104	81	107	0	1	1	1	1	1	0	M	
1024	161528857	YP_001582683.1	sensor protein	574	32	78	0	1	0	0	0	0	1		B=>T // B=>EA
1025	161528858	YP_001582684.1	hypothetical protein Nmar_1350	116	113	4	1	0	0		0	0	0		
1026	161528860	YP_001582686.1	ArsR family transcriptional regulator	266	248	7	1	1	0	0	0	0	0		
1027	161528861	YP_001582687.1	hypothetical protein Nmar_1353	180	143	13	0	1	1	0	0	0	1		
1028	161528862	YP_001582688.1	multicopper oxidase type 3	444	87	200	0	1	0	1	0	1	1		B=>T // B=>EA
1029	161528864	YP_001582690.1	flavodoxin FldA	174	115	141	0	0	0	0	0	1	1		B=>T
1030	161528865	YP_001582691.1	nitroreductase	214	123	42	0	1	1	0	1	0	1		
1031	161528866	YP_001582692.1	beta-lactamase domain-containing protein	479	76	35	0	0	0	0	0	1	1		E<=>T // B=>T
1032	161528867	YP_001582693.1	ArsR family transcriptional regulator	225	221	5	1	1	0	1	1	0	0		
1033	161528868	YP_001582694.1	hypothetical protein Nmar_1360	258	55	135	0	1	0	1	0	1	1		
1034	161528870	YP_001582696.1	hypothetical protein Nmar_1362	113	105	7	1	0	0	0	0	0	0		
1035	161528875	YP_001582701.1	hypothetical protein Nmar_1367	398	110	6	1	0	0	0	0	0	0		
1036	161528876	YP_001582702.1	hypothetical protein Nmar_1368	272	99	92	0	0	0	0	0	1	1		
1037	161528877	YP_001582703.1	hypothetical protein Nmar_1369	56	56	4	1	0	0	0	0	0	0		
1038	161528879	YP_001582705.1	cation efflux protein	206	159	88	0	0	0	0	0	0	1		B=>T
1039	161528880	YP_001582706.1	ArsR family transcriptional regulator	109	86	23	0	1	1	0	0	0	1		B=>T
1040	161528886	YP_001582712.1	metallophosphoesterase	252	69	91	0	1	1	1	0	1	1		
1041	161528887	YP_001582713.1	isocitrate dehydrogenase (NAD(+))	343	230	200	0	1	1	1	1	1	1	M2	
1042	161528888	YP_001582714.1	putative methylase	178	100	74	0	1	1	0	0	1	0	M3	
1043	161528889	YP_001582715.1	ribosomal RNA adenine methylase transferase	234	124	200	0	1	1	0	1	1	1	M3	
1044	161528890	YP_001582716.1	hypothetical protein Nmar_1382	188	128	96	1	1	1	1	1	0	0	M	
1045	161528891	YP_001582717.1	RNA polymerase Rpb4	108	108	5	1	0	0	0	0	0	0		
1046	161528892	YP_001582718.1	ribosomal protein L21e	99	86	73	0	1	1	1	1	1	0	M	
1047	161528893	YP_001582719.1	radical SAM protein	351	197	200	0	0	0	0	0	1	1		
1048	161528894	YP_001582720.1	DNA repair and recombination protein RadA	388	246	200	0	1	1	1	1	1	0	M	
1049	161528896	YP_001582722.1	radical SAM domain-containing protein	574	302	71	0	1	1	1	1	0	1	M3	
1050	161528897	YP_001582723.1	oxidoreductase FAD/NAD(P)-binding subunit	270	111	191	0	1	1	1	1	0	1	M3	B=>A
1051	161528898	YP_001582724.1	dihydroorotate dehydrogenase family protein	303	149	200	0	1	1	0	1	1	1	M2	
1052	161528899	YP_001582725.1	Pre-mRNA processing ribonucleoprotein	470	249	200	0	1	1	1	1	1	0	M	
1053	161528900	YP_001582726.1	fibrillarin	224	188	200	0	1	1	1	1	1	0	M	
1054	161528901	YP_001582727.1	ribonuclease HII	205	71	200	0	1	1	1	1	1	1	M3	
1055	161528902	YP_001582728.1	tRNA (guanine-N(2)-)-methyltransferase	382	149	200	0	1	1	1	1	1	1	M	
1056	161528903	YP_001582729.1	ribosomal RNA methyltransferase RrmJ/FtsJ	197	130	200	0	1	0	0	0	1	1	M2	
1057	161528904	YP_001582730.1	hypothetical protein Nmar_1396	291	100	14	1	0	1	1	0	0	0		
1058	161528906	YP_001582732.1	hypothetical protein Nmar_1398	697	315	14	0	1	0	0	0	0	1		B<=>T
1059	161528907	YP_001582733.1	hypothetical protein Nmar_1399	722	698	5	1	0	0	0	0	0	0		
1060	161528908	YP_001582734.1	adenine phosphoribosyltransferase	170	125	200	0	1	0	0	0	1	1		B=>T
1061	161528909	YP_001582735.1	methylthioadenosine phosphorylase	263	166	200	0	1	1	0	1	1	1	M3	A=>B
1062	161528911	YP_001582737.1	hypothetical protein Nmar_1403	132	132	4	1	0	0	0	0	0	0		
1063	161528912	YP_001582738.1	hypothetical protein Nmar_1404	116	84	4	1	0	0	0	0	0	0		
1064	161528914	YP_001582740.1	hypothetical protein Nmar_1406	170	152	5	1	0	0	0	0	0	0		
1065	161528915	YP_001582741.1	DNA topoisomerase VI subunit A	370	162	174	0	1	1	1	1	1	1	M	A=>B
1066	161528916	YP_001582742.1	DNA topoisomerase VI B subunit	626	307	130	0	1	1	1	1	1	1	M	A=>B
1067	161528917	YP_001582743.1	putative RNA-processing protein	190	133	89	0	1	1	1	1	1	0	M	
1068	161528918	YP_001582744.1	non-specific serine/threonine protein kinase	259	143	200	0	1	1	1	1	1	0	M	
1069	161528919	YP_001582745.1	hypothetical protein Nmar_1411	98	91	28	1	1	1	0	0	0	0	M3	
1070	161528920	YP_001582746.1	translation initiation factor IF2/IF5	140	88	200	0	1	1	1	1	1	0	M	
1071	161528921	YP_001582747.1	hypothetical protein Nmar_1413	53	45	8	0	0	1	0	0	0	1		
1072	161528922	YP_001582748.1	hypothetical protein Nmar_1414	165	128	4	1	0	0	0	0	0	0		
1073	161528923	YP_001582749.1	hypothetical protein Nmar_1415	369	184	95	1	1	1	1	1	0	0	M	
1074	161528924	YP_001582750.1	undecaprenyl diphosphate synthase	263	130	200	0	1	1	1	1	1	1	M3	A=>B
1075	161528925	YP_001582751.1	NUDIX hydrolase	139	57	71	0	1	1	1	0	1	1		
1076	161528926	YP_001582752.1	orotidine 5p-phosphate decarboxylase	246	105	124	0	1	1	1	1	1	1	M3	A=>B
1077	161528927	YP_001582753.1	hypothetical protein Nmar_1419	227	206	7	1	0	0	0	0	0	0		
1078	161528928	YP_001582754.1	adenylosuccinate synthetase	333	250	200	0	1	1	0	0	1	1	M3	A=>B
1079	161528931	YP_001582757.1	hypothetical protein Nmar_1423	141	104	9	1	0	0	0	0	0	0		
1080	161528933	YP_001582759.1	hypothetical protein Nmar_1425	213	188	4	1	0	0	0	0	0	0		
1081	161528934	YP_001582760.1	hypothetical protein Nmar_1426	95	89	12	1	1	0	1	0	0	0		
1082	161528935	YP_001582761.1	SuaS/YciO/YrdC/YwlC family protein	198	101	200	0	1	1	1	1	1	1		
1083	161528936	YP_001582762.1	THUMP domain-containing protein	173	168	13	0	1	1	0	0	1	0		
1084	161528938	YP_001582764.1	hypothetical protein Nmar_1430	129	62	8	0	1	0	0		1	0		
1085	161528939	YP_001582765.1	TatD-related deoxyribonuclease	272	90	200	0	1	1	1	1	1	1		A=>B
1086	161528940	YP_001582766.1	transcription factor CBF/NF-Y/histone domain-containing protein	78	62	91	1	1	1	1	1	0	0		EA=>T
1087	161528946	YP_001582772.1	alkyl hydroperoxide reductase/ Thiol specific antioxidant/ Mal allergen	189	167	18	0	1	0	0	0	1	1		
1088	161528947	YP_001582773.1	glyoxalase/bleomycin resistance protein/dioxygenase	136	78	21	0	1	0	0	0	0	1		
1089	161528948	YP_001582774.1	histidine kinase	427	77	200	0	1	0	0	0	1	1		B=>T
1090	161528951	YP_001582777.1	blue (type1) copper domain-containing protein	152	47	59	0	1	0	1	0	0	1		
1091	161528952	YP_001582778.1	hypothetical protein Nmar_1444	583	104	45	0	1	0	0	0	0	1		B=>T // B=>EA
1092	161528954	YP_001582780.1	hypothetical protein Nmar_1446	93	83	13	1	0	1	1	0	0	0		
1093	161528955	YP_001582781.1	phosphoglucomutase/phosphomannomutase alpha/beta/subunit	432	222	200	0	1	1	0	1	1	1	M3	
1094	161528956	YP_001582782.1	thiamine-monophosphate kinase	310	144	200	0	1	1	1	1	0	1	M3	A<=>B
1095	161528958	YP_001582784.1	ribosomal protein S11	108	125	200	0	1	1	1	0	1	0	M	
1096	161528960	YP_001582786.1	hypothetical protein Nmar_1452	134	109	6	1	0	0	1	0	0	0		
1097	161528961	YP_001582787.1	hypothetical protein Nmar_1453	482	315	200	0	1	1	1	1	1	1	M3	A=B
1098	161528963	YP_001582789.1	adenylosuccinate lyase	453	333	200	0	1	1	0	1	0	1	M2	
1099	161528964	YP_001582790.1	hypothetical protein Nmar_1456	334	32	62	0	0	0	0	0	0	1		B<=>T
1100	161528965	YP_001582791.1	methyltransferase type 11	344	21	200	0	1	0	0	0	1	1		
1101	161528970	YP_001582796.1	hypothetical protein Nmar_1462	621	361	15	1	0	0	0	0	0	0		
1102	161528971	YP_001582797.1	glyoxalase/bleomycin resistance protein/dioxygenase	127	61	32	0	1	0	0	0	1	1		B=>T
1103	161528972	YP_001582798.1	hypothetical protein Nmar_1464	130	102	9	1	0	0	0	0	0	0		
1104	161528974	YP_001582800.1	asparagine synthase	251	146	144	0	1	1	1	0	1	1		EA<>T // B<=>A
1105	161528975	YP_001582801.1	Excalibur domain-containing protein	259	73	40	0	1	0	0	0	0	1		
1106	161528976	YP_001582802.1	transcriptional coactivator/pterin dehydratase	91	49	180	0	1	0	0	0	1	1		B=>T
1107	161528977	YP_001582803.1	hypothetical protein Nmar_1469	272	118	18	0	1	0	1	0	0	1		
1108	161528979	YP_001582805.1	CMP/dCMP deaminase zinc-binding	120	84	46	0	1	0	0	0	1	1		B=>A
1109	161528981	YP_001582807.1	DNA-cytosine methyltransferase	360	93	200	0	1	1	0	0	1	1		B<=>T
1110	161528982	YP_001582808.1	hypothetical protein Nmar_1474	290	266	5	0	0	0	0	0	0	1		B<=>T
1111	161528985	YP_001582811.1	valyl-tRNA synthetase	771	434	200	0	1	1	1	1	1	1	M	A=>B
1112	161528986	YP_001582812.1	SNO glutamine amidotransferase	205	112	200	0	1	1	0	1	1	1	M3	
1113	161528987	YP_001582813.1	pyridoxine biosynthesis protein	322	261	200	0	1	1	0	1	1	1	M	
1114	161528989	YP_001582815.1	signal transduction protein	282	94	61	0	1	1	0	0	0	1		
1115	161528990	YP_001582816.1	aconitate hydratase	754	441	200	0	1	1	1	0	1	1		B=>T // B=>EA
1116	161528993	YP_001582819.1	AsnC family transcriptional regulator	79	69	72	0	1	1	1	0	0	1		A=>B
1117	161528994	YP_001582820.1	hypothetical protein Nmar_1486	139	82	162	0	0	0	0	0	1	1		B=>T
1118	161528996	YP_001582822.1	phenylalanyl-tRNA synthetase subunit beta	547	257	200	0	1	1	1	1	1	1	M2	A=>B
1119	161528997	YP_001582823.1	phenylalanyl-tRNA synthetase alpha subunit	462	229	200	0	1	1	11	1	1	1	M2	A=>B
1120	161528998	YP_001582824.1	tryptophanyl-tRNA synthetase	370	253	200	0	1	1	1	1	1	0	M2	
1121	161529000	YP_001582826.1	hypothetical protein Nmar_1492	336	113	5	0	0	0	0	0	0	1		
1122	161529002	YP_001582828.1	cytidyltransferase-like protein	158	72	149	0	1	1	1	1	1	1	M3	A=>B
1123	161529004	YP_001582830.1	alkyl hydroperoxide reductase/ Thiol specific antioxidant/ Mal allergen	184	100	154	0	1	0	0	0	1	1		B=>A
1124	161529005	YP_001582831.1	putative transcriptional regulator	295	274	4	1	0	0	0	0	0	0		
1125	161529006	YP_001582832.1	hypothetical protein Nmar_1498	127	127	4	1	0	0	0	0	0	0		
1126	161529007	YP_001582833.1	DNA adenine methylase	280	96	178	0	1	0	0	0	1	1		
1127	161529008	YP_001582834.1	hypothetical protein Nmar_1500	216	184	31	1	0	0	0	0	0	0		
1128	161529009	YP_001582835.1	hypothetical protein Nmar_1501	120	118	5	1	0	0	0	0	0	0		
1129	161529010	YP_001582836.1	ammonia monooxygenase/methane monooxygenase subunit C	190	163	10	0	0	0	0	0	0	1		T<=>B
1130	161529011	YP_001582837.1	hypothetical protein Nmar_1503	189	125	16	1	0	0	0	0	0	0		
1131	161529012	YP_001582838.1	hypothetical protein Nmar_1504	104	100	4	1	0	0	0	0	0	0		
1132	161529013	YP_001582839.1	hypothetical protein Nmar_1505	172	172	4	1	0	0	0	0	0	0		
1133	161529014	YP_001582840.1	hypothetical protein Nmar_1506	89	89	4	1	0	0	0	0	0	0		
1134	161529016	YP_001582842.1	30S ribosomal protein S15P	149	148	200	0	1	1	1	1	1	0	M	
1135	161529017	YP_001582843.1	phosphoesterase DHHA1	473	205	110	0	1	1	1	1	0	0	M2	
1136	161529019	YP_001582845.1	seryl-tRNA synthetase	421	311	200	0	1	1	1	1	1	1	M3	A<=>E
1137	161529020	YP_001582846.1	30S ribosomal protein S3Ae	203	155	200	0	1	1	1	1	1	0	M	
1138	161529021	YP_001582847.1	hypothetical protein Nmar_1513	141	129	14	1	1	0	0	0	0	0		
1139	161529022	YP_001582848.1	geranylgeranyl reductase	485	156	118	0	1	1	1	1	0	1	M2	
1140	161529023	YP_001582849.1	ATPase	264	94	200	0	1	1	1	1	1	1		
1141	161529024	YP_001582850.1	von Willebrand factor type A	520	520	4	1	0	0	0	0	0	0		
1142	161529025	YP_001582851.1	pyridoxal-5p-phosphate-dependent protein beta subunit	514	277	10	0	0	0	0	0	0	1		
1143	161529026	YP_001582852.1	hypothetical protein Nmar_1518	151	115	10	0	1	0	0	0	1	1		
1144	161529027	YP_001582853.1	TATA-box binding family protein	187	135	200	0	1	1	1	1	1	0	M	
1145	161529030	YP_001582856.1	hypothetical protein Nmar_1522	236	197	6	1	0	0	0	0	0	0		
1146	161529032	YP_001582858.1	AsnC family transcriptional regulator	79	69	72	0	1	1	1	0	0	1	M2	
1147	161529033	YP_001582859.1	nifR3 family TIM-barrel protein	320	175	200	0	1	0	0	0	0	1		B=>A
1148	161529034	YP_001582860.1	thrombospondin type 3 repeat-containing protein	722	126	40	0	0	0	0	0	0	1		B=>T
1149	161529035	YP_001582861.1	thrombospondin type 3 repeat-containing protein	722	120	55	0	0	0	0	0	1	1		B=>T
1150	161529036	YP_001582862.1	ATPase	792	88	56	0	0	0	0	0	1	1		B=>T
1151	161529037	YP_001582863.1	tRNA-modifying enzyme	341	203	161	0	1	1	1	1	1	1	M2	
1152	161529038	YP_001582864.1	RdgBIHAM1 family non-canonical purine NTP pyrophosphatase	188	93	200	0	1	1	1	1	1	1	M2	A<=>B
1153	161529039	YP_001582865.1	Mn2+-dependent serine/threonine protein kinase	206	87	174	0	1	1	1	1	1	0	M2	A<=>E
1154	161529041	YP_001582867.1	metalloendopeptidase glycoprotease family	327	224	200	0	1	1	1	1	1	1	M2	
1155	161529042	YP_001582868.1	translation-associated GTPase	398	182	200	0	1	1	1	1	1	1	M3	
1156	161529043	YP_001582869.1	hypothetical protein Nmar_1535	182	89	107	0	1	1	1	1	1	0	M	
1157	161529044	YP_001582870.1	tRNA-guanine transglycosylase	507	210	200	0	1	1	1	1	0	1	M	
1158	161529045	YP_001582871.1	4Fe-4S ferredoxin iron-sulfur binding domain-containing protein	100	99	10	1	0	0	0	0	0	0		
1159	161529046	YP_001582872.1	hypothetical protein Nmar_1538	70	68	5	1	0	0	0	0	0	0		
1160	161529047	YP_001582873.1	beta-lactamase domain-containing protein	421	220	200	0	1	1	1	1	0	1	M	
1161	161529048	YP_001582874.1	HhH-GPD family protein	287	92	131	0	1	1	0	0	1	1		A<=>B
1162	161529050	YP_001582876.1	blue (type1) copper domain-containing protein	154	95	8	1	0	0	1	0	0	0		
1163	161529051	YP_001582877.1	cytochrome b/b6 domain-containing protein	531	159	200	0	1	1	1	1	1	1		B=>A
1164	161529052	YP_001582878.1	Rieske (2Fe-2S) domain-containing protein	201	180	7	1	0	0	1	1	0	0		
1165	161529053	YP_001582879.1	peptidase S8 and S53 subtilisin kexin sedolisin	1269	114	120	0	1	1	0	0	1	1		
1166	161529054	YP_001582880.1	hypothetical protein Nmar_1546	334	319	5	1	0	0	0	0	0	0		
1167	161529055	YP_001582881.1	hypothetical protein Nmar_1547	1734	108	23	1	0	1	0	0	0	0		
1168	161529056	YP_001582882.1	UBA/THF-type NAD/FAD binding protein	443	166	200	0	1	1	0	1	1	1		
1169	161529057	YP_001582883.1	threonine synthase	404	254	200	0	1	1	1	1	0	1		
1170	161529059	YP_001582885.1	phosphoribosyl-AMP cyclohydrolase	108	85	200	0	1	0	1	0	1	1		B=>T
1171	161529060	YP_001582886.1	imidazoleglycerol phosphate synthase cyclase subunit	266	234	200	0	1	1	1	0	0	1		A<=>B
1172	161529061	YP_001582887.1	1-(5-phosphoribosyl)-5-[(5-phosphoribosylamino)methylideneamino]imidazole-4-carboxamideisomerase	236	147	200	0	1	0	0	0	0	1		A<=>B
1173	161529062	YP_001582888.1	imidazole glycerol phosphate synthase glutamine amidotransferase subunit	200	119	200	0	1	0	1	0	1	1		B=>T+AA // A<=>B
1174	161529063	YP_001582889.1	imidazoleglycerol-phosphate dehydratase	195	153	200	0	1	1	1	0	1	1		
1175	161529064	YP_001582890.1	hydrolase	306	258	17	1	1	0	1	0	0	0		
1176	161529065	YP_001582891.1	aminotransferase class I and II	356	135	200	0	1	1	1	0	1	1		B=>T
1177	161529066	YP_001582892.1	histidinol dehydrogenase	420	265	200	0	1	0	1	0	1	1	M2	
1178	161529067	YP_001582893.1	ATP phosphoribosyltransferase	325	147	200	0	1	1	1	0	1	1		B=>T // B=>EA
1179	161529068	YP_001582894.1	hydroxymethylglutaryl-CoA reductase degradative	418	280	70	0	1	1	1	1	1	1		

1180	161529069	YP_001582895.1	glucose sorbosone dehydrogenase	363	118	200	0	1	1	1	0	1	1		B=>T // B=>EA
1181	161529070	YP_001582896.1	hypothetical protein Nmar_1562	450	204	96	1	1	1	1	0	0	0	M	
1182	161529071	YP_001582897.1	diphthamide biosynthesis protein	331	136	165	0	1	1	1	0	1	0	M	
1183	161529072	YP_001582898.1	TPR repeat-containing protein	199	29	200	0	1	0	0	0	1	1		
1184	161529073	YP_001582899.1	alcohol dehydrogenase	361	121	200	0	1	1	1	1	1	1		B/T/EA/E
1185	161529074	YP_001582900.1	RNA-binding protein	183	119	31	0	1	1	0	0	1	0	M3	B=>T // B=>A
1186	161529075	YP_001582901.1	prephenate dehydratase	271	151	200	0	1	1	1	1	1	1		
1187	161529077	YP_001582903.1	inosine-5p-monophosphate dehydrogenase	476	415	200	0	1	0	1	0	1	1		
1188	161529078	YP_001582904.1	dihydropteroate synthase	281	140	200	0	1	1	0	1	1	1		
1189	161529079	YP_001582905.1	hypothetical protein Nmar_1571	244	106	88	1	1	1	1	1	0	0	M2	
1190	161529080	YP_001582906.1	hypothetical protein Nmar_1572	183	148	14	1	0	1	1	0	0	0		
1191	161529081	YP_001582907.1	bifunctional GMP synthase/glutamine amidotransferase protein	507	403	200	0	1	1	1	0	1	1		
1192	161529082	YP_001582908.1	hypothetical protein Nmar_1574	118	98	12	1	0	0	0	0	0	0		
1193	161529083	YP_001582909.1	hypothetical protein Nmar_1575	86	84	9	0	0	0	0	0	0	1		T<=>B
1194	161529084	YP_001582910.1	ROK family protein	289	116	200	0	1	0	0	0	0	1		B=>T
1195	161529085	YP_001582911.1	thermosome	540	484	200	1	1	1	1	1	0	0	M	
1196	161529086	YP_001582912.1	adenine deaminase	571	291	133	0	1	1	0	0	1	1		B=>A
1197	161529087	YP_001582913.1	50S ribosomal protein L44e	93	95	200	0	0	1	1	1	1	0	M	
1198	161529088	YP_001582914.1	ribosomal protein S27E	65	56	43	1	1	1	1	0	0	0	M	
1199	161529089	YP_001582915.1	5-deoxyadenosylcobinamide phosphate nucleotidyltransferase	194	90	64	0	1	1	0	0	0	1	M3	A=>B
1200	161529090	YP_001582916.1	cobalamin 5p-phosphate synthase	241	141	46	0	1	0	0	0	0	1	M3	
1201	161529091	YP_001582917.1	cobalamin biosynthesis protein CobD	323	126	200	0	1	1	0	0	1	1		
1202	161529092	YP_001582918.1	cobyrinic acid ac-diamide synthase	279	289	200	0	1	0	0	0	0	1		EA=>T
1203	161529093	YP_001582919.1	aminotransferase class I and II	359	115	200	0	1	1	1	0	1	1		
1204	161529094	YP_001582920.1	aspartate-semialdehyde dehydrogenase	358	191	200	0	1	1	1	1	1	1		
1205	161529095	YP_001582921.1	AsnC family transcriptional regulator	124	115	66	0	1	1	0	0	0	1		A<=>B
1206	161529096	YP_001582922.1	protoheme IX farnesyltransferase	309	201	200	0	1	1	1	1	0	1		
1207	161529099	YP_001582925.1	hypothetical protein Nmar_1591	581	43	10	0	0	1	1	0	0	1		
1208	161529100	YP_001582926.1	hypothetical protein Nmar_1592	127	117	5	1	0	1	0	0	0	0		
1209	161529102	YP_001582928.1	hypothetical protein Nmar_1594	221	206	7	1	0	0	0	0	0	0		
1210	161529104	YP_001582930.1	thioesterase superfamily protein	161	72	200	0	1	0	1	0	1	1		
1211	161529105	YP_001582931.1	UbiD family decarboxylase	443	250	200	0	1	1	1	1	1	1	M3	
1212	161529106	YP_001582932.1	hypothetical protein Nmar_1598	113	112	4	1	0	0	0	0	0	0		
1213	161529107	YP_001582933.1	hypothetical protein Nmar_1599	148	70	6	1	0	0	0	0	0	0		
1214	161529108	YP_001582934.1	hypothetical protein Nmar_1600	220	161	5	1	1	0	0	0	0	0		
1215	161529110	YP_001582936.1	ubiquitin-associated domain-containing protein	113	100	29	1	1	1	1	0	0	0	M3	
1216	161529111	YP_001582937.1	PUA domain-containing protein	157	113	75	1	1	1	1	0	0	0	M2	
1217	161529112	YP_001582938.1	phosphate uptake regulator PhoU	348	88	147	1	1	1	1	1	0	0	M2	
1218	161529113	YP_001582939.1	GTP1/OBG protein	371	201	200	0	1	1	1	0	1	1	M	
1219	161529114	YP_001582940.1	hypothetical protein Nmar_1606	177	133	93	1	1	1	1	0	0	0	M2	EA=>T
1220	161529115	YP_001582941.1	transcription factor TFIIE alpha subunit	171	90	84	1	1	1	1	0	0	0	M2	
1221	161529116	YP_001582942.1	aldehyde dehydrogenase	443	276	200	1	0	0	0	0	1	1		
1222	161529118	YP_001582944.1	hemerythrin HHE cation binding domain-containing protein	185	88	32	0	1	1	1	0	0	1		A<=>B
1223	161529120	YP_001582946.1	deaminase-reductase domain-containing protein	218	95	200	0	1	1	0	0	1	1		
1224	161529121	YP_001582947.1	amidohydrolase	396	127	200	0	1	1	1	0	1	1	M3	
1225	161529122	YP_001582948.1	GTP cyclohydrolase IIA	251	143	48	1	1	1	0	0	0	0	M	
1226	161529123	YP_001582949.1	6 7-dimethyl-8-ribityllumazine synthase	138	98	200	0	1	1	0	1	1	1	M3	
1227	161529124	YP_001582950.1	3 4-dihydroxy-2-butanone 4-phosphate synthase	221	147	200	0	1	1	0	0	1	1	M2	
1228	161529125	YP_001582951.1	riboflavin synthase alpha subunit	197	112	200	0	1	0	0	0	1	1		B=>T // B=>EA
1229	161529126	YP_001582952.1	TPR repeat-containing protein	241	28	200	0	1	0	0	0	1	1		
1230	161529127	YP_001582953.1	response regulator receiver protein	116	62	200	0	1	0	0	0	1	1		
1231	161529129	YP_001582955.1	Cityl-CoA lyase	284	128	200	0	1	1	0	0	1	0		B=>T // B=>A
1232	161529130	YP_001582956.1	alcohol dehydrogenase	358	132	200	0	1	1	1	1	1	1		
1233	161529131	YP_001582957.1	DNA-binding TFAR19-related protein	93	79	113	0	1	1	1	1	1	0	M	
1234	161529132	YP_001582958.1	30S ribosomal protein S19e	150	124	200	0	1	1	1	1	1	0	M	
1235	161529133	YP_001582959.1	ribosome biogenesis protein	222	122	67	0	1	1	1	1	1	0	M2	
1236	161529135	YP_001582961.1	ribonucleoside-diphosphate reductase adenosylcobalamin-dependent	884	303	200	0	1	1	1	1	1	1		A<=>B
1237	161529136	YP_001582962.1	regulatory protein ArsR	723	459	56	0	1	0	1	0	0	1	M2	
1238	161529137	YP_001582963.1	hypothetical protein Nmar_1629	57	53	7	1	0	0	0	0	0	0		
1239	161529138	YP_001582964.1	hypothetical protein Nmar_1630	128	110	28	1	1	1	1	0	0	0		
1240	161529139	YP_001582965.1	Propanoyl-CoA C-acyltransferase	387	231	200	0	1	1	0	0	1	1		EA=>T
1241	161529140	YP_001582966.1	DNA topoisomerase type IA Zn finger domain-containing protein	152	152	8	1	0	0	0	0	0	1		
1242	161529141	YP_001582967.1	Nitrilase/cyanide hydratase and apolipoprotein N-acyltransferase	268	139	200	0	1	1	0	0	1	1		B=>T
1243	161529142	YP_001582968.1	4-oxalocrotonate tautomerase	58	58	4	1	0	0	0	0	0	0		
1244	161529143	YP_001582969.1	rhomboid family protein	252	144	112	1	0	1	1	1	0	1		
1245	161529144	YP_001582970.1	hypothetical protein Nmar_1636	517	79	47	1	0	0	0	0	0	0		
1246	161529145	YP_001582971.1	blue (type1) copper domain-containing protein	623	0	1									
1247	161529146	YP_001582972.1	hypothetical protein Nmar_1638	623	66	54	1	0	0	0	0	0	0		
1248	161529147	YP_001582973.1	hypothetical protein Nmar_1639	443	81	18	1	0	0	0	0	0	0		
1249	161529149	YP_001582975.1	von Willebrand factor type A	316	93	128	0	1	0	0	0	1	1		
1250	161529150	YP_001582976.1	hypothetical protein Nmar_1642	290	56	200	0	1	0	0	0	0	1		
1251	161529151	YP_001582977.1	ATPase	340	211	200	0	1	1	0	0	0	1		
1252	161529152	YP_001582978.1	hypothetical protein Nmar_1644	386	80	41	1	0	0	0	0	0	0		
1253	161529153	YP_001582979.1	ArsR family transcriptional regulator	166	83	30	1	1	0	0	0	0	0		
1254	161529155	YP_001582981.1	hypothetical protein Nmar_1647	531	72	47	1	0	0	0	0	0	0		
1255	161529156	YP_001582982.1	hypothetical protein Nmar_1648	572	77	48	1	0	0	0	0	0	0		
1256	161529157	YP_001582983.1	hypothetical protein Nmar_1649	275	118	31	0	0	0	0	1	0	1		
1257	161529158	YP_001582984.1	hypothetical protein Nmar_1650	270	78	30	1	1	0	0	0	0	0		
1258	161529159	YP_001582985.1	hypothetical protein Nmar_1651	230	211	7	1	0	0	0	0	0	0		
1259	161529160	YP_001582986.1	copper resistance D domain-containing protein	927	103	35	0	1	0	0	0	1	1		
1260	161529161	YP_001582987.1	ABC-3 protein	272	174	200	0	1	0	1	0	1	1		B=>T // B=>A
1261	161529162	YP_001582988.1	ABC transporter related	240	121	200	0	1	0	1	0	0	1		B=>T // B=>A
1262	161529163	YP_001582989.1	periplasmic solute binding protein	306	106	200	0	1	0	1	0	0	1		B=>T // B=>A
1263	161529164	YP_001582990.1	CopG family transcriptional regulator	127	95	86	0	1	1	0	1	0	1		EA<=>T
1264	161529166	YP_001582992.1	DSBA oxidoreductase	263	36	200	0	1	0	0	0	1	1		B=>T
1265	161529167	YP_001582993.1	UspA domain-containing protein	140	39	200	0	1	1	1	1	1	1		
1266	161529168	YP_001582994.1	glycosyl hydrolase BNR repeat-containing glycosyl hydrolase	322	248	7	0	0	0	0	0	0	1		
1267	161529169	YP_001582995.1	HxlR family transcriptional regulator	130	50	200	0	1	1	1	0	1	1		
1268	161529170	YP_001582996.1	putative metal cation transporter	392	183	23	0	1	1	0	0	0	1		
1269	161529171	YP_001582997.1	multicopper oxidase type 3	353	80	200	0	1	0	1	0	1	1		B=>T
1270	161529172	YP_001582998.1	DtxR family iron dependent repressor	163	61	156	1	1	1	1	0	0	1		B=>T
1271	161529173	YP_001582999.1	blue (type1) copper domain-containing protein	167	51	39	0	1	0	0	0	0	1		
1272	161529174	YP_001583000.1	kelch repeat-containing protein	344	36	200	0	0	0	1	0	1	1		
1273	161529175	YP_001583001.1	hypothetical protein Nmar_1667	470	121	92	0	1	0	1	0	1	1		B=>T
1274	161529178	YP_001583004.1	DSBA oxidoreductase	240	49	200	0	1	0	0	0	1	1		B=>T
1275	161529179	YP_001583005.1	hypothetical protein Nmar_1671	524	258	18	0	1	0	0	0	0	1		
1276	161529181	YP_001583007.1	2-alkenal reductase	381	166	200	0	1	0	1	0	1	1		B=>A
1277	161529182	YP_001583008.1	transcriptional regulator TrmB	247	103	20	1	1	0	0	0	0	0		
1278	161529184	YP_001583010.1	hypothetical protein Nmar_1676	573	448	16	1	0	0	0	0	0	0		
1279	161529186	YP_001583012.1	blue (type1) copper domain-containing protein	336	45	45	0	1	0	1	0	0	1		
1280	161529187	YP_001583013.1	hypothetical protein Nmar_1679	208	216	6	1	0	0	0	0	0	0		
1281	161529188	YP_001583014.1	hypothetical protein Nmar_1680	293	276	15	0	1	0	0	0	0	1		T<=>EA
1282	161529189	YP_001583015.1	hypothetical protein Nmar_1681	121	99	88	0	1	0	0	0	1	1		
1283	161529190	YP_001583016.1	hypothetical protein Nmar_1682	109	99	87	0	1	0	0	0	1	1		
1284	161529191	YP_001583017.1	cell division control protein 6 family protein	439	158	200	0	1	1	1	1	1	0	M2	
1285	161529192	YP_001583018.1	hypothetical protein Nmar_1684	457	424	17	1	0	0	0	0	0	0		
1286	161529194	YP_001583020.1	aspartate carbamoyltransferase	309	220	200	0	1	1	1	1	1	1	M	
1287	161529195	YP_001583021.1	aspartate carbamoyltransferase regulatory subunit	153	103	126	0	1	1	1	1	0	1	M3	A<=>B
1288	161529197	YP_001583023.1	V-type ATPase D subunit	209	114	131	0	1	1	1	0	1	1	M3	A<=>B
1289	161529198	YP_001583024.1	V-type ATP synthase subunit B	461	443	200	0	1	1	1	1	1	1	M2	A=>B
1290	161529199	YP_001583025.1	H+transporting two-sector ATPase alpha/beta subunit central region	592	455	200	0	1	1	1	1	1	1	M2	A=>B
1291	161529200	YP_001583026.1	H+transporting two-sector ATPase E subunit	198	135	36	0	1	1	1	0	1	0		A=>EK
1292	161529201	YP_001583027.1	V-type ATPase 116 kDa subunit	699	198	73	0	1	1	1	0	1	1		
1293	161529203	YP_001583029.1	putative GAF sensor protein	330	241	15	0	1	0	0	0	0	1		
1294	161529204	YP_001583030.1	nitroreductase	213	58	200	0	1	1	0	0	1	1		B=>A
1295	161529205	YP_001583031.1	rhodanese domain-containing protein	258	108	200	0	1	1	1	0	1	1		B=>T // B=>A
1296	161529206	YP_001583032.1	ammonium transporter	521	308	200	0	1	1	0	0	1	1		B=>T // B=>A
1297	161529208	YP_001583034.1	H(+)-transporting two-sector ATPase	344	331	17	1	1	1	1	0	0	0		
1298	161529212	YP_001583038.1	hypothetical protein Nmar_1704	116	104	6	1	0	0	1	0	0	0		
1299	161529213	YP_001583039.1	uridylate kinase putative	227	126	200	0	1	1	1	1	1	1	M	
1300	161529214	YP_001583040.1	metal dependent phosphohydrolase	411	112	200	0	1	1	1	1	1	1		
1301	161529215	YP_001583041.1	thymidylate kinase	193	75	170	0	1	1	1	0	1	1		A=>B
1302	161529216	YP_001583042.1	heat shock protein Hsp20	126	83	19	0	1	1	0	0	0	1		
1303	161529217	YP_001583043.1	hypothetical protein Nmar_1709	514	225	200	0	1	1	0	0	1	1		
1304	161529218	YP_001583044.1	hypothetical protein Nmar_1710	127	121	5	1	1	0	0	0	0	0		
1305	161529220	YP_001583046.1	hypothetical protein Nmar_1712	430	415	5	1	0	1	0	0	0	0		
1306	161529221	YP_001583047.1	ABC transporter related	207	142	200	0	1	1	1	0	1	1		
1307	161529222	YP_001583048.1	hypothetical protein Nmar_1714	395	167	33	0	1	1	0	0	0	1		
1308	161529223	YP_001583049.1	AsnC family transcriptional regulator	300	93	39	0	1	1	0	0	0	1		
1309	161529224	YP_001583050.1	NUDIX hydrolase	171	62	200	0	1	1	0	1	0	1		B=>T // B=>A
1310	161529225	YP_001583051.1	phosphate uptake regulator PhoU	326	84	146	1	1	1	1	1	0	0	M2	
1311	161529226	YP_001583052.1	GHMP kinase	303	128	73	1	1	1	1	0	0	0	M2	EA=>CA
1312	161529227	YP_001583053.1	hypothetical protein Nmar_1719	254	201	84	1	1	1	0	0	0	0	M	
1313	161529228	YP_001583054.1	3-methyl-2-oxobutanoate hydroxymethyltransferase	279	201	200	0	1	1	0	0	1	1		
1314	161529229	YP_001583055.1	phosphopantothenoylcysteine decarboxylase/phosphopantothenate--cysteine ligase	414	222	200	0	1	1	0	0	0	1	M	
1315	161529230	YP_001583056.1	methyltransferase type 11	280	50	179	0	1	0	0	0	1	1		
1316	161529232	YP_001583058.1	peptidase M24	354	140	200	0	1	1	0	1	0	1		
1317	161529233	YP_001583059.1	tRNA (5-methylaminomethyl-2-thiouridylate)-methyltransferase	337	227	39	0	0	0	0	0	0	1		B=>T
1318	161529234	YP_001583060.1	hypothetical protein Nmar_1726	92	92	5	1	1	0	0	0	0	0		
1319	161529235	YP_001583061.1	beta-lactamase domain-containing protein	646	225	200	0	1	1	1	1	1	1	M	
1320	161529236	YP_001583062.1	proteasome endopeptidase complex	210	129	200	0	1	1	1	1	1	0	M	
1321	161529237	YP_001583063.1	3-dehydroquinate synthase	354	127	200	0	1	1	1	1	1	1	M2	
1322	161529238	YP_001583064.1	peptidylprolyl isomerase FKBP-type	248	89	161	0	1	1	1	1	0	1	M2	
1323	161529239	YP_001583065.1	Serine--pyruvate transaminase	381	225	200	0	1	1	1	0	1	1		
1324	161529240	YP_001583066.1	cytidyltransferase-like protein	164	99	133	0	1	1	1	1	0	1	M2	
1325	161529241	YP_001583067.1	RNA-3p-phosphate cyclase	338	125	177	0	1	1	1	1	1	1		B=>T+AA
1326	161529242	YP_001583068.1	hypothetical protein Nmar_1734	165	164	5	1	0	0	0	0	0	1		
1327	161529243	YP_001583069.1	hypothetical protein Nmar_1735	164	141	4	1	0	0	0	0	0	0		
1328	161529244	YP_001583070.1	cob(II)yrinic acid a c-diamide reductase	236	66	148	0	1	1	0	1	1	1		
1329	161529245	YP_001583071.1	NADPH-dependent FMN reductase	174	88	21	0	1	0	0	0	1	1		
1330	161529246	YP_001583072.1	hypothetical protein Nmar_1738	153	83	5	1	0	0	0	0	0	0		
1331	161529247	YP_001583073.1	hypothetical protein Nmar_1739	76	72	5	1	0	0	0	0	0	0		
1332	161529248	YP_001583074.1	peptidylprolyl isomerase	158	92	200	0	1	0	0	0	1	1		
1333	161529251	YP_001583077.1	hypothetical protein Nmar_1743	160	155	4	1	0	0	0	0	0	0		
1334	161529253	YP_001583079.1	ribokinase-like domain-containing protein	297	166	58	0	1	1	0	0	0	1		
1335	161529254	YP_001583080.1	dUTPase	115	109	6	0	0	0	0	0	0	1		
1336	161529255	YP_001583081.1	tetrahydromethanopterin S-methyltransferase 23 kD subunit	155	146	60	0	1	0	0	0	0	0	M3	
1337	161529256	YP_001583082.1	glutamine--scyllo-inositol transaminase	377	210	200	0	1	1	0	1	0	1		A=>B
1338	161529257	YP_001583083.1	hypothetical protein Nmar_1749	190	185	4	1	0	0	0	0	0	0		
1339	161529259	YP_001583085.1	tyrosyl-tRNA synthetase	348	186	200	0	1	1	1	1	1	0	M3	
1340	161529260	YP_001583086.1	precorrin-3B C17-methyltransferase	264	162	200	0	1	1	1	0	1	1		B=>A
1341	161529261	YP_001583087.1	hypothetical protein Nmar_1753	488	123	8	0	0	0	0	0	0	1		
1342	161529262	YP_001583088.1	aspartate kinase	467	236	200	0	1	1	1	1	1	1	M3	A=>B
1343	161529263	YP_001583089.1	proliferating cell nuclear antigen PcnA	248	103	200	0	1	1	1	1	1	0	M2	
1344	161529264	YP_001583090.1	transcription termination factor Tfs	105	52	137	0	1	1	1	1	1	0	M2	
1345	161529266	YP_001583092.1	hypothetical protein Nmar_1758	178	81	4	1	0	0	0	0	0	0		
1346	161529268	YP_001583094.1	dual specificity protein phosphatase	169	29	162	0	1	1	1	1	1	1		
1347	161529269	YP_001583095.1	cyclase family protein	213	54	200	0	1	0	1	0	1	1		
1348	161529270	YP_001583096.1	AsnC family transcriptional regulator	85	73	51	1	0	1	1	0	0	0		

1349	161529271	YP_001583097.1	aminotransferase class V	381	244	200	0	1	1	1	1	1	1		B=>T
1350	161529272	YP_001583098.1	hypothetical protein Nmar_1764	198	192	5	1	0	0	0	0	0	0		
1351	161529273	YP_001583099.1	4Fe-4S ferredoxin iron-sulfur binding domain-containing protein	99	99	10	1	0	0	0	0	0	0		
1352	161529274	YP_001583100.1	adenylate/guanylate cyclase with integral membrane sensor	591	82	35	0	1	0	0	0	0	1		
1353	161529275	YP_001583101.1	heat shock protein DnaJ domain-containing protein	384	234	7	0	1	0	0	0	1	0		
1354	161529276	YP_001583102.1	UspA domain-containing protein	140	42	200	0	1	1	0	1	1	1		
1355	161529277	YP_001583103.1	amino acid permease-associated region	439	257	200	0	1	1	0	0	1	1		
1356	161529278	YP_001583104.1	hypothetical protein Nmar_1770	218	206	7	1	0	0	0	0	0	0		
1357	161529279	YP_001583105.1	glutamine synthetase, type I	481	407	200	0	1	1	1	0	0	1		
1358	161529280	YP_001583106.1	lysine 2,3-aminomutase related protein	448	166	200	0	1	0	0	0	1	1		
1359	161529281	YP_001583107.1	MscS mechanosensitive ion channel	547	162	32	0	1	0	0	0	0	1		
1360	161529282	YP_001583108.1	ribose-phosphate pyrophosphokinase	292	217	200	0	1	1	1	0	1	1	M3	
1361	161529283	YP_001583109.1	putative RNA methylase	315	99	106	0	1	1	1	1	1	0		
1362	161529284	YP_001583110.1	ribonuclease Z	299	137	200	0	1	1	1	1	1	1	M3	A=>B
1363	161529285	YP_001583111.1	silent information regulator protein Sir2	242	101	200	0	1	1	0	1	1	1		
1364	161529287	YP_001583113.1	hypothetical protein Nmar_1779	134	88	43	1	1	1	1	0	0	0	M3	
1365	161529288	YP_001583114.1	dephospho-CoA kinase, CoaE	197	101	105	0	1	1	0	0	1	0		
1366	161529289	YP_001583115.1	2p-5p RNA ligase	180	57	176	0	1	1	1	1	0	1	M3	
1367	161529290	YP_001583116.1	tRNA cytidylyltransferase	443	193	98	1	1	1	1	1	0	0	M	
1368	161529291	YP_001583117.1	putative serine/threonine protein kinase	253	91	52	0	1	1	1	0	1	1	M3	
1369	161529293	YP_001583119.1	FAD-dependent pyridine nucleotide-disulphide oxidoreductase	382	75	200	0	1	1	1	1	1	1		
1370	161529294	YP_001583120.1	short-chain dehydrogenase/reductase SDR	246	136	200	0	1	1	0	0	1	1		
1371	161529295	YP_001583121.1	ANL-type Zinc finger protein	73	73	10	1	1	0	0	0	0	0		
1372	161529296	YP_001583122.1	hypothetical protein Nmar_1788	94	89	4	1	0	0	0	0	0	0		
1373	161529297	YP_001583123.1	hypothetical protein Nmar_1789	185	126	5	1	0	0	1	0	0	0		
1374	161529298	YP_001583124.1	glutamine synthetase, type I	490	409	200	0	1	1	1	0	0	1		
1375	161529300	YP_001583126.1	thermosome	570	481	200	1	1	1	1	1	0	0	M2	
1376	161529301	YP_001583127.1	serine hydroxymethyltransferase	440	313	200	0	1	1	1	1	1	1	M2	
1377	161529302	YP_001583128.1	DNA polymerase II large subunit	1125	900	75	1	1	0	1	1	0	0	M	
1378	161529303	YP_001583129.1	geranylgeranylglyceryl phosphate synthase-like protein	250	174	75	0	1	1	1	1	0	1	M2	
1379	161529304	YP_001583130.1	hypothetical protein Nmar_1796	120	119	4	1	0	0	0	0	0	0		
1380	161529305	YP_001583131.1	hypothetical protein Nmar_1797	200	123	6	1	0	0	1	0	0	0		
1381	161529307	YP_001583133.1	hypothetical protein Nmar_1799	128	125	5	1	0	0	1	0	0	0		

b.

	Sequence de départ				Alignement		Présence (1) /Absence (0) d'homologues dans les phylogénies							Observations	
							Archaea								
	Référence GI	Référence RefSeq	annotation	taille (aa)	Nombre de positions	Nombre de sequences	spéciale archée	Eury-archées	Cren-archées	Aig-archée	Kor-archée	Eucarya	Bacteria	Marqueur potentiel	THG potentiel
1	118575216	YP_874959.1	copper binding protein plastocyanin/azurin family	154	94	21	1	1	0	1	0	0	0		T<=>EA
2	118575219	YP_874962.1	prefoldin chaperonin cofactor	172	111	37	1	1	1	1	0	0	0	M2	
3	118575223	YP_874966.1	nucleic acid-binding protein	195	85	85	0	1	1	1	0	1	0	M2	
4	118575232	YP_874975.1	ribonuclease HI	117	118	6	0	0	0	0	0	0	1		
5	118575236	YP_874979.1	hypothetical protein CENSYa_0024	267	103	7	1	0	0	0	0	0	0		
6	118575238	YP_874981.1	hypothetical protein CENSYa_0026	183	169	4	1	0	0	0	0	0	0		
7	118575239	YP_874982.1	hypothetical protein CENSYa_0027	119	118	4	1	0	0	0	0	0	0		
8	118575244	YP_874987.1	copper binding protein plastocyanin/azurin family	210	42	64	0	1	0	1	0	1	1		
9	118575245	YP_874988.1	ribose 5-phosphate isomerase	225	132	200	0	1	1	0	0	1	1		
10	118575246	YP_874989.1	hypothetical protein CENSYa_0035	195	178	4	1	0	0	0	0	0	0		
11	118575249	YP_874992.1	conserved hypothetical protein putative amidase	236	68	200	0	1	1	0	1	1	1		
12	118575253	YP_874996.1	hypothetical protein CENSYa_0043	311	134	98	0	1	0	0	0	1	1		B=>T+EA
13	118575254	YP_874997.1	hypothetical protein CENSYa_0044	276	59	175	0	1	1	0	0	0	1		T<=>EA
14	118575255	YP_874998.1	MoxR-like ATPases	335	211	200	0	1	1	0	0	1	1		T<=>EA
15	118575258	YP_875001.1	hypothetical protein CENSYa_0048	154	53	53	1	1	0	0	0	0	0		T<=>EA
16	118575259	YP_875002.1	hypothetical protein CENSYa_0049	502	426	12	1	0	0	0	0	0	0		
17	118575260	YP_875003.1	signal peptide protein	519	50	65	0	0	0	0	0	0	1		T<=>B
18	118575261	YP_875004.1	cytochrome c biogenesis protein	381	25	200	0	0	0	0	0	1	1		B=>T
19	118575263	YP_875006.1	Mg-dependent DNase	249	93	200	0	1	1	1	1	1	1	M3	A<=>B // A<=>E
20	118575271	YP_875014.1	hypothetical membrane protein	127	119	9	1	0	0	0	0	0	0		
21	118575272	YP_875015.1	pyridoxine biosynthesis protein/glutamine amidotransferase	201	109	200	0	1	1	0	1	1	1		
22	118575281	YP_875024.1	transcriptional regulator	79	68	72	0	1	1	1	0	0	1	M3	
23	118575283	YP_875026.1	phenylalanyl-tRNA synthetase beta subunit	403	251	200	0	1	1	1	1	1	1	M	A=>B // T=> B
24	118575284	YP_875027.1	phenylalanyl-tRNA synthetase beta subunit	113	379	64	1	1	1	1	0	0	0	M	
25	118575285	YP_875028.1	phenylalanyl-tRNA synthetase alpha subunit	457	244	200	0	1	1	1	1	1	1	M	A=>B
26	118575288	YP_875031.1	pantetheine-phosphate nucleotidyltransferase	152	76	125	0	1	1	1	0	1	0	M2	
27	118575296	YP_875039.1	type I membrane protein	209	199	4	1	0	0	0	0	0	0		
28	118575297	YP_875040.1	hypothetical protein CENSYa_0091	118	117	4	1	0	1	0	0	0	0		
29	118575300	YP_875043.1	cysteine synthase	627	216	25	0	0	0	0	0	1	1		
30	118575301	YP_875044.1	hypothetical protein CENSYa_0095	521	520	5	0	0	0	0	0	0	1		
31	118575304	YP_875047.1	exosome subunit	137	141	5	1	1	0	0	0	0	0		
32	118575306	YP_875049.1	seryl-tRNA synthetase	418	302	200	0	1	1	1	1	1	1	M3	
33	118575308	YP_875051.1	single-stranded DNA-specific exonuclease	443	214	109	1	1	1	1	1	0	0	M	
34	118575313	YP_875056.1	glycosyltransferase involved in cell wall biogenesis	1069	37	200	0	1	1	1	1	0	1		
35	118575317	YP_875060.1	hypothetical protein CENSYa_0111	111	109	5	1	0	0	0	0	0	0		
36	118575320	YP_875063.1	5-formyltetrahydrofolate cyclo-ligase	173	59	200	0	1	0	0	0	1	1		B=>T
37	118575326	YP_875069.1	transcriptional regulator	104	75	39	1	0	1	1	0	0	0	M2	
38	118575329	YP_875072.1	hypothetical protein CENSYa_0123	55	53	7	1	1	1	0	0	0	0		
39	118575330	YP_875073.1	glutathione synthase/ribosomal protein S6 modification enzyme (glutaminyl transferase)	282	120	200	0	1	1	1	0	1	1	M3	
40	118575332	YP_875075.1	diphthamide biosynthesis methyltransferase	345	180	190	0	1	1	1	0	1	0	M2	
41	118575334	YP_875077.1	transcriptional regulator of a riboflavin/FAD biosynthetic operon	193	86	86	1	1	1	1	1	0	0	M2	
42	118575352	YP_875095.1	bacterial-type DNA primase	274	348	12	1	0	1	1	0	0	0		
43	118575354	YP_875097.1	bacterial-type DNA primase	194	338	14	1	1	1	1	0	0	0		
44	118575365	YP_875108.1	hypothetical protein CENSYa_0159	1680	93	24	1	0	1	0	0	0	0		
45	118575367	YP_875110.1	hypothetical protein CENSYa_0161	1737	100	24	1	0	1	0	0	0	0		
46	118575369	YP_875112.1	bacterial-type DNA primase	509	348	10	1	1	1	1	1	0	0		
47	118575374	YP_875117.1	GMP synthase glutamine amidotransferase domain	210	59	200	0	1	1	0	0	1	1		B=>T
48	118575381	YP_875124.1	20S proteasome alpha and beta subunit	238	174	200	0	1	1	1	1	1	0	M	
49	118575384	YP_875127.1	Sec-independent protein secretion pathway component	112	101	4	1	1	0	0	0	0	0		
50	118575386	YP_875129.1	helicase	820	222	200	0	1	1	0	1	1	1	M2	
51	118575388	YP_875131.1	DNA topoisomerase IB	519	185	128	0	0	0	1	0	1	1		
52	118575390	YP_875133.1	hypothetical protein CENSYa_0185	398	110	6	1	0	1	0	0	0	0		
53	118575391	YP_875134.1	hypothetical protein CENSYa_0186	263	81	94	0	0	0	0	0	1	1		B=>T
54	118575394	YP_875137.1	glycerol uptake facilitator	248	153	200	0	1	0	0	0	1	1		
55	118575399	YP_875142.1	histone acetyltransferase	269	185	14	1	1	1	0	0	0	0		
56	118575403	YP_875146.1	thiamine biosynthesis enzyme	368	221	200	0	1	1	0	0	1	1		
57	118575406	YP_875149.1	3-dehydroquinate dehydratase	217	72	105	0	1	1	1	1	1	1		
58	118575407	YP_875150.1	shikimate 5-dehydrogenase	268	133	200	0	1	0	0	1	1	1		
59	118575408	YP_875151.1	archaeal shikimate kinase	281	139	76	1	1	1	1	1	0	0	M2	
60	118575409	YP_875152.1	5-enolpyruvylshikimate-3-phosphate synthase	414	240	200	0	1	1	1	1	1	1		
61	118575412	YP_875155.1	Rad3-related DNA helicase	559	262	41	0	1	0	0	0	0	1		T<=>EA
62	118575417	YP_875160.1	aspartate/tyrosine/aromatic aminotransferase	449	199	200	0	1	1	1	1	0	1		
63	118575419	YP_875162.1	translation initiation factor eIF 2B alpha subunit	101	103	7	0	1	0	0	0	0	1		
64	118575425	YP_875168.1	hydroxymethylpyrimidine phosphate kinase	438	149	200	0	1	1	1	1	1	1		
65	118575426	YP_875169.1	small membrane protein	135	134	5	1	0	0	0	0	0	0		
66	118575427	YP_875170.1	protein-disulfide isomerase	212	47	173	0	1	0	0	0	1	1		B=>T // B=>EA
67	118575432	YP_875175.1	hypothetical protein CENSYa_0228	128	128	4	1	0	0	0	0	0	0		
68	118575434	YP_875177.1	3-hydroxyisobutyrate dehydrogenase	282	149	200	0	1	1	0	0	1	1		
69	118575435	YP_875178.1	hypothetical protein CENSYa_0231	207	110	200	0	1	1	1	0	1	1		B=>T
70	118575437	YP_875180.1	HD superfamily hydrolase	268	77	142	0	1	1	1	0	1	1		A<=>B
71	118575442	YP_875185.1	Asp-tRNAAsn/Glu-tRNAGln amidotransferase A subunit	479	360	200	0	1	1	0	0	1	1		B=>A // A<=>A
72	118575444	YP_875187.1	aspartyl/asparaginyl-tRNA synthetase	522	260	200	0	1	1	1	1	1	1	M	
73	118575460	YP_875203.1	sugar kinase	326	122	200	0	1	1	1	1	1	1	M2	
74	118575463	YP_875206.1	hypothetical protein CENSYa_0260	143	113	14	0	0	0	0	0	0	1		B<=>T
75	118575468	YP_875211.1	peptide methionine sulfoxide reductase	125	89	200	0	1	0	0	0	1	1		B=>T
76	118575469	YP_875212.1	thymidylate synthase	526	345	21	0	1	0	0	0	0	1		T<=>EA // A<=>B
77	118575471	YP_875214.1	hypothetical protein CENSYa_0268	273	197	17	1	0	1	0	0	0	0		
78	118575474	YP_875217.1	hypothetical protein CENSYa_0273	110	113	9	1	0	0	0	0	0	0		
79	118575483	YP_875226.1	leucyl aminopeptidase	488	247	200	0	0	1	0	0	1	1		B=>A
80	118575485	YP_875228.1	metal-dependent hydrolase	209	58	200	0	1	1	0	0	1	1		
81	118575486	YP_875229.1	transcriptional regulator	86	75	45	1	1	0	1	0	0	0	M	
82	118575487	YP_875230.1	cysteine desulfurase/selenocysteine lyase	374	239	200	0	1	1	1	1	1	1		B=>T
83	118575488	YP_875231.1	hypothetical protein CENSYa_0288	187	192	5	1	0	0	0	0	0	0		
84	118575491	YP_875234.1	phosphomannomutase	423	206	200	0	1	1	1	1	0	1	M2	A=>B
85	118575497	YP_875240.1	flavin-nucleotide-binding protein	123	97	15	0	0	0	0	0	0	1		T<=>B
86	118575498	YP_875241.1	asparagine synthase (glutamine-hydrolyzing)	383	269	200	0	1	1	1	0	1	1		
87	118575501	YP_875244.1	DNA-directed RNA polymerase subunit K/omega	155	70	41	0	1	1	1	0	1	0	M3	
88	118575504	YP_875247.1	hypothetical protein CENSYa_0306	260	170	108	0	1	0	1	1	0	1	M3	
89	118575506	YP_875249.1	hypothetical protein CENSYa_0308	168	172	6	1	0	0	0	0	0	0		
90	118575511	YP_875254.1	2 lactate dehydrogenase	348	190	200	0	1	1	0	1	1	1		
91	118575512	YP_875255.1	hypothetical protein CENSYa_0314	133	123	4	1	0	0	0	0	0	0		
92	118575513	YP_875256.1	hypothetical protein CENSYa_0315	93	86	7	1	0	0	0	0	0	0		
93	118575517	YP_875260.1	hypothetical protein CENSYa_0320	145	147	7	1	0	0	0	0	0	0		
94	118575521	YP_875264.1	archaeal adenylate kinase	178	135	66	1	1	1	1	0	0	0	M2	
95	118575524	YP_875267.1	ribosomal protein L30/L7E	155	134	98	1	1	1	1	1	0	0	M	
96	118575529	YP_875272.1	prefoldin molecular chaperone	173	129	18	1	1	0	0	0	0	0		
97	118575533	YP_875276.1	superoxide dismutase	206	150	200	0	1	0	1	0	1	1		
98	118575537	YP_875280.1	methylated DNA-protein cysteine methyltransferase	104	59	107	0	1	1	0	0	1	1		
99	118575538	YP_875281.1	hypothetical protein CENSYa_0341	258	103	94	0	1	1	1	1	1	0	M2	A<=>E
100	118575542	YP_875285.1	transcriptional regulator	159	114	47	0	1	1	1	0	0	1		
101	118575544	YP_875287.1	arginyl-tRNA synthetase	607	220	200	0	1	1	1	1	1	1	M	A=>B
102	118575545	YP_875288.1	hydrolase of the HAD superfamily	267	91	88	0	1	1	0	0	0	1	M2	A=>B
103	118575549	YP_875292.1	collagen type XI alpha 2	468	189	18	0	0	0	0	0	1	1		
104	118575550	YP_875293.1	hypothetical protein CENSYa_0353	86	83	7	1	0	1	0	0	0	0		
105	118575552	YP_875295.1	hypothetical protein CENSYa_0355	221	111	36	1	0	0	0	0	0	0		
106	118575554	YP_875297.1	TPR repeat protein	369	40	200	0	1	0	0	0	1	1		
107	118575556	YP_875299.1	rhodanese-related sulfurtransferase	264	89	200	0	1	1	1	0	1	1		
108	118575559	YP_875302.1	hypothetical protein CENSYa_0362	236	215	9	0	1	0	0	0	0	0		
109	118575562	YP_875305.1	metal-dependent phosphoesterase	305	82	106	0	1	1	0	0	0	1	M2	A=>B // T=>B
110	118575563	YP_875306.1	hypothetical protein CENSYa_0366	184	166	108	0	1	0	1	1	0	1	M2	B=>T // B=>EA
111	118575567	YP_875310.1	hypothetical protein CENSYa_0370	211	111	32	1	0	0	0	0	0	0		
112	118575569	YP_875312.1	hypothetical protein CENSYa_0372	346	146	146	0	1	0	0	0	1	1		B=>T
113	118575574	YP_875317.1	ABC-type oligopeptide transport system ATPase component	318	231	200	0	1	1	1	1	0	1		B=>T
114	118575575	YP_875318.1	chromosome segregation ATPase	1175	283	200	0	1	0	1	1	1	1		A=>B
115	118575580	YP_875323.1	archaeal enzyme of ATP-grasp superfamily	235	150	46	0	1	1	1	0	1	0	M2	
116	118575581	YP_875324.1	archaeal DNA-binding protein	158	77	69	1	1	1	1	1	0	0	M	
117	118575582	YP_875325.1	hypothetical protein CENSYa_0385	81	80	5	1	0	0	0	0	0	0		
118	118575585	YP_875328.1	secreted protein	272	108	9	0	0	0	0	0	0	1		
119	118575586	YP_875329.1	hypothetical protein CENSYa_0389	116	100	10	1	0	0	0	0	0	0		
120	118575589	YP_875332.1	hypothetical protein CENSYa_0392	261	172	4	1	0	0	0	0	0	0		
121	118575600	YP_875343.1	site-specific DNA methylase	286	105	170	0	1	0	0	0	1	1		B=>T
122	118575608	YP_875351.1	Rossmann fold nucleotide-binding protein	168	101	69	0	1	0	1	0	1	1		
123	118575611	YP_875354.1	hypothetical protein CENSYa_0415	56	53	6	1	0	0	0	0	0	0		
124	118575612	YP_875355.1	hypothetical protein CENSYa_0416	95	92	8	1	0	0	0	0	0	0		
125	118575613	YP_875356.1	diadenosine tetraphosphate hydrolase	138	55	200	0	1	1	0	0	1	1		
126	118575614	YP_875357.1	DnaJ-class molecular chaperone	247	64	23	0	0	0	0	0	1	1		
127	118575616	YP_875359.1	ATPase of the PP-loop superfamily	268	110	168	0	1	1	1	0	1	1	M2	A=>B
128	118575618	YP_875361.1	pseudouridylate synthase	402	255	63	1	1	1	1	1	0	0	M2	
129	118575621	YP_875364.1	ribosomal protein S27AE	55	45	42	0	1	0	0	0	1	0	M	
130	118575622	YP_875365.1	hypothetical protein CENSYa_0426	204	131	4	0	0	0	0	0	1	0		
131	118575626	YP_875369.1	hypothetical protein CENSYa_0430	179	69	16	0	1	1	1	0	0	1		
132	118575630	YP_875373.1	histone acetyltransferase	139	80	114	0	1	0	0	0	1	1		
133	118575631	YP_875374.1	nucleoside-diphosphate-sugar pyrophosphorylase	215	136	200	0	1	1	1	0	0	1		
134	118575632	YP_875375.1	ABC-type Fe3 -hydroxamate transport system periplasmic component	299	151	68	0	1	0	1	0	1	1		
135	118575634	YP_875377.1	hypothetical protein CENSYa_0438	374	222	116	1	1	1	1	1	0	0	M2	
136	118575635	YP_875378.1	hypothetical protein CENSYa_0439	284	265	6	1	0	0	0	0	0	0		
137	118575642	YP_875385.1	ribosomal protein L30E	101	103	5	1	0	0	0	0	0	0		
138	118575643	YP_875386.1	O-methyltransferase	187	91	200	0	1	0	0	0	1	1		B=>T
139	118575656	YP_875399.1	hypothetical protein CENSYa_0461	130	98	10	1	0	0	0	0	0	0		
140	118575663	YP_875406.1	RNA-binding protein	191	128	96	1	1	1	1	1	0	0	M	
141	118575664	YP_875407.1	dimethyladenosine transferase (rRNA methylation)	221	122	200	0	1	1	0	1	1	1	M3	

142	118575665	YP_875408.1	methylase of polypeptide chain release factor	165	103	72	0	1	1	1	1	1	0	M	
143	118575667	YP_875410.1	Ca2 /Na antiporter	294	176	200	0	1	0	0	0	1	1		
144	118575672	YP_875415.1	periplasmic protein kinase ArgK	307	145	200	0	1	1	0	0	1	1		
145	118575673	YP_875416.1	hypothetical protein CENSYa_0478	129	129	4	1	0	0	0	0	0	0		
146	118575674	YP_875417.1	aminopeptidase N	846	287	200	0	1	1	0	0	1	1		
147	118575678	YP_875421.1	ABC-type branched-chain amino acid transport system periplasmic component	422	156	77	0	1	1	1	1	1	1		
148	118575682	YP_875425.1	hypothetical protein CENSYa_0487	370	323	4	1	0	0	0	0	0	0		
149	118575692	YP_875435.1	hypothetical protein CENSYa_0498	160	159	5	1	0	0	0	0	0	0		
150	118575694	YP_875437.1	protein-disulfide isomerase	246	42	200	0	1	1	0	0	1	1		
152	118575702	YP_875445.1	hypothetical protein CENSYa_0508	6862	92	50	0	1	0	0	0	1	1		
153	118575704	YP_875447.1	Zn-dependent alcohol dehydrogenase	335	151	200	0	1	1	0	0	1	1		
154	118575709	YP_875452.1	tRNA pseudouridine synthase D	395	141	147	0	1	1	0	1	1	1	M2	A=>B
155	118575710	YP_875453.1	streptogramin lyase	508	393	12	0	0	0	0	0	0	1		
156	118575714	YP_875457.1	dehydrogenase	373	164	101	0	1	1	1	1	0	1		
157	118575715	YP_875458.1	polyprenyltransferase	271	201	200	0	1	1	1	1	0	1		A<=>B
158	118575722	YP_875465.1	hypothetical protein CENSYa_0528	210	107	33	1	0	0	0	0	0	0		
159	118575724	YP_875467.1	copper binding protein plastocyanin/azurin family	281	101	10	1	0	0	0	0	0	0		
160	118575725	YP_875468.1	metal-dependent protease of the PAD1/JA B1 superfamily	132	81	29	0	1	1	1	1	0	1		T<=>EA
161	118575737	YP_875480.1	excinuclease nuclease subunit	531	321	200	0	1	0	0	0	1	1		T<=>EA
162	118575741	YP_875484.1	flavin-nucleotide-binding protein	210	78	37	0	1	0	0	1	0	1		T<=>B
163	118575742	YP_875485.1	flavin-dependent oxidoreductase	390	65	200	0	1	0	1	0	0	1		A<=>B
164	118575743	YP_875486.1	alcohol dehydrogenase class IV	314	160	177	0	1	1	0	0	1	1		B=>T
165	118575748	YP_875491.1	Fe-S oxidoreductase	574	284	74	1	1	1	1	1	0	0	M3	
166	118575751	YP_875494.1	ribosomal biogenesis protein/nucleolar protein	392	248	200	0	1	1	1	1	1	0	M	
167	118575752	YP_875495.1	fibrillarin-like rRNA methylase	201	192	200	0	1	1	1	1	1	0	M	
168	118575753	YP_875496.1	ribonuclease HII	205	72	200	0	1	1	0	1	1	1	M2	
169	118575754	YP_875497.1	site-specific DNA methylase	303	138	65	0	1	0	0	0	0	1		B=>T
170	118575755	YP_875498.1	N2 N2-dimethylguanosine tRNA methyltransferase	377	143	200	0	1	1	1	1	1	1	M2	
171	118575759	YP_875502.1	hypothetical protein CENSYa_0565	1566	17	200	0	1	0	0	0	1	1		
172	118575763	YP_875506.1	transcriptional activator adenine-specific DNA methyltransferase	216	111	72	0	0	0	0	0	1	1		B=>T
173	118575775	YP_875518.1	hypothetical protein CENSYa_0581	2393	182	39	0	1	0	0	0	0	1		
174	118575776	YP_875519.1	hypothetical protein CENSYa_0582	3077	78	52	0	1	0	0	0	1	1		
175	118575781	YP_875524.1	3-ketoacyl-CoA thiolase/acetyl-CoA acetyltransferase	367	225	200	0	1	1	1	0	1	1	M3	
176	118575782	YP_875525.1	flavin-dependent oxidoreductase	320	111	109	0	1	0	1	0	0	1	M3	
177	118575783	YP_875526.1	hypothetical protein CENSYa_0589	290	222	4	1	0	1	0	0	0	0		
178	118575784	YP_875527.1	glucose-6-phosphate isomerase	364	131	45	0	1	1	1	0	0	1		A=>B
179	118575786	YP_875529.1	hypothetical protein CENSYa_0592	223	122	29	1	0	0	0	0	0	0		
180	118575787	YP_875530.1	hypothetical protein CENSYa_0593	219	115	32	1	0	0	0	0	0	0		
181	118575789	YP_875532.1	transcriptional activator adenine-specific DNA methyltransferase	216	103	69	0	0	0	0	0	1	1		B=>T
182	118575792	YP_875535.1	uracil-DNA glycosylase	216	101	200	0	1	1	1	1	0	1	M2	A<=>B
183	118575793	YP_875536.1	3-methyladenine DNA glycosylase	182	85	151	0	1	1	1	1	1	1		B=>T
184	118575794	YP_875537.1	uncharacterized membrane-associated protein	210	154	17	1	1	1	1	1	0	0	M3	
185	118575802	YP_875545.1	flavin-nucleotide-binding protein	166	65	24	0	1	0	1	0	0	1		
186	118575810	YP_875553.1	ABC-type branched-chain amino acid transport system periplasmic component	423	171	59	0	1	1	1	1	0	1		
187	118575814	YP_875557.1	hypothetical protein CENSYa_0620	127	109	4	1	0	0	0	0	0	0		
188	118575822	YP_875565.1	site-specific DNA methylase	289	120	128	0	1	0	0	0	1	1		B=>T
189	118575823	YP_875566.1	L-iditol 2-dehydrogenase/threonine dehydrogenase	359	136	200	0	1	1	1	0	1	1		
190	118575831	YP_875574.1	hypothetical protein CENSYa_0637	1623	12	67	0	1	0	0	0	0	1		
191	118575842	YP_875585.1	Ser/Thr protein kinase	143	143	22	0	1	1	1	0	0	1		A=>B
192	118575843	YP_875586.1	hypothetical protein CENSYa_0649	137	74	15	0	1	0	0	0	0	1		B<=>T
193	118575848	YP_875591.1	hypothetical protein CENSYa_0654	1485	0	1									
194	118575865	YP_875608.1	hypothetical protein CENSYa_0671	205	175	9	0	0	0	0	0	1	1		
195	118575882	YP_875625.1	micrococcal nuclease-like protein	637	118	8	0	1	0	0	0	0	1		
196	118575956	YP_875699.1	hypothetical protein CENSYa_0762	210	113	31	1	0	0	0	0	0	0		
197	118575960	YP_875703.1	hypothetical protein CENSYa_0766	124	109	11	1	0	0	0	0	0	0		
198	118575961	YP_875704.1	hypothetical protein CENSYa_0767	120	91	4	1	0	0	0	0	0	0		
199	118575962	YP_875705.1	ICC-like phosphoesterase	230	93	71	1	1	1	1	1	0	0	M2	
200	118575973	YP_875716.1	translation initiation factor 1	185	84	86	1	1	1	1	0	1	0	M2	
201	118575974	YP_875717.1	serine/threonine protein kinase	224	142	200	0	1	1	0	1	1	0	M	
202	118575975	YP_875718.1	RNA-binding protein	194	134	89	0	1	1	1	1	1	0	M	
203	118575981	YP_875724.1	hypothetical protein CENSYa_0787	215	110	34	1	0	0	0	0	0	0		
204	118575986	YP_875729.1	adenine/guanine phosphoribosyltransferase	170	123	200	0	1	0	0	0	1	1		
205	118575987	YP_875730.1	hypothetical protein CENSYa_0793	710	698	5	1	0	0	0	0	0	0		
206	118575993	YP_875736.1	hydroxypyruvate reductase	419	228	171	0	1	1	1	1	1	1		B=>T
207	118576012	YP_875755.1	translation initiation inhibitor	157	113	159	0	0	0	0	0	1	1		B=>T
208	118576015	YP_875758.1	transcriptional regulator	385	54	99	0	1	1	1	0	0	1		
209	118576041	YP_875784.1	phosphoribosylaminoimidazole carboxylase (NCAIR synthetase)	378	185	200	0	1	1	0	0	1	1		
210	118576046	YP_875789.1	uncharacterized membrane-associated protein	203	108	128	0	1	1	0	0	0	1		
211	118576047	YP_875790.1	nucleoside-diphosphate-sugar pyrophosphorylase	219	133	200	0	1	1	1	0	1	1		B<=>A
212	118576053	YP_875796.1	ribosomal protein L4	271	189	200	0	1	1	1	1	1	0	M	
213	118576057	YP_875800.1	ribosomal protein S3	261	148	200	0	1	1	1	1	1	1	M	
214	118576062	YP_875805.1	ribosomal protein L24	167	108	62	0	1	1	1	1	1	0	M	
215	118576078	YP_875821.1	transcriptional regulator	173	164	8	1	1	1	0	0	0	0	0	
216	118576080	YP_875823.1	phosphoribosylaminoimidazole-succinocarboxamide (SAICAR) synthase	273	160	200	0	1	1	0	0	1	1		B=>A
217	118576089	YP_875832.1	hypothetical protein CENSYa_0896	213	109	37	1	0	0	0	0	0	0		
218	118576090	YP_875833.1	hypothetical protein CENSYa_0897	10044	73	51	0	1	0	0	0	1	1		
219	118576097	YP_875840.1	hypothetical protein CENSYa_0904	184	140	14	0	1	0	0	0	0	0		T=>EA
220	118576099	YP_875842.1	hypothetical protein CENSYa_0906	217	110	34	1	0	0	0	0	0	0		
221	118576102	YP_875845.1	Trk-type K transport system membrane component	609	36	78	0	1	1	0	0	0	1		A<=>B
222	118576103	YP_875846.1	hypothetical protein CENSYa_0910	253	269	5	1	0	0	0	0	0	0		
223	118576107	YP_875850.1	hypothetical protein CENSYa_0914	161	117	8	1	0	0	0	0	0	0		
224	118576109	YP_875852.1	hypothetical protein CENSYa_0916	310	262	9	1	0	1	0	0	0	0		
225	118576113	YP_875856.1	universal stress protein	139	53	114	0	1	1	0	1	1	1		A=>B
226	118576118	YP_875861.1	DnaJ-class molecular chaperone	400	180	54	0	1	1	0	0	1	1		
227	118576119	YP_875862.1	adenylate cyclase	550	47	70	0	1	0	0	0	0	1		B=>T
228	118576120	YP_875863.1	glutathione synthase/ribosomal protein S6 modification enzyme (glutaminyl transferase)	281	121	200	0	1	1	1	1	1	1	M3	
229	118576122	YP_875865.1	ABC-type nitrate/sulfonate/bicarbonate transport system periplasmic component	324	159	42	0	1	0	0	0	0	1		B<=>A
230	118576125	YP_875868.1	permease	381	337	12	0	1	0	0		0	1		T<=>EA
231	118576126	YP_875869.1	SAM dependent methyltransferase	270	201	13	0	0	0	0	0	0	1		B=>T
232	118576130	YP_875873.1	NTP pyrophosphohydrolase	201	73	200	0	1	1	0	0	1	1		B=>T // B=>EA
233	118576133	YP_875876.1	cell division GTPase	310	310	5	1	0	0	0	0	0	0		
234	118576136	YP_875879.1	phosphoglycerate dehydrogenase	310	261	200	0	1	1	1	0	1	1		
235	118576142	YP_875885.1	ribosomal protein L15E	152	171	200	0	1	1	1	1	1	0	M	
236	118576147	YP_875890.1	HEAT repeat	148	107	22	0	1	0	1	0	1	1		
237	118576148	YP_875891.1	ATPase involved in DNA repair	777	179	51	0	1	1	0	1	0	1		EA=>T // A<=>A
238	118576149	YP_875892.1	DNA repair exonuclease	417	68	107	0	1	1	1	1	0	1		
239	118576152	YP_875895.1	dinucleotide-utilizing enzyme	276	147	76	0	1	1	0	0	1	1		A=>B
240	118576154	YP_875897.1	nicotinate-nucleotide pyrophosphorylase (carboxylating)	271	155	200	0	1	1	0	0	1	1	M3	
241	118576159	YP_875902.1	hypothetical protein CENSYa_0966	151	82	81	1	1	1	1	1	0	0	M	
242	118576163	YP_875906.1	hypothetical protein CENSYa_0970	85	84	9	0	0	0	0	0	0	1		T<=>B
243	118576164	YP_875907.1	hemolysin	418	158	200	0	1	1	0	0	1	1		
244	118576166	YP_875909.1	hypothetical protein CENSYa_0973	164	785	11	1	0	0	0	0	0	0		
245	118576169	YP_875912.1	protein-disulfide isomerase	226	47	200	0	1	0	0	0	0	1		T<=>B
246	118576172	YP_875915.1	cobyrinic acid a c-diamide synthase/dethiobiotin synthetase	258	87	200	0	1	0	0	0	0	1		
247	118576173	YP_875916.1	adenosylmethionine-8-amino-7-oxononanoate aminotransferase	439	242	200	0	1	0	0	0	1	1		B=>T
248	118576175	YP_875918.1	8-amino-7-oxononanoate synthase	340	262	200	0	1	1	0	0	1	1		
249	118576176	YP_875919.1	hypothetical protein CENSYa_0983	222	112	33	0	0	0	0	0	0	0		
250	118576177	YP_875920.1	cytochrome c biogenesis protein	243	125	119	0	1	0	0	0	1	1		B=>T
251	118576178	YP_875921.1	thiol-disulfide isomerase	370	63	91	0	1	0	0	0	1	1		
252	118576179	YP_875922.1	conserved protein implicated in secretion	209	124	41	1	0	1	1	0	0	0	M	
253	118576182	YP_875925.1	hypothetical protein CENSYa_0989	137	97	65	1	1	1	0	1	0	0	M2	
254	118576183	YP_875926.1	hypothetical protein CENSYa_0990	248	116	29	1	0	0	0	0	0	0		
255	118576184	YP_875927.1	20S proteasome alpha and beta subunit	245	172	200	0	1	1	1	1	1	0	M	
256	118576185	YP_875928.1	hypothetical protein CENSYa_0992	265	135	111	0	1	1	1	0	1	0	M2	
257	118576195	YP_875938.1	glycosyltransferase	327	182	23	0	1	0	1	0	0	1		
258	118576196	YP_875939.1	nucleoside-diphosphate-sugar epimerase	299	137	200	0	1	1	1	1	1	1		
259	118576208	YP_875951.1	uncharacterized protein involved in tolerance to divalent cations	109	52	200	0	1	1	1	1	1	1		
260	118576209	YP_875952.1	RNA-binding protein	343	237	57	1	1	1	1	0	1	0	M	
261	118576215	YP_875958.1	hypothetical protein CENSYa_1022	140	135	6	1	0	0	0	0	0	0		
262	118576218	YP_875961.1	hypothetical protein CENSYa_1025	167	145	11	1	0	0	0	0	0	0		
263	118576222	YP_875965.1	hypothetical protein CENSYa_1030	214	107	35	1	0	0	0	0	0	0		
264	118576223	YP_875966.1	copper-binding protein	113	94	20	1	0	0	0	0	0	0		
265	118576225	YP_875968.1	hypothetical protein CENSYa_1033	465	74	18	1	0	0	0	0	0	0		
266	118576245	YP_875988.1	ABC-type Mn2 /Zn2 transport system permease component	277	215	71	0	1	0	1	0	0	1		B=>T
267	118576246	YP_875989.1	ABC-type Mn2 /Zn2 transport system ATPase component	261	111	200	0	1	0	1	0	0	1		
268	118576247	YP_875990.1	ABC-type metal ion transport system periplasmic component/surface adhesin	560	81	150	0	1	0	1	0	0	1		
269	118576248	YP_875991.1	transcriptional regulator	131	116	27	0	1	0	0	0	0	0		
270	118576249	YP_875992.1	5-methyltetrahydrofolate--homocysteine S-methyltransferase	162	200	6	0	1	0	0	0	0	1		B<=>T
271	118576251	YP_875994.1	secreted periplasmic Zn-dependent protease	279	78	46	1	0	0	0	0	0	0		
272	118576254	YP_875997.1	6-pyruvoyl-tetrahydropterin synthase	270	145	7	0	1	0	1	0	0	0		
273	118576256	YP_875999.1	transcriptional regulator	267	138	35	0	0	0	0	0	0	1		B<=>T
274	118576259	YP_876002.1	asparagine synthase (glutamine-hydrolyzing)	261	143	79	0	1	1	1	0	1	1	M2	
275	118576265	YP_876008.1	hypothetical protein CENSYa_1074	128	109	7	0	0	1	1	0	0	0		
276	118576266	YP_876009.1	hypothetical protein CENSYa_1075	103	98	4	1	0	0	0	0	0	0		
277	118576268	YP_876011.1	thiamine monophosphate kinase	400	145	159	0	1	1	1	1	0	1	M	
278	118576269	YP_876012.1	phosphomannomutase	421	218	200	0	1	1	1	1	0	1	M2	A=>B
279	118576272	YP_876015.1	transcriptional regulator	92	92	10	0	0	0	1	0	0	0		
280	118576290	YP_876033.1	hypothetical protein CENSYa_1100	226	112	31	1	0	0	0	0	0	0		
281	118576306	YP_876049.1	superfamily II helicase	604	512	26	1	1	1	1	1	0	0	M3	
282	118576307	YP_876050.1	hypothetical protein CENSYa_1117	358	274	4	1	0	0	0	0	0	0		
283	118576313	YP_876056.1	methyl-accepting chemotaxis protein	660	185	4	1	0	0	0	0	0	0		
284	118576314	YP_876057.1	hypothetical protein CENSYa_1124	228	152	9	0	0	0	1	0	0	0		
285	118576316	YP_876059.1	ferredoxin	309	98	40	0	1	1	1	0	0	0		A<=>A
286	118576326	YP_876069.1	DNA repair helicase	219	323	21	0	1	0	0	0	0	1		B=>T
287	118576327	YP_876070.1	hypothetical protein CENSYa_1138	296	66	124	0	1	0	1	0	1	1		

288	118576333	YP_876076.1	secreted periplasmic Zn-dependent protease	483	81	42	1	0	0	0	0	0	0		
289	118576344	YP_876087.1	histone acetyltransferase	522	391	200	0	1	1	1	1	1	1	M2	
290	118576348	YP_876091.1	flavin-nucleotide-binding protein	148	124	11	0	1	0	0	0	0	1		
291	118576355	YP_876098.1	TPR repeat protein	268	35	200	0	1	0	0	0	1	1		
292	118576356	YP_876099.1	hypothetical protein CENSYa_1170	231	231	13	1	0	0	0	0	0	0		
293	118576358	YP_876101.1	double stranded beta helix fold enzyme	120	140	11	1	0	0	0	0	0	0		
294	118576361	YP_876104.1	transcriptional regulator	254	161	19	0	0	0	0	0	0	1		T<=>B
295	118576362	YP_876105.1	triosephosphate isomerase	217	173	96	1	1	1	1	1	0	0	M	
296	118576363	YP_876106.1	TATA-box binding protein (TBP) component of TFIID and TFIIIB	182	140	106	0	1	1	1	1	1	0	M	
297	118576364	YP_876107.1	hypothetical protein CENSYa_1179	101	83	4	0	1	0	0	0	0	1		
298	118576370	YP_876113.1	1 4-dihydroxy-2-naphthoate octaprenyltransferase	292	156	56	0	1	1	1	0	1	1		
299	118576379	YP_876122.1	oxidoreductase	372	158	44	0	1	1	0	0	1	1		
300	118576382	YP_876125.1	short-chain alcohol dehydrogenase	584	102	200	0	1	1	0	0	1	1		
301	118576383	YP_876126.1	permease	249	248	6	1	0	0	0	0	0	0		
302	118576385	YP_876128.1	hypothetical protein CENSYa_1200	118	116	5	1	0	0	0	0	0	0		
303	118576386	YP_876129.1	Mn-dependent transcriptional regulator	122	62	148	0	1	1	1	1	0	1		B=>T
304	118576391	YP_876134.1	transcriptional regulator	111	134	17	1	1	1	1	1	0	0		
305	118576392	YP_876135.1	2-isopropylmalate synthase	398	255	200	0	1	1	1	0	1	1	M3	
306	118576393	YP_876136.1	transcriptional regulator	266	171	55	0	0	0	0	0	0	1		B=>T
307	118576394	YP_876137.1	hypothetical protein CENSYa_1209	164	97	4	1	0	0	0	0	0	0		
308	118576398	YP_876141.1	hypothetical protein CENSYa_1213	310	282	4	1	0	0	0	0	0	0		
309	118576401	YP_876144.1	hypothetical protein CENSYa_1216	177	195	10	0	1	0	0	0	0	1		
310	118576403	YP_876146.1	hypothetical protein CENSYa_1218	449	395	5	1	0	0	0	0	0	0		
311	118576404	YP_876147.1	hypothetical protein CENSYa_1219	304	111	33	1	0	0	0	0	0	0		
312	118576405	YP_876148.1	hypothetical protein CENSYa_1220	172	99	161	0	1	1	0	1	1	1	M	
313	118576406	YP_876149.1	hypothetical protein CENSYa_1221	1079	150	6	1	1	0	0	0	0	0		
314	118576412	YP_876155.1	ABC-type branched-chain amino acid transport system periplasmic component	416	113	124	0	1	1	1	1	1	1		
315	118576417	YP_876160.1	hypothetical protein CENSYa_1233	356	256	33	1	1	1	1	1	0	0	M2	
316	118576419	YP_876162.1	hypothetical protein CENSYa_1235	173	126	7	1	1	1	0	0	0	0		
317	118576420	YP_876163.1	glycosyltransferase involved in cell wall biogenesis	385	40	200	0	1	1	1	1	1	1		
318	118576422	YP_876165.1	trypsin-like serine protease	385	154	200	0	1	0	1	0	1	1		
319	118576423	YP_876166.1	thiamine biosynthesis enzyme	350	209	107	0	1	1	0	0	1	1		
320	118576438	YP_876181.1	alanyl-tRNA synthetase	894	415	200	0	1	1	1	1	1	1	M	
321	118576440	YP_876183.1	leucyl-tRNA synthetase	578	463	200	0	1	1	1	1	1	1	M	A=>B
322	118576441	YP_876184.1	hypothetical protein CENSYa_1257	224	233	15	0	0	0	0	0	1	0		T<=>E
323	118576442	YP_876185.1	hypothetical protein CENSYa_1258	506	48	70	0	1	0	0	0	1	1		T<=>E
324	118576449	YP_876192.1	hypothetical protein CENSYa_1265	112	106	7	1	0	0	0	0	0	0		
325	118576450	YP_876193.1	hypothetical protein CENSYa_1266	109	105	7	1	0	0	0	0	0	0		
326	118576451	YP_876194.1	hypothetical protein CENSYa_1267	237	228	8	0	0	0	0	0	0	1		T<=>B
327	118576453	YP_876196.1	hypothetical protein CENSYa_1269	208	173	11	1	1	1	1	1	0	0		
328	118576454	YP_876197.1	protein-disulfide isomerase	177	56	138	0	1	0	0	0	1	1		B=>T
329	118576459	YP_876202.1	tryptophan synthase alpha chain	259	143	200	0	1	0	0	0	1	1		B=>T
330	118576461	YP_876204.1	indole-3-glycerol phosphate synthase	250	165	200	0	1	0	0	0	1	1		
331	118576462	YP_876205.1	anthranilate phosphoribosyltransferase	343	190	200	0	1	1	0	0	1	1		B=>T
332	118576471	YP_876214.1	hypothetical protein CENSYa_1287	126	168	7	1	0	0	0	0	0	0		
333	118576473	YP_876216.1	asymmetrical bis(5p-nucleosyl)-tetraphosphatase	141	46	185	0	1	1	0	1	1	1		
334	118576475	YP_876218.1	hypothetical protein CENSYa_1291	128	123	4	1	0	0	0	0	0	0		
335	118576481	YP_876224.1	protein-tyrosine phosphatase	166	58	36	0	1	1	1	1	1	1		
336	118576487	YP_876230.1	hypothetical protein CENSYa_1303	431	427	9	1	0	0	0	0	0	0		
337	118576490	YP_876233.1	hypothetical protein CENSYa_1306	101	103	8	1	0	0	0	0	0	0		
338	118576491	YP_876234.1	ABC-type dipeptide/oligopeptide/nickel transport system permease component	431	174	200	0	1	1	1	1	1	1		
339	118576495	YP_876238.1	membrane protein	578	222	40	1	1	1	1	0	0	0	M2	
340	118576496	YP_876239.1	hypothetical protein CENSYa_1312	207	126	16	1	1	1	0	1	0	0		
341	118576497	YP_876240.1	solute-binding protein	821	166	50	0	1	1	1	0	0	1		A=>B
342	118576498	YP_876241.1	SAM dependent methyltransferase	248	138	24	0	1	0	0	0	1	1		A<=>B
343	118576501	YP_876244.1	dehydrogenase	262	291	37	1	1	1	1	0	0	0	M2	
344	118576507	YP_876250.1	short-chain alcohol dehydrogenase	252	118	200	0	1	0	0	1	1	1		
345	118576509	YP_876252.1	universal stress protein	148	87	30	0	1	1	0	1	1	1	M3	
346	118576510	YP_876253.1	acyltransferase	261	64	200	0	1	0	1	0	1	1		
347	118576512	YP_876255.1	hypothetical protein CENSYa_1328	141	134	4	1	0	0	0	0	0	0		
348	118576515	YP_876258.1	hypothetical protein CENSYa_1331	131	95	12	1	0	0	0	0	0	0		
349	118576516	YP_876259.1	transcription initiation factor TFIIB	328	250	200	0	1	1	1	1	1	0	M	
350	118576518	YP_876261.1	hypothetical protein CENSYa_1334	276	110	32	1	0	0	0	0	0	0		
351	118576519	YP_876262.1	TPR repeat protein	467	132	15	1	0	0	0	0	0	0		
352	118576521	YP_876264.1	biotin-(acetyl-CoA carboxylase) ligase	323	82	200	0	1	1	0	0	0	1		B=>T // B=>A
353	118576522	YP_876265.1	hypothetical protein CENSYa_1338	172	72	9	1	0	0	0	0	0	0		
354	118576525	YP_876268.1	hypothetical protein CENSYa_1341	118	118	6	1	0	0	0	0	0	0		
355	118576528	YP_876271.1	cysteinyl-tRNA synthetase	455	328	200	0	1	1	0	0	1	1		
356	118576529	YP_876272.1	NH(3)-dependent NAD synthase	271	136	200	0	1	1	1	1	1	1	M2	
357	118576533	YP_876276.1	Fe-S cluster formation protein	135	84	184	0	1	1	1	1	1	1		A<=>B
358	118576538	YP_876281.1	uroporphyrinogen-III synthase/methyltransferase	259	83	115	0	1	1	0	1	1	1		
359	118576540	YP_876283.1	porphobilinogen deaminase	311	195	200	0	1	1	0	0	1	1	M3	
360	118576541	YP_876284.1	glutamate-1-semialdehyde aminotransferase	456	341	200	0	1	1	0	0	1	1	M3	
361	118576542	YP_876285.1	ABC-type branched-chain amino acid transport system periplasmic component	410	182	56	0	1	1	1	1	0	1		
362	118576543	YP_876286.1	permease	220	197	18	0	0	0	0	0	1	1		
363	118576544	YP_876287.1	short-chain alcohol dehydrogenase	276	107	200	0	1	0	0	0	1	1		B=>T
364	118576554	YP_876297.1	UDP-glucuronosyltransferase	338	314	6	0	1	0	0	0	0	1		
365	118576555	YP_876298.1	aspartokinase	471	201	200	0	1	1	1	1	1	1		
366	118576565	YP_876308.1	CDP-diacylglycerol--glycerol-3-phosphate 3-phosphatidyltransferase	190	109	131	0	1	1	1	1	0	1		
367	118576568	YP_876311.1	RNA-binding protein	176	171	11	1	0	1	0	0	0	0	M3	
368	118576570	YP_876313.1	nicotinamide mononucleotide adenylyltransferase	176	111	120	0	1	1	1	1	0	1	M2	
369	118576571	YP_876314.1	selenocysteine-specific translation elongation factor Tu domain 2	310	199	40	1	1	1	0	0	0	0	M	
370	118576572	YP_876315.1	hypothetical protein CENSYa_1388	233	211	4	1	0	0	0	0	0	0		
371	118576575	YP_876318.1	dioxygenase ferredoxin protein	111	93	14	0	1	0	0	0	0	1		A<=>B
372	118576582	YP_876325.1	single-stranded DNA-binding replication protein A large (70 kD) subunit	132	96	11	1	1	0	1	0	0	0		
373	118576584	YP_876327.1	hypothetical protein CENSYa_1401	85	84	4	1	0	0	0	0	0	0		
374	118576586	YP_876329.1	glutamyl-tRNA reductase	417	199	200	0	1	1	0	0	1	1	M3	
375	118576587	YP_876330.1	siroheme synthase/precorrin-2 oxidase	217	75	121	0	1	1	0	0	1	1		
376	118576594	YP_876337.1	signal recognition particle 19 kDa protein	101	97	5	1	1	0	0	0	0	0		
377	118576604	YP_876347.1	DNA-directed RNA polymerase subunit M/Transcription elongation factor TFIIS	103	53	159	0	1	1	1	1	1	0	M	
378	118576607	YP_876350.1	hypothetical protein CENSYa_1424	516	293	6	1	0	0	0	0	0	0		
379	118576608	YP_876351.1	DNA modification methylase	359	66	156	0	1	1	0	0	1	1		
380	118576612	YP_876355.1	transcriptional regulator	76	76	33	1	0	1	1	0	0	0	M	
381	118576614	YP_876357.1	pyridoxal phosphate-dependent enzyme	360	203	200	0	1	1	0	0	0	1		
382	118576615	YP_876358.1	tetrahydromethanopterin S-methyltransferase subunit A	141	165	51	0	1	0	0	0	0	1	M2	
383	118576621	YP_876364.1	peptidyl-prolyl cis-trans isomerase (rotamase)	159	89	200	0	1	0	0	0	1	1		T<=>B
384	118576630	YP_876373.1	archaeal/vacuolar-type H -ATPase subunit E	198	170	24	0	1	0	1	0	0	1		A=>B
385	118576632	YP_876375.1	hypothetical protein CENSYa_1449	224	112	33	1	0	0	0	0	0	0		
386	118576634	YP_876377.1	archaeal/vacuolar-type H -ATPase subunit I	691	202	61	0	1	1	1	0	0	1		A=>B
387	118576635	YP_876378.1	rhodanese-related sulfurtransferase	263	104	200	0	1	1	1	0	1	1		
388	118576638	YP_876381.1	lysine biosynthesis protein	55	54	8	1	1	1	0	0	0	0		
389	118576639	YP_876382.1	archaeal/vacuolar-type H -ATPase subunit C	346	325	14	1	1	1	1	0	0	0		
390	118576642	YP_876385.1	hypothetical protein CENSYa_1459	94	91	8	1	0	0	1	0	0	0		
391	118576650	YP_876393.1	hypothetical protein CENSYa_1467	180	206	7	1	0	0	0	0	0	0		
392	118576651	YP_876394.1	orotidine-5p-phosphate decarboxylase	242	127	98	0	1	1	1	1	1	1	M2	
393	118576652	YP_876395.1	diadenosine 5p5ppp-P1 P4-tetraphosphate pyrophosphohydrolase	171	84	47	0	1	1	1	0	1	1		
394	118576660	YP_876403.1	hypothetical protein CENSYa_1477	178	119	9	1	0	0	0	0	0	0		
395	118576664	YP_876407.1	hypothetical protein CENSYa_1481	161	128	4	1	0	0	0	0	0	0		
396	118576665	YP_876408.1	hypothetical protein CENSYa_1482	47	45	8	0	0	1	0	0	0	1		
397	118576670	YP_876413.1	3-isopropylmalate dehydratase small subunit	174	141	200	0	0	0	1	0	1	1		B=>T/AA
398	118576671	YP_876414.1	mannose-1-phosphate guanyltransferase	239	106	200	0	1	1	1	0	1	1		
399	118576674	YP_876417.1	ribosomal protein L18E	126	102	94	1	1	1	1	1	0	0	M	
400	118576681	YP_876424.1	SAM dependent methyltransferase	273	189	11	0	1	0	0	0	1	1		
401	118576682	YP_876425.1	phosphopantothenoylcysteine synthetase/decarboxylase	395	221	200	0	1	1	0	0	0	1	M	
402	118576683	YP_876426.1	3-methyl-2-oxobutanoate hydroxymethyltransferase	262	193	200	0	1	1	0	0	1	1		
403	118576684	YP_876427.1	hypothetical protein CENSYa_1501	252	201	84	1	1	1	0	0	0	0	M	
404	118576685	YP_876428.1	archaeal suger kinase	290	103	88	0	1	1	1	0	0	1	M	
405	118576686	YP_876429.1	phosphate uptake regulator	325	94	127	1	1	1	1	1	0	0	M2	
406	118576687	YP_876430.1	transcriptional regulator	314	93	49	0	1	1	0	0	0	1	M3	
407	118576689	YP_876432.1	ABC-type transport system involved in lipoprotein release ATPase component	226	153	200	0	1	1	1	1	1	1		B=>T
408	118576691	YP_876434.1	hypothetical protein CENSYa_1508	177	121	5	1	0	0	0	0	0	0		
409	118576693	YP_876436.1	molecular chaperone (small heat shock protein)	127	107	14	0	1	1	0	0	0	0		
410	118576698	YP_876441.1	hypothetical protein CENSYa_1515	198	118	28	1	0	0	0	0	0	0		
411	118576701	YP_876444.1	hypothetical protein CENSYa_1518	209	106	34	1	0	0	0	0	0	0		
412	118576702	YP_876445.1	hypothetical protein CENSYa_1519	218	114	31	1	0	0	0	0	0	0		
413	118576703	YP_876446.1	hypothetical protein CENSYa_1520	218	111	32	1	0	0	0	0	0	0		
414	118576704	YP_876447.1	hypothetical protein CENSYa_1521	219	112	33	1	0	0	0	0	0	0		
415	118576712	YP_876455.1	thymidylate kinase	186	99	128	0	1	1	0	0	1		M3	
416	118576713	YP_876456.1	HD superfamily phosphohydrolases	407	111	200	0	1	1	1	1	1	1	M3	
417	118576723	YP_876466.1	hypothetical protein CENSYa_1540	223	105	35	1	0	0	0	0	0	0		
418	118576726	YP_876469.1	hypothetical protein CENSYa_1543	233	108	34	1	0	0	0	0	0	0		
419	118576731	YP_876474.1	FKBP-type peptidyl-prolyl cis-trans isomerase/rotamase	261	120	102	1	1	1	1	1	0	0	M	
420	118576733	YP_876476.1	nicotinamide mononucleotide adenylyltransferase	111	148	61	1	1	1	1	1	0	0	M	
421	118576735	YP_876478.1	RNA 3p-terminal phosphate cyclase	339	149	148	0	1	1	1	1	1	1	M3	
422	118576736	YP_876479.1	copper export protein	169	164	4	1	0	0	0	0	0	0		
423	118576737	YP_876480.1	hypothetical protein CENSYa_1554	141	141	4	1	0	0	0	0	0	0		
424	118576739	YP_876482.1	NADH-dependent FMN reductase	174	169	6	0	0	0	0	0	0	1		
425	118576746	YP_876489.1	hypothetical protein CENSYa_1565	388	346	17	1	0	0	0	0	0	0		
426	118576750	YP_876493.1	hypothetical protein CENSYa_1569	267	148	22	1	0	0	0	0	0	0		
427	118576751	YP_876494.1	hypothetical protein CENSYa_1570	287	192	18	1	0	0	0	0	0	0		
428	118576752	YP_876495.1	hypothetical protein CENSYa_1571	281	235	11	1	0	0	0	0	0	0		
429	118576753	YP_876496.1	cysteine sulfinate desulfinase/cysteine desulfurase	374	311	200	0	1	0	0	0	1	1		
430	118576755	YP_876498.1	hypothetical protein CENSYa_1574	405	166	153	0	1	1	0	1	1	1		
431	118576757	YP_876500.1	NCAIR mutase	267	142	145	0	1	1	0	1	1	1		
432	118576758	YP_876501.1	Fe-S oxidoreductase	511	295	73	0	1	1	1	1	0	1		

433	118576760	YP_876503.1	hypothetical protein CENSYa_1579	330	106	36	0	1	0	0	1	0	1			
434	118576761	YP_876504.1	secreted periplasmic Zn-dependent protease	1022	71	51	1	0	0	0	0	0	0			
435	118576762	YP_876505.1	divalent heavy-metal cation transporter	374	354	12	0	1	1	0	0	0	1			
436	118576763	YP_876506.1	multicopper oxidase	346	93	200	0	1	0	1	0	1	1			
437	118576765	YP_876508.1	hypothetical protein CENSYa_1584	100	99	4	1	0	0	0	0	0	0			
438	118576781	YP_876524.1	glucosamine 6-phosphate synthetase	585	387	200	0	1	1	1	1	1	1			
439	118576782	YP_876525.1	phosphoribosyltransferase	346	185	115	0	1	1	0	0	0	1	M3	A=>B	
440	118576784	YP_876527.1	hypothetical protein CENSYa_1603	213	229	130	0	1	1	1	1	1	1			
441	118576789	YP_876532.1	RNA-binding protein	211	125	45	0	1	1	0	1	0	1		A=>B	
442	118576791	YP_876534.1	enolase	412	337	200	0	1	1	1	1	1	1			
443	118576793	YP_876536.1	dioxygenase	273	110	174	0	1	1	1	1	1	1		A=>B	
444	118576794	YP_876537.1	mevalonate kinase	309	121	77	0	1	1	1	1	1	1			
445	118576795	YP_876538.1	archaeal gamma-glutamyl kinase	246	86	101	0	1	1	1	0	1	1			
446	118576796	YP_876539.1	isopentenyl-diphosphate delta-isomerase	210	75	157	0	1	0	0	0	1	1			
447	118576797	YP_876540.1	geranylgeranyl pyrophosphate synthase/geranyltranstransferase	321	148	200	0	1	1	0	0	1	1		T=>B // A=>B	
448	118576798	YP_876541.1	glutamyl-tRNA synthetase	562	246	200	0	1	1	1	1	1	1	M		
449	118576799	YP_876542.1	2-keto-4-pentenoate hydratase/2-oxohepta-3-ene-1 7-dioic acid hydratase	288	145	200	0	1	0	1	1	1	1			
450	118576801	YP_876544.1	SAM dependent methyltransferase	248	127	43	0	0	1	0	0	1	1			
451	118576812	YP_876555.1	Mg2 and Co2 transporter	334	118	193	0	1	0	0	0	1	1			
452	118576813	YP_876556.1	autotransporter adhesin	922	197	26	0	0	0	0	0	0	1		T<=>B	
453	118576815	YP_876558.1	autotransporter adhesin	611	134	41	0	0	0	0	0	1	1		T<=>B	
454	118576817	YP_876560.1	hypothetical protein CENSYa_1638	209	111	33	1	0	0	0	0	0	0			
455	118576819	YP_876562.1	xanthosine triphosphate pyrophosphatase	126	93	200	0	1	1	1	1	1	1			
456	118576820	YP_876563.1	Mn2 -dependent serine/threonine protein kinase	318	98	170	0	1	1	1	1	1	0	M		
457	118576824	YP_876567.1	nucleotide kinase	153	127	22	1	1	1	1	0	0	0			
458	118576825	YP_876568.1	queuine/archaeosine tRNA-ribosyltransferase	508	211	200	0	1	1	1	1	0	1	M		
459	118576830	YP_876573.1	3-methyladenine DNA glycosylase/8-oxoguanine DNA glycosylase	225	91	126	0	1	0	0	0	1	1			
460	118576831	YP_876574.1	archaeal fructose-1 6-bisphosphatase	271	97	162	0	1	1	0	1	1	1	M3		
461	118576832	YP_876575.1	hypothetical protein CENSYa_1653	374	357	5	1	0	0	0	0	0	0			
462	118576835	YP_876578.1	hypothetical protein CENSYa_1656	242	237	5	1	0	0	0	0	0	0			
463	118576836	YP_876579.1	glutamyl-tRNA reductase	356	194	200	0	1	1	0	0	1	1	M2		
464	118576841	YP_876584.1	acetyl/propionyl-CoA carboxylase alpha subunit	169	65	113	0	1	1	0	0	1	1			
465	118576859	YP_876602.1	hypothetical protein CENSYa_1681	225	111	36	1	0	0	0	0	0	0			
466	118576863	YP_876606.1	orotate phosphoribosyltransferase	200	95	200	0	1	1	0	1	1	1		A=>B	
467	118576864	YP_876607.1	hypothetical protein CENSYa_1687	357	360	5	1	0	0	0	0	0	0			
468	118576869	YP_876612.1	methyltransferase	275	123	200	0	1	1	1	1	1	1			
469	118576872	YP_876615.1	ribosomal protein S27E	224	60	43	1	1	1	1	1	0	0	M		
470	118576874	YP_876617.1	translation factor	189	86	200	0	1	1	0	0	1	1			
471	118576875	YP_876618.1	argininosuccinate lyase	503	320	200	0	1	0	0	0	1	1	M2		
472	118576886	YP_876629.1	archaeal DNA-binding protein	96	75	78	1	1	1	1	1	0	0	M		
473	118576890	YP_876633.1	branched-chain amino acid aminotransferase/4-amino-4-deoxychorismate lyase	305	214	200	0	1	1	1	0	1	1			
474	118576893	YP_876636.1	acetyltransferase	158	98	143	0	1	1	1	1	1	1		B=>T	
475	118576894	YP_876637.1	UDP-N-acetylglucosamine-1-phosphate transferase	329	160	43	0	1	1	1	1	0	1	M2		
476	118576895	YP_876638.1	superfamily II helicase	707	268	200	0	1	1	1	1	1	1	M2		
477	118576897	YP_876640.1	hypothetical protein CENSYa_1724	172	169	5	1	0	0	0	0	0	0			
478	118576899	YP_876642.1	ferredoxin	177	98	42	1	1	1	1	0	0	0			
479	118576900	YP_876643.1	ferredoxin	164	97	43	1	1	1	1	0	0	0			
480	118576901	YP_876644.1	translation initiation factor	221	165	200	0	1	1	1	1	1	0	M2		
481	118576902	YP_876645.1	histidyl-tRNA synthetase	417	163	200	0	1	1	1	1	1	1	M3		
482	118576904	YP_876647.1	surface antigen	723	259	4	1	0	0	0	0	0	0			
483	118576906	YP_876649.1	hypothetical protein CENSYa_1734	524	86	41	0	1	0	0	1	0	1		T/EA/B	
484	118576908	YP_876651.1	hypothetical protein CENSYa_1736	162	112	4	1	0	0	0	0	0	0			
485	118576920	YP_876663.1	hypothetical protein CENSYa_1748	1291	1026	7	1	0	0	0	0	0	0			
486	118576922	YP_876665.1	hypothetical protein CENSYa_1750	213	113	27	1	0	0	0	0	0	0			
487	118576924	YP_876667.1	citrate lyase beta subunit	288	123	200	0	1	1	0	0	1	1		B=>T	
488	118576925	YP_876668.1	TPR repeat protein	235	82	94	0	1	0	0	0	1	1			
489	118576926	YP_876669.1	riboflavin synthase alpha chain	195	112	200	0	1	0	0	0	1	1		B=>T	
490	118576929	YP_876672.1	GTP cyclohydrolase III	185	226	23	1	1	0	0	0	0	0			
491	118576930	YP_876673.1	cytosine deaminase	392	122	200	0	1	1	1	1	1	1	M3		
492	118576931	YP_876674.1	riboflavin specific deaminase	217	103	200	0	1	1	0	0	1	1	M3	A=>B	
493	118576933	YP_876676.1	hypothetical protein CENSYa_1761	183	95	29	0	1	1	1	0	0	1		B<=>A	
494	118576936	YP_876679.1	DNA-binding protein	95	73	115	0	1	1	1	0	1	0	M2		
495	118576939	YP_876682.1	hypothetical protein CENSYa_1768	222	113	122	0	1	1	1	1	1	0	M2		
496	118576942	YP_876685.1	oxygen-sensitive ribonucleoside-triphosphate reductase	184	692	9	1	0	0	0	0	0	0			
497	118576948	YP_876691.1	autotransporter adhesin	964	151	37	0	0	0	0	0	0	1		T<=>B	
498	118576952	YP_876695.1	hypothetical protein CENSYa_1781	184	114	25	1	0	0	0	0	0	0			
499	118576954	YP_876697.1	amidohydrolase	269	131	200	0	1	1	0	0	1	1			
500	118576959	YP_876702.1	hypothetical protein CENSYa_1788	162	108	16	0	1	0	0	0	0	0			
501	118576962	YP_876705.1	secreted periplasmic Zn-dependent protease	577	79	46	1	0	0	0	0	0	0			
502	118576964	YP_876707.1	hypothetical protein CENSYa_1793	275	96	34	0	0	0	0	1	0	1		A<=>B	
503	118576965	YP_876708.1	copper binding protein plastocyanin/azurin family	363	31	66	0	1	0	1	1	0	1		A<=>B	
504	118576966	YP_876709.1	copper binding protein plastocyanin/azurin family	291	0	1										
505	118576967	YP_876710.1	copper binding protein plastocyanin/azurin family	331	48	51	0	1	0	1	0	0	1		T<=>EA // T<=>B	
506	118576968	YP_876711.1	hypothetical protein CENSYa_1797	232	204	8	1	0	0	0	0	0	0			
507	118576969	YP_876712.1	streptogramin lyase	927	145	28	0	1	0	0	0	1	1		T<=>B	
508	118576970	YP_876713.1	hypothetical protein CENSYa_1799	445	422	17	1	0	0	0	0	0	0			
509	118576977	YP_876720.1	Mn2 and Fe2 transporter of the NRAMP family	426	348	29	0	0	0	0	0	0	1		B=>T	
510	118576978	YP_876721.1	hypothetical protein CENSYa_1807	141	345	15	1	0	0	0	0	0	0			
511	118576980	YP_876723.1	FAD/FMN-containing dehydrogenase	458	204	200	0	1	1	0	0	1	1		B=>T	
512	118576983	YP_876726.1	DnaJ-class molecular chaperone	171	77	200	0	1	0	0	0	1	1			
513	118576985	YP_876728.1	TPR repeat protein	73	66	18	1	1	0	0	0	0	0			
514	118576986	YP_876729.1	TPR repeat protein	157	43	200	0	1	0	0	0	1	1		B=>T	
515	118576989	YP_876732.1	hypothetical protein CENSYa_1818	104	76	53	0	1	1	0	1	1	1			
516	118576993	YP_876736.1	CBS domain	272	33	119	0	1	1	1	1	0	1			
517	118576994	YP_876737.1	transcriptional regulator	117	64	18	0	1	0	0	0	0	0			
518	118576995	YP_876738.1	nitrogen regulatory protein PII	109	104	91	0	1	0	0	0	0	1		B=>A (EA+T)	
519	118576996	YP_876739.1	transcription initiation factor TFIIB	304	249	200	0	1	1	1	1	1	0	M		
520	118576997	YP_876740.1	ferritin-like protein	181	95	156	0	1	1	0	0	1	1		B=>T	
521	118577002	YP_876745.1	hypothetical protein CENSYa_1831	399	86	21	1	0	0	0	0	0	0			
522	118577005	YP_876748.1	hypothetical protein CENSYa_1834	120	119	4	1	0	0	0	0	0	0			
523	118577006	YP_876749.1	hypothetical protein CENSYa_1835	215	163	4	1	0	0	0	0	0	0			
524	118577011	YP_876754.1	archaeal DNA polymerase II small subunit/DNA polymerase delta subunit B	476	238	80	0	1	0	1	1	1	0	M		
525	118577012	YP_876755.1	hypothetical protein CENSYa_1841	176	163	5	1	0	0	0	0	0	0			
526	118577014	YP_876757.1	hypothetical protein CENSYa_1843	162	161	4	1	0	0	0	0	0	0			
527	118577015	YP_876758.1	peptidyl-prolyl cis-trans isomerase (rotamase)	545	77	200	0	1	0	0	0	1	1			
528	118577017	YP_876760.1	hypothetical protein CENSYa_1846	219	212	4	1	0	0	0	0	0	0			
529	118577018	YP_876761.1	molecular chaperone (small heat shock protein)	164	108	54	0	1	1	1	1	0	1	M3		
530	118577019	YP_876762.1	AAA ATPase	709	569	200	0	1	1	1	1	1	1	M		
531	118577022	YP_876765.1	chaperonin GroEL (HSP60 family)	566	480	200	1	1	1	1	1	0	0	M		
532	118577024	YP_876767.1	glutamine synthetase	482	408	200	0	1	1	1	0	0	1			
533	118577025	YP_876768.1	hypothetical protein CENSYa_1854	198	126	5	1	0	0	1	0	0	0			
534	118577031	YP_876774.1	Ser/Thr protein kinase	229	98	49	0	1	1	1	0	1	1		A<=>A	
535	118577032	YP_876775.1	tRNA nucleotidyltransferase (CCA-adding enzyme)	442	191	98	1	1	1	1	1	0	0	M		
536	118577033	YP_876776.1	2p-5p RNA ligase	175	71	124	0	1	1	1	1	0	1	M2		
537	118577034	YP_876777.1	dephospho-CoA kinase	176	112	91	1	1	1	0	0	0	0	M2		
538	118577035	YP_876778.1	hypothetical protein CENSYa_1865	134	103	23	1	1	1	1	0	0	0	M3	A<=>A	
539	118577037	YP_876780.1	metal-dependent hydrolase of the beta-lactamase superfamily III	298	135	200	0	1	1	1	1	1	1		B=>T	
540	118577039	YP_876782.1	phosphoribosylpyrophosphate synthetase	283	220	200	0	1	1	1	0	1	1	M2		
541	118577042	YP_876785.1	glutamine synthetase	475	409	200	0	1	1	1	0	0	1	M3		
542	118577043	YP_876786.1	hypothetical protein CENSYa_1873	204	204	8	1	0	0	0	0	0	0			
543	118577044	YP_876787.1	secreted periplasmic Zn-dependent protease	516	94	42	1	0	0	0	0	0	0			
544	118577046	YP_876789.1	transcription initiation factor TFIIB	303	255	200	1	1	1	1	1	0	0	M		
545	118577049	YP_876792.1	hypothetical protein CENSYa_1879	181	169	7	0	0	0	0	0	0	1		T<=>B	
546	118577051	YP_876794.1	3p-phosphoadenosine 5p-phosphosulfate (PAPS) 3p-phosphatase	260	106	200	0	0	0	0	0	1	1		B=>T	
547	118577053	YP_876796.1	hypothetical protein CENSYa_1883	144	138	4	1	0	0	0	0	0	0			
548	118577054	YP_876797.1	transcription initiation factor TFIIB	303	253	200	1	1	1	1	1	0	0	M		
549	118577055	YP_876798.1	conserved protein implicated in secretion	195	130	42	1	0	1	1	0	0	0	M		
550	118577056	YP_876799.1	hypothetical protein CENSYa_1886	111	92	8	1	0	0	0	0	0	0			
551	118577057	YP_876800.1	uncharacterized membrane-associated protein	325	225	15	1	1	1	0	0	0	0			
552	118577062	YP_876805.1	organic radical activating enzyme	238	63	169	0	1	1	0	0	0	1			
553	118577066	YP_876809.1	homoserine kinase	257	132	177	0	1	1	1	1	1	1			
554	118577070	YP_876813.1	hypothetical protein CENSYa_1900	80	79	4	1	0	0	0	0	0	0			
555	118577072	YP_876815.1	thiamine biosynthesis protein	375	366	4	1	0	0	0	0	0	0			
556	118577075	YP_876818.1	hypothetical protein CENSYa_1905	237	216	6	1	0	0	0	0	0	0			
557	118577077	YP_876820.1	hypothetical protein CENSYa_1907	398	226	95	0	1	1	0	1	1	1	M3		
558	118577079	YP_876822.1	hypothetical protein CENSYa_1909	216	112	33	1	0	0	0	0	0	0			
559	118577081	YP_876824.1	phosphoribosylformylglycinamidine (FGAM) synthase synthetase domain	713	482	200	0	1	1	0	0	0	1	M		
560	118577083	YP_876826.1	uracil-DNA glycosylase	227	84	133	0	1	1	1	1	0	1			
561	118577085	YP_876828.1	conserved protein implicated in secretion	214	127	42	1	0	1	1	0	0	0	M		
562	118577090	YP_876833.1	RNA-binding protein	631	229	164	0	1	1	1	1	1	0	M		
563	118577093	YP_876836.1	hypothetical protein CENSYa_1923	199	91	162	0	1	1	1	1	1	1	M3	A=>B	
564	118577096	YP_876839.1	eukaryotic-type DNA primase large subunit	337	133	67	1	1	1	1	1	0	0	M		
565	118577097	YP_876840.1	eukaryotic-type DNA primase catalytic (small) subunit	372	167	68	1	1	1	1	1	0	0	M		
566	118577099	YP_876842.1	GMP synthase glutamine amidotransferase domain	179	137	19	0	1	0	0	0	0	1		A<=>B	
567	118577100	YP_876843.1	TPR repeat protein	319	138	27	0	0	0	0	0	1	1			
568	118577106	YP_876849.1	phospholipid-binding protein	150	70	200	0	1	1	0	1	1	1			
569	118577107	YP_876850.1	hypothetical protein CENSYa_1938	271	110	34	1	0	0	0	0	0	0			
570	118577108	YP_876851.1	uncharacterized membrane protein required for N-linked glycosylation	736	361	166	0	1	1	1	1	1	0	M2		
571	118577109	YP_876852.1	cobalt-precorrin-6A synthase	311	171	109	0	1	1	0	0	0	1		B=>T	
572	118577110	YP_876853.1	precorrin-3B C17-methyltransferase	347	123	135	0	1	1	0	0	0	1		B=>T	
573	118577113	YP_876856.1	cobyrinic acid a c-diamide synthase	423	244	200	0	1	1	0	0	0	1			
574	118577116	YP_876859.1	ATP corrinoid adenosyltransferase	176	107	200	0	1	1	0	0	0	1			
575	118577118	YP_876861.1	nucleoside-diphosphate-sugar epimerase	249	101	85	0	1	0	0	0	1	1		B<=>A	
576	118577121	YP_876864.1	RecA/RadA recombinase related protein	218	257	121	0	1	1	0	1	1	0	M		
577	118577123	YP_876866.1	single-stranded DNA-binding replication protein A large (70 kD) subunit	468	136	19	1	1	0	0	0	0	0			

578	118577124	YP_876867.1	ABC-type multidrug transport system ATPase component	313	168	200	0	1	1	1	1	0	0	0	
579	118577126	YP_876869.1	hypothetical protein CENSYa_1958	347	187	49	0	1	1	1	1	0	1	M2	A=>B
580	118577128	YP_876871.1	molecular chaperone GrpE	187	85	114	0	1	0	0	0	1	1		
581	118577130	YP_876873.1	DnaJ-class molecular chaperone	351	278	200	0	1	0	0	0	1	1		
582	118577134	YP_876877.1	archaeal Glu-tRNAGln amidotransferase subunit E	628	450	106	0	1	1	1	1	0	1	M	
583	118577135	YP_876878.1	archaeal Glu-tRNAGln amidotransferase subunit D/asparaginase	392	170	200	0	1	1	1	1	1	1	M	
584	118577136	YP_876879.1	AAA ATPase	728	568	200	0	1	1	1	1	1	1	M	
585	118577141	YP_876884.1	hypothetical protein CENSYa_1975	107	128	6	1	0	0	0	0	0	0		
586	118577143	YP_876886.1	dinucleotide-binding enzyme	206	158	61	0	1	0	1	0	0	1	M2	E=>B
587	118577144	YP_876887.1	fructose-2 6-bisphosphatase	228	42	200	0	1	1	0	0	1	1		
588	118577152	YP_876895.1	thiol-disulfide isomerase	135	56	200	0	1	1	1	0	1	1		
589	118577153	YP_876896.1	hypothetical protein CENSYa_1988	139	77	82	0	1	1	1	1	1	1	M2	E=>B
590	118577156	YP_876899.1	polyprenyltransferase	310	203	200	0	1	1	1	1	0	1	M3	
591	118577157	YP_876900.1	transcriptional regulator	152	53	200	0	1	1	1	1	0	0		
592	118577159	YP_876902.1	threonine-phosphate decarboxylase	385	113	200	0	1	1	1	0	1	1		
593	118577160	YP_876903.1	cobyric acid synthase	276	287	200	0	1	0	0	0	0	1		
594	118577161	YP_876904.1	cobalamin biosynthesis protein	320	122	200	0	1	1	0	0	1	1	M3	A=>B /E
595	118577162	YP_876905.1	cobalamin-5-phosphate synthase	219	175	35	0	1	0	0	0	0	1	M3	
596	118577163	YP_876906.1	4-diphosphocytidyl-2C-methyl-D-erythritol synthase	198	99	54	0	1	1	0	0	0	0	M2	
597	118577165	YP_876908.1	adenine deaminase	609	290	133	0	1	1	1	1	1	1		
598	118577167	YP_876910.1	ROK-family glucokinase	282	122	200	0	1	1	0	0	0	1		B=>T
599	118577169	YP_876912.1	hypothetical protein CENSYa_2004	207	141	12	1	0	1	1	0	0	0	M3	
600	118577171	YP_876914.1	uncharacterized Rossmann fold enzyme	239	104	89	1	1	1	1	1	0	0	M2	
601	118577172	YP_876915.1	dihydropteroate synthase	275	136	200	0	1	1	0	1	1	1		
602	118577179	YP_876922.1	hypothetical protein CENSYa_2014	362	162	9	1	1	1	0	0	0	0		
603	118577181	YP_876924.1	Zn-dependent oxidoreductase	442	151	200	0	1	1	1	1	1	1		T<=>B
604	118577182	YP_876925.1	TPR repeat protein	206	45	178	0	1	0	0	0	1	1		
605	118577188	YP_876931.1	DNA-binding protein containing a Zn-ribbon domain	508	202	98	0	1	1	1	0	0	0	M	
606	118577189	YP_876932.1	glucose/sorbosone dehydrogenase	390	114	200	0	1	1	1	0	1	0		
607	118577190	YP_876933.1	hydroxymethylglutaryl-CoA reductase	415	288	69	0	1	1	1	1	1	1		
608	118577191	YP_876934.1	hypothetical protein CENSYa_2027	191	129	15	0	1	0	0	0	0	1		T<=>EA/B
609	118577192	YP_876935.1	hypothetical protein CENSYa_2028	193	142	15	1	1	0	0	0	0	0		T<=>EA
610	118577193	YP_876936.1	hypothetical protein CENSYa_2029	208	142	15	1	1	0	0	0	0	0		T<=>EA
611	118577194	YP_876937.1	hypothetical protein CENSYa_2030	218	145	11	1	0	0	0	0	0	0		
612	118577195	YP_876938.1	hypothetical protein CENSYa_2031	183	122	16	0	1	0	0	0	0	1		T<=>EA/B
613	118577196	YP_876939.1	hypothetical protein CENSYa_2032	207	142	15	1	1	0	0	0	0	0		T<=>EA
614	118577197	YP_876940.1	hypothetical protein CENSYa_2033	189	142	15	1	1	0	0	0	0	0		T<=>EA
615	118577200	YP_876943.1	histidinol dehydrogenase	422	265	200	0	1	1	1	0	1	1		A=>B
616	118577201	YP_876944.1	histidinol-phosphate aminotransferase	354	120	200	0	1	1	1	0	1	1		
617	118577202	YP_876945.1	phosphatase	321	258	17	1	1	0	1	0	0	0		
618	118577204	YP_876947.1	glutamine amidotransferase	201	116	200	0	1	1	1	0	1	1		
619	118577205	YP_876948.1	phosphoribosyl formimino-5-aminoimidazole isomerase	234	149	200	0	1	0	1	0	0	1		B=>AA+T
620	118577210	YP_876953.1	4-methyl-5-(beta-hydroxyethyl)thiazole monophosphate synthesis protein	458	169	200	0	1	0	0	1	1	1		
621	118577215	YP_876958.1	hypothetical protein CENSYa_2051	204	119	16	0	1	0	0	0	0	1		T<=>EA/B
622	118577223	YP_876966.1	hypothetical protein CENSYa_2059	1846	93	24	1	0	1	0	0	0	0		
623	118577224	YP_876967.1	hypothetical protein CENSYa_2060	1232	167	22	1	0	1	0	0	0	0		
624	118577225	YP_876968.1	hypothetical protein CENSYa_2061	473	911	10	1	0	0	0	0	0	0		
625	118577228	YP_876971.1	hypothetical protein CENSYa_2064	523	84	9	1	0	0	0	0	0	0		
626	118577230	YP_876973.1	surface layer-associated STABLE protease	1047	122	110	0	1	1	0	0	1	1		

C.	Sequence de départ				Alignement		Présence (1) /Absence (0) d'homologues dans les phylogénies							Observations	
							Archaea								
	Référence GI	Référence RefSeq	annotation	taille (aa)	Nombre de positions	Nombre de sequences	spéciale archée	Eury-archées	Cren-archées	Thaum-archée	Kor-archée	Eucarya	Bacteria	Marqueur potentiel	THG potentiel
1	68445562	BAE03270.1	hypothetical conserved protein	77	77	5	1	0	0	0	0	0	0		
2	68445563	BAE03271.1	pyrophosphatase	97	84	41	0	1	1	0	1	0	1	M2	
3	68445564	BAE03272.1	hypothetical conserved protein	223	193	24	1	1	1	0	1	0	0		
4	68445565	BAE03273.1	aerobic-type carbon monoxide dehydrogenase large subunit CoxLCutL homologs	797	557	100	0	1	1	0	0	0	1		B=>AA
5	68445566	BAE03274.1	hypothetical protein HGI-05	142	142	5	1	0	0	0	0	0	0		
6	68445567	BAE03275.1	hypothetical protein HGI-06	384	384	4	1	0	0	0	0	0	0		
7	68445568	BAE03276.1	hypothetical protein HGI-07	440	440	5	1	0	0	0	0	0	0		
8	68445569	BAE03277.1	hypothetical conserved protein	512	512	5	1	0	0	0	0	0	0		
9	68445570	BAE03278.1	acetolactate synthase large subunit homolog ilvB-6	561	315	43	0	0	1	0	0	1	1		
10	68445571	BAE03279.1	hypothetical conserved protein	931	464	69	0	1	0	1	0	0	1		B=AA+T
11	68445573	BAE03281.1	hypothetical conserved protein	232	180	83	1	1	1	0	1	0	0	M	
12	68445575	BAE03283.1	hypothetical conserved protein	243	243	4	1	0	0	0	0	0	0		
13	68445576	BAE03284.1	hypothetical conserved protein	156	92	50	0	1	1	0	1	1	1		B<=>A
14	68445577	BAE03285.1	dehydrogenase	366	162	100	0	1	1	1	0	1	1		B=>AA
15	68445578	BAE03286.1	sorbitol dehydrogenase	339	172	100	0	1	1	1	0	1	1		B<=>A
16	68445579	BAE03287.1	hypothetical conserved protein	279	276	5	1	0	0	0	0	0	0		
17	68445580	BAE03288.1	ABC transporter	272	254	6	0	0	0	0	0	0	1		
18	68445581	BAE03289.1	ABC transporter ATP-binding protein	325	187	100	0	1	1	0	1	0	1		
19	68445582	BAE03290.1	V-type ATP synthase alpha chain	596	504	100	0	1	1	1	1	0	1	M	A=>B
20	68445584	BAE03292.1	permeases of the major facilitator superfamily	368	349	9	1	1	1	0	0	0	0		
21	68445586	BAE03294.1	hypothetical conserved protein	417	168	100	0	1	1	1	0	1	1		
22	68445587	BAE03295.1	ornithine carbamoyltransferase	297	236	100	0	1	1	1	1	1	1		A<=>B
23	68445588	BAE03296.1	O-acetyl-L-serine sulfhydrylase	342	282	100	0	1	1	0	0	1	1		B=>CA+AA
24	68445589	BAE03297.1	hypothetical protein HGI-28	98	98	4	1	0	0	0	0	0	0		
25	68445590	BAE03298.1	threonine synthase	414	316	100	0	1	1	1	0	0	1		A<=>B
26	68445591	BAE03299.1	carbamate kinase-like carbamoyl phosphate synthetase	310	221	100	0	1	1	0	1	0	1		
27	68445592	BAE03300.1	ornithine carbamoyltransferase	339	237	100	0	1	1	0	1	0	1		B=>CA+AA
28	68445594	BAE03302.1	aldehyde oxidase and xanthine dehydrogenase family protein	774	357	100	0	1	1	0	0	1	1		B=CA+AA
29	68445595	BAE03303.1	carbon-monoxide dehydrogenase small chain	146	114	100	0	1	1	0	0	0	1		B=>CA+AA
30	68445596	BAE03304.1	carbon monoxide dehydrogenase	277	157	54	0	1	1	0	0	0	1		B=>AA
31	315424919	BAJ46595.1	inosine-5-monophosphate dehydrogenase	440	426	100	0	1	0	1	0	1	1		
32	315424920	BAJ46596.1	hypothetical conserved protein	194	189	36	1	1	1	0	1	0	0	M3	
33	315424922	BAJ46598.1	endonuclease V	190	117	100	0	1	1	0	0	1	1		
34	315424923	BAJ46599.1	cell division control protein 6	371	283	100	1	1	1	1	1	0	0	M2	
35	315424925	BAJ46601.1	inteinrRNA intron-related DNA endonuclease	299	254	9	1	1	0	0	0	0	0		EA<=>AA
36	315424926	BAJ46602.1	hypothetical conserved protein	118	124	12	1	0	1	0	0	0	0		
37	315424927	BAJ46603.1	hypothetical protein HGMM_F01H02C10	134	134	7	1	0	0	0	0	0	0		
38	315424928	BAJ46604.1	hypothetical protein HGMM_F01H02C11	288	266	14	1	0	0	0	0	0	0		
39	315424933	BAJ46609.1	hypothetical conserved protein	237	127	10	0	1	1	0	0	0	1		
40	315424937	BAJ46613.1	hypothetical conserved protein	337	124	84	0	1	0	1	0	1	1		B<=>AA+T
41	315424939	BAJ46615.1	antibiotic transport system ATP-binding protein	314	191	100	0	1	1	0	1	0	1	M3	
42	315424943	BAJ46619.1	electron transport protein SCO1SenC	196	71	100	0	0	0	0	0	1	1		B=>AA
43	315424944	BAJ46620.1	multicopper oxidase	485	183	51	0	1	0	1	0	1	1		B=>AA
44	315424946	BAJ46622.1	cysteine synthase B	302	251	100	0	0	0	0	0	1	1		B=>AA
45	315424947	BAJ46623.1	hypothetical conserved protein	140	81	27	0	1	0	0	1	0	1		
46	315424948	BAJ46624.1	ThiS family protein	180	86	58	0	1	0	0	0	0	1		B=>AA
47	315424954	BAJ46630.1	glyoxalasebleomycin resistance proteindioxygenase	341	175	86	0	1	1	0	0	0	1		
48	315424955	BAJ46631.1	hypothetical conserved protein	179	176	5	0	1	1	0	0	0	1		
49	315424959	BAJ46635.1	hypothetical conserved protein	140	105	36	0	1	0	1	1	0	1		B<=>A
50	315424960	BAJ46636.1	hypothetical conserved protein	82	73	17	0	0	1	0	1	0	1		B=>AA
51	315424963	BAJ46638.1	hypothetical conserved protein	196	189	5	1	1	0	0	0	0	0		
52	315424964	BAJ46639.1	xanthine dehydrogenase	689	502	100	0	0	1	0	0	0	1		B=>AA
53	315424965	BAJ46640.1	N-ethylammeline chlorohydrolase	439	277	100	0	1	1	0	0	1	1		
54	315424966	BAJ46641.1	methylenetetrahydromethanopterin reductase	333	170	78	0	1	0	0	0	0	1		
55	315424967	BAJ46642.1	amidohydrolase	463	286	100	0	1	1	0	1	1	1		
56	315424970	BAJ46645.1	hypothetical conserved protein	168	124	28	1	1	1	0	0	0	0		
57	315424971	BAJ46646.1	acetyl-CoA C-acetyltransferase	394	336	100	0	1	1	1	0	0	1		A=>B
58	315424972	BAJ46647.1	ATPase	810	401	100	0	1	1	1	1	0	1	M	
59	315424975	BAJ46650.1	hypothetical conserved protein	209	168	4	1	0	1	0	0	0	0		
60	315424976	BAJ46651.1	phosphomannomutase	452	262	100	0	1	1	0	0	0	1		B<=>A
61	315424977	BAJ46652.1	F420-dependent NADP+ oxidoreductase	217	153	58	0	1	0	0	0	0	1		B=>AA
62	315424980	BAJ46655.1	phosphoribosyltransferase	201	80	100	0	1	1	0	1	1	1		
63	315424981	BAJ46656.1	glutamate dehydrogenase NADP+	417	316	100	0	1	1	1	1	1	1		
64	315424982	BAJ46657.1	hypothetical conserved protein	237	227	5	1	1	0	0	0	0	0		
65	315424984	BAJ46659.1	pyruvate dehydrogenase E1 component subunit alpha	341	214	100	0	1	1	0	0	1	1		
66	315424985	BAJ46660.1	hypothetical conserved protein	162	176	39	1	1	1	0	1	0	0	M3	
67	315424989	BAJ46664.1	fructose-16-bisphosphatase	376	344	85	0	1	1	1	1	0	1	M	A=>B
68	315424993	BAJ46668.1	hypothetical conserved protein	318	238	100	0	1	1	0	0	0	1		
69	315424997	BAJ46672.1	tRNA splicing endonuclease	181	153	24	1	1	0	1	0	0	0		
70	315424998	BAJ46673.1	4-hydroxybenzoate octaprenyltransferase	280	164	96	0	1	1	0	0	0	0		
71	315424999	BAJ46674.1	hypothetical conserved protein	217	171	15	0	1	0	1	0	0	1		
72	315425000	BAJ46675.1	F420-dependent glucose-6-phosphate dehydrogenase	360	130	100	0	1	0	1	0	0	1		
73	315425001	BAJ46676.1	hypothetical conserved protein	348	327	4	0	0	0	0	0	0	1		
74	315425004	BAJ46679.1	phenylalanyl-tRNA synthetase beta chain	557	331	100	0	1	1	1	1	1	0	M	
75	315425005	BAJ46680.1	methylmalonate-semialdehyde dehydrogenase	459	423	100	0	1	1	0	1	1	1		B=>A
76	315425007	BAJ46681.1	branched-chain alpha-keto acid dehydrogenase subunit E2	223	266	100	0	1	1	0	0	1	1		
77	315425009	BAJ46683.1	hypothetical conserved protein	692	375	100	0	1	1	1	0	1	1		
78	315425010	BAJ46684.1	tyrosyl-tRNA synthetase	363	258	100	0	1	1	1	0	1	0	M2	
79	315425011	BAJ46685.1	ribonucleoside-diphosphate reductase alpha chain	1044	474	100	0	1	1	1	1	0	1		
80	315425012	BAJ46686.1	multicomponent Na+H+ antiporter subunit D	486	279	100	0	1	1	1	0	0	1		
81	315425014	BAJ46688.1	metallophosphoesterase	261	87	70	1	1	1	1	1	0	0	M3	
82	315425016	BAJ46690.1	cupin domain protein	88	104	13	0	1	1	0	0	0	1		
83	315425017	BAJ46691.1	hydrolase	364	302	5	1	0	0	1	0	0	0		
84	315425019	BAJ46693.1	isopropylmalateisohomocitrate dehydrogenase	339	234	100	0	1	1	1	1	1	1	M2	A=>B
85	315425020	BAJ46694.1	radical SAM domain protein	235	206	71	0	1	1	0	1	1	1	M2	
86	315425021	BAJ46695.1	hypothetical conserved protein	193	192	10	1	0	1	0	0	0	0		
87	315425022	BAJ46696.1	hypothetical conserved protein	163	117	22	0	1	1	1	0	0	1		
88	315425024	BAJ46698.1	peptidenickel transport system substrate-binding protein	851	321	44	0	1	1	1	1	0	1		A=>B
89	315425025	BAJ46699.1	peptidenickel transport system permease protein	354	192	100	0	1	1	1	1	0	1		
90	315425026	BAJ46700.1	peptidenickel transport system permease protein	469	200	100	0	1	1	1	1	0	1		
91	315425028	BAJ46702.1	peptidenickel transport system ATP-binding protein	327	249	100	0	1	1	1	1	0	1		
92	315425029	BAJ46703.1	peptidenickel transport system ATP-binding protein	331	264	100	0	1	1	1	1	0	1		
93	315425032	BAJ46706.1	hypothetical protein HGMM_F11H10C26	460	428	5	1	0	0	0	0	0	0		
94	315425033	BAJ46707.1	NADH dehydrogenase I subunit L	550	422	100	0	1	0	1	0	0	1		B=>AA+T
95	315425036	BAJ46710.1	hypothetical conserved protein	179	144	13	1	1	1	1	0	0	0		
96	315425037	BAJ46711.1	hypothetical conserved protein	243	227	8	1	0	1	0	0	0	0		
97	315425038	BAJ46712.1	Fe-S cluster assembly ATP-binding protein SufC	257	213	100	0	1	1	1	0	1	1		
98	315425040	BAJ46713.1	isocitrate lyase	360	397	100	0	0	1	0	0	1	1		B=>CA
99	315425041	BAJ46714.1	malate synthase	527	439	100	0	0	1	0	0	1	1		B=>CA
100	315425042	BAJ46715.1	hypothetical conserved protein	377	150	100	0	1	1	0	0	0	0	M3	
101	315425045	BAJ46718.1	proteasome endopeptidase complex	187	168	100	1	1	1	1	1	0	0	M	
102	315425046	BAJ46719.1	proteasome alpha subunit	225	185	100	0	1	1	1	0	1	0	M	
103	315425047	BAJ46720.1	transcriptional regulator AsnC family	77	76	43	1	0	1	1	0	0	0	M2	
104	315425048	BAJ46721.1	metallo-beta-lactamase family protein	218	130	100	0	0	1	0	0	0	1		
105	315425049	BAJ46722.1	uracil-DNA glycosylase	231	117	92	0	1	1	1	0	0	1		
106	315425053	BAJ46726.1	threonine synthase	314	229	100	0	1	1	1	1	1	1		
107	315425055	BAJ46728.1	phosphate starvation-inducible protein PhoH	346	115	95	1	1	1	1	0	0	0	M3	
108	315425056	BAJ46729.1	hypothetical conserved protein	509	97	93	0	1	1	0	1	1	1		
109	315425057	BAJ46730.1	thiamine biosynthesis protein ThiI	398	181	100	0	1	1	0	0	0	1		
110	315425058	BAJ46731.1	aldehyde dehydrogenase	477	366	100	0	1	1	0	0	1	1		
111	315425060	BAJ46733.1	multiple antibiotic resistance protein MarC family	198	110	99	0	1	1	0	0	0	1		
112	315425062	BAJ46735.1	cytochrome c oxidase subunit III	278	203	15	0	1	0	0	0	0	1		T<=>AA
113	315425063	BAJ46736.1	translation initiation factor eIF-2B alpha subunit	279	250	6	1	1	0	0	0	0	0		EA<=>AA
114	315425064	BAJ46737.1	NADPHquinone reductase and related Zn-dependent oxidoreductases	344	214	100	0	1	1	1	1	0	1		
115	315425065	BAJ46738.1	nitrilasecyanide hydratase and apolipoprotein N-acyltransferase	270	93	82	0	1	1	1	1	1	1		
116	315425066	BAJ46739.1	apurinic endonuclease Apn1	283	189	100	0	1	0	1	0	1	1		
117	315425067	BAJ46740.1	glyoxalasebleomycin resistance proteindioxygenase	301	146	51	0	1	1	0	0	0	1		B=>AA
118	315425070	BAJ46743.1	hypothetical conserved protein	182	163	7	1	0	1	1	0	0	0		
119	315425071	BAJ46744.1	carbohydrate kinase FGGY	444	198	100	0	1	1	0	0	1	1		
120	315425072	BAJ46745.1	uracil phosphoribosyltransferase	220	147	100	0	1	1	0	0	1	1	M3	A=>B
121	315425073	BAJ46746.1	3-oxosteroid 1-dehydrogenase	551	194	100	0	1	0	0	0	1	1		B=>AA
122	315425075	BAJ46748.1	regulatory protein ArsR	143	132	6	1	0	1	0	0	0	0		
123	315425077	BAJ46750.1	Fe-S oxidoreductase	535	302	71	0	1	1	1	1	0	0		
124	315425079	BAJ46751.1	sulfonatenitratetaurine transport system permease protein	229	162	100	0	1	0	0	0	0	1		B<=>AA
125	315425083	BAJ46755.1	translation initiation inhibitor	128	115	100	0	1	1	0	1	1	1	M3	
126	315425084	BAJ46756.1	Fe-S cluster binding protein	708	292	100	0	1	0	0	0	0	1		B<=>AA
127	315425085	BAJ46757.1	hypothetical conserved protein	194	181	7	0	0	0	0	0	0	1		B<=>AA
128	315425086	BAJ46758.1	glycosyl transferase family protein	485	113	100	0	1	1	0	1	0	1		
129	315425087	BAJ46759.1	dimethyladenosine transferase	242	199	38	0	1	0	0	0	0	1	M2	A=>B

130	315425089	BAJ46761.1	appr-1-p processing domain-containing protein	180	117	100	0	1	1	0	1	1	1		
131	315425091	BAJ46763.1	Rieske 2Fe-2S family protein	103	86	22	0	1	0	1	0	0	1		
132	315425092	BAJ46764.1	thioredoxin reductase	324	257	100	0	1	0	1	0	1	1		B=>AA
133	315425093	BAJ46765.1	ferredoxin	85	97	12	0	0	0	1	0	0	1		B<=>AAT
134	315425094	BAJ46766.1	methyltransferase type 11	151	150	4	0	0	0	0	0	0	1		
135	315425095	BAJ46767.1	DNA polymerase II small subunit	492	240	80	0	1	0	1	1	1	0	M	
136	315425097	BAJ46769.1	hypothetical protein HGMM_F22F09C19	395	385	4	1	0	1	0	0	0	0		
137	315425099	BAJ46771.1	branched-chain amino acid transport system ATP-binding protein	236	196	100	0	1	1	0	1	0	1		
138	315425100	BAJ46772.1	branched-chain amino acid transport system ATP-binding protein	241	179	100	0	1	1	0	1	0	1		B=>AA
139	315425101	BAJ46773.1	branched-chain amino acid transport system permease protein	321	260	17	0	1	0	0	0	0	1		B=>AA
140	315425102	BAJ46774.1	branched-chain amino acid transport system permease protein	291	170	100	0	1	0	0	0	0	1		B=>AA
141	315425103	BAJ46775.1	extracellular solute-binding protein	417	135	100	0	1	1	0	1	0	1		B=>AA
142	315425104	BAJ46776.1	carbohydrate kinase YjeF related protein	432	303	67	0	1	1	1	0	0	1	M2	A=>B
143	315425105	BAJ46777.1	seryl-tRNA synthetase	427	339	100	0	1	1	1	1	1	1	M2	A=>B
144	315425106	BAJ46778.1	hypothetical conserved protein	416	156	97	0	1	1	0	1	0	1	M2	A=>B
145	315425107	BAJ46779.1	hypothetical protein HGMM_F22F09C29	489	397	4	1	1	0	0	0	0	0		
146	315425108	BAJ46780.1	ABC transport system ATP-binding protein	236	178	100	0	1	1	0	0	0	1		
147	315425110	BAJ46782.1	glutamate-1-semialdehyde 21-aminomutase	367	347	100	0	1	1	0	0	1	1	M	
148	315425111	BAJ46783.1	NUDIX hydrolase	177	104	23	0	1	1	0	0	0	0		
149	315425112	BAJ46784.1	Fe-S cluster assembly protein SufD	494	298	100	0	1	1	1	0	0	0		B=>A
150	315425114	BAJ46786.1	hypothetical conserved protein	258	173	12	1	0	1	1	0	0	0		
151	315425115	BAJ46787.1	ATPase	374	197	100	0	1	1	1	0	1	0		
152	315425116	BAJ46788.1	chromosome assembly protein homolog	548	170	85	0	1	0	1	0	1	1		E<=>A
153	315425119	BAJ46791.1	ATP-dependent helicase	930	509	100	1	1	1	1	1	0	0	M3	
154	315425120	BAJ46792.1	acyl-CoA dehydrogenase domain protein	401	311	100	0	1	1	0	0	1	1		A<=>B
155	315425122	BAJ46794.1	thermosome	459	510	100	1	1	1	0	1	0	0	M	
156	315425124	BAJ46795.1	chaperone protein DnaK	218	534	100	0	1	0	1	0	0	1		A=>B
157	315425125	BAJ46796.1	molecular chaperone GrpE	164	110	33	0	1	0	0	0	0	1		B=>A
158	315425127	BAJ46798.1	hypothetical conserved protein	355	215	38	0	1	1	1	1	0	1	M2	A=>B
159	315425128	BAJ46799.1	glycosyl transferase family protein	328	319	11	1	0	1	1	1	0	0		
160	315425131	BAJ46802.1	dehydrogenase	328	149	100	0	1	1	1	1	0	1		
161	315425133	BAJ46804.1	hypothetical protein HGMM_F29E04C10	281	266	14	1	0	0	0	0	0	0		
162	315425135	BAJ46806.1	type II restriction enzyme methylase subunit	1019	230	37	0	1	0	0	0	0	1		
163	315425136	BAJ46807.1	helicase	1026	167	86	0	1	0	1	0	0	1		
164	315425141	BAJ46812.1	transposase	157	159	31	1	1	0	0	0	0	0		
165	315425142	BAJ46813.1	ATPase	296	181	100	0	1	0	0	0	0	1		
166	315425147	BAJ46818.1	hypothetical conserved protein	431	171	21	0	1	0	0	0	0	1		
167	315425148	BAJ46819.1	hypothetical conserved protein	305	168	27	0	1	1	0	0	0	1		
168	315425149	BAJ46820.1	hypothetical conserved protein	388	73	100	0	1	1	0	0	0	1		
169	315425153	BAJ46824.1	threonine synthase	403	279	100	0	1	1	0	1	0	1	M	
170	315425154	BAJ46825.1	hypothetical conserved protein	959	412	100	0	1	1	1	1	1	1	M2	A=>B
171	315425159	BAJ46829.1	transcription initiation factor TFIIB	160	259	100	1	1	1	1	1	0	0	M	
172	315425161	BAJ46831.1	acetylacyl transferase related protein	235	159	38	1	1	1	0	1	0	0	M	
173	315425162	BAJ46832.1	hypothetical conserved protein	189	262	11	0	1	0	0	0	0	1		
174	315425163	BAJ46833.1	methylenetetrahydromethanopterin reductase	301	265	55	1	1	0	0	0	0	0	M	
175	315425165	BAJ46835.1	glycosyl transferase family 39	449	400	6	0	0	1	0	0	0	1		
176	315425166	BAJ46836.1	hypothetical conserved protein	240	162	55	1	1	1	0	0	0	0	M2	
177	315425167	BAJ46837.1	S-adenosylmethionine synthetase	409	321	100	0	1	1	0	1	0	1	M	
178	315425170	BAJ46840.1	isoleucyl-tRNA synthetase	1008	709	100	0	1	1	1	1	0	1	M	
179	315425171	BAJ46841.1	oxidoreductase molybdopterin binding protein	164	107	72	0	1	1	0	0	0	1		
180	315425172	BAJ46842.1	nitroreductase	169	91	71	0	1	1	0	0	0	1		
181	315425173	BAJ46843.1	hydrogenase expressionformation protein HypE	326	163	100	0	1	1	0	1	0	1		
182	315425175	BAJ46845.1	electron transfer flavoprotein alpha subunit	313	163	100	0	1	1	0	1	0	1		
183	315425176	BAJ46846.1	electron transfer flavoprotein beta subunit	261	155	100	0	1	1	0	1	1	1		
184	315425177	BAJ46847.1	Fe-S oxidoreductase	610	315	100	0	1	1	0	1	0	1		
185	315425178	BAJ46848.1	malate dehydrogenase oxaloacetate-decarboxylating	440	322	100	0	1	1	0	1	1	1		
186	315425179	BAJ46849.1	GCN5-related N-acetyltransferase	155	138	20	1	1	1	0	0	0	0		
187	315425180	BAJ46850.1	deoxyribodipyrimidine photo-lyase	360	235	100	0	1	0	0	0	1	1		
188	315425186	BAJ46856.1	formate dehydrogenase-O iron-sulfur subunit	336	142	100	0	0	1	0	0	0	1		B=>AA
189	315425187	BAJ46857.1	formate dehydrogenase alpha subunit	908	469	100	0	1	1	0	0	0	1		B=>A
190	315425190	BAJ46860.1	rhodanese domain protein	256	117	49	0	1	1	1	0	0	1		
191	315425191	BAJ46861.1	molybdopterin biosynthesis protein	652	271	100	0	1	1	0	1	0	1	M2	A=>B
192	315425192	BAJ46862.1	molybdopterin biosynthesis protein MoeA	415	262	100	0	1	1	0	1	0	1	M2	A=>B
193	315425193	BAJ46863.1	formate dehydrogenase family accessory protein FdhD	250	101	71	0	1	1	0	0	0	1		
194	315425196	BAJ46866.1	hypothetical conserved protein	182	177	17	1	0	1	0	1	0	0	M3	
195	315425206	BAJ46875.1	aldehyde dehydrogenase	258	309	100	0	1	1	0	0	1	1		
196	315425208	BAJ46877.1	branched-chain amino acid transport system substrate-binding protein	426	76	100	0	1	1	0	1	0	1		B<=>A
197	315425209	BAJ46878.1	branched-chain amino acid transport system permease protein	284	167	100	0	1	1	0	1	0	1		
198	315425210	BAJ46879.1	branched-chain amino acid transport system permease protein	315	181	100	0	1	1	0	1	0	1		
199	315425211	BAJ46880.1	branched-chain amino acid transport system ATP-binding protein	231	182	100	0	1	1	0	1	0	1		
200	315425212	BAJ46881.1	branched-chain amino acid transport system ATP-binding protein	250	172	100	0	1	1	0	1	0	1		
201	315425213	BAJ46882.1	HEPN domain-containing protein	120	122	9	1	1	1	0	0	0	0		
202	315425214	BAJ46883.1	DNA polymerase beta subunit	106	105	6	1	0	1	0	0	0	0		
203	315425219	BAJ46888.1	hypothetical conserved protein	95	92	40	0	1	1	0	1	0	0	M2	
204	315425220	BAJ46889.1	phytoene dehydrogenase	489	231	100	0	1	1	0	0	1	1		
205	315425221	BAJ46890.1	phytoene synthase	294	100	92	0	1	1	0	0	1	1		
206	315425222	BAJ46891.1	prenyltransferase	288	261	19	0	1	0	0	0	0	1		
207	315425223	BAJ46892.1	hypothetical conserved protein	175	182	7	1	0	1	0	0	0	0		
208	315425224	BAJ46893.1	hypothetical conserved protein	162	157	77	0	1	1	0	1	1	1	M	
209	315425225	BAJ46894.1	polyprenyl synthetase	353	141	100	0	1	1	0	1	1	1		
210	315425227	BAJ46896.1	cell division protein FtsZ	361	306	100	1	1	0	0	0	0	0	M	
211	315425230	BAJ46899.1	hypothetical conserved protein	681	59	100	0	1	1	0	0	0	1		EA=>B
212	315425231	BAJ46900.1	glutamate synthase NADPHNADH large chain	316	380	21	0	1	1	0	0	0	1	M3	EA=>B
213	315425232	BAJ46901.1	glutamate synthase NADPHNADH large chain	411	362	100	0	1	1	0	0	1	1	M3	EA=>B
214	315425233	BAJ46902.1	hypothetical conserved protein	275	220	57	0	1	1	0	1	0	1		
215	315425235	BAJ46904.1	Rieske 2Fe-2S family protein	272	163	8	1	0	1	1	1	0	0		
216	315425237	BAJ46906.1	DNA methylase	325	159	37	0	1	1	0	0	1	1		
217	315425240	BAJ46909.1	hydroxypyruvate reductase	398	309	100	0	1	1	1	1	1	1		
218	315425241	BAJ46910.1	deoxyhypusine synthase	345	171	100	0	1	1	1	1	1	1		
219	315425242	BAJ46911.1	oxidoreductase	184	138	26	0	1	1	0	0	0	1		A=>B
220	315425243	BAJ46912.1	reverse gyrase	1044	288	100	0	1	1	0	1	0	1	M2	A=>B
221	315425244	BAJ46913.1	isopropylmalate isomerase large subunit	470	437	100	0	0	0	1	0	1	1		B=>AA+T
222	315425250	BAJ46918.1	ornithine carbamoyltransferase	49	49	4	1	0	0	0	0	0	0		
223	315425251	BAJ46919.1	regulatory protein ArsR	526	560	24	1	1	0	1	0	0	0		
224	315425252	BAJ46920.1	nascent polypeptide-associated complex subunit alpha	122	100	32	1	1	1	0	1	0	0	M2	
225	315425253	BAJ46921.1	PUA domain containing protein	154	149	25	1	1	1	0	1	0	0		
226	315425258	BAJ46926.1	bacterioferritin	160	147	18	0	1	1	0	0	0	1		
227	315425259	BAJ46927.1	nucleotide binding protein containing PIN domain	121	112	9	1	1	1	0	0	0	0		
228	315425260	BAJ46928.1	hypothetical protein HGMM_F37B02C11	80	80	4	1	0	0	0	0	0	0		
229	315425262	BAJ46930.1	cysteine desulfurase SufS subfamily	409	362	100	0	1	1	1	0	1	1		
230	315425263	BAJ46931.1	3-hydroxyacyl-CoA dehydrogenaseenoyl-CoA hydratase	655	211	100	0	1	1	1	1	1	1		
231	315425264	BAJ46932.1	transcriptional regulator PadR family	105	99	18	1	1	1	0	1	0	0		
232	315425265	BAJ46933.1	myo-inositol-1-phosphate synthase	357	280	100	0	1	1	1	1	0	1		B=>AA
233	315425266	BAJ46934.1	hypothetical conserved protein	364	227	72	0	1	1	0	1	0	1	M2	
234	315425267	BAJ46935.1	DNA-directed RNA polymerase II subunit F	85	83	5	0	0	0	0	0	1	0		
235	315425268	BAJ46936.1	thermosome	551	512	100	1	1	1	1	1	0	0	M	
236	315425271	BAJ46939.1	3-hydroxyisobutyrate dehydrogenase	263	237	100	0	1	1	0	0	1	1		
237	315425272	BAJ46940.1	chorismate mutase prephenate dehydratase	226	186	100	0	1	1	1	1	1	1		
238	315425274	BAJ46942.1	citrate transporter	465	243	82	0	1	1	0	0	1	1		
239	315425276	BAJ46944.1	dehydratase family protein	534	506	100	0	1	1	1	0	0	1		
240	315425277	BAJ46945.1	dTMP kinase	200	91	100	0	1	1	0	1	1	1		
241	315425280	BAJ46948.1	glutamyl-tRNAGln amidotransferase subunit D	67	67	4	1	0	0	0	0	0	0		
242	315425281	BAJ46949.1	metal dependent phosphohydrolase	388	167	100	0	1	1	1	1	0	1	M2	
243	315425282	BAJ46950.1	nitrogen fixation protein NifU	131	95	100	0	1	1	1	1	0	1		B<=>A
244	315425283	BAJ46951.1	hypothetical conserved protein	291	249	13	0	1	1	0	0	0	1		A=>B
245	315425284	BAJ46952.1	hypothetical conserved protein	246	269	11	0	1	1	0	0	0	1		A=>B
246	315425285	BAJ46953.1	ABC-2 type transport system ATP-binding protein	238	179	100	0	1	1	0	0	1	1		A<=>B
247	315425286	BAJ46954.1	ABC transporter related protein	421	399	12	1	1	1	0	0	0	0		
248	315425287	BAJ46955.1	CTP synthase	546	452	100	0	1	1	1	0	0	1	M2	A=>B
249	315425288	BAJ46956.1	EmrBQacA family drug resistance transporter	466	288	100	0	1	0	0	0	1	1		B=>AA
250	315425289	BAJ46957.1	thiol-disulfide oxidoreductase	197	138	4	0	0	0	0	0	0	1		
251	315425291	BAJ46959.1	glutamine synthetase	437	417	100	0	1	1	1	0	0	1	M2	
252	315425293	BAJ46960.1	maltooligosaccharide transport system ATP-binding protein	292	258	100	0	1	1	0	0	0	1		
253	315425294	BAJ46961.1	multiple sugar transport system permease protein	289	170	100	0	1	0	0	0	0	1		
254	315425295	BAJ46962.1	inner membrane protein MalF	288	161	100	0	1	1	0	0	0	1		
255	315425297	BAJ46964.1	alanine racemase domain protein	360	150	100	0	1	0	0	0	1	1		
256	315425298	BAJ46965.1	luciferase family protein	332	119	89	0	1	0	1	0	0	1		
257	315425300	BAJ46967.1	N5N10-methylenetetrahydromethanopterin reductase	357	107	100	0	1	0	1	0	0	1		
258	315425301	BAJ46968.1	thermosome	528	512	100	1	1	1	1	1	0	0	M2	
259	315425302	BAJ46969.1	enolase	401	400	100	0	1	0	0	0	0	1		B=>AA
260	315425303	BAJ46970.1	aminoglycoside phosphotransferase	326	325	5	0	0		0	0	0	0		B=>AA
261	315425304	BAJ46971.1	oxidoreductase NAD-binding protein	331	127	100	0	1	1	0	0	1	1		
262	315425307	BAJ46974.1	membrane dipeptidase	279	146	100	0	1	1	0	0	1	1		

263	315425309	BAJ46976.1	multiple sugar transport system permease protein	295	172	100	0	1	1	0	0	1	1			B=>AA
264	315425310	BAJ46977.1	multiple sugar transport system permease protein	300	163	100	0	1	1	0	0	0	1			B=>AA
265	315425311	BAJ46978.1	maltooligosaccharide transport system ATP-binding protein	362	262	100	0	1	1	0	0	0	1			
266	315425312	BAJ46979.1	maltooligosaccharide transport system ATP-binding protein	364	260	100	0	1	1	0	1	0	1			
267	315425313	BAJ46980.1	oxidoreductase	363	93	100	0	1	1	0	1	1	1			
268	315425314	BAJ46981.1	oxidoreductase	339	126	100	0	1	1	0	1	1	1			
269	315425315	BAJ46982.1	NADH-dependent dehydrogenase	310	129	100	0	1	1	0	1	1	1			
270	315425316	BAJ46983.1	dimethylmenaquinone methyltransferase	225	122	37	0	1	1	0	1	1	1			
271	315425317	BAJ46984.1	ornithine cyclodeaminase	325	155	100	0	1	1	0	1	1	1			
272	315425341	BAJ47007.1	hypothetical conserved protein	418	164	100	0	1	1	0	1	0	1	M		
273	315425344	BAJ47010.1	molybdopterin biosynthesis protein MoeB	386	271	100	0	0	0	0	0	1	1			B=>AA
274	315425346	BAJ47012.1	NH3-dependent NAD+ synthetase	278	166	100	0	1	1	1	1	0	1	M2		A=>B
275	315425347	BAJ47013.1	acyl-CoA dehydrogenase	421	285	100	0	1	1	0	1	0	1			
276	315425348	BAJ47014.1	DNA-binding protein Alba	94	90	8	1	1	1	0	0	0	0			
277	315425350	BAJ47016.1	FAD dependent oxidoreductase	585	304	34	1	1	1	1	0	0	0	M2		
278	315425351	BAJ47017.1	mRNA 3-end processing factor	222	143	50	0	1	1	1	1	0	1	M2		
279	315425353	BAJ47019.1	circadian clock protein KaiC	260	144	100	0	1	1	0	1	0	1	M2		
280	315425354	BAJ47020.1	hypothetical conserved protein	313	212	18	0	1	0	0	0	1	1			
281	315425357	BAJ47023.1	phosphoribosyl formimino-5-aminoimidazole isomerase	241	134	100	0	1	1	1	0	0	1			
282	315425358	BAJ47024.1	imidazole glycerol phosphate synthase cyclase subunit HisF	254	240	100	0	1	1	1	0	0	1	M3		
283	315425359	BAJ47025.1	phosphoribosyl-AMP cyclohydrolase	106	87	100	0	1	1	1	0	0	1			
284	315425360	BAJ47026.1	imidazole glycerol phosphate synthase glutamine amidotransferase subunit HisH	200	127	100	0	1	1	1	0	1	1			
285	315425361	BAJ47027.1	ATP phosphoribosyltransferase	336	209	100	0	1	1	1	0	1	1	M		A=>B
286	315425362	BAJ47028.1	imidazole glycerol phosphate dehydratase	192	164	100	0	1	0	1	0	1	1			
287	315425366	BAJ47032.1	Mg-protoporphyrin IX monomethyl ester oxidative cyclase	459	137	100	0	1	1	0	0	0	1			
288	315425368	BAJ47034.1	glycosyl transferase family protein	602	324	100	0	0	0	1	0	1	1			B=>AA+T
289	315425373	BAJ47039.1	mRNA 3-end processing factor	638	508	100	1	1	1	1	1	0	0	M		
290	315425374	BAJ47040.1	proteasome beta subunit	223	165	100	1	1	1	1	1	0	0	M		
291	315425376	BAJ47042.1	N utilization substance protein A	144	107	99	1	1	1	1	1	0	0	M		
292	315425378	BAJ47044.1	DNA-directed RNA polymerase subunit A	1254	713	100	0	1	1	1	1	1	0	M		
293	315425379	BAJ47045.1	DNA-directed RNA polymerase subunit beta	250	544	100	0	1	1	1	1	1	0	M		
294	315425393	BAJ47058.1	molecular chaperone DnaK	592	535	100	0	1	0	1	0	1	1			
295	315425394	BAJ47059.1	molecular chaperone DnaJ	370	282	100	0	1	0	1	0	1	1			
296	315425395	BAJ47060.1	hypothetical protein HGMM_F51C10C15	158	86	35	1	1	1	0	0	0	0			EA<=>AA
297	315425398	BAJ47063.1	conserved hypothetical protein	143	131	5	1	0	0	1	0	0	0			
298	315425400	BAJ47065.1	conserved hypothetical protein	219	144	32	0	1	1	0	0	1	1			
299	315425402	BAJ47067.1	TatD DNase family protein	244	149	100	0	1	1	0	1	1	1			
300	315425405	BAJ47070.1	archaeal cell division control protein 6	396	254	100	1	1	1	1	1	0	0	M		
301	315425420	BAJ47084.1	glyceraldehyde-3-phosphate dehydrogenase NADP	487	288	100	0	1	1	0	0	1	1			
302	315425425	BAJ47089.1	transcriptional regulator	200	68	100	0	1	1	0	0	0	1			
303	315425427	BAJ47091.1	Cu2+-exporting ATPase	845	618	100	0	1	0	0	0	0	1			
304	315425428	BAJ47092.1	Cu2+-exporting ATPase	55	46	4	0	1	1	0	0	0	1			
305	315425431	BAJ47095.1	molybdopterin oxidoreductase molybdopterin binding subunit	1153	369	58	0	1	1	0	0	1	1			
306	315425432	BAJ47096.1	molybdopterin oxidoreductase membrane subunit	391	299	19	0	1	1	0	0	1	1			
307	315425433	BAJ47097.1	molybdopterin oxidoreductase iron-sulfur binding subunit	223	133	100	0	1	1	0	0	1	1			
308	315425434	BAJ47098.1	small subunit ribosomal protein S12	142	138	100	0	1	1	1	1	1	0	M		
309	315425435	BAJ47099.1	molybdenum cofactor biosynthesis protein A	347	213	100	0	0	0	0	0	1	1			B=>AA
310	315425436	BAJ47100.1	cytochrome bb6 domain-containing protein	449	201	94	0	1	1	1	1	1	1			
311	315425439	BAJ47103.1	pyruvate ferredoxin oxidoreductase beta subunit	324	234	100	0	1	1	0	1	0	1			
312	315425440	BAJ47104.1	pyruvate ferredoxin oxidoreductase alpha subunit	410	289	100	0	1	1	0	1	0	1			
313	315425441	BAJ47105.1	pyruvate ferredoxin oxidoreductase gamma subunit	294	144	100	0	1	1	0	1	0	0	M		
314	315425442	BAJ47106.1	acetyl-CoA synthetase ADP-forming alpha and beta subunit	703	441	100	0	1	1	1	1	1	1			
315	315425447	BAJ47111.1	lysine biosynthesis enzyme LysX	284	140	100	0	1	1	1	1	0	1			
316	315425448	BAJ47112.1	peptidase zinc-dependent	176	83	68	0	1	1	1	1	0	1			
317	315425451	BAJ47115.1	AAA family ATPase	224	632	100	1	1	1	1	1	0	0	M2		
318	315425454	BAJ47117.1	AAA family ATPase	726	614	100	1	1	1	1	1	0	0	M2		
319	315425456	BAJ47119.1	iron complex ABC transporter substrate-binding protein	295	106	91	0	1	1	1	0	1	1			
320	315425457	BAJ47120.1	undecaprenyl-diphosphatase	255	151	86	0	1	1	0	1	0	1			
321	315425463	BAJ47126.1	ribonucleoside-diphosphate reductase alpha chain	834	424	100	0	1	1	1	1	0	1			
322	315425464	BAJ47127.1	osmotically inducible protein OsmC	145	119	73	0	1	0	0	0	1	1			B=>AA
323	315425465	BAJ47128.1	conserved hypothetical protein	176	69	79	0	1	1	1	0	1	1	M		B=>A
324	315425466	BAJ47129.1	Fe-S cluster assembly protein SufB	466	419	100	0	1	1	1	0	0	1	M3		
325	315425467	BAJ47130.1	xanthine dehydrogenase	759	378	100	0	1	1	0	0	0	1			
326	315425468	BAJ47131.1	aminotransferase	382	288	100	0	1	1	1	0	0	1	M2		
327	315425469	BAJ47132.1	conserved hypothetical protein	145	123	23	0	1	1	0	0	0	1			
328	315425470	BAJ47133.1	threonylalanyl tRNA synthetase	239	143	100	0	1	1	0	1	0	1			
329	315425471	BAJ47134.1	GTPase	407	323	100	1	1	1	1	1	0	0	M		
330	315425472	BAJ47135.1	sugar kinase	290	140	100	0	1	1	1	0	1	1			
331	315425476	BAJ47139.1	hypothetical protein HGMM_F35A09C25	340	293	7	1	1	1	0	1	0	0			
332	315425478	BAJ47141.1	voltage-gated potassium channel	264	224	8	0	1	1	0	0	0	1			
333	315425479	BAJ47142.1	major facilitator superfamily MFS	374	281	44	0	1	1	0	1	0	1			CA=>EA
334	315425480	BAJ47143.1	threonine dehydratase	316	218	100	0	1	1	1	0	1	1			
335	315425482	BAJ47145.1	elongation factor EF-1 alpha subunit	526	322	100	0	1	1	0	1	1	0	M2		
336	315425499	BAJ47161.1	ornithine carbamoyltransferase	333	250	100	0	1	1	1	0	0	1			
337	315425500	BAJ47162.1	hypothetical protein HGMM_F42C08C15	302	124	55	1	1	1	1	0	0	0	M		
338	315425501	BAJ47163.1	large subunit ribosomal protein L11	162	145	100	1	1	1	1	1	0	0	R		
339	315425502	BAJ47164.1	large subunit ribosomal protein L1	218	169	99	0	1	1	1	1	1	0	R		
340	315425503	BAJ47165.1	large subunit ribosomal protein L10	287	213	100	0	1	1	1	1	1	0	R		
341	315425506	BAJ47168.1	ABC-2 type transport system ATP-binding protein	245	154	100	0	1	1	0	1	1	1			
342	315425507	BAJ47169.1	large subunit ribosomal protein L12	103	102	24	1	1	1	0	0	0	0	M3		
343	315425508	BAJ47170.1	conserved hypothetical protein	366	227	73	0	1	1	0	1	0	1	M3		
344	315425510	BAJ47172.1	conserved hypothetical protein	299	188	7	1	1	1	0	0	0	0			
345	315425511	BAJ47173.1	molybdopterin-guanine dinucleotide biosynthesis protein B	175	130	20	0	1	0	0	0	0	1			
346	315425513	BAJ47175.1	ribose 5-phosphate isomerase A	224	151	100	0	1	1	0	0	1	1	M2		EA=>T // EA=>E
347	315425516	BAJ47178.1	ubiquinonemenaquinone biosynthesis methyltransferase	234	105	45	0	1	1	1	1	1	1	M3		
348	315425521	BAJ47183.1	conserved hypothetical protein	220	243	22	1	1	0	0	0	0	0	M3		
349	315425524	BAJ47185.1	short-chain dehydrogenasereductase	259	108	100	0	1	1	1	0	1	1			
350	315425525	BAJ47186.1	quinoprotein alcohol dehydrogenase	738	298	74	0	1	0	0	0	0	1			B=>AA
351	315425527	BAJ47188.1	kelch repeat protein	400	139	100	0	0	0	1	0	1	1			
352	315425528	BAJ47189.1	multiple sugar ABC transporter permease	279	211	23	0	1	0	0	0	0	1			B=>AA
353	315425533	BAJ47194.1	pyrroloquinoline-quinone synthase	208	218	40	0	0	0	0	0	0	1			B=>AA
354	315425535	BAJ47196.1	iron complex ABC transporter substrate-binding protein	449	121	63	0	1	0	0	0	1	1			B=>AA
355	315425536	BAJ47197.1	iron complex ABC transporter permease	564	240	85	0	1	1	0	0	0	1			B=>AA
356	315425537	BAJ47198.1	conserved hypothetical protein	314	293	5	1	1	0	0	0	0	0			
357	315425538	BAJ47199.1	ABC transporter ATP-binding protein	370	230	100	0	1	1	0	0	0	1			B=>AA
358	315425543	BAJ47204.1	methylated-DNA-	223	91	100	0	1	0	0	0	1	1			B=>AA
359	315425544	BAJ47205.1	3-oxoacyl-	266	142	100	0	1	1	0	1	1	1			
360	315425546	BAJ47207.1	DNA topoisomerase III	606	300	100	0	1	1	0	0	1	0	M		
361	315425551	BAJ47212.1	gylcosyl transferase group 1	456	75	100	0	1	1	0	0	1	1			
362	315425553	BAJ47214.1	conserved hypothetical protein	250	219	5	0	0	0	0	0	0	1			
363	315425554	BAJ47215.1	UDP-glucose 4-epimerase	318	148	100	0	1	1	1	1	0	1			
364	315425557	BAJ47218.1	large subunit ribosomal protein L21e	95	94	24	1	1	1	0	1	0	0	M3		
365	315425560	BAJ47221.1	3-hydroxybutyryl-CoA dehydratase	265	186	100	0	1	1	0	0	1	1			B=>AA
366	315425561	BAJ47222.1	formyl-CoA transferase	400	296	100	0	1	1	0	0	1	1			
367	315425562	BAJ47223.1	cyclase family protein	264	110	42	0	1	1	0	0	1	1			B=>AA
368	315425567	BAJ47227.1	acyl-CoA dehydrogenase	244	251	100	0	1	1	0	0	1	1			
369	315425568	BAJ47228.1	major facilitator superfamily MFS	373	268	32	0	1	0	0	0	0	1			
370	315425570	BAJ47230.1	hypothetical protein HGMM_F15E11C04	759	214	23	0	1	0	0	0	1	1			
371	315425571	BAJ47231.1	ornithine cyclodeaminase	329	170	100	0	1	1	0	0	1	1			
372	315425573	BAJ47233.1	conserved hypothetical protein	249	109	53	0	1	0	1	0	1	1			
373	315425574	BAJ47234.1	molybdopterin binding oxidoreductase large subunit	768	520	100	0	1	1	0	0	0	1			
374	315425575	BAJ47235.1	amidohydrolase	394	379	14	0	0	0	0	0	0	1			B=>AA
375	315425576	BAJ47236.1	luciferase family protein	328	102	100	0	1	1	0	0	0	1			
376	315425577	BAJ47237.1	branched-chain amino acid ABC transporter substrate-binding protein	405	151	85	0	1	0	0	1	1	1			
377	315425578	BAJ47238.1	branched-chain amino acid ABC transporter ATP-binding protein	248	172	100	0	1	1	0	1	0	1			
378	315425579	BAJ47239.1	branched-chain amino acid ABC transporter ATP-binding protein	239	194	100	0	1	1	0	1	0	1			
379	315425581	BAJ47241.1	branched-chain amino acid ABC transporter permease	322	261	11	0	1	0	0	0	0	1			B=>AA
380	315425582	BAJ47242.1	conserved hypothetical protein	202	158	36	0	0	0	0	0	0	1			B=>AA
381	315425588	BAJ47248.1	ABC transporter ATP-binding protein	304	136	100	0	1	1	0	1	0	1			
382	315425590	BAJ47250.1	major facilitator superfamily MFS	451	268	100	0	1	1	0	1	0	1			
383	315425592	BAJ47252.1	multiple sugar ABC transporter substrate-binding protein	492	241	18	0	1	0	0	0	1	1			
384	315425609	BAJ47268.1	signal recognition particle subunit SRP19	95	89	4	1	0	1	0	0	0	0			
385	315425611	BAJ47270.1	transcription initiation factor TFIIB	235	287	20	0	1	0	1	1	1	0			
386	315425612	BAJ47271.1	conserved hypothetical protein	115	111	7	0	0	1	1	0	0	0			
387	315425613	BAJ47272.1	ribosomal RNA large subunit methyltransferase E FtsJ	224	135	100	0	1	0	1	0	1	1			
388	315425615	BAJ47274.1	nucleotide kinase	174	87	68	0	1	1	0	1	1	1	M3		
389	315425616	BAJ47275.1	ribonuclease HII	189	151	35	1	1	1	1	1	0	0	M3		
390	315425623	BAJ47282.1	DNA polymerase beta domain protein	173	156	13	0	1	1	0	0	0	1			
391	315425624	BAJ47283.1	conserved hypothetical protein	97	137	8	0	1	1	0	0	0	0			
392	315425625	BAJ47284.1	NAD-dependent epimerasedehydratase	321	211	100	0	1	0	0	0	1	1			
393	315425626	BAJ47285.1	glycosyl transferases group 1	367	85	100	0	1	1	0	0	0	1			
394	315425627	BAJ47286.1	conserved hypothetical protein	288	88	49	0	1	1	1	1	0	1			
395	315425628	BAJ47287.1	conserved hypothetical protein	496	466	4	0	0	1	1	0	0	0			

396	315425632	BAJ47291.1	glycosyl transferase family 1	321	85	100	0	1	1	0	0	0	1		
397	315425633	BAJ47292.1	UDP-glucose 6-dehydrogenase	437	318	100	0	1	1	0	0	1	1	M2	A=>T
398	315425634	BAJ47293.1	UTP--glucose-1-phosphate uridylyltransferase	288	202	100	0	1	1	0	0	0	1	M2	A=>B
399	315425636	BAJ47295.1	conserved hypothetical protein	135	107	23	1	1	1	0	0	0	0		
400	315425637	BAJ47296.1	conserved hypothetical protein	136	115	8	1	0	1	0	0	0	0		
401	315425639	BAJ47298.1	V-type H+-transporting ATPase subunit D	227	139	100	0	1	1	1	0	1	1	M	A=>B
402	315425642	BAJ47301.1	prolyl-tRNA synthetase	484	339	100	0	1	1	1	1	1	1	M	
403	315425643	BAJ47302.1	peptidase M48	322	146	100	0	1	1	0	1	0	1		
404	315425644	BAJ47303.1	small subunit ribosomal protein S26e	94	94	18	1	0	1	1	0	0	0		
405	315425645	BAJ47304.1	D-3-phosphoglycerate dehydrogenase	310	284	100	0	1	1	1	0	1	1	M2	A=>B
406	315425646	BAJ47305.1	conserved hypothetical protein	286	179	9	1	0	1	0	0	0	0		
407	315425648	BAJ47307.1	carboxyltransferase	401	307	100	0	1	1	1	1	0	1	M	
408	315425649	BAJ47308.1	fumarylacetoacetate hydrolase family protein	321	174	100	0	1	1	1	0	1	1	M3	A=>B
409	315425650	BAJ47309.1	phenylacetate-CoA ligase	436	372	100	0	1	0	0	0	0	1		
410	315425651	BAJ47310.1	ornithine cyclodeaminase	325	121	100	0	1	1	0	0	1	1		
411	315425661	BAJ47319.1	large subunit ribosomal protein L2	154	217	100	0	1	1	0	1	1	0	R	
412	315425662	BAJ47320.1	hypothetical protein HGMM_F36E03C02	95	84	5	1	0	1	0	0	0	0		
413	315425663	BAJ47321.1	small subunit ribosomal protein S19	139	116	100	0	1	1	1	1	1	0	R	
414	315425665	BAJ47323.1	hypothetical protein HGMM_F36E03C05	136	123	16	0	1	1	0	0	0	1		A=>B
415	315425666	BAJ47324.1	DNA polymerase beta subunit	129	126	5	1	0	1	0	0	0	0		
416	315425668	BAJ47326.1	undecaprenyl pyrophosphate synthetase	260	173	100	0	1	1	1	1	0	1	M3	
417	315425669	BAJ47327.1	hypothetical protein HGMM_F36E03C09	287	161	70	0	1	1	0	1	0	1		B=>AA
418	315425670	BAJ47328.1	chromosome segregation protein SMC	1119	380	100	0	1	0	1	0	1	1	M2	
419	315425672	BAJ47330.1	segregation and condensation protein B	198	109	40	0	1	0	1	1	0	1	M2	
420	315425673	BAJ47331.1	small subunit ribosomal protein S8e	129	108	52	1	1	1	1	1	0	0	R	
421	315425676	BAJ47333.1	DNA primase large subunit	357	140	68	1	1	1	1	1	0	0	Prismase	
422	315425677	BAJ47334.1	conserved hypothetical protein	155	96	93	0	1	1	1	0	1	1		
423	315425678	BAJ47335.1	hypothetical protein HGMM_F27H06C04	399	207	57	1	1	0	1	0	0	0	M	
424	315425681	BAJ47338.1	large subunit ribosomal protein L4e	272	199	100	0	1	1	1	1	1	0	R	
425	315425682	BAJ47339.1	large subunit ribosomal protein L23	86	81	100	0	1	1	0	1	1	0	R	
426	315425702	BAJ47358.1	5-carboxymethyl-2-hydroxymuconate isomerase	308	182	100	0	1	1	1	0	1	1		
427	315425703	BAJ47359.1	2-alkenal reductase	382	200	100	0	1	0	1	0	1	1		
428	315425704	BAJ47360.1	histidyl-tRNA synthetase	349	212	100	0	1	1	0	0	0	1	M	
429	315425705	BAJ47361.1	conserved hypothetical protein	267	215	9	0	1	0	0	1	0	1		
430	315425706	BAJ47362.1	peptidase M24	390	159	97	0	1	1	0	1	0	1	M2	A=>B
431	315425710	BAJ47366.1	conserved hypothetical protein	180	167	18	0	0	0	0	0	0	1		B=>C
432	315425711	BAJ47367.1	conserved hypothetical protein	140	95	100	0	1	1	0	1	1	1		
433	315425714	BAJ47370.1	PilT protein domain protein	150	135	12	1	1	1	0	0	0	0		
434	315425715	BAJ47371.1	conserved hypothetical protein	308	110	100	0	1	0	1	0	0	1		
435	315425716	BAJ47372.1	DNA polymerase I	715	277	100	0	1	1	1	1	0	1	M3	
436	315425717	BAJ47373.1	small subunit ribosomal protein S2	208	167	100	0	1	1	0	1	1	0	R	
437	315425719	BAJ47375.1	DNA-directed RNA polymerase subunit N	72	58	100	0	1	1	1	1	1	0	Polymerase	
438	315425720	BAJ47376.1	small subunit ribosomal protein S9	131	124	100	0	1	1	0	0	1	0	R	
439	315425721	BAJ47377.1	large subunit ribosomal protein L13	145	103	85	0	1	1	1	1	0	0	R	
440	315425723	BAJ47379.1	DNA-directed RNA polymerase subunit D	215	112	100	0	1	1	1	1	1	0	Polymerase	
441	315425724	BAJ47380.1	iron-sulfur cluster assembly accessory protein	117	98	100	0	0	1	1	0	0	1		B=>CA+T
442	315425725	BAJ47381.1	peroxiredoxin QBCP	152	107	100	0	1	1	1	0	1	1		
443	315425726	BAJ47382.1	endonuclease III	217	121	100	0	1	1	1	0	1	1		
444	315425729	BAJ47385.1	superoxide dismutase	203	150	100	0	1	1	1	0	1	1		
445	315425735	BAJ47391.1	large subunit ribosomal protein L3	289	251	100	0	1	1	1	1	1	0	R	
446	315425737	BAJ47392.1	DNA-directed RNA polymerase subunit M	110	61	100	0	1	1	1	1	1	0	Polymerase	
447	315425739	BAJ47394.1	exosome complex component CSL4	190	170	11	1	1	1	0	0	0	0		
448	315425741	BAJ47396.1	glycineD-amino acid oxidase deaminating	423	294	22	0	1	1	0	1	1	1		
449	315425742	BAJ47397.1	conserved hypothetical protein	147	140	9	0	1	0	0	0	0	1		
450	315425743	BAJ47398.1	DEADDEAH box helicase domain protein	1041	175	79	0	1	1	0	1	1	1		
451	315425744	BAJ47399.1	conserved hypothetical protein	184	183	5	1	1	0	1	0	0	0		
452	315425746	BAJ47401.1	conserved hypothetical protein	204	93	100	0	1	1	0	0	0	1	M3	A=>B // CA=>EA
453	315425748	BAJ47403.1	inorganic pyrophosphatase	720	548	100	0	1	1	0	1	1	1		
454	315425749	BAJ47404.1	sugar phosphate isomeraseepimerase	322	196	88	0	1	0	0	0	1	1		B=>AA // B=>EA
455	315425750	BAJ47405.1	conserved hypothetical protein	576	261	16	0	1	1	0	0	1	0		
456	315425754	BAJ47409.1	ABC transporter ATP-binding protein	475	208	18	0	1	1	0	0	0	1		
457	315425755	BAJ47410.1	hypothetical protein HGMM_F29F08C19	426	358	4	0	0	0	0	0	0	1		
458	315425760	BAJ47415.1	prephenate dehydrogenase	277	261	5	1	0	0	1	1	0	0		
459	315425761	BAJ47416.1	aspartate aminotransferase	464	219	100	0	1	1	1	1	1	1		
460	315425763	BAJ47418.1	ABC-2 type transport system ATP-binding protein	314	172	100	0	1	1	0	0	0	1		
461	315425764	BAJ47419.1	ABC-2 type transport system permease	253	251	8	1	0	1	0	0	0	0		
462	315425765	BAJ47420.1	small subunit ribosomal protein S10	105	95	100	0	1	1	1	1	1	0	R	
463	315425768	BAJ47423.1	F420-dependent N5N10-methenyltetrahydromethanopterin reductase	317	149	100	0	1	0	1	0	0	1		
464	315425770	BAJ47425.1	cobaltnickel ABC transporter ATP-binding protein	288	152	100	0	1	1	0	1	0	1		
465	315425771	BAJ47426.1	cobaltnickel ABC transporter ATP-binding protein	281	173	100	0	1	1	0	1	0	1		
466	315425789	BAJ47443.1	molybdenum hydroxylase family protein medium subunit	289	139	93	0	1	1	0	0	0	1		
467	315425790	BAJ47444.1	dihydropyrimidinase	480	315	100	0	1	1	0	0	1	1		
468	315425791	BAJ47445.1	conserved hypothetical protein	284	159	100	0	1	1	1	1	0	1	M3	A=>B
469	315425792	BAJ47446.1	mevalonate kinase	319	147	53	0	1	1	1	1	1	1		
470	315425793	BAJ47447.1	acetylglutamate kinase	260	101	100	0	1	1	1	1	1	1		
471	315425794	BAJ47448.1	isopentenyl-diphosphate delta-isomerase	358	251	100	0	1	1	0	0	0	1	M	
472	315425796	BAJ47450.1	large subunit ribosomal protein L7Ae	123	117	100	0	1	1	1	1	1	0	R	
473	315425798	BAJ47452.1	large subunit ribosomal protein L24e	57	61	10	1	1	1	0	0	0	0	R	
474	315425799	BAJ47453.1	nucleoside diphosphate kinase	138	128	100	0	1	1	1	0	1	1		
475	315425802	BAJ47456.1	DNA-directed RNA polymerase subunit E	187	147	97	1	1	1	1	1	0	0	Polymerase	
476	315425803	BAJ47457.1	DNA-directed RNA polymerase subunit E	63	60	8	1	1	0	1	0	0	0	Polymerase	
477	315425807	BAJ47461.1	small subunit ribosomal protein S6e	141	112	18	1	1	1	0	0	0	0	R	
478	315425808	BAJ47462.1	translation initiation factor IF-2 unclassified subunit	748	464	100	0	1	1	1	1	1	0	IFM2	
479	315425812	BAJ47466.1	small subunit ribosomal protein S15	144	146	100	0	1	1	1	1	1	0	R	
480	315425813	BAJ47467.1	archaea-specific RecJ-like exonuclease	488	225	100	1	1	1	1	1	0	0	M2	
481	315425815	BAJ47469.1	small subunit ribosomal protein S3Ae	206	143	95	0	1	1	1	1	1	0	R	
482	315425816	BAJ47470.1	conserved hypothetical protein	109	102	15	1	1	1	0	0	0	0		
483	315425817	BAJ47471.1	translation-associated GTPase	402	316	100	1	1	1	1	1	0	0	M	
484	315425818	BAJ47472.1	large subunit ribosomal protein L15e	143	173	100	0	1	1	1	1	1	0	R	
485	315425847	BAJ47500.1	sec-independent protein translocase protein TatC	287	171	100	0	1	1	0	1	0	1		
486	315425849	BAJ47502.1	3-oxoacyl-	256	138	100	0	1	1	1	1	1	1		
487	315425850	BAJ47503.1	small subunit ribosomal protein S13	161	124	100	0	1	1	1	1	1	0	R	
488	315425851	BAJ47504.1	5-methyltetrahydropteroyltriglutamate-- homocysteine methyltransferase	356	210	100	0	1	1	0	1	1	1		
489	315425852	BAJ47505.1	methylcobalaminhomocysteine methyltransferase	315	149	53	0	1	1	0	0	0	1		
490	315425853	BAJ47506.1	thioredoxin reductase NADPH	336	180	100	0	1	1	0	0	0	1		
491	315425854	BAJ47507.1	phosphate uptake regulator	330	123	100	1	1	1	1	1	0	0		
492	315425855	BAJ47508.1	large subunit ribosomal protein L22	152	120	100	0	1	1	1	1	1	0	R	
493	315425856	BAJ47509.1	small subunit ribosomal protein S3	203	175	100	0	1	1	1	1	1	0	R	
494	315425859	BAJ47512.1	small subunit ribosomal protein S17	109	87	100	0	1	1	1	1	1	0	R	
495	315425860	BAJ47513.1	large subunit ribosomal protein L14	144	140	100	0	1	1	1	0	1	0	R	
496	315425861	BAJ47514.1	large subunit ribosomal protein L24	119	109	100	0	1	1	1	1	1	0	R	
497	315425862	BAJ47515.1	small subunit ribosomal protein S4e	238	188	100	0	1	1	1	1	1	0	R	
498	315425863	BAJ47516.1	large subunit ribosomal protein L5	174	161	100	0	1	1	1	0	1	0	R	
499	315425864	BAJ47517.1	small subunit ribosomal protein S14	53	52	4	1	0	1	0	0	0	0	R	
500	315425865	BAJ47518.1	small subunit ribosomal protein S8	127	118	100	0	1	1	1	1	1	0	R	
501	315425866	BAJ47519.1	large subunit ribosomal protein L6	180	143	100	0	1	1	1	1	1	0	R	
502	315425867	BAJ47520.1	large subunit ribosomal protein L32e	128	96	100	0	1	1	1	0	1	0	R	
503	315425868	BAJ47521.1	large subunit ribosomal protein L19e	146	132	92	0	1	1	1	1	1	0	R	
504	315425869	BAJ47522.1	large subunit ribosomal protein L18	192	157	100	1	1	1	1	1	0	0	R	
505	315425870	BAJ47523.1	small subunit ribosomal protein S5	204	183	100	0	1	1	1	1	1	0	R	
506	315425871	BAJ47524.1	large subunit ribosomal protein L30	136	142	60	1	1	1	1	1	0	0	R	
507	315425872	BAJ47525.1	large subunit ribosomal protein L15	140	133	13	1	0	1	1	0	0	0	R	
508	315425876	BAJ47529.1	large subunit ribosomal protein L14e	107	91	24	1	1	1	0	0	0	0	R	
509	315425878	BAJ47531.1	signal-transduction protein	128	69	81	0	1	1	0	1	0	1		
510	315425879	BAJ47532.1	peptidenickel ABC transporter ATPase subunit	318	252	100	0	1	1	1	1	0	1		
511	315425883	BAJ47536.1	peptidenickel ABC transporter substrate-binding protein	838	321	44	0	1	1	1	1	0	1		
512	315425885	BAJ47538.1	peptidenickel ABC transporter permease	203	201	100	0	1	1	1	1	0	1		A=>B
513	315425887	BAJ47539.1	peptidenickel ABC transporter permease	171	191	100	0	1	1	1	1	0	1		
514	315425888	BAJ47540.1	peptidenickel ABC transporter ATP-binding protein	321	245	100	0	1	1	1	1	0	1		
515	315425894	BAJ47545.1	peptidenickel ABC transporter permease	145	348	6	1	0	1	0	0	0	0		
516	315425900	BAJ47551.1	FAD dependent oxidoreductase	392	153	100	0	1	0	0	0	1	1		
517	315425902	BAJ47553.1	conserved hypothetical protein	186	125	88	1	1	1	0	1	0	0	M	
518	315425904	BAJ47555.1	adenosylhomocysteinase	358	382	100	0	1	1	1	1	0	1	M	EA=>B
519	315425906	BAJ47556.1	conserved hypothetical protein	156	72	100	0	1	1	0	1	1	1		
520	315425907	BAJ47557.1	conserved hypothetical protein	145	79	100	0	1	1	0	0	1	1		
521	315425908	BAJ47558.1	glutamate dehydrogenase NADP+	401	296	100	0	1	1	0	1	1	1		
522	315425909	BAJ47559.1	cytosine deaminase	422	222	100	0	1	1	0	0	1	1		B=>A
523	315425910	BAJ47560.1	sugar ABC transporter permease	715	169	100	0	1	0	0	0	1	1		B=>AA
524	315425911	BAJ47561.1	sugar ABC transporter ATP-binding protein	510	346	100	0	1	1	0	1	0	1		
525	315425912	BAJ47562.1	basic membrane lipoprotein	340	237	71	0	1	1	0	0	1	1		
526	315425913	BAJ47563.1	conserved hypothetical protein	469	206	92	0	1	1	0	1	1	1		
527	315425914	BAJ47564.1	sugar ABC transporter permease	297	203	100	0	1	1	0	1	0	1		
528	315425915	BAJ47565.1	sugar ABC transporter permease	358	206	100	0	1	1	0	1	0	1		

529	315425916	BAJ47566.1	sugar ABC transporter ATP-binding protein	519	368	100	0	1	1	0	1	0	1		
530	315425919	BAJ47569.1	molybdenum hydroxylase family protein small subunit	155	120	100	0	1	1	0	0	0	1		
531	315425920	BAJ47570.1	molybdenum hydroxylase family protein medium subunit	289	137	100	0	1	1	0	0	0	1		
532	315425922	BAJ47572.1	propionyl-CoA carboxylase beta chain	467	472	100	0	1	1	0	0	0	1		
533	315425923	BAJ47573.1	molybdenum hydroxylase family protein large subunit	740	547	100	0	1	1	0	0	0	1		
534	315425927	BAJ47577.1	conserved hypothetical protein	516	420	15	0	0	0	0	0	0	1		
535	315425928	BAJ47578.1	cyclase family protein	288	193	13	0	1	0	0	0	0	1		
536	315425929	BAJ47579.1	conserved hypothetical protein	417	378	20	0	1	0	0	0	0	1		
537	315425933	BAJ47583.1	signal recognition particle receptor	302	232	100	0	1	1	1	0	0	1	M	
538	315425934	BAJ47584.1	conserved hypothetical protein	206	174	11	0	1	1	1	0	0	0		
539	315425935	BAJ47585.1	3-hydroxy-3-methylglutaryl-CoA reductase	432	294	67	0	1	1	1	0	1	1		
540	315425937	BAJ47587.1	conserved hypothetical protein	95	98	9	1	1	0	0	0	0	0		
541	315425943	BAJ47593.1	hypothetical protein HGMM_F52E01C38	468	222	18	0	1	1	0	0	0	1		
542	315425944	BAJ47594.1	6-pyruvoyl tetrahydrobiopterin synthase	156	154	5	1	0	0	1	0	0	0		
543	315425945	BAJ47595.1	transcription regulator ExsB related protein	213	141	100	0	1	1	1	0	1	1		
544	315425954	BAJ47603.1	conserved hypothetical protein	410	237	100	1	1	1	1	1	0	0	M3	
545	315425957	BAJ47606.1	ATP-binding protein involved in chromosome partitioning	357	252	100	0	1	0	1	0	1	1		B=>AA+T
546	315425958	BAJ47607.1	conserved hypothetical protein	387	109	76	0	1	1	0	0	0	1	M	
547	315425960	BAJ47609.1	Zn-dependent alcohol dehydrogenases	337	140	100	0	1	1	1	0	1	1		
548	315425963	BAJ47612.1	conserved hypothetical protein	140	94	54	0	1	1	0	0	1	0	M	
549	315425964	BAJ47613.1	conserved hypothetical protein	122	115	9	1	1	1	0	0	0	0		
550	315425971	BAJ47620.1	conserved hypothetical protein	130	128	5	0	1	1	0	0	0	1		
551	315425973	BAJ47622.1	site-specific DNA-methyltransferase adenine-specific	398	193	16	0	1	0	0	0	0	1		
552	315425974	BAJ47623.1	phosphoribosyl transferase domain protein	190	165	13	0	0	1	0	1	0	1		
553	315425977	BAJ47626.1	2-dehydro-3-deoxygluconokinase	320	138	100	0	1	1	0	1	1	1		
554	315425978	BAJ47627.1	dihydrodipicolinate synthetase	295	136	100	0	1	1	0	1	1	1		
555	315425979	BAJ47628.1	3-oxoacyl-	258	164	100	0	1	1	0	1	0	1		
556	315425980	BAJ47629.1	conserved hypothetical protein	429	86	78	0	1	1	1	0	1	1		
557	315425981	BAJ47630.1	molybdenum hydroxylase family protein large subunit	779	545	100	0	1	1	0	0	0	1		B=>A
558	315425985	BAJ47634.1	coenzyme F420-dependent N5N10-methenyltetrahydromethanopterin reductase	155	272	22	0	1	0	0	0	0	1		
559	315425987	BAJ47635.1	multiple sugar ABC transporter substrate-binding protein	218	427	9	0	0	1	0	0		1		
560	315425988	BAJ47636.1	multiple sugar ABC transporter permease	291	169	100	0	1	0	0	0	0	1		
561	315425989	BAJ47637.1	multiple sugar ABC transporter permease	302	179	100	0	1	0	0	0	1	1		
562	315425992	BAJ47640.1	maltooligosaccharide ABC transporter ATP-binding protein	363	252	100	0	1	1	0	0	0	1		
563	315425995	BAJ47643.1	metal-dependent hydrolase	270	148	100	0	1	1	0	1	1	1	M	EA=>B
564	315425996	BAJ47644.1	molybdenum hydroxylase family protein large subunit	336	377	100	0	1	1	0	0	0	1		
565	315425997	BAJ47645.1	molybdenum hydroxylase family protein large subunit	451	368	100	0	1	1	0	0	0	1		
566	315425998	BAJ47646.1	molybdenum hydroxylase family protein small subunit	151	117	100	0	1	1	0	0	0	1		
567	315425999	BAJ47647.1	molybdenum hydroxylase family protein medium subunit	278	164	59	0	1	1	0	0	0	1		
568	315426000	BAJ47648.1	iron complex ABC transporter substrate-binding protein	412	181	50	0	1	1	0	0	0	1		
569	315426001	BAJ47649.1	iron complex ABC transporter permease	333	207	100	0	1	0	0	0	0	1		
570	315426002	BAJ47650.1	iron ABC transporter ATP-binding protein	403	138	100	0	1	0	0	0	0	1		
571	315426003	BAJ47651.1	nucleotidyl transferase	243	167	100	0	1	1	1	0	0	1		
572	315426005	BAJ47653.1	conserved hypothetical protein	129	113	19	0	1	0	1	0	0	1		
573	315426007	BAJ47655.1	transcriptional regulator PadR family	174	78	21	0	1	1	0	1	0	1		B=>AA
574	315426008	BAJ47656.1	antibiotic ABC transporter ATP-binding protein	332	221	100	0	1	1	1	1	0	1		
575	315426009	BAJ47657.1	ABC-2 type transport system permease	290	155	100	0	1	1	1	0	0	1		
576	315426010	BAJ47658.1	oxidoreductase	281	187	100	0	1	1	0	1	1	1		
577	315426011	BAJ47659.1	prophage MuMc02 terminase ATPase subunit	439	249	15	0	1	0	0	0	0	1		
578	315426017	BAJ47665.1	nuclease	187	118	13	1	1	1	0	0	0	0		
579	315426021	BAJ47669.1	serine protease	307	287	13	1	1	0	0	0	0	0		
580	315426049	BAJ47696.1	aldehyde dehydrogenase	492	351	100	0	1	1	0	0	1	1		
581	315426050	BAJ47697.1	fumarylacetoacetate FAA hydrolase	308	114	72	0	1	1	0	0	1	1		
582	315426051	BAJ47698.1	NAD-dependent epimerasedehydratase	320	187	100	0	1	1	1	0	1	1		
583	315426052	BAJ47699.1	24-dihydroxyhept-2-ene-17-dioic acid aldolase	251	115	100	0	1	1	0	0	1	1		
584	315426056	BAJ47703.1	conserved hypothetical protein	703	461	100	0	0	1	1	1	1	0		
585	315426057	BAJ47704.1	hypothetical protein HGMM_F09F10C26	571	377	12	0	1	0	0	0	0	0		
586	315426059	BAJ47706.1	conserved hypothetical protein	377	180	100	0	1	1	1	1	0	1		
587	315426070	BAJ47716.1	isocitrate dehydrogenase	345	266	100	0	1	1	1	0	1	1		
588	315426071	BAJ47717.1	molybdenum hydroxylase family protein large subunit	764	345	100	0	1	1	0	0	0	1		B=>A
589	315426072	BAJ47718.1	molybdenum hydroxylase family protein large subunit	770	333	100	0	1	1	0	0	1	1		B=>A
590	315426073	BAJ47719.1	dihydroorotase	441	255	100	0	1	1	0	0	1	1		
591	315426074	BAJ47720.1	protoheme IX farnesyltransferase	290	216	100	0	1	1	1	1	0	1		
592	315426081	BAJ47727.1	hypothetical protein HGMM_F13A09C12	266	166	10	0	0	1	0	0	0	1		
593	315426086	BAJ47732.1	hydroxymethylpyrimidine transporter	428	348	19	0	1	1	0	1	0	1		
594	315426089	BAJ47735.1	conserved hypothetical protein	127	109	20	0	1	1	0	1	0	1		B=>A
595	315426090	BAJ47736.1	pyruvate ferredoxin oxidoreductase beta subunit	311	235	100	0	1	1	0	1	0	1	M	
596	315426091	BAJ47737.1	pyruvate ferredoxin oxidoreductase alpha subunit	390	309	100	0	1	1	0	1	0	1	M2	
597	315426092	BAJ47738.1	pyruvate ferredoxin oxidoreductase delta subunit	86	61	100	0	1	1	0	1	0	1	M3	
598	315426093	BAJ47739.1	pyruvate ferredoxin oxidoreductase gamma subunit	186	139	100	0	1	1	0	1	0	1	M2	
599	315426094	BAJ47740.1	conserved hypothetical protein	299	293	4	1	1	0	0	0	0	0		
600	315426095	BAJ47741.1	conserved hypothetical protein	341	268	23	1	1	1	0	0	0	0		
601	315426097	BAJ47743.1	zinc-dependent protease TIdDPmbA family	444	104	100	0	1	1	1	0	0	1	M2	
602	315426098	BAJ47744.1	adenylosuccinate lyase	454	381	100	0	1	1	0	0	1	1		B=>A
603	315426108	BAJ47753.1	succinate dehydrogenase flavoprotein subunit	571	441	100	0	1	1	1	1	0	1		
604	315426109	BAJ47754.1	succinate dehydrogenase cytochrome b-556 subunit	156	141	6	0	0	0	1	0	0	1		
605	315426111	BAJ47756.1	succinate dehydrogenase iron-sulfur protein	237	178	100	0	1	1	1	1	1	1		
606	315426113	BAJ47758.1	phosphopantetheine adenylyltransferase	170	105	92	0	1	1	1	0	1	0	M	
607	315426116	BAJ47761.1	conserved hypothetical protein	290	278	6	1	1	0	0	0	0	0		
608	315426117	BAJ47762.1	actinactin family protein	422	417	11	1	0	1	0	1	0	0		
609	315426119	BAJ47764.1	conserved hypothetical protein	309	194	23	1	1	1	0	0	0	0		
610	315426120	BAJ47765.1	conserved hypothetical protein	206	190	4	0	1	0	0	0	0	1		
611	315426121	BAJ47766.1	methylthioribose-1-phosphate isomerase	345	251	100	0	1	1	1	0	0	1	M2	
612	315426122	BAJ47767.1	FAD-dependent pyridine nucleotide-disulphide oxidoreductase	380	115	100	0	1	1	1	1	1	1		
613	315426123	BAJ47768.1	aconitate hydratase	909	743	100	0	1	1	0	0	0	1		B<=>A
614	315426127	BAJ47772.1	ribosome biogenesis protein	221	124	100	0	1	1	1	1	1	0	M	
615	315426129	BAJ47774.1	small subunit ribosomal protein S19e	143	123	100	0	1	1	0	1	1	0	R	
616	315426130	BAJ47775.1	phosphoribosyltransferase	214	125	100	0	1	0	0	0	1	1		B=>AA
617	315426131	BAJ47776.1	spermine synthase	545	131	28	0	1	1	1	0	0	1		B=>A
618	315426137	BAJ47782.1	pyruvate dehydrogenase E1 component subunit beta	332	290	100	0	1	1	0	0	1	1		
619	315426138	BAJ47783.1	pyruvate dehydrogenase E1 component subunit alpha	327	246	100	0	1	1	0	0	1	1		
620	315426139	BAJ47784.1	UDP-glucose 4-epimerase	306	172	100	0	1	1	1	1	1	1		
621	315426140	BAJ47785.1	glycosyl transferase family 1	343	86	100	0	1	1	1	1	1	1		
622	315426148	BAJ47792.1	large subunit ribosomal protein L10e	167	139	100	0	1	1	1	0	1	0	R	
623	315426149	BAJ47793.1	zinc-dependent protease TIdDPmbA family	430	248	100	0	1	1	1	0	0	1	M2	
624	315426150	BAJ47794.1	conserved hypothetical protein	110	109	6	1	0	0	1	0	0	0		
625	315426151	BAJ47795.1	conserved hypothetical protein	207	189	15	0	1	1	0	0	0	1		
626	315426154	BAJ47798.1	glutamine--scyllo-inositol transaminase	398	245	100	0	1	1	0	1	0	1		A<=>B
627	315426155	BAJ47799.1	conserved hypothetical protein	219	217	6	0	0	0	0	0	0	1		
628	315426156	BAJ47800.1	acetylornithine deacetylase	373	98	100	0	1	1	0	1	1	1		
629	315426158	BAJ47802.1	glycine cleavage system H protein	140	106	100	0	1	1	0	1	1	1		
630	315426160	BAJ47804.1	conserved hypothetical protein	155	70	100	0	1	1	0	1	1	1		
631	315426161	BAJ47805.1	CDP-diacylglycerol--glycerol-3-phosphate 3-phosphatidyltransferase	199	111	62	0	1	1	1	0	0	1		A=>B
632	315426162	BAJ47806.1	sterol-binding domain protein	134	120	5	1	0	1	0	0	0	0		
633	315426163	BAJ47807.1	peroxiredoxin QBCP	156	107	100	0	1	1	0	0	1	1		
634	315426164	BAJ47808.1	DNA polymerase bacteriophage-type	200	126	100	0	1	1	1	1	0	1		
635	315426165	BAJ47809.1	DNA-3-methyladenine glycosylase	187	106	100	0	1	1	1	0	1	1		
636	315426166	BAJ47810.1	histidinol dehydrogenase	456	295	100	0	1	1	1	0	1	1	M3	A=>B
637	315426169	BAJ47813.1	methyltransferase	318	151	100	1	1	1	1	0	0	0	M3	
638	315426170	BAJ47814.1	citrate transporter	434	241	86	0	1	1	0	0	1	1		A<=>B
639	315426171	BAJ47815.1	nuclear ribonucleoprotein	143	138	10	0	0	1	0	0	0			
640	315426172	BAJ47816.1	DNA topoisomerase I	539	384	100	0	0	0	1	0	1	0		
641	315426173	BAJ47817.1	conserved hypothetical protein	166	95	58	0	1	1	0	1	0	1		A<=>B
642	315426174	BAJ47818.1	conserved hypothetical protein	81	74	22	0	1	1	0	1	0	1		
643	315426175	BAJ47819.1	FAD-dependent pyridine nucleotide-disulphide oxidoreductase	397	170	100	0	1	1	1	1	1	1		
644	315426176	BAJ47820.1	transcriptional regulator ArsR family	105	63	31	0	1	1	0	0	0	1		
645	315426177	BAJ47821.1	6-phosphogluconate dehydrogenase	466	417	100	0	0	0	0	0	1	1		
646	315426178	BAJ47822.1	AAA family ATPase	324	186	74	1	1	1	0	1	0	0		
647	315426180	BAJ47824.1	2-isopropylmalate synthase	507	315	100	0	1	1	1	1	0	1	M2	
648	315426184	BAJ47828.1	conserved hypothetical protein	154	96	65	0	1	1	0	0	1	1		
649	315426185	BAJ47829.1	3-isopropylmalateR-2-methylmalate dehydratase small subunit	214	157	100	0	0	0	1	0	1	1		B=>AA
650	315426186	BAJ47830.1	DNA polymerase beta subunit	102	98	8	1	1	1	0	1	0	0		
651	315426187	BAJ47831.1	conserved hypothetical protein	143	83	55	0	1	1	0	1	0	1		
652	315426188	BAJ47832.1	ABC transporter ATP-binding protein	316	166	100	0	1	1	0	1	0	1		
653	315426189	BAJ47833.1	ABC-2 type transport system permease	414	235	9	0	1	0	0	0	0	1		
654	315426191	BAJ47835.1	3-hydroxybutyryl-CoA dehydrogenase	317	189	100	0	1	1	0	0	1	1		
655	315426192	BAJ47836.1	pyruvate kinase	453	311	100	0	1	1	0	0	1	1		
656	315426195	BAJ47838.1	basic membrane protein A	494	183	100	0	1	1	0	1	0	1		
657	315426196	BAJ47839.1	ABC transporter permease	358	214	100	0	1	1	0	1	0	1		
658	315426197	BAJ47840.1	simple sugar transporter permease	336	204	100	0	1	1	0	1	0	1		
659	315426199	BAJ47842.1	5-methylthioadenosine phosphorylase	237	149	100	0	1	1	0	1	1	1		CA=>EA
660	315426200	BAJ47843.1	Biotin operon repressor BirA	239	231	7	0	1	0	0	0	0	1		
661	315426202	BAJ47845.1	methionyl-tRNA synthetase	514	374	100	0	1	1	0	1	1	1	M	A=>B

662	315426203	BAJ47846.1	2-haloacid dehalogenase	232	109	39	0	1	1	0	0	1	1		
663	315426204	BAJ47847.1	pyruvate formate-lyase activating enzyme	378	175	56	0	1	1	0	1	1	1		
664	315426205	BAJ47848.1	fructose-bisphosphate aldolase class I	247	216	100	0	1	1	1	0	0	1		
665	315426208	BAJ47851.1	conserved hypothetical protein	137	125	12	1	1	1	0	0	0	0		
666	315426210	BAJ47853.1	Ca2+-transporting ATPase	891	626	100	0	1	1	0	0	0	1		B=>A
667	315426218	BAJ47861.1	conserved hypothetical protein	495	428	10	1	0	1	0	1	0	0		
668	315426221	BAJ47864.1	pyruvate dehydrogenase E2 component dihydrolipoamide acetyltransferase	258	254	100	0	1	1	0	0	1	1		
669	315426222	BAJ47865.1	ribonuclease HI	134	73	97	0	1	1	0	0	1	1		
670	315426248	BAJ47890.1	conserved hypothetical protein	297	176	89	0	1	1	0	1	0	1	M	
671	315426251	BAJ47893.1	conserved hypothetical protein	246	230	6	1	0	1	0	1	0	0		
672	315426254	BAJ47896.1	3-phosphoshikimate 1-carboxyvinyltransferase	429	249	100	0	1	1	1	1	1	1		
673	315426255	BAJ47897.1	shikimate kinase	287	144	74	1	1	1	1	1	0	0	M	
674	315426256	BAJ47898.1	shikimate 5-dehydrogenase	260	158	100	0	1	1	1	1	0	1		
675	315426257	BAJ47899.1	3-dehydroquinate dehydratase I	230	95	78	0	1	1	1	1	1	1		
676	315426258	BAJ47900.1	3-dehydroquinate synthase	333	194	97	0	1	0	1	0	1	1	M	A=>B
677	315426259	BAJ47901.1	fructose-bisphosphate aldolase class I	262	215	100	0	1	0	1	1	0	1	M2	A=>B
678	315426260	BAJ47902.1	aspartate-semialdehyde dehydrogenase	350	274	100	0	1	1	1	1	1	1	M3	
679	315426261	BAJ47903.1	molybdate transport system regulatory protein	182	173	6	1	1	0	0	0	0	0		
680	315426262	BAJ47904.1	conserved hypothetical protein	295	268	8	1	1	0	0	0	0	0		
681	315426263	BAJ47905.1	sulfate ABC transporter permease	231	223	7	1	1	0	0	0	0	0		
682	315426264	BAJ47906.1	sulfate ABC transporter ATP-binding protein	223	123	100	0	1	1	0	1	0	1		
683	315426265	BAJ47907.1	tryptophanyl-tRNA synthetase	386	330	100	0	1	1	1	1	1	0	M3	
684	315426266	BAJ47908.1	citrate synthase	364	273	100	0	1	1	0	0	1	1		
685	315426267	BAJ47909.1	2-methylcitrate dehydratase	454	235	100	0	1	1	0	0	1	1		
686	315426268	BAJ47910.1	simple sugar ABC transporter ATP-binding protein	511	365	100	0	1	1	0	0	0	1		
687	315426269	BAJ47911.1	conserved hypothetical protein	273	174	81	0	1	1	0	0	1	1		
688	315426270	BAJ47912.1	uridine phosphorylase	272	178	100	0	1	1	0	0	1	1		
689	315426271	BAJ47913.1	ribokinase	307	167	100	0	0	1	0	0	1	1		
690	315426272	BAJ47914.1	deoxyribose-phosphate aldolase	229	184	100	0	1	1	0	0	1	1		
691	315426273	BAJ47915.1	inosineuridine-preferring nucleoside hydrolase	311	164	100	0	1	1	0	0	1	1		
692	315426274	BAJ47916.1	cytidine deaminase	132	77	100	0	1	1	0	0	1	1		
693	315426275	BAJ47917.1	ribokinase	294	95	100	0	1	1	0	0	1	1		
694	315426294	BAJ47935.1	tryptophan synthase beta subunit	425	396	100	0	1	1	0	0	1	1		
695	315426295	BAJ47936.1	phosphoribosylanthranilate isomerase	207	99	41	0	1	0	0	0	1	1		
696	315426296	BAJ47937.1	indole-3-glycerol-phosphate synthase	232	135	100	0	1	1	0	0	0	1		
697	315426297	BAJ47938.1	anthranilate phosphoribosyltransferase	358	250	100	0	1	0	0	0	0	1		
698	315426298	BAJ47939.1	4a-hydroxytetrahydrobiopterin dehydratase	120	59	81	0	1	1	0	0	1	1		
699	315426305	BAJ47946.1	DNA repair protein RadASms	377	314	41	0	0	0	0	0	1	1		
700	315426306	BAJ47947.1	Fe-S oxidoreductase	564	318	100	0	1	1	1	1	0	1	M2	A=>B
701	315426307	BAJ47948.1	conserved hypothetical protein	100	98	7	1	0	1	0	0	0	0		
702	315426308	BAJ47949.1	conserved hypothetical protein	134	100	20	1	1	1	0	0	0	0		
703	315426311	BAJ47952.1	transcriptional regulator TrmB-like	263	64	74	0	1	1	1	0	0	1	M3	A=>B
704	315426338	BAJ47978.1	conserved hypothetical protein	105	99	4	1	1	0	0	0	0	0		
705	315426345	BAJ47985.1	conserved hypothetical protein	496	193	100	0	1	1	1	1	0	0		
706	315426346	BAJ47986.1	metallophosphoesterase	377	138	45	0	1	1	1	1	0	1	M2	
707	315426347	BAJ47987.1	exonuclease SbcC	758	132	88	0	1	1	1	1	0	1	M2	A=>B
708	315426351	BAJ47991.1	conserved hypothetical protein	467	230	13	1	0	1	0	0	0	0		
709	315426355	BAJ47995.1	conserved hypothetical protein	249	235	6	1	1	0	0	1	0	0		
710	315426356	BAJ47996.1	conserved hypothetical protein	614	483	5	0	0	0	0	0	0	1		
711	315426357	BAJ47997.1	conserved hypothetical protein	126	127	7	1	1	1	0	0	0	0		
712	315426359	BAJ47999.1	conserved hypothetical protein	553	281	5	1	1	0	0	0	0	0		
713	315426360	BAJ48000.1	tryptophan synthase alpha subunit	254	116	100	0	1	1	1	0	1	1		
714	315426373	BAJ48012.1	hypothetical protein HGMM_F28E01C13	508	165	100	0	1	0	0	0	1	1		
715	315426379	BAJ48018.1	1H-3-hydroxy-4-oxoquinaldine 24-dioxygenase	265	254	9	0	0	0	0	0	0	1		
716	315426381	BAJ48020.1	conserved hypothetical protein	140	132	12	1	1	1	0	0	0	0		
717	315426382	BAJ48021.1	hypothetical protein HGMM_F28E01C22	128	322	14	0	0	1	0	0	0	1		
718	315426384	BAJ48023.1	HEPN domain-containing protein	128	104	23	0	1	1	0	1	0	1		
719	315426386	BAJ48025.1	D-alanine-D-alanine ligase	640	35	100	0	1	1	1	0	1	1	M	
720	315426389	BAJ48028.1	conserved hypothetical protein	226	95	47	0	1	1	0	0	1	1		
721	315426392	BAJ48031.1	molybdopterin oxidoreductase Fe-S binding subunit	384	290	100	0	1	1	0	0	0	1		B=>A
722	315426393	BAJ48032.1	molybdopterin oxidoreductase molybdopterin binding subunit C-terminal	707	736	14	0	1	0	0	0	0	1		
723	315426394	BAJ48033.1	molybdopterin oxidoreductase molybdopterin binding subunit N-terminal	446	776	100	0	1	1	0	0	0	1		B=>A
724	315426395	BAJ48034.1	bacterio-opsin activator HTH domain protein	225	188	5	1	1	1	0	0	0	0		
725	315426396	BAJ48035.1	mercuric reductase	506	280	100	0	1	1	0	0	1	1		
726	315426397	BAJ48036.1	transcriptional regulator HxlR family	128	61	84	0	1	1	0	0	0	1		
727	315426398	BAJ48037.1	hypothetical protein HGMM_F28E01C38	472	407	6	1	1	1	0	0	0	0		
728	315426399	BAJ48038.1	conserved hypothetical protein	495	188	100	0	1	1	1	1	0	1		
729	315426400	BAJ48039.1	exonuclease SbcCD D subunit	383	142	42	0	1	1	0	1	0	1		
730	315426401	BAJ48040.1	exonuclease SbcCD C subunit	767	453	12	0	1	1	1	1	0	1		
731	315426403	BAJ48042.1	CRISPR-associated helicase Cas3	537	115	58	0	1	1	0	1	0	1		
732	315426418	BAJ48056.1	CRISPR-associated autoregulator DevR family	360	202	26	1	1	1	0	1	0	0	M3	
733	315426420	BAJ48058.1	CRISPR-associated protein Cas3 family	218	188	7	1	1	1	0	0	0	0		
734	315426422	BAJ48060.1	CRISPR-associated protein Cas1	320	104	100	0	1	1	0	1	0	1		
735	315426427	BAJ48065.1	CRISPR-associated RAMP family protein	213	199	7	1	0	1	0	0	0	0		
736	315426428	BAJ48066.1	conserved hypothetical protein	861	805	6	1	0	1	0	0	0	0		
737	315426429	BAJ48067.1	conserved hypothetical protein	316	242	4	1	0	1	0	0	0	0		
738	315426430	BAJ48068.1	CRISPR-associated RAMP family protein	296	128	24	0	0	1	0	0	0	1		
739	315426435	BAJ48073.1	thiamine-phosphate pyrophosphorylase	209	108	100	0	1	0	0	0	1	1		
740	315426436	BAJ48074.1	thiamine biosynthesis protein ThiG	232	228	100	0	0	0	0	0	1	1		B=>AA
741	315426438	BAJ48076.1	DsbA oxidoreductase	299	268	4	0	0	0	0	0	0	1		
742	315426439	BAJ48077.1	dithiol-disulfide isomerase	244	153	17	0	0	1	0	0	0	1		B=>AA
743	315426440	BAJ48078.1	hypothetical protein HGMM_F34A01C34	263	182	6	0	1	0	0	0	1	1		
744	315426441	BAJ48079.1	hydrogenase nickel incorporation protein HypA	142	129	21	0	1	1	0	1	0	1	M3	A=>B
745	315426442	BAJ48080.1	ATP-binding protein involved in chromosome partitioning	254	170	100	0	1	1	1	1	1	1		
746	315426443	BAJ48081.1	4Fe-4S ferredoxin iron-sulfur binding protein	375	148	68	0	1	0	0	0	0	1		
747	315426444	BAJ48082.1	transcriptional regulator CrpFnr family	162	153	6	0	0	0	0	0	0	1		
748	315426445	BAJ48083.1	oxidoreductase FADNADP-binding subunit	284	139	100	0	1	0	1	1	0	1		
749	315426446	BAJ48084.1	NADH ubiquinone oxidoreductase 20 kDa subunit	253	130	100	0	1	0	0	1	0	1		
750	315426447	BAJ48085.1	nickel-dependent hydrogenase large subunit	428	253	100	0	1	0	0	1	0	1		
751	315426449	BAJ48087.1	hydrogenase maturation protein HypF	248	450	97	0	1	1	0	1	0	1	M	
752	315426461	BAJ48094.1	hydrogenase maturation protein HypF	771	453	100	0	1	1	0	1	0	1	M	
753	315426462	BAJ48095.1	hydrogenase expressionformation protein HypD	384	253	100	0	1	1	0	1	0	1	M	
754	315426464	BAJ48097.1	hydrogenase expressionformation protein HypE	341	265	100	0	1	1	0	1	0	1	M	
755	315426466	BAJ48099.1	purine nucleosidase	308	173	100	0	1	0	0	0	1	1		
756	315426468	BAJ48101.1	cobaltnickel ABC transporter permease	269	178	23	0	1	0	0	1	0	1		
757	315426471	BAJ48104.1	ribokinase	320	172	100	0	0	1	0	0	1	1		B=>AA
758	315426472	BAJ48105.1	2-isopropylmalate synthase	524	424	100	0	1	1	0	0	1	1	M3	
759	315426475	BAJ48108.1	succinyl-CoA synthetase beta subunit	374	307	100	0	1	1	0	0	0	1	M2	
760	315426476	BAJ48109.1	succinyl-CoA synthetase alpha subunit	291	253	100	0	1	1	0	1	0	1	M2	
761	315426486	BAJ48118.1	conserved hypothetical protein	330	278	13	1	1	0	1	0	0	0		
762	315426487	BAJ48119.1	threonine synthase	424	317	100	0	1	1	1	0	0	1	M2	
763	315426491	BAJ48123.1	beta-lactamase domain protein	322	83	69	0	1	1	0	0	1	1		
764	315426492	BAJ48124.1	4-aminobutyrate aminotransferase	433	316	100	0	1	1	0	1	0	1		
765	315426493	BAJ48125.1	glutamate-1-semialdehyde-21-aminomutase	445	370	100	0	1	1	0	1	0	1	M3	
766	315426494	BAJ48126.1	aminotransferase	445	229	100	0	1	1	0	1	1	1		
767	315426495	BAJ48127.1	molybdenum hydroxylase family protein large subunit	763	559	100	0	1	1	0	0	0	1		
768	315426496	BAJ48128.1	saccharopine dehydrogenase	387	280	69	0	1	1	0	1	0	1		
769	315426497	BAJ48129.1	spermidineputrescine ABC transporter substrate-binding protein	386	193	100	0	1	1	0	0	0	1		
770	315426498	BAJ48130.1	spermidineputrescine ABC transporter permease	251	191	100	0	1	1	0	0	0	1		
771	315426499	BAJ48131.1	spermidineputrescine ABC transporter permease	286	213	100	0	1	0	0	0	0	1		
772	315426500	BAJ48132.1	spermidineputrescine ABC transporter ATP-binding protein	356	218	100	0	1	1	0	0	0	1		
773	315426501	BAJ48133.1	glutamate-1-semialdehyde-21-aminomutase	462	347	100	0	1	1	0	1	1	1		
774	315426502	BAJ48134.1	conserved hypothetical protein	373	114	100	0	1	1	1	0	1	1		
775	315426503	BAJ48135.1	saccharopine dehydrogenase NAD+ L-lysine forming	390	322	70	0	1	1	0	1	0	1		
776	315426504	BAJ48136.1	conserved hypothetical protein	300	179	55	0	1	1	0	1	0	1		
777	315426505	BAJ48137.1	peptidase M20	448	215	100	0	1	1	1	0	1	1		
778	315426507	BAJ48139.1	malate dehydrogenase	309	272	100	0	1	1	1	0	1	1		
779	315426508	BAJ48140.1	metallo-beta-lactamase protein	292	135	67	0	1	1	0	0	0	1		
780	315426509	BAJ48141.1	protein phosphatase	263	273	100	0	1	1	0	1	1	0		
781	315426510	BAJ48142.1	succinyl-CoA synthetase alpha subunit	162	270	100	0	1	1	0	1	1	1		
782	315426514	BAJ48145.1	dual specificity protein phosphatase	163	140	12	0	1	0	1	0	1	0		
783	315426518	BAJ48149.1	methyltransferase	338	132	88	0	1	1	0	1	1	0	M	
784	315426519	BAJ48150.1	ribonuclease Z	301	219	100	0	1	1	1	0	0	1	M	
785	315426520	BAJ48151.1	conserved hypothetical protein	126	101	23	1	1	1	1	0	0	0	M3	
786	315426521	BAJ48152.1	signal-transduction protein	138	81	53	0	1	1	1	0	0	1		
787	315426525	BAJ48156.1	serinethreonine protein kinase	234	118	28	0	1	1	1	0	0	1		
788	315426526	BAJ48157.1	glycolate oxidase	464	267	100	0	1	1	0	1	1	1		
789	315426527	BAJ48158.1	glucokinase	296	143	99	0	1	1	0	0	1	1	M3	
790	315426529	BAJ48160.1	conserved hypothetical protein	157	73	100	0	0	1	0	0	1	1		
791	315426550	BAJ48180.1	large subunit ribosomal protein L37e	55	55	8	0		0	0	0	1	0	R	
792	315426560	BAJ48190.1	DNA-directed RNA polymerase subunit B	1117	892	100	0	1	1	1	1	1	0	Polymerase	
793	315426561	BAJ48191.1	DNA-directed RNA polymerase subunit H	76	73	96	0	1	1	1	0	1	0	Polymerase	
794	315426565	BAJ48195.1	hypothetical protein HGMM_F33A05C34	373	243	17	1	1	1	0	0	0	0		

795	315426567	BAJ48197.1	signal-transduction protein	152	83	31	0	1	1	0	0	1	1		
796	315426570	BAJ48200.1	archaea-specific DNA-binding protein AlbA	94	75	76	1	1	1	0	1	0	0	M	
797	315426571	BAJ48201.1	transcription elongation factor NusA	178	150	16	1	1	1	0	0	0	0		
798	315426572	BAJ48202.1	aspartyl-tRNA synthetase	440	340	100	1	1	1	1	1	0	0	M	
799	315426593	BAJ48222.1	phosphate uptake regulator	330	149	70	1	1	1	1	0	0	0	M3	
800	315426639	BAJ48265.1	conserved hypothetical protein	78	179	32	0	1	1	0	0	0	1		
801	315426640	BAJ48266.1	branched-chain amino acid aminotransferase	310	251	100	0	1	1	1	0	0	1		
802	315426641	BAJ48267.1	translation factor	194	121	100	0	1	1	1	0	0	1		
803	315426644	BAJ48270.1	heat repeat-containing PBS lyase	172	160	7	0	1	0	1	0	0	1		
804	315426645	BAJ48271.1	carbon-monoxide dehydrogenase small subunit	156	106	100	0	1	1	0	0	0	1		
805	315426646	BAJ48272.1	molybdopterin dehydrogenase FAD-binding	290	145	100	0	1	1	0	0	0	1		
806	315426647	BAJ48273.1	aldehyde oxidase and xanthine dehydrogenase molybdopterin binding	756	350	100	0	1	1	0	0	0	1		
807	315426648	BAJ48274.1	aspartate kinase	406	223	100	0	1	1	1	1	0	1		
808	315426649	BAJ48275.1	deoxyhypusine synthase	317	228	100	0	1	1	1	1	1	0	M	
809	315426655	BAJ48281.1	nitroreductase	213	107	28	0	1	1	0	0	0	1		
810	315426658	BAJ48284.1	proliferating cell nuclear antigen PCNA	256	124	100	0	1	1	1	0	1	0	M	
811	315426659	BAJ48285.1	conserved hypothetical protein	138	125	4	1	0	0	1	0	0	0		
812	315426660	BAJ48286.1	pyruvate phosphate dikinase	887	724	100	0	1	1	1	1	1	1		
813	315426661	BAJ48287.1	FMN-dependent monooxygenase	331	113	100	0	1	0	1	0	0	1		
814	315426662	BAJ48288.1	thioesterase superfamily protein	333	73	100	0	1	1	0	0	1	1		
815	315426663	BAJ48289.1	metallophosphoesterase	235	193	13	0	0	1	0	1	0	1		
816	315426664	BAJ48290.1	ribosomal-protein-alanine N-acetyltransferase	170	107	100	0	0	1	0	1	1	1	M2	
817	315426665	BAJ48291.1	nucleotidyl transferase	236	156	100	0	1	1	1	0	1	1		
818	315426666	BAJ48292.1	cysteinyl-tRNA synthetase	475	356	100	0	1	1	0	1	1	1		
819	315426702	BAJ48327.1	beta-ribofuranosylaminobenzene 5-phosphate synthase	311	155	92	0	1	1	0	0	0	1		
820	315426703	BAJ48328.1	conserved hypothetical protein	368	262	29	1	1	1	0	0	0	0		
821	315426704	BAJ48329.1	conserved hypothetical protein	397	359	16	0	1	1	0	0	0	1		
822	315426705	BAJ48330.1	SirA family protein	103	79	6	1	0	1	0	0	0	0		
823	315426706	BAJ48331.1	V-type H+-transporting ATPase subunit C	351	334	12	0	1	0	1	0	0	0		
824	315426709	BAJ48334.1	aminotransferase	395	245	100	0	1	0	1	0	1	1		
825	315426710	BAJ48335.1	conserved hypothetical protein	214	118	71	0	1	1	0	1	0	1		
826	315426740	BAJ48364.1	conserved hypothetical protein	467	212	28	0	1	0	1	1	0	1		
827	315426741	BAJ48365.1	conserved hypothetical protein	231	108	100	0	1	1	0	1	1	1		
828	315426742	BAJ48366.1	arginyl-tRNA synthetase	624	291	100	0	1	1	1	0	0	1	M	
829	315426743	BAJ48367.1	conserved hypothetical protein	179	171	24	0	1	0	0	0	0	0		
830	315426744	BAJ48368.1	GTP-binding protein	385	286	100	0	1	1	1	0	1	1	M	A=>B
831	315426747	BAJ48371.1	5-nucleotidase SurE	267	152	100	0	1	1	0	1	0	1	M	
832	315426749	BAJ48373.1	transcriptional regulator AsnC family	143	81	100	0	1	1	1	1	1	1		
833	315426750	BAJ48374.1	N-acetyl-gamma-glutamyl-phosphate reductase	355	243	100	0	1	1	1	1	0	1	M3	A=>B
834	315426751	BAJ48375.1	acetylglutamateacetylaminoadipate kinase	264	181	100	0	1	1	1	1	1	1		A=>B
835	315426753	BAJ48377.1	lysine biosynthesis protein LysX	281	135	100	0	1	1	1	1	0	1		
836	315426754	BAJ48378.1	acetylornithine aminotransferase	393	256	100	0	1	1	1	1	1	1		A=>B
837	315426755	BAJ48379.1	conserved hypothetical protein	403	410	56	0	1	1	0	1	1	1		
838	315426756	BAJ48380.1	acetyl-lysine deacetylase	354	139	66	0	1	1	1	1	1	1		
839	315426757	BAJ48381.1	argininosuccinate lyase	477	302	100	0	1	1	1	0	1	1	M2	
840	315426759	BAJ48383.1	conserved hypothetical protein	136	132	4	1	0	0	1	0	0	0		
841	315426761	BAJ48385.1	conserved hypothetical protein	233	185	67	1	1	1	0	1	0	0	M2	
842	315426772	BAJ48395.1	acylphosphatase	288	73	41	0	1	1	0	0	0	1		
843	315426777	BAJ48400.1	type II secretion system protein E	525	267	100	1	1	1	0	0	0	0	M	
844	315426779	BAJ48402.1	UspA domain protein	146	138	12	0	0	1	0	0	0	1		
845	315426780	BAJ48403.1	conserved hypothetical protein	150	124	21	1	0	1	0	0	0	0		
846	315426781	BAJ48404.1	zinc ABC transporter permease	267	177	100	0	1	0	1	0	1	1		
847	315426782	BAJ48405.1	zinc ABC transporter ATP-binding protein	251	133	100	0	1	1	1	0	0	1		
848	315426783	BAJ48406.1	hypothetical protein HGMM_F08E07C20	306	111	100	0	1	1	1	0	0	1		
849	315426784	BAJ48407.1	metal dependent repressor DtxR family	142	121	13	0	1	0	0	0	0	1		
850	315426785	BAJ48408.1	ADP-ribose pyrophosphatase	147	122	55	0	0	1	0	0	0	1		
851	315426787	BAJ48410.1	dolichol-phosphate mannosyltransferase	240	169	100	0	1	1	0	0	1	1		
852	315426790	BAJ48413.1	conserved hypothetical protein	137	76	53	0	1	0	0	0	0	1		
853	315426791	BAJ48414.1	thiolase	393	238	100	0	1	1	0	0	0	1	M2	
854	315426793	BAJ48416.1	acyl-CoA dehydrogenase	389	283	100	0	1	1	0	0	1	1		
855	315426794	BAJ48417.1	tryptophan synthase beta subunit	466	413	100	0	1	1	0	1	1	1	M3	A=>B
856	315426795	BAJ48418.1	ERCC4 domain-containing protein	229	203	16	1	1	1	0	0	0	0		
857	315426796	BAJ48419.1	branched-chain amino acid ABC transporter substrate-binding protein	442	326	13	0	1	0	0	0	0	1		
858	315426797	BAJ48420.1	branched-chain amino acid ABC transporter permease	288	168	68	0	1	1	0	1	0	0		
859	315426798	BAJ48421.1	branched-chain amino acid ABC transporter permease	311	165	92	0	1	0	0	1	0	1		
860	315426799	BAJ48422.1	branched-chain amino acid ABC transporter ATP-binding protein	238	167	100	0	1	1	0	1	0	1		
861	315426813	BAJ48435.1	branched-chain amino acid ABC transporter ATP-binding protein	238	204	100	0	1	1	0	1	1	1		
862	315426818	BAJ48440.1	dehydrogenase flavoprotein-like protein	460	168	91	0	1	1	1	1	0	1		
863	315426820	BAJ48442.1	conserved hypothetical protein	135	124	18	1	1	1	0	0	0	0		
864	315426821	BAJ48443.1	conserved hypothetical protein	211	125	28	0	1	1	0	0	0	1		
865	315426822	BAJ48444.1	acetyl-CoA C-acetyltransferase	384	259	100	0	1	1	1	0	0	1		
866	315426823	BAJ48445.1	acetyl-CoA C-acetyltransferase	366	306	36	0	1	1	0	0	0	1		
867	315426824	BAJ48446.1	acyl-CoA dehydrogenase	380	285	100	0	1	1	0	0	1	1		
868	315426826	BAJ48448.1	acetoacetyl-CoA synthetase	666	446	100	0	1	1	0	0	1	1		
869	315426827	BAJ48449.1	hypothetical protein HGMM_F29A12C25	203	156	4	1	0	0	0	0	0	0		
870	315426841	BAJ48462.1	conserved hypothetical protein	142	131	8	1	0	1	0	0	0	0		
871	315426842	BAJ48463.1	conserved hypothetical protein	109	117	10	1	0	1	0	0	0	0		
872	315426843	BAJ48464.1	sulfonatenitratetaurine ABC transporter substrate-binding protein	353	98	68	0	1	0	0	0	1	1		
873	315426844	BAJ48465.1	sulfonatenitratetaurine ABC transporter ATP-binding protein	261	166	100	0	1	0	0	0	0	1		
874	315426851	BAJ48472.1	conserved hypothetical protein	122	106	23	1	0	1	0	0	0	0	M3	
875	315426888	BAJ48508.1	archaeal histone	79	67	11	1	1	0	1	0	0	0		
876	315426893	BAJ48513.1	chorismate synthase	360	283	100	0	1	1	1	1	1	1	M	
877	315426894	BAJ48514.1	4a-hydroxytetrahydrobiopterin dehydratase	92	61	100	0	0	1	1	0	1	1		
878	315426895	BAJ48515.1	phosphate uptake regulator	336	138	96	0	1	1	1	1	0	0	M2	
879	315426896	BAJ48516.1	phosphate ABC transporter substrate-binding protein	347	189	100	0	1	0	0	0	0	1		B=>AA
880	315426897	BAJ48517.1	phosphate ABC transporter permease	308	228	100	0	1	0	0	0	0	1		
881	315426898	BAJ48518.1	phosphate ABC transporter permease	224	211	100	0	1	0	0	0	0	1		
882	315426899	BAJ48519.1	phosphate ABC transporter ATP-binding protein	263	226	100	0	1	0	0	0	0	1		
883	315426900	BAJ48520.1	phosphate uptake regulator	210	101	33	0	1	1	1	1	1	1		
884	315426901	BAJ48521.1	metal-dependent protease PAD1JA B1 superfamily	142	123	14	1	1	0	1	0	0	0		
885	315426903	BAJ48523.1	glycine dehydrogenase subunit 1	456	310	100	0	1	1	0	1	0	1	M2	
886	315426904	BAJ48524.1	glycine dehydrogenase subunit 2	508	393	100	0	1	1	0	1	0	1	M2	
887	315426905	BAJ48525.1	glycine cleavage system aminomethyltransferase	360	232	100	0	1	1	0	1	0	0		
888	315426908	BAJ48528.1	electron transfer flavoprotein beta subunit	410	219	15	0	1	1	0	0	0	1		
889	315426909	BAJ48529.1	electron transfer flavoprotein alpha subunit	364	225	100	0	1	1	0	0	0	1		
890	315426910	BAJ48530.1	electron transfer flavoprotein-quinone oxidoreductase	562	228	90	0	1	1	0	0	1	1		
891	315426911	BAJ48531.1	thiosulfate sulfurtransferase	283	174	100	0	1	1	1	0	0	1		
892	315426913	BAJ48533.1	tRNA-intron endonuclease	178	106	64	1	1	1	1	1	0	0	M	
893	315426914	BAJ48534.1	conserved hypothetical protein	221	212	10	1	0	1	1	0	0	0		
894	315426916	BAJ48535.1	ubiquitin-activating enzyme E1-like protein	157	128	100	0	1	1	0	0	1	1		
895	315426917	BAJ48536.1	ubiquitin-conjugating enzyme E2-like protein	168	145	100	0	0	0	0	0	1	0		
896	315426919	BAJ48538.1	26S proteasome regulatory subunit N11-like protein	202	258	100	0	0	0	0	0	1	0		
897	315426921	BAJ48540.1	nucleic-acid-binding protein	140	109	79	0	1	1	0	1	0	1	M	
898	315426922	BAJ48541.1	acetyl-CoA C-acetyltransferase	384	312	100	0	1	1	1	1	0	1	M2	
899	315426923	BAJ48542.1	conserved hypothetical protein	350	278	100	0	1	1	0	1	0	1	M	
900	315426924	BAJ48543.1	transcriptional regulator lrp family	149	94	59	0	1	1	0	1	0	1		
901	315426925	BAJ48544.1	4-nitrophenyl phosphatase	264	135	100	0	1	1	0	0	1	1		
902	315426926	BAJ48545.1	signal-transduction protein	132	98	30	0	1	1	0	1	0	1		
903	315426927	BAJ48546.1	aminopeptidase N	822	217	100	0	1	1	0	0	1	1		
904	315426930	BAJ48549.1	glucose sorbosone dehydrogenase	373	198	100	0	1	1	1	0	0	0		
905	315426933	BAJ48552.1	HAD-superfamily hydrolase	206	199	6	0	0	0	0	0	0	1		
906	315426934	BAJ48553.1	branched-chain amino acid ABC transporter ATP-binding protein	233	168	100	0	1	1	0	1	0	1		
907	315426935	BAJ48554.1	branched-chain amino acid ABC transporter ATP-binding protein	253	181	100	0	1	1	1	1	0	1		
908	315426936	BAJ48555.1	branched-chain amino acid ABC transporter substrate-binding protein	443	221	47	0	1	1	1	1	0	1		
909	315426937	BAJ48556.1	branched-chain amino acid ABC transporter permease	312	263	30	0	1	1	1	1	0	1		
910	315426938	BAJ48557.1	branched-chain amino acid ABC transporter permease	375	196	61	0	1	1	1	1	0	1		
911	315426939	BAJ48558.1	ATP--cobalamin adenosyltransferase	170	117	100	0	1	1	1	0	1	1		
912	315426941	BAJ48560.1	archaeal flagellin FlaB	169	158	19	1	1	1	0	0	0	0		
913	315426945	BAJ48564.1	archaeal flagellar protein FlaH	233	207	28	1	1	1	0	0	0	0	M	
914	315426946	BAJ48565.1	archaeal flagellar protein FlaI	553	358	100	1	1	1	0	0	0	0	M3	
915	315426949	BAJ48568.1	asparaginyl-tRNA synthetase	347	253	100	0	1	1	0	1	0	1	M	
916	315426950	BAJ48569.1	ferredoxin-dependent glutamate synthase	453	386	21	0	1	1	0	1	0	1		
917	315426952	BAJ48571.1	glucosamine--fructose-6-phosphate aminotransferase isomerizing	605	423	100	0	1	1	0	0	1	1	M	A=>B
918	315426954	BAJ48573.1	molybdopterin-guanine dinucleotide biosynthesis protein A	198	162	11	0	1	1	0	0	0	1		
919	315427005	BAJ48623.1	NADH dehydrogenase I subunit H	343	288	100	0	1	0	1	0	1	1		
920	315427014	BAJ48631.1	hypothetical protein HGMM_F12C01C05	302	157	8	1	0	0	1	0	0	0		
921	315427015	BAJ48632.1	cytochrome c oxidase subunit I	524	466	100	0	1	0	0	0	0	1		
922	315427016	BAJ48633.1	cytochrome c oxidase subunit II	138	105	100	0	1	1	1	0	0	1		
923	315427019	BAJ48636.1	luciferase family protein	295	98	100	0	1	1	0	0	0	1		
924	315427020	BAJ48637.1	phosphomethylpyrimidine kinase	453	178	100	0	1	1	1	1	1	1		
925	315427021	BAJ48638.1	conserved hypothetical protein	421	204	75	0	1	1	0	1	0	1	M	
926	315427022	BAJ48639.1	conserved hypothetical protein	184	167	5	1	0	1	0	1	0	0		
927	315427023	BAJ48640.1	conserved hypothetical protein	193	152	10	0	1	1	0	1	0	1		

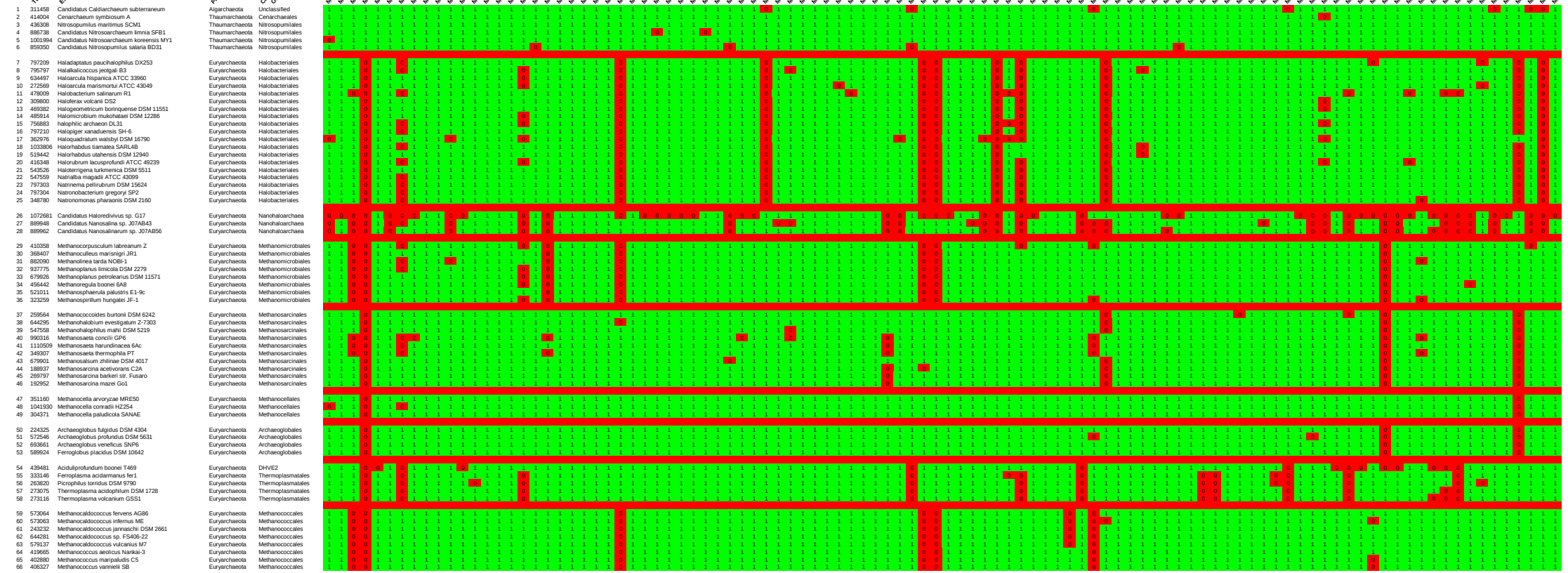
928	315427025	BAJ48642.1	DNA polymerase IV	363	177	100	0	1	1	1	0	1	1		
929	315427026	BAJ48643.1	acetoin utilization protein AcuC	317	148	100	0	1	1	0	1	1	1		
930	315427027	BAJ48644.1	conserved hypothetical protein	210	121	53	1	1	1	1	0	0	0	M3	
931	315427028	BAJ48645.1	bifunctional protein GlmU	349	221	100	0	1	1	0	0	0	1		
932	315427031	BAJ48648.1	riboflavin kinase archaea type	199	147	69	1	1	1	1	1	0	0	M2	
933	315427034	BAJ48651.1	nicotinamide-nucleotide adenylyltransferase	171	136	100	1	1	1	1	1	0	0	M	
934	315427035	BAJ48652.1	conserved hypothetical protein	263	139	100	0	1	1	0	1	0	1		
935	315427036	BAJ48653.1	peptidyl-tRNA hydrolase PTH2 family	118	95	100	0	1	1	1	1	1	0		
936	315427037	BAJ48654.1	rhomboid family protein	227	129	85	0	1	1	1	1	0	1		
937	315427038	BAJ48655.1	NADH dehydrogenase I subunit B	168	136	100	0	0	1	1	1	1	1		
938	315427039	BAJ48656.1	NADH dehydrogenase I subunit C	161	65	100	0	1	1	1	1	0	1		
939	315427040	BAJ48657.1	NADH dehydrogenase I subunit D	405	327	100	0	1	1	1	1	1	1		CA=>EA
940	315427041	BAJ48658.1	DNA polymerase I	827	360	100	0	1	1	1	1	1	1	M3	
941	315427045	BAJ48662.1	tRNA and rRNA cytosine-C5-methylases	320	178	100	0	1	1	0	1	1	1		
942	315427046	BAJ48663.1	N-glycosylaseDNA lyase	207	135	32	0	1	1	0	1	0	1		
943	315427054	BAJ48671.1	coenzyme F420-dependent hydrogenase beta subunit	399	160	100	0	1	0	0	0	0	1		
944	315427056	BAJ48673.1	ferredoxin	141	92	37	1	1	1	1	0	0	0		
945	315427057	BAJ48674.1	conserved hypothetical protein	226	124	41	0	1	1	0	1	0	0	M3	
946	315427058	BAJ48675.1	conserved hypothetical protein	145	74	74	0	1	1	0	1	0	1		
947	315427062	BAJ48678.1	aldehydeferredoxin oxidoreductase	590	405	100	0	1	1	0	1	0	1		
948	315427063	BAJ48679.1	aldoketo reductase	323	166	100	0	1	1	0	1	1	1		
949	315427069	BAJ48685.1	conserved hypothetical protein	249	131	74	1	1	1	1	1	0	0	M	
950	315427070	BAJ48686.1	acetyl-CoA synthetase	667	531	100	0	1	1	1	0	0	1		
951	315427071	BAJ48687.1	5-formyltetrahydrofolate cyclo-ligase	241	156	78	0	1	1	0	1	1	1	M	
952	315427072	BAJ48688.1	mannose-6-phosphate isomerase glucose-6-phosphate isomerase	338	274	11	0	0	1	0	0	0	1		
953	315427073	BAJ48689.1	competencedamage-inducible protein CinA	264	100	100	0	1	1	0	0	0	1	M	A=>B
954	315427074	BAJ48690.1	conserved hypothetical protein	217	119	70	0	1	1	0	1	0	1	M	
955	315427075	BAJ48691.1	hypothetical protein HGMM_F40F12C16	192	64	100	0	1	1	1	1	0	1		
956	315427076	BAJ48692.1	D-isomer specific 2-hydroxyacid dehydrogenase NAD-binding	314	119	100	0	1	1	0	0	1	0		
957	315427077	BAJ48693.1	transcriptional regulator MarR family	116	64	100	0	1	1	1	0	1	1		
958	315427079	BAJ48695.1	DtxR family iron dependent repressor	168	97	27	0	1	0	1	0	0	1		B<=>C/T
959	315427082	BAJ48698.1	large subunit ribosomal protein L44e	100	93	12	1	0	1	0	0	0	0		
960	315427083	BAJ48699.1	small subunit ribosomal protein S27e	66	64	9	1	1	1	0	0	0	0		
961	315427087	BAJ48703.1	peptidase M50	210	122	31	0	1	1	0	0	0	1		
962	315427089	BAJ48705.1	DNA ligase 1	529	357	100	0	1	1	1	1	0	1		
963	315427091	BAJ48707.1	methionyl-tRNA synthetase	121	90	100	0	1	1	0	1	0	1		
964	315427092	BAJ48708.1	asparaginyl-tRNA synthetase	429	327	100	0	1	1	0	1	1	1		
965	315427094	BAJ48710.1	basic amino acidpolyamine antiporter APA family	624	234	49	0	1	1	1	1	1	1		
966	315427096	BAJ48712.1	molybdopterin biosynthesis protein MoeA	412	261	100	0	1	1	0	1	0	1		
967	315427097	BAJ48713.1	carbon monoxide dehydrogenase F protein	283	84	100	0	1	1	0	0	0	1		
968	315427098	BAJ48714.1	VWA containing CoxE family protein	382	209	46	0	1	1	0	0	0	1		
969	315427099	BAJ48715.1	ATPase	289	210	100	0	1	1	0	0	0	1		
970	315427100	BAJ48716.1	molybdopterin dehydrogenase FAD-binding	291	177	100	0	1	1	0	0	1	1		
971	315427101	BAJ48717.1	2Fe-2S-binding domain protein	179	122	100	0	1	1	0	0	0	1		
972	315427136	BAJ48751.1	conserved hypothetical protein	256	107	100	0	1	1	1	1	1	0		
973	315427137	BAJ48752.1	aspartyl-tRNAAsnglutamyl-tRNA Gln amidotransferase subunit A	443	279	100	0	1	1	0	0	1	1		
974	315427143	BAJ48758.1	large subunit ribosomal protein L37Ae	76	73	11	1	1	1	0	0	0	0		
975	315427146	BAJ48761.1	conserved hypothetical protein	335	214	31	1	1	1	0	0	0	0	M3	
976	315427147	BAJ48762.1	cobaltnickel ABC transporter ATP-binding protein	282	156	100	0	1	1	0	1	0	1		
977	315427149	BAJ48764.1	cobaltnickel ABC transporter permease	265	218	13	0	1	1	0	1	0	1		
978	315427151	BAJ48766.1	lipoyloctanoyl transferase	225	116	100	0	0	1	0	0	1	1		
979	315427152	BAJ48767.1	dihydrolipoamide dehydrogenase	156	347	100	0	1	1	0	0	1	1		
980	315427171	BAJ48785.1	2-oxoisovalerate dehydrogenase E1 component beta subunit	324	288	100	0	1	1	0	0	1	1		
981	315427172	BAJ48786.1	lipoic acid synthetase	291	246	100	0	1	1	0	0	1	1		
982	315427173	BAJ48787.1	NUDIX hydrolase	147	109	39	0	1	1	1	0	1	0		
983	315427176	BAJ48790.1	F420-0gamma-glutamyl ligase	254	178	100	0	1	0	1	1	0	1		
984	315427178	BAJ48792.1	histidine triad HIT protein hit	163	113	100	0	1	1	0	1	0	1		
985	315427180	BAJ48794.1	conserved hypothetical protein	289	233	9	1	1	1	1	1	0	0		
986	315427182	BAJ48796.1	S-adenosylmethionine decarboxylase	150	83	100	0	1	1	0	0	0	1	M3	
987	315427183	BAJ48797.1	molybdenum cofactor biosynthesis protein E	146	126	10	0	1	0	0	0	1	1		
988	315427185	BAJ48799.1	dCTP deaminase	172	132	100	0	1	1	0	1	0	1	M	EA=>B
989	315427186	BAJ48800.1	conserved hypothetical protein	262	87	95	0	1	1	0	0	1	1		
990	315427187	BAJ48801.1	molybdenum cofactor biosynthesis protein B	174	106	100	0	1	1	0	1	0	1		
991	315427188	BAJ48802.1	molybdenum cofactor biosynthesis protein C	142	124	100	0	1	1	0	1	1	1	M	
992	315427189	BAJ48803.1	small subunit ribosomal protein S11	147	119	100	0	1	1	0	1	1	0	R	
993	315427190	BAJ48804.1	small subunit ribosomal protein S4	170	140	100	0	1	1	1	1	1	0	R	
994	315427191	BAJ48805.1	geranylgeranyl reductase	379	202	100	0	1	1	1	1	0	1	M3	
995	315427193	BAJ48807.1	isochorismatase	196	91	100	0	1	1	0	0	1	1		
996	315427194	BAJ48808.1	nicotinate phosphoribosyltransferase	403	253	100	0	1	1	0	1	0	1	M2	
997	315427195	BAJ48809.1	fumarate hydratase class II	465	395	100	0	1	0	0	0	1	1		
998	315427197	BAJ48811.1	conserved hypothetical protein	276	137	56	1	1	1	0	0	0	0	M2	
999	315427198	BAJ48812.1	conserved hypothetical protein	123	123	5	0	0	0	1	0	0	0		
1000	315427200	BAJ48814.1	SAM-dependent methyltransferase	283	168	46	0	1	1	0	0	1	1	M3	
1001	315427201	BAJ48815.1	lysyl-tRNA synthetase class I	542	290	100	0	1	0	1	1	0	1		
1002	315427202	BAJ48816.1	replication factor A1	142	129	11	1	1	1	0	0	0	0		
1003	315427203	BAJ48817.1	membrane-bound serine protease ClpP class	422	263	93	0	1	1	0	1	0	1	M2	
1004	315427204	BAJ48818.1	membrane protease subunit	270	221	100	0	1	1	0	1	1	1	M3	
1005	315427207	BAJ48820.1	DNA repair protein RadA	320	291	100	0	1	1	1	1	1	0	M	
1006	315427208	BAJ48821.1	phosphoglycerate mutase	409	262	100	0	1	1	0	1	1	1	M3	
1007	315427211	BAJ48824.1	SAM-dependent methyltransferase	260	137	100	0	1	1	0	1	1	0		
1008	315427213	BAJ48826.1	long-chain-fatty-acid--CoA ligase	568	349	100	0	1	1	0	0	0	1		
1009	315427214	BAJ48827.1	citrate synthase	380	283	100	0	1	1	1	0	0	1		
1010	315427215	BAJ48828.1	sugar fermentation stimulation protein A	236	104	100	0	1	1	0	0	0	1		
1011	315427216	BAJ48829.1	protoheme IX farnesyltransferase	281	219	100	0	1	1	1	1	0	1		
1012	315427219	BAJ48832.1	conserved hypothetical protein	162	129	10	0	0	0	0	0	0	1		
1013	315427220	BAJ48833.1	valyl tRNA-synthetase	775	543	100	0	1	1	1	1	0	1	M2	
1014	315427221	BAJ48834.1	conserved hypothetical protein	117	69	100	0	1	1	0	0	0	1		
1015	315427222	BAJ48835.1	protease I	170	129	100	0	1	1	0	1	0	1		
1016	315427223	BAJ48836.1	modification methyltransferase	315	303	7	0	0	0	1	0	0	1		
1017	315427224	BAJ48837.1	glycosyl transferase family 2	384	123	87	0	1	1	1	1	0	1		
1018	315427225	BAJ48838.1	small heat shock protein HSP20	191	92	67	0	1	1	1	1	0	1	M2	
1019	315427226	BAJ48839.1	2-oxoglutarate ferredoxin oxidoreductase subunit alpha	651	338	100	0	1	1	1	1	0	1	M3	EA=>B
1020	315427227	BAJ48840.1	2-oxoglutarate ferredoxin oxidoreductase subunit beta	322	206	100	0	1	1	1	1	0	1		EA=>B
1021	315427232	BAJ48845.1	diphosphomevalonate decarboxylase	320	242	30	0	1	0	0	0	0	1		
1022	315427233	BAJ48846.1	serine hydroxymethyltransferase	391	295	100	0	1	1	1	1	0	1	M	
1023	315427234	BAJ48847.1	transcriptional regulator AsnC family	148	95	100	0	1	1	1	1	0	1		
1024	315427276	BAJ48888.1	CBS-domain-containing protein	196	109	56	1	1	1	0	0	0	0	M	
1025	315427277	BAJ48889.1	conserved hypothetical protein	145	123	31	0	1	1	0	0	0	1	M2	
1026	315427278	BAJ48890.1	conserved hypothetical protein	216	129	100	0	1	1	1	1	0	1	M3	
1027	315427279	BAJ48891.1	glycosyltransferase family 28	382	268	24	1	1	0	0	1	0	0		
1028	315427280	BAJ48892.1	conserved hypothetical protein	172	154	4	1	0	1	0	1	0	0		
1029	315427281	BAJ48893.1	sulfonatenitratetaurine ABC transporter ATP binding protein	253	177	100	0	1	1	0	1	1	1		
1030	315427282	BAJ48894.1	sulfonatenitratetaurine ABC transporter permease	517	345	73	0	1	1	0	1	1	1		
1031	315427283	BAJ48895.1	octaprenyl-diphosphate synthase	312	136	100	0	1	1	1	0	0	1		
1032	315427284	BAJ48896.1	thioredoxin 1	135	86	100	0	1	1	0	0	1	1		
1033	315427286	BAJ48898.1	histidinol-phosphate aminotransferase	360	162	100	0	1	0	0	0	1	1		
1034	315427289	BAJ48901.1	V-type H+-transporting ATPase subunit I	679	193	85	0	1	1	1	1	0	1	M3	A=>B
1035	315427294	BAJ48906.1	calcineurin superfamily phosphohydrolase	252	162	39	0	1	0	1	0	0	1		
1036	315427295	BAJ48907.1	coenzyme A-binding protein	147	102	100	0	1	1	1	0	1	1		
1037	315427296	BAJ48908.1	pyruvate formate-lyase activating enzyme	360	276	100	0	1	1	1	1	0	1	M	CA=>EA
1038	315427304	BAJ48915.1	conserved hypothetical protein	184	174	10	0	1	1	0	0	0	0		
1039	315427307	BAJ48918.1	GMP synthase glutamine-hydrolysing	510	430	100	0	0	1	1	0	1	1		
1040	315427308	BAJ48919.1	enoyl-CoA hydratase	268	185	100	0	1	1	0	1	1	1		
1041	315427310	BAJ48921.1	S-adenosylmethionine decarboxylase	132	82	100	0	1	1	0	0	0	1		
1042	315427311	BAJ48922.1	spermidine synthase	303	163	100	0	1	1	0	0	1	1		
1043	315427312	BAJ48923.1	phosphonate ABC transporter permese	214	156	100	0	1	1	1	0	1	1		
1044	315427313	BAJ48924.1	phosphonate ABC transporter ATP-binding protein	238	142	100	0	1	1	1	0	1	1		
1045	315427314	BAJ48925.1	phosphonate ABC transporter substrate-binding protein	318	125	56	0	0	1	1	0	0	1		B=>AA+T
1046	315427316	BAJ48927.1	manganeseiron ABC transporter permease	277	241	100	0	0	0	0	0	1	1		B=>AA
1047	315427317	BAJ48928.1	ABC transporter ATP-binding protein	251	186	100	0	0	0	1	0	1	1		B=>AA
1048	315427318	BAJ48929.1	manganeseiron ABC transporter substrate-binding protein	307	208	100	0	0	0	0	0	0	1		B=>AA
1049	315427321	BAJ48932.1	maleate isomerase	232	132	76	0	1	1	0	0	1	1		
1050	315427322	BAJ48933.1	N-methylhydantoinase A	683	406	100	0	1	1	0	0	0	1		
1051	315427323	BAJ48934.1	N-methylhydantoinase B	522	366	100	0	1	1	0	0	1	1		
1052	315427324	BAJ48935.1	small subunit ribosomal protein S7	202	180	100	0	1	1	1	1	1	0	R	
1053	315427325	BAJ48936.1	phosphoenolpyruvate carboxylase	522	274	40	0	1	1	0	0	0	1		
1054	315427326	BAJ48937.1	simple sugar ABC transporter permease	324	234	100	0	1	1	0	1	1	1		
1055	315427327	BAJ48938.1	simple sugar transporter permease	354	227	100	0	1	1	0	1	0	1		
1056	315427329	BAJ48940.1	simple sugar ABC transporter substrate-binding protein	449	191	100	0	1	1	0	1	1	1		
1057	315427330	BAJ48941.1	adenine deaminase	596	355	100	0	1	1	1	1	0	1		
1058	315427332	BAJ48943.1	conserved hypothetical protein	410	250	100	0	1	1	0	0	1	1		
1059	315427594	BAJ49193.1	hypothetical protein HGMM_F51A06C36	448	448	6	0	0	0	0	0	0	1		
1060	315427597	BAJ49196.1	conserved hypothetical protein	236	225	7	0	0	0	0	0	0	1		

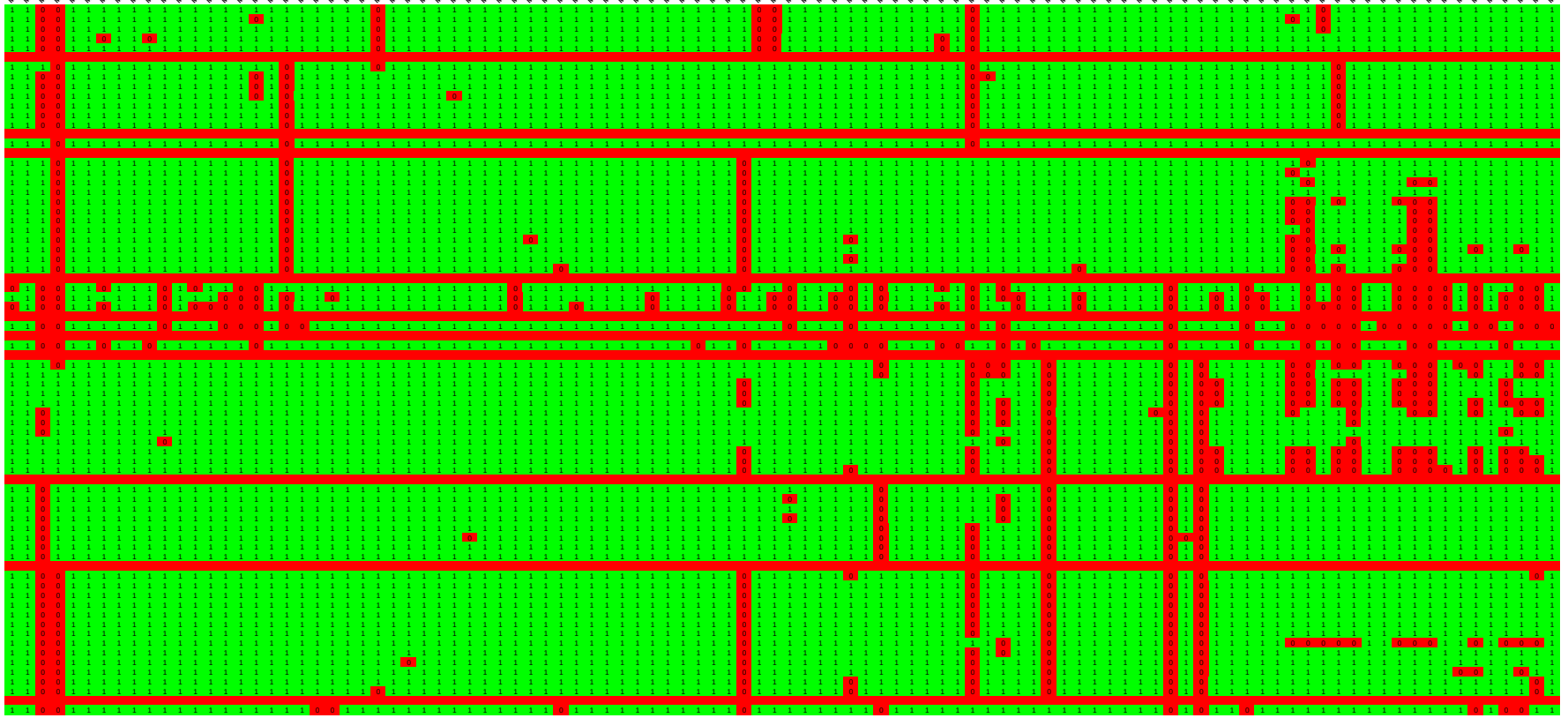
1061	315427598	BAJ49197.1	UDP-glucose 4-epimerase	315	173	100	0	1	1	0	1	0	1		
1062	315427603	BAJ49202.1	NAD-dependent epimerasedehydratase	347	150	100	0	1	1	1	1	0	1		
1063	315427739	BAJ49334.1	conserved hypothetical protein	87	194	12	1	0	1	0	1	0	0		
1064	315427751	BAJ49346.1	hypothetical protein HGMM_F01D06C13	695	557	6	0	1	0	0	0	0	1		
1065	315427767	BAJ49362.1	arsenic transporting ATPase	327	131	100	0	1	1	0	0	1	1		
1066	315427770	BAJ49365.1	carbon starvation protein CstA	596	376	100	0	1	1	0	0	0	1		
1067	315427781	BAJ49375.1	enoyl-CoA hydratase	254	192	100	0	1	1	0	1	1	1		
1068	315427782	BAJ49376.1	alcohol dehydrogenase	339	178	100	0	1	1	0	0	1	1		
1069	315427783	BAJ49377.1	naphthoate synthase	312	220	100	0	1	1	0	0	1	1		
1070	315427784	BAJ49378.1	aldehydeferredoxin oxidoreductase	627	373	100	0	1	1	0	1	0	1		
1071	315427786	BAJ49380.1	phenylacetate-CoA ligase	458	375	100	0	1	1	0	0	0	1		
1072	315427790	BAJ49384.1	branched-chain amino acid ABC transporter substrate- binding protein	494	222	22	0	1	0	0	0	0	1		
1073	315427795	BAJ49389.1	phenylacetyl-CoAacceptor oxidoreductase subunit	211	132	100	0	1	1	0	0	1	1		
1074	315427796	BAJ49390.1	phenylacetyl-CoAacceptor oxidoreductase subunit	904	199	100	0	1	1	0	0	0	1		
1075	315427797	BAJ49391.1	benzoyl-CoA reductase subunit C	394	205	35	0	1	0	0	0	0	1		
1076	315427798	BAJ49392.1	CoA-substrate-specific enzyme activase	249	137	100	0	1	0	0	0	0	1		
1077	315427799	BAJ49393.1	CoA-substrate-specific enzyme activase	275	138	100	0	1	0	0	0	0	1		
1078	315427800	BAJ49394.1	benzoyl-CoA reductase subunit B	420	171	29	0	1	0	0	0	0	1		
1079	315427969	BAJ49559.1	dihydroorotate oxidase	332	207	100	0	1	0	0	0	1	1		
1080	315427983	BAJ49572.1	5-oxoprolinase ATP-hydrolysing	498	399	100	0	1	1	0	0	0	1		
1081	315427984	BAJ49573.1	N-methylhydantoinase B	566	304	100	0	1	1	0	0	0	1		
1082	315427986	BAJ49575.1	N-methylhydantoinase A	692	398	100	0	1	1	0	0	0	1		
1083	315427989	BAJ49578.1	cobaltnickel ABC transporter ATP-binding protein	586	236	100	0	1	1	0	1	0	1		

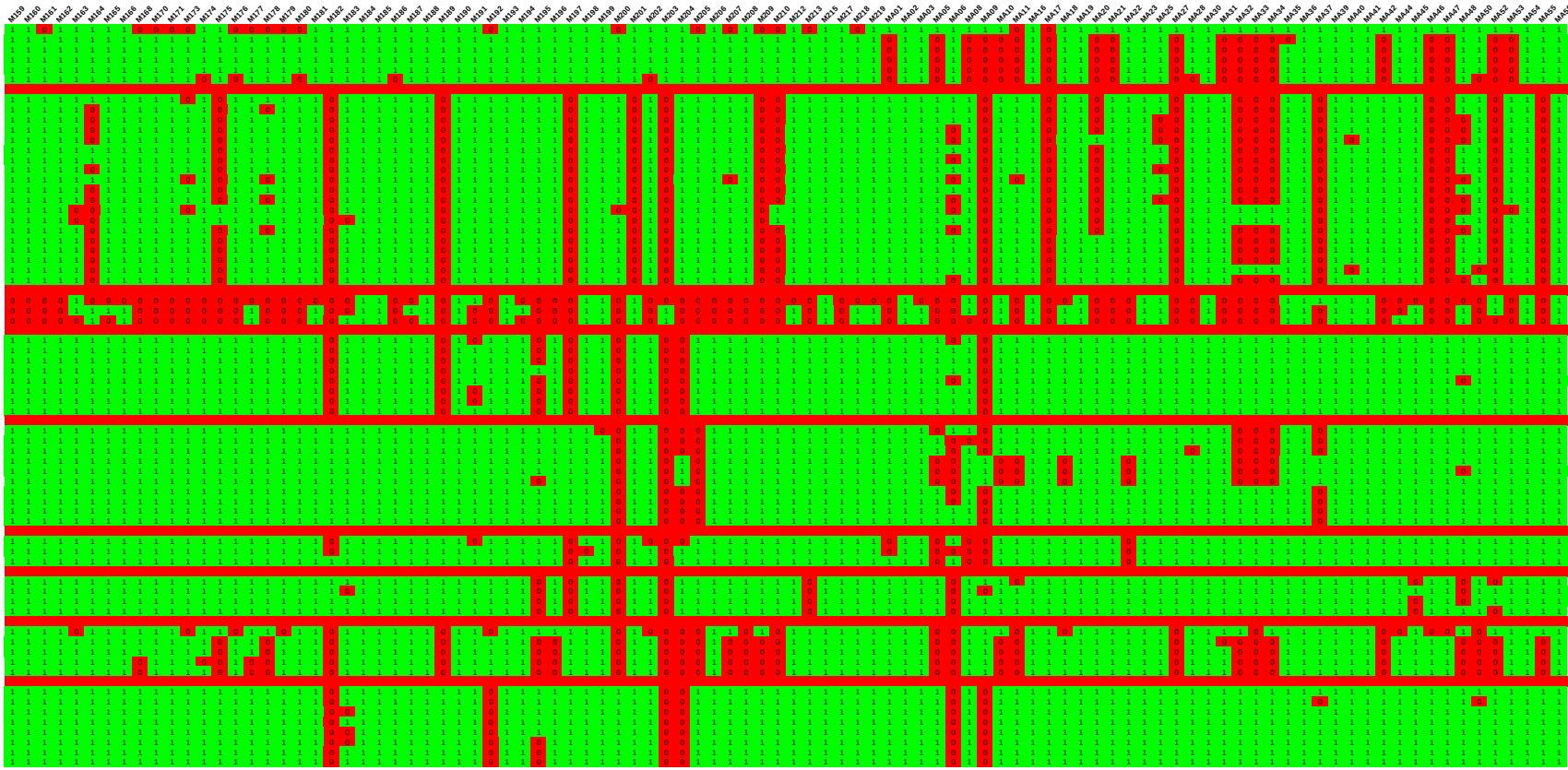
Annexe 3 : Répartition taxonomique des 200 nouveaux marqueurs chez les archées.

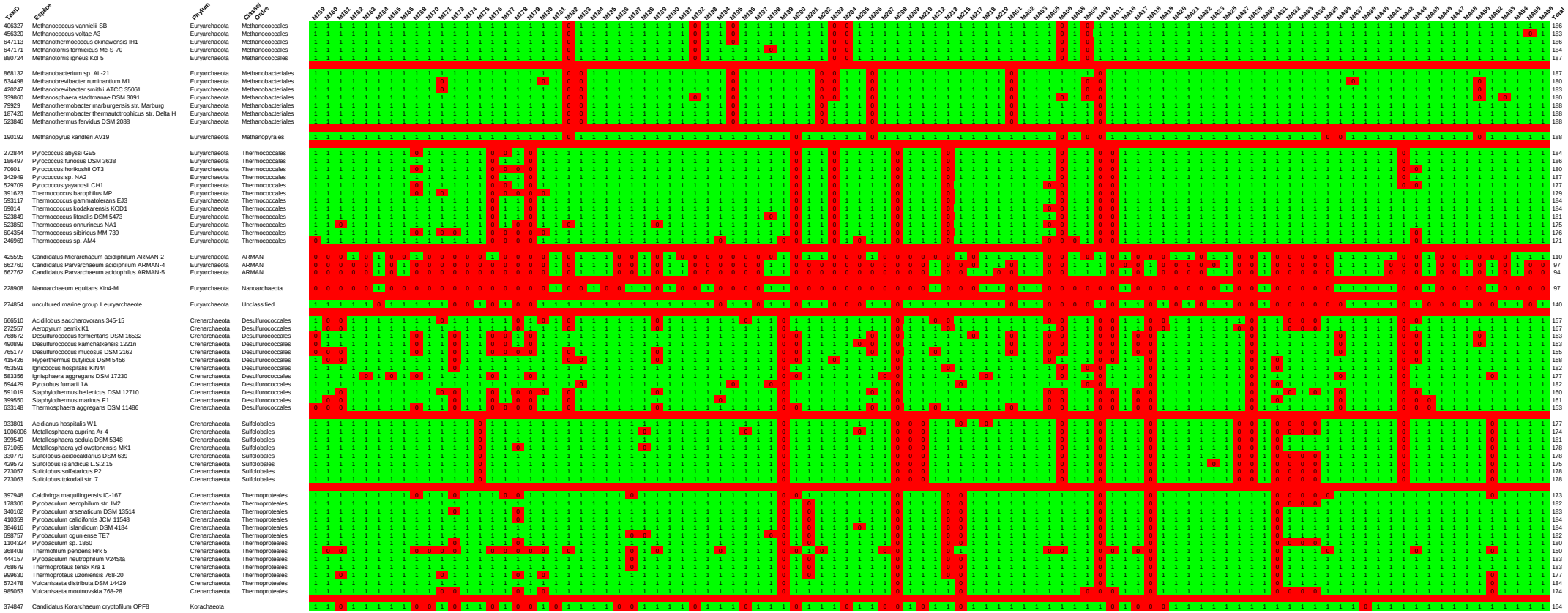
Chaque ligne correspond à une espèce, chaque colonne à un jeu de données, le croisement entre les deux correspond à la présence (« 1 », case verte), ou à l'absence (« 0 », case rouge) d'homologue de l'espèce dans le jeu de données. La dernière colonne donne le nombre de jeux de données dans lesquels l'espèce est présente.

La liste des archées correspond uniquement aux 129 génomes utilisés dans ce travail.




TaxID
Espèce
Phylum
Classe/Ordre
M002 M003 M004 M005 M010 M011 M012 M014 M018 M021 M022 M025 M026 M027 M028 M029 M030 M031 M032 M033 M034 M035 M036 M037 M038 M039 M040 M041 M042 M043 M044 M045 M046 M047 M048 M049 M050 M051 M052 M053 M054 M055 M056 M057 M058 M059 M060 M061 M064 M066 M067 M068 M069 M070 M071 M075 M076 M077 M078 M080 M085 M086 M092 M093 M112 M115 M116 M117 M118 M119 M121 M122 M123 M124 M125 M126 M127 M128 M129 M130 M132 M134 M135 M136 M137 M138 M140 M142 M143 M144 M145 M146 M147 M148 M149 M150 M151 M153 M154 M156 M157 M158
66 406327 Methanococcus vannielii SB Euryarchaeota Methanococcales
67 456320 Methanococcus voltae A3 Euryarchaeota Methanococcales
68 647113 Methanothermococcus okinawensis IH1 Euryarchaeota Methanococcales
69 647171 Methanotorris formicicus Mc-S-70 Euryarchaeota Methanococcales
70 880724 Methanotorris igneus Kol 5 Euryarchaeota Methanococcales
71 868132 Methanobacterium sp. AL-21 Euryarchaeota Methanobacteriales
72 634498 Methanobrevibacter ruminantium M1 Euryarchaeota Methanobacteriales
73 420247 Methanobrevibacter smithii ATCC 35061 Euryarchaeota Methanobacteriales
74 339860 Methanosphaera stadtmanae DSM 3091 Euryarchaeota Methanobacteriales
75 79929 Methanothermobacter marburgensis str. Marburg Euryarchaeota Methanobacteriales
76 187420 Methanothermobacter thermautotrophicus str. Delta H Euryarchaeota Methanobacteriales
77 523846 Methanothermus fervidus DSM 2088 Euryarchaeota Methanobacteriales
78 190192 Methanopyrus kandleri AV19 Euryarchaeota Methanopyrales
79 272844 Pyrococcus abyssi GE5 Euryarchaeota Thermococcales
80 186497 Pyrococcus furiosus DSM 3638 Euryarchaeota Thermococcales
81 70601 Pyrococcus horikoshii OT3 Euryarchaeota Thermococcales
82 342949 Pyrococcus sp. NA2 Euryarchaeota Thermococcales
83 529709 Pyrococcus yayanosii CH1 Euryarchaeota Thermococcales
84 391623 Thermococcus barophilus MP Euryarchaeota Thermococcales
85 593117 Thermococcus gammatolerans EJ3 Euryarchaeota Thermococcales
86 69014 Thermococcus kodakarensis KOD1 Euryarchaeota Thermococcales
87 523849 Thermococcus litoralis DSM 5473 Euryarchaeota Thermococcales
88 523850 Thermococcus onnurineus NA1 Euryarchaeota Thermococcales
89 604354 Thermococcus sibiricus MM 739 Euryarchaeota Thermococcales
90 246969 Thermococcus sp. AM4 Euryarchaeota Thermococcales
91 425595 Candidatus Micrarchaeum acidiphilum ARMAN-2 Euryarchaeota ARMAN
92 662760 Candidatus Parvarchaeum acidiphilum ARMAN-4 Euryarchaeota ARMAN
93 662762 Candidatus Parvarchaeum acidophilus ARMAN-5 Euryarchaeota ARMAN
94 228908 Nanoarchaeum equitans Kin4-M Euryarchaeota Nanoarchaeota
95 274854 uncultured marine group II euryarchaeote Euryarchaeota Unclassified
96 666510 Acidilobus saccharovorans 345-15 Crenarchaeota Desulfurococcales
97 272557 Aeropyrum pernix K1 Crenarchaeota Desulfurococcales
98 768672 Desulfurococcus fermentans DSM 16532 Crenarchaeota Desulfurococcales
99 490899 Desulfurococcus kamchatkensis 1221n Crenarchaeota Desulfurococcales
100 765177 Desulfurococcus mucosus DSM 2162 Crenarchaeota Desulfurococcales
101 415426 Hyperthermus butylicus DSM 5456 Crenarchaeota Desulfurococcales
102 453591 Ignicoccus hospitalis KIN4/I Crenarchaeota Desulfurococcales
103 583356 Ignisphaera aggregans DSM 17230 Crenarchaeota Desulfurococcales
104 694429 Pyrolobus fumarii 1A Crenarchaeota Desulfurococcales
105 591019 Staphylothermus hellenicus DSM 12710 Crenarchaeota Desulfurococcales
106 399550 Staphylothermus marinus F1 Crenarchaeota Desulfurococcales
107 633148 Thermosphaera aggregans DSM 11486 Crenarchaeota Desulfurococcales
108 933801 Acidianus hospitalis W1 Crenarchaeota Sulfolobales
109 1006006 Metallosphaera cuprina Ar-4 Crenarchaeota Sulfolobales
110 399549 Metallosphaera sedula DSM 5348 Crenarchaeota Sulfolobales
111 671065 Metallosphaera yellowstonensis MK1 Crenarchaeota Sulfolobales
112 330779 Sulfolobus acidocaldarius DSM 639 Crenarchaeota Sulfolobales
113 429572 Sulfolobus islandicus L.S.2.15 Crenarchaeota Sulfolobales
114 273057 Sulfolobus solfataricus P2 Crenarchaeota Sulfolobales
115 273063 Sulfolobus tokodaii str. 7 Crenarchaeota Sulfolobales
116 397948 Caldivirga maquilingensis IC-167 Crenarchaeota Thermoproteales
117 178306 Pyrobaculum aerophilum str. IM2 Crenarchaeota Thermoproteales
118 340102 Pyrobaculum arsenaticum DSM 13514 Crenarchaeota Thermoproteales
119 410359 Pyrobaculum calidifontis JCM 11548 Crenarchaeota Thermoproteales
120 384616 Pyrobaculum islandicum DSM 4184 Crenarchaeota Thermoproteales
121 698757 Pyrobaculum oguniense TE7 Crenarchaeota Thermoproteales
122 1104324 Pyrobaculum sp. 1860 Crenarchaeota Thermoproteales
123 368408 Thermofilum pendens Hrk 5 Crenarchaeota Thermoproteales
124 444157 Pyrobaculum neutrophilum V24Sta Crenarchaeota Thermoproteales
125 768679 Thermoproteus tenax Kra 1 Crenarchaeota Thermoproteales
126 999630 Thermoproteus uzoniensis 768-20 Crenarchaeota Thermoproteales
127 572478 Vulcanisaeta distributa DSM 14429 Crenarchaeota Thermoproteales
128 985053 Vulcanisaeta moutnovskia 768-28 Crenarchaeota Thermoproteales
129 374847 Candidatus Korarchaeum cryptofilum OPF8 Korachaeota





	TaxID	Espèce	Phylum	Classe/ Ordre	Total
66	406327	Methanococcus vannielii SB	Euryarchaeota	Methanococcales	186
67	456320	Methanococcus voltae A3	Euryarchaeota	Methanococcales	183
68	647113	Methanothermococcus okinawensis IH1	Euryarchaeota	Methanococcales	186
69	647171	Methanotorris formicicus Mc-S-70	Euryarchaeota	Methanococcales	184
70	880724	Methanotorris igneus Kol 5	Euryarchaeota	Methanococcales	187
71	868132	Methanobacterium sp. AL-21	Euryarchaeota	Methanobacteriales	187
72	634498	Methanobrevibacter ruminantium M1	Euryarchaeota	Methanobacteriales	180
73	420247	Methanobrevibacter smithii ATCC 35061	Euryarchaeota	Methanobacteriales	183
74	339860	Methanosphaera stadtmanae DSM 3091	Euryarchaeota	Methanobacteriales	180
75	79929	Methanothermobacter marburgensis str. Marburg	Euryarchaeota	Methanobacteriales	188
76	187420	Methanothermobacter thermautotrophicus str. Delta H	Euryarchaeota	Methanobacteriales	188
77	523846	Methanothermus fervidus DSM 2088	Euryarchaeota	Methanobacteriales	188
78	190192	Methanopyrus kandleri AV19	Euryarchaeota	Methanopyrales	188
79	272844	Pyrococcus abyssi GE5	Euryarchaeota	Thermococcales	184
80	186497	Pyrococcus furiosus DSM 3638	Euryarchaeota	Thermococcales	186
81	70601	Pyrococcus horikoshii OT3	Euryarchaeota	Thermococcales	180
82	342949	Pyrococcus sp. NA2	Euryarchaeota	Thermococcales	187
83	529709	Pyrococcus yayanosii CH1	Euryarchaeota	Thermococcales	177
84	391623	Thermococcus barophilus MP	Euryarchaeota	Thermococcales	179
85	593117	Thermococcus gammatolerans EJ3	Euryarchaeota	Thermococcales	184
86	69014	Thermococcus kodakarensis KOD1	Euryarchaeota	Thermococcales	184
87	523849	Thermococcus litoralis DSM 5473	Euryarchaeota	Thermococcales	181
88	523850	Thermococcus onnurineus NA1	Euryarchaeota	Thermococcales	175
89	604354	Thermococcus sibiricus MM 739	Euryarchaeota	Thermococcales	176
90	246969	Thermococcus sp. AM4	Euryarchaeota	Thermococcales	171
91	425595	Candidatus Micrarchaeum acidiphilum ARMAN-2	Euryarchaeota	ARMAN	110
92	662760	Candidatus Parvarchaeum acidiphilum ARMAN-4	Euryarchaeota	ARMAN	97
93	662762	Candidatus Parvarchaeum acidophilus ARMAN-5	Euryarchaeota	ARMAN	94
94	228908	Nanoarchaeum equitans Kin4-M	Euryarchaeota	Nanoarchaeota	97
95	274854	uncultured marine group II euryarchaeote	Euryarchaeota	Unclassified	140
96	666510	Acidilobus saccharovorans 345-15	Crenarchaeota	Desulfurococcales	157
97	272557	Aeropyrum pernix K1	Crenarchaeota	Desulfurococcales	167
98	768672	Desulfurococcus fermentans DSM 16532	Crenarchaeota	Desulfurococcales	163
99	490899	Desulfurococcus kamchatkensis 1221n	Crenarchaeota	Desulfurococcales	163
100	765177	Desulfurococcus mucosus DSM 2162	Crenarchaeota	Desulfurococcales	155
101	415426	Hyperthermus butylicus DSM 5456	Crenarchaeota	Desulfurococcales	168
102	453591	Ignicoccus hospitalis KIN4/I	Crenarchaeota	Desulfurococcales	182
103	583356	Ignisphaera aggregans DSM 17230	Crenarchaeota	Desulfurococcales	177
104	694429	Pyrolobus fumarii 1A	Crenarchaeota	Desulfurococcales	182
105	591019	Staphylothermus hellenicus DSM 12710	Crenarchaeota	Desulfurococcales	160
106	399550	Staphylothermus marinus F1	Crenarchaeota	Desulfurococcales	161
107	633148	Thermosphaera aggregans DSM 11486	Crenarchaeota	Desulfurococcales	153
108	933801	Acidianus hospitalis W1	Crenarchaeota	Sulfolobales	177
109	1006006	Metallosphaera cuprina Ar-4	Crenarchaeota	Sulfolobales	174
110	399549	Metallosphaera sedula DSM 5348	Crenarchaeota	Sulfolobales	181
111	671065	Metallosphaera yellowstonensis MK1	Crenarchaeota	Sulfolobales	178
112	330779	Sulfolobus acidocaldarius DSM 639	Crenarchaeota	Sulfolobales	178
113	429572	Sulfolobus islandicus L.S.2.15	Crenarchaeota	Sulfolobales	175
114	273057	Sulfolobus solfataricus P2	Crenarchaeota	Sulfolobales	178
115	273063	Sulfolobus tokodaii str. 7	Crenarchaeota	Sulfolobales	178
116	397948	Caldivirga maquilingensis IC-167	Crenarchaeota	Thermoproteales	173
117	178306	Pyrobaculum aerophilum str. IM2	Crenarchaeota	Thermoproteales	182
118	340102	Pyrobaculum arsenaticum DSM 13514	Crenarchaeota	Thermoproteales	183
119	410359	Pyrobaculum calidifontis JCM 11548	Crenarchaeota	Thermoproteales	184
120	384616	Pyrobaculum islandicum DSM 4184	Crenarchaeota	Thermoproteales	184
121	698757	Pyrobaculum oguniense TE7	Crenarchaeota	Thermoproteales	182
122	1104324	Pyrobaculum sp. 1860	Crenarchaeota	Thermoproteales	180
123	369408	Thermofilum pendens Hrk 5	Crenarchaeota	Thermoproteales	150
124	444157	Pyrobaculum neutrophilum V24Sta	Crenarchaeota	Thermoproteales	183
125	768679	Thermoproteus tenax Kra 1	Crenarchaeota	Thermoproteales	183
126	999630	Thermoproteus uzoniensis 768-20	Crenarchaeota	Thermoproteales	177
127	572478	Vulcanisaeta distributa DSM 14429	Crenarchaeota	Thermoproteales	184
128	985053	Vulcanisaeta moutnovskia 768-28	Crenarchaeota	Thermoproteales	174
129	374847	Candidatus Korarchaeum cryptofilum OPF8	Korachaeota		164

Annexe 4 : Liste des protéines informationnelles mises à jour pour le Chapitre 1.

Liste des 57 protéines ribosomiques, 14 sous-unités de l'ARN polymérase et 2 facteurs de transcription mis à jours pour l'étude exposée dans le Chapitre 1.

Annotation	Réference gi	Réference accession	séquences
50S ribosomal protein L1	161527890	YP_001581716	127
50S ribosomal protein L2	161527614	YP_001581440	128
50S ribosomal protein L3P	161528317	YP_001582143	127
50S ribosomal protein L4P	161528316	YP_001582142	127
50S ribosomal protein L5	161528305	YP_001582131	128
50S ribosomal protein L6	161528302	YP_001582128	128
50S ribosomal protein L7Ae	161527733	YP_001581559	129
acidic ribosomal protein P0 - L10	161527889	YP_001581715	128
50S ribosomal protein L10e	161527957	YP_001581783	129
50S ribosomal protein L11	161527893	YP_001581719	127
50S ribosomal protein L12 - L12e	161527882	YP_001581708	128
50S ribosomal protein L13	161527934	YP_001581760	129
50S ribosomal protein L14	161528308	YP_001582134	127
50S ribosomal protein L15	161527909	YP_001581735	129
50S ribosomal protein L15e	161528762	YP_001582588	129
50S ribosomal protein L18	161527906	YP_001581732	128
50S ribosomal protein L15 – L18e	161527935	YP_001581761	127
50S ribosomal protein L19e	161527898	YP_001581724	128
50S ribosomal protein L21e	161528892	YP_001582718	128
50S ribosomal protein L22	161528313	YP_001582139	128
50S ribosomal protein L25 -L23	161528315	YP_001582141	127
KOW domain-containing protein L24	161528307	YP_001582133	128
50S ribosomal protein L24 - L24e	161527735	YP_001581561	127
50S ribosomal protein L29	161528311	YP_001582137	125
50S ribosomal protein L30	161527908	YP_001581734	129
50S ribosomal protein L31e	161528771	YP_001582597	126
50S ribosomal protein L32e	161527899	YP_001581725	128
hypothetical protein Nmar_0434 -L37ae	161527942	YP_001581768	129
50S ribosomal protein L37e	161527728	YP_001581554	122
50S ribosomal protein L39e	161528772	YP_001582598	121
50S ribosomal protein L40e	161528633	YP_001582459	126
50S ribosomal protein L44e	161529087	YP_001582913	127
30S ribosomal protein S2	161527824	YP_001581650	129
30S ribosomal protein S3	161528312	YP_001582138	128
30S ribosomal protein S3Ae	161529020	YP_001582846	128
30S ribosomal protein S4	161527832	YP_001581658	129
30S ribosomal protein S4e	161528306	YP_001582132	128
30S ribosomal protein S5	161527907	YP_001581733	129
30S ribosomal protein S6e	161527583	YP_001581409	129
30S ribosomal protein S7	161527863	YP_001581689	129
30S ribosomal protein S8	161528303	YP_001582129	128
30S ribosomal protein S8e	161528028	YP_001581854	129
30S ribosomal protein S9	161527933	YP_001581759	129
30S ribosomal protein S10	161528541	YP_001582367	129
30S ribosomal protein S11	161528958	YP_001582784	129
30S ribosomal protein S12	161527862	YP_001581688	129
30S ribosomal protein S13	161527833	YP_001581659	129
30S ribosomal protein S14	161528304	YP_001582130	125
30S ribosomal protein S15	161529016	YP_001582842	128
30S ribosomal protein S17	161528309	YP_001582135	127
hypothetical protein Nmar_1430 -S17e	161528938	YP_001582764	128
30S ribosomal protein S19	161528314	YP_001582140	128
30S ribosomal protein S19e	161529132	YP_001582958	129
hypothetical protein Nmar_0535 – S24e	161528043	YP_001581869	127
30S ribosomal protein S27ae	161528042	YP_001581868	123
30S ribosomal protein S27E	161529088	YP_001582914	127
30S ribosomal protein S28e	161527734	YP_001581560	129
RNA polymerase Rbp10 – RpoP	161528540	YP_001582366	126
DNA-directed RNA polymerase subunit N – RpoN	161527826	YP_001581652	129
transcription termination factor Tfs – RpoM	161529264	YP_001583090	125
hypothetical protein Nmar_1757 – RpoL	161529265	YP_001583091	127
RNA polymerase Rpb6 – RpoK	161528440	YP_001582266	127
RNA polymerase Rpb5 – RpoH	161527854	YP_001581680	128
RNA polymerase Rpb4 – RpoF	161528891	YP_001582717	127
DNA-directed RNA polymerase subunit E RpoE2	161528744	YP_001582570	127
DNA-directed RNA polymerase subunit E' – RpoE1	161528743	YP_001582569	129
RNA polymerase insert RpoD	161527936	YP_001581762	129
DNA-directed RNA polymerase subunit B – RpoB2	161527855	YP_001581681	128
DNA-directed RNA polymerase subunit B – RpoB1	161527856	YP_001581681	128
bifunctional DNA-directed RNA polymerase subunit A'/A" – RpoA2	161527856	YP_001581682	128
bifunctional DNA-directed RNA polymerase subunit A'/A" – RpoA1	161527857	YP_001581683	129
NusG antitermination factor	161527894	YP_001581720	127
NusA family protein	161527861	YP_001581687	127

Annexe 5 : Liste des matrices utilisées pour l'analyse en désaturation par sélection de sites pour l'étude exposée dans le Chapitre 1.

Matrice	Nombre de positions
S_0	1958
S_1	3884
S_2	5439
S_3	6929
S_4	8373
S_5	9742
S_6	11089
S_7	12465
S_8	13806
S_9	15176
S_{10}	16508
S_{11}	17845
S_{12}	19167
S_{13}	20458
S_{14}	21746
S_{15}	23000
S_{16}	24143
S_{17}	25237
S_{19}	27217
S_{20}	28146
S_{22}	29825
S_{23}	30561
S_{24}	31225
S_{26}	32472
S_{27}	32999
S_{29}	33799
S_{31}	34449
S_{34}	35102
Totale	35589

Annexe 6 : Liste des matrices utilisées pour l'analyse en désaturation par sélection de gènes pour l'étude exposée dans le Chapitre 1.

Pour chaque marqueur, la matrice indiquée est la plus grande dans laquelle ce marqueur a été inclus. Par exemple, les marqueurs de la matrice DT30 sont aussi dans les matrices DT20 et DT10.

Le nombre de distance calculées correspond au nombre de couples d'espèces pour lesquels une distance a été mesurée, et qui sert de diviseur pour le calcul de la distance moyenne (maximum = 10).

Marqueur	Distance moyenne	Nombre de distances calculées	Matrice	Nombre de positions
s28e	0,81314591	10	DT10	1242
s19	0,87391316	10		
l12e	0,93431001	10		
rpoM	0,97764542	10		
l37e	0,99762814	10		
M036	1,05088415	10		
M046	1,05354258	10		
l7ae	1,0705633	10		
M066	1,07546077	6		
M197	1,12731502	6	DT20	2672
MA09	1,14941092	6		
rpoK	1,15097537	10		
s12	1,15257442	10		
s2	1,16728252	10		
s7	1,18210045	10		
s11	1,18845449	10		
l39e	1,19038229	10		
l5	1,2145064	10		
M128	1,24681426	10		
M039	1,25275645	10	DT30	4701
MA23	1,27057753	10		
M086	1,27195784	10		
M201	1,29563716	10		
M207	1,29925848	10		
s19e	1,32980492	10		
M171	1,33274006	10		
M136	1,33417573	10		
MA55	1,34222622	10		
l3	1,34508504	10		
rpoN	1,35059768	10	DT40	6330
M070	1,38272673	10		
MA42	1,39090948	6		
M147	1,39397627	10		
M194	1,41037915	10		
MA44	1,41982895	10		
s5	1,42965314	10		
M159	1,43083271	10		
nusA	1,43475249	10		
M122	1,44850164	10		

M053	1,45048367	10	DT50	8733
MA22	1,45300464	10		
M187	1,45823695	10		
s15	1,45895579	10		
MA52	1,45955437	6		
MA56	1,46319707	10		
M138	1,46852341	10		
M219	1,47410339	10		
rpoA1	1,47653898	10		
l15e	1,4770368	10		
M173	1,48723625	10	DT60	10963
M213	1,48769448	6		
MA54	1,48959821	10		
rpoB2	1,49780334	10		
s14	1,50123284	10		
s6e	1,50692094	10		
M184	1,5087578	10		
MA02	1,52833851	10		
l6	1,5326468	10		
M148	1,53295574	10		
M215	1,53507611	10	DT70	12840
l18	1,53508518	10		
M210	1,54374317	6		
l21e	1,54800109	10		
M126	1,55528883	10		
M037	1,55583385	10		
MA33	1,55617389	10		
s9	1,56276877	10		
l40e	1,57209471	10		
s17e	1,57961288	10		
s27ae	1,58240479	10	DT80	16340
M212	1,58492593	10		
M182	1,58702595	6		
M164	1,58786794	10		
M178	1,59166906	10		
MA20	1,59818404	10		
rpoB1	1,6056481	10		
M134	1,60834848	10		
M028	1,60895551	10		
M125	1,60974439	10		
s8e	1,61139908	10	DT90	18552
MA37	1,61157562	10		
MA08	1,61504253	10		
s10	1,62061821	10		
l11	1,62389996	10		
s27e	1,62453301	10		
M190	1,6253055	10		
s8	1,62783873	10		
M032	1,63652623	3		
M040	1,6390425	10		

M175	1,64378792	6	DT100	20629
M048	1,64612471	10		
l19e	1,64926715	10		
MA48	1,6518101	10		
s17	1,66160866	10		
M112	1,66920497	10		
M014	1,67635768	10		
MA53	1,67638445	10		
M145	1,67818156	10		
M115	1,68043672	10		
M193	1,68251642	10	DT110	22987
rpoE2	1,6835943	10		
M199	1,68750593	3		
rpoH	1,69432393	10		
M150	1,69511077	10		
M067	1,69860747	6		
M054	1,70618695	10		
l2	1,70618973	10		
M047	1,70769043	10		
M183	1,71158673	10		
M179	1,71368038	6	DT120	25510
M132	1,7166913	10		
M038	1,72678923	6		
M055	1,72787748	10		
M149	1,72921748	10		
M203	1,73160633	3		
M137	1,73293096	10		
s3ae	1,73610174	10		
MA05	1,73957769	10		
l22	1,74150042	10		
M043	1,74290963	10	DT130	27863
M205	1,75559127	10		
M002	1,75609693	10		
MA21	1,7635065	10		
MA34	1,76999781	10		
M166	1,77032753	10		
M069	1,77593921	10		
M160	1,77648761	10		
M198	1,77753262	10		
MA47	1,78662984	10		
M154	1,78957167	10	DT140	29880
l24e	1,79649302	10		
l24	1,79701265	10		
MA32	1,8027434	10		
l32e	1,82365418	10		
M218	1,82441005	10		
M140	1,82609143	10		
M121	1,83672235	10		
rpoA2	1,83942755	10		
M170	1,83986158	10		

M010	1,84349775	10	DT150	32350
M144	1,84549012	10		
M119	1,84695156	10		
M033	1,84957126	10		
M050	1,85102774	10		
MA06	1,85697567	3		
M189	1,8603604	6		
M041	1,86730981	10		
rpoD	1,87969418	10	DT160	34383
MA27	1,88091723	6		
M180	1,88631171	10		
l14	1,88733091	10		
M042	1,89380325	10		
l4	1,89637119	10		
M176	1,89869073	6		
MA17	1,90345972	6		
l1	1,915029	10		
M064	1,9263975	3	DT170	37838
MA10	1,9276212	1		
l37ae	1,92827006	10		
M045	1,94010169	10		
M151	1,94109423	10		
l10e	1,94242958	10		
M031	1,94342384	10		
M158	1,94502309	10		
M018	1,95293228	10		
M118	1,95612123	3		
M056	1,97297299	10	DT180	40222
M044	1,97307409	10		
MA25	1,97636206	10		
M029	1,9846848	10		
MA11	1,9968463	3		
M163	2,00887608	10		
M188	2,01565706	10		
l23	2,0258275	10		
M208	2,0295969	1		
M085	2,02974738	10		
s3	2,03084497	10	DT190	41985
M011	2,03361637	10		
MA01	2,03673252	10		
M186	2,03939054	10		
M021	2,05283813	10		
M135	2,05393557	10		
M153	2,06065383	10		
M143	2,06608771	10		
M206	2,06818987	10		
rpoE1	2,06850103	10		
M076	2,06999741	10		
M004	2,0786757	1		
MA03	2,07922682	10		

M027	2,08069095	10	DT200	44233
M057	2,09163333	10		
M181	2,0924164	10		
M078	2,09626543	10		
M129	2,0979118	3		
MA41	2,09916525	10		
M080	2,10355579	10		
M165	2,10695921	10	DT210	47068
M058	2,10953107	10		
M093	2,11022341	10		
M195	2,11249604	10		
M077	2,11973943	6		
M071	2,12595264	10		
MA30	2,13438893	10		
M035	2,13569968	10		
M025	2,13917813	10		
M030	2,14694397	10		
s4e	2,15617214	10	DT220	48803
M061	2,16334797	10		
MA36	2,16627175	10		
l30	2,17080781	10		
M052	2,1814731	10		
M142	2,18244907	6		
M204	2,18913913	6		
M117	2,19676109	10		
M162	2,22253408	10	DT230	50694
M156	2,22317808	10		
l31e	2,23807342	10		
MA45	2,24108395	10		
M123	2,24461605	10		
l18e	2,25657231	10		
M168	2,26454572	10		
M177	2,26865965	10		
l13	2,30213907	10		
M185	2,30378868	10		
M026	2,30489167	10	DT240	52179
rpoF	2,30735296	10		
M191	2,31472467	10		
nusG	2,31501369	10		
MA35	2,330827	10		
M192	2,33192669	10		
MA31	2,33737627	3		
l10	2,33958011	10		
M012	2,34153314	10		
l29	2,34431401	10		
M049	2,35663201	10	DT250	54165
MA50	2,36046493	10		
M209	2,38441758	6		
MA46	2,38842761	10		
M005	2,403694	1		

M051	2,42672475	10		
M124	2,45930511	10		
M161	2,46849158	10		
M157	2,47051129	10		
l15	2,47296853	10	DT260	55565
MA19	2,50098472	10		
MA18	2,5013873	3		
M127	2,5057969	3		
M060	2,51569924	10		
M202	2,51595733	10		
l44e	2,51782973	10		
M075	2,55022911	10		
MA40	2,58724886	10		
MA28	2,58987802	6		
M022	2,5985988	10	DT270	57148
M116	2,60849567	10		
M196	2,67002588	10		
M003	2,67553257	10		
M059	2,68817926	10		
rpoP	2,7186661	10		
MA39	2,7404339	10		
M146	2,75250332	10		
M130	2,79218596	10		
M092	2,82906282	10	DT280	58433
M174	2,83338458	10		
M034	2,84542223	10		
M217	2,87419791	10		
M200	2,93473747	3		
MA16	2,9804462	10		
s24e	3,03426218	10		
M068	3,08430638	10		
rpoL	3,55880201	10		

Annexe 7 : Matériels supplémentaires de l'Article 2 (Chapitre 2).

Supplementary material

Supplementary table S1. List of archaeal and bacterial complete genome sequences used in this work.

Supplementary table S2. List of all protein markers used in this work.

Supplementary table S3. List of species-pairs used for evolutionary estimations for gene-by-gene desaturation analysis.

Supplementary fig. S1. Maximum likelihood tree based on the complete set of markers (32 ribosomal proteins and 38 diverse conserved proteins, 9450 sites) and rooted using bacterial sequences as outgroup. Numbers at branches are bootstrap proportions. The scale bar indicates the number of substitutions per position.

Supplementary fig. S2. Unrooted maximum likelihood tree based on the complete set of markers (32 ribosomal proteins and 38 diverse conserved proteins, 9450 sites). Numbers at branches are bootstrap proportions. The scale bar indicates the number of substitutions per position.

Supplementary figs. S3-S12. Rooted maximum likelihood trees corresponding to the site-by-site desaturation analysis. Numbers at branches are bootstrap proportions. The scale bar indicates the number of substitutions per position.

Supplementary figs. S13-S22. Rooted maximum likelihood trees corresponding to the gene-by-gene desaturation analysis. Numbers at branches are bootstrap proportions. The scale bar indicates the number of substitutions per position.

	TaxID	Species		Phylum	Class/Order
		* : species retained for the construction of the final datasets (108 species)			
		ARCHAEA			
1	311458	*Candidatus* Caldiarchaeum subterraneum	*	Aigarchaeota	Unclassified
2	414004	*Cenarchaeum symbiosum A*	*	Thaumarchaeota	Cenarchaeales
3	436308	*Nitrosopumilus maritimus SCM1*	*	Thaumarchaeota	Nitrosopumilales
4	886738	*Candidatus* Nitrosoarchaeum limnia SFB1	*	Thaumarchaeota	Nitrosopumilales
5	1001994	*Candidatus* Nitrosoarchaeum koreensis MY1	*	Thaumarchaeota	Nitrosopumilales
6	859350	*Candidatus* Nitrosopumilus salaria BD31	*	Thaumarchaeota	Nitrosopumilales
7	797209	*Haladaptatus paucihalophilus DX253*	*	Euryarchaeota	Halobacteriales
8	795797	*Halalkalicoccus jeotgali B3*		Euryarchaeota	Halobacteriales
9	634497	*Haloarcula hispanica ATCC 33960*		Euryarchaeota	Halobacteriales
10	272569	*Haloarcula marismortui ATCC 43049*		Euryarchaeota	Halobacteriales
11	478009	*Halobacterium salinarum R1*		Euryarchaeota	Halobacteriales
12	309800	*Haloferax volcanii DS2*		Euryarchaeota	Halobacteriales
13	469382	*Halogeometricum borinquense DSM 11551*	*	Euryarchaeota	Halobacteriales
14	485914	*Halomicrobium mukohataei DSM 12286*		Euryarchaeota	Halobacteriales
15	756883	*halophilic archaeon DL31*		Euryarchaeota	Halobacteriales
16	797210	*Halopiger xanaduensis SH-6*		Euryarchaeota	Halobacteriales
17	362976	*Haloquadratum walsbyi DSM 16790*		Euryarchaeota	Halobacteriales
18	1033806	*Halorhabdus tiamatea SARL4B*		Euryarchaeota	Halobacteriales
19	519442	*Halorhabdus utahensis DSM 12940*	*	Euryarchaeota	Halobacteriales
20	416348	*Halorubrum lacusprofundi ATCC 49239*	*	Euryarchaeota	Halobacteriales
21	543526	*Haloterrigena turkmenica DSM 5511*		Euryarchaeota	Halobacteriales
22	547559	*Natrialba magadii ATCC 43099*	*	Euryarchaeota	Halobacteriales
23	797303	*Natrinema pellirubrum DSM 15624*		Euryarchaeota	Halobacteriales
24	797304	*Natronobacterium gregoryi SP2*		Euryarchaeota	Halobacteriales
25	348780	*Natronomonas pharaonis DSM 2160*	*	Euryarchaeota	Halobacteriales
26	1072681	*Candidatus* Haloredivivus sp. G17	*	Euryarchaeota	Nanohaloarchaea
27	889948	*Candidatus* Nanosalina sp. J07AB43	*	Euryarchaeota	Nanohaloarchaea
28	889962	*Candidatus* Nanosalinarum sp. J07AB56	*	Euryarchaeota	Nanohaloarchaea
29	410358	*Methanocorpusculum labreanum Z*	*	Euryarchaeota	Methanomicrobiales
30	368407	*Methanoculleus marisnigri JR1*	*	Euryarchaeota	Methanomicrobiales
31	882090	*Methanolinea tarda NOBI-1*		Euryarchaeota	Methanomicrobiales
32	937775	*Methanoplanus limicola DSM 2279*	*	Euryarchaeota	Methanomicrobiales
33	679926	*Methanoplanus petrolearius DSM 11571*		Euryarchaeota	Methanomicrobiales
34	456442	*Methanoregula boonei 6A8*		Euryarchaeota	Methanomicrobiales
35	521011	*Methanosphaerula palustris E1-9c*	*	Euryarchaeota	Methanomicrobiales
36	323259	*Methanospirillum hungatei JF-1*	*	Euryarchaeota	Methanomicrobiales
37	259564	*Methanococcoides burtonii DSM 6242*	*	Euryarchaeota	Methanosarcinales
38	644295	*Methanohalobium evestigatum Z-7303*		Euryarchaeota	Methanosarcinales
39	547558	*Methanohalophilus mahii DSM 5219*		Euryarchaeota	Methanosarcinales
40	990316	*Methanosaeta concilii GP6*		Euryarchaeota	Methanosarcinales
41	1110509	*Methanosaeta harundinacea 6Ac*	*	Euryarchaeota	Methanosarcinales
42	349307	*Methanosaeta thermophila PT*	*	Euryarchaeota	Methanosarcinales
43	679901	*Methanosalsum zhilinae DSM 4017*	*	Euryarchaeota	Methanosarcinales
44	188937	*Methanosarcina acetivorans C2A*		Euryarchaeota	Methanosarcinales
45	269797	*Methanosarcina barkeri str. Fusaro*		Euryarchaeota	Methanosarcinales
46	192952	*Methanosarcina mazei Go1*	*	Euryarchaeota	Methanosarcinales
47	351160	*Methanocella arvoryzae MRE50*	*	Euryarchaeota	Methanocellales
48	1041930	*Methanocella conradii HZ254*	*	Euryarchaeota	Methanocellales
49	304371	*Methanocella paludicola SANAE*	*	Euryarchaeota	Methanocellales
50	224325	*Archaeoglobus fulgidus DSM 4304*	*	Euryarchaeota	Archaeoglobales
51	572546	*Archaeoglobus profundus DSM 5631*	*	Euryarchaeota	Archaeoglobales
52	693661	*Archaeoglobus veneficus SNP6*	*	Euryarchaeota	Archaeoglobales
53	589924	*Ferroglobus placidus DSM 10642*	*	Euryarchaeota	Archaeoglobales
54	439481	*Aciduliprofundum boonei T469*	*	Euryarchaeota	Thermoplasmatales
55	333146	*Ferroplasma acidarmanus fer1*	*	Euryarchaeota	Thermoplasmatales
56	263820	*Picrophilus torridus DSM 9790*	*	Euryarchaeota	Thermoplasmatales
57	273075	*Thermoplasma acidophilum DSM 1728*	*	Euryarchaeota	Thermoplasmatales
58	273116	*Thermoplasma volcanium GSS1*	*	Euryarchaeota	Thermoplasmatales
59	573064	*Methanocaldococcus fervens AG86*		Euryarchaeota	Methanococcales
60	573063	*Methanocaldococcus infernus ME*	*	Euryarchaeota	Methanococcales
61	243232	*Methanocaldococcus jannaschii DSM 2661*	*	Euryarchaeota	Methanococcales
62	644281	*Methanocaldococcus sp. FS406-22*		Euryarchaeota	Methanococcales
63	579137	*Methanocaldococcus vulcanius M7*		Euryarchaeota	Methanococcales
64	419665	*Methanococcus aeolicus Nankai-3*	*	Euryarchaeota	Methanococcales
65	402880	*Methanococcus maripaludis C5*		Euryarchaeota	Methanococcales
66	406327	*Methanococcus vannielii SB*	*	Euryarchaeota	Methanococcales
67	456320	*Methanococcus voltae A3*		Euryarchaeota	Methanococcales
68	647113	*Methanothermococcus okinawensis IH1*		Euryarchaeota	Methanococcales
69	647171	*Methanotorris formicicus Mc-S-70*		Euryarchaeota	Methanococcales
70	880724	*Methanotorris igneus Kol 5*	*	Euryarchaeota	Methanococcales
71	868132	*Methanobacterium sp. AL-21*	*	Euryarchaeota	Methanobacteriales
72	634498	*Methanobrevibacter ruminantium M1*		Euryarchaeota	Methanobacteriales
73	420247	*Methanobrevibacter smithii ATCC 35061*	*	Euryarchaeota	Methanobacteriales
74	339860	*Methanosphaera stadtmanae DSM 3091*	*	Euryarchaeota	Methanobacteriales
75	79929	*Methanothermobacter marburgensis str. Marburg*		Euryarchaeota	Methanobacteriales
76	187420	*Methanothermobacter thermautotrophicus str. Delta H*	*	Euryarchaeota	Methanobacteriales

77	523846	*Methanothermus fervidus DSM 2088*	*	Euryarchaeota	Methanobacteriales
78	190192	*Methanopyrus kandleri AV19*	*	Euryarchaeota	Methanopyrales
79	272844	*Pyrococcus abyssi GE5*	*	Euryarchaeota	Thermococcales
80	186497	*Pyrococcus furiosus DSM 3638*		Euryarchaeota	Thermococcales
81	70601	*Pyrococcus horikoshii OT3*		Euryarchaeota	Thermococcales
82	342949	*Pyrococcus sp. NA2*		Euryarchaeota	Thermococcales
83	529709	*Pyrococcus yayanosii CH1*	*	Euryarchaeota	Thermococcales
84	391623	*Thermococcus barophilus MP*	*	Euryarchaeota	Thermococcales
85	593117	*Thermococcus gammatolerans EJ3*	*	Euryarchaeota	Thermococcales
86	69014	*Thermococcus kodakarensis KOD1*		Euryarchaeota	Thermococcales
87	523849	*Thermococcus litoralis DSM 5473*	*	Euryarchaeota	Thermococcales
88	523850	*Thermococcus onnurineus NA1*		Euryarchaeota	Thermococcales
89	604354	*Thermococcus sibiricus MM 739*		Euryarchaeota	Thermococcales
90	246969	*Thermococcus sp. AM4*		Euryarchaeota	Thermococcales
91	425595	*Candidatus* Micrarchaeum acidiphilum ARMAN-2	*	Euryarchaeota	ARMAN
92	662760	*Candidatus* Parvarchaeum acidiphilum ARMAN-4	*	Euryarchaeota	ARMAN
93	662762	*Candidatus* Parvarchaeum acidophilus ARMAN-5	*	Euryarchaeota	ARMAN
94	228908	*Nanoarchaeum equitans Kin4-M*	*	Euryarchaeota	Nanoarchaeota
95	274854	uncultured marine group II euryarchaeote	*	Euryarchaeota	Unclassified
96	666510	*Acidilobus saccharovorans 345-15*	*	Crenarchaeota	Desulfurococcales
97	272557	*Aeropyrum pernix K1*	*	Crenarchaeota	Desulfurococcales
98	768672	*Desulfurococcus fermentans DSM 16532*		Crenarchaeota	Desulfurococcales
99	490899	*Desulfurococcus kamchatkensis 1221n*	*	Crenarchaeota	Desulfurococcales
100	765177	*Desulfurococcus mucosus DSM 2162*		Crenarchaeota	Desulfurococcales
101	415426	*Hyperthermus butylicus DSM 5456*	*	Crenarchaeota	Desulfurococcales
102	453591	*Ignicoccus hospitalis KIN4/I*	*	Crenarchaeota	Desulfurococcales
103	583356	*Ignisphaera aggregans DSM 17230*	*	Crenarchaeota	Desulfurococcales
104	694429	*Pyrolobus fumarii 1A*	*	Crenarchaeota	Desulfurococcales
105	591019	*Staphylothermus hellenicus DSM 12710*	*	Crenarchaeota	Desulfurococcales
106	399550	*Staphylothermus marinus F1*		Crenarchaeota	Desulfurococcales
107	633148	*Thermosphaera aggregans DSM 11486*	*	Crenarchaeota	Desulfurococcales
108	933801	*Acidianus hospitalis W1*	*	Crenarchaeota	Sulfolobales
109	1006006	*Metallosphaera cuprina Ar-4*		Crenarchaeota	Sulfolobales
110	399549	*Metallosphaera sedula DSM 5348*	*	Crenarchaeota	Sulfolobales
111	671065	*Metallosphaera yellowstonensis MK1*	*	Crenarchaeota	Sulfolobales
112	330779	*Sulfolobus acidocaldarius DSM 639*	*	Crenarchaeota	Sulfolobales
113	429572	*Sulfolobus islandicus L.S.2.15*		Crenarchaeota	Sulfolobales
114	273057	*Sulfolobus solfataricus P2*	*	Crenarchaeota	Sulfolobales
115	273063	*Sulfolobus tokodaii str. 7*	*	Crenarchaeota	Sulfolobales
116	397948	*Caldivirga maquilingensis IC-167*	*	Crenarchaeota	Thermoproteales
117	178306	*Pyrobaculum aerophilum str. IM2*	*	Crenarchaeota	Thermoproteales
118	340102	*Pyrobaculum arsenaticum DSM 13514*	*	Crenarchaeota	Thermoproteales
119	410359	*Pyrobaculum calidifontis JCM 11548*		Crenarchaeota	Thermoproteales
120	384616	*Pyrobaculum islandicum DSM 4184*	*	Crenarchaeota	Thermoproteales
121	698757	*Pyrobaculum neutrophilum V24Sta*		Crenarchaeota	Thermoproteales
122	1104324	*Pyrobaculum oguniense TE7*		Crenarchaeota	Thermoproteales
123	368408	*Pyrobaculum sp. 1860*		Crenarchaeota	Thermoproteales
124	444157	*Thermofilum pendens Hrk 5*	*	Crenarchaeota	Thermoproteales
125	768679	*Thermoproteus tenax Kra 1*	*	Crenarchaeota	Thermoproteales
126	999630	*Thermoproteus uzoniensis 768-20*		Crenarchaeota	Thermoproteales
127	572478	*Vulcanisaeta distributa DSM 14429*	*	Crenarchaeota	Thermoproteales
128	985053	*Vulcanisaeta moutnovskia 768-28*		Crenarchaeota	Thermoproteales
129	374847	*Candidatus* Korarchaeum cryptofilum OPF8	*	Korachaeota	

BACTERIA

130	134676	*Actinoplanes sp. SE50/110*		Actinobacteria	Actinobacteria
131	1133849	*Nocardia brasiliensis ATCC 700358*	*	Actinobacteria	Actinobacteria
132	101510	*Rhodococcus jostii RHA1*		Actinobacteria	Actinobacteria
133	1179773	*Saccharothrix espanaensis DSM 44229*		Actinobacteria	Actinobacteria
134	463191	*Streptomyces sviceus ATCC 29083*		Actinobacteria	Actinobacteria
135	224324	*Aquifex aeolicus VF5*		Aquificae	Aquificae
136	608538	*Hydrogenobacter thermophilus TK-6*		Aquificae	Aquificae
137	123214	*Persephonella marina EX-H1*	*	Aquificae	Aquificae
138	436114	*Sulfurihydrogenibium sp. YO3AOP1*		Aquificae	Aquificae
139	648996	*Thermovibrio ammonificans HB-1*		Aquificae	Aquificae
140	485918	*Chitinophaga pinensis DSM 2588*		Bacteroidetes/Chlorobi group	Bacteroidetes
141	760192	*Haliscomenobacter hydrossis DSM 1100*		Bacteroidetes/Chlorobi group	Bacteroidetes
142	700598	*Niastella koreensis GR20-10*		Bacteroidetes/Chlorobi group	Bacteroidetes
143	290317	*Chlorobium phaeobacteroides DSM 266*	*	Bacteroidetes/Chlorobi group	Chlorobi
144	517418	*Chloroherpeton thalassium ATCC 35110*		Bacteroidetes/Chlorobi group	Chlorobi
145	945713	*Ignavibacterium album JCM 16511*		Bacteroidetes/Chlorobi group	Ignavibacteria
146	511051	*Caldisericum exile AZM16c01*		Caldiserica	Caldisericia
147	765952	*Parachlamydia acanthamoebae UV-7*		Chlamydiae/Verrucomicrobia group	Chlamydiae
148	331113	*Simkania negevensis Z*		Chlamydiae/Verrucomicrobia group	Chlamydiae
149	716544	*Waddlia chondrophila WSU 86-1044*		Chlamydiae/Verrucomicrobia group	Chlamydiae

150	583355	*Coraliomargarita akajimensis DSM 45221*		Chlamydiae/Verrucomicrobia group	Verrucomicrobia
151	452637	*Opitutus terrae PB90-1*	*	Chlamydiae/Verrucomicrobia group	Verrucomicrobia
152	926569	*Anaerolinea thermophila UNI-1*		Chloroflexi	Anaerolineae
153	926550	*Caldilinea aerophila DSM 14535 = NBRC 104270*		Chloroflexi	Caldilineae
154	326427	*Chloroflexus aggregans DSM 9485*		Chloroflexi	Chloroflexi
155	552811	*Dehalogenimonas lykanthroporepellens BL-DC-9*	*	Chloroflexi	Dehalococcoidetes
156	309801	*Thermomicrobium roseum DSM 5159*		Chloroflexi	Thermomicrobia
157	653733	*Desulfurispirillum indicum S5*	*	Chrysiogenetes	Chrysiogenetes
158	329726	*Acaryochloris marina MBIC11017*		Cyanobacteria	Chroococcales
159	251221	*Gloeobacter violaceus PCC 7421*		Cyanobacteria	Gloeobacteria
160	63737	*Nostoc punctiforme PCC 73102*		Cyanobacteria	Nostocales
161	179408	*Oscillatoria nigro-viridis PCC 7112*	*	Cyanobacteria	Oscillatoriales
162	251229	*Chroococcidiopsis thermalis PCC 7203*		Cyanobacteria	Pleurocapsales
163	59922	*Prochlorococcus marinus str. MIT 9303*		Cyanobacteria	Prochlorales
164	768670	*Calditerrivibrio nitroreducens DSM 19672*	*	Deferribacteres	Deferribacteres
165	639282	*Deferribacter desulfuricans SSM1*		Deferribacteres	Deferribacteres
166	522772	*Denitrovibrio acetiphilus DSM 12809*		Deferribacteres	Deferribacteres
167	717231	*Flexistipes sinusarabici DSM 4947*		Deferribacteres	Deferribacteres
168	937777	*Deinococcus peraridilitoris DSM 19664*		Deinococcus-Thermus	Deinococci
169	526227	*Meiothermus silvanus DSM 9946*	*	Deinococcus-Thermus	Deinococci
170	670487	*Oceanithermus profundus DSM 14977*		Deinococcus-Thermus	Deinococci
171	743525	*Thermus scotoductus SA-01*		Deinococcus-Thermus	Deinococci
172	649638	*Truepera radiovictrix DSM 17093*		Deinococcus-Thermus	Deinococci
173	309799	*Dictyoglomus thermophilum H-6-12*		Dictyoglomi	Dictyoglomia
174	515635	*Dictyoglomus turgidum DSM 6724*	*	Dictyoglomi	Dictyoglomia
175	445932	*Elusimicrobium minutum Pei191*		Elusimicrobia	Elusimicrobia
176	471821	uncultured Termite group 1 bacterium phylotype Rs-D17	*	Elusimicrobia	environmental samples
177	240015	*Acidobacterium capsulatum ATCC 51196*		Fibrobacteres/Acidobacteria group	Acidobacteria
178	234267	*Candidatus* Solibacter usitatus Ellin6076	*	Fibrobacteres/Acidobacteria group	Acidobacteria
179	682795	*Granulicella mallensis MP5ACTX8*		Fibrobacteres/Acidobacteria group	Acidobacteria
180	926566	*Terriglobus roseus DSM 18391*		Fibrobacteres/Acidobacteria group	Acidobacteria
181	59374	*Fibrobacter succinogenes subsp. succinogenes S85*		Fibrobacteres/Acidobacteria group	Fibrobacteres
182	1195464	*Bacillus thuringiensis MC28*		Firmicutes	Bacilli
183	997761	*Paenibacillus mucilaginosus K02*		Firmicutes	Bacilli
184	573061	*Clostridium cellulovorans 743B*		Firmicutes	Clostridia
185	768706	*Desulfosporosinus orientis DSM 765*	*	Firmicutes	Clostridia
186	650150	*Erysipelothrix rhusiopathiae*		Firmicutes	Erysipelotrichi
187	479436	*Veillonella parvula DSM 2008*		Firmicutes	Negativicutes
188	190304	*Fusobacterium nucleatum subsp. nucleatum ATCC 25586*		Fusobacteria	Fusobacteriia
189	572544	*Ilyobacter polytropus DSM 2926*		Fusobacteria	Fusobacteriia
190	523794	*Leptotrichia buccalis C-1013-b*		Fusobacteria	Fusobacteriia
191	526218	*Sebaldella termitidis ATCC 33386*	*	Fusobacteria	Fusobacteriia
192	519441	*Streptobacillus moniliformis DSM 12112*		Fusobacteria	Fusobacteriia
193	379066	*Gemmatimonas aurantiaca T-27*		Gemmatimonadetes	Gemmatimonadetes
194	330214	*Candidatus* Nitrospira defluvii		Nitrospirae	Nitrospira
195	1048260	*Leptospirillum ferriphilum ML-04*		Nitrospirae	Nitrospira
196	1162668	*Leptospirillum ferrooxidans C2-3*	*	Nitrospirae	Nitrospira
197	289376	*Thermodesulfovibrio yellowstonii DSM 11347*		Nitrospirae	Nitrospira
198	1142394	*Phycisphaera mikurensis NBRC 102666*		Planctomycetes	Phycisphaerae
199	530564	*Pirellula staleyi DSM 6068*		Planctomycetes	Planctomycetia
200	756272	*Planctomyces brasiliensis DSM 5305*	*	Planctomycetes	Planctomycetia
201	243090	*Rhodopirellula baltica SH 1*		Planctomycetes	Planctomycetia
202	886293	*Singulisphaera acidiphila DSM 18658*		Planctomycetes	Planctomycetia
203	311403	*Agrobacterium radiobacter K84*		Proteobacteria	Alphaproteobacteria
204	137722	*Azospirillum sp. B510*	*	Proteobacteria	Alphaproteobacteria
205	1037409	*Bradyrhizobium japonicum USDA 6*		Proteobacteria	Alphaproteobacteria
206	460265	*Methylobacterium nodulans ORS 2060*		Proteobacteria	Alphaproteobacteria
207	216596	*Rhizobium leguminosarum bv. viciae 3841*		Proteobacteria	Alphaproteobacteria
208	762376	*Achromobacter xylosoxidans A8*		Proteobacteria	Betaproteobacteria
209	266265	*Burkholderia xenovorans LB400*	*	Proteobacteria	Betaproteobacteria
210	1042878	*Cupriavidus necator N-1*		Proteobacteria	Betaproteobacteria
211	398578	*Delftia acidovorans SPH-1*		Proteobacteria	Betaproteobacteria
212	381666	*Ralstonia eutropha H16*		Proteobacteria	Betaproteobacteria
213	572480	*Arcobacter nitrofigilis DSM 7299*	*	Proteobacteria	delta/epsilon subdivisions
214	706587	*Desulfomonile tiedjei DSM 6799*	*	Proteobacteria	delta/epsilon subdivisions
215	502025	*Haliangium ochraceum DSM 14365*		Proteobacteria	delta/epsilon subdivisions
216	246197	*Myxococcus xanthus DK 1622*		Proteobacteria	delta/epsilon subdivisions
217	709032	*Sulfuricurvum kujiense DSM 16994*		Proteobacteria	delta/epsilon subdivisions
218	349521	*Hahella chejuensis KCTC 2396*	*	Proteobacteria	Gammaproteobacteria
219	1191061	*Klebsiella oxytoca E718*		Proteobacteria	Gammaproteobacteria
220	592316	*Pantoea sp. At-9b*		Proteobacteria	Gammaproteobacteria
221	220664	*Pseudomonas protegens Pf-5*		Proteobacteria	Gammaproteobacteria

222	338187	*Vibrio harveyi ATCC BAA-1116*		Proteobacteria	Gammaproteobacteria
223	573825	*Leptospira interrogans serovar Lai str. IPAV*	*	Spirochaetes	Spirochaetia
224	158190	*Sphaerochaeta pleomorpha str. Grapes*		Spirochaetes	Spirochaetia
225	573413	*Spirochaeta smaragdinae DSM 11293*		Spirochaetes	Spirochaetia
226	545694	*Treponema primitia ZAS-2*		Spirochaetes	Spirochaetia
227	869212	*Turneriella parva DSM 21527*		Spirochaetes	Spirochaetia
228	572547	*Aminobacterium colombiense DSM 12261*		Synergistetes	Synergistia
229	584708	*Aminomonas paucivorans DSM 12260*		Synergistetes	Synergistia
230	891968	*Anaerobaculum mobile DSM 13181*	*	Synergistetes	Synergistia
231	580340	*Thermovirga lienii DSM 17291*		Synergistetes	Synergistia
232	441768	*Acholeplasma laidlawii PG-8A*		Tenericutes	Mollicutes
233	322098	*Aster yellows witches'-broom phytoplasma AYWB*		Tenericutes	Mollicutes
234	265311	*Mesoplasma florum L1*		Tenericutes	Mollicutes
235	272633	*Mycoplasma penetrans HF-2*		Tenericutes	Mollicutes
236	565575	*Ureaplasma urealyticum serovar 10 str. ATCC 33699*		Tenericutes	Mollicutes
237	667014	*Thermodesulfatator indicus DSM 15286*	*	Thermodesulfobacteria	Thermodesulfobacteria
238	795359	*Thermodesulfobacterium sp. OPB45*		Thermodesulfobacteria	Thermodesulfobacteria
239	521045	*Kosmotoga olearia TBF 19.5.1*		Thermotogae	Thermotogae
240	443254	*Marinitoga piezophila KA3*		Thermotogae	Thermotogae
241	660470	*Mesotoga prima MesG1.Ag.4.2*		Thermotogae	Thermotogae
242	403833	*Petrotoga mobilis SJ95*		Thermotogae	Thermotogae
243	126740	*Thermotoga sp. RQ2*	*	Thermotogae	Thermotogae
244	880073	*Caldithrix abyssi DSM 13497*		unclassified Bacteria	Caldithrix
245	671143	*Candidatus* Methylomirabilis oxyfera	*	unclassified Bacteria	candidate division NC10
246	525904	*Thermobaculum terrenum ATCC BAA-798*	*	unclassified Bacteria	Thermobaculum

Annotation	Reference gi	Reference accession	COG number	COG class	COG definition
argininosuccinate synthase	161528794	YP_001582620.1	COG0137	E	Amino acid transport and metabolism
homoserine kinase	161527553	YP_001581379.1	COG0083	E	Amino acid transport and metabolism
homoserine dehydrogenase	161528204	YP_001582030.1	COG0460	E	Amino acid transport and metabolism
aspartate kinase	161529262	YP_001583088.1	COG0527	E	Amino acid transport and metabolism
aspartate carbamoyltransferase	161529194	YP_001583020.1	COG0540	F	Nucleotide transport and metabolism
phosphoribosylformylglycinamidine cyclo-ligase	161528090	YP_001581916.1	COG0150	F	Nucleotide transport and metabolism
uridylate kinase putative	161529213	YP_001583039.1	COG0528	F	Nucleotide transport and metabolism
phosphoribosylformylglycinamidine synthase II	161528288	YP_001582114.1	COG0046	F	Nucleotide transport and metabolism
adenylosuccinate lyase	161528963	YP_001582789.1	COG0015	F	Nucleotide transport and metabolism
RdgB/HAM1 family non-canonical purine NTP pyrophosphatase	161529038	YP_001582864.1	COG0127	F	Nucleotide transport and metabolism
glutamine amidotransferase class-II	161528289	YP_001582115.1	COG0034	F	Nucleotide transport and metabolism
phosphoglycerate kinase	161528008	YP_001581834.1	COG0126	G	Carbohydrate transport and metabolism
phosphopantothenoylcysteine decarboxylase/phosphopantothenate--cysteine ligase	161529229	YP_001583055.1	COG0452	H	Coenzyme transport and metabolism
pyridoxine biosynthesis protein	161528987	YP_001582813.1	COG0214	H	Coenzyme transport and metabolism
UbiD family decarboxylase	161529105	YP_001582931.1	COG0043	H	Coenzyme transport and metabolism
porphobilinogen deaminase	118576540	YP_876283.1	COG0181	H	Coenzyme transport and metabolism
glutamate-1-semialdehyde-2 1-aminomutase	161527998	YP_001581824.1	COG0001	H	Coenzyme transport and metabolism
molybdenum cofactor biosynthesis protein C	315427188	BAJ48802.1	COG0315	H	Coenzyme transport and metabolism
conserved hypothetical protein	315426923	BAJ48542.1	COG3425	I	Lipid transport and metabolism
beta-lactamase domain-containing protein	161529047	YP_001582873.1	COG1236	J	Translation, ribosomal structure and biogenesis
glutamyl-tRNA(Gln) amidotransferase subunit E	161527611	YP_001581437.1	COG2511	J	Translation, ribosomal structure and biogenesis
tRNA-guanine transglycosylase	161529044	YP_001582870.1	COG0343	J	Translation, ribosomal structure and biogenesis
glutamyl-tRNA(Gln) amidotransferase B subunit	161528376	YP_001582202.1	COG0064	J	Translation, ribosomal structure and biogenesis
exosome complex exonuclease 1	161527940	YP_001581766.1	COG0689	J	Translation, ribosomal structure and biogenesis
MiaB-like tRNA modifying enzyme	161528769	YP_001582595.1	COG0621	J	Translation, ribosomal structure and biogenesis
phenylalanyl-tRNA synthetase alpha subunit	161528997	YP_001582823.1	COG0016	J	Translation, ribosomal structure and biogenesis
prolyl-tRNA synthetase	315425642	BAJ47301.1	COG0442	J	Translation, ribosomal structure and biogenesis
histidyl-tRNA synthetase	315425704	BAJ47360.1	COG0124	J	Translation, ribosomal structure and biogenesis
5-nucleotidase SurE	315426747	BAJ48371.1	COG0522	J	Translation, ribosomal structure and biogenesis
peptidase M50	161527799	YP_001581625.1	COG0750	M	Cell wall/membrane/envelope biogenesis
cytidyltransferase-like protein	161528806	YP_001582632.1	COG0615	MI	Cell wall/membrane/envelope biogenesis
metalloendopeptidase glycoprotease family	161529041	YP_001582867.1	COG0533	O	Post-translational modification, protein turnover, and chaperones
hydrogenase maturation protein HypF	315426449	BAJ48087.1	COG0068	O	Post-translational modification, protein turnover, and chaperones
hydrogenase expression formation protein HypD	315426462	BAJ48095.1	COG0409	O	Post-translational modification, protein turnover, and chaperones
GTP1/OBG protein	161529113	YP_001582939.1	COG2262	R	General function prediction only
LPPG:FO 2-phospho-L-lactate transferase	161528134	YP_001581960.1	COG0391	S	Function unknown
GTP-binding signal recognition particle	161528039	YP_001581865.1	COG0541	U	Intracellular trafficking, secretion, and vesicular transport
signal recognition particle receptor	315425933	BAJ47583.1	COG0552	U	Intracellular trafficking, secretion, and vesicular transport
50S ribosomal protein L1	161527890	YP_001581716			
50S ribosomal protein L2	161527614	YP_001581440			
50S ribosomal protein L3P	161528317	YP_001582143			
50S ribosomal protein L4P	161528316	YP_001582142			
50S ribosomal protein L5	161528305	YP_001582131			
50S ribosomal protein L6	161528302	YP_001582128			
acidic ribosomal protein P0 – L10	161527889	YP_001581715			
50S ribosomal protein L10e	161527957	YP_001581783			
50S ribosomal protein L11	161527893	YP_001581719			
50S ribosomal protein L13	161527934	YP_001581760			
50S ribosomal protein L14	161528308	YP_001582134			
50S ribosomal protein L15	161527909	YP_001581735			
50S ribosomal protein L18	161527906	YP_001581732			
50S ribosomal protein L22	161528313	YP_001582139			
50S ribosomal protein L25	161528315	YP_001582141			

KOW domain-containing protein – L24	161528307	YP_001582133
50S ribosomal protein L29	161528311	YP_001582137
50S ribosomal protein L30	161527908	YP_001581734
30S ribosomal protein S2	161527824	YP_001581650
30S ribosomal protein S3	161528312	YP_001582138
30S ribosomal protein S4	161527832	YP_001581658
30S ribosomal protein S5	161527907	YP_001581733
30S ribosomal protein S7	161527863	YP_001581689
30S ribosomal protein S8	161528303	YP_001582129
30S ribosomal protein S9	161527933	YP_001581759
30S ribosomal protein S10	161528541	YP_001582367
30S ribosomal protein S11	161528958	YP_001582784
30S ribosomal protein S12	161527862	YP_001581688
30S ribosomal protein S13	161527833	YP_001581659
30S ribosomal protein S15	161529016	YP_001582842
30S ribosomal protein S17	161528309	YP_001582135
30S ribosomal protein S19	161528314	YP_001582140

Nitrosopumilus maritimus SCM1 versus *Methanoculleus marisnigri JR1*
Nitrosopumilus maritimus SCM1 versus *Thermococcus gammatolerans EJ3*
Nitrosopumilus maritimus SCM1 versus *Metallosphaera sedula DSM 5348*
Nitrosopumilus maritimus SCM1 versus *Pyrobaculum islandicum DSM 4184*
Methanoculleus marisnigri JR1 versus *Thermococcus gammatolerans EJ3*
Methanoculleus marisnigri JR1 versus *Metallosphaera sedula DSM 5348*
Methanoculleus marisnigri JR1 versus Pyrobaculum islandicum DSM 4184
Thermococcus gammatolerans EJ3 versus *Metallosphaera sedula DSM 5348*
Thermococcus gammatolerans EJ3 versus *Pyrobaculum islandicum DSM 4184*
Metallosphaera sedula DSM 5348 versus *Pyrobaculum islandicum DSM 4184*
Nitrosopumilus maritimus SCM1 versus *Persephonella marina EX-H1*
Nitrosopumilus maritimus SCM1 versus *Azospirillum sp. B510*
Nitrosopumilus maritimus SCM1 versus *Oscillatoria nigro-viridis PCC 7112*
Methanoculleus marisnigri JR1 versus *Persephonella marina EX-H1*
Methanoculleus marisnigri JR1 versus *Azospirillum sp. B510*
Methanoculleus marisnigri JR1 versus *Oscillatoria nigro-viridis PCC 7112*
Metallosphaera sedula DSM 5348 versus *Persephonella marina EX-H1*
Metallosphaera sedula DSM 5348 versus *Azospirillum sp. B510*
Metallosphaera sedula DSM 5348 versus *Oscillatoria nigro-viridis PCC 7112*
Persephonella marina EX-H1 versus *Azospirillum sp. B510*
Persephonella marina EX-H1 versus *Oscillatoria nigro-viridis PCC 7112*
Azospirillum sp. B510 versus *Oscillatoria nigro-viridis PCC 7112*

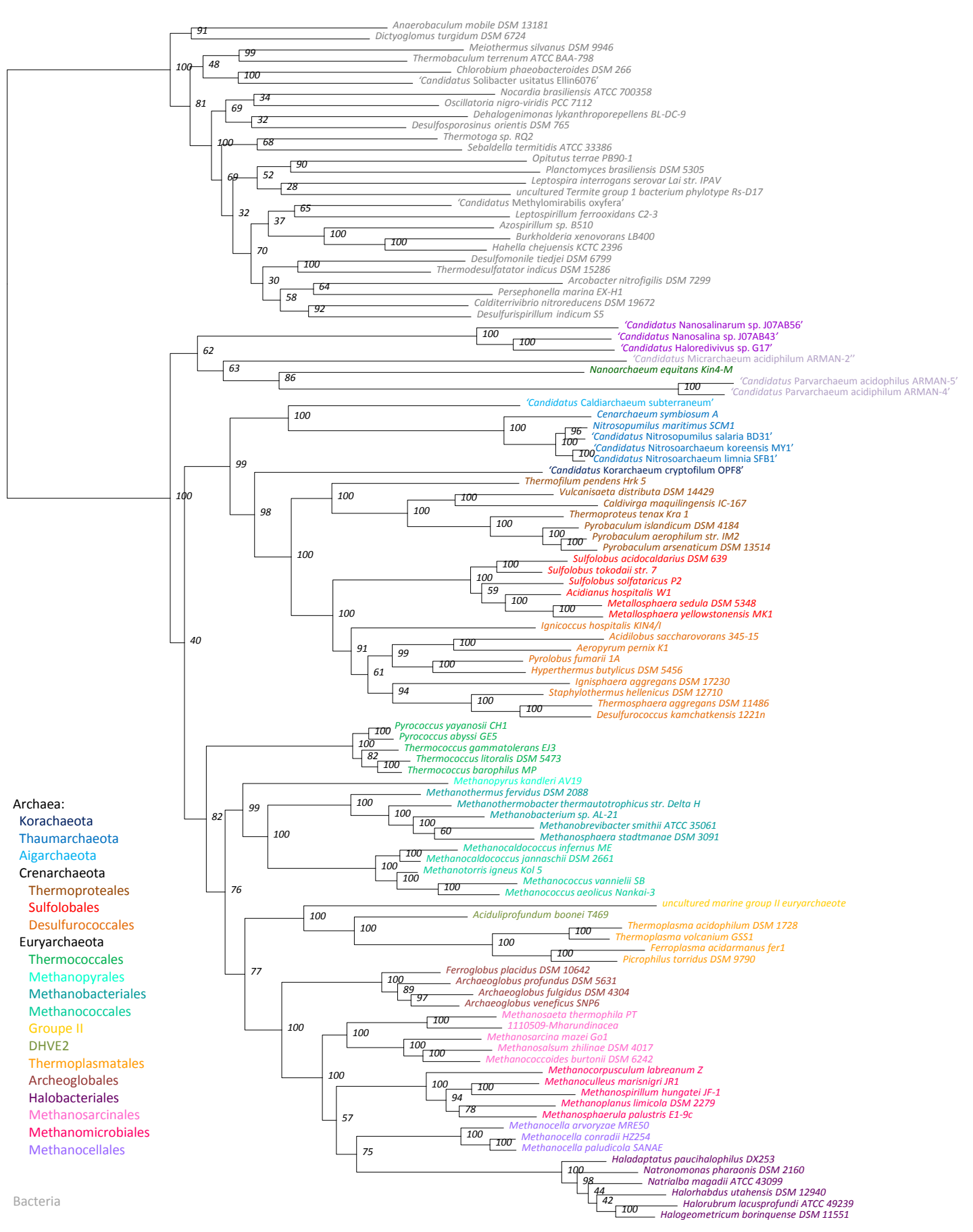


Figure S1
108 species – 9450 positions

Figure S2

81 species – 9450 positions

Figure S3
101 species - 939 positions

Figure S4
101 species - 2032 positions

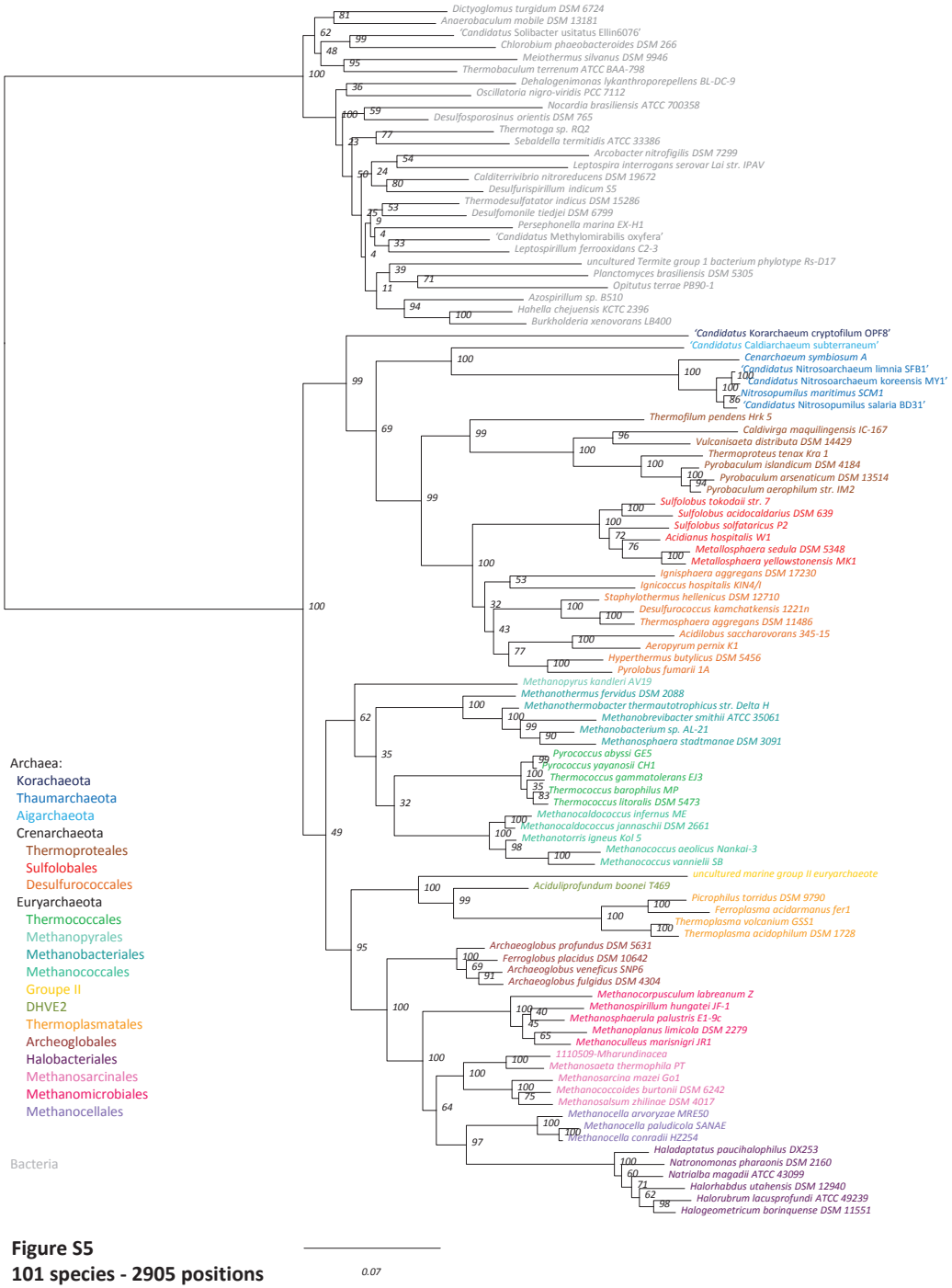


Figure S5
101 species - 2905 positions

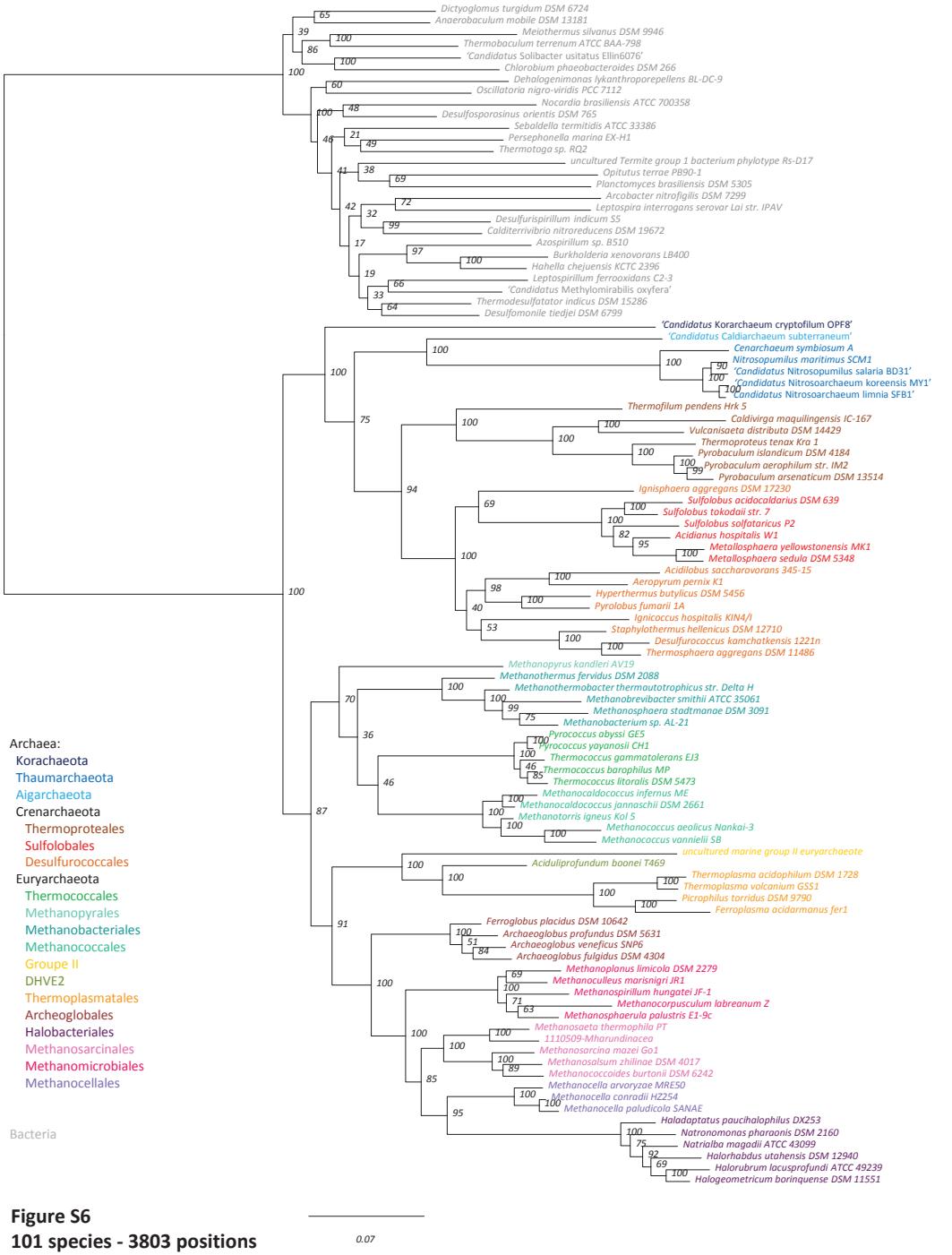


Figure S6
101 species - 3803 positions

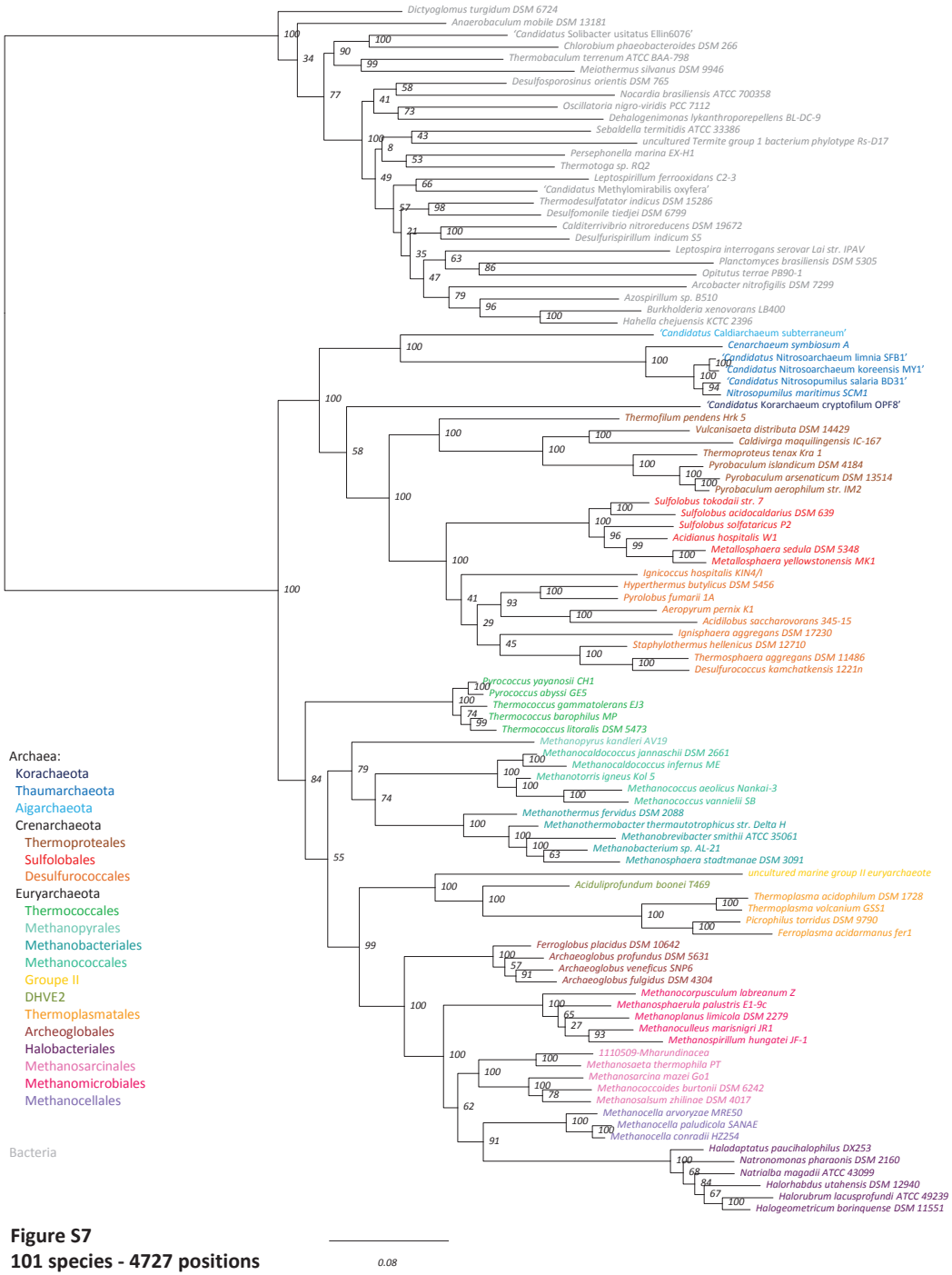


Figure S7
101 species - 4727 positions

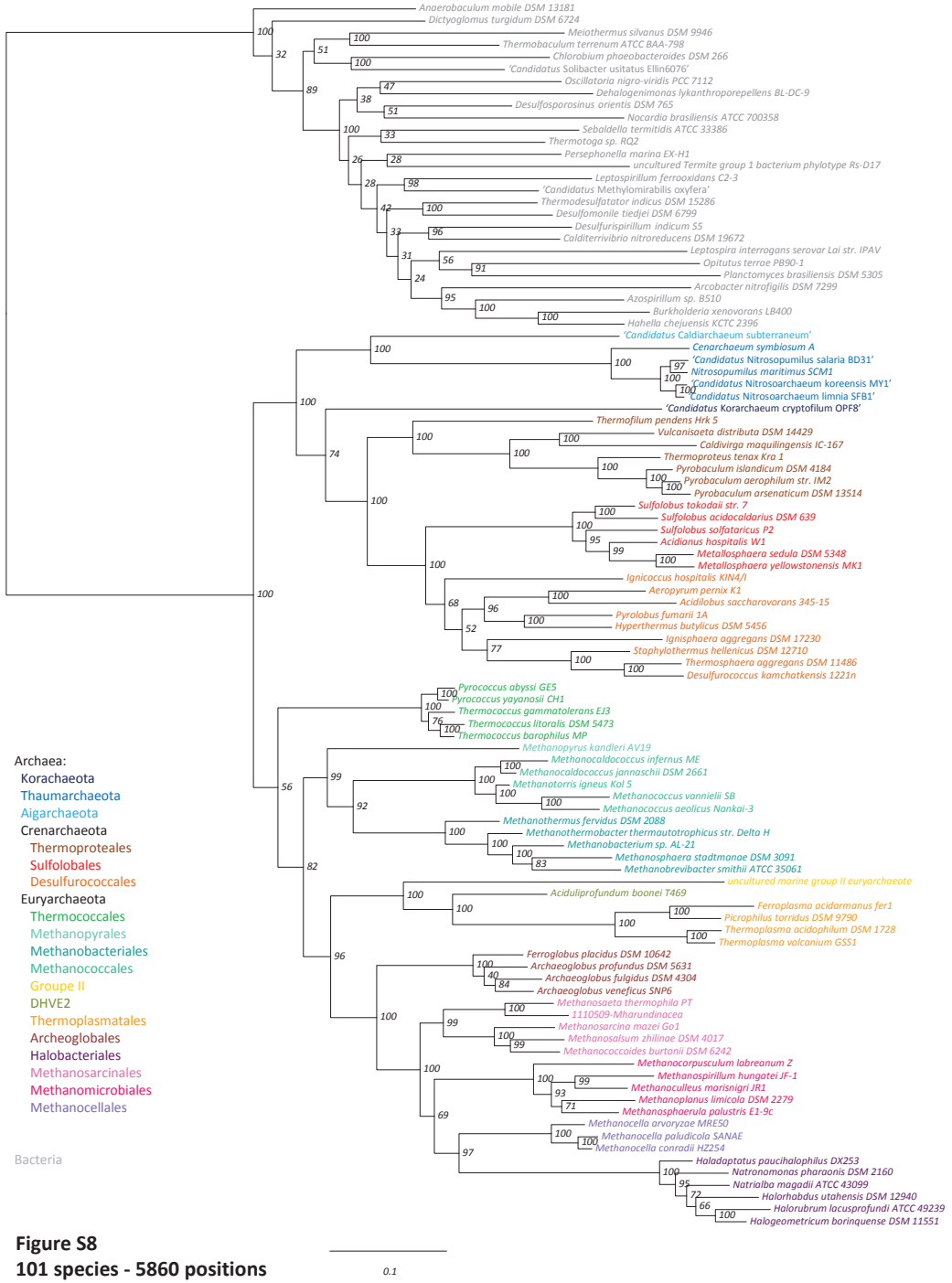


Figure S8
101 species - 5860 positions

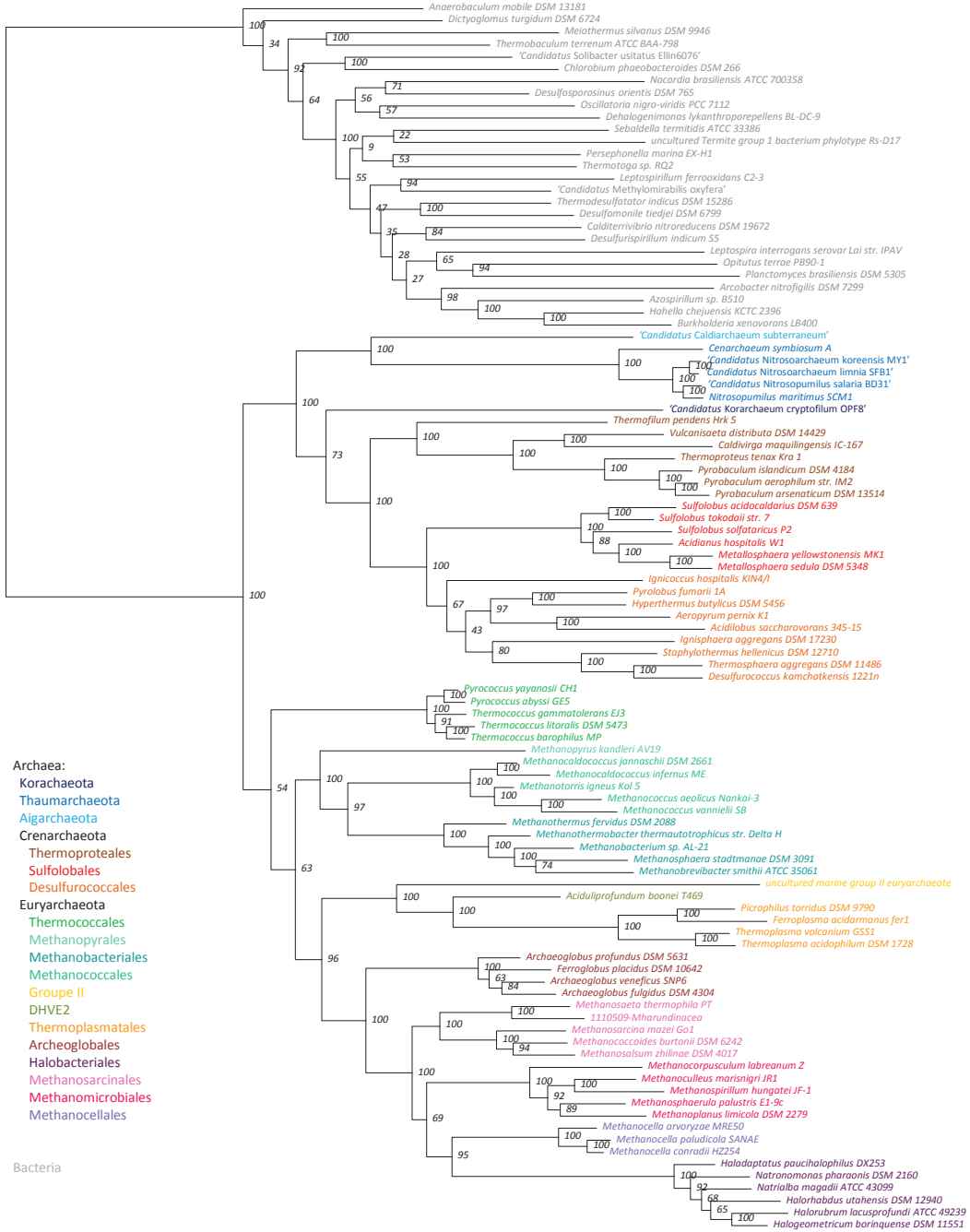


Figure S9
101 species - 6939 positions

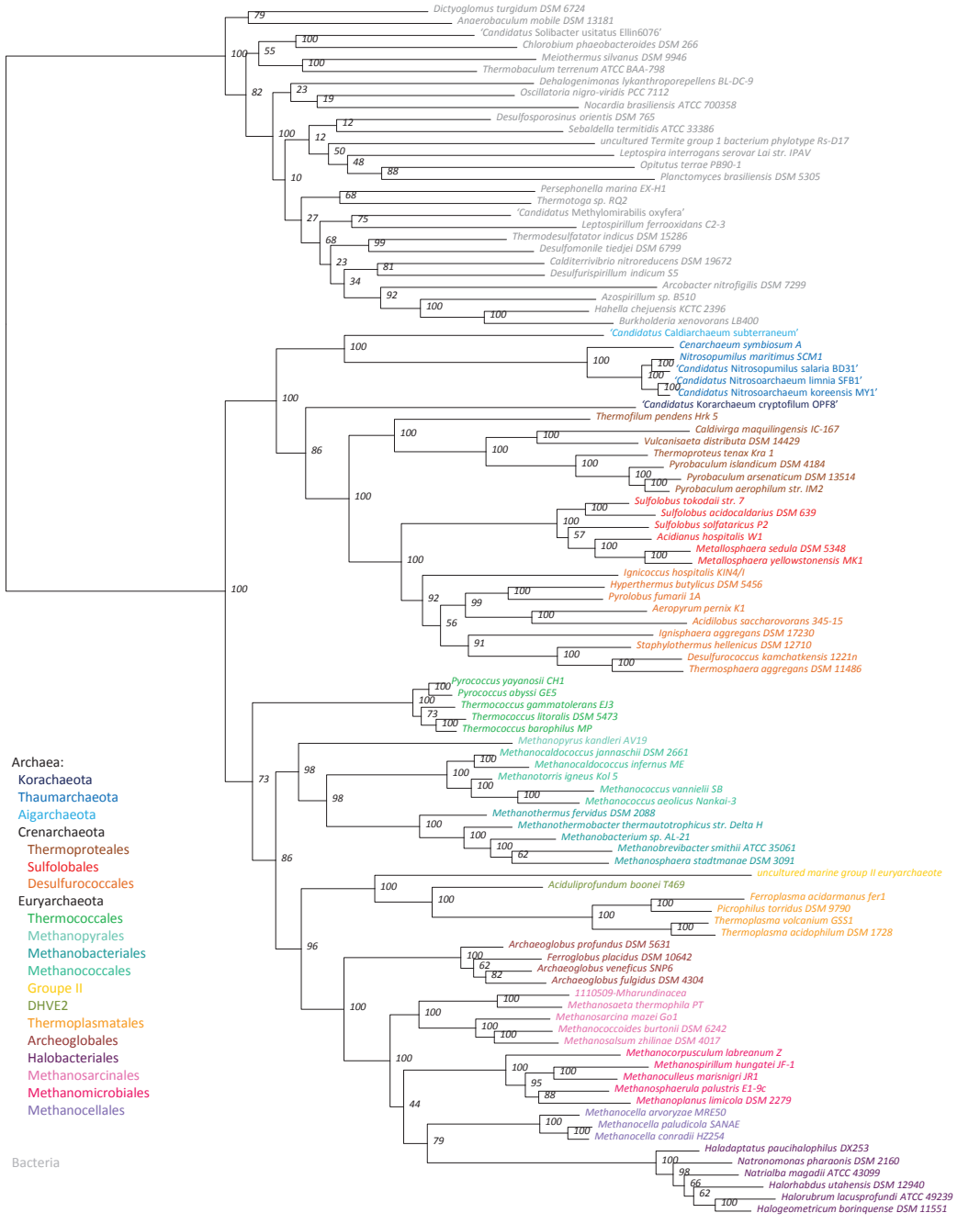


Figure S10
101 species - 7978 positions

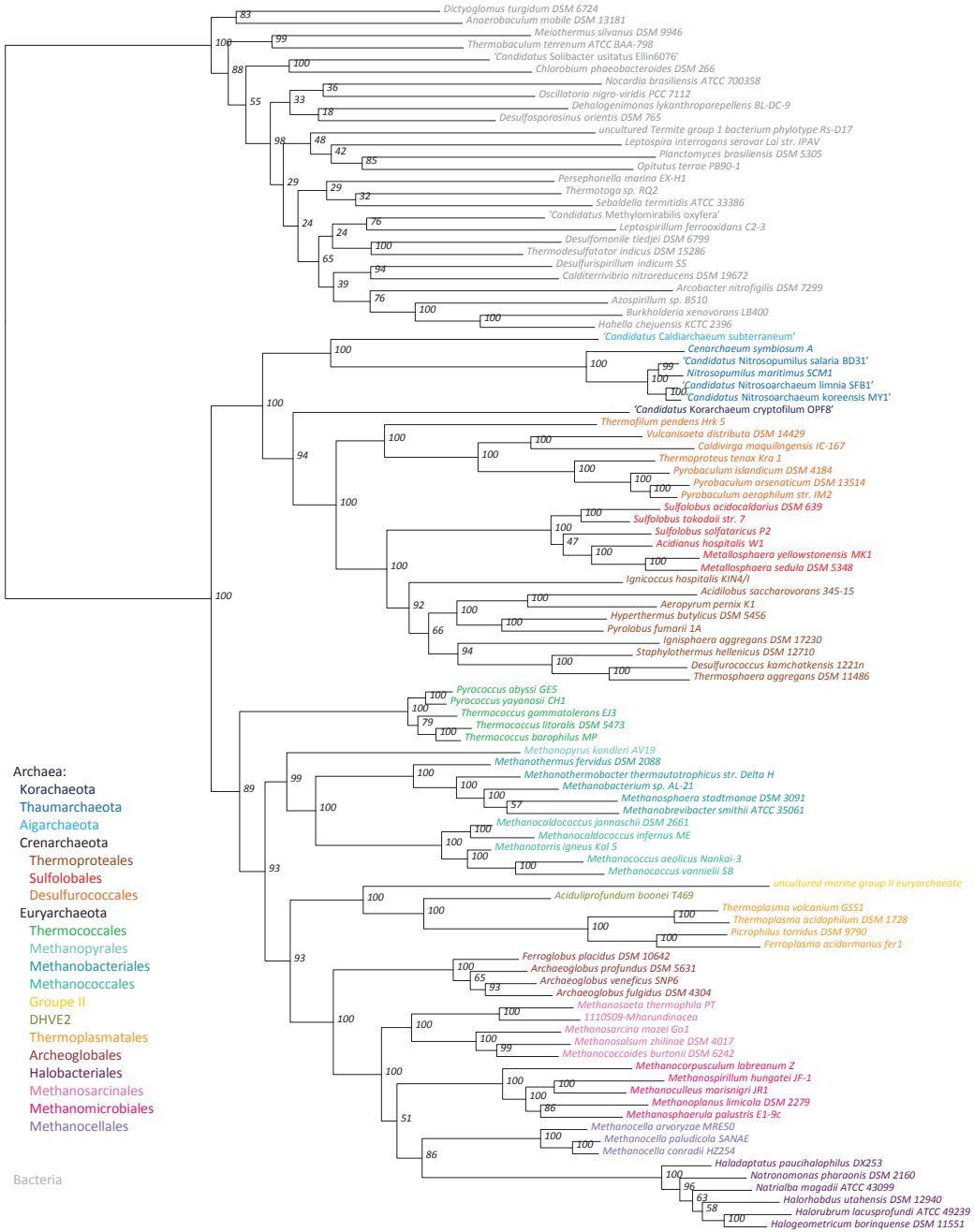


Figure S11
101 species - 8980 positions

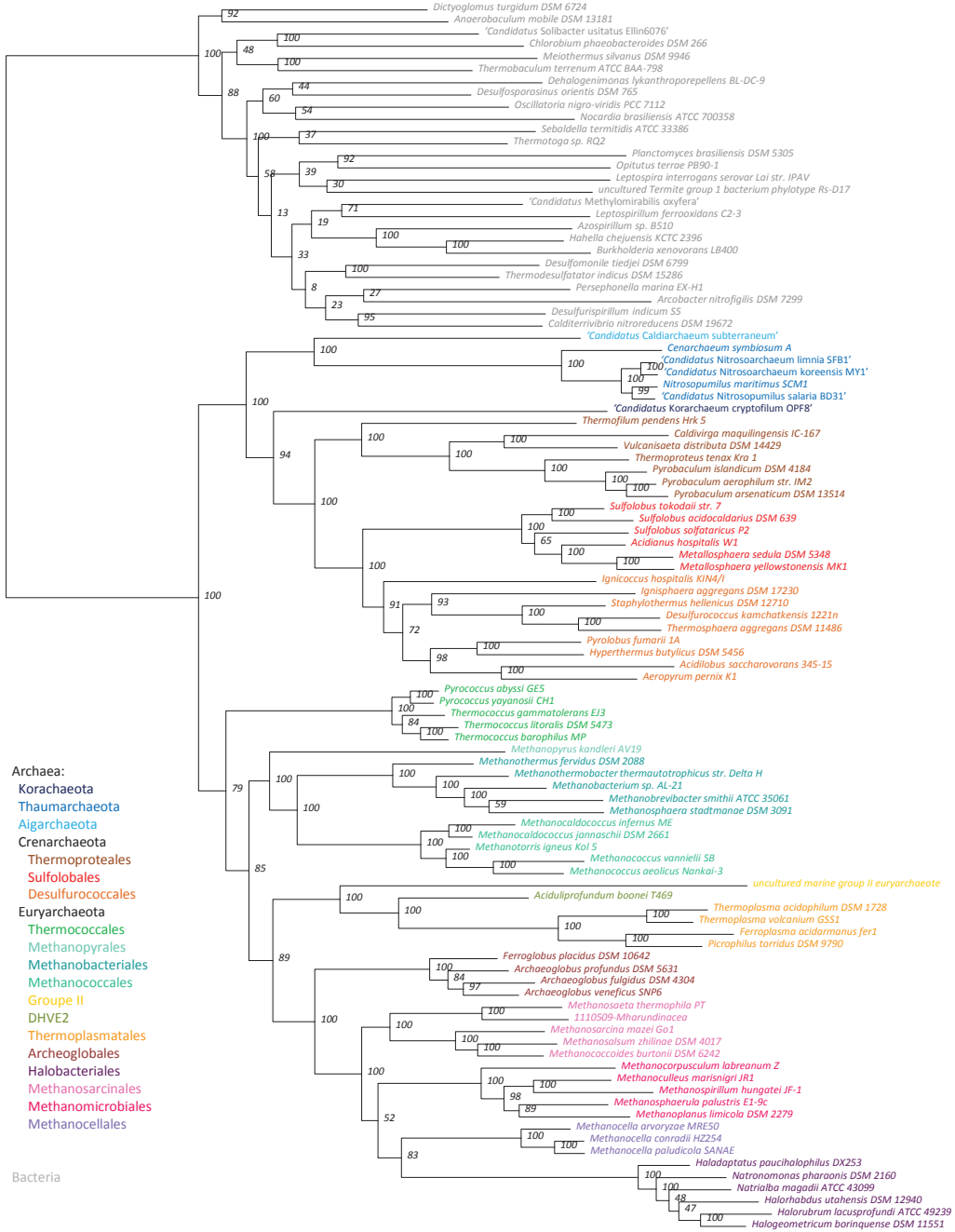


Figure S12
101 species - 9450 positions

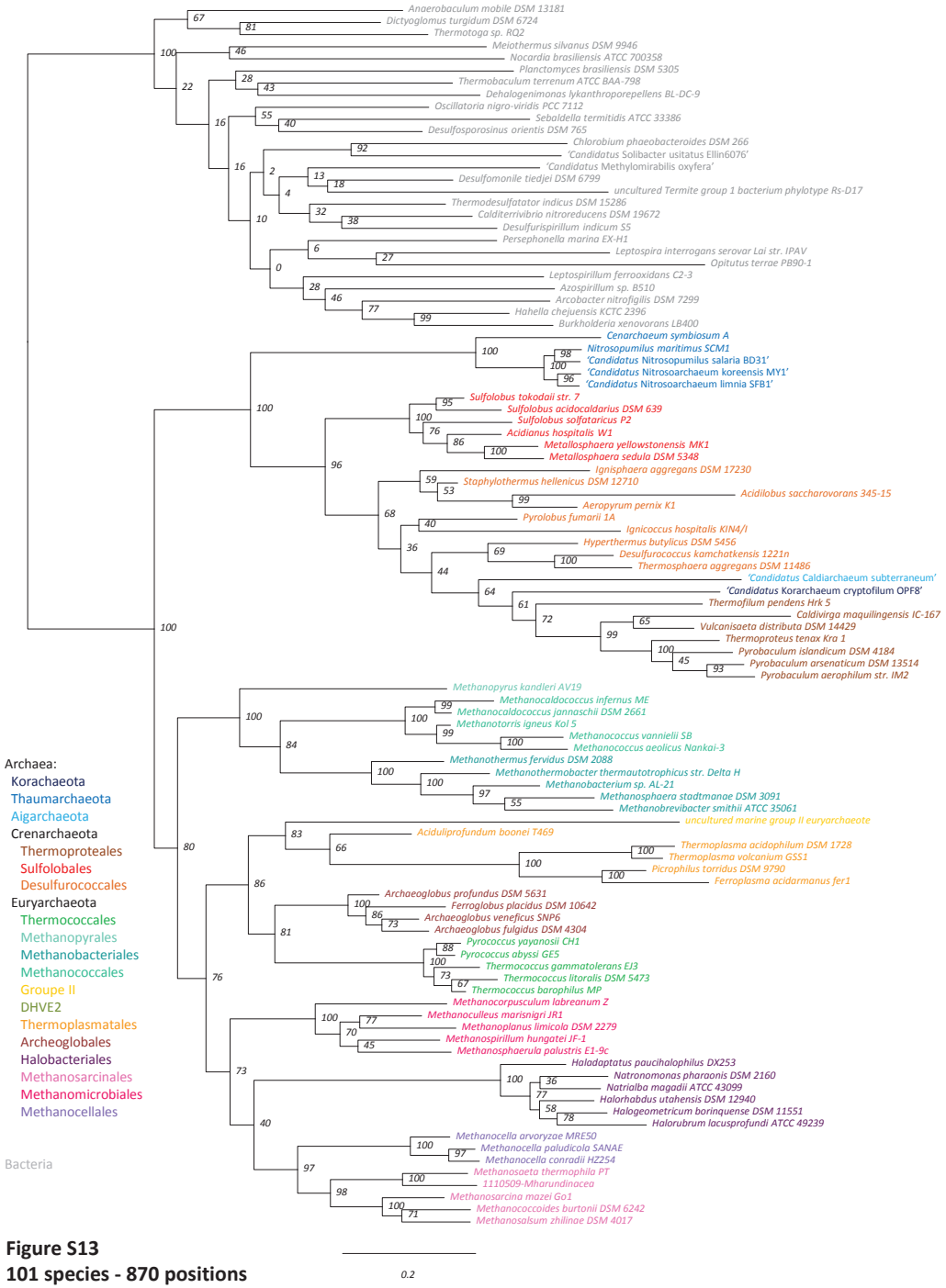


Figure S13
101 species - 870 positions

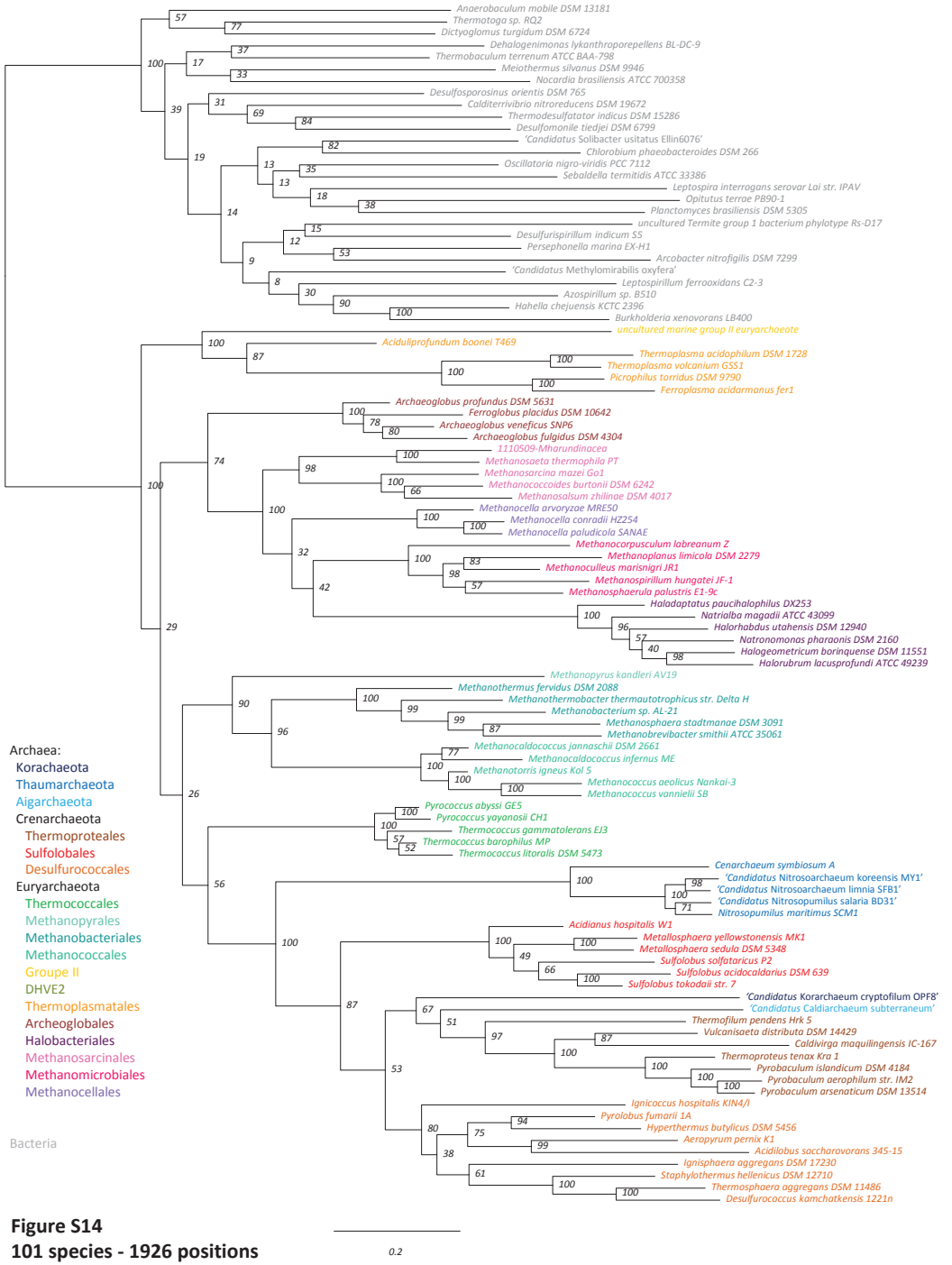


Figure S14
101 species - 1926 positions

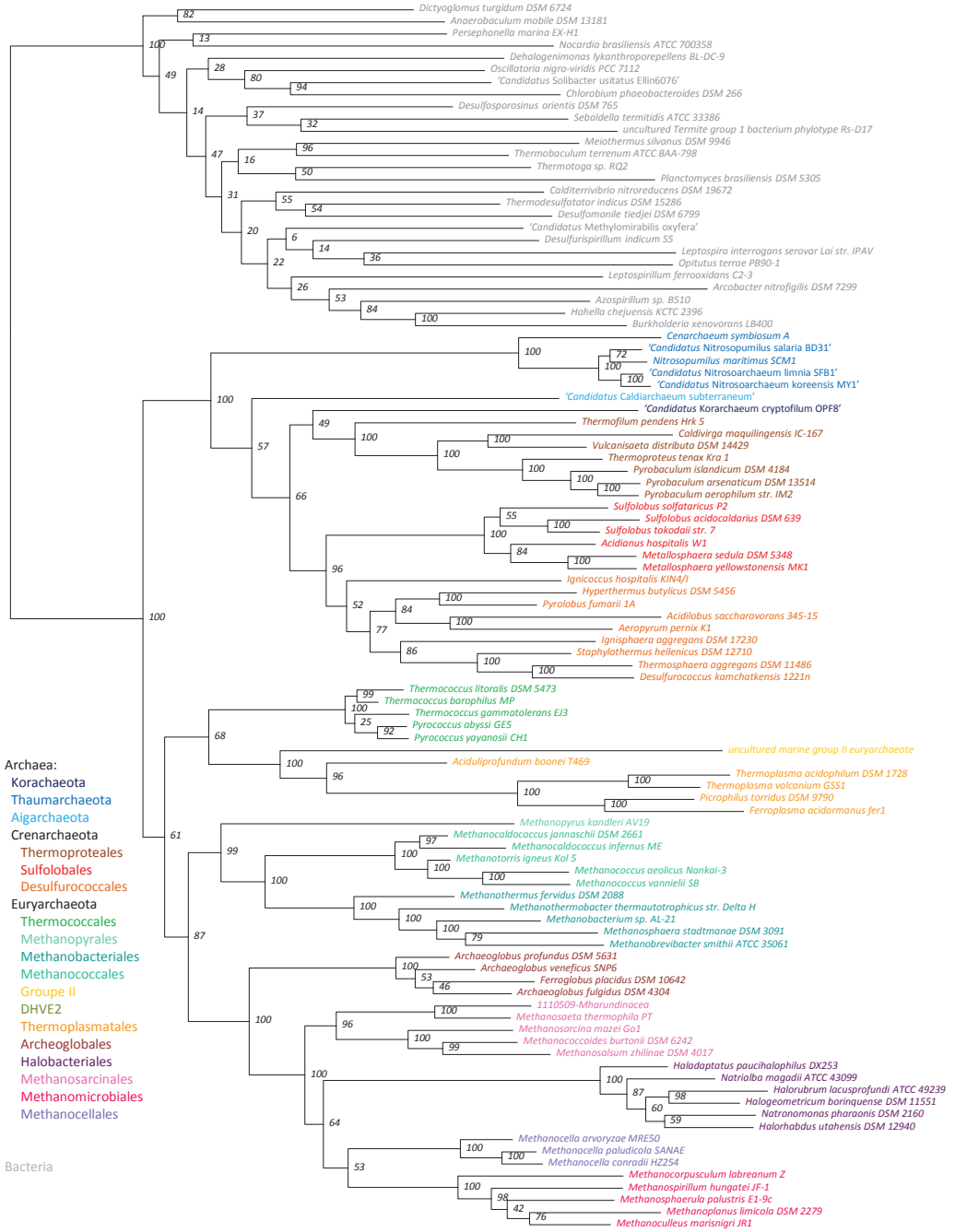


Figure S15
101 species - 3172 positions

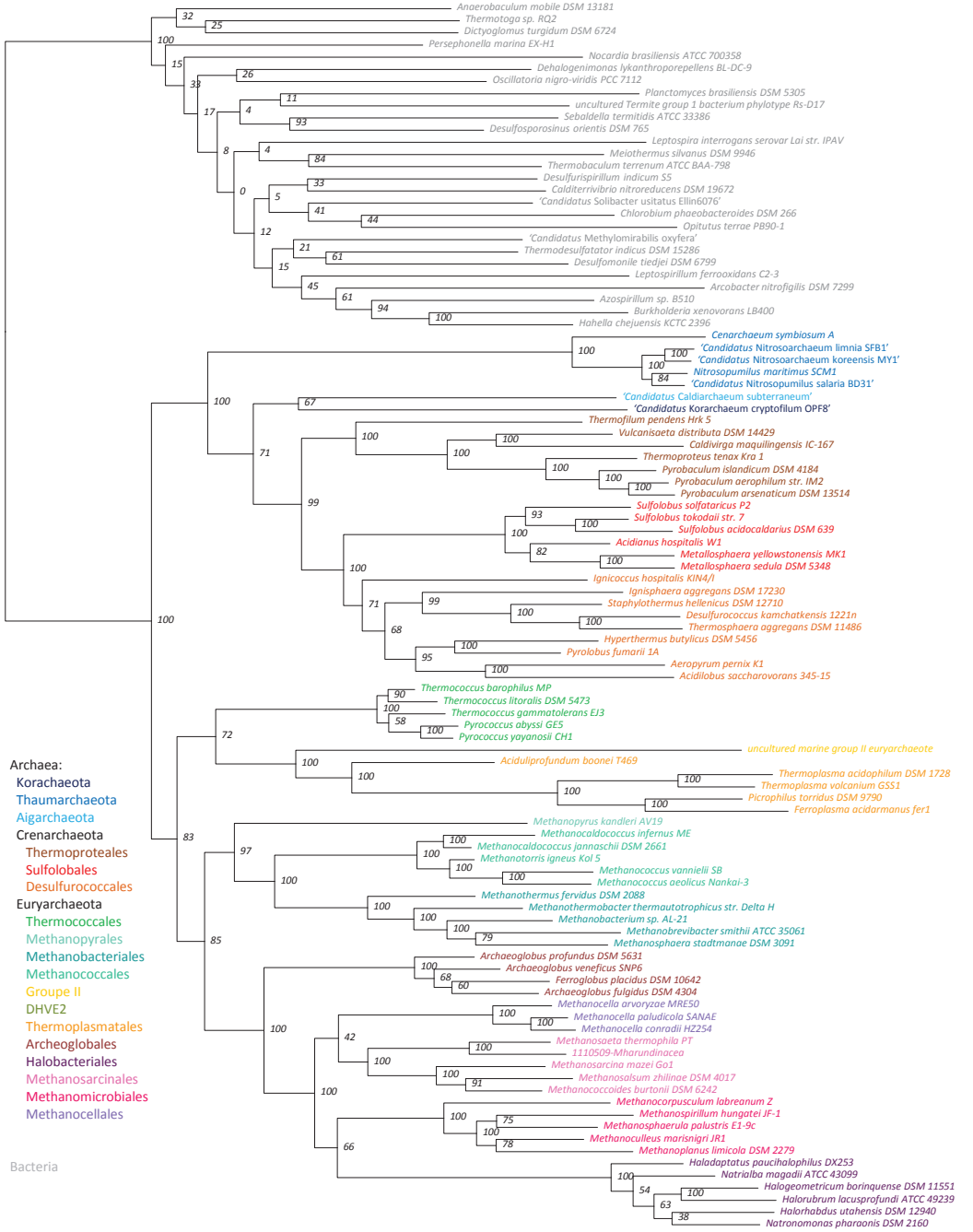


Figure S16
101 species - 4175 positions

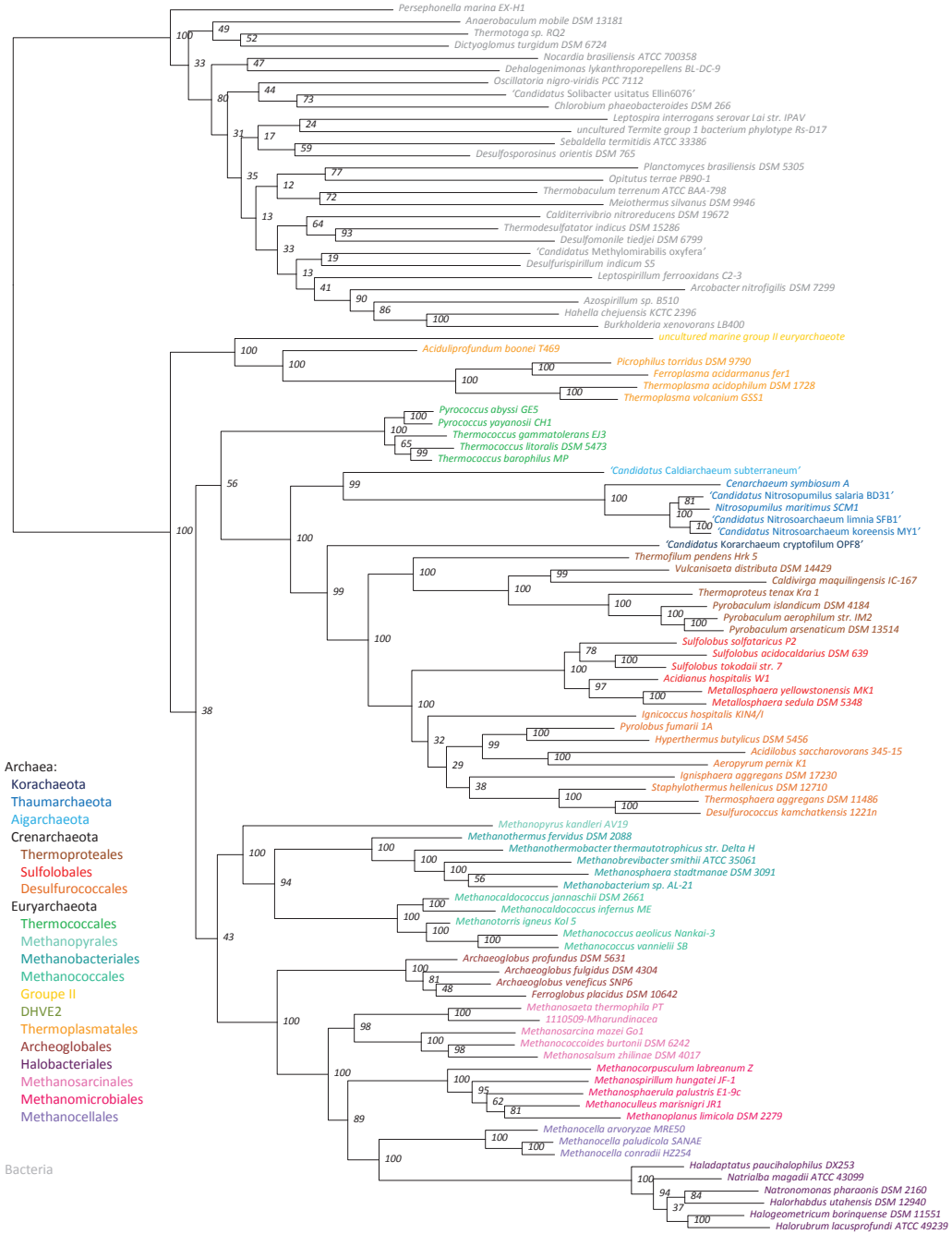


Figure S17
101 species - 4902 positions

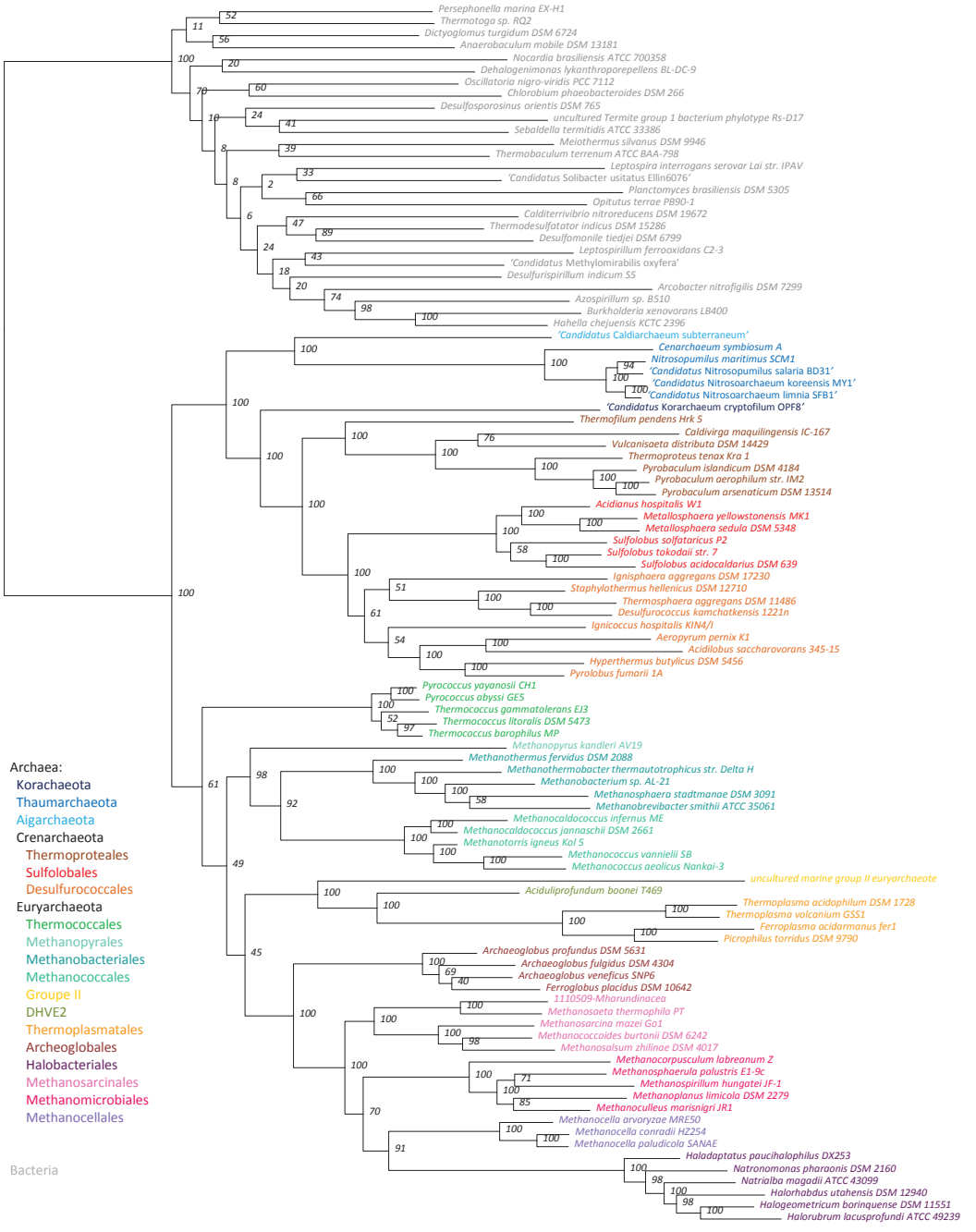


Figure S18
101 species - 5633 positions

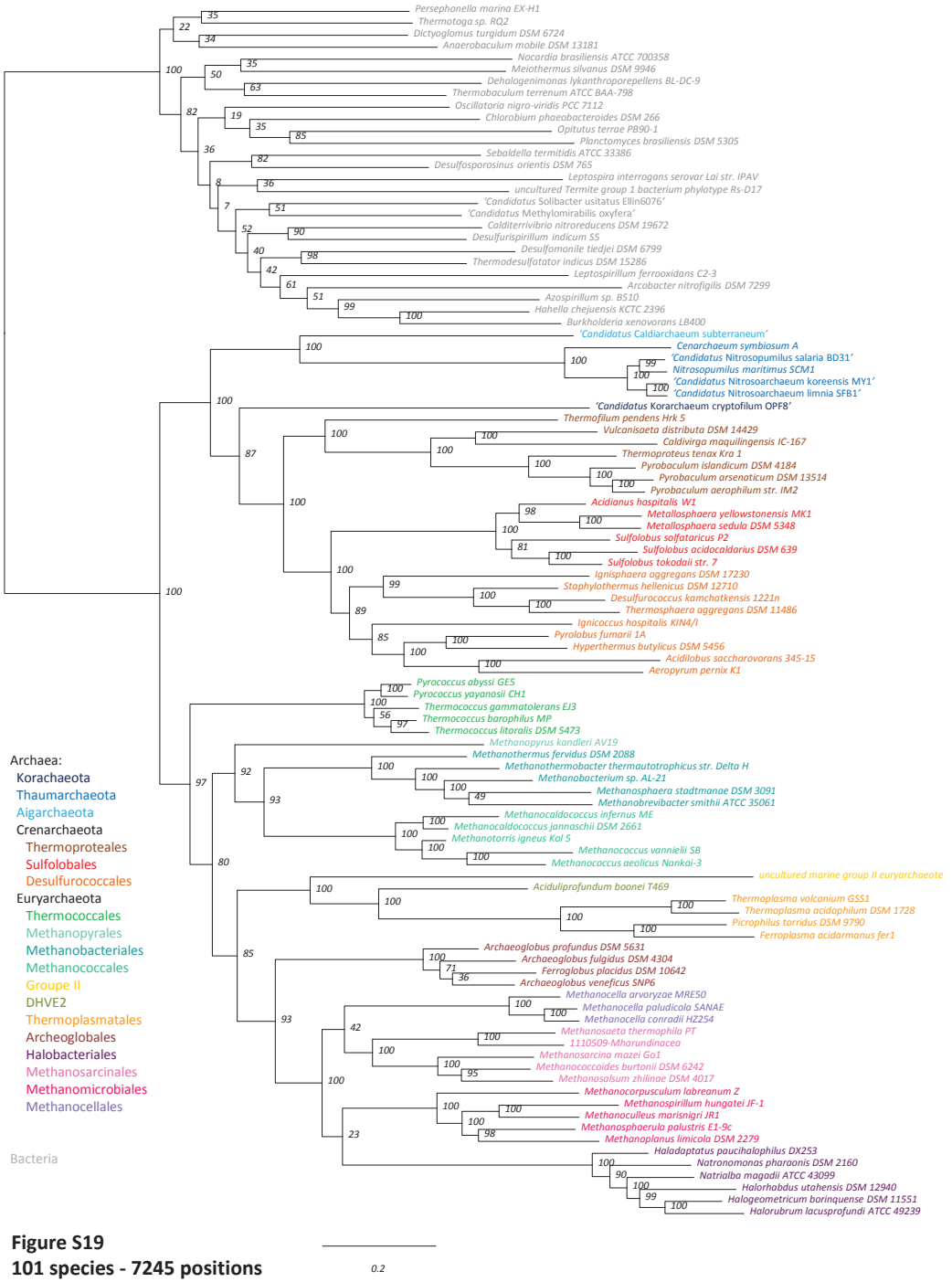


Figure S19
101 species - 7245 positions

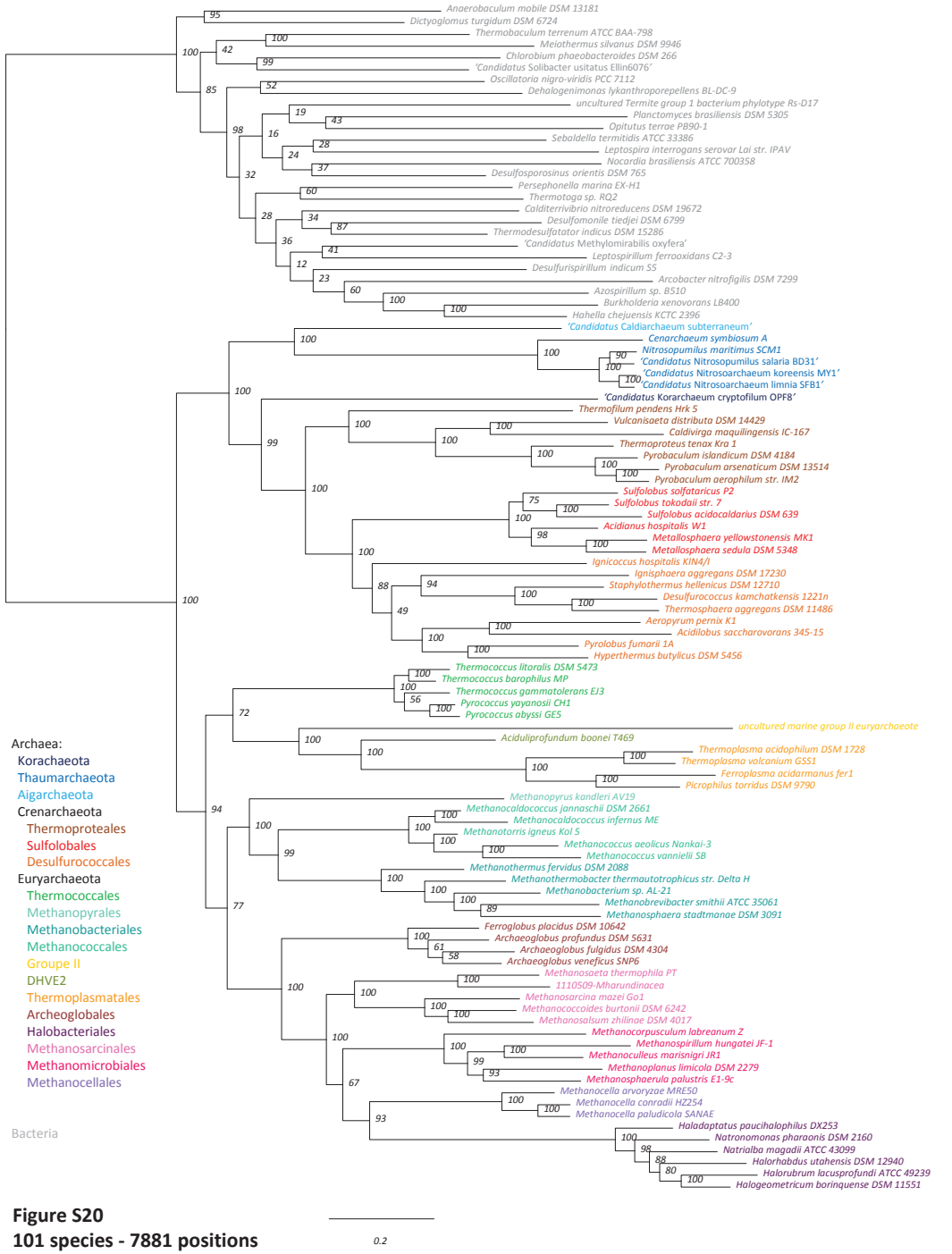


Figure S20
101 species - 7881 positions

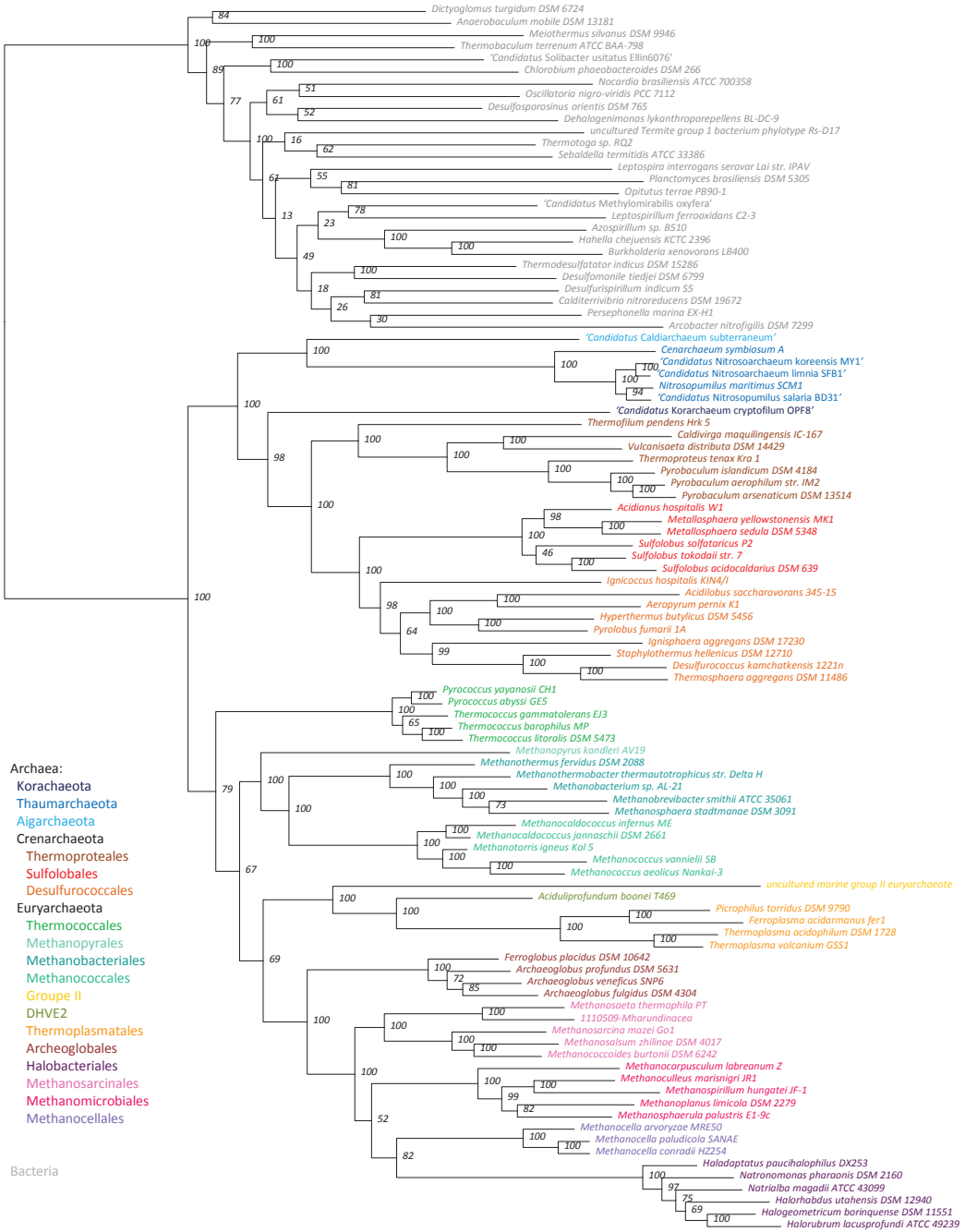


Figure S21
101 species - 8827 positions

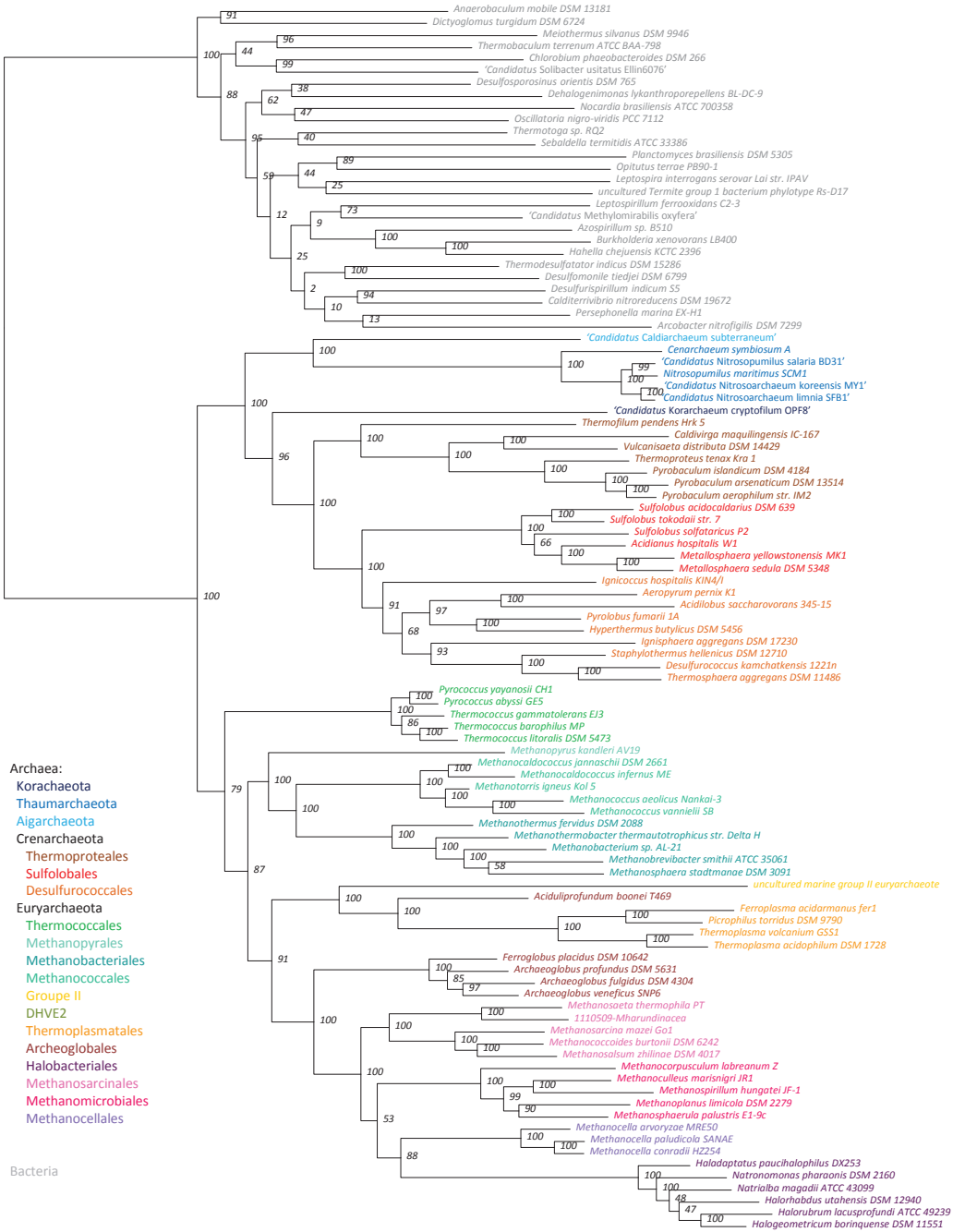


Figure S22
101 species - 9450 positions

Annexe 8 : Liste des 38 nouveaux marqueurs sélectionnés pour l'étude présentée dans le Chapitre 2.

La colonne « Référence thèse » correspond à la référence assignée aux différentes protéines dans cette thèse, utilisée dans les Annexes 9, 10 et 11.

Annotation	Référence thèse	Référence gi	Référence RefSeq
peptidase M50	M004R	161527799	YP_001581625.1
metalloendopeptidase glycoprotease family	M010R	161529041	YP_001582867.1
GTP-binding signal recognition particle	M028R	161528039	YP_001581865.1
cytidyltransferase-like protein	M029R	161528806	YP_001582632.1
beta-lactamase domain-containing protein	M032R	161529047	YP_001582873.1
glutamyl-tRNA(Gln) amidotransferase subunit E	M041R	161527611	YP_001581437.1
tRNA-guanine transglycosylase	M061R	161529044	YP_001582870.1
glutamyl-tRNA(Gln) amidotransferase B subunit	M064R	161528376	YP_001582202.1
exosome complex exonuclease 1	M066R	161527940	YP_001581766.1
MiaB-like tRNA modifying enzyme	M071R	161528769	YP_001582595.1
phenylalanyl-tRNA synthetase alpha subunit	M078R	161528997	YP_001582823.1
argininosuccinate synthase	M138R	161528794	YP_001582620.1
homoserine kinase	M142R	161527553	YP_001581379.1
homoserine dehydrogenase	M143R	161528204	YP_001582030.1
aspartate kinase	M146R	161529262	YP_001583088.1
aspartate carbamoyltransferase	M149R	161529194	YP_001583020.1
phosphoribosylformylglycinamidine cyclo-ligase	M151R	161528090	YP_001581916.1
uridylate kinase putative	M153R	161529213	YP_001583039.1
phosphoribosylformylglycinamidine synthase II	M156R	161528288	YP_001582114.1
adenylosuccinate lyase	M157R	161528963	YP_001582789.1
RdgB/HAM1 family non-canonical purine NTP pyrophosphatase	M158R	161529038	YP_001582864.1
glutamine amidotransferase class-II	M161R	161528289	YP_001582115.1
phosphoglycerate kinase	M165R	161528008	YP_001581834.1
phosphopantothenoylcysteine decarboxylase/phosphopantothenate--cysteine ligase	M170R	161529229	YP_001583055.1
pyridoxine biosynthesis protein	M171R	161528987	YP_001582813.1
UbiD family decarboxylase	M175R	161529105	YP_001582931.1
porphobilinogen deaminase	M176R	118576540	YP_876283.1
glutamate-1-semialdehyde-2 1-aminomutase	M179R	161527998	YP_001581824.1
GTP1/OBG protein	M195R	161529113	YP_001582939.1
LPPG:FO 2-phospho-L-lactate transferase	M208R	161528134	YP_001581960.1
prolyl-tRNA synthetase	MA17R	315425642	BAJ47301.1
histidyl-tRNA synthetase	MA19R	315425704	BAJ47360.1
signal recognition particle receptor	MA23R	315425933	BAJ47583.1
hydrogenase maturation protein HypF	MA32R	315426449	BAJ48087.1
hydrogenase expression formation protein HypD	MA33R	315426462	BAJ48095.1
5-nucleotidase SurE	MA42R	315426747	BAJ48371.1
conserved hypothetical protein	MA47R	315426923	BAJ48542.1
molybdenum cofactor biosynthesis protein C	MA53R	315427188	BAJ48802.1

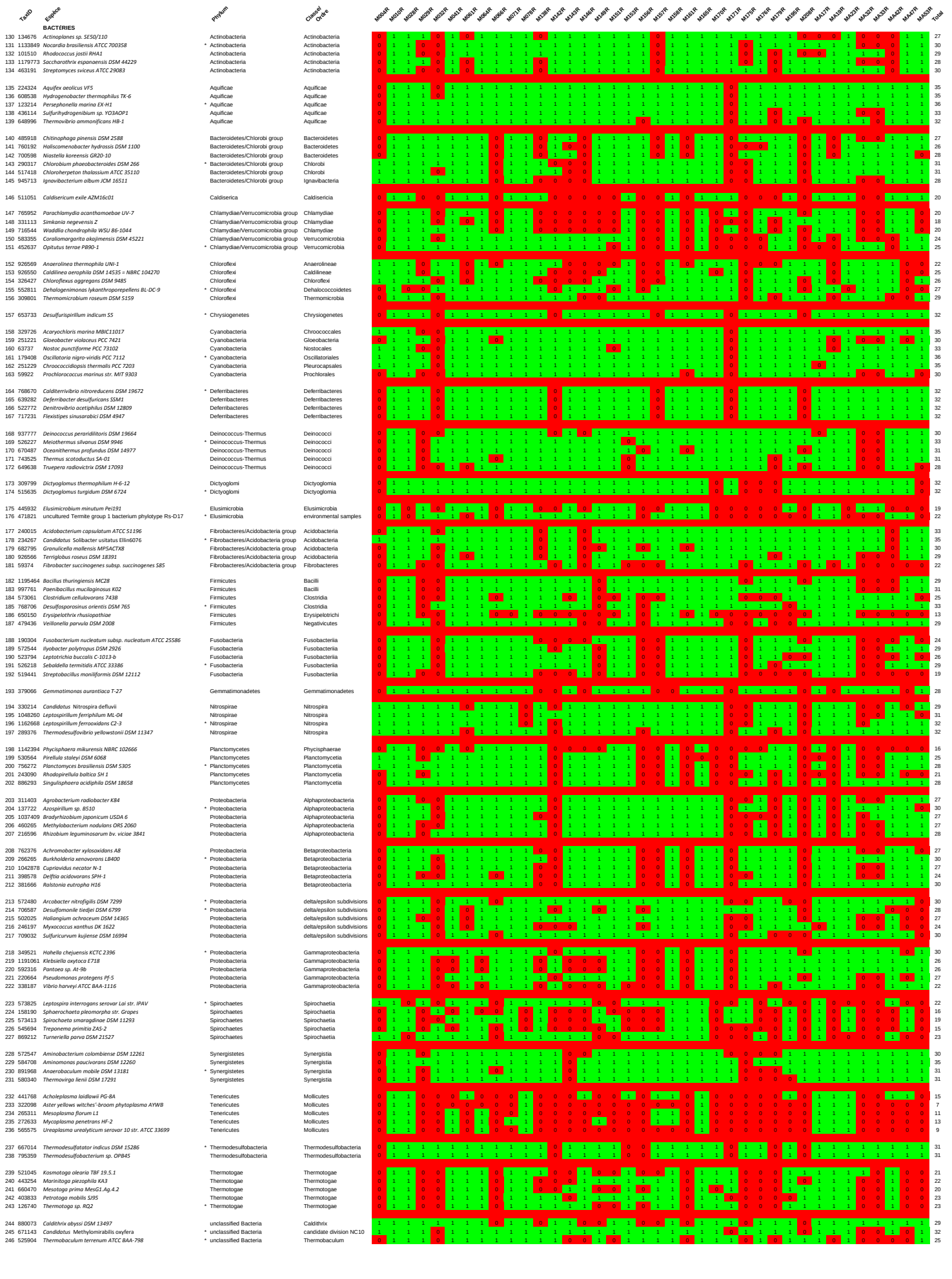
Annexe 9 : Répartition taxonomique des 38 nouveaux marqueurs chez les archées et chez les bactéries.

Chaque ligne correspond à une espèce, chaque colonne à un jeu de données, le croisement entre les deux correspond à la présence (« 1 », case verte), ou à l'absence (« 0 », case rouge) d'homologue de l'espèce dans le jeu de données. La dernière colonne donne le nombre de jeux de données dans lesquels l'espèce est présente.

La liste des archées et des bactéries correspond uniquement aux 129 et 117 génomes utilisés dans ce travail.

Les espèces utilisées pour la construction des jeux de données finaux sont notées par un astérisque.

	TaxID	Espèce	Phylum	Classe/ Ordre	M004R	M010R	M028R	M029R	M032R	M041R	M061R	M064R	M066R	M071R	M078R	M138R	M142R	M143R	M146R	M149R	M151R	M153R	M156R	M157R	M158R	M161R	M165R	M170R	M171R	M175R	M176R	M179R	M195R	M208R	MA17R	MA19R	MA23R	MA32R	MA33R	MA42R	MA47R	MA53R	Total
		ARCHÉES																																									
1	311458	*Candidatus* Caldiarchaeum subterraneum	* Aigarchaeota	Unclassified	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	0	0	1	0	0	1	0	1	0	0	1	0	0	1	1	0	1	1	1	1	1	1	1	27
2	414004	*Cenarchaeum symbiosum A*	* Thaumarchaeota	Cenarchaeales	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	0	0	0	0	0	31
3	436308	*Nitrosopumilus maritimus SCM1*	* Thaumarchaeota	Nitrosopumilales	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	0	0	0	0	0	31
4	886738	*Candidatus* Nitrosoarchaeum limnia SFB1	* Thaumarchaeota	Nitrosopumilales	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1	1	1	0	0	1	0	0	0	0	0	28
5	1001994	*Candidatus* Nitrosoarchaeum koreensis MY1	* Thaumarchaeota	Nitrosopumilales	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	0	0	0	0	0	31
6	859350	*Candidatus* Nitrosopumilus salaria BD31	* Thaumarchaeota	Nitrosopumilales	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	0	0	1	1	0	0	1	0	0	0	0	0	27
7	797209	*Haladaptatus paucihalophilus DX253*	* Euryarchaeota	Halobacteriales	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	0	1	1	0	0	1	0	1	31
8	795797	*Halalkalicoccus jeotgali B3*	Euryarchaeota	Halobacteriales	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	0	1	1	0	0	1	0	1	31
9	634497	*Haloarcula hispanica ATCC 33960*	Euryarchaeota	Halobacteriales	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	0	1	1	0	0	1	0	1	31
10	272569	*Haloarcula marismortui ATCC 43049*	Euryarchaeota	Halobacteriales	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	0	1	1	0	0	1	0	1	31
11	478009	*Halobacterium salinarum R1*	Euryarchaeota	Halobacteriales	0	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	0	1	1	0	0	1	0	1	30
12	309800	*Haloferax volcanii DS2*	Euryarchaeota	Halobacteriales	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	0	1	1	0	0	1	0	1	31
13	469382	*Halogeometricum borinquense DSM 11551*	* Euryarchaeota	Halobacteriales	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	0	1	1	0	0	1	0	1	31
14	485914	*Halomicrobium mukohataei DSM 12286*	Euryarchaeota	Halobacteriales	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	0	1	1	0	0	1	0	1	31
15	756883	*halophilic archaeon DL31*	Euryarchaeota	Halobacteriales	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	0	1	1	0	0	1	0	1	31
16	797210	*Halopiger xanaduensis SH-6*	Euryarchaeota	Halobacteriales	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	0	1	1	0	0	1	0	1	31
17	362976	*Haloquadratum walsbyi DSM 16790*	Euryarchaeota	Halobacteriales	1	1	1	1	1	1	0	1	0	0	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	0	1	1	0	0	1	0	1	29
18	1033806	*Halorhabdus tiamatea SARL4B*	Euryarchaeota	Halobacteriales	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0	0	33
19	519442	*Halorhabdus utahensis DSM 12940*	* Euryarchaeota	Halobacteriales	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0	1	34
20	416348	*Halorubrum lacusprofundi ATCC 49239*	* Euryarchaeota	Halobacteriales	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	0	1	1	0	0	1	0	1	31
21	543526	*Haloterrigena turkmenica DSM 5511*	Euryarchaeota	Halobacteriales	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	0	1	1	0	0	1	0	1	31
22	547559	*Natrialba magadii ATCC 43099*	* Euryarchaeota	Halobacteriales	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	0	1	1	0	0	1	0	1	31
23	797303	*Natrinema pellirubrum DSM 15624*	Euryarchaeota	Halobacteriales	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	0	1	1	0	0	1	0	1	31
24	797304	*Natronobacterium gregoryi SP2*	Euryarchaeota	Halobacteriales	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	0	1	1	1	1	1	0	1	33
25	348780	*Natronomonas pharaonis DSM 2160*	* Euryarchaeota	Halobacteriales	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	0	1	1	0	0	1	0	1	31
26	1072681	*Candidatus* Haloredivivus sp. G17	* Euryarchaeota	Nanohaloarchaea	0	1	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	6
27	889948	*Candidatus* Nanosalina sp. J07AB43	* Euryarchaeota	Nanohaloarchaea	0	1	1	1	0	1	0	1	1	0	1	1	0	0	1	0	0	1	0	0	1	0	1	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	14
28	889962	*Candidatus* Nanosalinarum sp. J07AB56	* Euryarchaeota	Nanohaloarchaea	0	1	1	1	0	1	0	1	1	0	1	1	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	12
29	410358	*Methanocorpusculum labreanum Z*	* Euryarchaeota	Methanomicrobiales	0	1	1	1	0	1	1	1	0	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	0	1	1	1	1	1	1	1	1	1	31
30	368407	*Methanoculleus marisnigri JR1*	* Euryarchaeota	Methanomicrobiales	0	1	1	1	0	1	1	1	0	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	33
31	882090	*Methanolinea tarda NOBI-1*	Euryarchaeota	Methanomicrobiales	0	1	1	1	0	1	1	1	0	1	1	1	0	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	32
32	937775	*Methanoplanus limicola DSM 2279*	* Euryarchaeota	Methanomicrobiales	0	1	1	1	0	1	1	1	0	1	1	1	0	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	33
33	679926	*Methanoplanus petrolearius DSM 11571*	Euryarchaeota	Methanomicrobiales	0	1	1	1	0	1	1	1	0	1	1	1	0	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	32
34	456442	*Methanoregula boonei 6A8*	Euryarchaeota	Methanomicrobiales	0	1	1	1	0	1	1	1	0	1	1	1	0	1	1	0	1	1	1	0	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	31
35	521011	*Methanosphaerula palustris E1-9c*	* Euryarchaeota	Methanomicrobiales	0	1	1	1	0	1	1	1	0	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	33
36	323259	*Methanospirillum hungatei JF-1*	* Euryarchaeota	Methanomicrobiales	0	1	1	1	0	1	1	1	0	1	1	1	0	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	32
37	259564	*Methanococcoides burtonii DSM 6242*	* Euryarchaeota	Methanosarcinales	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	35
38	644295	*Methanohalobium evestigatum Z-7303*	Euryarchaeota	Methanosarcinales	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	34
39	547558	*Methanohalophilus mahii DSM 5219*	Euryarchaeota	Methanosarcinales	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	34
40	990316	*Methanosaeta concilii GP6*	Euryarchaeota	Methanosarcinales	0	1	1	1	0	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	33
41	1110509	*Methanosaeta harundinacea 6Ac*	* Euryarchaeota	Methanosarcinales	0	1	1	1	0	1	1	1	1	1	1	1	0	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	32
42	349307	*Methanosaeta thermophila PT*	* Euryarchaeota	Methanosarcinales	0	1	1	1	0	1	1	1	1	1	1	1	0	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	0	1	1	1	1	0	0	1	1	1	31
43	679901	*Methanosalsum zhilinae DSM 4017*	* Euryarchaeota	Methanosarcinales	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	37
44	188937	*Methanosarcina acetivorans C2A*	Euryarchaeota	Methanosarcinales	1	1	1	1	1	1	1	1	0	1	1	1	0	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	35
45	269797	*Methanosarcina barkeri str. Fusaro*	Euryarchaeota	Methanosarcinales	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	36
46	192952	*Methanosarcina mazei Go1*	* Euryarchaeota	Methanosarcinales	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	36
47	351160	*Methanocella arvoryzae MRE50*	* Euryarchaeota	Methanocellales	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	37
48	1041930	*Methanocella conradii HZ254*	* Euryarchaeota	Methanocellales	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	37
49	304371	*Methanocella paludicola SANAE*	* Euryarchaeota	Methanocellales	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	37
50	224325	*Archaeoglobus fulgidus DSM 4304*	* Euryarchaeota	Archaeoglobales	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	36
51	572546	*Archaeoglobus profundus DSM 5631*	* Euryarchaeota	Archaeoglobales	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	35
52	693661	*Archaeoglobus veneficus SNP6*	* Euryarchaeota	Archaeoglobales	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	
53	589924	*Ferroglobus placidus DSM 10642*	* Euryarchaeota	Archaeoglobales	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	35
54	439481	*Aciduliprofundum boonei T469*	* Euryarchaeota	DHVE2	1	0	1	1	1	1	1	0	1	1	1	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	0	0	1	0	1	1	1	1	0	0	0	1	26
55	333146	*Ferroplasma acidarmanus fer1*	* Euryarchaeota	Thermoplasmatales	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	0	1	1	1	0	0	0	1	1	31
56	263820	*Picrophilus torridus DSM 9790*	* Euryarchaeota	Thermoplasmatales	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0	1	1	0	0	1	1	1	0	0	0	1	1	30
57	273075	*Thermoplasma acidophilum DSM 1728*	* Euryarchaeota	Thermoplasmatales	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0	1	1	0	0	1	1	1	0	0	0	1	1	30
58	273116	*Thermoplasma volcanium GSS1*	* Euryarchaeota	Thermoplasmatales	1	1	1	1	1	1	1	0	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	1	1	0	0	1	1	1	0	0	0	1	1	30
59	573064	*Methanocaldococcus fervens AG86*	Euryarchaeota	Methanococcales	0	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	35
60	573063	*Methanocaldococcus infernus ME*	* Euryarchaeota	Methanococcales	0	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	35
61	243232	*Methanocaldococcus jannaschii DSM 2661*	* Euryarchaeota	Methanococcales	0	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	36
62	644281	*Methanocaldococcus sp. FS406-22*	Euryarchaeota	Methanococcales	0	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	35
63	579137	*Methanocaldococcus vulcanius M7*	Euryarchaeota	Methanococcales	0	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	36
64	419665	*Methanococcus aeolicus Nankai-3*	* Euryarchaeota	Methanococcales	0	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	34
65	402880	*Methanococcus maripaludis C5*	Euryarchaeota	Methanococcales	0	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	34
66	406327	*Methanococcus vannielii SB*	* Euryarchaeota	Methanococcales	0	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	35
67	456320	*Methanococcus voltae A3*	Euryarchaeota	Methanococcales	0	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	34
68	647113	*Methanothermococcus okinawensis IH1*	Euryarchaeota	Methanococcales	0	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	35
69	647171	*Methanotorris formicicus Mc-S-70*	Euryarchaeota	Methanococcales	0	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	35
70	880724	*Methanotorris igneus Kol 5*	* Euryarchaeota	Methanococcales	0	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	36
71	868132	*Methanobacterium sp. AL-21*	* Euryarchaeota	Methanobacteriales	1	1	1	1	0	1	1	1	1	1	1	1	0	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	34
72	634498	*Methanobrevibacter ruminantium M1*	Euryarchaeota	Methanobacteriales	0	1	1	1	0	1	1	1	1	1	1	1	0	1	1	1	1	1	1	0	1	1	1	1	0	1	1	1	0	1	1	1	1	1	1	1	1	1	32
73	420247	*Methanobrevibacter smithii ATCC 35061*	* Euryarchaeota	Methanobacteriales	0	1	1	1	0	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	1	1	1	1	1	1	1	1	1	33
74	339860	*Methanosphaera stadtmanae DSM 3091*	* Euryarchaeota	Methanobacteriales	0	1	1	1	0	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	0	33
75	79929	*Methanothermobacter marburgensis str. Marburg*	Euryarchaeota	Methanobacteriales	0	1	1	1	0	1	1	1	1	1	1	1	0	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	33
76	187420	*Methanothermobacter thermautotrophicus str. Delta H*	* Euryarchaeota	Methanobacteriales	0	1	1	1	0	1	1	1	1	1	1	1	0	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	33
77	523846	*Methanothermus fervidus DSM 2088*	* Euryarchaeota	Methanobacteriales	0	1	1	1	0	1	1	1	1	1	1	1	0	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	33
78	190192	*Methanopyrus kandleri AV19*	* Euryarchaeota	Methanopyrales	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	36
79	272844	*Pyrococcus abyssi GE5*	* Euryarchaeota	Thermococcales	1	1	1	1	0	1	1	0	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	0	1	1	1	1	1	0	1	1	31
80	186497	*Pyrococcus furiosus DSM 3638*	Euryarchaeota	Thermococcales	1	1	1	1	0	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	0	1	1	1	1	1	0	1	1	32
81	70601	*Pyrococcus horikoshii OT3*	Euryarchaeota	Thermococcales	1	1	1	1	0	1	1	0	1	1	1	0	1	1	1	1	1	1	1	0	1	1	1	1	1	1	0	0	1	0	1	1	1	1	1	0	1	1	30
82	342949	*Pyrococcus sp. NA2*	Euryarchaeota	Thermococcales	1	1	1	1	0	1	1	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	0	0	1	0	1	1	1	1	1	0	1	1	31
83	529709	*Pyrococcus yayanosii CH1*	* Euryarchaeota	Thermococcales	1	1	1	1	0	1	1	0	1	1	1	0	0	1	0	1	1	1	1	0	1	1	1	1	1	1	0	0	1	0	1	1	1	1	1	0	1	1	28
84	391623	*Thermococcus barophilus MP*	* Euryarchaeota	Thermococcales	1	1	1	1	0	1	1	0	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	0	1	0	1	1	1	1	1	1	1	1	31
85	593117	*Thermococcus gammatolerans EJ3*	* Euryarchaeota	Thermococcales	1	1	1	1	0	1	1	0	1	1	1	0	1	1	1	1	1	1	1	0	1	1	1	1	1	1	0	0	1	0	1	1	1	1	1	1	1	1	31
86	69014	*Thermococcus kodakarensis KOD1*	Euryarchaeota	Thermococcales	1	1	1	1	0	1	1	0	1	1	1	0	1	1	1	1	1	1	1	0	1	1	1	1	1	1	0	0	1	0	1	1	1	1	1	1	1	1	31
87	523849	*Thermococcus litoralis DSM 5473*	* Euryarchaeota	Thermococcales	1	1	1	1	0	1	1	0	1	1	1	0	1	1	1	1	1	1	1	0	1	1	1	1	1	1	0	0	1	0	1	1	1	1	1	1	1	1	31
88	523850	*Thermococcus onnurineus NA1*	Euryarchaeota	Thermococcales	1	1	1	1	0	1	1	0	1	1	1	0	0	1	0	1	0	1	0	1	1	0	1	1	1	1	0	0	1	0	1	1	1	1	1	1	1	1	27
89	604354	*Thermococcus sibiricus MM 739*	Euryarchaeota	Thermococcales	1	1	1	1	0	1	1	0	1	1	1	0	1	1	1	1	1	1	1	0	1	1	1	1	0	1	0	0	1	0	1	1	1	1	1	1	1	1	30
90	246969	*Thermococcus sp. AM4*	Euryarchaeota	Thermococcales	1	1	1	1	0	1	1	0	1	1	1	0	0	1	0	1	1	1	1	0	1	1	1	1	1	1	0	0	1	0	1	1	1	1	1	1	1	1	29
91	425595	*Candidatus* Micrarchaeum acidiphilum ARMAN-2	* Euryarchaeota	ARMAN	0	1	1	0	1	1	0	0	1	1	1	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	10
92	662760	*Candidatus* Parvarchaeum acidiphilum ARMAN-4	* Euryarchaeota	ARMAN	0	1	0	0	0	1	0	1	1	0	1	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7
93	662762	*Candidatus* Parvarchaeum acidophilus ARMAN-5	* Euryarchaeota	ARMAN	0	1	0	0	0	1	0	1	1	0	1	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	8
94	228908	*Nanoarchaeum equitans Kin4-M*	* Euryarchaeota	Nanoarchaeota	0	1	0	0	0	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7
95	274854	uncultured marine group II euryarchaeote	* Euryarchaeota	Unclassified	0	1	1	1	1	1	1	0	1	0	1	0	0	0	1	1	1	1	1	0	1	1	1	1	1	1	0	0	1	0	1	0	1	0	0	0	0	1	23
96	666510	*Acidilobus saccharovorans 345-15*	* Crenarchaeota	Desulfurococcales	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	1	0	1	0	0	1	0	1	1	0	1	1	1	1	0	1	0	1	0	0	0	1	1	24
97	272557	*Aeropyrum pernix K1*	* Crenarchaeota	Desulfurococcales	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	0	1	0	0	1	0	1	1	1	1	1	1	1	0	1	0	1	0	0	0	1	1	28
98	768672	*Desulfurococcus fermentans DSM 16532*	Crenarchaeota	Desulfurococcales	1	1	1	1	1	1	1	0	1	1	1	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	0	0	1	0	0	1	1	1	1	0	1	1	28
99	490899	*Desulfurococcus kamchatkensis 1221n*	* Crenarchaeota	Desulfurococcales	1	1	1	1	1	1	1	0	1	1	1	0	0	0	0	1	1	1	1	0	1	1	1	1	1	1	0	0	1	0	0	1	1	1	1	0	1	1	27
100	765177	*Desulfurococcus mucosus DSM 2162*	Crenarchaeota	Desulfurococcales	1	1	1	1	1	1	1	0	1	1	1	0	0	0	0	1	0	1	0	0	1	0	1	1	1	1	0	0	1	0	0	1	1	1	1	0	1	1	24
101	415426	*Hyperthermus butylicus DSM 5456*	* Crenarchaeota	Desulfurococcales	0	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	1	0	0	1	0	1	1	1	1	1	1	1	0	1	1	1	1	1	0	1	1	30
102	453591	*Ignicoccus hospitalis KIN4/I*	* Crenarchaeota	Desulfurococcales	0	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0	1	1	33
103	583356	*Ignisphaera aggregans DSM 17230*	* Crenarchaeota	Desulfurococcales	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	1	1	1	0	0	1	0	1	1	1	1	1	0	1	1	31
104	694429	*Pyrolobus fumarii 1A*	* Crenarchaeota	Desulfurococcales	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1	0	1	1	33
105	591019	*Staphylothermus hellenicus DSM 12710*	* Crenarchaeota	Desulfurococcales	1	1	1	1	1	1	1	0	1	1	1	0	0	0	0	1	0	1	0	0	1	0	1	1	0	1	0	0	1	0	0	1	1	0	1	0	1	1	22
106	399550	*Staphylothermus marinus F1*	Crenarchaeota	Desulfurococcales	1	1	1	1	1	1	1	0	1	1	1	0	0	0	0	1	0	1	0	0	1	0	1	1	1	1	0	0	1	0	0	1	1	1	1	0	1	1	24
107	633148	*Thermosphaera aggregans DSM 11486*	* Crenarchaeota	Desulfurococcales	1	1	1	1	1	1	1	0	1	1	1	0	0	0	0	0	0	1	0	0	1	0	1	1	1	1	0	0	1	0	0	1	1	1	1	0	1	1	23
108	933801	*Acidianus hospitalis W1*	* Crenarchaeota	Sulfolobales	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	1	1	1	0	0	0	1	1	32
109	1006006	*Metallosphaera cuprina Ar-4*	Crenarchaeota	Sulfolobales	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	1	1	1	0	0	0	1	1	32
110	399549	*Metallosphaera sedula DSM 5348*	* Crenarchaeota	Sulfolobales	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	0	1	1	1	1	1	0	1	1	33
111	671065	*Metallosphaera yellowstonensis MK1*	* Crenarchaeota	Sulfolobales	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	0	1	1	1	1	1	0	1	1	33
112	330779	*Sulfolobus acidocaldarius DSM 639*	* Crenarchaeota	Sulfolobales	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	0	1	1	1	0	0	0	1	1	31
113	429572	*Sulfolobus islandicus L.S.2.15*	Crenarchaeota	Sulfolobales	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	0	1	1	0	0	0	0	1	1	30
114	273057	*Sulfolobus solfataricus P2*	* Crenarchaeota	Sulfolobales	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	1	1	1	0	0	0	1	1	32
115	273063	*Sulfolobus tokodaii str. 7*	* Crenarchaeota	Sulfolobales	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	0	1	1	1	0	0	0	1	1	31
116	397948	*Caldivirga maquilingensis IC-167*	* Crenarchaeota	Thermoproteales	0	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	0	1	1	1	0	0	1	1	1	32
117	178306	*Pyrobaculum aerophilum str. IM2*	* Crenarchaeota	Thermoproteales	0	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	0	1	1	1	0	0	1	1	1	32
118	340102	*Pyrobaculum arsenaticum DSM 13514*	* Crenarchaeota	Thermoproteales	0	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	34
119	410359	*Pyrobaculum calidifontis JCM 11548*	Crenarchaeota	Thermoproteales	0	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	35
120	384616	*Pyrobaculum islandicum DSM 4184*	* Crenarchaeota	Thermoproteales	0	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	34
121	698757	*Pyrobaculum oguniense TE7*	Crenarchaeota	Thermoproteales	0	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	34
122	1104324	*Pyrobaculum sp. 1860*	Crenarchaeota	Thermoproteales	0	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	0	1	1	1	33
123	368408	*Thermofilum pendens Hrk 5*	* Crenarchaeota	Thermoproteales	0	1	1	1	1	1	1	0	1	1	1	0	0	0	0	1	0	1	0	0	1	0	1	0	0	1	0	0	1	0	0	1	1	1	1	1	1	1	22
124	444157	*Thermoproteus neutrophilus V24Sta*	Crenarchaeota	Thermoproteales	0	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	34
125	768679	*Thermoproteus tenax Kra 1*	* Crenarchaeota	Thermoproteales	0	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	35
126	999630	*Thermoproteus uzoniensis 768-20*	Crenarchaeota	Thermoproteales	0	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	0	1	0	0	1	0	1	1	0	1	1	1	1	0	1	1	1	1	1	1	1	1	30
127	572478	*Vulcanisaeta distributa DSM 14429*	* Crenarchaeota	Thermoproteales	0	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	34
128	985053	*Vulcanisaeta moutnovskia 768-28*	Crenarchaeota	Thermoproteales	0	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	0	1	1	1	1	0	1	1	1	0	0	1	1	1	31
129	374847	*Candidatus* Korarchaeum cryptofilum OPF8	* Korachaeota		0	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	0	1	0	1	1	0	1	0	1	1	0	0	1	0	0	0	1	1	1	1	1	1	27


TaxID
Espèce
Phylum
Classe/ Ordre
M004R
M010R
M020R
M029R
M032R
M041R
M061R
M064R
M066R
M071R
M072R
M138R
M142R
M145R
M146R
M148R
M151R
M153R
M156R
M157R
M158R
M161R
M165R
M170R
M171R
M175R
M176R
M179R
M185R
M208R
MA17R
MA18R
MA23R
MA3R
MA3SR
MA4SR
MA47R
MA5SR
Total
BACTÉRIES
130 134676 Actinoplanes sp. SE50/110 Actinobacteria Actinobacteria 27
131 1133849 Nocardia brasiliensis ATCC 700358 * Actinobacteria Actinobacteria 30
132 101510 Rhodococcus jostii RHA1 Actinobacteria Actinobacteria 29
133 1179773 Saccharothrix espanaensis DSM 44229 Actinobacteria Actinobacteria 28
134 463191 Streptomyces sviceus ATCC 29083 Actinobacteria Actinobacteria 30
135 224324 Aquifex aeolicus VF5 Aquificae Aquificae 35
136 608538 Hydrogenobacter thermophilus TK-6 Aquificae Aquificae 35
137 123214 Persephonella marina EX-H1 * Aquificae Aquificae 36
138 436114 Sulfurihydrogenibium sp. YO3AOP1 Aquificae Aquificae 33
139 648996 Thermovibrio ammonificans HB-1 Aquificae Aquificae 32
140 485918 Chitinophaga pinensis DSM 2588 Bacteroidetes/Chlorobi group Bacteroidetes 27
141 760192 Haliscomenobacter hydrossis DSM 1100 Bacteroidetes/Chlorobi group Bacteroidetes 26
142 700598 Niastella koreensis GR20-10 Bacteroidetes/Chlorobi group Bacteroidetes 28
143 290317 Chlorobium phaeobacteroides DSM 266 * Bacteroidetes/Chlorobi group Chlorobi 31
144 517418 Chloroherpeton thalassium ATCC 35110 Bacteroidetes/Chlorobi group Chlorobi 31
145 945713 Ignavibacterium album JCM 16511 Bacteroidetes/Chlorobi group Ignavibacteria 28
146 511051 Caldisericum exile AZM16c01 Caldiserica Caldisericia 20
147 765952 Parachlamydia acanthamoebae UV-7 Chlamydiae/Verrucomicrobia group Chlamydiae 20
148 331113 Simkania negevensis Z Chlamydiae/Verrucomicrobia group Chlamydiae 18
149 716544 Waddlia chondrophila WSU 86-1044 Chlamydiae/Verrucomicrobia group Chlamydiae 20
150 583355 Coraliomargarita akajimensis DSM 45221 Chlamydiae/Verrucomicrobia group Verrucomicrobia 24
151 452637 Opitutus terrae PB90-1 * Chlamydiae/Verrucomicrobia group Verrucomicrobia 25
152 926569 Anaerolinea thermophila UNI-1 Chloroflexi Anaerolineae 22
153 926550 Caldilinea aerophila DSM 14535 = NBRC 104270 Chloroflexi Caldilineae 25
154 326427 Chloroflexus aggregans DSM 9485 Chloroflexi Chloroflexi 26
155 552811 Dehalogenimonas lykanthroporepellens BL-DC-9 * Chloroflexi Dehalococcoidetes 27
156 309801 Thermomicrobium roseum DSM 5159 Chloroflexi Thermomicrobia 29
157 653733 Desulfurispirillum indicum S5 * Chrysiogenetes Chrysiogenetes 32
158 329726 Acaryochloris marina MBIC11017 Cyanobacteria Chroococcales 35
159 251221 Gloeobacter violaceus PCC 7421 Cyanobacteria Gloeobacteria 30
160 63737 Nostoc punctiforme PCC 73102 Cyanobacteria Nostocales 33
161 179408 Oscillatoria nigro-viridis PCC 7112 * Cyanobacteria Oscillatoriales 36
162 251229 Chroococcidiopsis thermalis PCC 7203 Cyanobacteria Pleurocapsales 35
163 59922 Prochlorococcus marinus str. MIT 9303 Cyanobacteria Prochlorales 30
164 768670 Calditerrivibrio nitroreducens DSM 19672 * Deferribacteres Deferribacteres 32
165 639282 Deferribacter desulfuricans SSM1 Deferribacteres Deferribacteres 32
166 522772 Denitrovibrio acetiphilus DSM 12809 Deferribacteres Deferribacteres 32
167 717231 Flexistipes sinusarabici DSM 4947 Deferribacteres Deferribacteres 32
168 937777 Deinococcus peraridilitoris DSM 19664 Deinococcus-Thermus Deinococci 30
169 526227 Meiothermus silvanus DSM 9946 * Deinococcus-Thermus Deinococci 33
170 670487 Oceanithermus profundus DSM 14977 Deinococcus-Thermus Deinococci 31
171 743525 Thermus scotoductus SA-01 Deinococcus-Thermus Deinococci 31
172 649638 Truepera radiovictrix DSM 17093 Deinococcus-Thermus Deinococci 28
173 309799 Dictyoglomus thermophilum H-6-12 Dictyoglomi Dictyoglomia 32
174 515635 Dictyoglomus turgidum DSM 6724 * Dictyoglomi Dictyoglomia 32
175 445932 Elusimicrobium minutum Pei191 Elusimicrobia Elusimicrobia 19
176 471821 uncultured Termite group 1 bacterium phylotype Rs-D17 * Elusimicrobia environmental samples 22
177 240015 Acidobacterium capsulatum ATCC 51196 Fibrobacteres/Acidobacteria group Acidobacteria 33
178 234267 Candidatus Solibacter usitatus Ellin6076 * Fibrobacteres/Acidobacteria group Acidobacteria 35
179 682795 Granulicella mallensis MP5ACTX8 Fibrobacteres/Acidobacteria group Acidobacteria 30
180 926566 Terriglobus roseus DSM 18391 Fibrobacteres/Acidobacteria group Acidobacteria 29
181 59374 Fibrobacter succinogenes subsp. succinogenes S85 Fibrobacteres/Acidobacteria group Fibrobacteres 22
182 1195464 Bacillus thuringiensis MC28 Firmicutes Bacilli 29
183 997761 Paenibacillus mucilaginosus K02 Firmicutes Bacilli 31
184 573061 Clostridium cellulovorans 743B Firmicutes Clostridia 25
185 768706 Desulfosporosinus orientis DSM 765 * Firmicutes Clostridia 33
186 650150 Erysipelothrix rhusiopathiae Firmicutes Erysipelotrichi 13
187 479436 Veillonella parvula DSM 2008 Firmicutes Negativicutes 29
188 190304 Fusobacterium nucleatum subsp. nucleatum ATCC 25586 Fusobacteria Fusobacteriia 24
189 572544 Ilyobacter polytropus DSM 2926 Fusobacteria Fusobacteriia 29
190 523794 Leptotrichia buccalis C-1013-b Fusobacteria Fusobacteriia 26
191 526218 Sebaldella termitidis ATCC 33386 * Fusobacteria Fusobacteriia 29
192 519441 Streptobacillus moniliformis DSM 12112 Fusobacteria Fusobacteriia 19
193 379066 Gemmatimonas aurantiaca T-27 Gemmatimonadetes Gemmatimonadetes 28
194 330214 Candidatus Nitrospira defluvii Nitrospirae Nitrospira 29
195 1048260 Leptospirillum ferriphilum ML-04 Nitrospirae Nitrospira 31
196 1162668 Leptospirillum ferrooxidans C2-3 * Nitrospirae Nitrospira 32
197 289376 Thermodesulfovibrio yellowstonii DSM 11347 Nitrospirae Nitrospira 32
198 1142394 Physisphaera mikurensis NBRC 102666 Planctomycetes Physisphaerae 16
199 530564 Pirellula staleyi DSM 6068 Planctomycetes Planctomycetia 25
200 756272 Planctomyces brasiliensis DSM 5305 * Planctomycetes Planctomycetia 28
201 243090 Rhodopirellula baltica SH 1 Planctomycetes Planctomycetia 21
202 886293 Singulisphaera acidiphila DSM 18658 Planctomycetes Planctomycetia 28
203 311403 Agrobacterium radiobacter K84 Proteobacteria Alphaproteobacteria 27
204 137722 Azospirillum sp. B510 * Proteobacteria Alphaproteobacteria 30
205 1037409 Bradyrhizobium japonicum USDA 6 Proteobacteria Alphaproteobacteria 27
206 460265 Methylobacterium nodulans ORS 2060 Proteobacteria Alphaproteobacteria 27
207 216596 Rhizobium leguminosarum bv. viciae 3841 Proteobacteria Alphaproteobacteria 28
208 762376 Achromobacter xylosoxidans A8 Proteobacteria Betaproteobacteria 27
209 266265 Burkholderia xenovorans LB400 * Proteobacteria Betaproteobacteria 30
210 1042878 Cupriavidus necator N-1 Proteobacteria Betaproteobacteria 27
211 398578 Delftia acidovorans SPH-1 Proteobacteria Betaproteobacteria 24
212 381666 Ralstonia eutropha H16 Proteobacteria Betaproteobacteria 30
213 572480 Arcobacter nitrofigilis DSM 7299 * Proteobacteria delta/epsilon subdivisions 30
214 706587 Desulfomonile tiedjei DSM 6799 * Proteobacteria delta/epsilon subdivisions 28
215 502025 Haliangium ochraceum DSM 14365 Proteobacteria delta/epsilon subdivisions 27
216 246197 Myxococcus xanthus DK 1622 Proteobacteria delta/epsilon subdivisions 24
217 709032 Sulfuricurvum kujiense DSM 16994 Proteobacteria delta/epsilon subdivisions 30
218 349521 Hahella chejuensis KCTC 2396 * Proteobacteria Gammaproteobacteria 30
219 1191061 Klebsiella oxytoca E718 Proteobacteria Gammaproteobacteria 26
220 592316 Pantoea sp. At-9b Proteobacteria Gammaproteobacteria 26
221 220664 Pseudomonas protegens Pf-5 Proteobacteria Gammaproteobacteria 27
222 338187 Vibrio harveyi ATCC BAA-1116 Proteobacteria Gammaproteobacteria 22
223 573825 Leptospira interrogans serovar Lai str. IPAV * Spirochaetes Spirochaetia 22
224 158190 Sphaerochaeta pleomorpha str. Grapes Spirochaetes Spirochaetia 16
225 573413 Spirochaeta smaragdinae DSM 11293 Spirochaetes Spirochaetia 19
226 545694 Treponema primitia ZAS-2 Spirochaetes Spirochaetia 15
227 869212 Turneriella parva DSM 21527 Spirochaetes Spirochaetia 23
228 572547 Aminobacterium colombiense DSM 12261 Synergistetes Synergistia 30
229 584708 Aminomonas paucivorans DSM 12260 Synergistetes Synergistia 35
230 891968 Anaerobaculum mobile DSM 13181 * Synergistetes Synergistia 31
231 580340 Thermovirga lienii DSM 17291 Synergistetes Synergistia 31
232 441768 Acholeplasma laidlawii PG-8A Tenericutes Mollicutes 15
233 322098 Aster yellows witches'-broom phytoplasma AYWB Tenericutes Mollicutes 7
234 265311 Mesoplasma florum L1 Tenericutes Mollicutes 11
235 272633 Mycoplasma penetrans HF-2 Tenericutes Mollicutes 13
236 565575 Ureaplasma urealyticum serovar 10 str. ATCC 33699 Tenericutes Mollicutes 9
237 667014 Thermodesulfatator indicus DSM 15286 * Thermodesulfobacteria Thermodesulfobacteria 31
238 795359 Thermodesulfobacterium sp. OPB45 Thermodesulfobacteria Thermodesulfobacteria 31
239 521045 Kosmotoga olearia TBF 19.5.1 Thermotogae Thermotogae 21
240 443254 Marinitoga piezophila KA3 Thermotogae Thermotogae 20
241 660470 Mesotoga prima MesG1.Ag.4.2 Thermotogae Thermotogae 20
242 403833 Petrotoga mobilis SJ95 Thermotogae Thermotogae 23
243 126740 Thermotoga sp. RQ2 * Thermotogae Thermotogae 23
244 880073 Caldithrix abyssi DSM 13497 unclassified Bacteria Caldithrix 29
245 671143 Candidatus Methylomirabilis oxyfera * unclassified Bacteria candidate division NC10 32
246 525904 Thermobaculum terrenum ATCC BAA-798 * unclassified Bacteria Thermobaculum 25

Annexe 10 : Liste des matrices utilisées pour l'analyse en désaturation par sélection de sites pour l'étude exposée dans le Chapitre 2.

Matrice	Nombre de positions
S_5	939
S_{10}	2032
S_{13}	2905
S_{16}	3803
S_{19}	4727
S_{23}	5860
S_{27}	6939
S_{32}	7978
S_{40}	8980
Totale	9450

Annexe 11 : Liste des matrices utilisées pour l'analyse en désaturation par sélection de gènes pour l'étude exposée dans le Chapitre 2.

Pour chaque marqueur, la matrice indiquée est la plus grande dans laquelle ce marqueur a été inclus. Par exemple, les marqueurs de la matrice DTR21 sont aussi dans les matrices DTR07 et DTR14.
Le nombre de distance calculées correspond au nombre de couples d'espèces pour lesquels une distance a été mesurée, et qui sert de diviseur pour le calcul de la distance moyenne (maximum = 10).

Marqueur	Distance moyenne	Nombre de distances calculées	Matrice	Nombre de positions
s4	0,96050367	22	DTR07	870
s12	0,9814734	22		
s19	1,02444372	22		
M151R	1,10259481	22		
M171R	1,11171562	22		
MA42R	1,12795002	15		
M064R	1,23301883	15		
M158R	1,37887247	22	DTR14	1926
M004R	1,40398118	4		
MA23R	1,43100849	22		
l5	1,4362529	22		
l6	1,44130399	22		
M179R	1,44385348	18		
MA33R	1,44517545	22		
MA32R	1,46667928	22	DTR21	3172
s11	1,47497957	22		
M138R	1,51568416	18		
M176R	1,52872207	18		
l23	1,54398756	22		
M028R	1,55431829	22		
s15	1,56131733	22		
M143R	1,57662369	22	DTR28	4175
M157R	1,58667373	22		
M208R	1,62469153	6		
s8	1,63179175	22		
s3	1,64072369	22		
M195R	1,64326337	22		
M153R	1,66182908	22		
s17	1,67199516	22	DTR35	4902
s7	1,67240291	22		
M078R	1,68421457	22		
l11	1,69115896	22		
l24	1,70010601	22		
l3	1,71632184	22		
M066R	1,73363966	15		
l4	1,73706915	22	DTR42	5633
M170R	1,75706427	22		
l10	1,76798183	22		
l13	1,8150769	22		
l29	1,81841707	22		
l2	1,8385409	22		
M165R	1,86667573	22		

M156R	1,93674863	22	DTR49	7245
s5	1,93978399	22		
MA53R	1,94381957	22		
M142R	1,94681564	15		
M029R	1,94934403	22		
MA17R	1,95739969	15		
M175R	1,95802418	15		
s13	1,96731478	22	DTR56	7881
M071R	1,97633225	22		
MA19R	1,98073354	22		
M032R	1,98385148	12		
M146R	1,98393917	22		
M041R	1,99327988	22		
M149R	2,06404805	22		
s9	2,06570095	22	DTR63	8827
M161R	2,11260774	22		
l1	2,1819755	22		
l10e	2,19431942	22		
M061R	2,2061242	22		
l14	2,22846266	22		
MA47R	2,30049458	22		
l22	2,32519427	22	Totale	9450
M010R	2,35635182	22		
s10	2,4054802	22		
s2	2,56797706	22		
l18	2,5921461	22		
l30	2,88936107	22		
l15	3,06879313	22		

Annexe 12 : Matériels supplémentaires de l'Article 3 (Chapitre 3).

Additional files

Additional file 1: Table showing the taxonomic distribution of DnaJ-Fer, DnaK, DnaJ and GrpE proteins in Archaea. Numbers correspond to accession numbers in the NCBI Genpep database in the 92 complete genomes available in July 2011 and that of, 'Ca. Nitrosoarchaeum koreensis', a thaumarchaeotal genome available more recently. The two divergent sequences of DnaK and GrpE found in Methanococcus vannielii SB are underlined.

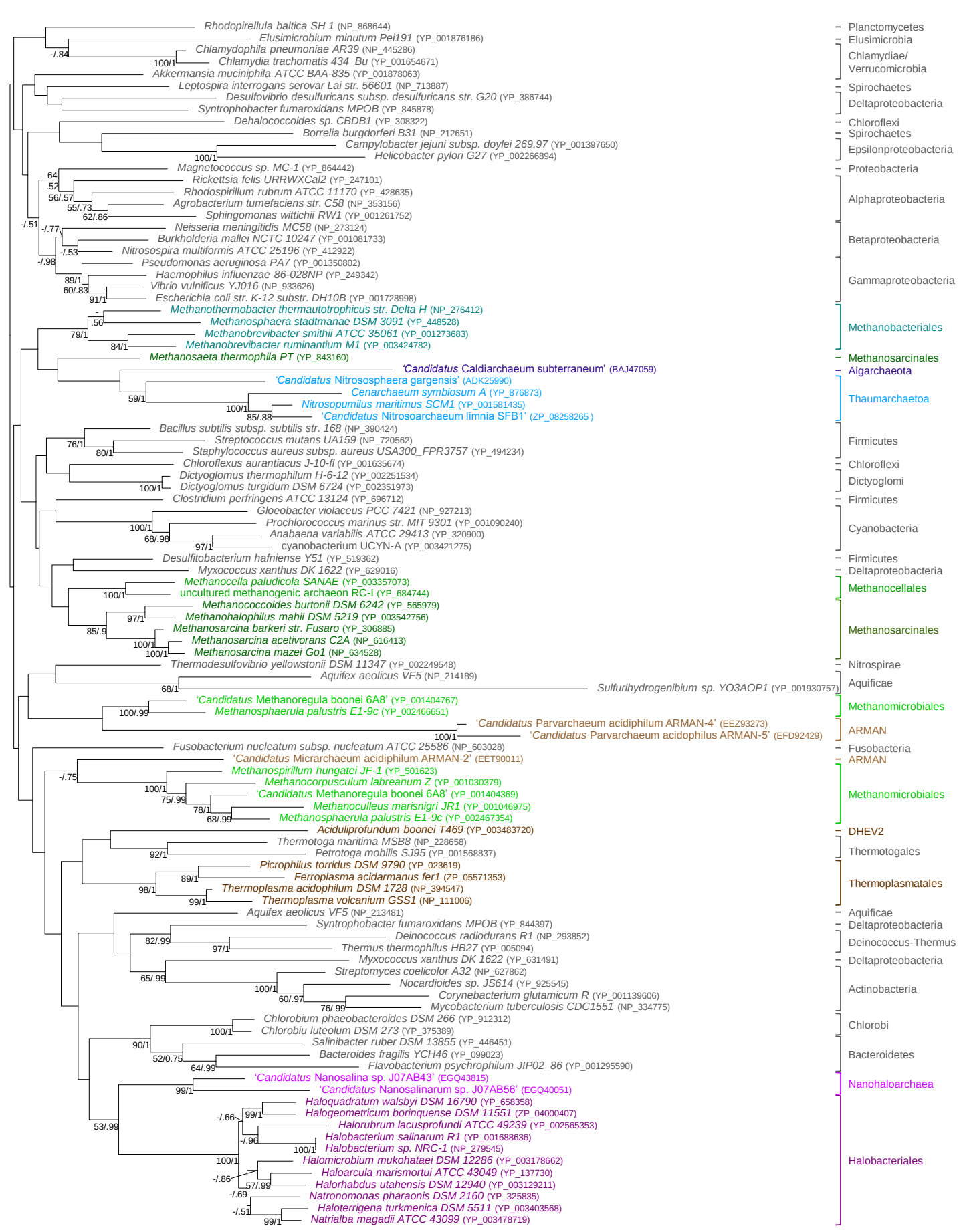
Additional file 2: Unrooted ML tree of the DnaK protein (136 sequences and 444 positions) inferred with TreeFinder and the LG + Γ4 model. Numbers at nodes represent bootstrap values and Bayesian posterior probabilities computed with TreeFinder and MrBayes, respectively (only values >50% and 0.5 are shown, dashes indicate the corresponding support is inferior to the threshold, whereas when both supports are inferior to the thresholds no support values are indicated). Archaeal sequences are shown with colours according to their taxonomic classification. The scale bar represents the average number of substitutions per site.

Additional file 3: Unrooted ML tree of the DnaJ protein (102 sequences and 227 positions) inferred with TreeFinder and the LG + Γ4. Numbers at nodes represent bootstrap values and Bayesian posterior probabilities computed with TreeFinder and MrBayes, respectively (only values >50% and 0.5 are shown, dashes indicate that the corresponding support is inferior to the threshold, whereas when both supports are inferior to the thresholds no support values are indicated). Archaeal sequences are shown with colours according to their taxonomic classification. The scale bar represents the average number of substitutions per site.

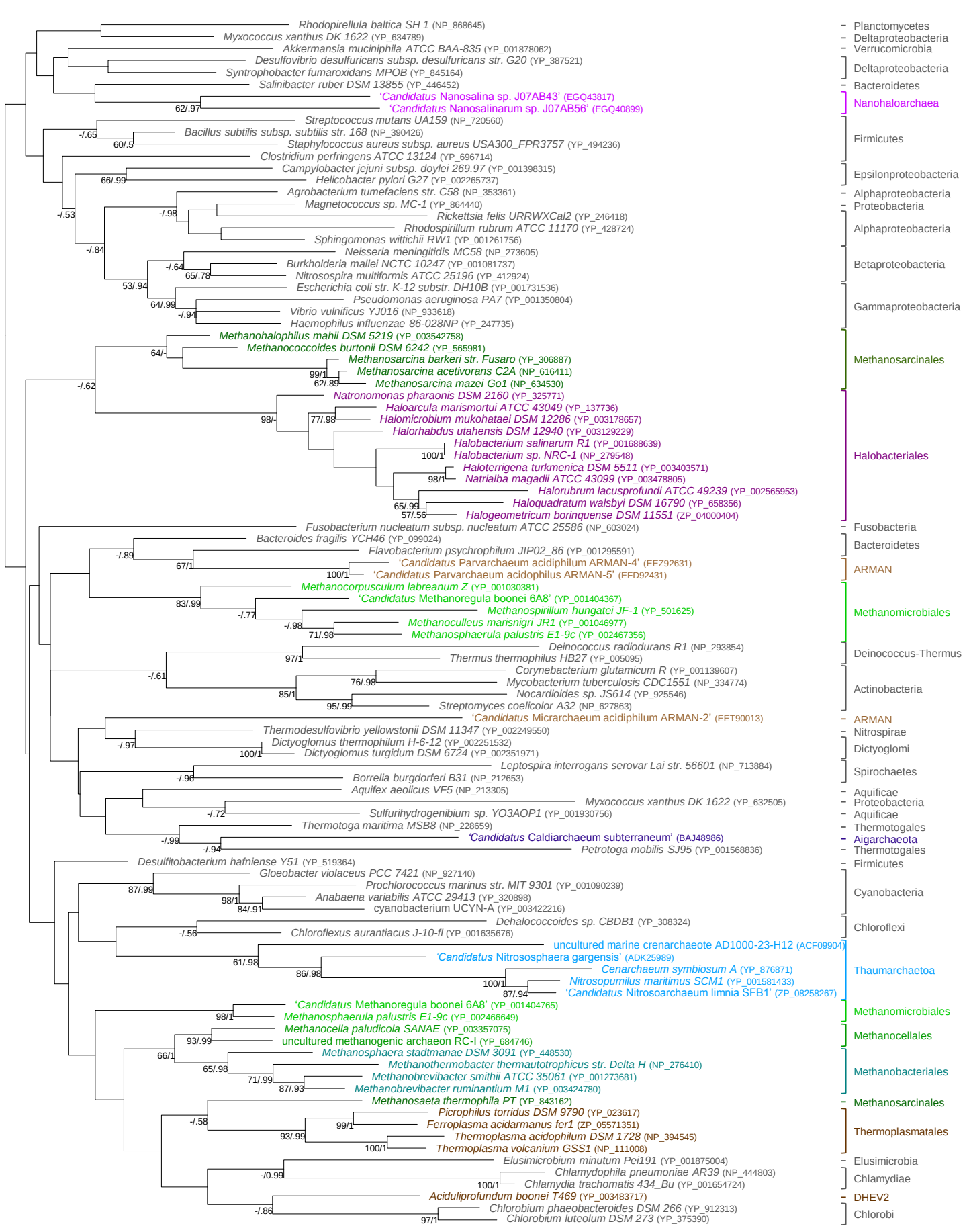
Additional file 4: Unrooted ML tree of the GrpE protein (101 sequences and 105 positions) inferred with TreeFinder and the LG + Γ4. Numbers at nodes represent bootstrap values and Bayesian posterior probabilities computed with TreeFinder and MrBayes, respectively (only values >50% and 0.5 are shown, dashes indicate the corresponding value is inferior to the threshold, whereas when both supports are inferior to the thresholds no support values are indicated). Archaeal sequences are shown with colours according to their taxonomic classification. The scale bar represents the average number of substitutions per site.

TaxID	Organism	Group		DnaJ-Fer	GrpE	DnaK	DnaJ	Cluster
414004	*Cenarchaeum symbiosum A*	Group I.1a	Thaumarchaetoa	YP_875357	YP_876871	YP_876872	YP_876873	EKJ
436308	*Nitrosopumilus maritimus SCM1*	Group I.1a	Thaumarchaetoa	YP_001582358	YP_001581433	YP_001581434	YP_001581435	EKJ
886738	'*Candidatus* Nitrosoarchaeum limnia SFB1'	Group I.1a	Thaumarchaeota	ZP_08256939	ZP_08258267	ZP_08258266	ZP_08258265	EKJ
1088740	'*Candidatus* Nitrosoarchaeum koreensis'	Group I.1a	Thaumarchaeota	ZP_08668202	ZP_08667092	ZP_08667093	ZP_08667094	EKJ
926571	*Nitrososphaera viennensis EN76*	Group I.1b	Thaumarchaeota	(unpublished)	NA	NA	NA	NA
497727	'*Candidatus* Nitrososphaera gargensis'	Group I.1b	Thaumarchaeota	(unpublished)	ADK25989	ADK25988	ADK25990.1	EKJ
311458	'*Candidatus* Caldiarchaeum subterraneum'		'Aigarchaeota'		BAJ47057	BAJ47058	BAJ47059	EKJ
64091	*Halobacterium sp. NRC-1*	Halobacteriales	Euryarchaeota		NP_279548	NP_279546	NP_279545	E-KJ
272569	*Haloarcula marismortui ATCC 43049*	Halobacteriales	Euryarchaeota		YP_137736	YP_137735	YP_137730	EK----J
348780	*Natronomonas pharaonis DSM 2160*	Halobacteriales	Euryarchaeota		YP_325771	YP_325772	YP_325835	EK / J
362976	*Haloquadratum walsbyi DSM 16790*	Halobacteriales	Euryarchaeota		YP_658356	YP_658357	YP_658358	EKJ
416348	*Halorubrum lacusprofundi ATCC 49239*	Halobacteriales	Euryarchaeota		YP_002565953	YP_002565354	YP_002565353	E / KJ
469382	*Halogeometricum borinquense DSM 11551*	Halobacteriales	Euryarchaeota		ZP_04000404	ZP_04000406	ZP_04000407	E-KJ
478009	*Halobacterium salinarum R1*	Halobacteriales	Euryarchaeota		YP_001688639	YP_001688637	YP_001688636	E-KJ
485914	*Halomicrobium mukohataei DSM 12286*	Halobacteriales	Euryarchaeota		YP_003178657	YP_003178658	YP_003178662	EK---J
519442	*Halorhabdus utahensis DSM 12940*	Halobacteriales	Euryarchaeota		YP_003129229	YP_003129230	YP_003129211	EK / J
543526	*Haloterrigena turkmenica DSM 5511*	Halobacteriales	Euryarchaeota		YP_003403571	YP_003403570	YP_003403568	EK-J
547559	*Natrialba magadii ATCC 43099*	Halobacteriales	Euryarchaeota		YP_003478805	YP_003478720	YP_003478719	E / KJ
889948	'*Candidatus* Nanosalina sp. J07AB43'	Nanohaloarchaea	Euryarchaeota		EGQ43817	EGQ43816	EGQ43815	EKJ
889962	'*Candidatus* Nanosalinarum sp. J07AB56'	Nanohaloarchaea	Euryarchaeota		EGQ40899	EGQ40898	EGQ40051	EK / J
323259	*Methanospirillum hungatei JF-1*	Methanomicrobiales	Euryarchaeota		YP_501625	YP_501624	YP_501623	EKJ
368407	*Methanoculleus marisnigri JR1*	Methanomicrobiales	Euryarchaeota		YP_001046977	YP_001046976	YP_001046975	EKJ
410358	*Methanocorpusculum labreanum Z*	Methanomicrobiales	Euryarchaeota		YP_001030381	YP_001030380	YP_001030379	EKJ
456442	*'Candidatus Methanoregula boonei 6A8'*	Methanomicrobiales	Euryarchaeota		YP_001404367	YP_001404368	YP_001404369	EKJ
456442	*'Candidatus Methanoregula boonei 6A8'*	Methanomicrobiales	Euryarchaeota		YP_001404765	YP_001404766	YP_001404767	EKJ
521011	*Methanosphaerula palustris E1-9c*	Methanomicrobiales	Euryarchaeota		YP_002466649	YP_002466650	YP_002466651	EKJ
521011	*Methanosphaerula palustris E1-9c*	Methanomicrobiales	Euryarchaeota		YP_002467356	YP_002467355	YP_002467354	EKJ
304371	*Methanocella paludicola SANAE*	Methanocellales	Euryarchaeota		YP_003357075	YP_003357074	YP_003357073	EKJ
351160	*uncultured methanogenic archaeon RC-I*	Methanocellales	Euryarchaeota		YP_684746	YP_684745	YP_684744	EKJ
188937	*Methanosarcina acetivorans C2A*	Methanosarcinales	Euryarchaeota		NP_616411	NP_616412	NP_616413	EKJ
192952	*Methanosarcina mazei Go1 (NP_634529)*	Methanosarcinales	Euryarchaeota		NP_634530	NP_634529	NP_634528	EKJ
259564	*Methanococcoides burtonii DSM 6242*	Methanosarcinales	Euryarchaeota		YP_565981	YP_565980	YP_565979	EKJ
269797	*Methanosarcina barkeri str. Fusaro*	Methanosarcinales	Euryarchaeota		YP_306887	YP_306886	YP_306885	EKJ
349307	*Methanosaeta thermophila PT*	Methanosarcinales	Euryarchaeota		YP_843162	YP_843161	YP_843160	EKJ
547558	*Methanohalophilus mahii DSM 5219*	Methanosarcinales	Euryarchaeota		YP_003542758	YP_003542757	YP_003542756	EKJ
224325	*Archaeoglobus fulgidus DSM 4304*	Archaeoglobales	Euryarchaeota					
572546	*Archaeoglobus profundus DSM 5631*	Archaeoglobales	Euryarchaeota					
589924	*Ferroglobus placidus DSM 10642*	Archaeoglobales	Euryarchaeota					
439481	*Aciduliprofundum boonei T469 (YP_003483718)*	DHEV2	Euryarchaeota		YP_003483717.1	YP_003483718	YP_003483720.1	EK-J
263820	*Picrophilus torridus DSM 9790*	Thermoplasmatales	Euryarchaeota		YP_023617	YP_023618	YP_023619	EKJ
273075	*Thermoplasma acidophilum DSM 1728*	Thermoplasmatales	Euryarchaeota		NP_394545	NP_394546	NP_394547	EKJ
273116	*Thermoplasma volcanium GSS1*	Thermoplasmatales	Euryarchaeota		NP_111008	NP_111007	NP_111006	EKJ
333146	*Ferroplasma acidarmanus fer1*	Thermoplasmatales	Euryarchaeota		ZP_05571351	ZP_05571352	ZP_05571353	EKJ
425595	'*Candidatus* Micrarchaeum acidiphilum ARMAN-2'	ARMAN	Euryarchaeota		EET90013	EET90012	EET90011	EKJ
662760	'*Candidatus* Parvarchaeum acidiphilum ARMAN-4'	ARMAN	Euryarchaeota		EEZ92631	EEZ93272	EEZ93273	E / KJ
662762	'*Candidatus* Parvarchaeum acidophilus ARMAN-5'	ARMAN	Euryarchaeota		EFD92431	EFD92430	EFD92429	EKJ
243232	*Methanocaldococcus jannaschii DSM 2661*	Methanococcales	Euryarchaeota					
267377	*Methanococcus maripaludis S2*	Methanococcales	Euryarchaeota					
402880	*Methanococcus maripaludis C5*	Methanococcales	Euryarchaeota					
406327	*Methanococcus vannielii SB*	Methanococcales	Euryarchaeota		YP_001322618	YP_001322619		EK
419665	*Methanococcus aeolicus Nankai-3*	Methanococcales	Euryarchaeota					
426368	*Methanococcus maripaludis C7*	Methanococcales	Euryarchaeota					
444158	*Methanococcus maripaludis C6*	Methanococcales	Euryarchaeota					
456320	*Methanococcus voltae A3*	Methanococcales	Euryarchaeota					
573063	*Methanocaldococcus infernus ME*	Methanococcales	Euryarchaeota					
573064	*Methanocaldococcus fervens AG86*	Methanococcales	Euryarchaeota					
579137	*Methanocaldococcus vulcanius M7*	Methanococcales	Euryarchaeota					
187420	*Methanothermobacter thermautotrophicus str. Delta H*	Methanobacteriales	Euryarchaeota		NP_276410	NP_276411	NP_276412	EKJ
339860	*Methanosphaera stadtmanae DSM 3091*	Methanobacteriales	Euryarchaeota		YP_448530	YP_448529	YP_448528	EKJ
420247	*Methanobrevibacter smithii ATCC 35061*	Methanobacteriales	Euryarchaeota		YP_001273681	YP_001273682	YP_001273683	EKJ
634498	*Methanobrevibacter ruminantium M1*	Methanobacteriales	Euryarchaeota		YP_003424780	YP_003424781	YP_003424782	EKJ
190192	*Methanopyrus kandleri AV19*	Methanopyrales	Euryarchaeota					
69014	Thermococcus kodakarensis KOD1	Thermococcales	Euryarchaeota					
70601	Pyrococcus horikoshii OT3	Thermococcales	Euryarchaeota					
186497	Pyrococcus furiosus DSM 3638	Thermococcales	Euryarchaeota					
246969	Thermococcus sp. AM4	Thermococcales	Euryarchaeota					
272844	Pyrococcus abyssi GE5	Thermococcales	Euryarchaeota					
523850	Thermococcus onnurineus NA1	Thermococcales	Euryarchaeota					
593117	Thermococcus gammatolerans EJ3	Thermococcales	Euryarchaeota					
604354	Thermococcus sibiricus MM 739	Thermococcales	Euryarchaeota					
228908	Nanoarchaeum equitans Kin4-M	Nanoarchaeota	Euryarchaeota					
272557	Aeropyrum pernix K1	Desulfurococcales	Crenarchaeota					
399550	Staphylothermus marinus F1	Desulfurococcales	Crenarchaeota					
415426	Hyperthermus butylicus DSM 5456	Desulfurococcales	Crenarchaeota					
453591	Ignicoccus hospitalis KIN4/I	Desulfurococcales	Crenarchaeota					
490899	Desulfurococcus kamchatkensis 1221n	Desulfurococcales	Crenarchaeota					
591019	Staphylothermus hellenicus DSM 12710	Desulfurococcales	Crenarchaeota					
633148	Thermosphaera aggregans DSM 11486	Desulfurococcales	Crenarchaeota					
273057	*Sulfolobus solfataricus P2*	Sulfolobales	Crenarchaeota					
273063	*Sulfolobus tokodaii str. 7*	Sulfolobales	Crenarchaeota					
330779	*Sulfolobus acidocaldarius DSM 639*	Sulfolobales	Crenarchaeota					
399549	*Metallosphaera sedula DSM 5348*	Sulfolobales	Crenarchaeota					
419942	*Sulfolobus islandicus Y.N.15.51*	Sulfolobales	Crenarchaeota					
425944	*Sulfolobus islandicus L.D.8.5*	Sulfolobales	Crenarchaeota					
426118	*Sulfolobus islandicus M.16.4*	Sulfolobales	Crenarchaeota					
427317	*Sulfolobus islandicus M.14.25*	Sulfolobales	Crenarchaeota					
427318	*Sulfolobus islandicus M.16.27*	Sulfolobales	Crenarchaeota					
429572	*Sulfolobus islandicus L.S.2.15*	Sulfolobales	Crenarchaeota					
439386	*Sulfolobus islandicus Y.G.57.14*	Sulfolobales	Crenarchaeota					
178306	*Pyrobaculum aerophilum str. IM2*	Thermoproteales	Crenarchaeota					
340102	*Pyrobaculum arsenaticum DSM 13514*	Thermoproteales	Crenarchaeota					
368408	*Thermofilum pendens Hrk 5*	Thermoproteales	Crenarchaeota					
384616	*Pyrobaculum islandicum DSM 4184*	Thermoproteales	Crenarchaeota					
397948	*Caldivirga maquilingensis IC-167*	Thermoproteales	Crenarchaeota					
410359	*Pyrobaculum calidifontis JCM 11548*	Thermoproteales	Crenarchaeota					
444157	*Thermoproteus neutrophilus V24Sta*	Thermoproteales	Crenarchaeota					
374847	'*Candidatus* Korarchaeum cryptofilum OPF8'		Korarchaeota					

(Petitjean et al., Suppl. Figure S1)



(Petitjean et al., Suppl. Figure S2)



(Petitjean et al., Suppl. Figure S3)