UNIVERSITE DE LIEGE

Faculté des Sciences
InBioS (Integrative Biological Sciences from molecules to systems)
Unit of Eukaryotic Phylogenomics

PHYLOGENOMIQUE DES ARCHEES

&

RELATIONS AVEC LES EUCARYOTES

THESE DE DOCTORAT
Présentée en vue de l'obtention du grade de Docteur en Sciences
Soutenue par

Richard Thierry GOUY

Financement : F.R.S.-FNRS FRIA

Thèse soutenue le 21 novembre 2024 devant un jury composé de :

Directeur de thèse
Prof. Denis BAURAIN, InBioS, ULiège
Président du jury
Prof. Patrick MEYER, InBioS, ULiège
Examinateurs
Prof. Emmanuelle JAVAUX, ASTROBIOLOGY, ULiège
Dr. Alice MOUTON, SEED, ULiège
Dr. Henner BRINKMANN, ex-DSMZ, Braunschweig/ULiège (coll. sci.)
Dr. Damien P. DEVOS, Pablo de Olavide University, Séville
Dr. Ugo CENCI, UGSF, Université de Lille

A ma famille qui m'a vu réaliser cette thèse,

qui m'a supporté et soutenu

tout au long de ce rêve de devenir chercheur

REMERCIEMENTS

Je tiens à remercier le plus profondément possible et témoigner mon plus grand respect à mon Directeur de thèse, le Professeur Denis Baurain, pour avoir cru en moi et répondu présent au cours des nombreuses péripéties qui ont pu jalonner cette aventure, ô combien stressante et éprouvante moralement, mais tout autant prenante, passionnante et stimulante.

Je remercie également Hervé Philippe pour m'avoir permis de publier mon premier article et pour ses conseils dans la construction du jeu de données.

Je tiens à exprimer toute ma reconnaissance aux membres de mon comité de thèse pour avoir toujours cru en ce projet et qui m'ont permis de le mener à terme.

J'aimerais ensuite remercier les membres du laboratoire : Amandine, Raphaël, Mick, Luc, Loïc, Catherine et Valérian. Ça a toujours été un plaisir d'être avec vous et je garderai un souvenir impérissable de tous ces moments passés ensemble.

Je remercie également Rosa et Damien pour leur soutien technique et logistique.

Je remercie les membres du jury. Merci d'avoir accepté de prendre le temps d'évaluer mon travail.

Enfin, je ne pourrai jamais remercier assez ma famille, mes parents et mon frère. Merci pour tout votre amour et votre soutien infaillible.

ABSTRACT

The origin of the eukaryotic cell remains one of the most contentious puzzles in evolutionary biology. In the late 1970s, by discovering the domain Archaea, Woese put an end to the dichotomous view of life (eukaryotes *vs* prokaryotes) (C. R. Woese & Fox, 1977). His work, dubbed "Woese's Revolution" shows that the living world is divided into three domains: bacteria, archaea and eukaryota. First, bacteria were considered the ancestral line, which gave birth to archaea and eukaryota. However, this initial view, from simple to complex, is still reassessed, especially since we know that some species can evolve by secondary simplification. Accordingly, rooting the tree of life has become a problem, relationships among these three domains being not reliable. Moreover, the eukaryotic cell seems to have both bacterial operational genes and archaeal informational genes, so that it could have originated from a fusion event between a bacterium and an archaeon. Then, since the discovery of the Asgard group, it has been suggested that eukaryota originate from archaea, making them paraphyletic. Nevertheless, things are perhaps not as simple. Indeed, many artifacts can affect phylogenetics reconstructions, such as long branch attraction phenomenon, contaminations etc.

Two main types of competing scenarii explain the origin of the eukaryotic cell. The first posits a symbiotic fusion between an archaea (whose nature varies) and an α-proteobacterium at the origin of the mitochondrion, while the second considers the eukaryotes as an independent lineage of both Archaea and Bacteria that would have, during its evolution, phagocyted an α-proteobacterium. The recent discoveries on the Archaea, in particular with the highlighting of the Asgard group, have thus revived the debate between the supporters of a two-domain life (the eukaryotes being descended from the Archaea, therefore paraphyletic) and the supporters of a three-domain life.

The aim of this thesis is to revisit the question of the relationship between Archaea and Eukaryotes, based on an original and rigorous methodology. The Archaea, which were very poorly represented until recently, are now appearing more and more as a very diverse domain of life. The discovery of new Archaea, the super-phylum Asgard, possessing genes encoding proteins previously considered specific to eukaryotes, suggests that the latter could be directly derived from the former and would thus be the result of a fusion between an Asgard Archaea and a Bacterium. However, the reliability of the datasets as well as the phylogenetic inference methods can sometimes be questionable. Phylogenetic inference models struggle to avoid artifacts that often go unnoticed. The considerations of this thesis manuscript focus on these methodological biases in order to minimize systematic errors in phylogenomic reconstruction and provide a more reliable phylogeny between Archaea and Eukaryotes.

In a first chapter, we revisited the question of the root of the tree of life through the use of the Elongation Factor (EF) gene. Then we took a critical look at the methods used in papers dealing with deep phylogenies, focusing on the consideration of phylogenetic reconstruction artifacts and the identification of methodological biases.

In a second study, we established a phylogeny of the Archaea through a « quadratic jackknife » procedure resampling both genes and species, in order to evaluate the robustness of the phylogenetic results in the face of these variations in the data, but also in the methods (super-

matrices and super-trees) and models used (LG4X, C20, C60 and PMSF models). We then performed Slow-Fast analyses to compare tree topologies based on supermatrices of sites featuring different substitution rates. Our analyses favor the Korarchaeota rather than the SCGC group as a sister group of Sulfolobales + Desulfurococcales + Thermofilaceae + Thermoproteaceae. Furthermore, we recover Hadesarchaeota as the sister group to Thermococci, Theionarchaea and Methanomicrobia_Arc. Finally, our results struggle to systematically recover the monophyly of Euryarchaeota. Indeed, the DPANN group, whose position is itself uncertain, tends to attract the Altiarchaeaota, making the euryarchaeota paraphyletic. It is impossible for us to decide for one or the other solution. On the other hand, we also observe some polyphyly at the genus level. i.e. archaea classified in the same genus while belonging to different clades.

In the third study, we included Eukaryotes in our datasets in order to test the hypotheses concerning a relationship between Asgard Archaea and Eukaryotes. To do so, we consider multiple approaches to control the systematic error. Thus, we controlled problems related to paralogy by examining topologies for each gene and performed sites removal to test heterotachy and heteropecilly. For the slowest genes, the clustering of Asgard with eukaryotes is only minimally supported, in favor of a clustering of Asgard with TACKs. Only a strong addition of genes with fast mutation rates systematically groups Asgard with eukaryotes. We can therefore hypothesize that the grouping of Asgard with eukaryotes is the result of a phylogenetic reconstruction bias due to the use of genes with a too fast substitution rate.

We also conducted work aiming to root the archaeal tree, both with and without eukaryotes. Our findings support a root within the SANT group. Euryarchaeota appear to us as a paraphyletic group. The uncertainty lies in the group at the base of the Ouranosarchaea: DPANN, Altiarchaeota, or Hadesarchaeota. While the PMSF method of IQ-TREE favors a connection between DPANN and Altiarchaeota as the basal group, Bayesian inference analyses with PhyloBayes rather support Hadesarchaeota in this position.

We therefore conclude that it is difficult to promote with certainty a model of eukaryogenesis based on a 2-domain model where eukaryotes would be the sister group of Asgard. Without excluding an archaeal origin of eukaryotes, a 3-domain (or even 1-domain) scenario is still possible as long as doubts remain about the phylogenetic methods used. We debate the possibility that the grouping of Asgards with eukaryotes is, as the history of the conception of the tree of Life has too often shown us, the result of a lack of reliable data and analysis artifacts due to the particular biology of these species.

RESUME

L'origine des eucaryotes demeure l'une des questions les plus controversées de la biologie évolutive. A la fin des années 1970, en découvrant le domaine des archées, Carl Woese met fin à la vision dichotomique de la vie (eucaryotes *vs* procaryotes) (C. R. Woese & Fox, 1977). Ses travaux, qualifiés de « Révolution woesienne », démontrent que la vie est dès lors divisée en trois domaines : bactéries, archées et eucaryotes. Dans un premier temps, les bactéries furent considérées comme la lignée ancestrale, dont aurait pu dériver les archées et eucaryotes. Mais cette vision de la Vie, du simple vers le complexe, est aujourd'hui remise en question, notamment depuis le rejet des Archezoa. L'enracinement de l'arbre de la vie pose alors problème, les relations de parenté entre les trois domaines étant encore incertaines. De plus, la cellule eucaryote semble posséder à la fois des gènes opérationnels bactériens et des gènes informationnels archéens, laissant supposer une évolution par fusion entre bactéries et archées. Cependant, les choses ne sont peut-être pas aussi simples.

Deux grands types de scenarii en compétition permettent d'expliquer l'origine de la cellule eucaryote. Les premiers y voient une fusion symbiotique entre une archée (dont la nature varie) et une α-protéobactérie à l'origine de la mitochondrie, tandis que les seconds considèrent les eucaryotes comme une lignée indépendante des Archaea et des Bactéries qui aurait, au cours de son évolution, phagocyté une α-protéobactérie. Les récentes découvertes sur les Archaea, en particulier avec la mise en évidence du groupe Asgard, ont ainsi ravivé le débat entre les partisans d'une vie à deux domaines (les eucaryotes étant issus des Archaea, dès lors paraphylétiques) et les partisans d'une vie à trois domaines.

L'objet de cette thèse est notamment de revisiter la question des relations de parenté entre Archaea et Eucaryotes selon une méthodologie originale et systématique. Les Archaea, très peu représentées jusqu'à encore récemment, apparaissent de plus en plus comme un domaine de la vie très diversifié. La découverte de nouvelles Archaea, le super-phylum Asgard, possédant des gènes encodant des protéines auparavant considérées comme spécifiques aux eucaryotes, suggère que ces derniers pourraient en être directement issus et seraient donc le résultat d'une fusion entre une Archaea et une Bactérie. Cependant, la fiabilité des jeux de données ainsi que les méthodes d'inférence phylogénétique peuvent parfois être sujettes à caution. Les modèles d'inférence phylogénétique peinent à éviter des artefacts qui passent alors souvent inaperçus. Les considérations de ce manuscrit de thèse se focalisent sur ces biais méthodologiques afin de minimiser les erreurs de reconstruction phylogénomique et fournir une phylogénie des plus fiables entre Archaea et Eucaryotes.

Premièrement, nous avons revisité la question de la racine de l'arbre de la vie via l'utilisation du gène du Facteur d'Elongation (EF). Puis nous avons dressé un regard critique sur les méthodes employées ces dernières années dans les articles traitant de phylogénies profondes, en mettant l'accent sur la prise en compte des artefacts de reconstruction phylogénétique et la mise en évidence de biais méthodologiques.

Deuxièmement, nous avons établi une phylogénie des Archaea aux travers d'un jackknife à la fois de gènes et d'espèces, afin de comparer plusieurs jeux de données et ainsi étudier la stabilité des résultats phylogénétiques face à ces variations dans les données, mais aussi dans les

méthodes employées (super-matrices et super-arbres, modèles plus ou moins sophistiqués). Après des analyses de Slow-Fast afin de comparer des topologies d'arbres basées sur des super-matrices de sites ayant des vitesses de substitution différentes, nos analyses favorisent les Korarchaeota plutôt que le groupe SCGC comme groupe frère de Sulfolobales + Desulfurococcales + Thermofilaceae + Thermoproteaceae. Les Hadesarchaeota privilégient comme groupe frère des Thermococci, des Theionarchaea et des Methanomicrobia_Arc. De plus, nos résultats peinent à retrouver systématiquement la monophylie des Euryarchaeota. En effet, le groupe DPANN, dont sa position est elle-même incertaine, tend à attirer vers lui les Altiarchaeaota, rendant paraphylétique les euryarchaeota. Il nous est impossible de trancher pour l'une ou l'autre solution. En revanche, On notera également une attention particulière concernant des cas de synonymie à l'échelle du genre pour des archées appartenant à des clades différents.

Enfin troisièmement, nous avons inclus les Eucaryotes dans nos jeux de données afin de vérifier les hypothèses concernant une relation de parenté entre Archaea et Eucaryotes. Pour ce faire, nous avons envisagé plusieurs approches pour contrôler l'erreur systématique. Ainsi, nous avons contrôlé les problèmes liés à la paralogie en examinant les topologies pour chaque gène et nous avons procédé à du retrait de site afin de tester à la fois l'hétérotachie et l'hétéropécilie. Nos résultats indiquent qu'il est fort possible que la supposée relation de parenté entre les archées du groupe Asgard et les eucaryotes soit le résultat d'artefact de reconstruction phylogénétique. Pour les gènes les plus lents, le regroupement des Asgards auprès des eucaryotes n'est que très peu représenté, au profit d'un regroupement des Asgards auprès des TACK. Seul un fort ajout de gènes à taux de mutation rapide regroupe les Asgards systématiquement avec les eucaryotes. On peut dès lors supposer que le regroupement des Asgards avec les eucaryotes soit le fruit d'un biais de reconstruction phylogénétiques lié à l'utilisation de gènes à taux de substitution trop rapide qui fausse le signal phylogénétique. Autrement dit, ce regroupement fortuit est la conséquence d'une erreur systématique produisant un signal non-phylogénétique à la fois faux et statistiquement supportés entrant en compétition avec le véritable signal phylogénétique.

Nous avons également procédé à des travaux visant à enraciner l'arbre des archées, avec et sans eucaryotes. Nos travaux vont dans le sens d'une racine au sein du groupe SANT. Les Euryarchaeota nous apparaissent comme un groupe paraphylétique. L'incertitude réside sur le groupe à la base des Ouranosarchaea : DPANN, Altiarchaeota ou Hadesarchaeota. Alors que la méthode PMSF de IQ-TREE privilégie un rapprochement des DPANN et Altiarchaeota comme groupe à leur base, les analyse d'inférence bayésienne avec PhyloBayes soutiennent plutôt les Hadesarchaeota à cette place.

Nous concluons dès lors qu'il est difficile de promouvoir avec certitude un modèle d'eucaryogenèse basé sur un modèle à 2 domaines où les eucaryotes seraient le groupe frère des Asgards. Sans exclure une origine archéenne des eucaryotes, un scénario à 3 domaines (voire 1) reste encore envisageable tant que des doutes subsisteront sur les méthodes phylogénétiques employées. Nous débattons de la possibilité que le groupement des Asgards avec les eucaryotes est, comme l'histoire de la conception de l'arbre du vivant nous l'a trop souvent montré, le résultat d'un manque de données fiables et d'artéfacts d'analyses dus à la biologie particulière de ces espèces.

TABLE DES MATIERES

INTRODUCTION

1 DEFINIR LE VIVANT

« Si personne ne me le demande, je le sais bien ; mais si on me le demande, et que j'entreprenne de l'expliquer, je trouve que je l'ignore. »

Saint Augustin d'Hippone, IVe siècle.

Cette pensée de saint Augustin d'Hippone concernait la définition du temps. Mais il serait aujourd'hui tout à fait possible de la paraphraser pour la vie, tant en donner une définition est difficile, le passage entre l'inerte et la vie étant flou. L'une des principales questions concernant les origines de la vie est de comprendre les étapes de l'évolution ayant permis le passage d'une chimie prébiotique complexe à une biologie simple. De nos jours, trois types de molécules géantes sont utilisées pour encoder et transmettre l'information génétique : deux familles d'acides nucléiques (l'acide ribonucléique (ARN) et l'acide désoxyribonucléique (ADN)) et les protéines (composées d'une vingtaine d'acides aminés différents). Le dogme central de la biologie moléculaire, introduit dans les années 1950 par Francis Crick (co-découvreur de la structure de l'ADN qui lui valut le prix Nobel de physiologie et de médecine en 1962) établit le lien qui existe entre le matériel génétique (ADN et ARN) contenu dans la cellule et les protéines que cette cellule synthétise. Ce dogme stipule que chez tous les êtres vivants (au moins les êtres vivants actuels), l'information génétique est transmise dans un seul sens : de l'ADN aux protéines en passant par l'ARN, structure transitoire permettant la transmission de l'information à une machine de traduction pour produire les protéines, constituants de base qui font fonctionner la cellule, les tissus, les organes et l'organisme entier (Watson & Crick F H C, 1953). Ce transfert d'information repose sur le code génétique, universel chez tous les êtres vivants.

Finalement, qu'est-ce que la vie ? On peut regrouper les définitions de la vie en trois grandes catégories :

- La première est basée sur les capacités d'auto-réplication et d'évolution. C'est la définition adoptée officiellement par la NASA : « La vie est un système chimique auto-entretenu capable d'évolution darwinienne » (Joyce, 2002; Podolsky & Tauber, 1994). Une critique possible est que cette définition ne considère donc pas comme vivant des êtres incapables de se reproduire (ex. hybrides stériles comme une mule, fourmis ouvrières…).
- La seconde, plus fonctionnelle, est axée sur la présence de caractéristiques nécessaires et suffisantes. Le « chemoton » (Ganti Tibor, 2003c, 2003a, 2003b), modèle formalisé par Tibor Ganti, stipule que la vie repose sur trois propriétés : le métabolisme, l'auto-catalyse/auto-réplication et une membrane formée d'une double couche lipidique. Dans ce modèle, la notion d'une structure confinée séparée par une enveloppe et capable de développer un métabolisme est essentielle. Ainsi, selon P. L. Luisi : « La forme de vie minimale est un système circonscrit par un compartiment semi-perméable qui s'auto-organise et s'auto-entretient en produisant ses propres éléments constitutifs par la transformation de l'énergie et des nutriments extérieurs à l'aide de ses propres mécanismes de production » (Cleland & Chyba, 2002). C'est ce que l'on essaie de reproduire dans les manipulations de chimie prébiotique : on explore toutes les voies

possibles menant des substances inorganiques à des substances organiques, des biomolécules ou d'autres biomolécules analogues.

- La troisième est une vision purement physico-chimique, émise par Schrödinger (Schrodinger, 2012), qui donne aux êtres vivants un « point d'entropie négative » ou « systèmes dissipatifs d'entropie ». Un être vivant serait un être capable de soutirer de l'énergie à son environnement pour organiser la matière, l'ordonner, et donc faire reculer pour un temps le deuxième principe de la thermodynamique. Les êtres vivants seraient donc des zones d'entropie négative extrêmement localisées dans l'espace et le temps. L'intérêt de cette définition est qu'elle s'affranchit des conditions purement terrestres de la vie et ne s'appuie sur aucun préjugé quant aux structures et comportements de cette Vie. Cette définition pourrait être une bonne référence pour un exobiologiste qui chercherait à savoir si la vie existe ailleurs dans l'Univers à partir d'une chimie différente de celle de la Terre.

Notons que, bien que tentant d'englober les principales caractéristiques de ce que nous considérons actuellement comme vivant, un certain nombre de systèmes généralement considérés comme non vivants répondent en partie à ces propriétés : virus (parasites obligatoires), certaines protéines comme les prions, liposomes et autres protocellules artificielles, ribozymes ou ARN auto-réplicants, et même certains cristaux et argiles, virus et programmes informatiques ou encore… le feu !

2 DECRIRE LE VIVANT

2.1 DE LA SYSTEMATIQUE…

La systématique a pour objectif de nommer, décrire et classer l'ensemble des êtres vivants. Elle permet l'étude de la diversité de la Vie. Le besoin de classer les organismes est considéré comme le reflet d'un ordre naturel. Les premières tentatives de classification de la vie sont très anciennes et remontent au moins à l'Antiquité, où l'on essaie d'y voir un véritable système organisé. Ainsi, Aristote (384-322 av. J-C), considéré comme le père de la zoologie, fut le premier à construire un système zoologique (*Histoire des animaux* et *Traité des parties des animaux*) en tentant de décrire et classer le monde animal (Sapp, 2005, 2006). Il y distingue dans un premier temps les animaux « sanguins » (vertébrés) et « non-sanguins » (invertébrés) et décrit l'homme comme un « être à part doué de raison ». Il peut être considéré comme le père du fixisme car il estime les espèces éternelles et immuables. Il classe les espèces selon une échelle de complexité des êtres allant des organismes les plus simples et moins organisés jusqu'aux organismes les plus complexes, avec l'homme au sommet. Cette conception hiérarchique de l'organisation de la Vie, l'étageant selon un axe linéaire de propriétés cumulées, donna naissance au scalisme, (*Scala naturae*, littéralement « l'échelle de la nature ») déjà suggéré auparavant par Démocrite et Platon. Théophraste (372-287 av. J-C), élève d'Aristote, sera quant à lui considéré comme le père de la botanique en tentant de classer les végétaux au travers de ses deux ouvrages (*Histoire des plantes* et *Recherche sur les plantes*). Au premier siècle de notre ère, à Rome, Pline l'ancien (23-79) (surnommé Pline le naturaliste) va rédiger une *Histoire Naturelle* en 37 volumes qui traite aussi bien du règne animal que végétal. Il complète ce travail par des critères utilitaires (agricoles, alimentaires, thérapeutiques...) mais sans véritable esprit de classification.

Le Moyen-Âge n'apporte que peu de nouveautés et d'originalité dans la classification du monde vivant. Les textes de l'Antiquité vont acquérir au Moyen-Âge un aspect mystique et faire autorité, à la fois au sein des communautés religieuse et scientifique. En effet, ces textes sont en parfaite adéquation avec la vision créationniste et fixiste de la Vie, prônée par les religions monothéistes de l'époque. Ils contribuent à figer le vivant dans le principe de fixité des espèces, tout en gardant les principes utilitaires conformes à la vision religieuse.

Fin XVIIème siècle, Joseph Pitton de Tournefort présente un système de classification plus clair et plus précis que ses prédécesseurs : « *D'abord nommer, ensuite classer* ». Il souhaite « *éviter que les noms de plantes atteignent le nombre même des plantes !* ». A la même époque, en Angleterre, John Ray s'attaque au problème de la définition d'espèce en la fondant sur une nécessité de ressemblance entre les parents et leurs descendants. Il est également le premier à postuler que deux individus d'espèces différentes ne peuvent avoir une descendance viable et fertile. Il est aussi un précurseur dans l'étude des fossiles, remettant en doute la vision biblique du Déluge Universel en affirmant qu'il s'agit de restes d'organismes vivants ayant disparu. Ce n'est qu'au XVIIIème siècle que Carl von Linné (1707-1778) va rationaliser la classification des espèces, en publiant en 1735 sa première édition des *Systèmes de la Nature* (*Systema Naturae*), dans laquelle il établit une classification hiérarchique de la vie : classe, ordre, genre et espèce. Il devient ainsi le fondateur de la classification moderne, basée sur la ressemblance entre espèces. Il souligne ainsi que de nombreux êtres vivants partagent beaucoup de caractères communs, et il place les humains, avec les singes, dans le groupe des primates. En 1758, Linné va ainsi généraliser dans sa 10ème édition la dénomination binomiale des espèces (déjà amorcée par Guillaume Rondelet et Pierre Belon au XVIe siècle), avec deux noms latins : le premier correspondant au genre, et le second à l'espèce, unité de base de son système hiérarchique.

2.2 … A LA THEORIE DE L'EVOLUTION

Pour Linné, les espèces étaient considérées comme fixes et immuables. Sa classification repose d'ailleurs sur la fixité des caractères. Linné, dont le père était pasteur, avait étudié la théologie et était donc un fixiste résolu, croyant à un « ordre souverain de la nature ». Il classe l'homme et le singe dans le même ordre des *Anthropomorpha*, à l'encontre des instances religieuses qui voient l'homme comme l'aboutissement ultime de la création divine. Malgré tout, pour lui, le but n'était pas de remettre en question la supériorité intellectuelle et morale de l'homme sur le singe, tous les deux étant le résultat d'une Création et demeurant donc aussi parfaits qu'immuables. Or, à la même époque, va se développer l'idée que les espèces ne sont pas fixes et immuables, mais au contraire se transforment et évoluent. Georges-Louis Leclerc, Comte de Buffon, utilise les critères d'interfécondité et interstérilité pour décrire les espèces. Il considère comme appartenant à la même espèce les individus capables de produire entre eux des descendants fertiles, et comme n'appartenant pas à la même espèce des individus interstériles. Les espèces actuelles sont alors la continuation d'espèces antérieures. En élaborant cette idée, Buffon est l'un des précurseurs du transformisme. Avec son *Histoire Naturelle* en 44 volumes (le premier volume est publié en 1744), il tente de rédiger une encyclopédie décrivant toute la biologie et la géologie de son époque. Il propose que les éléphants d'Afrique et d'Asie soient les descendants des mammouths, dont des fossiles viennent d'être découverts, à son époque, en Sibérie. Il disserte sur la formation de la Terre et sur l'origine de la vie et estime l'âge de la Terre à 74 000 ans. Bien que l'on sache aujourd'hui que cet âge est encore loin de la réalité, son travail

permet non seulement l'émancipation des idées par rapport à la Bible (6000 ans), mais également d'inclure un nouveau paramètre à prendre en compte pour la transformation des espèces : le temps.

En 1789, Antoine-Laurent de Jussieu publie une classification botanique en mettant au point le principe de hiérarchisation des caractères. Pour identifier un taxon donné, une classe par exemple, l'idéal est d'avoir un (ou plusieurs) caractère(s) constant(s) à l'intérieur de cette classe et différent(s) dans toutes les autres. Un type de caractère est donc utile à un niveau précis de la classification, certains au niveau de l'ordre, d'autres au niveau du genre etc. Les caractères sont donc « subordonnés ».

En 1809, Jean-Baptiste Pierre Antoine de Monet, chevalier de Lamarck, dans sa *Philosophie zoologique*, introduit le concept d'une évolution graduelle des espèces : c'est le transformisme. Il est notamment, avec l'Allemand Treviranus, celui à qui on doit la création du terme « biologie ». Sa théorie transformiste de la transmission des caractères acquis est fondée sur deux principes :

1. la complexification croissante de l'organisation des êtres vivants, du plus simple au plus complexe ;
2. leur diversification, ou spéciation, en espèces, à la suite d'une adaptation de leur comportement ou de leurs organes à leur milieu.

Pour expliquer ce second principe, il avance deux lois :

1. la « loi d'usage et de non-usage », souvent résumée par la formule « la fonction crée l'organe », stipule qu'à la suite de l'emploi plus fréquent et soutenu d'un organe quelconque, celui-ci se développe peu à peu en fonction de l'emploi qu'on lui réserve, et à l'opposé, détériore progressivement les facultés d'un organe si ce dernier n'est pas utilisé.
2. La transmission des caractères acquis, qui consiste en la possibilité de transmettre à la descendance les changements organiques ou morphologiques acquis au cours de la vie, en rapport avec la première loi.

Cet expert de la classification des mollusques, des vers et des insectes remarqua un ensemble de changements (une évolution) entre des fossiles de mollusques et les mollusques vivants qu'il avait regroupés dans sa classification. Il proposa que de nouvelles espèces se forment lorsque les animaux et les plantes s'adaptent à un milieu de vie changeant, modifiant de génération en génération leurs organes et leur apparence par renforcement. Sa vision du vivant est basée sur le postulat de l'hérédité des caractères acquis. Il crut voir un sens dans cette évolution, qui se serait faite des formes les plus simples vers les plus complexes, mais sans proposer d'explication quant au mécanisme qui permettait aux êtres vivants de s'adapter à leur milieu, sinon qu'en postulant qu'un organe très utilisé se développe alors qu'un autre, peu utilisé, régresse et disparait. Lamarck fut très critiqué par ses contemporains, mais il fut le premier à publier clairement l'idée, appuyée par des observations des êtres vivants et des fossiles, que les espèces étaient sujettes à une évolution dans le temps.

En 1796, Etienne Geoffroy Saint-Hilaire émet l'idée que la nature aurait formé tous les êtres vivants sur un plan d'organisation unique, aux variations nombreuses, ce qui suppose une origine commune de toutes les espèces. Cuvier, élève de ce dernier, s'oppose à l'idée du transformisme de Lamarck et de Geoffroy Saint-Hilaire. Pour lui, l'apparition et la disparition des

espèces sont le résultat de catastrophes soudaines qui anéantissent les espèces que l'on retrouve sous forme fossiles afin de laisser la place à de nouvelles espèces. Cette théorie du catastrophisme permet d'expliquer l'apparition et la disparition des espèces sans toucher à leur immuabilité. Il reprend également la méthode des travaux de Jussieu, qu'il applique aux animaux avec sa loi de subordination des organes. Fondateur de l'anatomie comparée, il part du postulat que les organes d'un animal sont fonctionnellement dépendants et en fait un principe de classification qu'il applique à l'étude des fossiles.

En 1859, dans *L'Origine des Espèces*, Charles Darwin popularise la représentation du vivant sous la forme d'un arbre phylogénétique (**Figure 1**), remettant ainsi en cause le scalisme d'Aristote, en développant l'idée d'une évolution biologique buissonnante basée sur la sélection naturelle (Ragan, 2009). A la même période, Alfred R. Wallace (père de la biogéographie) arrive aux mêmes conclusions que Darwin, en mettant en avant la sélection sexuelle comme processus évolutif. C'est une lettre de Wallace qui poussera Darwin à « précipiter » la publication de sa théorie de l'évolution. Ainsi, pour Darwin et Wallace, les caractères de ressemblance ou de différence devraient être considérés comme le résultat de phénomènes évolutifs.

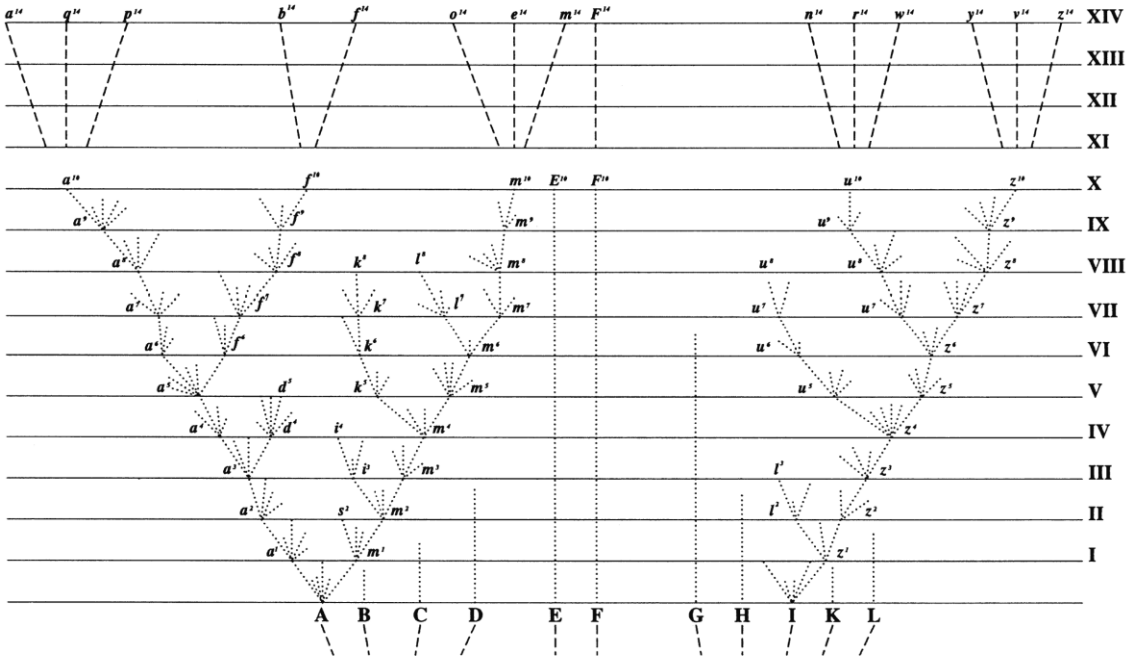


Figure 1. L'unique figure de Darwin, dans son livre L'Origine des Espèces par Voie de Sélection Naturelle (1859).

Cette illustration met en lumière le processus de descendance avec modification. Ce processus établit des liens généalogiques entre les êtres vivants. Les espèces et formes de vie ancestrales sont situées à la base de cette représentation. Sous l'effet de la sélection naturelle, les membres d'une espèce divergent au fil des générations en accumulant des variations, donnant naissance à de nouvelles lignées. Seules les variations qui ont pu être avantageuses auront été conservées par sélection naturelle. Les intervalles entre les lignes horizontales représentent des nombres importants de générations. C'est pourquoi les espèces contemporaines (en haut du schéma) sont reliées à leurs ancêtres par une longue série de lignées intermédiaires désormais disparues. L'ensemble forme un arbre évolutif.

En 1866, dans Generelle Morphologie, Ernst Haeckel, fervent défenseur des idées de Darwin, a également abordé un sujet que ce dernier avait ignoré dans L'Origine des Espèces, mais essentiel pour une image complète de l'évolution : l'origine de la vie. Haeckel postule que la vie sur Terre est née d'une "archegonia", c'est-à-dire de la génération spontanée des organismes les plus primitifs sans structure (monères) à partir de la matière inorganique. Ainsi, selon Haeckel, l'apparition initiale de toute vie était polyphylétique et la matière vivante est apparue directement à partir de substances chimiques inorganiques et non à partir de substances organiques préalablement générées (Kutschera et al., 2019; Levit & Hossfeld, 2019; Schmitt, 2009). Il publie ainsi les premiers arbres phylogénétiques (**Figure 2**), même s'il place encore l'homme au sommet de ses arbres. Comme pour ses contemporains, l'héritage de la pensée d'Aristote est toujours présent. Ses arbres se dessinent avec les êtres les plus simples vers la racine, jugés comme primitifs, jusqu'aux êtres les plus complexes dans les branches terminales, en accord avec sa « loi biogénétique », selon laquelle l'ontogenèse récapitule la phylogenèse. L'homme est ainsi le « produit terminal » d'une évolution. Il a proposé que les *Monera*, étaient au stade le plus bas d'un troisième royaume qu'il nomme *Protista* (Haeckel, 1866). Haeckel va ainsi diviser la vie en trois

règnes : les plantes, les animaux et les protistes, ces derniers correspondant aux formes unicellulaires.

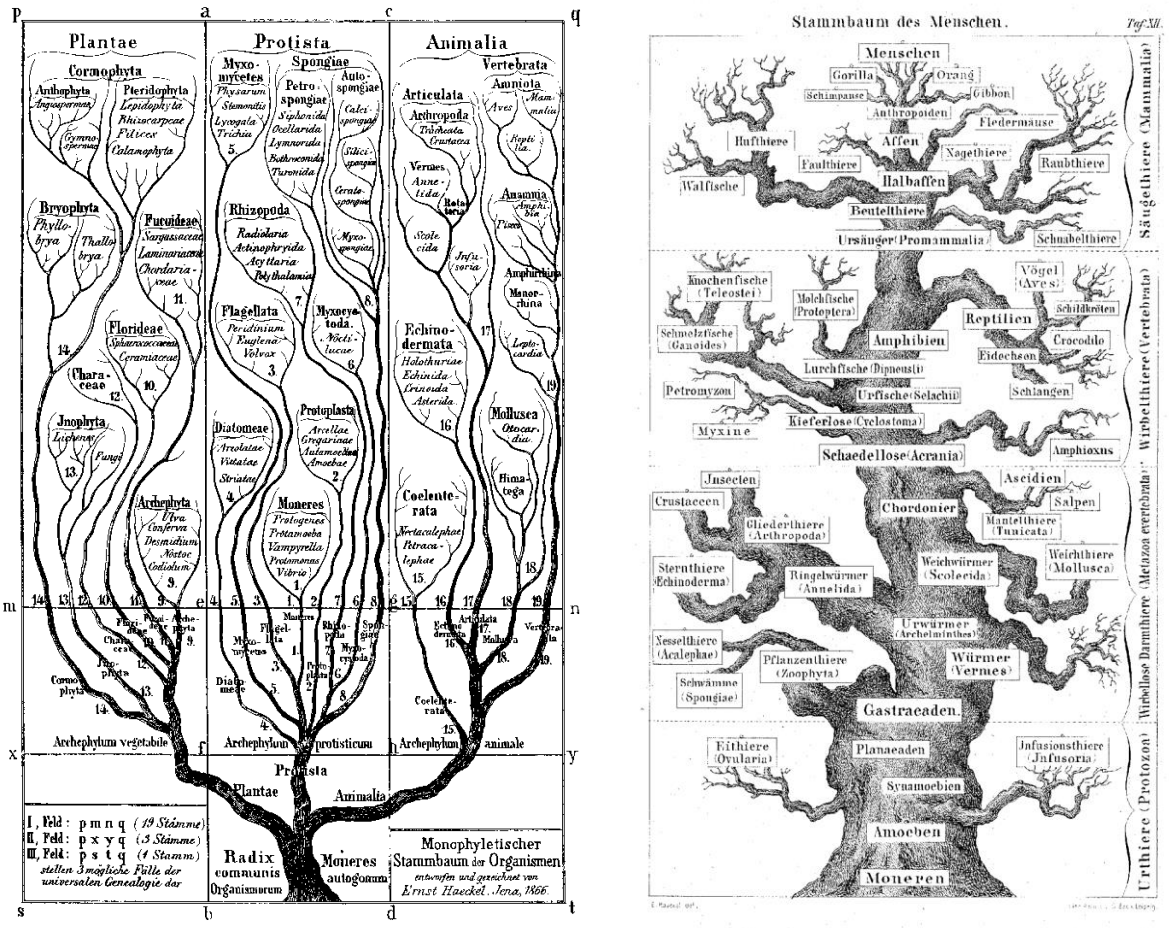


Figure 2. Arbres de Haeckel (a) Version de 1866 montrant les plantes, les protistes et les animaux ; (b) Version de 1874 conduisant à l'espèce humaine.

En 1866, le naturaliste allemand Ernst Haeckel a créé un arbre évolutif unique qui rassemblait l'ensemble des organismes vivants connus en trois grands règnes : les animaux, les végétaux, et également les protistes. Sous ce dernier terme, il a regroupé les organismes unicellulaires tels que les flagellés, les amibes, les diatomées, les éponges et les bactéries (alors appelées monères). Cet arbre, qui est considéré comme l'un des premiers arbres phylogénétiques modernes, illustre la diversité du vivant et les liens de parenté entre les différentes formes de vie. En 1874, Haeckel a continué à explorer ces idées dans son ouvrage *Anthropogenie*, où il a esquissé un arbre similaire pour illustrer l'évolution des êtres humains et leur lien avec d'autres formes de vie. Ces arbres ont contribué à populariser la théorie de l'évolution et ont ouvert la voie à de futures recherches sur la diversité et l'histoire de la vie sur Terre.

2.3 DE LA MICROBIOLOGIE…

Depuis l'Antiquité, on explique que la vie peut apparaitre presque partout, non seulement par des êtres de leur espèce (reproduction), mais également dans la boue, les végétaux ou la matière en décomposition : c'est la théorie de la génération spontanée d'organismes vivants. Cette idée héritée des Grecs estime que la vie est une propriété de la matière et qu'elle peut apparaitre

à chaque fois que les conditions y sont propices, répondant ainsi à une volonté créatrice d'ordre divin. Jusqu'au XVIIème siècle, le monde microscopique (invisible) n'était pas encore connu. Les scientifiques ne proposent pas d'explication de la diversité du monde vivant, pour la simple raison qu'ils en ont pour la plupart une conception fixiste. En 1668, Francesco Redi démontre que les asticots ne se forment pas spontanément dans la viande en décomposition, contrairement à ce que tout le monde pensait jusqu'alors, mais qu'ils sont en fait des larves de mouches qui se sont développées sur un milieu favorable et non le résultat d'une hypothétique conséquence de génération spontanée. A la même période, Antoni van Leeuwenhoek, drapier et naturaliste amateur hollandais, commence à perfectionner les lentilles de microscope pour vérifier la pureté des étoffes (Gest, 2004; Wollman et al., 2015). Il fut le premier à découvrir les bactéries et divers protozoaires, et réalise les premières descriptions de globules rouges et de spermatozoïdes. Redi et van Leeuwenhoek font partie des premiers à remettre en cause la théorie de la génération spontanée. Robert Hooke, quant à lui, construit également son propre microscope et donne le nom de « cellules » aux structures qu'il observe. Il est l'un des premiers à émettre l'hypothèse selon laquelle la surface de la Terre aurait subi des bouleversements au cours du temps et que les fossiles seraient les reliques d'espèces disparues appartenant à des temps plus anciens. Il s'ensuivra de nombreuses controverses sur cette théorie au sein de la communauté scientifique de l'époque, que l'on surnomme « guerre des bouillons ». Si on cesse peu à peu de croire à la génération spontanée pour les êtres vivants visibles à l'œil nu, pour le domaine de l'invisible, la controverse entre scientifiques, philosophes et religieux va perdurer et s'exacerber jusqu'au XIXème siècle. Il faudra attendre Louis Pasteur qui, en mettant au point un protocole expérimental rigoureux de stérilisation, bouleversera la vision du monde vivant et portera le coup de grâce à la théorie de la génération spontanée au profit de la théorie cellulaire. Cette théorie stipule que la cellule est l'unité structurale, l'unité fonctionnelle et l'unité reproductrice du vivant. Aujourd'hui, les points consensus de la théorie cellulaire sont les suivants :

- Tous les êtres vivants sont constitués d'une ou plusieurs cellules (ce qui définit les êtres vivants et exclut les virus de cette catégorie).
- Toute cellule provient d'une autre cellule par division cellulaire.
- La cellule est une unité vivante et l'unité de base du vivant, c'est-à-dire qu'une cellule est une entité plus ou moins autonome capable de réaliser certaines fonctions nécessaires et suffisantes à sa vie.
- Il y a individualité cellulaire grâce à la membrane plasmique qui règle les échanges entre la cellule et son environnement.
- La cellule renferme sous forme d'ADN l'information nécessaire à son fonctionnement et à sa reproduction.

Bien que la découverte des unicellulaires date du XVIIème siècle (Adoutte et al., 1996; Philippe et al., 1995), et que la première description du monde microbien date du XIXème siècle, la distinction entre procaryotes et eucaryotes est relativement récente en biologie et étroitement liée aux progrès de la microscopie, photonique d'abord puis surtout électronique. C'est Edouard Chatton qui, en 1925 fut le premier à différencier deux grands types cellulaires (cellule avec ou sans noyau), qu'il nomme respectivement « eucaryote » et « procaryote » (E. Chatton, 1925; É. P.

L. Chatton, 1938; Sapp, 2005; Scamardella, 1999). Toutefois, son idée fut dans un premier temps rejetée (Sapp, 2005).

L'année suivante, en 1938, Herbert Copeland propose que les Monera de Haeckel aient leur propre règne, sur la base du fait qu'ils sont « les descendants relativement peu modifiés de n'importe quelle forme unique de vie apparue sur terre, et qu'ils se distinguent nettement des protistes par l'absence de noyaux » (Copeland, 1938, 1956). Il introduit un système de classification à quatre règnes : Monera, Protista, Plantae et Animalia.

En 1957, André Lwoff, élève d'Edouard Chatton, fait la distinction entre les virus et les bactéries, en montrant que les virus ne contiennent qu'un seul type d'acide nucléique (ARN ou ADN) et qu'ils ne se reproduisent pas par division comme une cellule (Lwoff, 1957).

En 1962, Roger Yate Stanier et Cornelius Bernardus Van Niel revisitent l'idée de Chatton dans « *The Concept of a Bacterium* » et montrent qu'ils ne peuvent définir les bactéries que de manière négative par rapport aux eucaryotes, leur principale caractéristique étant l'absence de membranes internes et de matériel génétique entouré d'une membrane nucléaire (Stanier & Van Niel, 1962). La seule définition positive d'une bactérie est la présence d'une paroi cellulaire constituée de peptidoglycane (les bactéries sans paroi comme les Mollicutes sont le résultat d'une simplification secondaire).

En 1969, Robert Harding Whittaker publie une nouvelle classification de la vie en cinq règnes (Whittaker, 1969) : Monera (bactéries + algues bleu-vert (= Cyanobactéries)), Fungi (= champignons), Protista, Plantae et Animalia (**Figure 3**). Sa classification prend en compte le type et niveau d'organisation et la complexité cellulaire (procaryote, eucaryote uni- ou pluri-cellulaire), ainsi que son type de nutrition (photosynthèse, absorption et ingestion). Il considère les bactéries comme les êtres les plus inférieurs à partir desquels se sont développées progressivement des formes de vie plus complexes jusqu'aux organismes multicellulaires. Il fut également le premier à considérer les Fungi comme un groupe à part, en raison de leur nutrition par absorption. Le schéma de Whittaker contient des incertitudes et des ambiguïtés dues au fait que l'auteur a choisi de ne pas former des groupes strictement monophylétiques mais plutôt de regrouper les êtres vivants en cinq grands « règnes », définis en fonction de caractéristiques écologiques ou de grades d'organisation (Adoutte et al., 1996). Les groupes sont représentés dans un empilement vertical, comme dérivant les uns des autres avec certaines lignées néanmoins qui se prolongent dans un autre groupe. Bien que sa classification fît naître de graves ambiguïtés en créant une classification non naturelle (c'est-à-dire non phylogénétique), elle connut néanmoins un important succès.

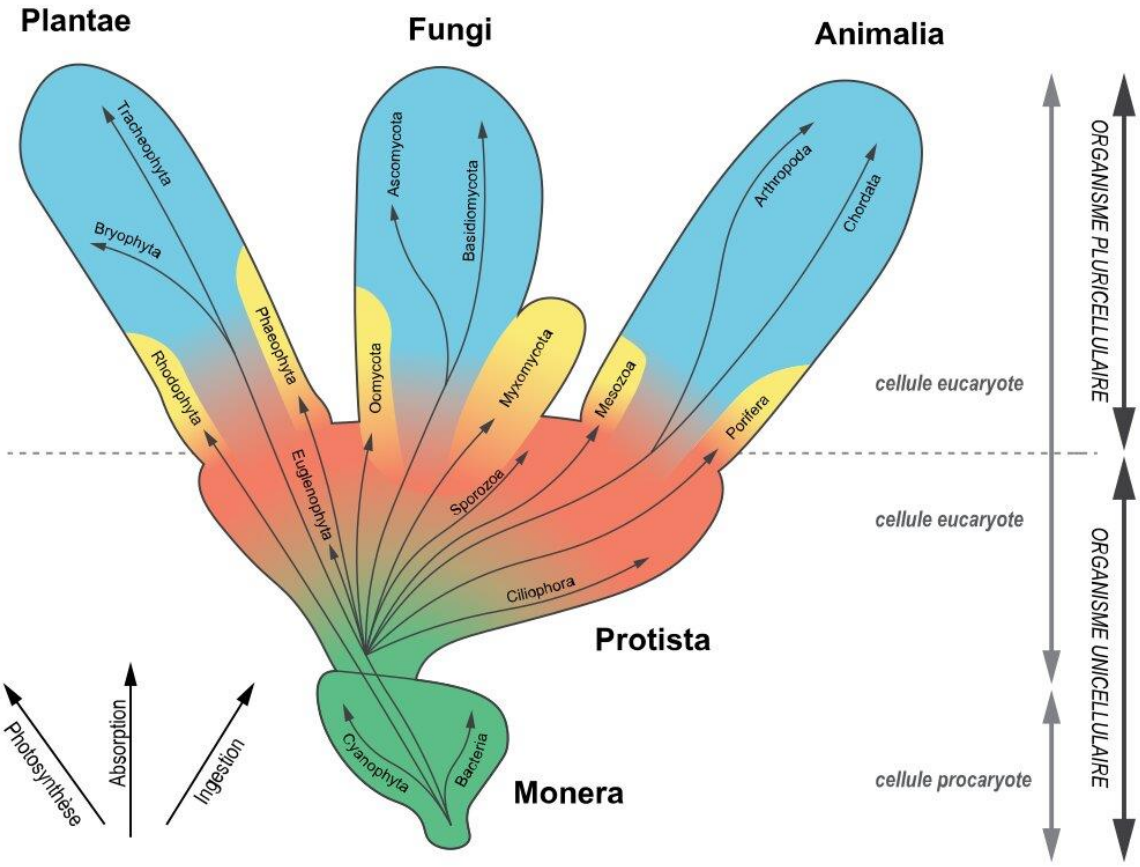


Figure 3. Système à 5 règnes selon Robert Whittaker (Whittaker, 1969).
En 1969, Robert Harding Whittaker, un éminent écologiste américain, a proposé une classification novatrice du vivant en cinq règnes : monera, protiste, fungi, plantes, animaux. Sa classification prend en compte le type et niveau d'organisation et la complexité cellulaire, ainsi que son type de nutrition. Il considère les monera comme les êtres les plus inférieurs à partir desquels se sont développées progressivement des formes de vie plus complexes, les protistes, puis les organismes multicellulaires (fungi, plantes et animaux).

2.4 … A LA REVOLUTION MOLECULAIRE

La révolution moléculaire commence en 1953 par la découverte de la structure en double hélice de l'ADN par Rosalind Franklin (Franklin & Gosling, 1953), repris par James Dewey Watson et Francis Crick (qui leur valut le prix Nobel de physiologie et de médecine en 1962) (Watson & Crick F H C, 1953). Crick introduit le dogme central de la biologie moléculaire (**Box 1**) en établissant le lien qui existe entre le matériel génétique (ADN et ARN) contenu dans la cellule et les protéines que cette cellule synthétise (Crick, 1968). Ainsi l'information génétique est transmise dans un sens unique : ADN → ARN → protéine. Il a toutefois mentionné une exception au dogme : les rétrovirus qui peuvent aller de l'ARN vers l'ADN grâce à une enzyme, la transcriptase inverse. À ce moment-là, il n'a pas pu étayer ses propos avec des preuves, mais elle s'est en effet avérée vraie.

Box 1. Le dogme central de la biologie moléculaire

Les cellules contiennent de nombreux constituants, dont les principaux sont l'eau, les ions inorganiques, des petites molécules et leurs précurseurs (sucres, acides aminés, nucléotides), des acides gras et leurs précurseurs, des macromolécules (polysaccharides, protéines, acides nucléiques…). Les protéines et les acides nucléiques sont de longues chaînes linéaires de petites molécules (polymères) : des acides aminés pour les protéines, des nucléotides pour les acides nucléiques. Parmi les acides nucléiques, on distingue deux catégories : l'acide désoxyribonucléique (l'ADN) et les acides ribonucléiques (les ARN). ADN et ARN sont formés d'un enchaînement de β-D-nucléosides 5'-phosphates. Dit autrement, il s'agit de quatre molécules appelées nucléotides monophosphates reliés entre eux par des liaisons 3'-5' phosphodiester. Les nucléotides sont formés par l'association d'un sucre (β-D-ribofuranose pour l'ARN, β-D-2'-désoxyribofuranose pour l'ADN), d'un acide phosphorique et d'une base hétérocyclique azotée. Les bases azotées sont des hétérocycles dérivant de la purine (adénine (A) et guanine (G)) ou de la pyrimidine (cytosine (C), thymine (T) et uracile (U)). Trois d'entre elles sont communes à l'ADN et à l'ARN : adénine, guanine et cytosine. La quatrième diffère selon le type d'acide nucléique : uracile pour l'ARN et thymine pour l'ADN (qui est en fait une forme méthylée de l'uracile).

Certaines de ces bases ont la particularité de s'apparier entre elles. Ainsi, la guanine s'apparie à la cytosine par trois ponts hydrogènes, tandis que l'adénine s'apparie avec la thymine (dans le cas de l'ADN) et l'uracile (dans le cas de l'ARN) par deux ponts hydrogènes. Néanmoins, les ARN possèdent une très grande diversité de nucléotides puisqu'on dénombre aujourd'hui une centaine de nucléotides modifiés. De plus, alors que l'ADN est présent sous forme de double brin apparié, les ARN sont (à de rares exceptions près, dont certains virus) simple brin. Sur un même brin d'ARN, les bases peuvent s'apparier et ainsi former des structures secondaires qui jouent un rôle central dans la fonction de tous les ARN, en leur imposant une structure tridimensionnelle dictée par la séquence des bases.

C'est l'ordre de ces nucléotides qui détermine, sous forme de code, toutes les informations nécessaires qui confèrent aux espèces leurs propriétés et aux individus leurs caractères uniques. L'information contenue dans l'ADN est organisée en unités, que l'on appelle des gènes. Le gène peut être considéré comme une séquence d'ADN qui spécifie la synthèse d'une chaîne polypeptidique (protéine) ou d'un acide ribonucléique (ARN) fonctionnels. Plus précisément, un gène est défini comme un segment nucléotidique d'ADN fonctionnel (ou ARN chez certains virus) contenant de l'information génétique qui détermine la structure d'une chaine polypeptidique (voire de plusieurs isomorphes) ou la structure d'un ARN. Un gène encode des molécules fonctionnelles (protéines ou ARN) et qui contiennent les signaux nécessaires à la régulation de son expression en fonction des conditions cellulaires. Ainsi, chaque gène correspond donc à une séquence de nucléotides qui détermine l'ordre d'enchaînement des acides aminés dans une protéine particulière, et donc la fonction de celle-ci.

Le code génétique établit les règles permettant de traduire ces informations pour produire des protéines. L'information portée par un gène est dans un premier temps copiée sous forme d'ARN, appelé ARN messager (ARNm) : c'est la transcription. Cet ARNm, à durée de vie temporaire, assure soit des fonctions structurales, soit des fonctions enzymatiques, soit des fonctions de transport de l'information génétique. Dans le dernier cas, celui-ci dirige ensuite la synthèse de protéines, via un complexe ribonucléoprotéique (c'est-à-dire composé d'ARN et de protéines) appelé le ribosome : c'est la traduction de l'information, où chaque triplet de nucléotides de l'ARNm induit l'incorporation

d'un acide aminé via un ARN de transfert (ARNt) complémentaire. Les ARN sont donc des intermédiaires fonctionnels facilitant l'expression de l'ADN sous forme de protéine. C'est ce que l'on appelle le flux d'information génétique, ou dogme central de la biologie moléculaire, tel que présenté par Francis Crick en 1958. L'ADN est alors considéré comme le support stable et transmissible de l'information génétique, qui définit les fonctions biologiques d'un organisme. Sa réplication garantit également une continuité génétique d'une génération à l'autre.
Toutefois, le dogme central ne prédit pas que l'on puisse revenir à l'ADN à partir d'un ARN. Or, il a été mis en évidence chez les rétrovirus et les rétrotransposons une enzyme ADN polymérase dépendante de l'ARN, la transcriptase inverse (ou reverse transcriptase), capable d'utiliser un brin d'ARN comme matrice et de catalyser sa rétrotranscription en un ADN complémentaire (ADNc). De plus, certains ARN ont également la capacité de s'auto-répliquer. En effet, certains virus à ARN se répliquent via un intermédiaire à ARN. Pour ce faire, ils encodent une ARN polymérase dépendante de l'ARN. Sans remettre totalement en cause le dogme central de la biologie moléculaire, cette découverte a apporté un nouveau regard sur les possibilités du vivant et ouvert de nouvelles perspectives de recherche concernant les origines du système génétique actuel.

Références

Crick, F.H.C. (1968) "The origin of the genetic code," *Journal of Molecular Biology*, 38(3), pp. 367–379.

En 1959, Frederick Sanger détermine que la double chaine polypeptidique de l'insuline est constituée d'un enchainement précis d'acides aminés, le conduisant à formuler l'idée que toutes les protéines sont constituées d'une séquence d'acides aminés qui leur est propre (Sanger, 1959). Il décroche une première fois le prix Nobel en 1958 pour ses travaux sur la structure des protéines. D'autre part, une méthode de séquençage de l'ADN porte son nom : la méthode de Sanger. Elle lui valut un deuxième prix Nobel en 1980. Aujourd'hui, sa technique est moins utilisée pour les études génomiques au profit du *High-Throughput Sequencing* (HTS) et *Third-generation sequencing* (TGS). Elle reste cependant encore largement répandue dans les approches de biologie moléculaire.

En 1964, Emile Zuckerkandl et Linus Pauling ont montré que la séquence des protéines contient une très grande quantité d'informations sur l'histoire évolutive ancienne (Zuckerkandl & Pauling, 1965). Elle est en effet modifiée par les mutations qui mènent d'un organisme ancestral à ses descendants. La séquence d'une même protéine diffère donc d'autant plus d'un organisme à l'autre que ces derniers ne sont pas étroitement apparentés. Cette découverte est à la base de la phylogénie, qui est l'étude des liens de parenté entre les êtres vivants (actuels et/ou disparus). Ils furent les pionniers de la comparaison d'acides aminés en étudiant la phylogénie des primates sur base des séquences d'hémoglobine, les conduisant en outre à introduire le concept d'« horloge moléculaire ». Ces auteurs ont également suggéré, avec raison, qu'il en allait probablement de même pour les séquences des acides nucléiques (ADN et ARN), techniquement impossibles à « lire » à cette époque. La phylogénie moléculaire était née (**Box 2**). Cette nouvelle discipline, qui reconstruisait les relations de parenté entre les organismes à partir de la comparaison des séquences, a bouleversé notre regard sur l'origine des phyla actuels. On tente ainsi, grâce aux gènes, de retracer l'histoire des organismes.

Box 2. Evolution moléculaire & phylogénie

Forces évolutives & variation moléculaire

Le biologiste autrichien Emile Zuckerkandl et le biochimiste américain Linus Pauling furent les premiers à se pencher sur les modalités de la variation moléculaire. Bien que très stable d'une génération à la suivante, le patrimoine génétique varie, comme en attestent les différences observées entre les génomes de diverses espèces. Comparer les séquences génétiques de deux espèces permet donc d'établir les relations de parenté entre elles. L'étude des variations de taux d'évolution moléculaire est un formidable outil pour mesurer l'évolution des génomes et des espèces. Historiquement, les premières phylogénies représentant l'évolution des espèces étaient fondées sur des caractères morphologiques. Mais de nos jours, les progrès de la génétique permettent de transcender cette approche en comparant leurs séquences protéiques ou nucléiques. Les caractères des individus (donc des espèces) sont liés aux gènes. Les variations des caractères, qui sont le matériel de base de l'évolution, sont liées aux variations des gènes. Les variations moléculaires sont observables entre individus d'une même espèce (polymorphisme) ou bien entre espèces distinctes (divergence). L'analyse des données de polymorphisme est du domaine de la génétique des populations, tandis que l'étude des divergences relève de la phylogénie.

Diverses forces évolutives génèrent et influencent les variations moléculaires. La première de ces variations peut être obtenue par de brusques changements survenant sur un ou plusieurs gènes : les mutations. A chaque génération, dans une population, un individu peut présenter, à une position donnée de son génome, un état différent de celui de ses parents. On dit qu'un nouvel allèle apparaît, créant un polymorphisme au locus correspondant. Divers types de mutations sont reconnaissables : des changements de nucléotides, des insertions ou délétions, des élongations/raccourcissements de motifs répétés, des duplications de gènes ou de segments génomiques, des transpositions et des inversions. Ces mutations peuvent être non-codantes, synonymes (la séquence codante est modifiée mais pas la protéine codée) ou non-synonymes (la protéine codée diffère). Les mutations entre purines (A <-> G) ou pyrimidines (C <-> T) sont appelées transitions, tandis que les mutations faisant passer d'une purine à une pyrimidine (ou réciproquement) sont appelées des transversions. Les mutations qui donnent aux individus un avantage reproductif se transmettent et se répandent, tandis que les autres disparaissent ; il s'agit de notre deuxième force évolutive : la sélection naturelle. Notre troisième force évolutive est la dérive génétique, correspondant aux variations aléatoires de fréquences alléliques, liées au hasard de la reproduction. La dérive sélective est la seule force qui s'applique aux locus neutres, c'est-à-dire dont les variations n'exercent aucune influence sur le succès reproducteur des individus. De plus, les gènes peuvent être plus ou moins liés physiquement sur les chromosomes. Cette liaison physique entraîne une liaison génétique, c'est-à-dire que certains allèles à différents locus sont préférentiellement associés entre eux, une combinaison d'allèles formant ce que l'on appelle un haplotype. De ce fait, les événements sélectifs se produisant sur différents gènes voisins vont interférer entre eux. La recombinaison, et plus précisément les crossing-over qu'elle engendre, est le processus qui permet de « casser » ces associations entre gènes, et ainsi rendre leur évolution moins dépendante les uns des autres. Enfin, la démographie et la structure des populations vont également jouer un rôle dans les variations moléculaires à l'échelle intraspécifique. Les effets démographiques (goulot d'étranglement, balayage sélectif, migration, panmixie…) s'appliquent à l'ensemble du génome, tandis que les effets sélectifs restent localisés.

La phylogénie dite « moléculaire » présente de nombreux avantages par rapport à la comparaison de caractères morphologiques : (i) d'abord le nombre d'états de caractère est fixe (quatre nucléotides et vingt acides aminés) et leur universalité fait que ceux-ci peuvent être comparés à travers tous les organismes vivants ; (ii) de plus, l'évolution des protéines et des séquences nucléotidiques suit un patron plus ou moins régulier qui permet l'utilisation de modèles mathématiques pour formaliser leurs changements ; (iii) enfin, les génomes de tous les organismes sont composés de très longues séquences nucléiques dont on peut déduire les séquences protéiques, fournissant une immense quantité d'informations surpassant celles fournies par les caractères morphologiques.

Recherche de caractères homologues

Afin d'évaluer les ressemblances moléculaires (et donc déterminer les relations de parenté entre les espèces), il convient en phylogénie de travailler sur des séquences dites homologues, c'est-à-dire héritées d'un ancêtre commun. Après avoir aligné des séquences pour tenir compte des événements d'insertion et de délétion de nucléotides, il est possible d'évaluer leur ressemblance. Plus deux séquences se ressemblent, plus il y a de chances qu'elles partagent un ancêtre commun proche. La recherche d'homologie avec le programme BLAST donne également une E-value, qui est l'espérance mathématique. Cette E-value correspond au nombre de séquences d'une banque de données de même taille et même composition qui s'aligneraient par chance avec la séquence de référence (requête)avec un score de similarité supérieur ou égal au score obtenu.

On peut distinguer trois grands types de gènes homologues :

- les gènes orthologues (du grec *ortho*, "droit"), présents dans des espèces différentes et dont la divergence remonte aux spéciations qui les ont produites. Ainsi les gènes α dérivés de la copie α chez l'ancêtre commun, de même pour les gènes β dérivés de la copie β. Les gènes orthologues reflètent l'histoire des spéciations entre les espèces. Ce sont eux que l'on va utiliser dans les phylogénies moléculaires ;
- les gènes paralogues (du grec *para*, « en parallèle »), issus de duplications génétiques au sein d'une même espèce. Par conséquent, de nombreux exemplaires de ces gènes divergeant les uns des autres peuvent exister pour cette espèce. Ainsi les copies α et β d'un même gène chez un même organisme sont paralogues. On distingue également trois cas de paralogie :
    - in-paralogie : deux séquences paralogues au sein d'une espèce sont in-paralogues si l'événement de duplication a eu lieu après leur spéciation ;
    - out-paralogie : deux séquences paralogues au sein d'une espèce sont out-paralogues si l'événement de duplication a eu lieu avant leur spéciation ;
    - ohnologie : deux séquences paralogues sont ohnologues si elles résultent d'un événement de duplication complète du génome. Le terme ohnologue a été proposé par Ken Wolfe en hommage à Susumu Ohno.
- les gènes xénologues (du grec *xenos*, "étranger"), représentant des transferts de matériel génétique en provenance d'une autre espèce (on parle de transfert horizontal de gènes) et pour lequel le lien de parenté n'est pas direct. De tels gènes ne peuvent être utilisés pour retracer la phylogénie des espèces.

L'inférence de la phylogénie des espèces, dans ce contexte hérité de la morphologie de comparaison

de caractères homologues, requiert l'analyse de gènes strictement orthologues, seuls témoins des événements de spéciation. L'insertion de gènes paralogues et/ou xénologues dans ce type d'analyse est à éviter à tout prix car ils faussent le résultat de l'inférence.

Références

Owen, R. 1843. Lectures on the comparative anatomy and physiology of the invertebrate animals. Longman, Brown, Green & Longmans.
Walter M. Fitch, Homology: a personal view on some of the problems, Trends in Genetics, Volume 16, Issue 5, 2000, Pages 227-231, ISSN 0168-9525, https://doi.org/10.1016/S0168-9525(00)02005-9.
Wolfe, K. (2000). Robustness—it's not where you think it is. Nature Genetics, 25, 3-4.
Nei, M., et S. Kumar. 2000. Molecular Evolution and Phylogenetics. Oxford University Press, New York.
Gray, G. S., et W. M. Fitch. 1983. Evolution of antibiotic resistance genes: the DNA sequence of a kanamycin resistance gene from Staphylococcus aureus. Mol Biol Evol 1:57-66.
Boussau B, Daubin V. Genomes as documents of evolutionary history. Trends Ecol Evol. 2010 Apr;25(4):224-32. doi: 10.1016/j.tree.2009.09.007. Epub 2009 Oct 31. PMID: 19880211.
Turelli M, Barton NH, Coyne JA. Theory and speciation. Trends Ecol Evol. 2001 Jul 1;16(7):330-343. doi: 10.1016/s0169-5347(01)02177-2. PMID: 11403865.
Barraclough, Timothy Giles and Sean Nee. "Phylogenetics and speciation." Trends in ecology & evolution 16 7 (2001): 391-399.
Nichols R. Gene trees and species trees are not the same. Trends Ecol Evol. 2001 Jul 1;16(7):358-364. doi: 10.1016/s0169-5347(01)02203-0. PMID: 11403868.
Schluter D. Ecology and the origin of species. Trends Ecol Evol. 2001 Jul 1;16(7):372-380. doi: 10.1016/s0169-5347(01)02198-x. PMID: 11403870.

Durant les années 1980, au fur et à mesure des progrès technologiques du séquençage, un arbre universel du vivant a été progressivement élaboré, basé sur la séquence d'un acide nucléique particulier, l'ARN de la petite sous-unité du ribosome, responsable de la synthèse protéique. Cette molécule a été choisie pour plusieurs raisons. Tout d'abord parce que le ribosome est universel : on le trouve à des milliers d'exemplaires dans toutes les cellules, ce qui permet d'établir les liens de parenté de tous les êtres vivants (cela exclut les virus qui en sont dépourvus). Ensuite, parce que la taille de cet ARN est suffisante pour fournir de l'information tout en restant « manipulable ». Depuis les années 2000, notre connaissance de la diversité microbienne a été bouleversée grâce au développement des techniques de séquençage et la possibilité de séquencer de l'ADN environnemental (métagénomique). La possibilité de séquencer massivement tout l'ADN présent dans un environnement donné à moindre coût grâce à la technique du séquençage à haut débit a permis de reconstituer, par des méthodes informatiques, les génomes d'organismes (appelés MAG pour *metagenome-assembled genomes*) que l'on n'avait encore jamais réussi à cultiver.

Certains iront plus loin en refusant le modèle d'un arbre du vivant, remettant alors en cause les classifications phylogénétiques. Ainsi, en 1998, Ford Doolittle suggère que les gènes bactériens présents chez les eucaryotes ont été acquis via la nourriture (théorie de *you are what you eat* = « tu es ce que tu manges ») (Doolittle, 1998). Ses travaux démontrent l'importance des transferts horizontaux de gènes entre les êtres vivants. Pour lui, aucune classification

hiérarchique du vivant n'est valable. Les classifications phylogénétiques ne reflètent pas l'histoire des organismes mais celle des gènes (Doolittle, 1999; Hirt et al., 1999).

3 LE PARADIGME DE WOESE

3.1 PREMIERES PHYLOGENIES MOLECULAIRES ET DECOUVERTE D'UN NOUVEAU DOMAINE : LES « ARCHEOBACTERIES »

Quand il démarre ses recherches en 1969, Carl Woese décide de comparer la séquence de la petite sous-unité ribosomique de l'ARN ribosomique (SSU rRNA) des bactéries (que l'on nomme ARN-16S) connues à l'époque (G. Fox et al., 1977). Or, à cette époque, on ne sait pas encore séquencer rapidement de l'ADN ou de l'ARN. Pour contourner ce problème, Woese utilise une méthode développée par Frederick Sanger consistant à réaliser une « empreinte génétique » des molécules de SSU rRNA en les coupant de façon reproductible en une centaine de petits fragments (ou oligonucléotides) et à quantifier la divergence entre deux espèces A et B grâce à un calcul de coefficient de similarité S_{AB} (G. E. Fox et al., 1977). Concrètement, il cultive en présence d'un isotope radioactif du phosphore les organismes qu'il souhaite analyser afin d'isoler leur SSU rRNA. Puis, après digestion de l'ARN avec ribonucléase T1 qui clive la molécule après les résidus G, il génère des catalogues d'oligonucléotides qui sont composés d'une certaine collection de U, C et A avec un seul résidu G (en position terminale). Les différents oligonucléotides obtenus sont ensuite séparés sur une feuille de papier selon leurs propriétés chimiques par chromatographie en deux dimensions. La radioactivité fait apparaître sous forme de taches les différents fragments d'ARN : c'est l'empreinte génétique, propre à chaque espèce. Chaque oligonucléotide présent sur la feuille peut être élué et la séquence identifiée. Beaucoup de motifs de petite taille sont présents chez toutes les bactéries et ne sont donc pas pertinents pour les analyses. Woese décide dès lors de ne considérer comme pertinents que les oligonucléotides d'au moins six nucléotides. Par conséquent, plus deux espèces ont un nombre important de grands oligonucléotides en commun, plus elles seront apparentées. Les oligonucléotides résultants pouvant être séquencés, il a ensuite comparé les séquences et les fréquences de ces oligonucléotides pour définir une distance phylogénétique, *PD value* (*Phylogenetic Distance value*), entre les différentes paires d'organismes étudiés. Woese détermine, pour toutes les espèces prises deux à deux, leur coefficient de similarité S_{AB} selon la formule : $S_{AB} = 2\ N_{AB}/(N_A + N_B)$, avec N_{AB} le nombre d'oligonucléotides communs aux deux organismes et N_A et N_B les nombres d'oligonucléotides dans les séquences de taille 6 ou plus. Ainsi, le S_{AB} vaut 1 pour deux espèces identiques et tend vers 0 pour deux espèces fortement éloignées (G. Fox et al., 1977; G. E. Fox et al., 1977). 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 (**Tableau 1**).

Lorsque George Fox (C. R. Woese & Fox, 1977), un étudiant de Woese, analyse les coefficients de similarité des bactéries prises deux à deux, il s'aperçoit qu'un groupe de bactéries n'est pas plus apparenté aux autres bactéries qu'il ne l'est aux eucaryotes : il s'agit des bactéries dites « méthanogènes ». Ces résultats révèlent donc l'existence de non pas deux mais trois types de SSU rRNA, et donc de trois types de ribosomes dans le monde du vivant. C'est ainsi qu'émerge avec Woese la vision d'une vie à 3 domaines : bactéries, eucaryotes et archaea. Des études ultérieures montrèrent qu'on pouvait encore rattacher au moins deux autres groupes déjà connus

à ce nouveau domaine : les bactéries halophiles (Magrum et al., 1978) et les thermo-acidophiles (C. Woese et al., 1978).

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	—

Tableau 1. Table représentant les coefficients d'association (S_{AB}) entre les séquences de SSU rRNA de 13 représentants des trois domaines primaires (C. R. Woese & Fox, 1977).

Cette classification revêt une importance toute particulière d'un point de vue historique, car c'est la première fois que les scientifiques ont accès à une classification basée sur des critères phylogénétiques objectifs et mesurables, en accord avec la théorie de l'évolution telle que décrite par Darwin. Elle montre deux choses : tout d'abord que la classification doit se faire selon un système naturellement logique, le plus évident étant celui de l'histoire évolutive de la Vie, et que la phylogénie moléculaire est la méthode la plus efficace permettant d'accomplir cette tâche.

3.2 A LA RECHERCHE DE LA RACINE DE L'ARBRE UNIVERSEL DU VIVANT

Une solution pour comprendre l'évolution de la cellule eucaryote et relater son histoire est d'établir ses relations de parenté avec les autres domaines du vivant. Ces relations entre les trois domaines du vivant sont très débattues. Lequel est le groupe frère des deux autres ? Les trois domaines sont-ils mono-, para- ou polyphylétiques ? Où est située la racine de l'arbre du vivant ? Ces questions demeurent encore sans réponse évidente du fait de la difficulté de polariser les caractères moléculaires. Ainsi, il est difficile de déterminer si un état de caractère donné est ancestral ou dérivé.

On nomme LUCA (*Last Universal Commun Ancestor*) le dernier ancêtre commun à tous les organismes actuels. La « racine » d'un arbre phylogénétique désigne le nœud le plus ancien d'une phylogénie donnée. Par définition, la racine est également une représentation de l'ancêtre de l'ensemble des taxons étudiés, donnant ainsi une direction à l'ensemble des processus de transformation utilisés pour reconstruire la phylogénie. La méthode la plus répandue permettant d'enraciner l'arbre de la vie est d'utiliser des gènes paralogues universels, c'est-à-dire des gènes ayant subi une duplication ancestrale à LUCA, si bien qu'ils sont présents dans l'ensemble des trois domaines du vivant sous la forme d'au moins deux copies (**Figure 4**). L'idée est la suivante : prenons un gène ancestral X. Il subit au cours du temps une duplication menant à deux copies : une copie α et une copie β. On obtient ainsi deux enregistrements des processus évolutifs au cours du temps. Il suffit alors d'établir la phylogénie d'une des copies et de l'enraciner avec la phylogénie de l'autre copie afin d'obtenir un arbre en miroir que l'on replie sur lui-même afin de mettre en évidence le groupe externe. Habituellement, le choix d'un groupe externe pour enraciner un arbre se fait sur une sélection de taxons extérieurs au groupe. Or ici, l'enracinement dans un groupe

externe se fait alors sur base d'une des copies du paralogue. Cette technique fut d'abord employée avec des gènes de la ferrédoxine. En effet, chaque gène de ferrédoxine est composé de deux séquences paralogues attachées entre elles, de manière que la phylogénie de chaque moitié de gène peut permettre d'enraciner la phylogénie de l'autre moitié (Schwartz & Dayhoff, 1978). Cependant, la ferrédoxine est une protéine très courte et seulement une quinzaine de positions conservées avaient été utilisées, ce qui est très insuffisant pour construire une phylogénie des organismes fiable.

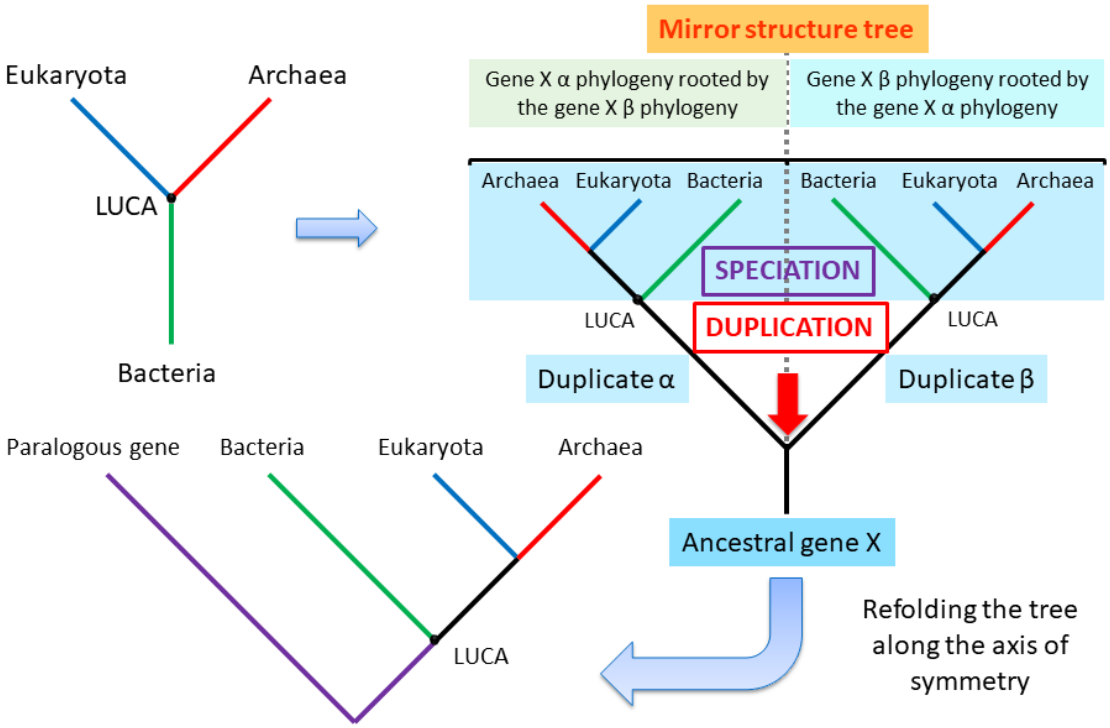


Figure 4. Méthode d'enracinement d'un arbre phylogénétique via des gènes paralogues.
La méthode des gènes paralogues est utilisée pour enraciner un arbre phylogénétique en se basant sur des gènes homologues qui ont subi une duplication au cours de l'évolution. Des séquences de gènes paralogues (issus de duplication) sont sélectionnées chez différentes espèces. Après alignement et construction d'un arbre phylogénétique, on peut identifier le gène paralogue ancestral à partir duquel ont eu lieu les duplications en enracinant la phylogénie de cette copie avec la phylogénie de l'autre copie.

C'est via cette méthode qu'au tournant des années 1990, les premiers arbres phylogénétiques du vivant furent enracinés dans les bactéries (Gogarten et al., 1989; Iwabe et al., 1989; C. R. Woese et al., 1990) en utilisant les ATPases et les facteurs d'élongation. Malheureusement, on sait aujourd'hui que ces arbres sont faux (Philippe & Forterre, 1999).

Par la suite, cette méthode fut utilisée sur d'autres gènes paralogues universels (**Tableau 2**). Le problème est que seules quelques familles de gènes ancestralement dupliqués sont connues comme étant de potentiels marqueurs phylogénétiques, et leurs analyses ont mené à des conclusions variées et sujettes à controverses.

Localisation de la racine	Marqueur phylogénétique utilisé	Référence(s)
Avant la trifurcation originale	Caractéristiques diverses	(C. Woese et al., 1978)
Sur la branche menant aux bactéries	Sous-unités régulatrices et catalytiques des ATPases de type V et F	(Gogarten et al., 1989)
Sur la branche menant aux bactéries	Protéines de facteurs d'élongation de la traduction, EF-tu/1 et EF-G/2	(Iwabe et al., 1989)
Sur la branche menant aux bactéries	Val/Ile amino-acyl tRNA synthétases	(J. Brown & Doolittle, 1995)
Sur la branche menant aux bactéries	Duplication interne dans le CPS (carbamoylphosphate synthétase)	(Lawson et al., 1996)
Sur la branche menant aux bactéries	Tyr/Trp amino-acyl tRNA synthétases	(J. R. Brown et al., 1997)
Résultats non concluants : trop faible signal phylogénétique, incongruence des gènes *his* individuels	Gènes de biosynthèse de l'histidine hisA/hisF	(Charlebois et al., 1997)
Sur la branche menant aux bactéries	Particule de reconnaissance de signal SRP54 et son récepteur SRα	(Gribaldo & Cammarano, 1998)
Sur la branche menant aux bactéries	Aspartate et Ornithine transcarbamoylases	(Labedan et al., 1999)
Sur la branche menant aux eucaryotes	Facteurs d'élongation de la traduction, EF-1α et EF-2	(Lopez et al., 1999; Philippe & Forterre, 1999)
Sur la branche menant aux bactéries, mais probablement dû à un artéfact d'attraction des longues branches	Facteurs d'élongation, ATPases, ARNt synthétases, CPS, protéines SRP	(Philippe & Forterre, 1999)
Monophylie des procaryotes avec les sites lents vs archées groupe frère des eucaryotes avec les sites rapides	Particule de reconnaissance de signal SRP54 et son récepteur SRα	(Brinkmann & Philippe, 1999)
Au sein des bactéries Gram-négatives	Architecture de la paroi bactérienne	(Cavalier-Smith, 2002)
Résultats non concluants	Structure secondaire de l'ARN	(Caetano-Anollés, 2002)
Difficultés conceptuelles	Non applicable	(Bapteste & Brochier, 2004)

Tableau 2. Différents points de vue concernant la localisation de la racine de l'arbre universel de la vie (adapté de (Zhaxybayeva et al., 2005)).

De plus, l'histoire d'un gène ne reflète pas nécessairement l'histoire d'une espèce. La découverte de transferts horizontaux de gènes a montré que certains gènes étaient inutilisables pour reconstruire des phylogénies, différents gènes pouvant raconter des histoires différentes.

3.3 LES 3 DOMAINES DU VIVANT : BACTERIES, EUCARYOTES ET ARCHEES

De nos jours, la classification des êtres vivants se fait via l'utilisation de caractères héritables, qu'ils soient visibles (critères anatomiques, embryologiques, morphologiques) ou non visibles (séquences d'ADN, d'ARN et de protéines). Elle fait l'objet d'une discipline à part entière, la phylogénie, qui est l'étude des relations de parenté entre les êtres vivants afin de comprendre leur histoire évolutive (la phylogenèse) (Haeckel, 1866). Cette discipline peut également prendre en compte des données issues de la paléontologie.

L'arbre que publie Woese en 1990, classant le vivant en trois domaines (archées, bactéries, eucaryotes) (C. R. Woese et al., 1990) (**Figure 5**) bouleverse la vision du vivant. Son arbre avait été enraciné dans la branche des bactéries, sur base de l'étude de gènes paralogues (Facteur d'Elongation et ATPases). Les archées et les eucaryotes y sont apparentés et partageraient un ancêtre commun. Le point d'embranchement basal représente la position du dernier ancêtre commun, que Woese appelle le « Progénote » (CarlR. Woese & Fox, 1977). Il le définit comme « *une entité ayant un lien rudimentaire, imprécis, entre son génotype et son phénotype* » (Doolittle & Brown, 1994; CarlR. Woese & Fox, 1977). Cependant, l'enracinement de l'arbre de la vie demeure encore un problème ouvert. Placer les bactéries en groupe externe n'est peut-être que la conséquence d'un artefact phylogénétique dû à l'attraction des longues branches (Philippe et al., 2000) (**Box 7**). C'est pourquoi les relations phylogénétiques entre les trois domaines de la vie étant incertaines (Philippe & Forterre, 1999), il est préférable de travailler sur des arbres universels dépourvus de racine.

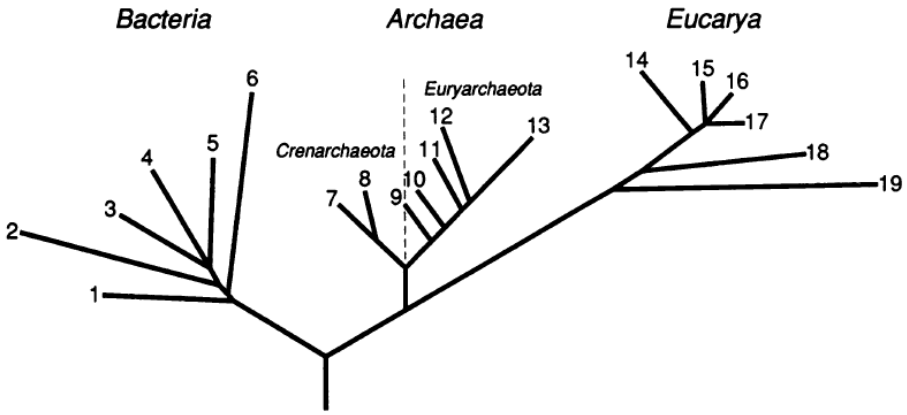


Figure 5. Arbre enraciné original de Woese publié en 1990 (C. R. Woese et al., 1990).
Cet arbre est basé sur l'analyse des gènes SSU rRNA. La position de la racine a été déterminée en comparant les séquences des paires de gènes paralogues des Facteurs d'Elongation et des ATPases ayant divergé avant l'émergence de LUCA (cf. **Tableau 2**). **Cette classification en trois domaine (bactéries, archées, eucaryotes) a bouleversé notre compréhension de la diversité du vivant.**

3.4 L'HYPOTHESE ARCHEZOA

Au sein des arbres des eucaryotes basés sur les ARNr et les Facteurs d'Elongation, un groupe d'organismes unicellulaires (et souvent parasites) simples et sans organites apparaît à la base. Ce groupe rassemble les Métamonades (ex. *Giardia intestinalis*), les Microsporidies (ex. *Encephalitozoon cuniculi*), les Parabasaliens (ex. *Trichomonas vaginalis*) et les Archéamibes (ex. *Entamoeba histolytica*). Thomas Cavalier-Smith les avait regroupés dans un règne à part sous le nom d'*Archézoa* (Cavalier-Smith, 1993), en les considérant comme des organismes ayant divergé

du reste des eucaryotes avant la capture d'une protéobactérie transformée par la suite en mitochondrie : c'est l'hypothèse Archezoa. La théorie de l'évolution des eucaryotes postulait alors que (1) les premiers eucaryotes étaient amitochondriés et vivaient en milieu anaérobie, puis par la suite (2) l'émergence d'un milieu aérobie aurait favorisé l'apparition et la diversification de lignées avec mitochondries capables d'utiliser l'oxygène pour la respiration (T. M. Embley & Martin, 2006; Martin Embley, 2006). En effet, l'oxygène est un produit très toxique pour les organismes anaérobies et cause l'extermination de la plupart de ceux qui ne savent pas s'en débarrasser. Or, les mitochondries sont en quelque sorte l'organe de résistance à l'oxygène d'une cellule, transformant celui-ci en eau par respiration. Les Archézoa se seraient donc différenciés plus tôt au cours de l'évolution des eucaryotes et étaient considérés comme des reliques d'un ancien monde dépourvu d'oxygène, vivant en conditions anaérobies avant l'Evénement de Grande Oxydation (2,4 Ga) (Bekker et al., 2010; Hannah et al., 2004; Kump, 2008; Lyons et al., 2014). Ils représentaient un stade « primitif » de l'évolution des eucaryotes, correspondant à des niveaux d'organisation ancestraux sur le chemin de la complexification progressive de la cellule eucaryote. L'origine de la mitochondrie était alors interprétée comme la conséquence de l'augmentation du taux d'oxygène, provoquant une crise chez les cellules dépourvues de mitochondries, et assurant la survie par sélection naturelle des cellules avec mitochondries capables de détoxifier de l'oxygène (Kurland & Andersson, 2000; Vellai et al., 1998).

Cependant, plusieurs observations viennent contredire ce scénario :

1. Des reconstructions phylogénétiques effectuées à partir d'autres gènes que l'ARNr (ex. α-tubuline (Keeling & Doolittle, 1996), grande sous-unité de l'ARN polymérase II (RBP1) (Hirt et al., 1999) positionnent dans les arbres les Archézoa au sein de diverses lignées eucaryotes, les rendant donc polyphylétiques. Autrement dit, des gènes différents racontent des histoires différentes : c'est ce que l'on appelle l'incongruence phylogénétique. Certains Archézoa ont donc une origine plus récente que ce que l'on pensait jusqu'alors, en particulier les Microsporidies aujourd'hui rattachées aux Fungi (Hirt et al., 1999; James et al., 2013; Keeling, 1998; Keeling & Doolittle, 1996).
2. Des analyses génétiques ont mis en évidence que beaucoup d'Archézoa possèdent des gènes d'origine mitochondriale (ex. *Heat shock proteins*) (Bui et al., 1996; T. M. Embley & Hirt, 1998; Germot et al., 1996, 1997; Hirt et al., 1997; Horner et al., 1996; Keeling & Doolittle, 1997; Peyretaillade et al., 1998; Roger, 1999; Roger et al., 1996), voire des mitochondries vestigiales dépourvues d'ADN (hydrogénosomes et mitosomes), prouvant dès lors qu'ils avaient, au cours de l'évolution, possédé une mitochondrie. Ainsi, on sait aujourd'hui que l'apparente absence de mitochondries typiques de ces eucaryotes est la conséquence d'une importante réduction (et très rarement d'une perte) secondaire, sans que cela ne traduise ni un réel état ancestral ni un apparentement entre eux (Vacek et al., 2018). On distingue ainsi deux types de réduction mitochondriale : l'hydrogénosome (réalisant une fermentation productrice de H2, par exemple chez certains Ciliés) et le mitosome dont la fonction est longtemps demeurée cryptique (M. Embley et al., 2003; Hjort et al., 2010; B. A. P. Williams et al., 2002) mais qui joue un rôle dans l'assemblage de clusters Fe-S (Karnkowska et al., 2016). En effet, ce dernier ne présente aucune des protéines impliquées dans d'autres fonctions mitochondriales majeures (respiration aérobie, biosynthèse de l'hème) et ils présentent des protéines nécessaires à la

biosynthèse des clusters Fe-S (comme la frataxine, la cystéine désulfurase, Isu1 et une Hsp70 mitochondriale) (Vacek et al., 2018). Dès lors, il semblerait que la symbiose mitochondriale ait pré-daté tous les Archézoa, et donc tous les eucaryotes actuellement connus. Par conséquent, l'adaptation des eucaryotes à des milieux anaérobies a eu lieu plusieurs fois indépendamment et rien ne prouve qu'ils soient apparus dans de telles conditions (Martin Embley, 2006). Cela montre simplement que l'ancêtre commun des eucaryotes actuels possédait déjà des mitochondries (Sogin, 1997) et devait donc probablement être aérobie.

3. L'attraction des longues branches est responsable de la position artéfactuelle des Archézoa dans l'arbre des eucaryotes. L'homoplasie (accumulation de substitutions convergentes) peut affecter seulement certaines séquences divergentes, conduisant au phénomène d'attraction des longues branches, en particulier pour les lignées anciennes ou évoluant rapidement. Les espèces qui, pour le gène considéré, évoluent plus vite que les autres, présentent des séquences très différentes de celles de leurs proches parents. En conséquence, elles apparaissent plus éloignées de ceux-ci qu'elles ne le sont en réalité. Si elles sont non détectées, ces multiples substitutions génèrent un signal non-phylogénétique qui empêche la reconstruction correcte des relations de parenté (Forterre & Philippe, 1998). En effet, la saturation en mutations à une position donnée efface le signal phylogénétique et laisse la place à des biais convergents, tels que la composition en bases, ce qui mène à l'erreur systématique.
4. Les études isotopiques indiquent que des environnements anoxiques ont persisté localement et globalement ces 2 derniers Ga. Bien qu'il soit généralement admis que l'oxygène est apparu pour la première fois dans l'atmosphère il y a 2 Ga (Kump, 2008; Lyons et al., 2014), on pense que jusqu'à environ –600 Ma, les océans se trouvaient dans un état d'oxydation intermédiaire avec des eaux de surface oxygénées (où la photosynthèse avait lieu) et des eaux profondes anoxiques et riches en sulfures (Poulton et al., 2004; Shen et al., 2003). L'oxygène n'est pas apparu brutalement au cours des temps géologiques, son abondance augmenta de manière progressive et était liée à l'évolution ou à l'adaptation des micro-organismes (Fenchel & Finlay, 1995). Les premiers « aérobes » étaient des microaérophiles pouvant se développer sous de très faibles pressions d'oxygène, dans une atmosphère saturée entre 0,1 et 1 % en oxygène. Par conséquent, l'événement de Grande Oxydation de l'atmosphère doit être découplé des environnements marins anoxiques, où les eucaryotes anaérobies vivant en marge d'un monde oxique auraient pu prospérer, comme ils le font encore aujourd'hui. Les mitochondries n'augmentent même pas la fréquence respiratoire : gramme pour gramme, de nombreux procaryotes respirent plus vite que les eucaryotes (Lane & Martin, 2010).

La phylogénie des eucaryotes semble donc moins facile à cerner que ce que l'on pensait à l'époque des Archézoa. De nombreuses convergences et/ou pertes ont eu lieu, rendant hasardeuses les conjectures sur la nature de leur ancêtre commun. Cette vision de l'évolution de la cellule eucaryote était influencée par l'hypothèse que la vie évoluait depuis les formes les plus « simples » ou « primitives » vers des formes plus « complexes » ou « évoluées », induisant une notion fausse de progrès continu dans l'évolution. Cette vision finaliste de l'évolution (cf. scalisme) a été abandonnée depuis qu'il a été démontré que de nombreuses formes de vie (notamment parasites) sont issues de simplifications secondaires. Il est ainsi acquis de nos jours que tous les eucaryotes actuels connus ont évolué à partir d'un ancêtre relativement complexe,

pourvu d'une mitochondrie (A. M. Poole & Penny, 2007; A. Poole & Penny, 2007) et qui vivait en milieu aérobie. Cet eucaryote est couramment appelé LECA (*Last Eukaryotic Common Ancestor*). Certaines lignées ont par la suite subi une simplification secondaire (ex. pertes d'introns et d'une partie du *spliceosome*, « mitosomisation » de la mitochondrie). Mais si LECA était déjà un organisme complexe, était-ce le cas des premiers eucaryotes ? La mitochondrie marque-t-elle la première étape de l'évolution des eucaryotes ou, au contraire, des eucaryotes sans mitochondries (de type Archézoa) existaient-ils déjà auparavant ? Quelle influence les facteurs du milieu ont-ils eu sur leur apparition ? Sont-ils apparus en conditions aérobies ou anaérobies ? Les données que nous avons aujourd'hui ne nous permettent pas de remonter plus loin que LECA. Or, il est fort possible que des proto-eucaryotes aient existé, sans qu'ils ne possèdent toutes les propriétés des eucaryotes actuels. L'énigme reste entière…

4 L'ORIGINE DE LA CELLULE EUCARYOTE

Plusieurs scenarii permettent d'expliquer l'origine de la cellule eucaryote et ses relations de parenté avec les deux domaines procaryotes mais ne s'accordent pas sur la nature de l'hôte et le nombre de partenaires impliqués dans l'origine de la cellule eucaryote. Sans être exhaustif, les modèles considérés ici visent plutôt à généraliser les principales caractéristiques d'un grand nombre de modèles (A. M. Poole & Penny, 2007; A. Poole & Penny, 2007). Ces scenarii peuvent être regroupés en plusieurs catégories, selon le statut de LUCA (ADN ou ARN) (Forterre, 2007), le nombre de domaines primaires de la vie (deux domaines vs trois domaines) (Forterre, 2011; Gribaldo et al., 2010) ou encore selon la nature du premier eucaryote (avec ou sans mitochondrie) (T. M. Embley & Martin, 2006) (**Figure 6**).

Si l'on considère un scénario à trois domaines, alors la lignée eucaryote ancestrale (groupe souche) ne correspond pas aux eucaryotes modernes, mais à une ancienne lignée conduisant à LECA. Cette lignée peut être aussi ancienne, voir plus ancienne, que les bactéries et archées. Il n'est alors pas exclu que des eucaryotes amitochondriés (groupe souche) aient existé avant l'endosymbiose de la bactérie. Si l'on applique la théorie de la coalescence aux eucaryotes actuels, c'est aussi à lui que nous arrivons. Or, il est possible que d'autres lignées eucaryotes postérieures à LECA aient existé, mais sans donner de descendance actuelle (**Box 2**). Par conséquent, ce manque d'informations rend difficile les inférences sur l'identité de LECA. De plus, il est possible que LECA ne soit qu'un seul des multiples représentants d'une parmi plusieurs lignées ayant pu émerger à partir de FECA. Dans notre cas, nous considérerons les eucaryotes dans leur ensemble, en incluant les lignées éteintes depuis FECA.

En revanche, les partisans des scenarii de fusion à l'origine des eucaryotes jouent souvent sur la confusion entre les eucaryotes au sens large (incluant les lignées éteintes et vivant avant l'acquisition des mitochondries) et les eucaryotes modernes appartenant aux lignées actuelles (Forterre, 2011). Dans ces scenarii, les eucaryotes sont récents car issus de la fusion d'une archée et d'une bactérie. Dès lors, LUCA est un procaryote et FECA n'apparaît seulement qu'après la diversification des eubactéries et des archées. L'hypothèse de fusion est basée sur le fait que les eucaryotes et les archées partagent une machinerie cellulaire similaire (traduction, transcription, réplication) qui est très différente (parfois non homologue) de celle des bactéries (Golding & Gupta, 1995) (cf. Paradoxe de Janus de Lake (Lake, 2007)) tandis que les gènes métaboliques sont d'origine bactérienne. Les scenarii de fusion proposés impliquent divers types d'associations

(symbiose, syntrophie…) souvent justifiées par des considérations métaboliques contemporaines pour expliquer l'association de deux partenaires (López-García & Moreira, 1999).

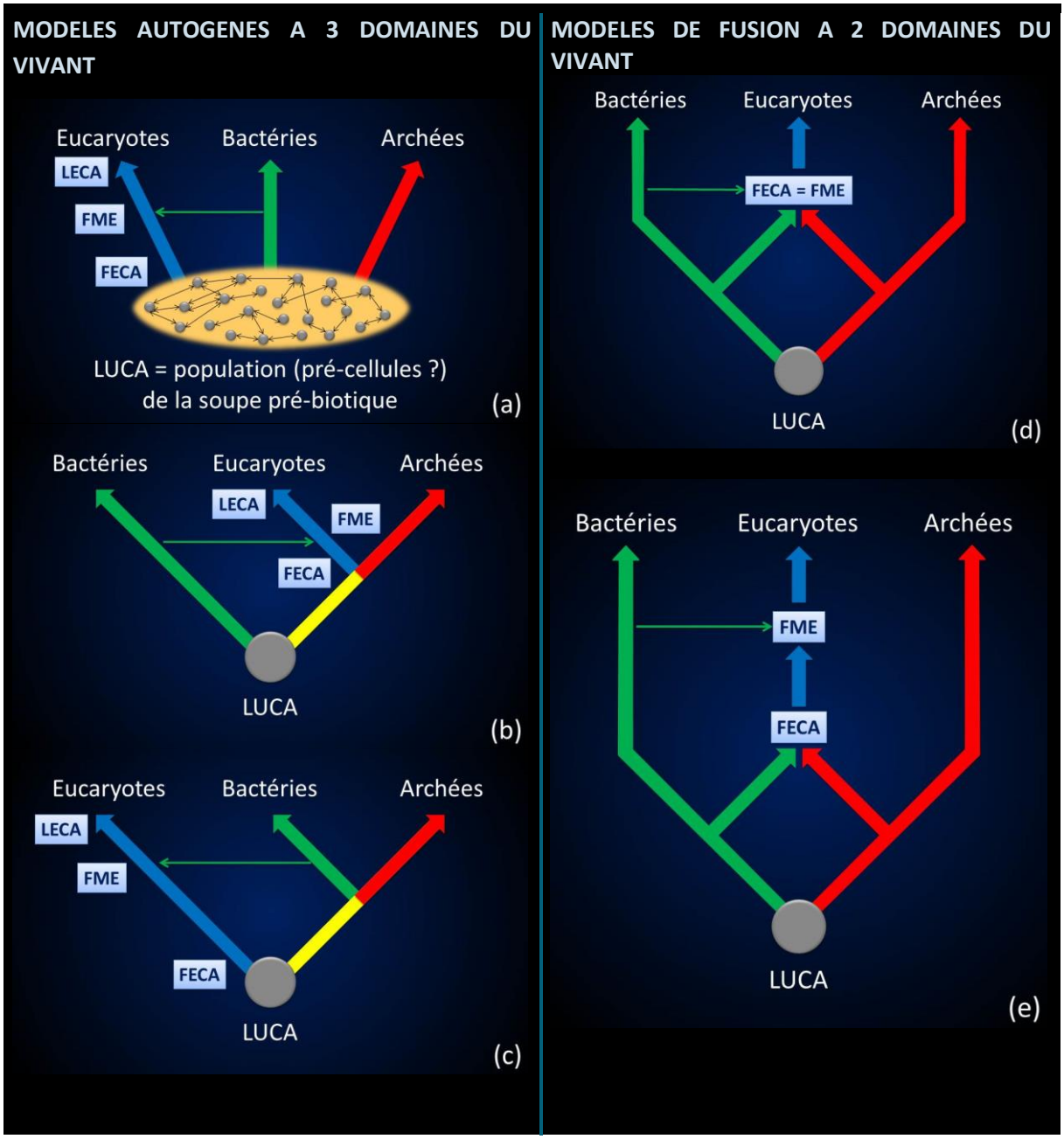


Figure 6. Les différents modèles proposés pour expliquer l'évolution des trois domaines du vivant. **(a)** Les trois domaines de la vie ont une origine indépendante et sont directement issus de la soupe prébiotique. Dans ce cas-là LUCA n'existe pas. **(b)** Les trois domaines dérivent d'un ancêtre commun simple, et l'évolution se serait faite par complexification chez les eucaryotes. LUCA est alors de type procaryote. **(c)** Les trois domaines dérivent d'un ancêtre commun complexe, et l'évolution se serait faite par simplification chez les procaryotes. LUCA peut alors être soit un procaryote, soit un eucaryote, soit quelque chose d'intermédiaire. **(d)** Les eucaryotes sont issus de la fusion d'une bactérie et d'une archée et apparaissent avec la mitochondrie (modèle *mitochondrion-early*) ou avec une proto-mitochondrie (modèle *mitochondrion-intermediate*), leurs autres caractéristiques évoluant après cet événement. **(e)** Les eucaryotes sont issus de la fusion d'une bactérie et d'une archée et apparaissent avant la mitochondrie (modèle *mitochondrion-late*).

Nous allons à présent passer ici en revue diverses hypothèses et modèles qui ont été formulés jusqu'à présent (de façon non exhaustive) et qui tentent d'expliquer au mieux l'origine de la cellule eucaryote.

4.1 SCENARII A 3 DOMAINES

Dans ces scenarii, les eucaryotes sont anciens, peut-être autant que les bactéries ou les archées, voire plus. Ces modèles constituent le modèle autogène des eucaryotes. Dans ce cas-là, la lignée des eucaryotes est indépendante des procaryotes, et on a un scénario à trois domaines. On aurait un « proto-eucaryote », amitochondrial, que l'on nomme FECA (*First Eukaryote Common Ancestor*). Puis, au cours de l'évolution, ces premiers eucaryotes auraient acquis une mitochondrie, donnant le FME (*First Mitochondriate Eukaryote*). Ces derniers auraient ensuite évolué jusqu'au LECA à l'origine de toutes les lignées d'eucaryotes actuelles. Trois hypothèses sont envisageables :

1. Les trois domaines de la vie ont une origine indépendante et sont directement issus de la soupe prébiotique. Dans ce cas-là, LUCA n'existe pas (**Figure 6 (a)**).
2. Les trois domaines dérivent d'un ancêtre commun simple, et l'évolution se serait faite par complexification chez les eucaryotes. Dans ce cas, on a un modèle à trois domaines issus d'un LUCA procaryote (**Figure 6 (b)**).
3. Les trois domaines dérivent d'un ancêtre commun complexe, et l'évolution se serait faite par simplification secondaire chez les procaryotes. Dans ce cas, on a un modèle à trois domaines issus d'un LUCA qui pourrait tout aussi bien être un procaryote qu'un proto-eucaryote (**Figure 6 (c)**).

Dans cette série d'hypothèses, la transition procaryote-eucaryote est expliquée par la complexification de structures ancestrales dans une lignée de procaryotes. L'acquisition de nouvelles structures (ex. cytosquelette, systèmes endomembranaires) aurait permis l'émergence de la phagocytose. Cette propriété aurait ensuite été à la base de l'incorporation d'une α-protéobactérie, qui serait devenue la mitochondrie (Muñoz-Gómez et al., 2022). Les gènes du métabolisme proviendraient donc d'un transfert horizontal à partir de l'ancêtre mitochondrial, tandis que le restant des gènes du proto-eucaryote serait proche des archées par héritage vertical (dans le cas où archées et eucaryotes partageraient un ancêtre commun) ou horizontal (dans le cas où archées et eucaryotes ne seraient pas apparentés). Le premier eucaryote serait donc

dépourvu de mitochondrie (FECA antérieur au FME). Tous ces modèles considèrent que la mitochondrie est tardive et est issue d'une α-protéobactérie de l'ordre des Rickettsiales. Or on sait aujourd'hui que ce dernier point n'est pas si évident, compte tenu de la nature en mosaïque du protéome mitochondrial et d'une faible proportion d'origine α-protéobactérienne (10-20%) (Gray, 2015). En effet, les génomes mitochondriaux montrent une grande diversité en termes de contenu et d'organisation des gènes. Cette diversité complique la reconstruction de l'ancêtre commun des mitochondries et rend difficile l'identification précise de son ancêtre bactérien. Les analyses phylogénétiques basées sur différentes protéines mitochondriales ou génomes entiers peuvent produire des résultats conflictuels, certaines suggérant une relation avec les α-protéobactéries et d'autres indiquant des affinités avec d'autres groupes bactériens (Roger et al., 2017). Certains gènes mitochondriaux montrent des similarités avec des gènes d'autres groupes de bactéries, pas seulement des alphaprotéobactéries, suggérant des transferts horizontaux de gènes ou des événements de recombinaison complexes.

En 1991, Mitchell Sogin suggère que les eucaryotes sont très anciens et appartiennent à une troisième lignée indépendante des archées et des bactéries. Il propose qu'une proto-cellule du monde à ARN, déjà complexe, ait englobé une archée, donnant ainsi les eucaryotes. Cette hypothèse fait suite à une série d'hypothèses lancée à la fin des années 70 par entre autres James Darnell, Hyman Hartman et Ford Doolittle, qui supposaient que les introns étaient des reliques d'assemblages de gènes pré-cellulaires. En effet, il semblerait que LECA était déjà complexe et riche en introns (théorie *Introns-First*) (Koonin, 2009; Penny et al., 2009). En 2002, Hyman Hartman et Alexis Fedorov comparent les génomes entièrement séquencés de cinq eucaryotes (*Saccharomyces cerevisiae*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Arabidopsis thaliana* et *Giardia lamblia*) et des 44 génomes complets d'archées et d'eubactéries présents sur GenBank (Hartman & Fedorov, 2002). Ils mettent ainsi en évidence 347 protéines propres aux eucaryotes, qu'ils nomment « protéines de signature eucaryote » (ESPs). Afin d'expliquer ces résultats, ils supposent l'intervention d'un troisième type de cellule à génome à ARN (qu'ils nomment « chronocyte ») qui n'était ni une archée ni une eubactérie pour expliquer la formation de la cellule eucaryote. Pour eux, l'hypothèse que l'endosymbionte soit dans une cellule hôte à génome à ARN suffit à expliquer pourquoi la transcription a lieu dans le noyau et la traduction dans le cytoplasme. Cette séparation des tâches serait alors le résultat de la mise en place d'un système de communication entre l'endosymbionte (une cellule à ADN) et la cellule hôte (une cellule à ARN). Ce chronocyte devait posséder, entre autres, un cytosquelette permettant la phagocytose, un système endomembranaire et un système complexe de voies de signalisation intracellulaire.

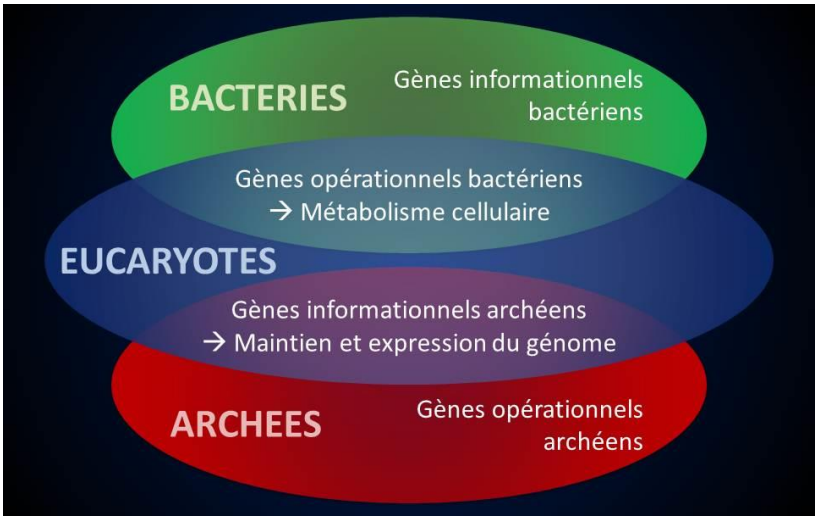
En 2000, Luis Villarreal et Victor DePhilippis émettent l'hypothèse d'une origine virale des protéines de réplication eucaryotes, d'après des études réalisées sur les ADN polymérases de phycodnavirus (virus infectant des microalgues) (Villarreal & DeFilippis, 2000). En effet, celles-ci montrent une grande similarité avec celles des eucaryotes et les reconstructions phylogénétiques de ces polymérases les placent proches de la racine de toutes les ADN polymérases delta des eucaryotes. En 2001, simultanément, Philip Bell et Masaharu Takemura vont même plus loin en proposant une origine virale des eucaryote (Bell, 2001, 2005, 2006, 2009; Takemura, 2001). Selon eux, le noyau eucaryote serait le résultat d'une infection d'une cellule par un virus à ADN complexe : c'est la théorie de l'eucaryogenèse virale. Celle-ci postule que les eucaryotes ont évolué à partir d'un ancêtre des archées actuelles dans lequel un virus à ADN s'est installé pour former

le noyau. Leur hypothèse se fonde sur les nombreuses similarités (coiffe des ARN messagers, chromosomes linéaires, séparation transcription/traduction…) qui existent entre le cycle intracellulaire de certains virus particuliers se répliquant dans le cytoplasme, les poxvirus et le cycle intracellulaire du noyau. Mais la plus forte similarité serait celle du recrutement du réticulum endoplasmique pour former, dans un cas, la membrane de l'usine virale et dans l'autre, la membrane nucléaire. Au vu de la complexité de la cellule eucaryote, Forterre suppose à partir de cette hypothèse que plusieurs virus aient contribué à la formation du noyau (Forterre, 2007). Cela expliquerait à ses yeux l'existence de plusieurs ADN et ARN polymérases chez les eucaryotes. En effet, les analyses phylogénétiques montrent que les différentes versions de ces enzymes ne sont pas issues de duplications qui se seraient produites dans une lignée eucaryote primordiale. Ainsi, les ADN polymérases α, δ et ε d'une part et les ARN polymérases I, II et III de l'autre ne forment pas de groupes monophylétiques, chaque version de la même enzyme étant séparée des autres, dans les arbres phylogénétiques, par des enzymes d'archées et par des enzymes d'origine virale (Filée et al., 2002; Forterre, 2007). L'ancêtre commun aux trois versions de ces enzymes eucaryotes aurait donc également été à l'origine des enzymes archéennes et virales, suggérant un FECA très ancien. L'ADN et les mécanismes de sa réplication seraient d'abord apparus et auraient évolué dans le monde des virus avant d'être transférés au monde cellulaire (Forterre, 2001).

4.2 SCENARII A 2 DOMAINES : LE PARADOXE DE JANUS, DEUX ORGANISMES EN UN ?

Afin de tester le modèle de Woese concernant une origine commune des archées et des eucaryotes, Brian Golding et Radhey Gupta entreprirent, en 1995, de comparer les séquences de génomes de bactéries, d'archées et d'eucaryotes. Ils montrèrent ainsi que le génome eucaryote était mi-archéen, mi-bactérien (de type Gram-négatif) (Golding & Gupta, 1995). Schématiquement, les gènes des eucaryotes peuvent être divisés en trois groupes (Andersson et al., 2003; Hartman & Fedorov, 2002; A. M. Poole & Penny, 2007; Rivera et al., 1998; Rivera & Lake, 2004; Yutin et al., 2008) : ceux qui leur sont spécifiques, ceux qui sont similaires à des gènes bactériens et ceux qui sont similaires à des gènes archéens. De plus, lorsque l'on considère la fonction de ces gènes, on se rend compte que les gènes bactériens sont ceux qui composent les gènes opérationnels en charge des processus de maintenance cellulaire et codent pour la biosynthèse de nucléotides et d'acides aminés des eucaryotes, tandis que les gènes archéens composent les gènes informationnels de la machinerie cellulaire (traduction, transcription, réplication) (**Figure 7**). Ainsi, contrairement aux arbres simples gènes basés sur l'ARN ribosomique qui placent les eucaryotes au sein des archées, l'analyse de génomes montre que la cellule eucaryote résulte probablement de la fusion de génomes archéens et bactériens, transformant ainsi l'arbre de la vie en un anneau de la vie (Rivera & Lake, 2004).

Il se pose alors une question fondamentale : pourquoi les gènes informationnels des eucaryotes ressemblent-ils à des gènes d'archées alors que les gènes opérationnels ressemblent à des gènes de bactéries ? Cette étrange corrélation a été nommée en 2007 « Paradoxe de Janus » par James Lake (Lake, 2007), en référence au dieu romain à une tête mais deux visages opposés (**Figure 7**). Ce paradoxe sous-tend deux questions qu'il convient alors d'expliquer.



Buste de Janus, Musée du Vatican

Figure 7. Dualité du génome eucaryote, entre bactérien et archéen, illustrée par le buste romain de Janus.

Les gènes bactériens sont ceux qui composent les gènes opérationnels en charge des processus de maintenance cellulaire et codent pour la biosynthèse de nucléotides et d'acides aminés des eucaryotes, tandis que les gènes archéens composent les gènes informationnels de la machinerie cellulaire (traduction, transcription, réplication).

1. **Pourquoi les gènes informationnels bactériens sont-ils absents des génomes eucaryotes ?**

Les gènes informationnels sont moins sujets aux transferts latéraux que les gènes opérationnels. Partant de l'observation que deux types de gènes ribosomiques ne peuvent coexister dans un même noyau, Lake a suggéré que le ribosome archéen aurait été, par chance, le seul survivant, tandis que les gènes des ribosomes bactériens auraient été inactivés par diverses protéines, conduisant à leur élimination (Lake, 2007). Cela aurait conduit à une centralisation de l'information dans le noyau hôte, facilitant ainsi les actions de coordination par rapport aux multiples copies de gènes mitochondriaux. De plus, la séparation de la production d'énergie (mitochondrie) et de l'information (noyau) permettrait une coopération équilibrée entre les deux composants (plutôt qu'une forme de parasitisme), chacun y tirant des bénéfices.

2. **Pourquoi les gènes opérationnels d'origine archéenne sont-ils absents des génomes eucaryotes ?**

Une des raisons invoquées serait liée à la structure de leur membrane, beaucoup moins perméable chez les archées que les bactéries. En effet, la composition et les voies de biosynthèse des phospholipides des archées sont très différentes de celles observées chez les bactéries et les eucaryotes (Peretó et al., 2004). Les membranes bactériennes et eucaryotes sont constituées d'acides gras linéaires attachés à du glycérol-3-phosphate par des liaisons esters. Les membranes archéennes, quant à elles, sont constituées de longues chaînes d'alcool isopréniques attachées au glycérol-1-phosphate par des liaisons éthers, beaucoup moins perméables aux ions. Celles-ci surpassent les bactéries dans les environnements soumis à des stress écologiques chroniques (Valentine, 2007). Or, il semblerait que l'écologie et l'évolution des archées soient dictées par leurs adaptations aux stress énergétiques chroniques (Valentine, 2007). Ainsi, les mécanismes biochimiques qui permettent aux archées de faire face à ces stress incluent la faible perméabilité de leur membrane et des voies métaboliques spécifiques adaptées à ces niches écologiques. Si les

membranes des archées offrent une barrière plus efficace contre les ions, pourquoi les eucaryotes ont-ils des membranes similaires à celles des bactéries ? En fait, il est fort possible que la faible perméabilité de la membrane des archées ait été un frein aux diffusions latérales des produits de la respiration et de la transduction bactérienne. Or, l'utilisation de membranes plus fluides et plus perméables pourrait représenter une adaptation visant à renforcer la production d'énergie au détriment de sa conservation, avec des membranes insaturées dans les cas les plus extrêmes. Enfin, des modifications chimiques, comme l'inclusion de stérols spécifiques (ex. Cholesterol) ont certainement pu jouer un rôle dans cette fluidité (Dufourc, 2008). Dès lors, au fur et à mesure que la proto-mitochondrie devenait l'organe énergétique (*power-house*) de la cellule proto-eucaryote, l'adaptation des archées, anaérobies, aux stress énergétiques serait devenue négligeable et aurait été surpassée par le métabolisme aérobie bactérien beaucoup plus énergétique (Davidov & Jurkevitch, 2007). La mitochondrie aurait ainsi mis fin au stress énergétique chronique que subissait l'archée en devenant un organite de conversion d'énergie plus efficace. Les adaptations des archées aux stress énergétiques chroniques n'étant plus avantageux et incompatibles avec un milieu riche en énergie, car surpassées par le métabolisme bactérien plus énergétique, auraient alors disparu petit à petit.

Ce paradoxe de Janus serait donc la métaphore de deux types d'incompatibilités entre archées et bactéries : incompatibilité structurelle entre les systèmes d'information et incompatibilité environnementale/écologique entre les systèmes métaboliques (Davidov & Jurkevitch, 2007). Ceci expliquerait (1) pourquoi les membranes eucaryotes sont de type bactérien alors que l'hôte est de type archéen et (2) le manque de gènes opérationnels archéens chez les eucaryotes (Davidov & Jurkevitch, 2009). La similarité entre la machinerie cellulaire des eucaryotes et des archées peut alors être interprétée comme un héritage vertical depuis un ancêtre commun (synapomorphie) ou, au contraire, comme un état ancestral (symplésiomorphie), les bactéries étant moins similaires à cause d'une accélération du taux d'évolution depuis leur divergence de LUCA. En revanche, bactéries et eucaryotes partagent de nombreuses protéines métaboliques dont l'origine est discutée. Toutefois, il reste encore difficile d'expliquer le tempo et les modalités de ces transferts. Ces constatations ne permettent pas d'expliquer si ces changements furent rapides et soudains, ou si au contraire ils furent lents et progressifs. En particulier, le passage d'une membrane archéenne à une membrane bactérienne est encore difficile à expliquer, surtout en l'absence d'un modèle fiable d'eucaryogenèse. Les membranes ont-elles co-existé un certain temps ensemble ou l'élimination de la membrane archéenne s'est-elle fait instantanément de façon mécanique ? La question est pour l'heure toujours ouverte, les observations actuelles n'expliquant pas la transition d'une membrane à une autre. Il resterait alors à envisager les scenarii permettant d'expliquer ce mélange de gènes. Quels sont-ils ?

4.2.1 CAS OU FECA EST AMITOCHONDRIE

Dans ces scenarii, les caractéristiques des eucaryotes apparaissent avant la mitochondrie (modèle *mitochondrion-late*). Dans ce cas, FECA est issu soit de la fusion physique de deux cellules (les membranes fusionnent), soit de l'endosymbiose d'une archée qui donnera le noyau par une autre cellule (qui ne peut être en aucun cas un eucaryote) (**Figure 6 (e)**). Ce type de modèle requiert trois partenaires. Pour la fusion, deux cellules sont nécessaires, suivie ensuite d'une

endosymbiose pour la mitochondrie. Pour les endosymbioses, il faut deux endosymbiotes (un pour le noyau et un pour la mitochondrie).

Parmi les modèles de fusion, de nombreux partenaires ont été proposés. La première hypothèse chimérique fut celle de l'endosymbiose en série de Lynn Margulis, formulée en 1970 (Margulis, 1970). Cette théorie de l'origine endosymbiotique des mitochondries et des plastes avait déjà été proposée par Mereschkowsky en 1905, avant d'être rejetée et oubliée (Mereschkowsky suggérait même que le noyau dérivait des bactéries). Dans son modèle, Margulis suggère que les mitochondries dérivent d'anciennes bactéries endosymbiotes. Après avoir été fortement critiquée, ses idées ont partiellement été acceptées grâce aux phylogénies moléculaires (Gray et al., 1999; Gray & Doolittle, 1982; Wallace, 1982) et à l'accumulation de données qui s'est poursuivie jusqu'au début des années 2000, lorsque les séquences génomiques des principaux organismes modèles sont devenues disponibles (Gray et al., 2001; Sato, 2021). Les analyses moléculaires effectuées à partir d'ARNr ont ainsi mis en évidence que les plastes sont issus des cyanobactéries et les mitochondries des α-protéobactéries. Par la suite, sur la base de critères morphologiques, Margulis a proposé que l'hôte ayant incorporé l'ancêtre de la mitochondrie (α-protéobactérie) était lui-même issu d'un événement de symbiose entre une archée sans paroi, semblable aux *Thermoplasma acidophilum* actuels, et un spirochète (bactérie) (Lynn et al., 2006). Cependant, aucune donnée génétique n'a pu confirmer l'implication d'un spirochète, Margulis ayant proposé ce partenaire simplement sur des bases morphologiques. En effet, *T. acidophilum* est une euryarchée sans paroi possédant des protéines histones biochimiquement semblables mais analogues à celles des eucaryotes, leur permettant de stabiliser l'ADN en formant des structures ressemblant aux nucléosomes. Quant aux spirochètes, ceux-ci possèdent des structures ressemblant aux microtubules du cytosquelette des eucaryotes. Lors de leur symbiose, la bactérie et l'archée aurait formé un proto-eucaryote sans noyau ni mitochondrie, que Margulis a appelé le « stade Thiodendron » (Margulis et al., 2000). Mais le séquençage de leur génome a montré que spirochètes et *T. acidophilum* n'avaient aucun lien direct d'apparentement avec les eucaryotes.

En 1984, James Lake propose un modèle à l'encontre de la vision de Woese, dans lequel il rend les archées paraphylétiques avec la création d'un quatrième domaine plus apparenté aux eucaryotes : les Crenarchaeota (qu'il appelle « éocyte ») (Lake et al., 1984). A l'origine, cette hypothèse est basée sur l'étude structurale des ribosomes des deux groupes, qui montrent des formes similaires. Par la suite, de nombreuses études sont venues conforter cette hypothèse (Cox et al., 2008), basées sur des séquences de la petite sous-unité des ARNr ou encore sur l'insertion de 11 acides aminés dans le domaine GTPase de la protéine du facteur d'élongation 1-α (EF-1α, aussi appelé EF-Tu) des éocytes et des eucaryotes (Rivera & Lake, 1992) ainsi que sur la phylogénie elle-même de ce gène, qui place les Crenarchaeota comme groupe frère des eucaryotes (Baldauf et al., 1996). Plus récemment, en 2010, les approches bayésiennes et les méthodes de maximum de vraisemblance ont identifié, au sein des Crenarchaeota, le groupe des thaumarchées comme groupe frère, voire comme groupe souche des eucaryotes (Kelly et al., 2011). Lake et Rivera proposent une vision de la vie basée sur un scénario de *ring of life* (cercle de la vie), où le génome des eucaryotes est issu de la fusion d'un génome bactérien (probablement une protéobactérie ou un membre d'un clade photosynthétique ancestral tel les cyanobactéries) et crénarchéen (Rivera, 2007; Rivera & Lake, 2004) suivie de nombreux transferts horizontaux de gènes.

En 1989, sur comparaison de 26 caractéristiques biochimiques, ainsi que sur la comparaison des ADN polymérases, Wolfram Zillig a élaboré un scénario de fusion entre une archée et une bactérie (Zillig, 1991; Zillig et al., 1989). Ce modèle suppose une fusion physique de deux cellules en une seule et nouvelle cellule, avec un seul génome. Ce processus étant indépendant de l'origine de la mitochondrie, ce modèle implique que les eucaryotes modernes aient émergé à partir de trois cellules : deux fusionnées en un proto-eucaryote amitochondrié et une troisième introduite par endosymbiose. Toutefois, dans ce scénario, le terme de fusion reste imprécis.

Cependant, tous les modèles n'expliquent pas forcément l'émergence d'une caractéristique fondamentale des eucaryotes : le noyau. Afin de justifier l'origine de celui-ci, Purificacion Lopez-Garcia et David Moreira émettent en 1998 une hypothèse syntrophique fondée sur le transfert de dihydrogène en milieu anaérobie entre une ancienne myxobactérie (δ-protéobactérie) réductrice de sulfate et une euryarchée méthanogène (futur noyau) (Moreira & López-García, 1998). Une seconde endosymbiose faisant intervenir une α-protéobactérie serait ensuite à l'origine de la mitochondrie. L'apparition du noyau se serait faite en deux étapes, sous l'influence de deux forces sélectives (López-García & Moreira, 2006). (1) Dans un premier temps, il y aurait eu une compartimentation métabolique permettant d'éviter la coexistence délétère des voies anaboliques (synthèses autotrophes par l'archée méthanogène) et cataboliques (fermentation par la myxobactérie). En effet, la fermentation bactérienne libère de l'hydrogène, du dioxyde de carbone et de l'acétate, tandis que l'archée méthanogène tire son énergie en réduisant ce dioxyde de carbone avec de l'hydrogène pour produire du méthane. Cette première phase de compartimentation permet donc d'éviter le recyclage délétère incessant des molécules par ces deux voies métaboliques opposées. (2) Par la suite, l'acquisition de la respiration oxygénique liée à l'arrivée de la mitochondrie aurait entraîné l'abandon de la méthanogenèse (cette dernière ayant un rendement énergétique beaucoup moins important). Une fois la capacité de méthanogenèse perdue, la membrane archéenne aurait donc été perdue à son tour. Il y aurait alors eu, dans le même temps, formation d'un système endomembranaire sécréteur par invagination de la membrane bactérienne, qui aurait supplanté la membrane archéenne originale, formant dès lors la membrane nucléaire (Lombard et al., 2012). L'apparition du noyau assurerait alors une protection contre la synthèse de protéines aberrantes due à l'invasion des introns après l'arrivée de la mitochondrie en découplant la transcription et la traduction.

4.2.2 CAS OU FECA EST MITOCHONDRIE (= FME)

Dans ces scenarii, les eucaryotes apparaissent avec la mitochondrie (modèle *mitochondrion-early*), leurs autres caractéristiques évoluant après cet événement, ou une proto-mitochondrie (modèle *mitochondrion-intermediate*) (Roger et al., 2017). Dans ce dernier modèle intermédiaire, la proto-mitochondrie est décrite comme un ancêtre des mitochondries modernes, une alphaprotéobactérie qui a été intégrée dans une cellule hôte. Contrairement au modèle *mitochondrion-early*, où cette endosymbiose est vue comme un événement déclencheur de la formation des eucaryotes, le modèle intermédiaire suggère que l'hôte possédait déjà quelques caractéristiques eucaryotes. Ainsi, la symbiose avec la proto-mitochondrie a eu lieu à un stade intermédiaire de l'évolution cellulaire. Cette cellule n'était pas encore un eucaryote complet mais avait commencé à évoluer vers cette direction. Cette acquisition symbiotique à un stade intermédiaire de l'évolution cellulaire qui a favorisé la complexité et la diversification des

eucaryotes. Dès lors, les eucaryotes sont issus directement de l'endosymbiose d'une bactérie (à l'origine de la mitochondrie) par une archée. Ici, seuls deux partenaires sont nécessaires (une cellule hôte et l'ancêtre de la mitochondrie). Dans ces scenarii, le FECA et le FME ne font qu'un (**Figure 6 (d)**).

De nombreux scenarii d'endosymbioses métaboliques ont été développés entre une archée et une bactérie évoluant par la suite en mitochondrie. Ainsi, en 1992, Dennis Searcy formula l'hypothèse que les eucaryotes dérivent d'une symbiose fondée sur le transfert de soufre entre une euryarchée sans paroi de type *Thermoplasma* et une protéobactérie qui sera la future mitochondrie. Bien que les partenaires soient les mêmes que dans le scénario de Margulis vu précédemment, la différence avec ce scénario repose sur la nature de la symbiose (fondée sur le soufre), le nombre de partenaires impliqués et la nature du premier eucaryote. Pour Searcy, les eucaryotes apparaissent directement après l'acquisition de la mitochondrie par endosymbiose, impliquant alors deux partenaires.

William Martin et Miklos Müller font, en 1998, du transfert de l'hydrogène la clé de la symbiose entre l'α-protéobactérie (fournissant l'hydrogène et le dioxyde de carbone) qui va devenir la mitochondrie et une euryarchée méthanogène autotrophe et anaérobie qui le consomme (Martin & Muller, 1998). Dans leur modèle, la membrane archéenne est remplacée progressivement par les phospholipides bactériens.

En 2006, Martin et Koonin proposent que l'apparition de la membrane nucléaire soit liée à la diffusion des introns du groupe II du génome mitochondrial dans le génome nucléaire et à la formation de *spliceosomes* (Martijn & Ettema, 2013; Martin & Koonin, 2006). L'apparition du noyau aurait ensuite été sélectionnée car elle découple la transcription (nucléaire) et la traduction (cytosolique), empêchant ainsi la coexistence de protéines délétères, après la diffusion dans le génome nucléaire d'introns d'origine mitochondriale (Koonin & Martin, 2005; López-García & Moreira, 2006). En effet, parce que la traduction ribosomique est plus rapide que le phénomène d'épissage, l'hôte aurait été incapable d'exprimer ses gènes dépourvus de ses introns, traduction et transcription se déroulant simultanément. Sans cela, il y aurait eu synthèse de protéines aberrantes à partir de gènes contenant encore les introns. La traduction aurait été trop rapide et aurait traduit les introns en même temps que les gènes, rendant les protéines défectueuses. Ainsi, une séparation entre l'épissage et la traduction devenait dès lors nécessaire. L'une des fonctions premières de la membrane nucléaire aurait ainsi été de permettre l'épissage des ARNm afin de les débarrasser de leurs introns, de telle sorte que la traduction se fasse sur un brin d'ARNm avec un cadre de lecture continu (Koonin & Martin, 2005).La membrane du noyau aurait alors été formée *de novo* par des vésicules de phospholipides bactériens qui auraient formé des vésicules encapsulant le génome archéen.

En 2013, Joran Martijn et Thijs Ettema ont développé un nouveau modèle, le « PhAT » (*phagocytosing archaeon theory*), supposant la possibilité de l'existence d'archées capables de phagocytose, pour expliquer la complexité de la cellule eucaryote (Martijn & Ettema, 2013). Parmi les modèles de fusion, de nombreuses études tendent à renforcer une origine archéenne des eucaryotes, probablement à partir d'un organisme appartenant au super-phylum TACK (Thaumarchées, Aigarchées, Crénarchées, Korarchées) (Guy & Ettema, 2011; Kelly et al., 2011; Martin, 2005; T. A. Williams et al., 2012). De plus, un nombre croissant d'ESPs (*eukaryotic signature proteins*) sont régulièrement découvertes parmi ces derniers (Hartman & Fedorov,

2002) laissant supposer que l'ancêtre archéen de la cellule eucaryote aurait pu être plus eucaryote que ce qui était imaginé jusqu'à présent. Peut-être possédait-il même déjà la capacité de phagocytose. Dans leur scénario en cinq étapes (Martijn & Ettema, 2013), (1) une archée (probablement du groupe TACK) possédant un cytosquelette de filaments d'actine et des ESPs (2) perd sa paroi et gagne ainsi en flexibilité. (3) Le cytosquelette aurait alors évolué en une machinerie primitive de phagocytose, lui permettant d'ingérer d'autres procaryotes. Il en résulterait une augmentation du taux de transfert horizontal de gènes, ce qui déstabiliserait le génome de l'hôte archéen. Il s'en suivrait alors une accélération du taux d'évolution du génome. (4) Une membrane protectrice se serait alors formée autour du génome par des mouvements d'invagination afin de le stabiliser, donnant un proto-noyau. Une ancienne α-protéobactérie aurait alors été ingérée, mais non digérée, lui permettant d'entretenir une relation symbiotique avec son hôte. (5) Finalement, cette α-protéobactérie aurait évolué par réduction pour donner la mitochondrie. Cette mitochondrie, fournissant à la cellule hôte un surplus d'énergie sous forme d'ATP, elle aurait facilité la maturation du noyau et des systèmes endomembranaires.

Et si le secret de l'origine des eucaryotes ne tenait pas dans la symbiose mais dans la prédation ? C'est l'idée développée par Yaacov Davidov et Edouard Jurkevitch en 2009 (Davidov & Jurkevitch, 2009). Si les procaryotes ne peuvent phagocyter, pouvait-on supposer qu'un petit parasite bactérien ait pénétré une grande archée ? En effet, jusqu'alors, tout laisser à croire que la phagocytose était apparue tardivement chez les eucaryotes (Jékely, 2003). Les petites GTPases de la superfamille Ras, propre aux eucaryotes, jouent un rôle essentiel dans la dynamique du cytosquelette et du trafic vésiculaire. Or, leur phylogénie montre que la première fonction des endomembranes était sécrétrice et que la phagocytose serait apparue plus tard (Jékely, 2003). De plus, une analyse des protéines impliquées dans la phagocytose chez différents groupes eucaryotes révèle que la phagocytose a évolué indépendamment au moins trois fois parmi les eucaryotes actuels (Yutin et al., 2009). Une origine tardive de la phagocytose contredit donc les scenarii postulant que la phagotrophie est un pré-requis de l'endosymbiose (Cavalier-Smith, 2009; A. Poole & Penny, 2007). Enfin, on connaît des interactions de prédation entre organismes unicellulaires non basées sur la phagocytose. C'est le cas par exemple des BALOs (*Bdellovibrio* et assimilés, δ-protéobactéries) qui parasitent d'autres bactéries Gram négatif, des *Rickettsia*-like (α-protéobactéries) qui parasitent les eucaryotes ou encore des Nanoarchaeota qui vivent en symbiose ou parasite avec *Ignicoccus hospitalis* (Davidov & Jurkevitch, 2009). Davidov et Jurkevitch proposent ainsi que la prédation ou une forme de parasitisme ait joué un rôle dans l'évolution de la cellule eucaryote. L'ancêtre de la mitochondrie serait ainsi capable de franchir la barrière cellulaire de la proie tout en conservant l'intégrité cellulaire de son hôte et de s'y joindre afin d'y développer des relations de prédation/parasitisme tout en déjouant ses mécanismes de défense.

4.3 SCENARII A 1 DOMAINE

Pour Thomas Cavalier-Smith, archées et eucaryotes sont deux groupes frères qu'il regroupe sous le terme *Neomura* (= Nouvelles Parois) (Cavalier-Smith, 2002). Selon lui, l'arbre de la vie est enraciné dans les bactéries, Gram négatives et les *Neomura* seraient issus de bactéries Gram positives capables de remplacer la muréine rigide (peptidoglycane) par des glycoprotéines flexibles permettant la phagocytose (Cavalier-Smith, 2006). Cela expliquerait alors la présence d'une seule membrane plasmique à la fois chez les archées, les eucaryotes, et les bactéries Gram

positives, alors que les bactéries Gram négatives en possèdent deux. La conversion d'une bactérie en eucaryote aurait alors nécessité une soixantaine d'innovations (Cavalier-Smith, 2009). Cavalier-Smith considère que les eucaryotes sont récents (800 à 850 Ma). Toutefois, cette hypothèse s'oppose au registre fossile qui suggère une origine des eucaryotes beaucoup plus ancienne (**Box 3**).

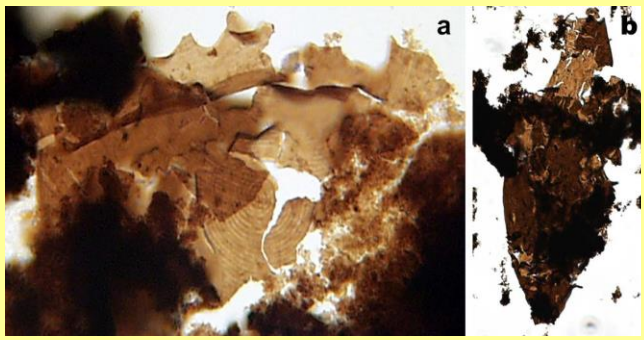
Certains auteurs vont même plus loin en considérant les eucaryotes comme ancestraux et les procaryotes comme en dérivant par réduction (= simplification secondaire). Ainsi, pour David Penny et Anthony Poole, le monde à ARN (s'il a bien existé) est à l'origine de toute vie cellulaire (Collins et al., 2009; A. Poole et al., 1999). Selon eux, si les ARN utilisés dans l'épissage des introns et les processus de stabilisation de l'ARN chez les eucaryotes modernes sont des reliques du monde à ARN, alors ces mécanismes devaient nécessairement être présents chez LUCA. Ainsi, cet ancêtre commun aurait été lui-même un eucaryote ou proto-eucaryote et les procaryotes en seraient dérivés, bien avant l'apparition de la mitochondrie (A. M. Poole & Penny, 2007; A. Poole & Penny, 2007).

Dans la même lignée, Patrick Forterre et Hervé Philippe ont eux aussi présenté, en 1999, un modèle selon lequel LUCA serait déjà un organisme complexe, avec un génome riche en gènes (Forterre & Philippe, 1999a). En effet, supposons qu'une phylogénie soit établie à partir d'un gène évoluant plus vite dans une lignée que dans une autre. Alors, dans ce cas, celle-ci sera placée à la base de l'arbre à proximité du groupe externe : c'est le phénomène d'attraction des longues branches. En tenant compte de ce biais et en éliminant les gènes ayant une vitesse d'évolution trop rapide, Forterre et Philippe ont mis en évidence que LUCA aurait ainsi pu être une cellule complexe comme un eucaryote (Forterre & Philippe, 1999a, 1999b). Les arbres sont alors racinés dans les eucaryotes, auquel cas l'évolution vers les archées et les bactéries se serait faite par simplification et réduction du génome. Forterre avait déjà formulé cette idée en 1995, en supposant que cette simplification aurait eu lieu lors du passage à une niche écologique thermophile (≈ 50–70 °C). C'est l'hypothèse de la thermo-réduction. En effet, dans un tel contexte, les organismes favorisés sont ceux qui se reproduisent le plus vite et qui possèdent l'ADN le plus stable en milieu chaud. Or, c'est précisément le cas des procaryotes, qui se reproduisent très vite et dont l'ADN circulaire est plus stable en milieu chaud que l'ADN linéaire des eucaryotes. Toutefois, à l'heure actuelle, les données fossiles sont en contradiction avec cette hypothèse, les plus anciennes traces procaryotes datant d'au moins 3,4 milliards d'années (Javaux, 2019) (**Box 3**). Cela dit, une absence de preuve n'est en rien une preuve d'absence ! De plus, cette simplification étant génétique, rien ne dit que cet ancêtre était morphologiquement un eucaryote. Ce pouvait très bien être un eucaryote ou proto-eucaryote ayant une autre structure et dont les membranes étaient différentes des eucaryotes actuels. Il a ainsi été suggéré que cette membrane était bifonctionnelle pour la production d'hopanoïdes et de stérols (Desmond & Gribaldo, 2009; Francis, 2021; Lombard et al., 2012; Peretó et al., 2004). Qui plus est, les fossiles moléculaires, en réalité, permettent de dater quand est apparue la membrane eucaryote, (**Box 3**) mais rien ne dit que ces derniers n'aient pas existé avant avec un autre type de membrane (Francis, 2021). Mais cela n'explique pas en soi l'origine de la cellule eucaryote qui pourrait être plus ancienne que les traces de membranes retrouvées qui correspondent en fait aux membranes observées actuellement.

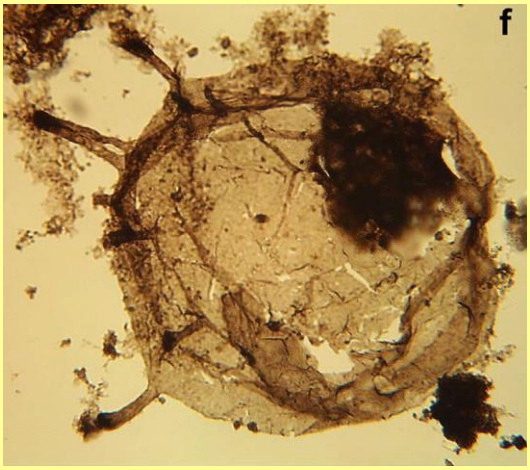
Box 3. Les microfossiles soutiennent une origine ancienne des eucaryotes

Les données moléculaires, biogéochimiques et paléontologiques semblent indiquer que les eucaryotes seraient apparus tôt dans l'histoire de la Terre. Il a été proposé que le plus ancien eucaryote fossile soit *Grypania spiralis*, découvert dans le Michigan (Etats-Unis) et âgé de 2,1 Ga. Son appartenance aux eucaryotes a été avancée uniquement sur base de sa grande taille. Toutefois, la taille n'est pas un critère fiable permettant de distinguer procaryote et eucaryote. De plus, sa morphologie pourrait correspondre à des filaments de cyanobactéries.

(A)



(B)



(C)

(D)

Figure 8. (A) en haut *Grypania spiralis*. (B) *Valeria lophostriata*, > 1,65 Ga Mallapunyah Fm, Australie, un eucaryote précoce dont la paroi est ornée de stries concentriques. (C) *Tappania plana*, protiste du Roger Group, Australie, 1.5 Ga. (D) *Bangiomorpha pubescens*, possible algue rouge, ce qui en ferait le plus ancien eucaryote appartenant à un groupe actuel.

Un caractère morphologique plus fiable permettant de différencier eucaryote et procaryote serait l'ornementation des surfaces cellulaires, propre aux eucaryotes. Les microfossiles ornementés les plus anciens sont des acritarches (du grec *akritos*, signifiant « incertain ou confus », et de *arche*, signifiant « origine », « ce qui est premier »). Ce sont des microfossiles d'affinité biologique incertaine, polyphylétiques.

Des fossiles de structures mycéliales datées à 2,4 Ga, découverts dans du basalte marin de la formation Ongeluk en Afrique du Sud et initialement attribués aux Fungi, restent controversés et nécessitent des preuves supplémentaires pour une identification certaine. Il pourrait en fait s'agir de biomorphes, c'est-à-dire des structures ressemblant à des organismes vivants mais qui ne sont pas nécessairement d'origine biologique, ou bien des fossiles de micro-organismes primitifs distincts des

eucaryotes classiques. Ces microfossiles se trouvent dans du basalte marin et ressemblent beaucoup à des mycéliums fongiques en raison de leur apparence filamenteuse et de leurs structures ramifiées et anastomosées. Cependant, leur grande ancienneté soulève des questions, car les champignons, en tant qu'eucaryotes, n'étaient pas censés exister à cette époque, bien avant l'apparition largement acceptée de ce groupe autour de 1 milliard d'années. Il est possible qu'il s'agisse de formes primitives d'organismes vivants qui auraient évolué dans la biosphère profonde des fonds marins, dans des environnements volcaniques. Il est également possible que ces fossiles représentent des formes de vie unicellulaires proches des procaryotes, capables de se développer dans ces habitats extrêmes.

Des micro-organismes anciens ont également été découverts dans le groupe paléoprotérozoïque de Hutuo, dans les montagnes de Wutai, dans la province de Shanxi, en Chine du Nord. Le sous-groupe de Dongye au sein du groupe de Hutuo est particulièrement intéressant. Il s'agit de *Dongyesphaera tenuispina* et *Dongyesphaera* gen. nov. caractérisé par une ornementation épineuse et *Dictyosphaera* qui possède un réseau d'ornementations. Des datations récentes au zircon situent le dépôt du groupe de Hutuo aux alentours de 2150-1950 Ma. Ces microfossiles fournissent des preuves cruciales du métabolisme eucaryote à une époque où la Terre subissait d'importants changements géologiques.

Les plus anciens eucaryotes fossiles avérés sont les microfossiles à parois organiques (OWM) provenant du groupe inférieur de Changcheng (1673-1638 Ma, formations de Changzhougou et Chuanlinggou) dans la chaîne de Yanshan, en Chine du Nord, qui se compose de 15 espèces. L'assemblage fossile est dominé par des sphéromorphes dont les vésicules porteuses de processus sont moins nombreuses, ainsi que par des formes coloniales et filamenteuses. Parmi celles-ci, 6 taxons morphologiquement complexes (*Dictyosphaera delicata*, 2 espèces de *Germinosphaera*, *Pterospermopsimorpha*, *Simia* et *Valeria lophostriata*) sont identifiés comme des eucaryotes unicellulaires non ambigus. Quatre espèces (*Cucumiforma*, *Navifusa*, *Schizofusa* et les grandes *Leiosphaeridia*) à la morphologie relativement simple mais de grande taille, à la paroi épaisse, et dont certaines présentent des structures d'exocytose à fente médiane, sont probablement d'affinité eucaryote. Cependant, divers microfossiles coloniaux pourraient être des eucaryotes ou des procaryotes. Le nouveau registre des OWM, morphologiquement disparates, représente l'une des premières occurrences d'eucaryotes en Chine et dans le monde, et indique que la vie eucaryote était déjà bien établie à la fin du Paléoprotérozoïque et présentait une diversité modérée, similaire à celle du Mésoprotérozoïque.

Un autre groupe de microfossiles eucaryotes datant du Paléoproterozoïque tardif (1642 Ma) fut trouvé dans le groupe Limbunya, situé dans le bassin de Birrindudu, au nord de l'Australie. Ces microfossiles, qui incluent des espèces nouvelles et bien conservées, témoignent d'une riche diversité eucaryote, ce qui remet en question l'idée que la diversification eucaryote n'a eu lieu qu'environ 800 millions d'années plus tard. 26 taxa ont été identifiés, dont 10 espèces précédemment non décrites. Parmi les espèces décrites, 12 sont considérées comme eucaryotes et 4 ont été nouvellement nommées. Ces eucaryotes prospéraient particulièrement dans des environnements marins marginaux, comme les estrans et les lagunes, notamment dans la formation Blue Hole, qui présente un assemblage très diversifié. Une autre découverte majeure est que ces fossiles montrent des signes de complexité morphologique et cellulaire, suggérant que les eucaryotes possédaient des caractéristiques sophistiquées, telles que des cytosquelettes et des structures internes. Cette diversité morphologique et l'abondance d'espèces dans ces anciens

écosystèmes indiquent que les ancêtres des eucaryotes modernes ont évolué bien plus tôt que prévu, posant de nouvelles questions sur l'évolution de ces organismes et leur adaptation à divers environnements.

On retrouve également *Valeria lophostriata* dans la formation Mallapunyah (super-groupe McArthur, Australie, 1.65 Ga) et dans de nombreuses successions siliciclastiques protérozoïques plus jeunes. *Valeria lophostriata* est un acritarche sphérique qui se distingue facilement par ses stries concentriques. Ces stries sont en fait des ornements de la paroi, comme le montrent les images de microscopie électronique à balayage (MEB). Ces crêtes sont espacées de 1µm et situées sur la face interne de la vésicule. *Valeria* possède une large distribution stratigraphique, s'étendant de la fin du Paléoprotérozoïque au Mésoprotérozoïque et Néoprotérozoïque.

D'autres fossiles, à l'image de *Tappania plana*, âgés de 1,5 Ga et découverts au nord de l'Australie dans les schistes du Roper Group, montrent par leur structure acanthomorphe (= présence de processus épineux) que le cytosquelette et les prérequis écologiques à la diversification des eucaryotes étaient déjà établis à cette époque. Ces acritarches pourraient constituer un groupe souche d'eucaryotes.

Les premiers fossiles pouvant éventuellement correspondre à un groupe actuel (groupe couronne) sont ceux de *Bangiomorpha pubesens*, trouvés dans la Hunting Formation au Canada et âgés de 1,047 Ga. Ceux-ci ont été interprétés comme des algues rouges bangiacées, sur base de figures de divisions cellulaires (aspect multicellulaire) et de reproduction sexuée, permettant de mettre une limite inférieure à l'âge d'apparition des végétaux à chloroplastes « primaires ».

Concernant les biomarqueurs d'eucaryotes (fossiles moléculaires de stéranes), ceux-ci sont abondants sans ambiguïté à partir de 1,7 Ga, appuyant l'idée que les eucaryotes sont anciens. Dans la formation de Barney Creek, dans le nord de l'Australie, près de Borroloola, le biote des protostérols est un groupe d'organismes composé de bactéries aquatiques productrices de protostérols et d'anciens eucaryotes souche, datant de 1,6 à 0,8 Ga (période du Tonien). Ces organismes étaient plus complexes que les bactéries actuelles et ont précédé les derniers ancêtres communs de tous les eucaryotes modernes. C'étaient des prédateurs se nourrissant de bactéries et peut-être d'autres eucaryotes. Ils étaient présents en grand nombre dans les milieux aquatiques des mers et ont sérieusement affecté l'écosystème terrestre de l'époque. Ces micro-organismes se sont adaptés aux niveaux d'oxygène beaucoup plus faibles de l'époque et auraient également produit des protostéroïdes. Des signaux chimiques suggèrent que les molécules pourraient provenir d'un ancêtre du dernier ancêtre commun eucaryote à partir duquel les champignons, les plantes et les animaux ont tous évolué. Les eucaryotes modernes seraient ainsi apparus lors de cette « transformation tonienne » suivi par la prolifération d'algues rouge il y a 800 Mo. Cette phase constitue l'un des plus profond tournant écologique de l'histoire de la Terre.

Références

Javaux, E. J., Knoll, A. H. & Walter, M. R. Morphological and ecological complexity in early eukaryotic ecosystems. Nature 412, 66–69 (2001).

Butterfield, N. J. Bangiomorpha pubescens n. gen., n. sp.: implications for the evolution of sex, multicellularity, and the Mesoproterozoic/Neoproterozoic radiation of eukaryotes. Paleobiology 26, 386–404 (2000).

Ke Pang, Qing Tang, Xun-Lai Yuan, Bin Wan, Shuhai Xiao. A biomechanical analysis of the early

eukaryotic fossil *Valeria* and new occurrence of organic-walled microfossils from the Paleo-Mesoproterozoic Ruyang Group. Palaeoworld Volume 24, Issue 3, 251-262 (2015).
Bengtson, S., Rasmussen, B., Ivarsson, M. et al. Fungus-like mycelial fossils in 2.4-billion-year-old vesicular basalt. Nat Ecol Evol 1, 0141 (2017). https://doi.org/10.1038/s41559-017-0141.
Timothy M. Gibson, Patrick M. Shih, Vivien M. Cumming, Woodward W. Fischer, Peter W. Crockford, Malcolm S.W. Hodgskiss, Sarah Wörndle, Robert A. Creaser, Robert H. Rainbird, Thomas M. Skulski, Galen P. Halverson; Precise age of Bangiomorpha pubescens dates the origin of eukaryotic photosynthesis. Geology 2017; 46 (2): 135–138. doi: https://doi.org/10.1130/G39829.1.
Leiming Yin, Fanwei Meng, Fanfan Kong, Changtai Niu. Microfossils from the Paleoproterozoic Hutuo Group, Shanxi, North China: Early evidence for eukaryotic metabolism. Precambrian Research, Volume 342 (2020).
Javaux, E.J., Knoll, A.H., Walter, M.R., 2004. TEM evidence for eukaryotic diversity in mid-Proterozoic oceans. Geobiology 2, 121–132.
Javaux, E.J., 2007. The early eukaryotic fossil record. In: Jekely, G. (Ed.), Origins and Evolution of Eukaryotic Endomembranes and Cytoskeleton. Landes Biosciences, TX, USA, pp. 1–19.
Javaux, E.J., Lepot, K., 2018. The Paleoproterozoic fossil record: Implications for the evolution of the biosphere during Earth's middle-age. Earth-Science Reviews, 176:68-86.
Riedman, L.A., Porter, S.M., Lechte, M.A., dos Santos, A. and Halverson, G.P. (2023), Early eukaryotic microfossils of the late Palaeoproterozoic Limbunya Group, Birrindudu Basin, northern Australia. Pap Palaeontol, 9: e1538. https://doi.org/10.1002/spp2.1538
Miao L, Moczydłowska M, Zhu S, Zhu M. New record of organic-walled, morphologically distinct microfossils from the late Paleoproterozoic Changcheng Group in the Yanshan Range, North China. Precambrian Research, (2019). doi: 10.1016/j.precamres.2018.11.019.
Brocks, J.J., Nettersheim, B.J., Adam, P. *et al.* Lost world of complex life and the late rise of the eukaryotic crown. *Nature* 618, 767–773 (2023). https://doi.org/10.1038/s41586-023-06170-w.

Cependant, un groupe monophylétique particulier d'organismes tend à renforcer une origine bactérienne des eucaryotes et des archées : les bactéries du groupe PVC. Ce groupe comprend les Chlamydiae, les Lentisphaerae, le phylum candidat Omnitrophica, les Planctomycetes, le phylum candidat Poribacteria, et les Verrucomicrobia. Les bactéries PVC présentent des caractéristiques génétiques et cellulaires inhabituelles pour les bactéries, mais spécifiques aux archées et/ou aux eucaryotes (D. P. Devos & Reynaud, 2010; Reynaud & Devos, 2011; Rivas-Marín & Devos, 2018; Santarella-Mellwig et al., 2010; Wiegand et al., 2018). Trois étapes majeures ont lieu conduisant à l'émergence des eucaryotes et des archées : (1) la perte de la paroi cellulaire à base de peptidoglycane des bactéries, (2) le remplacement de la protéine du cytosquelette bactérien FtsZ par la tubuline eucaryote et (3) le développement d'un système endomembranaire chez les eucaryotes. Les membres du groupe PVC semblent « intermédiaires » pour ces deux événements.

Sur la base de l'homologie entre de nombreux composants du système endomembranaire eucaryote et ceux du périplasme bactérien, il a été suggéré que le système endomembranaire eucaryote est le résultat de l'internalisation du périplasme bactérien (Blobel et al., 1986; de Duve, 2007). En outre, l'émergence et le développement du système endomembranaire eucaryote ont été basés sur les MCP (*membrane coat proteins*) (D. Devos et al., 2004). Les planctomycètes ont toujours été intéressants pour les scénarios d'eucaryogenèse en raison de leur système endomembranaire développé (Forterre, 2011; J. A. Fuerst & Nisbet, 2004), qui est sans doute l'un

des plus développés parmi les procaryotes (Acehan et al., 2014; Boedeker et al., 2017; D. P. Devos et al., 2014; Santarella-Mellwig et al., 2013). Chez ces bactéries, la membrane cytoplasmique envoie des invaginations vers l'intérieur du cytoplasme, formant une organisation complexe et représentant une « véritable » internalisation du périplasme (D. P. Devos et al., 2014; Santarella-Mellwig et al., 2013). Cependant, sans lien moléculaire entre les systèmes endomembranaires eucaryotes et bactériens, les Planctomycètes restaient une curiosité du monde procaryote. La détection de protéines présentant la même signature structurelle que les MPC eucaryotes dans les protéomes de divers Planctomycètes (et espèces apparentées), et la démonstration que ces protéines soutiennent leurs endomembranes, ont considérablement changé cette situation (Santarella-Mellwig et al., 2010). Bien qu'il n'y ait pas de signal de séquence entre les MCP des Planctomycètes et celles des eucaryotes, la présence de ces protéines chez les procaryotes est unique et représente le premier lien moléculaire entre les systèmes endomembranaires eucaryotes et procaryotes. Si leurs similitudes structurelles et fonctionnelles plaident en faveur d'une relation évolutive, la nature exacte de cette relation, convergente ou divergente, reste à déterminer (D. P. Devos, 2012).

Le développement de la phagocytose a été une étape cruciale lors du processus d'eucaryogenèse. Encore une fois, les Planctomycètes affichent des phénotypes apparentés. En effet, il a été démontré que le planctomycète *Gemmata obscuriglobus* était capable d'intérioriser des protéines avant de les dégrader en interne dans un processus rappelant de l'endocytose eucaryote (Lonhienne et al., 2010). Cette observation a ensuite été étendue à d'autres molécules, comme le dextrane, ainsi qu'à une autre espèce de Planctomycètes, *Planctopirus limnophila*, ce qui suggère que cette capacité est plus globale (Boedeker et al., 2017). De plus, il a été montré qu'un autre Planctomycète, *Candidatus Uab amorphum*, est capable de phagocyter d'autres bactéries de façon similaire aux eucaryotes (Shiratori et al., 2019).

D'autres caractéristiques que l'on ne trouve généralement pas chez les bactéries et qui sont plus souvent associées aux eucaryotes ou aux archées, ou aux deux, sont retrouvées au sein du groupe PVC. Parmi ces caractéristiques, nous pouvons citer :

- la synthèse des stérols, que l'on pensait auparavant liée à l'eucaryogenèse, s'avère être d'origine bactérienne (Santana-Molina et al., 2020). Ainsi, certaines bactéries telles le planctomycète *Gemmata obscuriglobus* sont capables de produire des stérols.
- des protéines liées à la tubuline ont été décrites chez les Verrucomicrobia (Pilhofer et al., 2011) et des protéines contenant un domaine de type tubuline ont été détectées chez les Planctomycètes (Makarova & Koonin, 2010). De même, le génome de *Candidatus Uab amorphum*, le planctomycète de type phagocytaire, code pour une protéine apparentée à l'actine (Shiratori et al., 2019). Cependant, les relations phylogénétiques de ces protéines de tubuline et actine par rapport à leurs homologues d'archées et eucaryotes sont encore indéfinis.
- les Planctomycètes anammox possèdent des chaînes d'hydrocarbures qui sont liées par un éther ou un ester au glycérol, appelées lipides membranaires ladderane, fournissant une solution possible à la question de la transition lipidique (Villanueva et al., 2021).
- diverses voies métaboliques jusqu'alors considérées comme eucaryotes ont été découvertes au sein du groupe PVC. C'est le cas des gènes du métabolisme de transfert C1 (Planctomycètes), qui pourraient être liés à l'origine de la méthanogenèse (Chistoserdova et al., 2004), ou encore des enzymes de la voie mévalonate de biosynthèse des isoprénoïdes des phylums Verrucomicrobia et Lentisphaerae, généralement associés aux

archées et eucaryotes (Hoshino & Gaucher, 2018). De même, des homologues des sérine/thréonine kinases eucaryotes et de l'ubiquitine E2 ont été détectés chez les Planctomycètes (Arcas et al., 2013).

- enfin, les Verrucomicrobia se divisent par fission binaire à l'aide de la protéine FtsZ, démontrant que le dernier ancêtre commun du groupe PVC avait le gène *ftsZ* et se divisaient par fission binaire, comme la plupart des bactéries. Cependant, tous les membres des Chlamydiae et des Planctomycètes ont perdu le gène *ftsZ*, ainsi que d'autres gènes issus de la division et de la paroi cellulaire (*dcw*), et se divisent par division asymétrique (Rivas-Marín & Devos, 2018; Santana-Molina et al., 2020). Leurs mécanismes de division sont actuellement inconnus, mais on sait que la perte de *ftsZ* et le développement d'un nouveau mode de division cellulaire s'est produit au cours de l'eucaryogenèse. L'ancêtre commun du groupe PVC et de LAECA (*Last Archaea-Eukaryote Common Ancestor*) pourrait donc être une bactérie qui a commencé à accumuler des caractéristiques qui seront plus tard reconnues comme spécifiques aux eucaryotes et/ou archées (Makarova & Koonin, 2010; Reynaud & Devos, 2011).

Malgré la présence de certaines de ces caractéristiques chez d'autres bactéries (McInerney et al., 2011), les membres du groupe PVC sont les seuls à en réunir autant dans un seul groupe d'organismes apparentés. Ce mélange de phénotypes chez les membres actuels du groupe PVC soulève l'hypothèse qu'ils peuvent être liés à des étapes intermédiaires faisant la transition entre les bactéries et un ancêtre de LAECA (D. P. Devos & Reynaud, 2010).. Dans ce scénario où les bactéries PVC et LAECA partagent un ancêtre commun, il existe une lignée, ou communauté d'organismes (O'Malley et al., 2019), liée aux bactéries PVC et intermédiaire entre l'ancêtre des archées et des eucaryotes en route vers LAECA, qui a déjà développé certaines de ses caractéristiques (celles partagées entre l'ancêtre des eucaryotes et des archées. Ainsi, dans ce scénario, les bactéries PVC sont les taxons frères de la lignée LAECA. La racine des archées et des eucaryotes est alors peut-être, mais pas nécessairement, située entre les lignées archées et eucaryotes. La question de la complexité de l'ancêtre de chaque domaine est également résolue dans ce scénario 1D où le LAECA diverge de bactéries allant vers plus de complexité. Il y a donc une augmentation continue de la complexité des bactéries aux eucaryotes, tandis qu'elle diminue chez les archées (simplification secondaire) (D. P. Devos, 2021).

4.4 DISCUSSION AUTOUR DES MODELES PROPOSES

Une multitude de modèles supposés expliquer l'origine de la cellule eucaryote ont ainsi vu le jour ces dernières années (**Figure 9**). Mais aucun ne semble réellement répondre à toutes les questions à la fois. De plus, la nature et les propriétés de FECA ainsi que sa parenté phylogénétique demeurent encore aujourd'hui une source d'interrogations (**Box 4**).

Au gré des récentes découvertes sur la physiologie, la biologie moléculaire et la génomique des archées et des bactéries, de nombreux modèles de fusion ont été remis en question (Forterre, 2011). Ainsi, les modèles de Margulis et Searcy impliquant des archées de type *Thermoplasma* à l'origine du noyau des eucaryotes ont depuis lors été réfutés. En effet, les analyses moléculaires ont montré que les « histones » des *Thermoplasma* ne sont pas homologues à celles des eucaryotes mais aux protéines HU bactériennes (DeLange et al., 1981). De même, l'idée que les eucaryotes puissent dériver d'archées méthanogènes (hypothèse de l'hydrogène de Martin & Müller, syntrophie de Moreira & Lopez-Garcia) est obsolète depuis la découverte de ces mêmes histones analogues aux eucaryotes chez des thaumarchées mésophiles (Brochier-Armanet, Gribaldo, et al.,

2008; Čuboňová et al., 2005). De plus, l'implication d'une δ-protéobactérie ou d'une spirochète, comme le suggère l'hypothèse syntrophique, est aussi écartée depuis que le séquençage de leur génome n'a pas réussi à révéler des affinités spécifiques avec les eucaryotes et après la découverte de protéines kinases et de protéines G dans de nombreux groupes d'archées et de bactéries (Dong et al., 2007; Tyagi et al., 2010).

Les modèles selon lesquels le noyau serait un endosymbionte (J A Lake and Rivera 1994) ont été fermement rejetés (Martin, 1999; A. Poole & Penny, 2001; Rotte & Martin, 2001). La découverte de protéines que l'on pensait jusqu'alors spécifiques aux eucaryotes (ESPs) a été un argument utilisé pour soutenir un modèle de fusion à trois partenaires (Hartman & Fedorov, 2002). Mais une explication plus simple est de dire que le domaine des eucaryotes s'était déjà séparé du domaine des archées avant l'acquisition de la mitochondrie (Amiri et al., 2003).

En fait, le choix des partenaires dans les modèles de fusion proposés a toujours été fonction de l'état des connaissances du moment. Ainsi, en 2011, Forterre a eu l'idée de proposer un nouveau modèle de fusion, le meilleur qui soit à cette époque, et de déceler en quoi son modèle est malgré tout incorrect (Forterre, 2011). Cet « exercice de style » avait pour but de révéler que l'on peut en fait proposer de nombreux partenaires de fusion. Il propose alors un modèle de fusion dans lequel FECA serait issu de la phagocytose d'une thaumarchée par une bactérie PVC (*Planctomycetes*, *Verrucomicrobia*, *Chlamydiae*), suivi par l'invasion massive de ses descendants par différentes lignées de virus (tels les NCLDV = Nucleo-Cytoplasmic Large DNA Viruses) et rétrovirus. En effet, des traits eucaryotes ont été mis en évidence chez les bactéries PVC (Santarella-Mellwig et al., 2010; Wagner & Horn, 2006) et les thaumarchées (Brochier-Armanet, Boussau, et al., 2008; Brochier-Armanet, Gribaldo, et al., 2008; Spang et al., 2010). Les thaumarchées auraient fourni des protéines informationnelles et opérationnelles tandis que les bactéries PVC auraient fourni les phospholipides, la tubuline et les protéines membranaires nécessaires à la formation du noyau. Les virus auraient ensuite apporté des protéines propres aux eucaryotes, conduisant à une complexification du proto-eucaryote. Cependant, son modèle n'expliquant ni l'origine du noyau, ni la répartition des traits archéens et bactériens au sein des eucaryotes, ni l'origine des virus eucaryotes, ni l'existence de trois lignées distinctes de ribosomes, il conclut que les eucaryotes et leurs virus ont probablement évolué à partir d'une lignée spécifique, selon le scénario à trois domaines proposé par Woese.

Box 4. A quoi ressemblait LECA, l'ancêtre commun des eucaryotes actuels ?

Il n'est pas certain que les caractéristiques des eucaryotes (enveloppe nucléaire, système endomembranaire, épissage des introns et chromosomes linéaires) aient précédé l'apparition de la mitochondrie. Cependant, en comparant les propriétés des différents groupes d'eucaryotes actuels, il est possible de dresser un portrait-robot de l'ancêtre commun des eucaryotes modernes (LECA) et d'en inférer ses caractéristiques minimales. Ainsi, LECA devait probablement être hétérotrophe, sans paroi cellulaire et à digestion extracellulaire. La première étape menant aux eucaryotes modernes a dû être le développement de la phagocytose. La formation de replis membranaires à l'intérieur de la cellule aurait permis une digestion intracellulaire, à l'origine des lysosomes. Cette faculté d'émettre de petits sacs intramembranaires est peut-être à l'origine de l'enveloppe nucléaire et du système endomembranaire. Ses chromosomes étaient sans doute déjà linéaires, exigeant dès lors la mise en œuvre d'une procédure de réparation permanente des extrémités *via* la constitution des télomères.

Les introns étaient déjà présents et devaient être excisés des transcrits ARN lors du processus de maturation d'excision-épissage par un *spliceosome*. C'est seulement après cette étape que les transcrits ARN sont transportés hors du noyau vers le cytoplasme pour servir de matrice lors de la synthèse protéique. Contrairement aux bactéries et aux archées, transcription et traduction n'étaient donc déjà pas synchrones. L'apparition d'un cytosquelette aurait été une réponse à la disparition de la paroi, afin de maintenir la membrane cellulaire et de structurer le cytoplasme. LECA était donc déjà un organisme complexe possédant toutes les caractéristiques eucaryotes et était donc une cellule eucaryote typique. Par conséquent, toutes les innovations clés de l'eucaryogenèse ont dû avoir lieu avant l'apparition de LECA, dans les groupes racines (FECA).

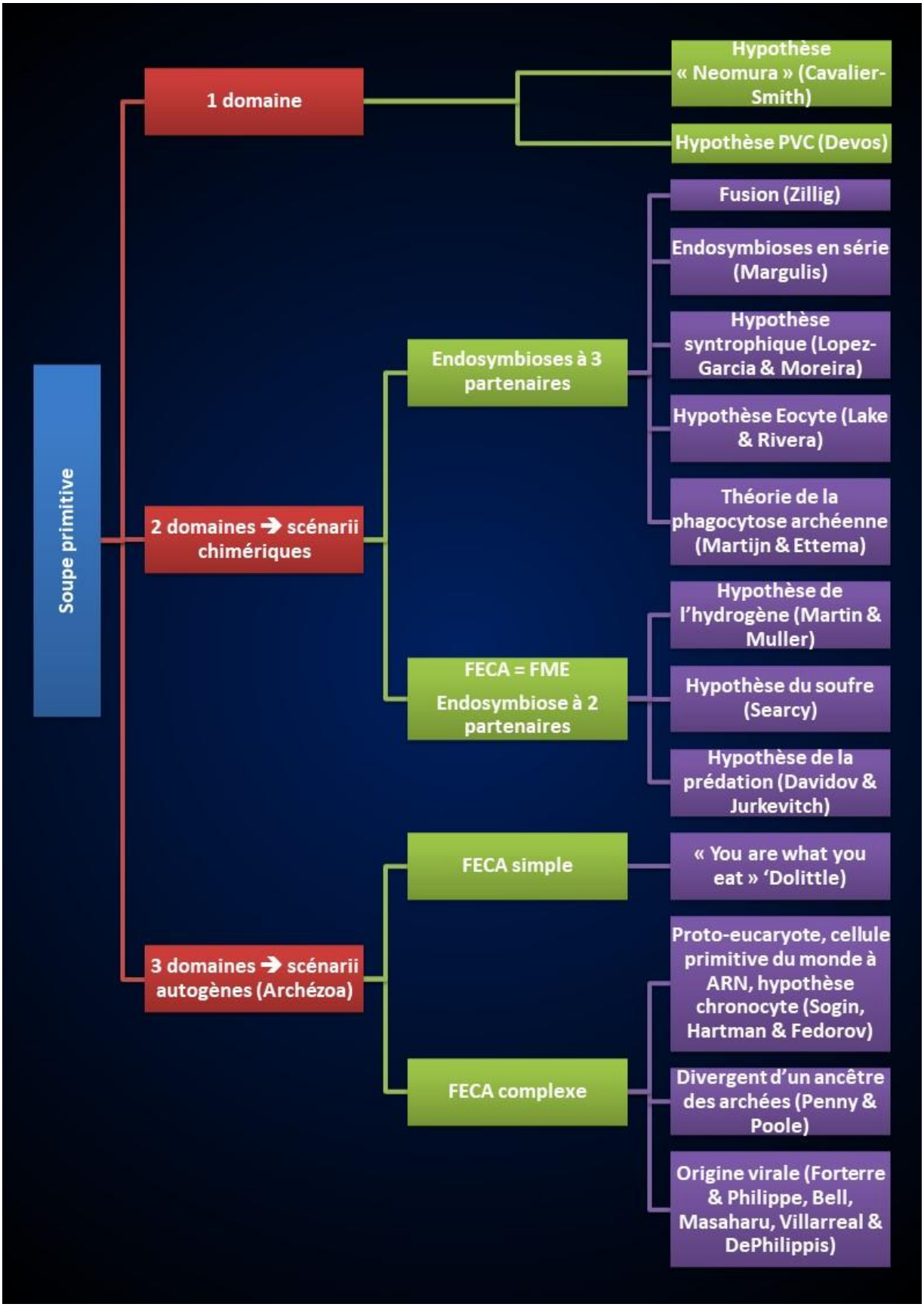

Soupe primitive
1 domaine
Hypothèse « Neomura » (Cavalier-Smith)
Hypothèse PVC (Devos)
2 domaines → scénarii chimériques
Endosymbioses à 3 partenaires
Fusion (Zillig)
Endosymbioses en série (Margulis)
Hypothèse syntrophique (Lopez-Garcia & Moreira)
Hypothèse Eocyte (Lake & Rivera)
Théorie de la phagocytose archéenne (Martijn & Ettema)
FECA = FME
Endosymbiose à 2 partenaires
Hypothèse de l'hydrogène (Martin & Muller)
Hypothèse du soufre (Searcy)
Hypothèse de la prédation (Davidov & Jurkevitch)
3 domaines → scénarii autogènes (Archézoa)
FECA simple
« You are what you eat » 'Dolittle)
FECA complexe
Proto-eucaryote, cellule primitive du monde à ARN, hypothèse chronocyte (Sogin, Hartman & Fedorov)
Divergent d'un ancêtre des archées (Penny & Poole)
Origine virale (Forterre & Philippe, Bell, Masaharu, Villarreal & DePhilippis)

Figure 9. Bilan des différents scenarii envisagés expliquant l'origine de la cellule eucaryote.
Ceux-ci peuvent être regroupés selon le nombre de domaines envisagé, le mécanisme de formation des eucaryotes (symbiose vs autogène) et selon la nature du premier eucaryote (avec ou sans mitochondrie).

Les eucaryotes actuels se distinguent des procaryotes par diverses caractéristiques (Vellai & Vida, 1999) :

- La cellule est compartimentée par un réseau de membranes internes délimitant le noyau et les organites. Ainsi, le noyau est constitué du matériel génétique entouré d'une enveloppe nucléaire, tandis que les organites comprennent le réticulum endoplasmique, l'appareil de Golgi et les lysosomes ;
- Le cytoplasme est structuré par un cytosquelette composé de filaments d'actine et de microtubules de tubuline ;
- La division cellulaire est une mitose. Lors de celle-ci, l'ADN est compacté en chromosomes avant d'être divisé ;
- La reproduction est véritablement sexuée, du moins chez de nombreux eucaryotes, chaque type sexuel apportant une part égale de matériel génétique à la génération suivante ;
- Le génome contient des séquences d'ADN répétées (microsatellites) ;
- Les introns subissent souvent un phénomène d'épissage, faisant du génome eucaryote un génome constitué de *genes in pieces* (Gilbert, 1978; Gilbert et al., 1986).
- Ils possèdent originellement des organites, délimités par des membranes, qui remplissent les fonctions énergétiques de la cellule. Ce sont les mitochondries (qui assurent entre autres la fonction de respiration) et, chez les organismes photosynthétiques, les chloroplastes (qui sont le siège de la photosynthèse).

Parallèlement aux eucaryotes, les procaryotes (bactéries et archées) ont longtemps été définis négativement par l'absence de caractères (Philippe et al., 1995). On y constate l'absence de réseaux membranaires internes (en particulier autour de l'ADN, c'est-à-dire l'absence d'un noyau bien délimité, exception faite du nucléoïde des bactéries du groupe PVC (JohnA. Fuerst, 2013; McInerney et al., 2011; Wagner & Horn, 2006), ainsi que l'absence d'organites cytoplasmiques individualisés. La division du matériel génétique s'y fait non pas par mitose mais par un processus de ségrégation dans lequel l'ADN bactérien demeure lié à l'enveloppe cellulaire. Enfin, chez les bactéries, une paroi mucopeptidique peut fournir une armature externe à la cellule. Le paradigme des procaryotes satisfaisait l'idée traditionnelle d'une évolution graduelle d'une forme de vie simple vers une forme de vie plus complexe (Brinkmann & Philippe, 2007). Ceux-ci regroupent les cellules appartenant à deux domaines différents : les bactéries et les archées. Classer les procaryotes selon un système phylogénétique s'avère compliqué car, contrairement aux animaux et végétaux sur qui on peut appliquer une méthodologie d'anatomie et de physiologie comparées, les bactéries n'ont pas de traits morphologiques complexes et affichent un haut niveau de diversité physiologique et biochimique difficile à interpréter (Philippe et al., 1995). C'est pourquoi on préfère utiliser maintenant des techniques de phylogénie moléculaire, basées sur des marqueurs génétiques.

La classification phylogénétique du vivant s'est d'abord faite sur des gènes individuels (notamment l'ARN ribosomique), avant de prendre en compte un plus grand nombre de gènes au fur et à mesure des progrès de la génomique. Aujourd'hui, les récentes avancées dans le développement de méthodes moléculaires (notamment le séquençage haut débit et la métagénomique) ont permis de faire des progrès importants dans la découverte de nouveaux organismes, comblant ainsi des lacunes dans les classifications phylogénétique

OBJECTIFS

Notre thèse a pour objectif d'apporter de nouveaux éléments concernant les discussions sur les domaines du vivant. Nous allons tenter de répondre successivement à 3 questions :

(1) Pourquoi le nombre de domaines et l'enracinement de l'arbre de la vie ne sont toujours pas résolus ?

(2) Quelle est la phylogénie des archées ?

(3) Quel est le lien entre eucaryotes et archées ?

Ces 3 questions sont majeures dans le domaine de l'évolution et impactent notre compréhension globale de l'évolution. Chacune de ces questions constituera un chapitre de cette thèse.

Ainsi, dans le premier chapitre, nous discuterons des biais dans la quête de ces réponses, ainsi que de l'importance accordée aux techniques et à nos biais de pensée. L'adéquation des méthodes utilisées sera abordée, celles-ci n'étant que des approximations globales. D'un point de vue méthodologique, ce chapitre abordera 6 points majeurs intrinsèques aux données et aux phénomènes biologiques : (1) l'hétérogénéité de substitution entre caractères, (2) l'hétérogénéité des taux de substitutions entre sites, (3) l'hétérogénéité des processus de substitutions entre sites, (4) l'hétérogénéité de composition au cours du temps, (5) l'hétérogénéité des processus de substitution d'un site au cours du temps et (6) l'hétérogénéité du taux de substitution d'un site au cours du temps au sein d'une lignée. Il sera également discuté des biais méthodologiques, des méthodes pour créer des alignements (aligner, sélectionner les positions...) et des modèles utilisés dans les reconstructions phylogénétiques qui peinent à passer outre tous ces problèmes dus à l'hétérogénéité des phénomènes biologiques. D'autre part, nous aborderons également notre propre conception de l'évolution et notre perception naturellement anthropocentrique. Dans ce premier chapitre, conceptuel, les raisonnements concernant l'évolution des archées et des eucaryotes défendus dans cette thèse seront établis sur des fondements solides. Ce premier chapitre sera crucial pour appréhender les questions d'évolution, en particulier celles liées aux questions d'évolutions profondes telles que l'origine des domaines du vivant ou l'apparition de la cellule eucaryotes. Il servira de base pour la suite de cette thèse.

Dans le deuxième chapitre, la thèse se focalisera sur l'évolution des archées et des relations entre les divers groupes qui les constituent. L'objectif principal sera de (1) créer un ensemble de données solide pour établir les liens de parenté entre les différentes archées, et (2) tester ces liens en effectuant des tests permettant de détecter les biais pouvant mener à de mauvaises interprétations. Nous allons d'abord sélectionner des protéomes d'intérêt, puis générer et sélectionner des groupes orthologues sur base de plusieurs critères (taxonomie, nombre d'espèces...). Puis, après ré-enrichissement de ces groupes, nous allons créer deux jeux de données : l'un contenant des OG strictement orthologues, l'autre des OG issus de gènes paralogues récoltés après une méthode de découpe phylogénétique. Le premier représentera un jeu de données très solide et le deuxième sera plus sujet à différents artefacts. Les OG seront encore filtrés selon les longueurs de branches, la diversité taxonomique (au minimum un Asgard, un DPANN, un Euryarchaetoa et un TACK). Ces données seront ensuite utilisées pour la reconstruction phylogénétique. Puis, en utilisant différents tests de solidité (*jackknife* de gènes, 5 réplicas d'espèces) avec différents modèles, les arbres seront comparés entre eux par distances de Robinson-Foulds. Nous déduirons alors les topologies majoritaires selon les jeux de données afin de déterminer les groupes avec des positions changeantes. Enfin, des tests *slow-fast* seront ensuite réalisés en utilisant des positions moins susceptibles au changement évolutif et donc

amenant moins de biais phylogénétiques, dus par exemple à des attractions des longues branches (LBA).

Enfin, dans le troisième chapitre, les résultats obtenus au chapitre 2 seront utilisés pour visualiser où se retrouvent les eucaryotes. Pour cela, les données de 11 espèces eucaryotes seront ajoutées à nos jeux de données obtenus précédemment. Les groupes orthologues formés seront ensuite évalués afin d'éliminer des doublons provenant en particulier des génomes d'organites (mitochondrie, plaste). Puis, en utilisant les différents réplicas générés au chapitre 2, des analyses ayant pour but d'identifier des biais potentiels évoqués au premier chapitre (en particulier les phénomènes d'hétérotatchie et d'hétéropécillie) seront réalisées. A l'issue de cela, afin d'inférer des scénarii d'évolution, nous allons effectuer des tentatives d'enracinement via les méthodes d'AU-test et de rootstrap sur notre réplica d'espèces semblant le plus fiable. Enfin, nous allons utiliser des méthodes bayésiennes afin de calculer un arbre final que nous enracinerons selon nos résultats précédemment obtenus.

Pour terminer, la discussion générale mettra en lien l'ensemble des résultats afin d'en tirer les conclusions appropriées concernant les relations de parenté entre les Asgard et les eucaryotes.

MATERIEL & METHODES

1 TELECHARGEMENT DES GENOMES

Les génomes et protéomes d'archées ont été récupérés à la fois de RefSeq et GenBank (les génomes de RefSeq sont un sous-ensemble de ceux de GenBank) à partir du portail du National Center for Biotechnology Information (NCBI). 584 protéomes conceptuels correspondant à autant de génomes furent téléchargés de RefSeq, tandis que 1077 génomes, pour lesquels seulement 770 protéomes étaient disponibles, venaient de GenBank. Le téléchargement a été effectué le 8 Février 2017.

2 EVALUATION DE LA QUALITE DES GENOMES AVEC QUAST

QUAST (v2.4) (Gurevich et al., 2013) a été utilisé pour estimer la qualité des assemblages génomiques afin de rassembler les génomes ayant la plus haute qualité possible. Nous avons ainsi récupéré pour chaque génome : le nombre de scaffolds, le nombre de scaffolds > 1000 nt, la taille totale du génome, la taille totale du génome basée sur les scaffolds > 1000 nt, la taille du plus grand scaffold, le taux de GC, les valeurs N50/N75/L50/L75 et le nombre de N pour 100 kpb (N étant le nombre de positions non assignées). Les assemblages sont évalués selon les trois dimensions clés : la contiguïté, la complétude et la pureté (incluant la présence de contaminations). La contiguïté est souvent mesurée en tant que contig N50. Étant donné un ensemble minimal de contigs classés par longueur décroissante, la valeur N50/N75 est définie comme la longueur du contig situé à 50%/75% de la longueur totale du génome dans la distribution, tandis que le L50/L75 est défini comme le rang de ce contig spécifique. Un contig N50 supérieur à 1 Mpb est généralement considéré comme très bon. Les résultats de QUAST sont donnés dans le fichier **quast.csv.** La complétude a été évaluée via la recherche des protéines ribosomiques pour chaque génome. Enfin, la mesure de l'exactitude s'est faite par l'évaluation de la contamination. Ainsi nous considérons un génome de haute qualité un génome ayant une bonne contiguïté d'assemblage, le plus complet possible et avec le moins de contaminations.

3 ANALYSE DES PROTEINES RIBOSOMIQUES ET ECHANTILLONNAGE TAXONOMIQUE PRELIMINAIRE

Afin d'établir une phylogénie préliminaire des Archées, un premier jeu de données a été assemblé en enrichissant 90 alignements de séquences multiples (MSA) de la base de données de protéines ribosomiques RiboDB 1.4.0131 (disponible sur https://bitbucket.org/phylogeno/42-ribo-msas/src/master/MSAs/prokaryotes/). Pour ce faire, nous allons générer des groupes orthologues à partir de notre sélection d'archées, puis nous allons les enrichir à partir de nos MSA.

Nous avons utilisé Forty-Two (42) (Irisarri et al., 2017; Simion et al., 2017), un programme dont le but est d'ajouter (et éventuellement d'aligner) des séquences à un MSA (*Multiple Sequence Alignment*) préexistant tout en contrôlant les relations d'orthologie et les séquences potentiellement contaminantes. En pratique, 42 recherche dans chaque génome les séquences qui sont homologues aux séquences de la MSA de l'alignement de référence et ensuite, grâce à une heuristique avancée, trie les orthologues des paralogues possibles. Nous avons exécuté 42 sur nos 1077 génomes de GenBank (qui comprenaient les 584 génomes disponibles dans RefSeq). Le nombre de génomes dans lesquels l'orthologue d'une protéine ribosomique donnée a été récupéré avec succès est fourni pour chaque MSA dans le TABLEAU 2 (**stats-completude.txt**).

Afin de vérifier si les génomes d'archées étaient contaminés (ou non) par des séquences bactériennes, nous avons effectué un test de contamination avec "42", en utilisant les bactéries

comme protéomes de référence pour effectuer le BRH (**Box 9**). Pour réaliser cet objectif, il utilise une heuristique dénommée BRH multiple. L'utilisateur sélectionne d'une part une liste d'organismes présents dans ses MSAs et d'autre part un ensemble de protéomes qui serviront de référence pour valider l'orthologie (**Figure 10**). Les séquences des organismes de la première liste (i.e., "*queries*") sont utilisées pour rechercher avec BLAST les homologues correspondants dans deux ensembles : (i) la base de données de recherche pouvant être constituée de génomes, protéomes ou transcriptomes et (ii) l'ensemble des protéomes de référence. Il s'ensuit une vérification de l'orthologie entre les meilleurs hits obtenus dans les différents protéomes de référence à partir des "*queries*" sur le principe du BRH. Cette étape permet de consolider une liste de protéomes (et de séquences) de référence vérifiant au mieux les relations d'orthologie pour le MSA concerné. D'un autre côté, les séquences homologues récupérées au point (i) sont également comparées aux protéomes de référence encore en lice. Si les meilleurs hits obtenus à partir d'un homologue font partie de la liste consolidée de séquences, ce dernier est alors considéré comme orthologue. Il est à noter que la définition de l'orthologie n'est donc pas limitée par la "*query*" de départ, une séquence définie comme homologue par une "*query*" pouvant être validée orthologue par des séquences de référence déterminées par une autre "*query*". En plus de cette heuristique de vérification de l'orthologie, 42 implémente aussi une vérification taxonomique des séquences ajoutées visant à identifier ou éviter les contaminations (**Figure 10**). Suite à la détermination des séquences orthologues, chacune de celles-ci est comparée aux séquences de l'alignement pour déterminer la séquence lui étant la plus similaire. Cette dernière sert de modèle pour réaliser l'alignement de la nouvelle séquence et permettre son insertion dans le MSA, mais également à lui affilier une taxonomie. Cette taxonomie est récupérée automatiquement à partir du nom de l'organisme de la séquence la plus similaire dans l'alignement de départ et correspond à celle implémentée dans la base de données du NCBI. Pour chaque nouvel organisme à ajouter, un filtre taxonomique peut être appliqué à ses séquences orthologues. La taxonomie affiliée est alors comparée à ce filtre pouvant être inclusif (la séquence doit appartenir au clade mentionné), exclusif (ne peut pas appartenir au clade mentionné) ou les deux. Cette dernière étape permet le retrait ou le marquage des séquences ne respectant pas le filtre taxonomique demandé par l'utilisateur. Elle vise à éviter des contaminations d'origine connue ou potentiellement inconnue ainsi qu'à limiter l'ajout de séquences transférées récemment (LGT).

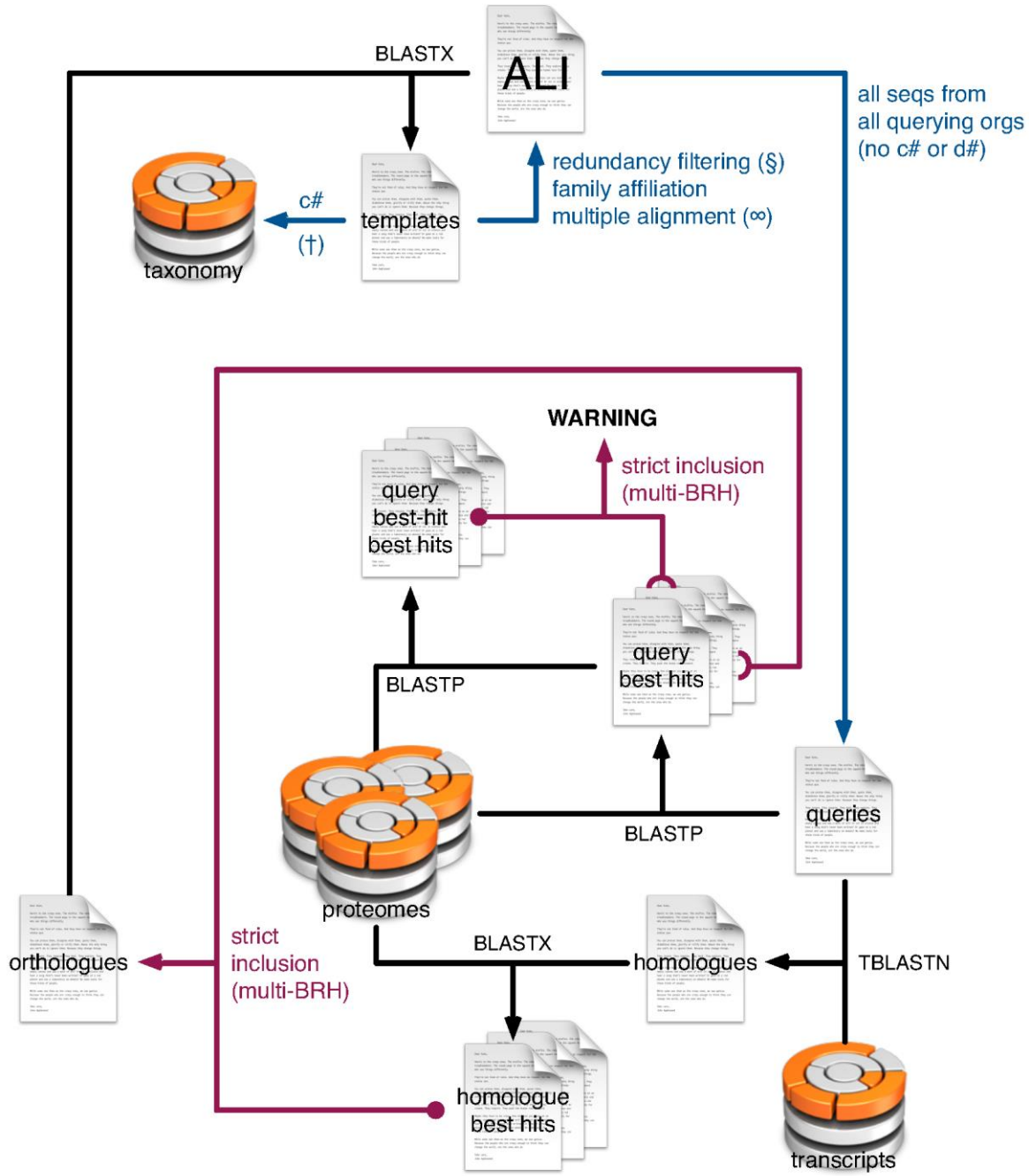

BLASTX
ALI
all seqs from
all querying orgs
(no c# or d#)
redundancy filtering (§)
family affiliation
multiple alignment (∞)
c#
(†)
templates
taxonomy
WARNING
strict inclusion
(multi-BRH)
query
best-hit
best hits
query
best hits
BLASTP
BLASTP
queries
proteomes
strict
inclusion
(multi-BRH)
BLASTX
orthologues
homologues
TBLASTN
homologue
best hits
transcripts

Figure 10. Schéma du fonctionnement de 42 (Denis Baurain, User Guide for 42)
`42` utilise le principe du Best Reciprocal Hit (BRH) afin de trouver au sein de génomes données des séquences orthologues à rajouter à des protéomes. Un BRH a lieu quand deux protéines encodées par deux gènes, chacun d'un génome différent, se retrouvent l'une et l'autre avec le meilleur score de correspondance dans l'autre génome. Dans le cas de `42`, l'ajout de séquences orthologues sur des alignements se fait via un BRH triangulaire impliquant une série de protéomes en témoins intermédiaires afin de valider le critère d'orthologie et éviter ainsi l'ajout de séquences paralogues. L'utilisateur dispose d'une série d'alignements de groupes orthologues qu'il souhaite enrichir de génomes traduits. Il extrait la liste de tous les organismes présents dans ses alignements, et en sélectionne un certain nombre représentant au mieux la diversité taxonomique de son échantillonnage : ceux sont les query_*orgs*. Il est également important que ces *query_orgs* soient présent dans un maximum d'alignements possible. Dans un premier temps, `42` extrait les séquences de tous *query_orgs* afin d'obtenir une série de *query_seqs*. Puis, il effectue un `BLASTP` de ces *query_seqs* contre des protéomes de référence. Idéalement, les protéomes de référence seront les mêmes que les *query_orgs* afin d'optimiser les scores de BLAST. Si des *query_orgs* ne sont pas disponibles en protéome, tenter de remplacer par une espèce phylogénétiquement proche. Dans un deuxième temps, un `TBLASTN` est effectué à partir des *query_seqs* mais cette fois sur les génomes de la banque à ajouter afin de trouver des séquences homologues, et donc de potentiels orthologues. Un rétro-contrôle est alors effectué à partir de ces potentiels orthologues, en effectuant un BLASTX de ces séquences sur les protéomes de référence. Cette stratégie permet d'éliminer les séquences paralogues afin de ne garder que des séquences orthologues. Pour être considéré comme orthologue, le meilleur score du BLASTX d'une séquence candidate sur les protéomes de référence doit correspondre au même que le meilleur score du BLASTP des *query_orgs* sur ces mêmes protéomes.

42 permet non seulement d'évaluer la qualité des données génomiques et protéomiques (que ce soit par l'estimation des contaminations ou par la complétude des données), mais aussi d'enrichir des groupes orthologues avec des séquences d'organismes supplémentaires. En bref, chaque séquence orthologue est classée en calculant le dernier ancêtre commun (LCA) des deux séquences du MSA qui sont les plus similaires à l'orthologue nouvellement ajouté, avec un seuil d'identité minimum de 90 %. L'acquisition d'un ensemble d'alignements de gènes orthologues sur base de données génomiques sert de point de départ à la construction d'un jeu de données phylogénomique. C'est à partir des séquences constituant ces alignements que sont recherchés les orthologues correspondants chez d'autres espèces. Par défaut, nous nous attendons à ce que le LCA inféré soit en accord avec la taxonomie du génome d'où provient l'orthologue, c'est-à-dire d'origine archéenne. Si ce n'est pas le cas, l'orthologue est considéré comme un contaminant de l'assemblage du génome (ou reste non classé s'il ne correspond à rien d'assez proche dans la base de données de référence RiboDB).

Notre but est de construire un arbre ribosomique afin de procéder à notre sélection d'espèces. Nous avons utilisé SCaFoS v1.30k (Roure et al., 2007) pour retenir 1070 génomes, tout en supprimant 7 génomes contaminés, ainsi que les génomes originaux (tous les doublons) de RiboDB. Nous avons ensuite concaténé nos MSAs de protéines ribosomiques en une supermatrice de 8344 positions AA x 1070 espèces, toujours en utilisant SCaFoS. Avant l'inférence phylogénétique, réalisée avec RAxML v8.1.17 (Stamatakis, 2014) avec une recherche par rapid-bootstrap (100 réplicas) sous le modèle de substitution rapide PROTCATLGF, nous avons supprimé les sites présentant >90% d'états de caractères manquants en utilisant ali2phylip.pl (option --max=0.1) de la distribution Bio-MUST-Core (D. Baurain ; disponible sur https://metacpan.org/pod/Bio::MUST::Core). A partir de l'arbre inféré (Tree-ribo-1070-sp), nous avons sélectionné manuellement 364 espèces représentant collectivement la diversité des

archées (archaea-364sp.lis). Sur ces 364 espèces, 305 avaient des protéomes disponibles (cf. **Supp.Mat selection-proteomes.txt**), tandis que les 59 espèces restantes n'existaient que sous forme d'assemblages de génomes non-annotés (cf. **Supp.Mat selection-genomes.txt**). Un nouvel arbre (plus petit) des protéines ribosomiques a ensuite été calculé sur la base de cette sélection d'espèces, en utilisant la même méthode et le même modèle que ceux décrits précédemment (cf. **Supp.Mat arbre-ribo-364-sp.pdf** & **arbre-racine.tre**).

4 CONSTRUCTION ET SELECTION DES GROUPES ORTHOLOGUES BASES SUR LA REPRESENTATION TAXONOMIQUE

En vue d'élargir notre sélection de gènes au-delà des protéines ribosomiques, nous nous avons utilisé USEARCH v8 (Edgar, 2010) et OrthoFinder v1.12 (Emms & Kelly, 2015, 2019) sur nos 305 protéomes sélectionnés pour générer des groupes de gènes orthologues (usearch64.8 -quiet -ublast ./Species0.fa -db ./Species0.fa.udb -evalue 1e-5 -accel 1 -threads 1 -blast6out Blast0_0.txt). Nous avons obtenu un total de 94 490 groupes orthologues (OGs) (dont 66 798 singletons), sur lesquels nous avons successivement appliqué deux filtres basés sur classify-ali.pl (Bio::MUST::Core).

5 IDENTIFICATION ET GENERATION DES GROUPES ORTHOLOGUES BASES SUR LE NOMBRE DE COPIES DU GENES PAR UNE METHODE DE DECOUPE PHYLOGENETIQUE

Nous avons eu recours à une méthode de découpe d'arbres phylogénétiques afin de trier et sélectionner nos gènes. Afin d'identifier les OGs à copie unique convenant à la concaténation, nous avons d'abord aligné nos 1007 OGs à l'aide de MAFFT (Katoh et al., 2002; Katoh & Standley, 2013) puis nous avons exécuté RAxML v8.1.17 avec une recherche par rapid-bootstrap (100 réplicas) sous le modèle de substitution PROTCATLGF (Stamatakis, 2014) sur les MSAs résultants.

Nous avons ensuite utilisé le programme root-max-div-taxon (H. Philippe, CNRS) sur les 1007 arbres obtenus. Ce programme scinde les arbres phylogénétiques selon quatre paramètres réglables par l'utilisateur :

- le nombre minimum d'espèces dans le clan 1;
- le nombre minimum d'espèces dans le clan 2 ;
- le pourcentage de branches internes plus longues que celle utilisée pour couper l'arbre ;
- le nombre minimum d'espèces partagées (= présentes dans chacun) par les deux sous-arbres.

Nous avons ainsi testé trois jeux de critères correspondant respectivement à chacun des paramètres listés précédemment :

- 40, 90, 100, 40 ;
- 90, 40, 100, 40 ;
- 90, 90, 100, 50.

Cette procédure peut ensuite être réitérée autant de fois que nécessaire tant qu'il y a moyen de couper nos arbres et extraire ainsi de nouveaux paralogues. Ainsi, un gène qui contient plusieurs paralogues est découpé au fur et à mesure en autant de séquences jusqu'à ce qu'ils soient orthologues entre eux. Par cette procédure, nous avons identifié 440 MSAs de gènes directement orthologues + 117 MSAs de gènes nouvellement orthologues (néo-orthologues).

6 ECHANTILLONNAGE TAXONOMIQUE FINAL ET SELECTION DES GROUPES ORTHOLOGUES

Pour affiner notre échantillonnage initial de taxons, nous avons utilisé SCaFoS (Roure et al., 2007) afin de créer une super-matrice de nos 440 gènes, totalisant 89 342 positions AA (cf. **Supp.Mat supermatrix-440genes.ali**). Sur cette super-matrice brute, nous avons utilisé ali2phylip.pl avec l'option avec --max=0.1 (voir ci-dessus) ce qui a donné une super-matrice de 89 342 positions d'AA alignées sans ambiguïté. Nous avons ensuite calculé un arbre en utilisant RAxML (Stamatakis, 2014) avec une recherche par rapid-bootstrap (100 réplicas) avec le modèle PROTCATLGF (cf**. Supp.Mat raxml-364-sp-440-genes**).

7 TEST DE CONGRUENCE, RETRAIT DE SEQUENCES ET CONSOLIDATION DES MSA

Après les premiers ajouts de séquences, nous avons appliqué une méthode permettant d'éliminer à la fois les mauvaises séquences (paralogues, xénologues) et les mauvais alignements (ceux pour lesquels l'orthologie de nombreuses séquences était ambiguë). Pour ce faire, nous avons comparé les longueurs de branches entre les arbres de gènes et l'arbre des espèces. Elle nécessite l'inférence d'un arbre des espèces intermédiaire qui servira de référence. Ce dernier s'obtient en réalisant l'inférence sur une super-matrice concaténant les différents alignements du jeu de données (**Figure 11**). Ensuite, on infère la phylogénie de chaque alignement utilisé pour la concaténation (une séquence par OTU) en fixant la topologie de l'arbre à celle de la super-matrice et en utilisant le même modèle. Finalement, on compare les branches de l'arbre de chaque gène à celui de l'arbre d'espèces réduit au même échantillonnage. À topologie fixe, si une séquence est forcée à une position incorrecte de l'arbre (i.e. la position taxonomique attendue, mais qui n'est pas la bonne à cause de paralogie ou de xénologie), le modèle d'évolution des séquences ne peut expliquer sa position que par un nombre important d'événements de substitution. La longueur de la branche de la séquence en question est donc augmentée de manière disproportionnée par le modèle. Par conséquent, on peut déterminer ce type de séquence en comparant la longueur de sa branche à celle attendue par l'arbre d'espèces de référence. De plus, si on regarde toutes les longueurs de branche simultanément, la présence de plusieurs longueurs de branches non-attendues dans un arbre de gène affecte la valeur de corrélation des longueurs de branches (coefficient de corrélation de Pearson R^2) entre la référence et le gène. Une faible valeur de R^2 permet donc également d'identifier des alignements contenant possiblement de nombreux paralogues ou présentant un signal phylogénétique opposé à celui de la majorité des gènes (xénologie ou grave incompatibilité de la phylogénie du gène avec la phylogénie des espèces).

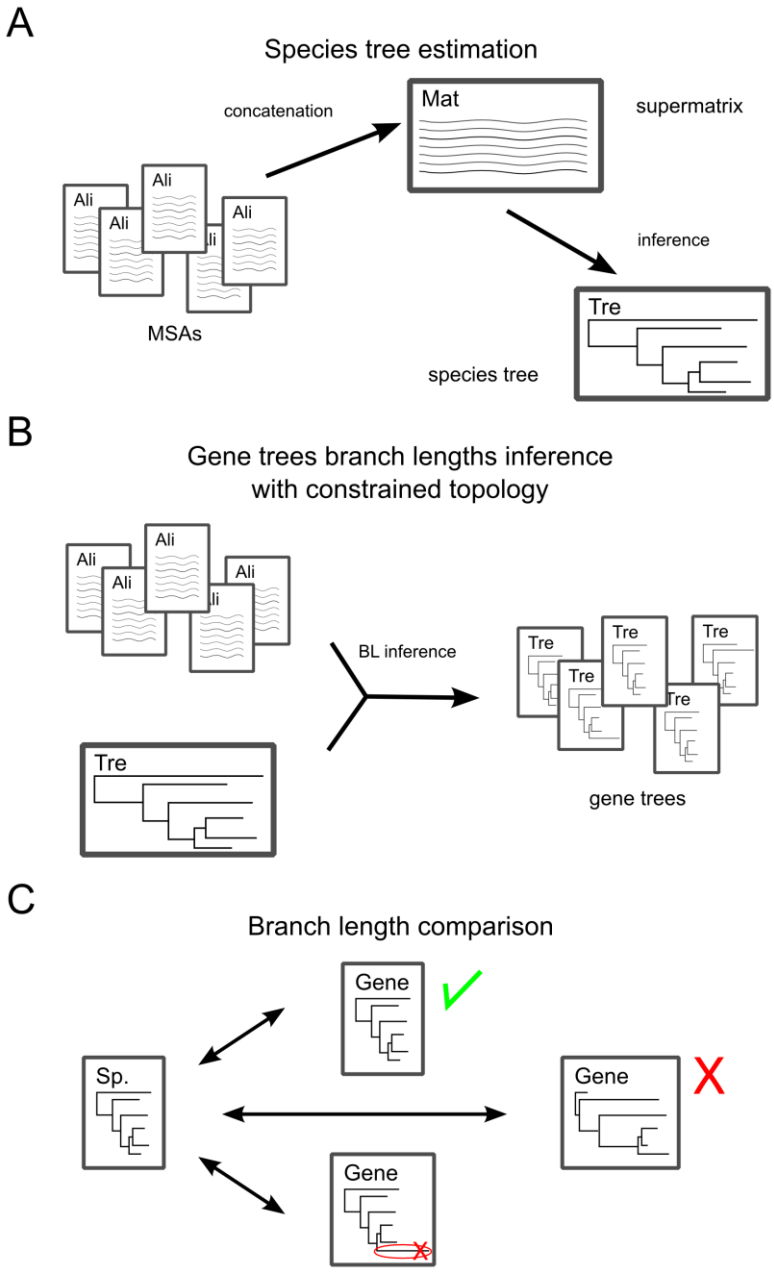


Figure 11. Protocole de décontamination par comparaison et corrélation des longueurs de branches.

A. Création d'un arbre d'espèces de référence. B. Estimation des longueurs de branches des arbres de gènes en forçant la topologie de référence. C. Comparaison de la longueur des branches entre les arbres de gènes et l'arbre d'espèces.

Concrètement, nous avons masqué les segments de séquences non-homologues avec HmmCleaner et retiré les positions de forte entropie avec BMGE (Criscuolo & Gribaldo, 2010) avant de les concaténer en utilisant SCaFoS (Roure et al., 2007), qui sélectionne par défaut la séquence la plus longue par OTU en cas de présence de paralogues résiduels. Nous avons utilisé la distance évolutive minimale comme critère de sélection parmi les séquences paralogues d'un même OTU (seuil d'élimination complète à 25% de distance entre les deux paralogues), le pourcentage maximal de sites manquants pour une séquence complète a été fixé à 10 et le nombre maximal d'OTUs manquants fixé à 25. Les arbres ont été inférés en utilisant le modèle PROTCATLGF avec RAxML (Stamatakis, 2014) avec une recherche par rapid-bootstrap (100

réplicas). Cela nous a permis de comparer les longueurs des branches terminales observées dans ces arbres de gènes individuels à celles de la super-matrice de référence correspondante afin de supprimer les séquences pour lesquelles le ratio de longueur de branche était 5 fois supérieur à celles correspondantes dans l'arbre de référence. On a ainsi enlevé 227 séquences (cf. **Supp.Mat remove_seq_branchlength_9**). Nous avons ensuite effectué le même protocole en recalculant des arbres, cette fois en calculant le coefficient de corrélation de Pearson R^2 (moyenne – 1,96 x écart-type = 0,626322348) des longueurs de branches entre les arbres de gènes individuels et la super-matrice correspondante (cf**. Supp.Mat congruence.csv**). Avec cette méthode, nous sommes passés de 440 à 416 gènes.

8 CREATION D'ORGANISMES CHIMERIQUES

Nous avons ensuite ré-évalué la complétude de nos 416 gènes. Certains organismes n'étaient pas assez représentés dans nos MSAs, mais parmi ceux-ci, plusieurs étaient étroitement apparentés. Nous avons donc créé des chimères par fusion de gènes in silico de façon que les séquences se complémentent, permettant d'avoir des organismes chimériques plus complets. Nous sommes passé ainsi de 364 à 352 espèces par fusion des espèces proches les moins représentées dans nos 416 gènes.

Suite à ces différentes étapes de nettoyage de nos MSA, la construction de la super-matrice a été finalisée en réalignant les séquences. Nous avons alors conservé les alignements possédant au moins un représentant Asgard, un DPANN, un Euryarchaeota et un TACK avant de les concaténer avec SCaFoS. J'obtins ainsi un jeu de données final de 343 gènes et 352 OTUs comprenant 94 485 positions. C'est à partir de ce dernier que j'ai évalué l'impact de l'échantillonnage taxonomique et du modèle d'évolution sur l'inférence de l'arbre des archées.

9 JACKKNIFE

Nous avons procédé à un jackknife à la fois d'espèces et de gènes. Pour ce faire, nous avons utilisé notre arbre ribosomique afin de définir des groupes monophylétiques qui représentent au mieux la diversité des archées (groups-uniq.txt). Nous avons défini nos groupes selon trois principes :

- Un groupe doit avoir un support statistique de 100 % ;
- Un groupe peut avoir au maximum 5 espèces ;
- Si un groupe a 5 espèces ou plus, on réduit ce groupe à 3 espèces.

Pour optimiser notre sélection, nous avons utilisé SCaFoS afin de favoriser les espèces qui sont présentes dans le plus grand nombre de nos OGs (autrement dit on a conservé les gènes les plus complets en organismes). Nous avons choisi de travailler sur 5 réplicas d'espèces.

10 CONSTRUCTION DES SUPER-MATRICES & INFERENCES PHYLOGENETIQUES

A partir de maintenant, les étapes suivantes ont toujours été réalisées de la même façon. Les jeux de données ont été construits de la façon suivante :

Nous avons systématiquement aligné tous nos alignements initiaux avec ali2phylip.pl implémentant le filtre BMGE avec l'option --bmge-mask = loose. Pour cela, nous avons testé dans ali2phylip.pl la meilleure valeur pour l'option --max entre 0.1, 0.2, 0.3, 0.4 et 0.5 (cf. **Supp.Mat a2p-test-max10-50.csv**). Nous avons constaté que nous supprimons un maximum de colonnes à --max = 0.1, puis qu'au-delà, peu de colonnes supplémentaires sont supprimées. Par conséquent,

nous avons gardé la valeur --max à 0.1 et la valeur --min à 0.1 afin d'éliminer les séquences de taille inférieure à 10 % de la séquence la plus longue. Par précaution, nous avons également appliqué avec Classify-ali.pl un filtre à 40 espèces afin d'être sûr de garder suffisamment d'espèces par alignement dans le cas où certaines espèces seraient venues à disparaitre. Dans le cas contraire, l'alignement est supprimé.

Nous avons également appliqué HmmCleaner. A partir des séquences des espèces cibles présentes dans les sous-arbres d'intérêt, nous avons procédé à deux grands types d'analyse : une analyse de la congruence du signal par concaténation des MSAs en super-matrice avec SCaFoS (Roure et al., 2007) et une analyse de super-arbre.

Pour chaque réplica, nous lançons plusieurs analyses (**Box 5**) :

- une analyse des 343 gènes par super-matrice et super-arbre selon les modèles LG4X, LG+C20+F+G et LG+C60+F+G (Le et al., 2012; Le & Gascuel, 2008; Si Quang et al., 2008) ;
- un jackknife de gènes pour 10 réplicas par super-matrice selon les modèles LG4X, LG+C20+F+G et LG+C60+F+G ainsi que par la méthode PMSF en utilisant l'arbre issu du modèle LG+C60+F+G comme arbre de référence.
- un jackknife de gènes pour 100 réplicas par super-matrices calculé selon la méthode PMSF en utilisant le même arbre que précédemment issu du modèle LG+C60+F+G comme arbre de référence et par super-arbres selon les modèles LG4X, LG+C20+F+G et LG+C60+F+G.

Les calculs des arbres en super-matrices ont été réalisés avec IQ-TREE 1.6.9 (Minh et al., 2020) avec Ultrafast-Bootstrap × 1000. En parallèle, les analyses de la congruence du signal basée sur la création de super-arbres ont été réalisées avec ASTRAL-III (v5.7.5) (Mirarab & Warnow, 2015), qui a produit un super-arbre pour chaque pipeline à partir d'arbres phylogénétiques simples gènes générés par IQ-TREE (LG4X, Ultrafast bootstrap) (Simion et al., 2017). Les distances de Robinson-Foulds normalisées ont été également calculées avec IQ-TREE et l'option -rf_all qui permet de calculer les distances de tous les arbres entre eux.

La recherche des partitions majoritaires de tous les arbres obtenus est facilitée par le programme parse_consense_out.pl (**Supp.Mat bilan-parse_consens_out.csv**). Pour ce faire, nous avons déterminé, à partir de nos résultats ribosomiques et de la littérature, des groupes d'espèces que l'on s'attend à retrouver (**Supp.Mat groupes.otu**).

Les arbres sélectionnés ont été automatiquement formatés par format-tree.pl (trouvé dans Bio::MUST::Core) et visualisés par iTOL (Letunic & Bork, 2021).

Box 5. Les modèles phylogénétiques

Afin de représenter le plus fidèlement possible les processus évolutifs des molécules, de nombreux modèles, relaxant telle ou telle hypothèse d'un modèle de base, ont été proposés. La pertinence de ces modèles est typiquement évaluée au travers du gain en vraisemblance qu'ils apportent, pénalisé par leur nombre de degrés de libertés, c'est-à-dire le nombre de paramètres supplémentaires qu'ils introduisent. C'est le principe sous-jacent aux tests de rapport de vraisemblance (ou LRTs pour *likelihood ratio tests* ; (Felsenstein, 1981)). Parmi ces modèles sophistiqués, nous pouvons citer :

- Les modèles sites-spécifiques qui prennent en compte le fait que la vitesse d'évolution, voire la nature des bases ou acides aminés utilisés, varie d'une position à l'autre de la molécule ;

- Les modèles non-stationnaires qui permettent de prendre en compte les variations de composition en base entre espèces et d'estimer les compositions en bases ancestrales ;
- Les modèles codons de l'ADN qui permettent notamment de détecter des gènes/sites/lignées sous un régime de sélection positive, caractérisé par un taux de substitutions non-synonymes plus élevé que le taux de substitutions synonymes ;
- Les modèles prenant en compte les différences de vitesse d'évolution entre lignées, utilisés notamment pour estimer des dates de divergence entre lignées sur la base de calibrations fossiles – on parle d'écart à l'hypothèse d'horloge moléculaire ;
- Les modèles représentant les variations du processus évolutif suivi par un site au cours du temps entre lignées (ou « covarion »), qui permettent de détecter des changements de contrainte fonctionnelle au cours de l'histoire d'une molécule.

Les modèles homogènes : le modèle GTR

Les modèles homogènes d'évolution des séquences diffèrent en fonction de deux composantes principales : la matrice d'échangeabilité R et le vecteur des fréquences d'équilibre des acides aminés Π. La matrice d'échangeabilité R décrit les taux relatifs auxquels un acide aminé change pour un autre. Dans les modèles homogènes, il est supposé que les taux de substitution sont uniformes entre tous les sites d'une séquence. Cela signifie que chaque paire de nucléotides (dans le cas de séquences d'ADN ou d'ARN) ou chaque paire d'acides aminés (dans le cas de séquences protéiques) a le même taux de substitution. Dans les modèles hétérogènes, il est admis que les taux de substitution peuvent varier entre les différents sites d'une séquence. Cela signifie que certains sites peuvent évoluer plus rapidement ou plus lentement que d'autres. Les modèles hétérogènes capturent mieux la complexité de l'évolution moléculaire en tenant compte des différences de taux de substitution entre les sites. Ils sont souvent utilisés lorsque l'on sait ou que l'on suspecte que l'évolution ne se produit pas uniformément dans toute la séquence.

Le modèle GTR (*General Time Reversible*), modèle le plus général, sous l'hypothèse d'homogénéité substitutionnelle entre sites, correspond à une matrice dont l'ensemble des paramètres, taux relatifs d'échange et probabilités stationnaires, sont considérés comme des inconnues, et sont donc inférés à partir des données (Lartillot et Philippe 2006). C'est un modèle général de substitution qui permet des taux de substitution spécifiques pour chaque paire de nucléotides ou acides aminés. Le modèle GTR utilise un seul ensemble de fréquences d'équilibre (π) pour tous les sites de la séquence. Ces fréquences d'équilibre représentent les proportions relatives des acides aminés ou nucléotides dans l'ensemble des séquences étudiées et sont constantes à travers tous les sites dans le modèle de base. GTR est un modèle flexible car il permet des taux de substitution différents pour chaque paire de nucléotides (ou acides aminés). Par exemple, les taux de substitution de A vers C, A vers G, etc., sont tous modélisés séparément.

Différence entre modèle de partition et modèle de mélange

Les modèles de mélange, comme les modèles de partition, permettent plus d'un modèle de substitution le long des séquences. Cependant, alors qu'un modèle de partition attribue à chaque site de l'alignement un modèle spécifique, les modèles de mélange n'ont pas besoin de cette information. Un modèle de mélange calcule pour chaque site sa probabilité (ou son poids) d'appartenir à chacune des classes du mélange (également appelées catégories ou composantes). Comme l'affectation site-classe est inconnue, la vraisemblance du site dans les modèles de mélange est la somme pondérée des vraisemblances du site par classe de mélange.

Par exemple, l'hétérogénéité discrète du taux Gamma est un modèle de mélange simple. Il comporte plusieurs catégories de taux avec un poids égal. IQ-TREE supporte également un certain nombre de modèles de mélange de protéines prédéfinis tels que les modèles de mélange de profil C10 à C60 (les variantes ML des modèles bayésiens CAT). Parmi ceux-ci, nous utiliserons dans cette thèse les modèles suivants :

- (C20, C60) : modèles de mélange à 10, 20, 30, 40, 50, 60 profils [Le et al., 2008a] comme variantes du modèle CAT [Lartillot et Philippe, 2004] pour la ML. Ces modèles supposent un remplacement AA Poisson et incluent implicitement une hétérogénéité de taux Gamma entre les sites.
- LG4X : Modèle à quatre matrices fusionnées avec l'hétérogénéité FreeRate [Le et al., 2012]
- LG+C20 : Application de la matrice LG au lieu de Poisson pour les 20 classes de profils AA et une hétérogénéité de taux Gamma.
- LG+C20+F : Application de la matrice LG pour les 20 classes plus la 21ème classe de profil AA empirique (évaluée à partir des données réelles) et une hétérogénéité de taux Gamma.
- +F : Fréquences empiriques des acides aminés évaluée à partir des données.
- +G : modèle Gamma discret ([Yang, 1994] avec 4 catégories de taux par défaut. Le nombre de catégories peut être changé, avec par exemple +G8.

Approches bayésiennes : PhyloBayes vs IQ-TREE

L'approche ML cherche à estimer les valeurs des paramètres du modèle qui maximisent la vraisemblance des données observées. Dans cette approche, les paramètres sont considérés comme des valeurs fixes et la meilleure estimation est celle qui rend les données observées les plus probables. Elle fournit une estimation ponctuelle des paramètres du modèle, mais ne fournit pas directement d'informations sur l'incertitude associée à ces estimations. L'approche ML n'utilise pas de priors sur les paramètres du modèle. Les estimations des paramètres sont uniquement basées sur la probabilité des données observées. En revanche, l'approche bayésienne considère les paramètres du modèle comme des variables aléatoires et les estime en calculant leur distribution postérieure conditionnelle aux données observées. Cette approche intègre à la fois les informations a priori sur les paramètres (priors) et les données observées pour estimer la distribution des paramètres. Elle fournit une estimation complète de l'incertitude des paramètres du modèle sous la forme de distributions postérieures. Cela permet de prendre en compte l'incertitude dans les prédictions et les décisions basées sur les estimations des paramètres.

Parmi les approches bayésiennes utilisant des modèles hétérogènes, le modèle CAT (pour CATégories de substitution) (Lartillot et Philippe 2004), qui est implémenté dans le logiciel PhyloBayes, est conçu pour capturer l'hétérogénéité des fréquences d'équilibre des acides aminés (ou nucléotides) à travers les sites (Lartillot et al. 2013). Chaque site de l'alignement est modélisé en utilisant un ensemble de profils de fréquences d'équilibre (π). Un profil π est un vecteur décrivant les fréquences d'équilibre des différents acides aminés (ou nucléotides) à ce site. Il y a plusieurs profils (ou vecteurs π), et chaque site de l'alignement est assigné de façon probabiliste à l'un de ces profils. Cela permet de modéliser les variations dans les pressions sélectives à travers les différents sites de la séquence. L'une des principales différences entre les modèles de substitution des acides aminés (par exemple, C60) et les modèles CAT hétérogènes au niveau du site est le nombre de catégories de fréquences d'équilibre. Le modèle CAT standard utilise un processus de Dirichlet avant d'inférer le nombre de catégories de fréquence d'équilibre, de sorte que le nombre de catégories est variable (Lartillot et

Philippe 2004, 2006). Ce modèle fonctionne comme un modèle de mélange, dont le nombre de classes, plutôt que d'être fixé a priori, est un des paramètres du problème. Par un traitement bayésien, le nombre de classes, ainsi que tous les autres paramètres du modèle, sont estimés par échantillonnage, en utilisant le principe des Chaînes de Markov Monte Carlo (MCMC). L'idée sous-jacente aux MCMC est qu'une chaîne de Markov, prenant la forme d'une marche guidée à travers l'espace multidimensionnel des paramètres, peut être utilisée pour estimer une distribution de probabilité en échantillonnant les valeurs de ces paramètres de façon périodique. L'approximation de la distribution sera d'autant plus exacte que le nombre de pas effectués par la chaîne de Markov sera élevé. Au contraire, IQ-TREE met en œuvre des modèles hétérogènes de substitution des acides aminés (Si Quang et al. 2008) avec un nombre fixe de catégories qui peut varier de 10 (C10) à 60 (C60). Lors de l'exécution d'IQ-TREE avec un modèle de catégories, le programme initialise plusieurs profils de substitution. IQ-TREE estime ensuite quel profil est le plus approprié pour chaque site en fonction des données d'alignement. Chaque site dans l'alignement des séquences est assigné à un profil particulier en fonction de sa probabilité de correspondance avec ce profil. Cette assignation est effectuée de manière itérative et probabiliste, en utilisant des algorithmes comme les algorithmes de Monte Carlo par chaîne de Markov (MCMC) ou des techniques de maximum de vraisemblance. En utilisant plusieurs profils, IQ-TREE peut modéliser plus précisément l'hétérogénéité des taux de substitution à travers les sites. Cela permet d'obtenir des estimations de phylogénie plus robustes et plus réalistes, particulièrement pour les séquences protéiques où les taux de substitution peuvent varier considérablement d'un site à l'autre.

PMSF

IQ-TREE fournit un nouveau modèle de « fréquence moyenne postérieure par site » (PMSF = *Posterior Mean Site Frequency*) comme approximation rapide des modèles de mélange de profils C10` à C60, qui demandent beaucoup de temps et de mémoire ([Le et al., 2008a] (https://doi.org/10.1093/bioinformatics/btn445) et qui peut être vue comme une variante du modèle CAT de PhyloBayes. Les PMSF sont les profils d'acides aminés pour chaque site de l'alignement calculés à partir d'un modèle de mélange d'entrée et d'un arbre guide. Le modèle PMSF est beaucoup plus rapide et nécessite beaucoup moins de RAM que les modèles C10 à C60 (voir tableau ci-dessous), quel que soit le nombre de classes de mélange. De plus, les simulations extensives et les analyses empiriques de données phylogénomiques démontrent que les modèles PMSF peuvent améliorer efficacement les artefacts d'attraction des longues branches.

Pour les nouvelles données phylogénomiques sans phylogénies encore bien acceptées, on peut d'abord estimer un arbre ML sous LG+F+Γ, qui est rapide à calculer, et utiliser cet arbre comme arbre guide pour ajuster un modèle LG+C20+F+Γ ou (encore mieux LG+C60+F+Γ) pour obtenir le profil PMSF. Ensuite, la recherche d'arbre peut être complétée sous LG+PMSF+Γ pour obtenir un arbre ML, ce qui est relativement plus rapide que d'estimer un arbre sous le mélange complet LG+C20+F+Γ. On peut aussi mettre à jour l'arbre guide de manière itérative. Comme ci-dessus, cette approche utiliserait PMSF dérivé de l'arbre guide LG+F+Γ pour obtenir un premier arbre LG+PMSF+Γ. Cet arbre LG+PMSF+Γ serait ensuite utilisé comme nouvel arbre guide, conduisant à de nouveaux PMSF et donc à un nouvel arbre LG+PMSF+Γ. Le processus peut être répété un nombre prédéterminé de fois ou bien jusqu'à ce qu'aucun changement topologique ne se produise plus. Comme on peut s'attendre à ce que l'arbre guide devienne plus précis à chaque itération, on peut

aussi s'attendre à ce que le profil PMSF se rapproche du vrai profil, ce qui devrait permettre d'estimer des phylogénies plus précises.

La PMSF désigne à la fois une méthode d'estimation et un modèle qui permet aux fréquences des acides aminés de varier selon les sites et pas nécessairement d'une manière compatible avec un mélange fini. Alors que les mélanges avec un nombre fini de classes sont nécessaires pour des raisons de calcul, l'unicité des contraintes structurelles et fonctionnelles des sites dans les protéines suggère que les vecteurs de fréquences sont mieux modélisés comme une variation continue. Les moyennes postérieures auront tendance à montrer une variation continue entre les sites ; un tel comportement a été montré pour les estimations de taux de moyenne postérieure calculées dans un mélange fini par Susko et al. (2003). La méthode PMSF fonctionne aussi bien, sinon mieux, que les modèles de mélange empiriques C20+F et C60+F pour estimer les phylogénies en présence d'hétérogénéité à travers les sites, à la fois dans les simulations et dans les études empiriques, tout en étant beaucoup plus efficace en termes de temps de calcul. En effet, un avantage de la méthode PMSF est qu'elle permet une variation continue des vecteurs de fréquences sur les sites.

Références

Felsenstein, J. Evolutionary trees from DNA sequences: A maximum likelihood approach. J Mol Evol 17, 368–376 (1981).

Nicolas Lartillot, Nicolas Rodrigue, Daniel Stubbs, Jacques Richer, PhyloBayes MPI: Phylogenetic Reconstruction with Infinite Mixtures of Profiles in a Parallel Environment, Systematic Biology, Volume 62, Issue 4, July 2013, Pages 611–615.

Nicolas Lartillot, Hervé Philippe, Computing Bayes Factors Using Thermodynamic Integration, Systematic Biology, Volume 55, Issue 2, April 2006, Pages 195–207

Lartillot N, Philippe H. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol Biol Evol. 2004 Jun;21(6):1095-109. doi: 10.1093/molbev/msh112. Epub 2004 Mar 10. PMID: 15014145.

Le Si Quang, Olivier Gascuel, Nicolas Lartillot, Empirical profile mixture models for phylogenetic reconstruction, Bioinformatics, Volume 24, Issue 20, October 2008, Pages 2317–2323

Si Quang Le, Olivier Gascuel, An Improved General Amino Acid Replacement Matrix, Molecular Biology and Evolution, Volume 25, Issue 7, July 2008, Pages 1307–1320

Si Quang Le, Cuong Cao Dang, Olivier Gascuel, Modeling Protein Evolution with Several Amino Acid Replacement Matrices Depending on Site Rates, Molecular Biology and Evolution, Volume 29, Issue 10, October 2012, Pages 2921–2936.

Nous avons également analysé en utilisant des modèles bayésiens (Lartillot & Philippe, 2004, 2006) nos super-matrices analysées précédemment avec IQ-TREE. Les analyses d'inférence bayésienne (BI) ont été réalisées à l'aide de PhyloBayes MPI 1.9a (Lartillot et al., 2013). Les modèles CAT-G et CAT-GTR-G ont été utilisés pour capturer à la fois l'hétérogénéité des taux de substitution spécifiques aux sites (modélisés par la distribution gamma qui divise les sites en différentes catégories de taux de substitution, permettant ainsi de tenir compte du fait que certains sites évoluent plus rapidement ou plus lentement que d'autres) et l'hétérogénéité des fréquences d'équilibre des acides aminés (modélisées par les multiples vecteurs de fréquences d'équilibre, ou profils, du modèle CAT). En raison de leur intensité de calcul (avec 40 ou 50 cœurs de CPU par chaîne, selon le modèle), les analyses BI ont été effectuées sur les super-matrices correspondant au réplica 2 analysées avec la méthode PMSF, avant et après ajout des eucaryotes. Toutes les analyses BI ont été réalisées avec quatre chaînes parallèles exécutées pendant 1000

cycles chacune (soit environ 12 ou 36 jours de calcul par chaîne, selon le modèle). L'examen manuel des fichiers « .trace » produits par PhyloBayes nous a permis d'écarter les 200 ou 500 premiers arbres de chaque chaîne en tant que « *burnin* », selon le modèle. Pour chaque analyse, nous avons calculé un arbre consensus par chaîne et un arbre consensus au travers des 4 chaînes, à chaque fois avec une règle de majorité de 25%, en échantillonnant 100 des arbres restants de chaque chaîne. Quel que soit le modèle, les valeurs de maxdiff calculées par PhyloBayes (bpcomp) à travers les 4 chaînes étaient égales à 1, ce qui indique que les chaînes n'avaient pas convergé selon cette mesure stricte. Néanmoins, les meandiff étaient respectivement de 0.06 et de 0.05. De plus, les échantillons de paramètres de chaque chaîne prise individuellement étaient très similaires, tout en présentant quelques différences topologiques clés.

11 SLOW-FAST

L'élimination progressive, site par site, des sites qui évoluent le plus rapidement a été réalisée en utilisant la méthode Slow-Fast (Brinkmann & Philippe, 1999). A chaque étape, un arbre phylogénétique a été calculé avec IQ-TREE (Minh et al., 2020) selon le modèle LG4X, permettant de suivre l'impact des positions supprimées sur les branches importantes. Cette approche nécessite la définition de groupes fiables pour estimer le taux d'évolution de chaque site. Les groupes correspondant à des classes ou des phyla monophylétiques indiscutables ont été considérés. Pour les Archaea, nous avons défini quatre groupes : DPANN, Euryarchaeota, Asgard, TACK. Pour la manipulation avec les Eucaryotes, nous avons séparé en deux groupes les archées et les eucaryotes.

12 ENRACINEMENT DES ARCHAEA

Box 6. Les méthodes d'enracinement

Lors de l'estimation des arbres phylogénétiques, on utilise souvent des arbres non enracinés pour calculer la vraisemblance des données. Une fois que l'arbre non enraciné optimal est déterminé, des méthodes supplémentaires peuvent être employées pour déterminer où la racine doit être placée sur cet arbre : enracinement via l'utilisation d'horloges moléculaires, distribution de la taille des branches (*midpoint rooting*, *minimal ancestor deviation*, *minimum variance rooting*), duplication de gènes, indels (insertion-délétion), distribution d'arbres de gènes non enracinés, coestimation probabiliste des arbres de gènes et d'espèces. Cependant, la méthode la plus utilisée pour enraciner les arbres en phylogénétique est la méthode du groupe extérieur (outgroup). Bien que l'utilisation d'un outgroup pour enraciner une phylogénie non enracinée soit généralement plus performante que les autres méthodes d'enracinement, la principale difficulté de cette méthode est de trouver un outgroup approprié. Les outgroups qui sont trop éloignés du groupe d'intérêt à enraciner (ingroup) peuvent avoir une évolution moléculaire sensiblement différente de celle de l'ingroup, ce qui peut compromettre la précision de la méthode.

Le principe de la poulie de Felsenstein est un concept important dans la reconstruction des arbres phylogénétiques, particulièrement lié aux probabilités des arbres enracinés et non enracinés. Ce principe indique que la position de la racine dans un arbre phylogénétique n'affecte pas les probabilités de la distribution des longueurs des branches lorsqu'on considère les relations entre les espèces terminales. En d'autres termes, le chemin de substitution d'une espèce à l'autre reste le même indépendamment de la position de la racine, ce qui permet de travailler avec des arbres non enracinés au début de l'analyse. L'utilisation des arbres non enracinés simplifie l'exploration de

l'espace des arbres possibles et accélère les calculs, car elle permet de ne pas se soucier de la position exacte de la racine initialement. Cependant, cela signifie également qu'il n'est pas possible d'imposer une ultramétrie (comme dans un chronogramme daté) car la direction du temps n'est pas définie dans un arbre non enraciné. Ce principe interagit également avec le type de modèle utilisé : il est particulièrement pertinent pour les modèles réversibles où les probabilités de transition sont symétriques.

La seule approche existante pour inclure le placement des racines dans le cadre d'une inférence probabiliste est l'application de modèles non réversibles. L'utilisation de modèles de substitution non réversibles relâche l'hypothèse fondamentale de réversibilité temporelle qui existe dans les modèles les plus largement utilisés dans l'inférence phylogénétique. En effet, cette hypothèse suppose que le processus évolutif des séquences d'ADN est réversible dans le temps. Autrement dit, les probabilités de mutation d'un nucléotide vers un autre sont les mêmes dans les deux sens. Cette condition de réversibilité permet de simplifier les modèles de substitution et facilite l'inférence des arbres phylogénétiques. Cependant, il est important de noter que cette hypothèse est une simplification et qu'elle peut ne pas toujours correspondre à la réalité biologique. En effet, des facteurs tels que la pression sélective, les biais mutationnels, et les mécanismes de correction des erreurs peuvent entraîner des taux de substitution asymétriques. L'hypothèse de réversibilité temporelle s'applique également aux phylogénies basées sur des séquences protéiques de manière similaire à ce qui est fait pour les séquences d'ADN. Dans ce contexte, les modèles de substitution des acides aminés prennent en compte les taux de remplacement entre les différents acides aminés plutôt que les bases nucléotidiques.

Références

John P. Huelsenbeck, Jonathan P. Bollback, Amy M. Levine, Inferring the Root of a Phylogenetic Tree, Systematic Biology, Volume 51, Issue 1, 1 January 2002, Pages 32–43.

Svetlana Cherlin, Sarah E Heaps, Tom M W Nye, Richard J Boys, Tom A Williams, T Martin Embley, The Effect of Nonreversibility on Inferring Rooted Phylogenies, Molecular Biology and Evolution, Volume 35, Issue 4, April 2018, Pages 984–1002, https://doi.org/10.1093/molbev/msx294

Drummond AJ, Ho SYW, Phillips MJ, Rambaut A (2006) Relaxed Phylogenetics and Dating with Confidence. PLOS Biology 4(5): e88.

James S. Farris, Estimating Phylogenetic Trees from Distance Matrices, The American Naturalist 1972 106:951, 645-668

Tria, F., Landan, G. & Dagan, T. Phylogenetic rooting using minimal ancestor deviation. Nat Ecol Evol 1, 0193 (2017).

Mai U, Sayyari E, Mirarab S. Minimum variance rooting of phylogenetic trees and implications for species tree reconstruction. PLoS One. 2017 Aug 11;12(8):e0182238. doi: 10.1371/journal.pone.0182238. PMID: 28800608; PMCID: PMC5553649.

Suha Naser-Khdour, Bui Quang Minh, Robert Lanfear, Assessing Confidence in Root Placement on Phylogenies: An Empirical Study Using Nonreversible Models for Mammals, Systematic Biology, Volume 71, Issue 4, July 2022, Pages 959–972.

Larry E. Watrous, Quentin D. Wheeler, The Out-Group Comparison Method of Character Analysis, Systematic Biology, Volume 30, Issue 1, March 1981, Pages 1–11.

Wayne P. Maddison, Michael J. Donoghue, David R. Maddison, Outgroup Analysis and Parsimony, Systematic Biology, Volume 33, Issue 1, March 1984, Pages 83–103.

Andrew B. Smith, Rooting molecular trees: problems and strategies, Biological Journal of the Linnean Society, Volume 51, Issue 3, March 1994, Pages 279–292.
James Lyons-Weiler, Guy A. Hoelzer, Robin J. Tausch, Optimal outgroup analysis, Biological Journal of the Linnean Society, Volume 64, Issue 4, August 1998, Pages 493–511.
Milinkovitch M.C., Lyons-Weiler J. 1998. Finding optimal ingroup topologies and convexities when the choice of outgroups is not obvious. Mol. Phylogenet. Evol. 9:348–357.
Philippe H, Brinkmann H, Lavrov DV, Littlewood DTJ, Manuel M, et al. (2011) Resolving Difficult Phylogenetic Questions: Why More Sequences Are Not Enough. PLOS Biology 9(3): e1000602.
Hervé Philippe, Elizabeth A. Snell, Eric Bapteste, Philippe Lopez, Peter W. H. Holland, Didier Casane, Phylogenomics of Eukaryotes: Impact of Missing Data on Large Alignments, Molecular Biology and Evolution, Volume 21, Issue 9, September 2004, Pages 1740–1752.
Delsuc, F., Brinkmann, H. & Philippe, H. Phylogenomics and the reconstruction of the tree of life. Nat Rev Genet 6, 361–375 (2005).

Nous avons mené nos analyses d'enracinement (**Box 6**) dans un cadre de maximum de vraisemblance (ML) en utilisant IQ-TREE (Minh et al., 2020). Un avantage de la ML par rapport à l'analyse bayésienne est qu'il n'est pas nécessaire d'avoir un a priori sur les distributions des paramètres, ce qui peut parfois affecter l'inférence des arbres (Cherlin et al., 2018; Huelsenbeck et al., 2002). Même si l'estimation des paramètres du modèle non réversible par maximisation de la fonction de vraisemblance semble plus intensive en termes de calcul que le calcul des probabilités postérieures (Huelsenbeck et al., 2002), l'algorithme d'IQ-TREE est suffisamment rapide pour nous permettre d'estimer le placement des racines pour des très grands ensembles de données. Ainsi, nous avons évalué la racine des archées en utilisant les valeurs de support rootstrap et l'AU-test, qui estiment la mesure dans laquelle les données soutiennent chaque branche en tant que possible racine d'un arbre phylogénétique. L'analyse rootstrap utilise des techniques de ré-échantillonnage pour évaluer la robustesse des différentes positions de la racine, tandis que l'AU-test compare statistiquement les vraisemblances de différentes hypothèses d'enracinement pour identifier la position de la racine la plus compatible avec les données. Ces analyses ont été effectuées sur notre réplica 2 d'espèces, avec et sans eucaryotes.

Nous avons calculé l'ensemble de confiance de toutes les branches susceptibles de contenir la racine de notre arbre issu du PMSF d'IQ-TREE à l'aide d'un AU-test (*Approximately Unbiased test*) (Shimodaira, 2002). L'AU-test est une méthode statistique permettant de comparer des arbres phylogénétiques et d'évaluer leur support par les données. Dans le contexte de l'enracinement des arbres, l'AU-test peut être utilisé pour comparer des arbres enracinés avec différentes positions de la racine. Les arbres non enracinés obtenus lors des analyses PMSF effectuées précédemment ont servi de topologie guide afin de guider la recherche de la racine. IQ-TREE teste toutes les positions possibles sur la topologie qu'on lui donne. Pour chaque arbre, sa vraisemblance (probabilité des données étant donné l'arbre) est calculée. Cela implique de mesurer à quel point les données supportent chaque arbre. L'AU-test compare les vraisemblances des différents arbres en tenant compte des variations dues à l'échantillonnage des données. Il utilise des méthodes de ré-échantillonnage comme le bootstrap pour évaluer la distribution des vraisemblances et calculer une p-value pour chaque arbre. Les p-values obtenues indiquent si une topologie est significativement rejetée par les données. Si cette valeur est inférieure à 0.05, la racine n'est pas sur la branche testée. Pour ce faire, nous réinitialisons notre arbre ML avec toutes les positions possibles de la racine (une position pour chaque branche) et nous calculons la vraisemblance de chaque arbre. En utilisant le AU-test, nous déterminons ensuite quels

placements de la racine peuvent être rejetés, en utilisant une valeur de seuil alpha de 5 %. Nous définissons l'ensemble de confiance des branches de la racine comme l'ensemble des branches qui ne sont pas rejetées en faveur du placement de la racine ML.

Nous avons également effectué des analyses en utilisant le rootstrap. La valeur du rootstrap est une mesure de la robustesse du placement de la racine, compte tenu du modèle et des données. Elle permet d'évaluer la stabilité et la confiance dans les différentes hypothèses concernant la position de la racine en utilisant des ré-échantillonnages bootstrap des données. L'analyse rootstrap est une méthode spécifique à IQ-TREE utilisée pour tester la position de la racine sur un arbre phylogénétique. IQ-TREE génère un grand nombre d'arbres bootstrap (ré-échantillonnés) à partir des données. Chaque arbre bootstrap est une reconstruction phylogénétique basée sur un sous-ensemble des données originales, ce qui permet de capturer la variabilité dans l'inférence des arbres. Cela permet de voir comment les données supportent chaque position possible de la racine. Les résultats montrent la fréquence avec laquelle chaque position de la racine est supportée par les arbres bootstrap. Cela donne une mesure de la robustesse et de la confiance dans les différentes positions de la racine. Pour calculer les supports rootstrap, nous avons effectué une analyse bootstrap, c'est-à-dire un ré-échantillonnage des sites de la super-matrice avec remplacement, afin d'obtenir un certain nombre de pseudo-réplicas d'arbres bootstrap. Le support rootstrap pour chaque branche de l'arbre ML est défini comme la proportion d'arbres bootstrap dont la racine se trouve sur cette branche. Comme la racine peut se trouver sur n'importe quelle branche d'un arbre non-enraciné, les valeurs du support de la racine sont calculées pour toutes les branches, y compris les branches externes. La somme des valeurs de support de la racine le long de l'arbre est toujours inférieure ou égale à 1. Une somme inférieure à un peut se rencontrer lorsqu'un ou plusieurs arbres bootstrap sont enracinés sur une branche qui n'apparaît pas dans l'arbre. En effet, certains des pseudo-réplicas d'arbres bootstrap peuvent avoir une topologie différente de celle de l'arbre de référence. Si une ou plusieurs des pseudo-réplicas bootstrap ont leur racine sur une branche qui n'existe pas dans l'arbre de référence, alors ces réplicas ne contribuent pas aux valeurs de support rootstrap des branches de l'arbre de référence, ce qui explique pourquoi la somme peut être inférieure à 1.

Une différence importante entre AU-test et le support rootstrap est que l'AU-test est conditionné à une seule topologie d'arbre de vraisemblance maximale (ML), tandis que le support rootstrap ne l'est pas. En raison de cela, ils fournissent des informations très différentes sur la position de la racine. Le test AU suppose que la topologie de l'arbre ML est vraie, puis cherche à déterminer l'ensemble de confiance des positions de la racine conditionné à cette topologie. Le rootstrap, quant à lui, ne suppose aucune topologie particulière et se demande combien de fois une position particulière de la racine apparaît dans un ensemble de pseudo-réplicas bootstrap.

CHAPITRE 1

ROOTING THE TREE OF LIFE: THE PHYLOGENETIC JURY IS STILL OUT

ROOTING THE TREE OF LIFE: THE PHYLOGENETIC JURY IS STILL OUT

*Richard Gouy[1,2], Denis Baurain[1] and Hervé Philippe[2,3]**

(1) Eukaryotic Phylogenomics, Department of Life Sciences & PhytoSYSTEMS, University of Liège, Liège 4000, Belgium.

(2) Centre for Biodiversity Theory and Modelling, USR CNRS 2936, Station d'Ecologie Expérimentale du CNRS, Moulis, 09200, France,

(3) Département de Biochimie, Centre Robert-Cedergren, Université de Montréal, Montréal, QC, Canada H3C 3J7.

* To whom correspondence should be addressed

Cite this article

Gouy R, Baurain D, Philippe H. 2015 Rooting the tree of life: the phylogenetic jury is still out. Phil. Trans. R. Soc. B 370: 20140329. http://dx.doi.org/10.1098/rstb.2014.0329

Accepted

3 July 2015

Subject Areas

evolution, taxonomy and systematics, theoretical biology

ABSTRACT

This article aims to shed light on difficulties in rooting the tree of life (ToL) and to explore the (sociological) reasons underlying the limited interest in accurately addressing this fundamental issue. First, we briefly review the difficulties plaguing phylogenetic inference and the ways to improve the modelling of the substitution process, which is highly heterogeneous, both across sites and over time. We further observe that enriched taxon samplings, better gene samplings and clever data removal strategies have led to numerous revisions of the ToL, and that these improved shallow phylogenies nearly always relocate simple organisms higher in the ToL provided that long branch attraction artefacts are kept at bay. Then, we note that, despite the flood of genomic data available since 2000, there has been a surprisingly low interest in inferring the root of the ToL. Furthermore, the rare studies dealing with this question were almost always based on methods dating from the 1990s that have been shown to be inaccurate for much more shallow issues! This leads us to argue that the current consensus about a bacterial root for the ToL can be traced back to the prejudice of Aristotle's Great Chain of Beings, in which simple organisms are ancestors of more complex life forms. Finally, we demonstrate that even the best models cannot yet handle the complexity of the evolutionary process encountered both at shallow depth, when the outgroup is too distant, and at the level of the inter-domain relationships. Altogether,

we conclude that the commonly accepted bacterial root is still unproven and that the root of the ToL should be revisited using phylogenomic supermatrices to ensure that new evidence for eukaryogenesis, such as the recently described Lokiarcheota, is interpreted in a sound phylogenetic framework.

Outline

- Introduction
- The complexity of the evolutionary process makes phylogenetic inference difficult
- Progress in modelling the heterogeneities of the substitution process
- Improved phylogenies support organismal simplification at shallow depth
- Deep phylogenetics and the prejudice of Aristotle's Great Chain of Beings
- On the persistent use of simple methods in deep phylogenetics
- Inability of current methods to prevent LBA artefacts
- Difficulty to root the ToL using anciently duplicated genes
- Conclusions

1 INTRODUCTION

Knowledge of the history of organisms is a prerequisite for the study of any evolutionary question. This explains why the evolutionary community has always been so committed to inferring phylogenies, resulting in a flood of species trees whenever new phylogenetic approaches were made available (e.g. cladistics in the 1960s; molecular data in the 1980s). More recently, the combined advances in sequencing technologies and computational methods have given a new impetus to the phylogenetic endeavour, as evidenced by the numerous studies trying to reconstruct (various parts of) the tree of life (ToL). At this point, it should be mentioned that phylogeny is only an approximation of the history of organisms. Several mechanisms are known to create full reticulations in species trees, including hybridization of related species, which is a recurrent phenomenon in numerous lineages such as flowering plants, and symbiogenesis (endosymbiosis of plastids and mitochondria), first suggested in 1905 by Mereschkowsky, albeit widely accepted only in the 1980s. Yet, exactly as Newton's law of universal gravitation is a very powerful approximation, phylogeny remains extremely useful, especially to display evolutionary relatedness, though taking into account major reticulations, such as the α-proteobacterial origin of the mitochondrion, inevitably leads to cycles (or 'rings') in the ToL. Another mechanism, horizontal gene transfer (HGT), probably plays an important role in evolution (e.g. by allowing rapid adaptation) while creating partial reticulations. Even if the latter is more difficult to display on bifurcating trees, HGT events are several orders of magnitude less frequent than vertical gene transmission (VGT). In our opinion, this justifies sticking to phylogeny as the best synthetic representation of the history of organisms [1], with horizontal gene flows shown as super imposed thin lines when really massive, such as probably for hyperthermophilic bacteria [2 – 4].

A surprising contrast appears when comparing scientific inquiries on shallow and deep phylogenetic questions. Obviously, there are many more publications on genus-level phylogenies than on domain-level phylogenies, simply because the former are much more numerous than the latter. Hence, there are more ongoing debates about, for example, the sister group of land plants or the root of the animal tree than about intra-domain phylogenies (in particular, Bacteria) and

the root of the ToL. Nevertheless, just like reconstructing the lifestyle of Magdalenians is more difficult than studying the habits of the Victorian era, inferring the deepest branches of the ToL is highly problematic. Consequently, this issue should still be a very hot topic, a topic that can be tackled only by the application of the most sophisticated and up-to-date methodology. Yet, a reality check shows that it is not the case. Instead, one of the most frequently cited references on the matter is a 25-year-old paper by Carl Woese and co-workers [6] (more than 3200 citations in Web of Science). For instance, the recent article describing the fascinatingly complex Lokiarchaeota [5] interprets them as an intermediate stage of the eukaryogenetic process, based on the bacterial rooting of the ToL that was inexplicably set in stone by that paper of Woese et al. [6]. Notably, Woese's tree (their fig. 1) also shows Microsporidia as the sister group of all remaining eukaryotes. Therefore, genomic data of the microsporidium Mitosporidium daphniae, especially its mitochondrial genome [7], could be, according to the same principle, interpreted as evidence that Mitosporidium represents an intermediate stage in the complexification of an ancestrally simple microsporidium into a complex eukaryote. However, thanks to their awareness of more recent references accounting for the heated debate that eventually led to recognize Microsporidia as Fungi [8,9], Haag and co-workers instead correctly interpret Mitosporidium as an intermediate stage in the simplification of a complex ancestral fungus into a simple microsporidium. Likewise, the complex Lokiarchaeota would be better interpreted as an intermediate stage in the simplification of a complex ancestral eukaryote into a simple prokaryote, provided that the root of the ToL turned out to lie on the branch leading to Eukaryota, an unorthodox hypothesis that has never been convincingly rejected [10]. Importantly, such a scenario would not imply that complex eukaryotic cells were created out of nowhere, but simply that all intermediates have disappeared. Put in another way, genuinely simple organisms did exist at some point in the past, but without leaving any extant offspring. Hence, a eukaryotic rooting is compatible with an ancestral (now extinct) prokaryotic life form.

In this paper, we first review the technical difficulties hindering phylogenetic inference as well as the recent methodological progresses on the matter, using the relatively recent (shallow) evolution of animals as our main case in point. Then, we explore the (sociological) reasons underlying the limited interest in accurately solving the rooting of the ToL, which is nonetheless fundamental to our understanding of prokaryogenesis and eukaryogenesis. Finally, we explore the potential avenues for a resolution of this issue.

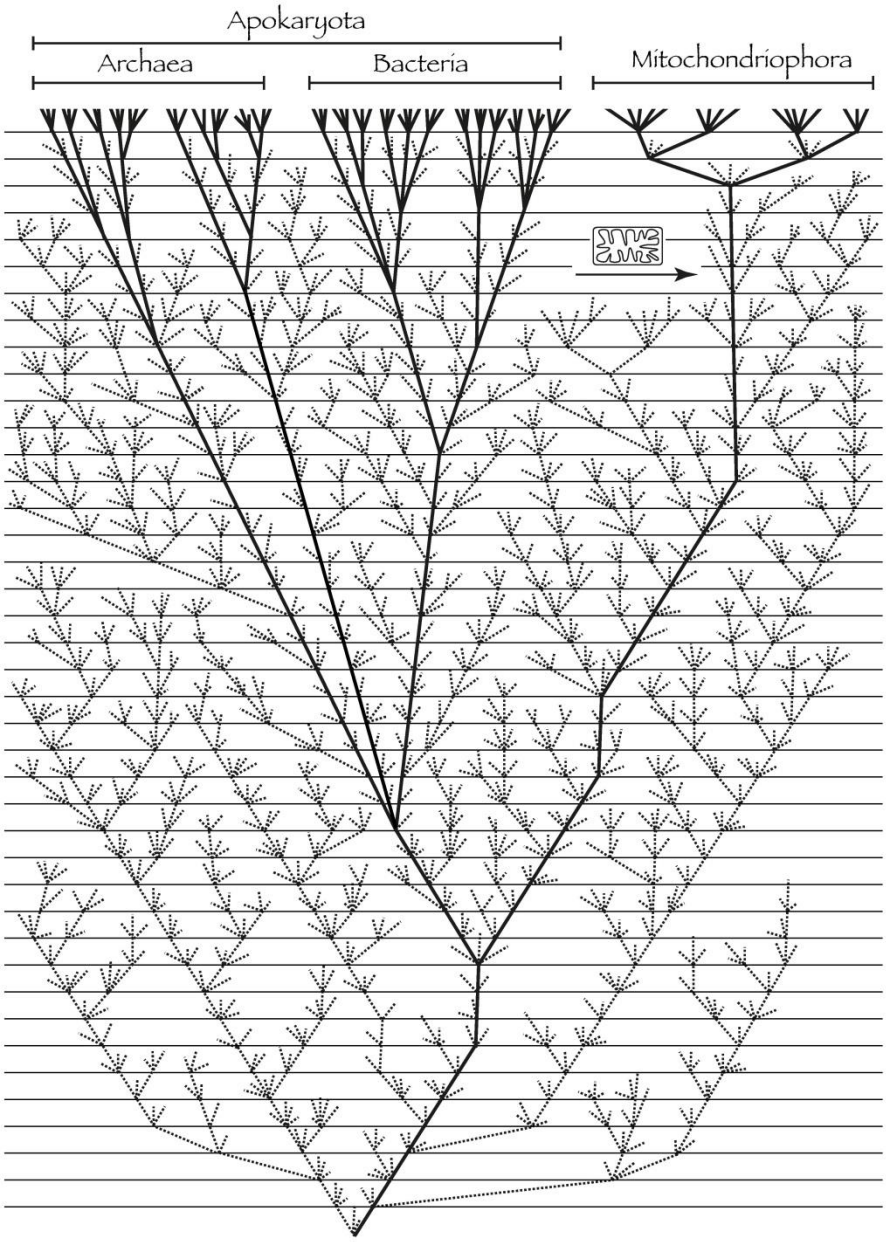

Apokaryota
Archaea
Bacteria
Mitochondriophora

Figure 12.

Words shape our minds. In this hypothetical ToL, Archaea and Bacteria form a monophyletic group (Apokaryota), derived from a nucleated ancestor through secondary simplification and concomitant loss of the nucleus. Present-day Eukaryota are named Mitochondriophora after their defining feature, the mitochondrion. Consequently, the last universal common ancestor (LUCA) would have belonged to Karyota (nucleated cells), whereas Prokaryota have probably existed before the advent of the nucleus. Even if apparently unorthodox, such a scenario is currently ruled out only by the power of Aristotle's prejudice and not by hard evidence. On the contrary, the shallow parts of the ToL are replete with secondary simplified lineages (e.g. Microsporidia, apicomplexans, acoelomorph worms, tunicates), which makes a eukaryotic root of the ToL rather more plausible than not. It is also important to note that the vast majority of ancient lineages probably went extinct [79], meaning that our sampling of biodiversity is highly biased. Figure drawn by Rosa Gago.

2 THE COMPLEXITY OF THE EVOLUTIONARY PROCESS MAKES PHYLOGENETIC INFERENCE DIFFICULT

A striking characteristic of phylogenetics, especially irritating for non-specialists, is that the ToL 'evolves' (i.e. the names and contents of clades change over time) and that several mutually incompatible solutions often coexist over long periods of time. The simplest explanation is that phylogenetics is an active field of science, which in itself is a positive fact. Importantly, contrary to the naive, yet commonly held view that open problems eventually get solved through the accumulation of more sequence data, incongruencies still persist in the genomic era (e.g. for streptophytes [11 – 16]) or Bilateria [17 – 22]). Indeed, while phylogenomics helps in decreasing stochastic error (due to small sample sizes), it actually makes systematic error more apparent. Systematic error stems from methodological biases (i.e. model violations in a probabilistic framework) that cause the inference to converge towards an incorrect solution as more and more data are added. The most wellknown case of this phenomenon is the infamous long branch attraction (LBA) artefact, which was originally formalized to demonstrate the inconsistency of maximum parsimony when branch lengths are sufficiently unequal [23]. And even today, in spite of the widespread use of sophisticated methods and evolutionary models, numerous incongruencies in phylogenomics are still associated with long branches, corresponding to fast-evolving lineages and/ or distant outgroups (e.g. Nematoda [24 – 29]), Ctenophora [22,30,31] or Zygnematales [11,12,15,16]).

This difficulty is due to the formidable complexity of the underlying evolutionary process. Hence, all existing models, even the most sophisticated and computationally demanding ones, remain dramatically oversimplified. Phylogenetic inference can be schematically separated into three steps: (1) homology assessment, i.e. identifying (1a) homologous genes through database similarity searches and (1b) homologous positions through multiple alignment; (2) modelling of the substitution process, in order to detect the multiple substitutional events falling at the same positions (i.e. estimating the probabilities of mutation and fixation) and to infer the gene tree; and (3) inference of the species tree from the gene trees, i.e. taking into account incomplete lineage sorting (ILS), HGT and gene duplication/conversion. In theory, the three steps should be performed simultaneously, but this is computationally intractable (see the article of N. Lartillot in this issue [32]). In practice, they are thus performed separately, even if a few software packages are available for the joint inference of steps (1) and (2) [33,34] or steps (2) and (3) (see the article

of B. Boussau and colleagues in this issue [35]). Nevertheless, computational limits imply that the joint evaluation of two or more steps is performed at the expense of using relatively simple methods within each step. For instance, the PHYLDOG software uses both a simplistic substitution model (homogeneous over time and across sites) and an incomplete gene history model (e.g. no gene conversion) [36]. To our knowledge, the relative performance of these joint approaches and of the well-established supermatrix methods (which assume that steps (1) and (3) have been already solved) has not yet been carefully evaluated, in particular for ancient questions. Our bet is that the assessment of homologous characters (especially thanks to the removal of ambiguously aligned regions) and of orthologous genes is relatively accurate and does not constitute the most important issue in deep phylogenetics. In addition, supermatrix-based inference appears to be robust to the inclusion of paralogous [37] and xenologous (i.e. horizontally transferred) sequences (unpublished results), but sensitive to the substitution model (see below).

Therefore, from now on, we focus on the supermatrix approach (which we consider as the best one currently available, even if we acknowledge its limitations) and on the modelling of the substitution process.

3 PROGRESS IN MODELLING THE HETEROGENEITIES OF THE SUBSTITUTION PROCESS

It is necessary to model the substitution process because, at geological timescales, successive substitutions at the same position are the rule. These multiple substitutions first blur then erase and rewrite the original phylogenetic signal, and the resulting homoplasy prevents naive methods, such as similarity-based distances and maximum parsimony, from being consistent. Unfortunately, the substitution process is highly heterogeneous, both across sites and over time, thus making its efficient modelling particularly difficult. First, the mutational process varies across positions (e.g. the hypermutable methylated CpG) and over time (due to e.g. evolutionary changes in the efficiency of the DNA repair machinery). Second, and probably more importantly, the fixation probability of any given possible mutation also varies across sites, owing to functional constraints on the encoded products, and over time, mainly because of variable effective population size, changes in epistasis and variable environment.

The very first substitution model ever developed [38] made numerous assumptions of homogeneity and independence that simplified computation, only branch lengths being heterogeneous (i.e. the global substitution rate was allowed to vary). Since then, three major and three minor, yet significant, improvements have been proposed:

1. Heterogeneity of substitutions among character states. Some substitutions are obviously easier than others (e.g. transitions versus transversions or Asp ! Glu versus Asp ! Trp) and exchangeability matrices were rapidly introduced [39]. The General-Time-Reversible (GTR) model is now widely used for nucleotides, where it only requires eight parameters, but much less for amino acids because then it requires 208 parameters. Yet, when datasets are large, an amino acid GTR matrix has a better fit than empirical matrices (e.g. WAG and LG) [40].
2. Heterogeneity of the substitution rate across sites. Following the seminal observations of Uzzell & Corbin [41], various methods have been developed to handle the fact that some sites are more susceptible than others to accumulate

substitutions, and thus to generate artefacts. The gamma distribution appears as a good compromise between computational efficiency and biological realism. That is why it is now widely used. More refined models (such as mixture of gamma or Dirichlet processes) might nevertheless prove to be useful for solving difficult questions.

3. Heterogeneity of the substitution process across sites. The fact that only a few amino acids are possible at a given position (e.g. charged or hydrophobic amino acids) was established by biochemists a long time ago, but it has attracted the attention of phylogeneticists only recently [42,43]. This is surprising because the efficiency of the detection of multiple substitutions is much higher when the number of possible character states is reduced [25]. CAT-like models [43] use a Dirichlet process to affiliate individual sites to different CATegories defined by their character state frequencies. With hundreds to thousands of categories usually inferred in a posteriori analyses, the observed heterogeneity is very high, demonstrating both the biological relevance and the statistical significance of accounting for this aspect of the evolutionary process. As expected, the CAT– GTR model, and to a lesser extent the CAT model, has a much better fit to data, provided that a few thousand sites are considered. Accordingly, these models are also less sensitive to homoplasy and LBA artefacts [22,25].
4. Separation of mutation and selection steps. Codon models were proposed as early as 1994 [44,45]. Owing to their mechanistic modelling that contrasts with the phenomenological modelling of all other protein models, they are biologically more realistic. Yet, their computational slowness (due to the 61 × 61 matrix), combined with numerous simplifying assumptions, so far has limited their usefulness for phylogenetic inference. Nevertheless, recent improvements, in particular their coupling with the CAT model [46], make them promising.
5. Heterogeneity of composition over time. The existence of a compositional bias and its implication in reconstruction artefacts was also identified more than 20 years ago, based on ribosomal RNA alignments [47 – 49]. Various modelling approaches [47,50,51] have been proposed, but these are often computationally demanding. However, since the compositional bias is dominant at large evolutionary scales, it is better to address it when inferring deep phylogenies [8,52].
6. Heterogeneity of rates within positions over time. Because of epistasis, the probability of accepting a mutation at any given position is expected to vary along the branches of the tree, as demonstrated early on by Fitch & Markovitz [53]. In the 1990s, a renewed interest in the so-called 'heterotachy' led to the development of multiple models [54 – 56]. Surprisingly, however, the increase in statistical fit, albeit systematic, is not very important, and their impact on topology rather marginal [57].

Despite these significant improvements, incorrect phylogenies keep being published due to uncontrolled artefacts. This is because many problems remain to be solved. First, not all these improvements are jointly incorporated into a single model, the best models combining at most four out of six improvements at the expense of being tractable only for small datasets [50]. Since

the first three are included in PHYLOBAYES [58], it is probably the most accurate and computationally tractable software available today. Second, numerous improvements are still needed to address the full spectrum of biological complexity. For instance, heteropecilly (the change of the substitution process at a position over time) is known to make the CAT model inconsistent [59]. Another example is the non-independence of sites, with the few models relaxing this assumption showing a better fit to data [60]. Importantly, future models should not try to account for all the subtleties of the evolutionary process but instead focus on the heterogeneities that are the most prone to generate phylogenetic artefacts.

4 IMPROVED PHYLOGENIES SUPPORT ORGANISMAL SIMPLIFICATION AT SHALLOW DEPTH

These methodological improvements, along with enriched taxon samplings (sometimes the only way to avoid artefacts), better gene samplings and clever data removal strategies, have led to numerous revisions of the ToL, especially at an intermediate evolutionary scale (e.g. within Metazoa [17,18,61,62]). Strikingly, a major trend is visible in these revised phylogenies: morphologically simple organisms, once considered as akin to ancestral intermediates ('living fossils') in a gradual rise towards complex organisms, are often relocated within groups of complex organisms, thus implying that their simplicity is not primitive but secondarily derived. In eukaryotes, 'Archezoa' (e.g. Microsporidia, Diplomonadida and Trichomonadida), which had been first recovered at the base of the rRNA tree [6], in apparent agreement with their lack of a mitochondrion, eventually turned out to be located (much) higher in the tree [8,9] and to possess degenerated mitochondria [63]. In animals, the very simple Myxozoa now appear to be closely related to Medusozoa [64], while acoelomate Platyhelminthes [24,25,27 – 29] and Acoelomorpha [65] have been shown to be closely related to Lophotrochozoa and Ambulacraria, respectively. Moreover, the mostly dull Urochordata are more closely related to Vertebrata than are the more complex Cephalochordata [66]. For all these phylogenetic errors, the methodological explanation is the same: morphological simplification is generally accompanied by an acceleration of the molecular evolutionary rate and by qualitative shifts in the substitution process. When simple models are used, this situation generates artefacts where the long branch of the (often distant) outgroup attracts the long branch of the simplified organisms, which erroneously results in a too basal location of the latter in the trees.

5 DEEP PHYLOGENETICS AND THE PREJUDICE OF ARISTOTLE'S GREAT CHAIN OF BEINGS

This rapid overview of relatively recent phylogenies (i.e. within Eukaryota, which corresponds to a sub-clade of a-Proteobacteria, itself a sub-clade of Proteobacteria, itself a sub-clade of Bacteria) demonstrates that sophisticated approaches (and especially substitution models handling multiple heterogeneities) are mandatory for accurate phylogenetic inference and that morphologically simple organisms are the most difficult to correctly locate. These results have profound implications for deep phylogenies, which are by essence much more difficult to infer due to increased noise (more multiple substitutions, HGTs and heterogeneities) and to decreased signal (less homologous positions). Consequently, artefacts are much more likely to

occur, especially when trying to position the simple prokaryotes (Archaea and Bacteria) with respect to Eukaryota.

Surprisingly, despite the flood of genomic data available since 2000, there has been almost no interest in inferring the root of the ToL (a dozen papers [67]) and only limited interest in the relationships within Bacteria and Archaea. More puzzlingly, with a few notable exceptions [52,68], these studies were almost always based on methods dating from the 1990s that have been shown to be inaccurate for much more recent questions! While a careful sociological study would be required to understand this baffling behaviour, our opinion is that it stems from the subliminal prevalence of Aristotle's Great Chain of Beings, reinforced by the progressivism of the Age of Enlightenment, and from humans' inclination for trends and 'stories that go somewhere', as pointed out by Gould [69]. An illustration of the strength of this prejudice is the recurrent use of scale-related wordings such as 'higher plants' or 'lower animals', a few per cent of manuscripts submitted to evolutionary journals comprising this inappropriate terminology (H. Philippe 2015, unpublished data). Another one is that assertions such as 'eukaryotes arose from prokaryotes' [70] are commonplace, whereas the evidence for this stance is both scarce and weak [10].

Aristotle's prejudice is constantly revived by the fact that language shapes thought [71], an idea also known as the linguistic relativity principle (or Sapir– Whorf hypothesis) and that can be traced back to Wilhelm von Humboldt [72]. In particular, the words 'prokaryotes' (before nucleus) and 'eukaryotes' (true nucleus) make us more prone to accept that the former have preceded the latter, and thus to focus our attention on the origin of eukaryotes. Pace has made much of the idea that the word 'prokaryote' imposes a certain temporal directionality on the prokaryote/eukaryote dichotomy [73,74]. Two concepts were initially distinguished within the prokaryote– eukaryote dichotomy when R. Y. Stanier and C. B. van Niel introduced the concept of prokaryote in the early 1960s. The first one was organizational and referred to comparative cell structure, whereas the second one was phylogenetic and referred to a natural classification of the living world [75,76]. Thus, the definition of prokaryote is blurred. Do prokaryotes lump extant organisms without nuclear membranes (Archaea and Bacteria)? Or do they refer to some long-gone ancestors of eukaryotes? These are two different matters [77]. The last one is misleading for it gives a direction to evolution and allows us to think that extant eukaryotes emerged from 'prokaryotes' that still exist, so that eukaryotes are more 'evolved' than prokaryotes. As a case in point, searching for 'eukaryogenesis' in PubMed returns 53 articles (as of May 2015), while the related terms 'prokaryogenesis', 'bacteriogenesis' and 'archaeogenesis' do not yield any result. This is significant because, whatever the correct theory is, both eukaryogenesis and prokaryogenesis (including bacteriogenesis and archaeogenesis) have occurred during the evolution of life on Earth. Therefore, only a scenario that adequately addresses the two issues would be completely satisfactory. Indeed, the temptation to justify the lack of research about prokaryogenesis by equating the latter to the origin of the living cell not only takes the prefix 'pro' of prokaryotes in the literal meaning, but also lends credit to the mistaken view that contemporaneous Bacteria and Archaea are long-standing intermediate stages (i.e. surviving stem groups) on the path to Eukaryota.

To become aware of how wording reinforces Aristotle's prejudice, it is insightful to fantasize an alternative history of science, in which Edouard Chatton would not have coined the name 'Prokaryota'. Instead, let us imagine that, impressed by the works of Mereschkowsky on

endosymbiosis and of Lwoff [78] on simplification in unicellular organisms, he would have proposed the evolutionary scheme shown in **Figure 12**. Assuming that simple cells devoid of nucleus were derived from complex nucleated cells, he would have named them 'Apokaryota'. Moreover, building on the idea that extant nucleated cells diversified after the mitochondrial endosymbiosis, he would have named the latter 'Mitochondriophora', reserving the names 'Karyota' for the common ancestor of all extant organisms and 'Prokaryota' for a hypothetical ancestor of Karyota devoid of nucleus. Had we used Apokaryota and Mitochondriophora instead of Prokaryota and Eukaryota, it is likely that our view of the evolution of life would have been quite different: 'Mitochondriophora arose from Apokaryota' being meaningless. Of course, this would not have prevented some researchers from arguing that Apokaryota are in fact ancestral to Mitochondriophora, exactly as some have proposed that Eukaryota actually preceded Prokaryota, the burden of the proof being just transferred on different shoulders.

no. studies	artefact awareness	taxon sampling	site removal	heterogeneous model
shallow phylogenies				
44				
6	y			
4	y			y
2	y		y	
4	y		y	y
2	y	y		
1	y	y		y
6	y	y	y	y
69	25	9	12	15
deep phylogenies				
41				
3	y			
7	y			y
2	y		y	
2	y		y	y
1	y	y	y	
1	y	y	y	y
57	16	2	6	10

Tableau 3.

Comparative bibliographical survey on tree reconstruction practices in studies dealing with shallow (i.e. metazoan evolution) versus deep (i.e. ToL root and archaeal/bacterial evolution) phylogenetic issues.

6 ON THE PERSISTENT USE OF SIMPLE METHODS IN DEEP PHYLOGENETICSON THE PERSISTENT USE OF SIMPLE METHODS IN DEEP PHYLOGENETICS

By looking at the phylogenetic studies published over the years, we are under the impression that the community shows a disproportionate interest in using ever more sequence data compared to using improved methods. Moreover, as aforementioned, this trend appears stronger for colleagues studying deep phylogenetic issues than for those interested in shallower questions. To flesh out this intuition, we searched Web of Science for phylogenetic studies

published since 2005 and addressing either shallow or deep evolutionary issues. Our exact queries were 'phylogenet* AND metazoa*' and 'phylogenom* AND (Bacteria OR Archaea)', respectively. After a first screening of the numerous irrelevant articles, this allowed us to download two sets of PDF files: 93 about shallow phylogenies and 137 about deep phylogenies. We then examined each paper in turn to determine: (1) whether it was relevant for establishing our statistics about tree reconstruction practices; (2) whether the authors demonstrated an awareness of possible phylogenetic artefacts (through the use of keywords such as 'long-branch attraction/LBA', 'artifact/artefact', 'non-phylogenetic signal', 'systematic error', 'homoplasy', 'saturation'); and (3) whether they had tried to reduce the systematic error by applying one of the three well-known approaches summarized in, for example, Philippe et al. [22]. As a reminder, these strategies are: (3a) varying the taxon sampling (e.g. inferring phylogenies with and without outgroups and/or fast-evolving lineages, replacing rogue organisms by slow-evolving relatives), (3b) removing fast-evolving (and/or biased) sites, based on preliminary rate or compositional analyses and (3c) using sophisticated substitution models (defined here as models heterogeneous across sites, such as CAT-like models, or over time, such as heterotachous/covarion models). The results of this quick bibliographic survey, limited to the relevant studies (69 'shallow studies' and 57 'deep studies'), are shown in **Tableau 3** (see also the electronic supplementary material, tables S1 and S2 for individual paper analyses).

Strikingly, less than half the studies showed awareness of possible artefacts. In particular, only 36% (25/69) of the publications dealing with shallow phylogenies mentioned any of the key words of our list, while the situation was slightly worse for papers about deep phylogenetic issues (16/57 ¼ 28%). Among the 'shallow studies' that effectively cared for artefacts, 76% (19/25) tried to do something to reduce the systematic error, a figure that was similar among 'deep studies' (13/16 ¼ 81%). In both cases, the most common strategy was to use a heterogeneous substitution model (15/25 ¼ 60% and 10/16 ¼ 62%), an efficient approach that is also the easiest to implement. By contrast, site removal strategies were more often applied in 'shallow studies' (12/25 ¼ 48%) than in 'deep studies' (6/16 ¼ 37%), whereas varying the taxon sampling was three times more explored in 'shallow studies' (9/25 ¼ 36%) than in 'deep studies' (2/16 ¼ 12%), a low figure that might be due to the lack of alternative outgroups at the domain level. Interestingly, six publications (24%) dealing with shallow phylogenies did use the three approaches for controlling the artefacts, while only one publication (6%) trying to infer the ToL [80] was equally comprehensive according to our criteria. Altogether, our modest survey confirmed our initial intuition and indicated that there was room for improvement in deep phylogenetic inference without the need for any additional methodological development. This is especially true for studies dealing with issues buried deeply in the ToL, where model violations, and thus artefacts, are expected to be much more frequent.

Rooting the ToL cannot be achieved using an outgroup. A clever way to get around this problem is to resort to universal duplicated paralogous genes, namely genes duplicated before the last universal common ancestor (LUCA), which are present in at least two copies in the three domains of life [81 – 83]. Half a dozen of such gene pairs were identified and put to use in the 1990s, most often with methods that we now consider as inaccurate. As a consequence, conflicting results were obtained (see table 1 in [67]). In 1999, one of us (H.P.) published several papers on the rooting of the ToL, one of them introducing a new method (the S/F method) that hinted at a

possible eukaryotic root [84]. When looking at the subsequent publications citing this work (see the electronic supplementary material, table S3 for individual paper analyses), an interesting pattern appears: the majority of the citations are due to the new method and not to the unorthodox result. Hence, for the 121 citations of Brinkmann & Philippe [84] that we analysed in detail, 83 (69%) referenced the S/F method (designed to remove fast-evolving sites in the hope of reducing artefacts), whereas 23 (19%) quoted it for a possible monophyly of prokaryotes associated with a eukaryotic root, and 16 (13%) for its point about the difficulty to root the ToL. This demonstrates that the S/F method is widely recognized as useful to avoid artefacts, even in shallow phylogenies. Therefore, it is surprising that the results of its application to deep phylogenies are ignored to the advantage of those obtained with very simplistic methods (e.g. without any heterogeneity across sites [81,82]). To make it clear, our point here is not to claim that the S/F method is adequate to locate the root of the ToL (see [59] for a recent criticism of fast site removal) nor that prokaryotes are indisputably monophyletic, but rather to emphasize the fact that many researchers have preferred results based on clearly inadequate methods over results based on improved methods. In our opinion, this paradox is to be attributed to the power of what we dubbed above 'Aristotle's prejudice' and that has permeated so much our way of thinking that claims in favour of simple ancestors are readily accepted, whereas opposite views betting on complex ancestors are swiftly discarded for the lack of very strong empirical evidence.

7 INABILITY OF CURRENT METHODS TO PREVENT LONG-BRANCH ATTRACTION ARTEFACTS

To show how sticking to simple methods in deep phylogenetics is doomed to failure, we illustrate that artefacts easily keep occurring with the sophisticated inference methods available today, even for shallow questions. Let us examine the tree of Bivalvia in the presence of Gastropoda, two molluscan groups whose monophylies are well established. To trigger the artefacts, we chose to study concatenated mitochondrial proteomes, because these of some Bivalvia (Pteriomorphia) have evolved much faster than those of others (Unionoida), and to include outgroups of decreasing relatedness (from Annelida to Fungi). As shown in figure 2, all models perform equally well as long as the outgroup is close (Annelida), but become sensitive to LBA when the outgroup distance gets larger, either due to old divergence (Fungi) or to fast evolutionary rate (Hymenoptera). As expected, site-heterogeneous models (CAT þ G and CATGTR þ G) perform slightly better than site-homogeneous models (LG þ G and GTR þ G).

However, the key difference here is not the substitution model used, but the taxon sampling (outgroup distance), which is precisely the parameter that is almost fully constrained when rooting the ToL (owing to the existence of only three domains and a few anciently duplicated genes). Several important model violations are known to affect mitochondrial genes: (i) heterogeneous amino acid composition across taxa [50], (ii) heterotachy [86] and (iii) heteropecilly [59]. These model violations are due to variations in the substitution process over time and initially stem from a change in functional constraints (e.g. relaxed selection). This means that long branches not only retain less phylogenetic signal but also bear a misleading signal, hence the observed LBA artefacts.

This illustrates how easily our best phylogenetic methods (here Bayesian inference under the CATGTR + Γ model) can still be misled when model violations are large. In fact, this is precisely

what happens when one tries to root the ToL [87]: the outgroup is incredibly distant (i.e. a paralogous gene with a very different function, which favours heterotachy and heteropecilly) while substitution rates for any marker are far from constant over billions of years. To this respect, we do expect major accelerations for informational genes on the branch connecting Mitochondriophora (Eukaryota) to Apokaryota (Archaea + Bacteria), at the very least because of the absence/presence of transcription/translation coupling. Other events, such as the adaptation to hyperthermophily or the (possible) loss of the nucleus, should also have led to major shifts of the functional constraints and thus to drastic changes in the evolutionary properties of each site over time. Considering that Bacteria always display an extremely long branch in unrooted gene trees [87] and that current methods are unable to resolve similar but much more recent issues (such as the monophyly of Bivalvia, **Figure 13**), it is rather perplexing that the traditional bacterial rooting is taken for granted by so many colleagues in the field.

outgroup	model topology	LG	GTRCAT	CAT	GTR
Annelida	Bivalvia	95	96	98	100
	LBA	5	4	0	0
Hymenoptera	Bivalvia	1	0	5	0
	LBA	99	100	95	99
Maxillopoda	Bivalvia	93	93	100	100
	LBA	7	7	0	0
Myriapoda	Bivalvia	81	79	98	100
	LBA	19	21	2	0
Echinodermata	Bivalvia	62	54	100	100
	LBA	38	46	0	0
Porifera	Bivalvia	39	39	98	99
	LBA	61	61	1	0
Fungi	Bivalvia	0	0	5	0
	LBA	100	100	94	100

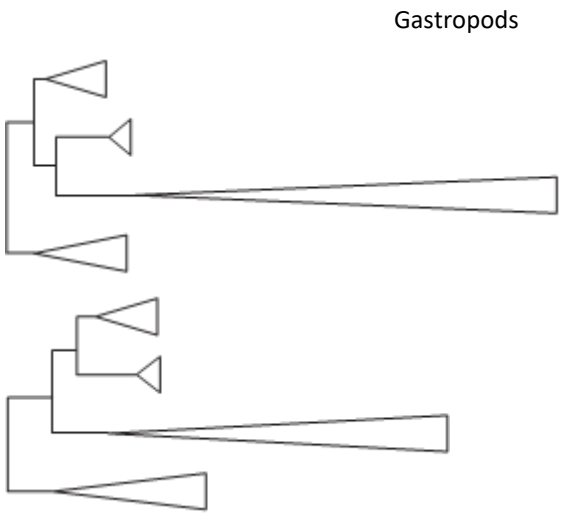


Figure 13.

Our best substitution models cannot yet address difficult phylogenetic issues, even at shallow depth. We assembled supermatrices by concatenating the translated mitochondrial genomes (12 genes) of nine slow-evolving bivalves (Unionoida), nine fast-evolving bivalves (Pteriomorphia), nine gastropods and nine outgroups. Seven different outgroups were considered, thus resulting in seven different supermatrices, each one containing 36 species and 2016 unambiguously aligned amino acid positions. We then analysed all supermatrices using RAXML [85] and PHYLOBAYES [59] under four different substitution models: LG + Γ, GTR + Γ, CAT + Γ and CATGTR + Γ. Bootstrap proportions (LG and GTR) and posterior probabilities (PPs; CAT and CATGTR) for the monophyly of bivalves (upper tree) and for an alternative (LBA) topology, in which fast-evolving bivalves are attracted by the outgroup (lower tree), were computed from 100 bootstrap pseudo-replicates or from two replicate chains per outgroup/model combination, each one run for 10 000 cycles. The burnin was set to 1000 cycles. In the associated table, outgroups are sorted by descending phylogenetic relatedness to molluscs, not evolutionary distance, to illustrate the fact that the latter parameter is the one that really drives the results. To see this, compare PPs for Hymenoptera versus Maxillopoda, two arthropod clades.

8 DIFFICULTY TO ROOT THE TREE OF LIFE USING ANCIENTLY DUPLICATED GENES

We re-examined the case of one anciently duplicated gene pair, the elongation factor: EF-Tu delivers aminoacyl-tRNAs to the A site of the ribosome, while EF-G catalyses the translocation of the peptidyl-tRNA. Even if these two functions are quite different, as shown by the fact that only the GTPase domain can be aligned, this disadvantage is compensated by the preservation of mitochondrial/plastid copies and, more importantly, by the absence of other inter-domain gene transfers. We used the CATGTR + Γ model, which appears to be the less sensitive to LBA [65], albeit the limited number of positions available in the EF alignment (198) prevents it from working at its best, not because of its large number of parameters that might cause over-fitting (see N. Lartillot's paper in this issue [32]), but because of the small amount of information available for defining the peaked amino acid profiles required to efficiently detect the multiple substitutions [25]. In spite of this reduced statistical power, the posterior mean number of categories (79 + 7) significantly rejected a site-homogeneous GTR model (which is a special case of CATGTR with a single category), thus confirming the need to take into account the heterogeneity of the substitution process across sites.

The salient features of the resulting tree (**Figure 14**) are the extremely long internal branches (i) interconnecting the two paralogous copies (3.5 substitutions per site), (ii) lying at the base of Bacteria in each subtree (1.2 and 1.8 for EF-Tu and EF-G, respectively) and (iii) leading to the eukaryotic additional paralogue U5 – 116 kD (1.1). The latter copy codes for a component of the 25S particle that is involved in splicing. While these multiple changes of function explain the length of the U5 – 116 kD branch and of the branch between EF-Tu and EF-G, to our knowledge, no scenario satisfyingly accounts for the very long branch observed at the base of Bacteria in each of the two subtrees. In any case, the length of these internal branches (more than 1 substitution per site) implies that their positioning in the EF tree is mainly determined by the substitution model, and not by a cladistic-like signal. Therefore, it is not really surprising that the two bacterial clades branch at different positions: as sister of Archaea þ Eukaryota for EF-Tu and as sister of Eukaryota for EF-G. In both subtrees, Archaea are highly paraphyletic, with Creanarcheota closer to Eukaryota, yet without any statistical support. Obviously, both stochastic and systematic errors

deeply affect this phylogeny based on duplicated elongation factors. Considering that the EF alignment hosts an average of 83 (+10) substitutions per site, this outcome was somewhat expected and indicates that the root of the ToL cannot yet be pinpointed.

To further study the importance of model violations, we modified the test for heteropecilly of Roure & Philippe [59] to simultaneously look for heterotachy and heteropecilly. This test consists of (i) dividing the dataset into predefined clades, (ii) computing the posterior probability of assigning a given site to a list of predefined CAT categories and (iii) computing the probability of identical profile (PIP) of each site as the sum over all categories of the product of that posterior probability over clades. Here, we did not use a gamma distribution for assigning sites to categories and used a total of 40 categories: the 20 categories defined by Le et al. [88], supplemented by 20 categories with only one non-null amino stationary frequency (one for each amino acid) to favour the assignment of constant sites to one of these 'singleton' categories. Consequently, if a site is heterotachous, i.e. constant in one clade but variable in others, it gets assigned to different categories and obtains a very low PIP value. This test thus estimates the level both of heteropecilly and of extreme heterotachy (i.e. constant versus variable), as it cannot distinguish between medium and fast rates. Interestingly, almost all sites of the EF alignment show a PIP value equal to 0 (161 out of 198 sites) or very small (less than 10210 : 30 sites). This indicates that the EF alignment violates the hypothesis of homogeneity of the substitution process over time assumed by the CATGTR model, a situation that makes very likely the occurrence of LBA artefacts. In this case, it is unfortunately not possible to alleviate the systematic error by removing heterotachous/pecillous sites [59]; too few sites would remain for phylogenetic inference!

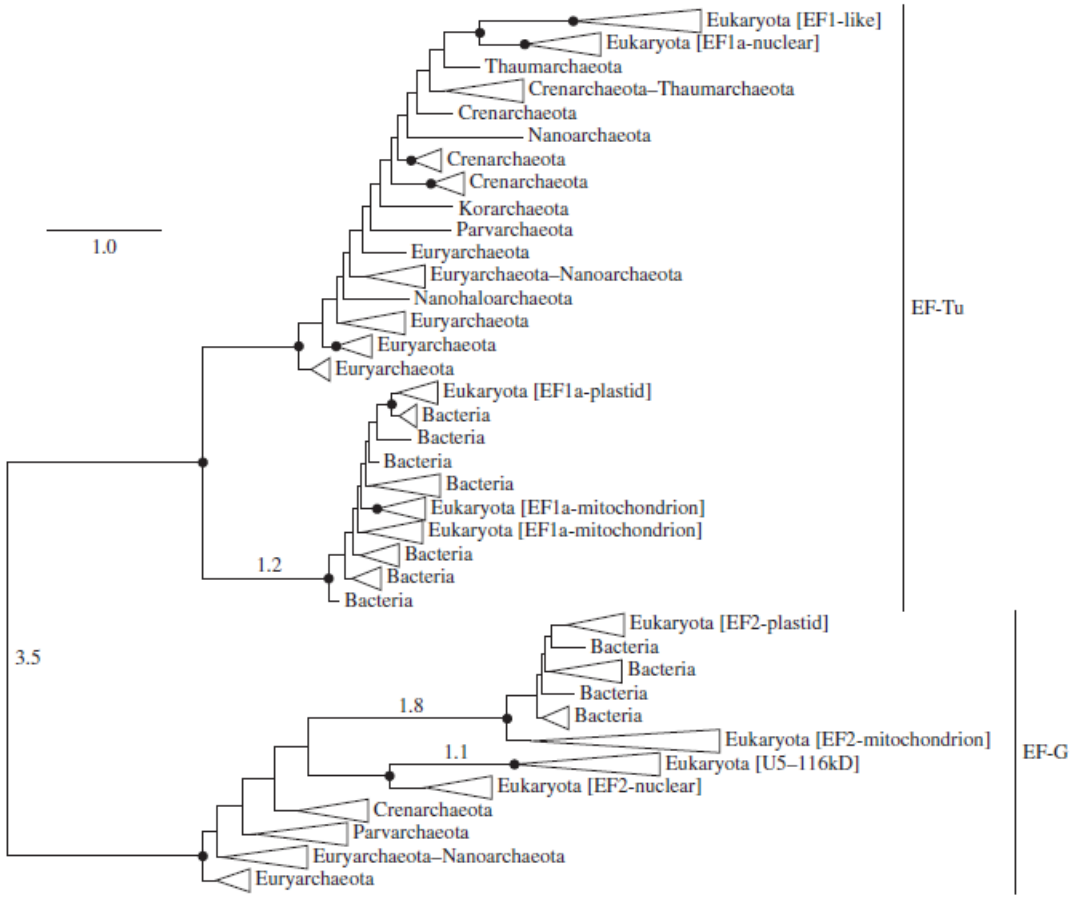


Figure 14.

The amount of model violations in alignments of anciently duplicated genes makes rooting the ToL very difficult. This elongation factor tree was inferred using PHYLOBAYES under the CATGTR + Γ model from an alignment of 211 sequences and 198 unambiguously aligned amino acid positions. Two replicate chains were run for 100 000 cycles and the burnin was set to 50 000 cycles. For clarity, subtrees were collapsed and named after their taxonomic contents. The scale bar corresponds to one substitution per site and the long internal branches discussed in the text are annotated with their length. Bullets indicate branches that are supported by PPs 0.98. In spite of a general lack of resolution, the EF-Tu and EF-G subtrees hint at two different roots for the ToL and suggest that Archaea are indeed paraphyletic, as repeatedly advocated in the literature.

9 CONCLUSION

Our results (figures 2 and 3) demonstrate that the root of the ToL is currently unknown, chiefly because published phylogenies are plagued by tremendous model violations and associated LBA artefacts. Nevertheless, properly addressing this issue is key to make progress in our understanding of archaeogenesis, bacteriogenesis and eukaryogenesis. Indeed, we argue that the current consensus about a bacterial root for the ToL is the product of the prejudice of Aristotle's Great Chain of Beings, in which simple organisms are ancestors of more complex life forms. By contrast, our Apokaryota/ Mitochondriophora stance builds on the many examples where advances in phylogenetic inference have relocated morphologically simple organisms higher in the ToL. However, we acknowledge that a non-bacterial rooting of the ToL would not

necessarily entail that our unorthodox scenario is correct. Indeed, an archaeal rooting or, probably more likely, an intra-domain (within Archaea or within Bacteria) rooting cannot yet be ruled out.

Since stochastic and systematic errors have more impact on rooting the ToL than on resolving any of its parts, rooting strategies should be first validated on shallower issues of similar difficulty, such as the monophyly of Bivalvia studied in **Figure 13**. In our opinion, it is unwise to directly apply new approaches, as clever as they might be, to locate the root of the ToL [89 – 92] without an extensive prior validation on difficult questions with known answers, in particular using very distant outgroups (or without outgroup in the case of nonreversible/non-stationary models). The needed test datasets are straightforward to assemble by subsampling already published complete datasets. Following this reasonable prerequisite, we argue that the supermatrix approach remains the method of choice for rooting the ToL, as it is the most widely used and validated strategy.

To take advantage of the best-fitting models, a relatively large number of characters are necessary, which cannot be obtained using single genes only (e.g. **Figure 14**). However, the concatenation of the few anciently duplicated genes (elongation factors, ATPases, SRP, tRNA-synthetases, etc.) should be possible, as long as the xenologous copies are removed, a task that is within reach thanks to the plethora of complete genomes available today.

While this phylogenetic approach is absolutely required, it will not provide us with a definitive answer, rather the opposite. In the best case, it will locate the root, probably with limited statistical support, which we will need to take into account when developing new evolutionary scenarii. However, beyond being compatible with a correctly rooted ToL, these scenarii will have to fulfil a number of additional constraints, such as:

1. to provide an explanation for the length heterogeneities observed between major branches (e.g. the long branches at the base of Bacteria and Eukaryota);
2. to accommodate palaeontological, genomic, biochemical and cellular knowledge;
3. to explain equally well the emergence of the three major cellular types (bacterial, eukaryotic and archaeal, the latter group likely being paraphyletic), instead of only addressing eukaryogenesis;
4. to provide transitional steps that are evolutionarily simple and plausible, rather than just proposing that simple organisms are ancestors of more complex ones.

In this respect, the study of the recently discovered, yet uncultured, Lokiarchaeota [5], an archaeal group featuring several eukaryotic-'specific' genes (many of them potentially involved in complex membrane remodelling), opens new avenues for completely rethinking the fascinating question of the origin of the three domains of life. Nevertheless, we hope that these will be pursued once freed from the prejudice of Aristotle's Great Chain of Beings.

10 REFERENCES

1. Philippe H, Douady CJ. 2003 Horizontal gene transfer and phylogenetics. Curr. Opin. Microbiol. 6, 498 –505. (doi:10.1016/j.mib.2003.09.008)
2. Eveleigh RJM, Meehan CJ, Archibald JM, Beiko RG. 2013 Being Aquifex aeolicus: untangling a hyperthermophile's checkered past. Genome Biol. Evol. 5, 2478 –2497. (doi:10.1093/gbe/evt195)

3. Zhaxybayeva O et al. 2009 On the chimeric nature, thermophilic origin, and phylogenetic placement of the Thermotogales. Proc. Natl Acad. Sci. USA 106, 5865 –5870. (doi:10.1073/pnas.0901260106)
4. Aravind L, Tatusov RL, Wolf YI, Walker DR, Koonin EV. 1998 Evidence for massive gene exchange between archaeal and bacterial hyperthermophiles. Trends Genet. 14, 442 – 444. (doi:10.1016/S01689525(98)01553-4)
5. Spang A et al. 2015 Complex archaea that bridge the gap between prokaryotes and eukaryotes. Nature 521, 173 –179. (doi:10.1038/nature14447)
6. Woese CR, Kandler O, Wheelis ML. 1990 Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc. Natl Acad. Sci. USA 87, 4576 – 4579. (doi:10.1073/pnas. 87.12.4576)
7. Haag KL, James TY, Pombert J-F, Larsson R, Schaer TMM, Refardt D, Ebert D. 2014 Evolution of a morphological novelty occurred before genome compaction in a lineage of extreme parasites. Proc. Natl Acad. Sci. USA 111, 15 480 –15 485. (doi:10. 1073/pnas.1410442111)
8. Hirt RP, Logsdon JM, Healy B, Dorey MW, Doolittle WF, Embley TM. 1999 Microsporidia are related to Fungi: evidence from the largest subunit of RNA polymerase II and other proteins. Proc. Natl Acad. Sci. USA 96, 580 –585. (doi:10. 1073/pnas.96.2.580)
9. James TY, Pelin A, Bonen L, Ahrendt S, Sain D, Corradi N, Stajich JE. 2013 Shared signatures of parasitism and phylogenomics unite Cryptomycota and Microsporidia. Curr. Biol. 23, 1548 – 1553. (doi:10.1016/j.cub.2013.06.057)
10. Forterre P, Philippe H. 1999 Where is the root of the universal tree of life? BioEssays 21, 871 –879. (doi:10.1002/(SICI)1521-1878(199910)21:10,871:: AID-BIES10.3.0.CO;2-Q)
11. Laurin-Lemay S, Brinkmann H, Philippe H. 2015 Origin of land plants revisited in the light of sequence contamination and missing data. Curr. Biol. 22, R593 –R594. (doi:10.1016/j.cub.2012.06.013)
12. Timme RE, Bachvaroff TR, Delwiche CF. 2012 Broad phylogenomic sampling and the sister lineage of land plants. PLoS ONE 7, e29696. (doi:10.1371/ journal.pone.0029696)
13. Wickett NJ et al. 2014 Phylotranscriptomic analysis of the origin and early diversification of land plants. Proc. Natl Acad. Sci. USA 111, E4859 –E4868. (doi:10.1073/pnas.1323926111)
14. Zhong B, Liu L, Yan Z, Penny D. 2013 Origin of land plants using the multispecies coalescent model. Trends Plant Sci. 18, 492 – 495. (doi:10.1016/j. tplants.2013.04.009)
15. Zhong B, Xi Z, Goremykin VV, Fong R, Mclenachan PA, Novis PM, Davis CC, Penny D. 2014 Streptophyte algae and the origin of land plants revisited using heterogeneous models with three new algal chloroplast genomes. Mol. Biol. Evol. 31, 177 – 183. (doi:10.1093/molbev/mst200)
16. Turmel M, Pombert J, Charlebois P, Otis C, Lemieux C. 2007 The green algal ancestry of land plants as revealed by the chloroplast genome. Int. J. Plant Sci. 168, 679 –689. (doi:10.1086/513470)
17. Bernt M et al. 2013 A comprehensive analysis of bilaterian mitochondrial genomes and phylogeny. Mol. Phylogenet. Evol. 69, 352 –364. (doi:10.1016/j. ympev.2013.05.002)
18. Nosenko T et al. 2013 Deep metazoan phylogeny: when different genes tell different stories. Mol. Phylogenet. Evol. 67, 223 –233. (doi:10.1016/j. ympev.2013.01.010)
19. Telford M. 2013 Field et al. Redux. EvoDevo 4, 5. (doi:10.1186/2041-9139-4-5)

20. Edgecombe G, Giribet G, Dunn C, Hejnol A, Kristensen R, Neves R, Rouse G, Worsaae K, Sørensen M. 2011 Higher-level metazoan relationships: recent progress and remaining questions. Organ. Divers. Evol. 11, 151 – 172. (doi:10.1007/s13127-011-0044-4)
21. Lartillot N, Philippe H. 2008 Improvement of molecular phylogenetic inference and the phylogeny of Bilateria. Phil. Trans. R. Soc. B 363, 1463 –1472. (doi:10.1098/rstb.2007.2236)
22. Philippe H, Brinkmann H, Lavrov DV, Littlewood DTJ, Manuel M, Wörheide G, Baurain D. 2011 Resolving difficult phylogenetic questions: why more sequences are not enough. PLoS Biol. 9, e1000602. (doi:10.1371/journal.pbio.1000602)
23. Felsenstein J. 1978 Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. 27, 401 –410. (doi:10.2307/2412923)
24. Aguinaldo AM, Turbeville JM, Linford LS, Rivera MC, Garey JR, Raff RA, Lake JA. 1997 Evidence for a clade of nematodes, arthropods and other moulting animals. Nature 387, 489 –493. (doi:10.1038/387489a0)
25. Lartillot N, Brinkmann H, Philippe H. 2007 Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. BMC Evol. Biol. 7(Suppl. 1), S4. (doi:10. 1186/1471-2148-7-S1-S4)
26. Philippe H, Germot A. 2000 Phylogeny of eukaryotes based on ribosomal RNA: long-branch attraction and models of sequence evolution. Mol. Biol. Evol. 17, 830 –834. (doi:10.1093/oxford journals.molbev.a026362)
27. Telford MJ. 2004 Animal phylogeny: back to the coelomata? Curr. Biol. 14, 274 –276. (doi:10.1016/j. cub.2004.03.022)
28. Telford MJ, Copley RR. 2005 Animal phylogeny: fatal attraction. Curr. Biol. 15, R296 –R299. (doi:10.1016/ j.cub.2005.04.001)
29. Philippe H, Lartillot N, Brinkmann H. 2005 Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. Mol. Biol. Evol. 22, 1246 –1253. (doi:10.1093/molbev/msi111)
30. Dunn CW et al. 2008 Broad phylogenomic sampling improves resolution of the animal tree of life. Nature 452, 745 – 749. (doi:10.1038/nature06614)
31. Whelan NV, Kocot KM, Moroz LL, Halanych KM. 2015 Error, signal, and the placement of Ctenophora sister to all other animals. Proc. Natl Acad. Sci. USA 112, 5773 – 5778. (doi:10.1073/pnas.1503453112)
32. Lartillot N. 2015 Probabilistic models of eukaryotic evolution: time for integration. Phil. Trans. R. Soc. B 370, 20140338. (doi:10.1098/rstb.2014.0338)
33. Westesson O, Barquist L, Holmes I. 2012 HandAlign: Bayesian multiple sequence alignment, phylogeny and ancestral reconstruction. Bioinformatics 28, 1170 – 1171. (doi:10.1093/bioinformatics/bts058)
34. Redelings BD, Suchard MA. 2005 Joint Bayesian estimation of alignment and phylogeny. Syst. Biol. 54, 401 –418. (doi:10.1080/10635150590947041)
35. Szöllősi GA, Davin AA, Tannier E, Daubin V, Boussau B. 2005 Genome-scale phylogenetic analysis finds extensive gene transfer among fungi. Phil. Trans. R. Soc. B 370, 20140335. (doi:10.1098/ rstb.2014.0335)
36. Boussau B, Szöllősi GJ, Duret L, Gouy M, Tannier E, Daubin V. 2013 Genome-scale coestimation of species and gene trees. Genome Res. 23, 323 –330. (doi:10.1101/gr.141978.112)

37. Struck TH. 2013 The impact of paralogy on phylogenomic studies – a case study on annelid relationships. PLoS ONE 8, e62892. (doi:10.1371/ journal.pone.0062892)
38. Jukes TH, Cantor CR. 1969 Evolution of protein molecules. Mamm. Protein Metab. 3, 21 –132. (doi:10.1016/B978-1-4832-3211-9.50009-7)
39. Dayhoff M, Schwartz R. 1978 A model of evolutionary change in proteins. In Atlas of protein sequence and structure (ed. M Dayhoff), pp. 345 – 352. Washington, DC: National Biomedical Research Foundation.
40. Le SQ, Gascuel O. 2008 An improved general amino acid replacement matrix. Mol. Biol. Evol. 25, 1307 –1320. (doi:10.1093/molbev/msn067)
41. Uzzell T, Corbin KW. 1971 Fitting discrete probability distributions to evolutionary events. Science 172, 1089 –1096. (doi:10.2307/1731831)
42. Halpern AL, Bruno WJ. 1998 Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. Mol. Biol. Evol. 15, 910 –917. (doi:10.1093/oxfordjournals.molbev.a025995)
43. Lartillot N, Philippe H. 2004 A Bayesian mixture model for across-site heterogeneities in the aminoacid replacement process. Mol. Biol. Evol. 21, 1095 –1109. (doi:10.1093/molbev/msh112)
44. Goldman N, Yang Z. 1994 A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol. 11, 725 –736.
45. Muse SV, Gaut BS. 1994 A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol. Biol. Evol. 11, 715 –724.
46. Rodrigue N, Philippe H, Lartillot N. 2010 Mutationselection models of coding sequence evolution with site-heterogeneous amino acid fitness profiles. Proc. Natl Acad. Sci. USA 107, 4629 –4634. (doi:10.1073/ pnas.0910915107)
47. Lockhart PJ, Steel MA, Hendy MD, Penny D. 1994 Recovering evolutionary trees under a more realistic model of sequence evolution. Mol. Biol. Evol. 11, 605 –612.
48. Woese CR, Kandler O, Wheelis ML. 1991 A natural classification. Nature 351, 528 –529. (doi:10.1038/ 351528c0)
49. Embley TM, Thomas RH, Williams RAD. 1993 Reduced thermophilic bias in the 16*S* rDNA sequence from Thermus ruber provides further support for a relationship between Thermus and Deinococcus. Syst. Appl. Microbiol. 16, 25 –29. (doi:10.1016/S0723-2020(11)80247-X)
50. Blanquart S, Lartillot N. 2008 A site- and time- heterogeneous model of amino acid replacement. Mol. Biol. Evol. 25, 842 – 858. (doi:10.1093/molbev/ msn018)
51. Foster PG. 2004 Modeling compositional heterogeneity. Syst. Biol. 53, 485 –495. (doi:10. 1080/10635150490445779)
52. Cox CJ, Foster PG, Hirt RP, Harris SR, Embley TM. 2008 The archaebacterial origin of eukaryotes. Proc. Natl Acad. Sci. USA 105, 20 356 –20 361. (doi:10. 1073/pnas.0810647105)
53. Fitch W, Markowitz E. 1970 An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. Biochem. Genet. 4, 579 – 593. (doi:10. 1007/BF00486096)
54. Galtier N, Jean-Marie A. 2004 Markov-modulated Markov chains and the covarion process of molecular evolution. J. Comp. Biol. 11, 727 –733. (doi:10.1089/cmb.2004.11.727)

55. Zhou Y, Brinkmann H, Rodrigue N, Lartillot N, Philippe H. 2010 A Dirichlet process covarion mixture model and its assessments using posterior predictive discrepancy tests. Mol. Biol. Evol. 27, 371 –384. (doi:10.1093/molbev/msp248)
56. Kolaczkowski B, Thornton JW. 2008 A mixed branch length model of heterotachy improves phylogenetic accuracy. Mol. Biol. Evol. 25, 1054 –1066. (doi:10. 1093/molbev/msn042)
57. Schwartz RS, Mueller RL. 2010 Limited effects of among-lineage rate variation on the phylogenetic performance of molecular markers. Mol. Phylogenet. Evol. 54, 849 –856. (doi:10.1016/j.ympev.2009. 12.025)
58. Lartillot N, Rodrigue N, Stubbs D, Richer J. 2013 PhyloBayes MPI: phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. Syst. Biol. 62, 611 –615. (doi:10.1093/ sysbio/syt022)
59. Roure B, Philippe H. 2011 Site-specific time heterogeneity of the substitution process and its impact on phylogenetic inference. BMC Evol. Biol. 11, 17. (doi:10.1186/1471-2148-11-17)
60. Rodrigue N, Kleinman CL, Philippe H, Lartillot N. 2009 Computational methods for evaluating phylogenetic models of coding sequence evolution with dependence between codons. Mol. Biol. Evol. 26, 1663 –1676. (doi:10.1093/molbev/msp078)
61. Pick KS et al. 2010 Improved phylogenomic taxon sampling noticeably affects nonbilaterian relationships. Mol. Biol. Evol. 27, 1983 –1987. (doi:10.1093/molbev/msq089)
62. Philippe H et al. 2009 Phylogenomics revives traditional views on deep animal relationships. Curr. Biol. 19, 706 –712. (doi:10.1016/j.cub.2009.02.052)
63. Williams BAP, Hirt RP, Lucocq JM, Embley TM. 2002 A mitochondrial remnant in the microsporidian Trachipleistophora hominis. Nature 418, 865 –869. (doi:10.1038/nature00949)
64. Jime´nez-Guri E, Philippe H, Okamura B, Holland PWH. 2007 Buddenbrockia is a cnidarian worm. Science 317, 116 –118. (doi:10.1126/science. 1142024)
65. Philippe H, Brinkmann H, Copley RR, Moroz LL, Nakano H, Poustka AJ, Wallberg A, Peterson KJ, Telford MJ. 2011 Acoelomorph flatworms are deuterostomes related to Xenoturbella. Nature 470, 255 –258. (doi:10.1038/nature09676)
66. Delsuc F, Brinkmann H, Chourrout D, Philippe H. 2006 Tunicates and not cephalochordates are the closest living relatives of vertebrates. Nature 439, 965 –968. (doi:10.1038/nature04336)
67. Zhaxybayeva O, Lapierre P, Gogarten JP. 2005 Ancient gene duplications and the root(s) of the tree of life. Protoplasma 227, 53 – 64. (doi:10.1007/ s00709-005-0135-1)
68. Williams TA, Foster PG, Nye TMW, Cox CJ, Embley TM. 2012 A congruent phylogenomic signal places eukaryotes within the Archaea. Proc. R. Soc. B 279, 4870 –4879. (doi:10.1098/ rspb.2012.1795)
69. Gould SJ. 1997 Full house: the spread of excellence from Plato to Darwin. New York, NY: Harmony Books.
70. Dagan T, Roettger M, Bryant D, Martin W. 2010 Genome networks root the tree of life between prokaryotic domains. Genome Biol. Evol. 2, 379 –392. (doi:10.1093/gbe/evq025)
71. Hage`ge C. 2012 Contre la pensée unique. Paris, France: Editions Odile Jacob.
72. Koerner EFK. 2000 Towards a 'full pedigree' of the 'Sapir-Whorf hypothesis': From Locke to Lucy. In Explorations in linguistic relativity (eds M Pütz, M Verspoor), p. 369. Amsterdam, The Netherlands: John Benjamins Publishing Company.

73. Pace NR. 2009 Problems with 'Procaryote'. J. Bacteriol. 191, 2008 –2010. (doi:10.1128/JB.01224-08)
74. Pace NR. 2006 Time for a change. Nature 441, 289. (doi:10.1038/441289a)
75. Sapp J. 2006 Two faces of the prokaryote concept. Int. Microbiol. 9, 163 –172.
76. Sapp J. 2005 The prokaryote– eukaryote dichotomy: meanings and mythology. Microbiol. Mol. Biol. Rev. 69, 292 –305. (doi:10.1128/MMBR.69.2.292305.2005)
77. Pace NR. 2008 The molecular tree of life changes how we see, teach microbial diversity. Microbe Mag. 3, 15 –20.
78. Lwoff A, Bordet JJBV. 1944 L'évolution physiologique: étude des pertes de fonctions chez les microorganismes. Paris, France: Hermann.
79. Zhaxybayeva O, Gogarten JP. 2004 Cladogenesis, coalescence and the evolution of the three domains of life. Trends Genet. 20, 182 –187. (doi:10.1016/j. tig.2004.02.004)
80. Lasek-Nesselquist E, Gogarten JP. 2013 The effects of model choice and mitigating bias on the ribosomal tree of life. Mol. Phylogenet. Evol. 69, 17 – 38. (doi:10.1016/j.ympev.2013.05.006)
81. Gogarten JP et al. 1989 Evolution of the vacuolar Hþ-ATPase: implications for the origin of eukaryotes. Proc. Natl Acad. Sci. USA 86, 6661 – 6665. (doi:10.1073/pnas.86.17.6661)
82. Iwabe N, Kuma K, Hasegawa M, Osawa S, Miyata T. 1989 Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. Proc. Natl Acad. Sci. USA 86, 9355–9359. (doi:10.1073/pnas.86.23.9355)
83. Schwartz RM, Dayhoff MO. 1978 Origins of prokaryotes, eukaryotes, mitochondria, and chloroplasts. Science 199, 395 –403. (doi:10.1126/ science.202030)
84. Brinkmann H, Philippe H. 1999 Archaea sister group of Bacteria? Indications from tree reconstruction artifacts in ancient phylogenies. Mol. Biol. Evol. 16, 817 –825. (doi:10.1093/oxfordjournals.molbev.a026166)
85. Stamatakis A. 2014 RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30, 1312 –1313. (doi:10.1093/bioinformatics/btu033)
86. Lopez P, Casane D, Philippe H. 2002 Heterotachy, an important process of protein evolution. Mol. Biol. Evol. 19, 1–7. (doi:10.1093/oxfordjournals.molbev.a003973)
87. Philippe H, Forterre P. 1999 The rooting of the universal tree of life is not reliable. J. Mol. Evol. 49, 509 – 523. (doi:10.1007/PL00006573)
88. Si Quang L, Gascuel O, Lartillot N. 2008 Empirical profile mixture models for phylogenetic reconstruction. Bioinformatics 24, 2317 – 2323. (doi:10.1093/bioinformatics/ btn445)
89. Fournier GP, Gogarten JP. 2010 Rooting the ribosomal tree of life. Mol. Biol. Evol. 27, 1792 – 1801. (doi:10.1093/molbev/msq057)
90. Harish A, Tunlid A, Kurland CG. 2013 Rooted phylogeny of the three superkingdoms. Biochimie 95, 1593 –1604. (doi:10.1016/j.biochi.2013.04.016)
91. Lake JA, Servin JA, Herbold CW, Skophammer RG. 2008 Evidence for a new root of the tree of life. Syst. Biol. 57, 835 –843. (doi:10.1080/1063515 0802555933)
92. Skophammer RG, Servin JA, Herbold CW, Lake JA. 2007 Evidence for a Gram-positive, eubacterial root of the tree of life. Mol. Biol. Evol. 24, 1761 –1768. (doi:10.1093/molbev/msm096)

MISE A JOUR

11 DES RELATIONS DE PARENTE COMPLEXES ENTRE ARCHEES ET EUCARYOTES

11.1 DECOUVERTE DU GROUPE ASGARD

En 2015, l'équipe de Thijs Ettema de l'université d'Uppsala en Suède a reconstitué un génome du clade δ du groupe benthique profond (*Deep Sea Archaeal Group*, DSAG) qui, selon eux, comble le fossé entre les procaryotes et les eucaryotes (Spang et al., 2015). Les études de métagénomique montrent qu'ils émergent dans le super-phylum TACK et seraient étroitement apparentés aux eucaryotes.

Baptisés « Lokiarchaeota », ils furent récoltés à 3 283 m de profondeur dans des sédiments marins de l'Océan Arctique entre le Groenland et la Norvège, proches de la zone d'activité hydrothermale de la dorsale de Gakkel, près du lieu-dit Loki's Castle (d'où leur nom) découvert en Juillet 2008. Ils sont représentés par un génome complet (5,1 Mpb pour 5 381 gènes encodant des protéines) et deux génomes presque complets nommés Loki2 et Loki3.

En 2017, les mêmes auteurs publient un second article décrivant les MAG de plusieurs nouvelles lignées d'archées apparentées aux Lokiarchées et auxquelles ils attribuent également des noms scandinaves : les Thorarchées, les Odinarchées et les Heimdallarchées (Eme et al., 2017; Spang et al., 2017). Ces nouvelles lignées ont été retrouvées dans de très nombreux environnements : sources chaudes, sédiments profonds pauvres en oxygène, sédiments côtiers, plancton océanique (Macleod et al., 2019).

Les noms de ces nouveaux génomes dérivant de dieux issus de la mythologie scandinave, ils ont été regroupés au sein d'un super-phylum nommé *Asgard*, du nom de la demeure où ces dieux sont supposés vivre. Depuis lors, de nouveaux échantillons issus de diverses régions du globe (Loki's Castle, Parc National du Yellowstone, Aarhus Bay au Danemark, un aquifère près de la rivière du Colorado, Radiata Pool en Nouvelle-Zélande, sources hydrothermales proches des îles Taketomi au Japon et estuaire du White Oak River) ont mis en évidence diverses souches de ce super-phylum, telles les Helarchées ou les Gerdarchées.

Il a été suggéré que l'étroite relation entre les Asgards et les Eucaryotes soit le résultat de possibles contaminations lors de l'assemblage de métagénomes, combiné avec des artefacts de reconstruction phylogénétique dus au choix des gènes utilisés et l'inclusion de groupes d'Archaea évoluant rapidement (Da Cunha et al., 2017). Mais depuis la découverte de nouveaux métagénomes d'Asgard, la faible probabilité que la contamination et/ou la recombinaison homologue avec des séquences eucaryotes se soit produite de la même manière dans l'ensemble des archées Asgards, dont les génomes ont été assemblés à partir de diverses sources et selon diverses méthodologies, discrédite le fait que l'étroit apparentement entre eucaryotes et Asgards soit le résultat d'une contamination avec des séquences eucaryotes.

De plus, en 2019, un organisme baptisé Candidatus *Prometheoarchaeum syntrophicum* (souche MK-D1) et apparenté aux Lokiarchaeota fut le seul de ce groupe à avoir pu être cultivé et totalement séquencé (Imachi et al., 2020). Quatre ans d'efforts ont encore été nécessaires pour caractériser cette archée Loki. En effet, celui-ci se divise très lentement : il lui faut soixante jours pour démarrer sa croissance et son temps de doublement est d'environ 14-25 jours.

Les chercheurs n'ont pas réussi à l'isoler en culture pure, mais en co-culture avec une bactérie sulfato-réductrice du genre Desulfovibrio et une archée productrice du méthane du genre Methanogenium. L'archée Loki produit de l'hydrogène qui est utilisé par la bactérie pour produire du sulfure d'hydrogène et par Methanogenium pour produire du méthane. Bien qu'il ait été possible de finalement éliminer la bactérie du système de culture, il semble en revanche que les Lokiarchaeota ne peuvent vivre seules et dépendent de leur étroite association avec l'archée Methanogenium. En retour, Methanogenium fournit aux Loki de nombreux composants, dont des acides aminés et des vitamines, que MK-D1 est incapable de produire lui-même.

D'après les auteurs de cette étude, toutes les archées Asgards pourraient également être dépendantes d'une association symbiotique pour leur développement, d'où la difficulté de les cultiver et l'absence dans leurs génomes d'un certain nombre de voies métaboliques importantes, en particulier les voies de biosynthèse des acides aminés.

Une deuxième espèce d'Asgard récoltée dans la boue d'un estuaire en Slovénie a également été cultivée après 7 ans d'effort (Rodrigues-Oliveira et al., 2023). Baptisée *Lokiarchaeum ossiferum*, celui-ci arbore plusieurs dizaines de fins tentacules comportant des épaississements et des excroissances ainsi qu'un cytosquelette s'étendant jusque dans les tentacules. Il s'agit d'un organisme anaérobie, trouvé dans des environnements marins profonds pauvres en oxygène, tels que les sédiments hydrothermaux. Il semble capable d'effectuer des processus de chimiosynthèse, obtenant de l'énergie à partir de la réduction des composés comme le soufre ou le méthane, typique des archées de ces environnements extrêmes. Son génome a pu être entièrement séquencé. Il possède des gènes codant des protéines semblables à celles que l'on trouve chez les eucaryotes. Ces protéines incluent notamment des composants du cytosquelette, des protéines impliquées dans le trafic membranaire et d'autres processus complexes que l'on pensait auparavant propres aux cellules eucaryotes. Cela suggère que cet ancêtre commun des eucaryotes aurait pu posséder une complexité cellulaire plus grande qu'on ne l'imaginait. Le génome de *Lokiarchaeum ossiferum* révèle une capacité à effectuer des processus métaboliques typiques des archées, mais avec une augmentation notable du répertoire de gènes associés à des fonctions eucaryotiques. Par exemple, on y trouve des gènes pour des protéines à domaine E2 ubiquitine, des protéines impliquées dans le remodelage de la membrane, et des GTPases associées au cytosquelette. Ces caractéristiques suggèrent une capacité accrue pour interagir avec des membranes intracellulaires et former des compartiments, ce qui serait une étape clé dans l'évolution vers des cellules plus complexes.

Lokiarchaeota est un organisme de petite taille (~550 nm), anaérobie, qui dégrade des acides aminés et des peptides par syntrophie avec l'archée Methanogenium et/ou avec la bactérie Halodesulfovibrio. La culture de cette souche confirme la présence de 80 protéines signatures des eucaryotes (ESPs), également trouvées chez les autres Asgards. De plus, elle produit des « bras »

ou « tentacules » qui lui permettent de se lier à d'autres microorganismes et donc éventuellement d'en « avaler » un autre.

À partir des caractéristiques identifiées chez la nouvelle archée MK-D1 et en partant de l'hypothèse que celle-ci pourrait être un hôte possible pour la théorie endosymbiotique, les auteurs de la publication ont établi un nouveau modèle nommé E3 pour Enchevêtrement – Engloutissement – Endogénéisation (= internalisation).

Ils ont alors proposé un récit des événements qui auraient amené une archée apparentée à MK-D1 à « avaler » un autre organisme et pouvant expliquer l'origine des Eucaryotes. Selon les auteurs, plusieurs étapes auraient été nécessaires (**Figure 15**) :

- La première étape est une transition entre un mode de vie sans oxygène vers un mode de vie tolérant la présence d'oxygène. Une relation de dépendance se serait installée entre les deux organismes : l'archée, futur hôte et la bactérie, futur endosymbiote (**Figure 15 (a)**).
- Ensuite, les structures externes de l'hôte (« bras ou tentacules ») auraient interagi avec la bactérie, permettant d'améliorer l'interaction physique et « d'engloutir » celle-ci. On pourrait dire que l'hôte a « avalé » la bactérie mais sans la « digérer ». (**Figure 15 (b)**).
- Suite à cette ingestion, l'hôte aurait partagé avec la bactérie symbiotique des molécules comme source d'énergie pour cette dernière. En retour, la bactérie aurait consommé l'oxygène environnant et fourni à son tour de l'énergie à l'hôte sous forme de la molécule ATP (**Figure 15 (c)**).
- La symbiose entre les deux organismes aurait évolué au cours du temps. Ainsi, à terme, l'hôte aurait délégué la production d'énergie sous forme d'ATP à la bactérie. Un canal, appelé AAC, qui permet le passage de l'ATP entre les deux cellules se serait mis en place (**Figure 15 (d)**). Ce canal est aujourd'hui essentiel pour les mitochondries actuelles des cellules eucaryotes.
- Finalement, au cours de l'évolution, des modifications dans la structure interne seraient survenues, donnant naissance aux mitochondries actuelles. Il est aussi possible que d'autres endosymbioses se soient produites en série pour parvenir aux Eucaryotes tels que nous les connaissons aujourd'hui.

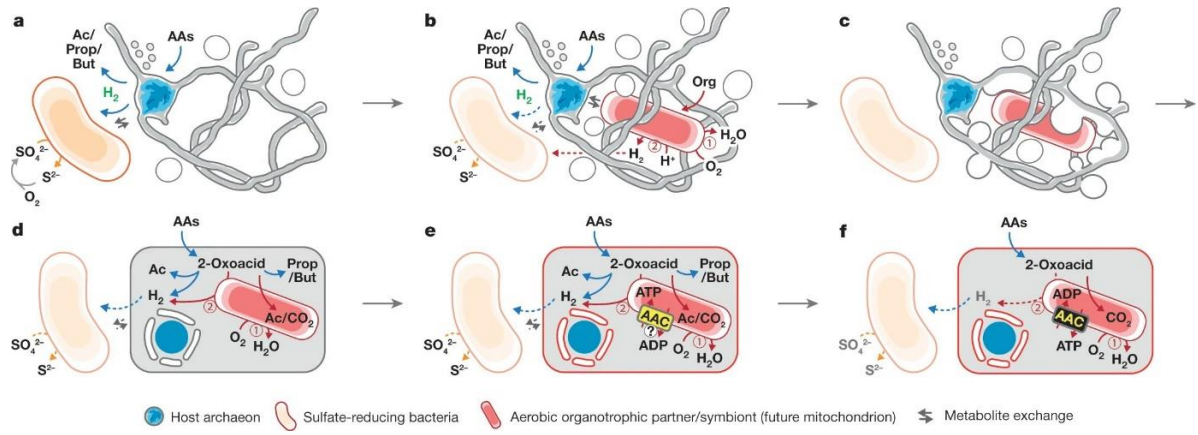


Figure 15. Modèle hypothétique expliquant l'apparition des Eucaryotes (Imachi et al., 2020).
a) Interaction entre l'archée hôte et la bactérie symbiotique. La bactérie symbiotique utilise le dioxygène (O_2) du milieu en échange de nutriments produits par l'hôte. **b**) Les « bras » de l'hôte interagissent avec la bactérie symbiotique, ce qui améliore l'interaction physique entre les deux organismes et mène à l'engloutissement de la bactérie. **c**) L'interaction se poursuit après l'engloutissement par le développement d'un canal (AAC) permettant le transfert de molécules énergétiques (ATP). **d**) Internalisation et délégation par l'hôte de la production de molécules énergétiques au symbiote.

Le nombre de représentants du phylum d'Asgard ne cesse de croître depuis 2015, avec à ce jour 18 lignées (Da Cunha, Gaïa, et al., 2022) (**Figure 16**).

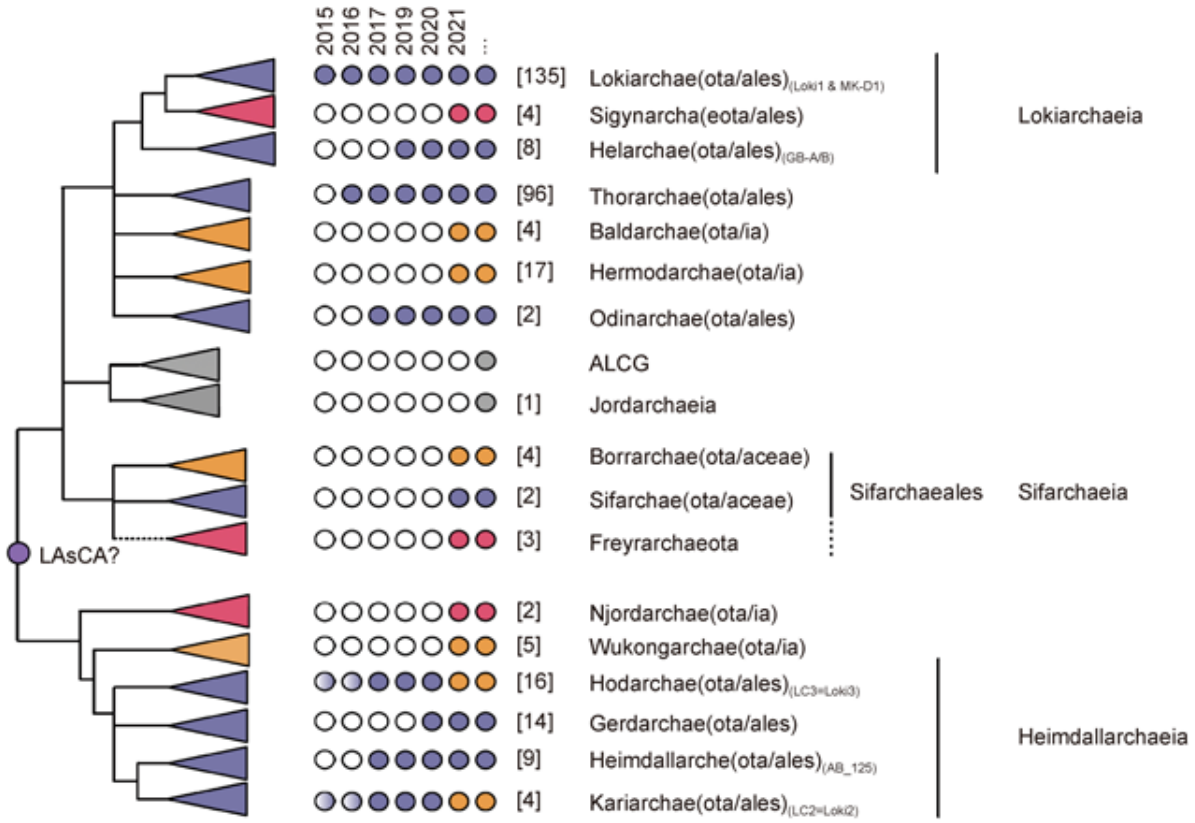


Figure 16. Représentation schématique de la diversité connue des archées d'Asgard, y compris les lignées récemment publiées. (Da Cunha, Gaïa, et al., 2022).
Les couleurs orange et en rouge correspondent aux lignées nouvellement découvertes introduites dans respectivement les deux publications de (Liu et al., 2021a) et (Xie et al., 2022) et en gris les nouvelles lignées non présentes dans ces publications correspondant aux Sifarchaea (F et al., 2021), aux Jordarchaeia (Sun et al., 2021) et à Asgard Lake Cootharaba (Sun et al., 2021). Le schéma a été conçu en combinant les arbres de toutes ces publications. Le suffixe entre parenthèses indique que, selon les auteurs, ces lignées sont considérées comme des Phylum (*ota*), des Familles (*aceae*), ou des Ordres (*ales*). La position du LAsCA (*Last Asgard archaeal Common Ancestor*) et localisée en se basant sur la racine observée dans la plupart des arbres phylogénétiques et utilisée par (Liu et al., 2021a) pour discuter de la distribution des ESPs. Pour chaque lignée Asgard, l'année de publication du génome de son premier représentant et indiquée, et les changements de taxonomie sont indiqués par des changements de couleur des points. Les points violet clair indiquent que les Asgards des lignées Karia et Hodar ont été décrits avant 2017 comme Loki 2 et Loki 3, respectivement (Spang et al., 2015). En outre, le nombre de génomes disponibles pour chaque lignée est indiqué entre crochets.

11.2 CARACTERISTIQUES DU GROUPE ASGARD

Ces nouveaux métagénomes ont beaucoup remis en question le modèle du vivant à 3 domaines, au profit d'un modèle du vivant à 2 domaines (hypothèse Eocyte). En effet, les analyses phylogénétiques basées sur 36 protéines universelles tendent à placer les eucaryotes au sein des archées Asgards (Spang et al., 2015). De plus, ces nouveaux génomes partagent de nombreuses caractéristiques que l'on pensait propres aux eucaryotes, laissant imaginer que ces derniers pourraient en être issus. Ainsi, un certain nombre de gènes codant pour des protéines de type eucaryote qui n'avaient jamais été identifiées jusque-là chez les archées ont été identifiés dans les génomes d'Asgards. On retrouve ainsi chez les Asgards des ESPs, dont des protéines impliquées dans les mécanismes de trafic membranaire, des protéines structurelles eucaryotes (cytosquelette), des protéines impliquées dans le système de modification par ubiquitination et des protéines ribosomiques eucaryotes. En particulier, les Asgards possèdent des protéines actines apparentées à celles des eucaryotes. De plus, chez les Odinarchées, deux gènes sont homologues au FtsZ un gène (OdinTubulin) possède une plus grande homologie avec la tubuline eucaryote qu'avec tout autres archées (Akıl et al., 2022). Ces deux protéines sont des composants majeurs du cytosquelette eucaryote. Le cytosquelette permettant aux cellules de se déplacer et de se déformer pour englober d'autres cellules (processus de phagocytose), ce qui aurait pu être un atout déterminant dans l'émergence de la cellule eucaryote. Plusieurs auteurs ont suggéré qu'un cytosquelette de ce type aurait pu apparaître chez les archées Asgards, permettant à l'une d'entre elles d'englober par phagocytose la bactérie à l'origine des mitochondries, initiant ainsi le processus qui aurait conduit à l'apparition des Eucaryotes. Si les archées Asgards ne sont pas nos ancêtres directs, la présence d'un grand nombre de protéines de type eucaryote chez ces dernières suggère qu'elles ont partagé pendant longtemps le même environnement que les proto-eucaryotes. Si les archées Asgards vivent bien toutes de préférence

en association avec d'autres organismes, il est même possible qu'elles aient un jour vécu en symbiose avec nos ancêtres.

La présence de systèmes semblables à ceux des eucaryotes dans le phylum Asgard suggère que les eucaryotes ont hérité de simples variations de la machinerie cellulaire d'un ancêtre archéen Cependant, plusieurs de ces systèmes eucaryotes codés de manière putative chez les Asgards sont incomplets ou doivent encore être caractérisés sur le plan fonctionnel. Comme ils possèdent des fonctions encore inconnues chez les archées, il est difficile d'en déduire le mécanisme par lequel les systèmes eucaryotes se sont développés au sein des archées.

Il a été suggéré que l'existence de transferts horizontaux de gènes entre les Asgards ancestraux et les proto-eucaryotes aurait pu aboutir aux topologies très diversifiées des arbres phylogénétiques de certains ESPs Asgards et de protéines marqueurs universels. Cette hypothèse est pertinente quel que soit le scénario envisagé concernant l'eucaryogenèse. Elle implique que les Asgards étaient déjà diversifiés avant le dernier ancêtre commun eucaryote et qu'ils ont partagé les mêmes biotopes avec les proto-eucaryotes. Les ESPs retrouvées chez les Asgards et les protéines universelles pourraient simplement avoir été recrutées par un proto-eucaryote, ce qui expliquerait la distribution éparse des ESPs et le placement atypique de certaines lignées Asgards dans les arbres universels basées sur certains gènes. Il est possible que certains Asgards pourraient encore vivre en symbiose aujourd'hui avec des eucaryotes modernes.

Ces génomes possèdent de nombreux gènes encodant des protéines à signature eucaryote, ce qui semble les relier fortement aux eucaryotes (**Figure 17** et **Figure 18**). Selon le modèle endosymbiotique, l'hôte ancestral possédait ainsi déjà certains composants clés qui ont régi l'émergence de la complexité cellulaire eucaryote après l'endosymbiose. Parmi ces gènes, nous pouvons entres autres citer (Eme et al., 2017; Spang et al., 2017) :

- le gène encodant la protéine eL22 de la grande sous-unité ribosomique. En revanche, le gène encodant la protéine eL41 est absent.
- des boîtes codant des domaines homologues à la gelsoline, protéine qui démantèle et coiffe les filaments d'actine en présence d'ions calcium, permettant la formation de vésicules d'exocytose.
- plusieurs gènes encodant le complexe ESCRT (*Endosomal Sorting Complexes Required for Transport*) et plusieurs protéines homologues aux composants de la voie eucaryotique de l'endosome multivésiculaire. A noter que le complexe ESCRT se retrouve déjà au sein des archées sous diverses formes (Frohn Béla P. AND Härtel 2022).
- Diverses sous-familles de la super-famille Ras de petites GTPases.
- Des protéines de la super-famille BAR/IMD impliquées dans le trafic cellulaire et le remodelage membranaire.
- MK-D1 exprime simultanément trois systèmes capables de participer à la division cellulaire : FtsZ, actine et le complexe ESCRT-II/III.

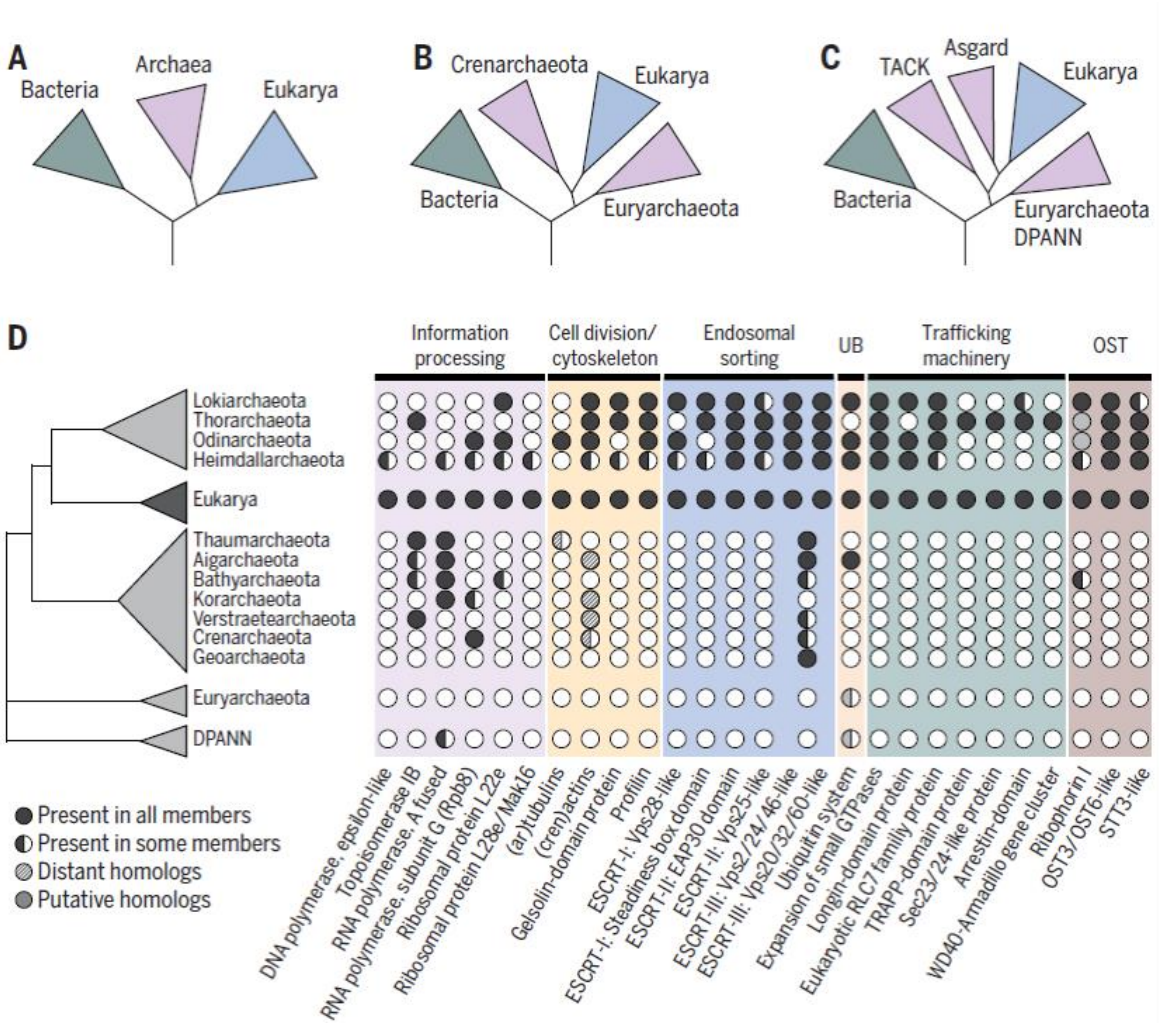


Figure 17. Les archées et l'évolution de l'arbre du vivant (Spang et al., 2017).

(A à C) Représentation schématique de la relation entre les archées, les bactéries et les eucaryotes selon la topologie à trois domaines (A) et à deux domaines (B), et mis à jour avec les Asgards (C). L'arbre à trois domaines suggère que les archées, les bactéries et les eucaryotes représentent des domaines primaires. En revanche, une topologie à deux domaines [(B) et (C)] est plus compatible avec l'idée que les eucaryotes représentent un domaine de vie secondaire issu de la fusion de deux procaryotes : un hôte archéen et un symbiote alphaprotéobactérien. (D) Un arbre schématique des archées et de leur relation avec les eucaryotes, soulignant le placement des eucaryotes avec Asgard et montrant la distribution d'ESPs dans les différentes lignées.

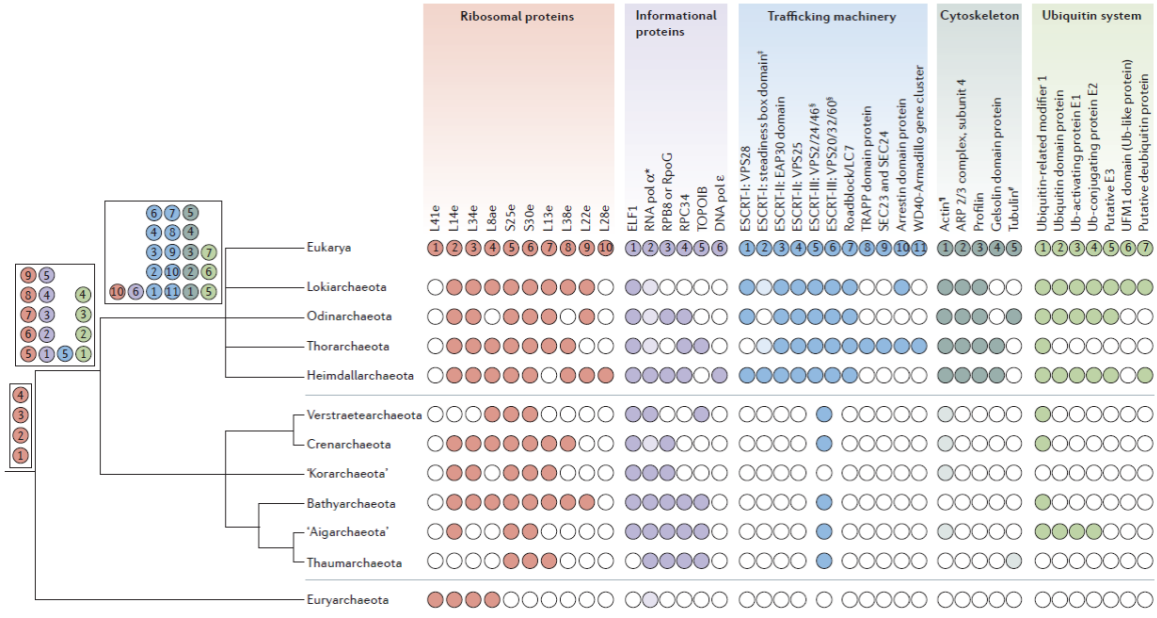


Figure 18. Origines des protéines de signatures eucaryotes (ESPs) présentes chez FECA (Eme et al., 2017).

Cette figure indique la présence d'homologues des protéines de signature eucaryotes (ESPs) dans diverses lignées d'archées (cercles remplis), ainsi que leur date d'émergence présumée le long de l'arbre schématique des archées. L'origine de chaque ESP est indiquée sur un arbre du vivant schématique (côté gauche). On notera que cette figure sous-entend une phylogénie des archées connue.

D'habitude, ces gènes sont présents de façon irrégulière au sein du groupe TACK. En revanche, ils semblent tous présents au sein du groupe Asgard, laissant supposer l'idée que les eucaryotes puissent être issus d'une fusion entre une de ces Archaea et une bactérie (Da Cunha, Gaïa, et al., 2022) laissant par la suite la possibilité de réaliser des endosymbioses via la présence d'un cytosquelette dynamique (ex. mitochondries, plastes) et de construire des systèmes membranaires (ex. membrane nucléaire, organites). C'est ainsi, grâce à ces gènes du remodelage membranaire et du trafic cellulaire, que cette cellule proto-eucaryote aurait pu acquérir la complexité interne qu'on lui connaît actuellement.

Toutefois, ces relations de parenté sont à ce jour encore sujet à débat. En effet, l'utilisation des eucaryotes dans les jeux de données actuels ne peut être fait que de façon très limitée, à cause de la difficulté de mettre en évidence une homologie détectable entre les génomes des différents domaines (Zhu et al., 2019).

L'analyse des arbres individuels concernant chacune des 36 protéines universelles a montré que des contradictions se cachent derrière l'analyse globale de ces mêmes 36 protéines. En effet, on note que la position des archées Asgards varie fortement dans les arbres individuels selon la protéine étudiée. Dans de nombreux cas, elles ne sont pas groupées dans les arbres individuels : certaines sont proches des eucaryotes tandis que d'autres se retrouvent au milieu des archées. Leur proximité avec les eucaryotes s'observe surtout avec les petites protéines qui

favorisent les modèles à 2 domaines (hypothèse Eocyte). Si on s'intéresse aux plus grosses protéines, qui portent plus de signaux pour placer correctement les organismes car elles permettent de comparer les positions d'un plus grand nombre d'acides aminés, les archées Asgards se retrouvent le plus souvent placées au milieu des autres archées, favorisant un scénario à 3 domaines. Dans les concaténations, les très nombreuses protéines universelles de petite taille masqueraient le signal présent dans les protéines de grande taille (Da Cunha, Gaïa, et al., 2022).

Actuellement, les Archaea incluent quatre supergroupes bien reconnus (Adam et al., 2017; Castelle et al., 2015; Castelle & Banfield, 2018; Spang et al., 2017; T. A. Williams et al., 2017) : les Euryarchaeota (qui comprennent une quinzaine de sous-groupes bien individualisés), le groupe TACK (Thaumarchaeota, Aigarchaeota, Crenarchaeota, Korarchaeota), le groupe DPANN (Diapherotrites, Parvarchaeota, Aenigmarchaeota, Nanoarchaeota, Nanohaloarchaeota) et les Asgards.

CHAPITRE 2

LA SUPER-MATRICE ET PHYLOGENIE DES ARCHAEA

1 INTRODUCTION

Les dernières études tendent à renforcer l'hypothèse Eocyte en considérant les eucaryotes comme un sous-groupe d'archées, probablement issues du groupe Asgard (Eme et al., 2017; Spang et al., 2017; T. A. Williams et al., 2017). Toutefois les jeux de données utilisés et les méthodes employées sont régulièrement remis en question, pouvant résulter d'artefacts de reconstruction phylogénétique (Brinkmann & Philippe, 2007; Da Cunha, Gaïa, et al., 2022; Forterre, 2011; Philippe & Forterre, 1999) (**Box 7**). Notre but est donc de construire un jeu de données le plus fiable possible afin d'éviter tout biais et d'effectuer des analyses phylogénomiques originales afin de le tester et vérifier la solidité des résultats obtenus. Pour ce faire, nous allons rassembler un jeu de données le plus représentatif de la diversité des archées, en vérifiant la complétude, la contamination (élimination des transferts horizontaux) et en gérant les problèmes de paralogie. Une fois ce jeu de données construit, nous allons le stresser de plusieurs façons.

Box 7. Les artefacts de reconstruction phylogénétique : hétérogénéités de substitution

Lorsque l'on compare des séquences homologues, si ces dernières ont divergé d'un ancêtre commun récent (*shallow phylogeny*), les séquences ne diffèrent probablement pas beaucoup, de quelques bases seulement. En revanche, plus on remonte dans le temps avec des espèces qui partagent un ancêtre commun très ancien (*deep phylogeny*), plus il est difficile de trouver des homologies. Les séquences d'ADN homologues diffèrent alors souvent de beaucoup de bases et sont de longueurs très différentes. Ces différences s'expliquent par l'accumulation de mutations au fil du temps (substitutions, insertions, délétions etc.). Les mutations sont les modifications de la séquence nucléotidique d'un génome. Ainsi, une des quatre bases (A, T, C, G) peut être remplacée par une autre base, voire supprimée (ou une nouvelle base insérée). Ces modifications se font de façon générale via deux processus : l'erreur de copie ou un dommage que subit l'ADN.

Le modèle d'évolution le plus simple, appelé Jukes-Cantor (JC), ne devrait s'appliquer qu'à des alignements dont les substitutions accumulées par les séquences se seraient produites au même rythme, quelle que soit la composition en bases des séquences et le type de mutation (transition vs transversion). Ce modèle suppose une évolution homogène et stationnaire le long de l'arbre du vivant. Or on sait maintenant que ces hypothèses sont très éloignées de ce qui est observé dans les séquences réelles. Il est donc nécessaire de corriger ces distances (similarités entre paires de séquences) en prenant en compte les particularités des modalités de substitution dans les séquences nucléotidiques.

L'un des problèmes majeurs lorsque l'on veut établir une phylogénie est de tenir compte des substitutions pour en extraire le bon signal. Bien que Darwin ait proposé que l'évolution soit un processus continu, rien n'indique que cette évolution se déroule de manière constante et uniforme au sein des différentes lignées. Simpson parle de "tempo et mode d'évolution" pour désigner la façon dont le rythme (tempo) et le type (mode) de changement évolutif peuvent varier à la fois entre les différentes lignées, au cours du temps et selon la biogéographie. Pour comparer les modes et tempos au travers des lignées et du temps, il est nécessaire d'estimer ces taux d'évolution. Malheureusement, le processus de substitution est très hétérogène, tant sur le site considéré qu'au cours du temps, rendant difficile la création de modèles complets et efficaces. Plusieurs types d'hétérogénéité ont été recensés :

- Hétérogénéité de substitution entre des états de caractères. Il existe des substitutions

beaucoup plus simples à faire et qui sont donc plus fréquentes. Par exemple, les substitutions de type transition (entre purines ou pyrimidines) sont plus fréquentes que les transversions (entre une purine et une pyrimidine ou inversement). L'estimation de ces taux de transition et transversion se fait via le modèle de Kimura à 2 paramètres (K2P). De plus, on pourrait s'attendre à trouver une proportion équivalente de chaque base azotée (25 %). Or, on observe très souvent une prédominance de certaines bases. Pour cela, il existe un modèle appelé Felsenstein-81 (F81). Il existe également un modèle appelé HKY qui prend en compte à la fois les taux de transition et transversion ainsi que les proportions de chaque base.

- Hétérogénéité des taux de substitution entre sites. Les premiers modèles utilisés les plus simples considéraient que les mutations se faisaient de façon équiprobable à travers tous les sites d'une séquence. Or, on sait qu'il n'en est rien. Certains sites sont plus susceptibles d'accumuler des substitutions, générant ainsi des artefacts. En effet, un site peut subir une très forte sélection avec de fortes contraintes, quand d'autres sites (par exemple la troisième base des codons) peuvent avoir moins de contraintes. D'autre part, certains sites ont une forte importance fonctionnelle et sont donc globalement moins tolérants aux mutations. Un site "rapide" connaît statistiquement plus de substitutions qu'un site "lent" pendant un même laps de temps. Si on pouvait séparer de façon fiable les sites lents des sites rapides et les analyser dans deux jeux de données séparés, on obtiendrait idéalement les mêmes topologies d'arbre, à la différence que la taille des branches serait beaucoup plus importante avec les sites rapides. Pour prendre en compte cette hétérogénéité, on peut considérer que les taux de substitutions sont variables entre sites et suivent une distribution du type gamma.
- Hétérogénéité des processus de substitution entre sites. Les premiers modèles phylogénétiques considéraient que n'importe quel acide aminé était possible à chaque site. Or, on sait par la biochimie que seuls certains sont acceptables pour une position donnée, compte tenu de leurs propriétés biochimiques (ex. chargé ou non, hydrophobe ou hydrophile). Deux séquences d'acides aminés très semblables l'une à l'autre mais prises chez deux espèces différentes sont bien souvent proches d'un point de vue fonctionnel ; une homologie de séquence reflète des fonctions biologiques similaires. Ces substitutions sont dites "conservatrices". Comme le nombre de paramètres requis pour décrire les préférences en acides aminés de chaque site est colossal, une solution est de regrouper par catégorie les sites ayant les mêmes propriétés. Pour éviter la question délicate du nombre de catégories à avoir, les modèles de type CAT utilisent le processus a priori de Dirichlet pour associer chaque site à une CATégorie définie par les fréquences d'états de leur caractère. Ces modèles CAT sont ainsi moins susceptibles à l'homoplasie et à l'artefact d'attraction des longues branches.
- Hétérogénéité de composition au cours du temps : il s'agit des fréquences stationnaires des nucléotides/acides aminés qui changent dans l'arbre.
- Hétérogénéité des processus de substitution d'un site au cours du temps (hétéropécillie). Il s'agit des profils d'acides aminés acceptables qui changent dans l'arbre, site par site, au cours du temps.
- Hétérogénéité du taux de substitution d'un site au cours du temps au sein d'une lignée (hétérotachie et covarion). A cause des changements de contraintes fonctionnelles des sites

d'une protéine et/ou du phénomène d'épistasie (interaction entre deux gènes ou plus, situés sur des loci différents, dont l'expression de l'un détermine l'expression de l'autre), la probabilité d'accepter une mutation à une position donnée varie le long des branches de l'arbre. Cela a notamment été montré pour le cas du cytochrome b (une protéine mitochondriale impliquée dans la respiration cellulaire) où la vitesse d'évolution de toutes les positions variables change au cours de l'évolution des vertébrés, sans que ces changements ne soient corrélés entre positions. De même, au sein d'une même protéine, les positions critiques pour son bon fonctionnement ne sont pas forcément les mêmes à travers le temps. Dans les années 70, Fitch et Markowitz (1970) montrèrent que, à un instant donné, seulement une petite fraction des positions sont susceptibles de varier, permettant la fixation de mutations. Ce comportement n'a biologiquement parlant rien de surprenant, dans le sens où il paraît logique que les contraintes fonctionnelles des sites d'une protéine (et donc leur vitesse d'évolution) vont changer au cours de l'évolution, et ce de façon indépendante selon les lignées. Par conséquent, cela peut mener à des artefacts phylogénétiques dans les cas où les proportions de sites invariables d'espèces non apparentées convergent.

Références

Gouy R, Baurain D, Philippe H. 2015 Rooting the tree of life: the phylogenetic jury is still out. Phil. Trans. R. Soc. B 370: 20140329. http://dx.doi.org/10.1098/rstb.2014.0329

Kapli P, Yang Z, Telford MJ. Phylogenetic tree building in the genomic age. Nat Rev Genet. 2020 Jul;21(7):428-444. doi: 10.1038/s41576-020-0233-0. Epub 2020 May 18. PMID: 32424311.

Olivier Jeffroy, Henner Brinkmann, Frédéric Delsuc, Hervé Philippe, Phylogenomics: the beginning of incongruence?, Trends in Genetics, Volume 22, Issue 4, 2006, Pages 225-231, ISSN 0168-9525.

Roure B, Philippe H. 2011. Site-specific time heterogeneity of the substitution process and its impact on phylogenetic inference. BMC Evol. Biol. 11:17.

Lartillot, N., et H. Philippe. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol Biol Evol 21:1095-109.

Lartillot, N., Brinkmann, H., & Philippe, H. (2007). Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. BMC evolutionary biology, 7 Suppl 1(Suppl 1), S4.

Lockhart, P. J., C. J. Howe, D. A. Bryant, T. J. Beanland, et A. W. Larkum. 1992. Substitutional bias confounds inference of cyanelle origins from sequence data. J Mol Evol 34:153-62.

Galtier, N. 2001. Maximum-likelihood phylogenetic analysis under a covarion-like model. Mol Biol Evol 18:866-73.

Miyamoto, M.M., & Fitch, W.M. (1996). Constraints on protein evolution and the age of the eubacteria/eukaryote split. Systematic biology, 45 4, 568-75.

Galtier, N., et M. Gouy. 1995. Inferring phylogenies from DNA sequences of unequal base compositions. Proc Natl Acad Sci U S A 92:11317-21.

Galtier, N., et M. Gouy. 1998. Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. Mol Biol Evol 15:871-9.

P. Lopez, D. Casane, H. Philippe, Heterotachy, an Important Process of Protein Evolution, Molecular Biology and Evolution, Volume 19, Issue 1, January 2002, Pages 1–7

Lopez, P., Casane, D. & Philippe, H. (2002). Bio-informatique (5) : phylogénie et

évolution moléculaires. M/S : médecine sciences, 18(11), 1146–1154.
Mooers, A. O., et E. C. Holmes. 2000. The evolution of base composition and phylogenetic inference. Trends Ecol Evol 15:365-369.
Fitch, W. M., et E. Margoliash. 1967. A method for estimating the number of invariant amino acid coding positions in a gene using cytochrome c as a model case. Biochem Genet 1:65-71.
Fitch, W. M., et E. Markowitz. 1970. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. Biochem Genet 4:579-93.
Fitch, W. M. 1971a. The nonidentity of invariable positions in the cytochromes c of different species. Biochem Genet 5:231–241.

La question de la validité d'un arbre est cruciale en phylogénie. Doit-on croire et tenir pour acquis l'arbre obtenu ? Un bon ajustement du modèle est nécessaire à la fois pour une estimation fiable de la phylogénie et pour l'interprétation des résultats, en particulier en présence de facteurs de confusion, c'est-à-dire de phénomènes méthodologiques et biologiques qui entraînent une violation du modèle (Young & Gillung, 2020) (**Tableau 4**). C'est l'impact de ces phénomènes que nous allons tenter d'évaluer dans le cadre de cette thèse. Notre but est d'évaluer la variabilité des arbres phylogénétiques en faisant varier les gènes et espèces, ainsi que les modèles et les méthodes de reconstruction phylogénétique employées. Diverses méthodes sont couramment employées afin de mesurer la solidité de chaque branche d'un arbre. Dans le cadre de cette thèse, l'évaluation se fera via une méthode de ré-échantillonnage sans remise (*jackknife*) et par le calcul des distances normalisées de Robinson-Foulds entre les arbres obtenus. Varier l'échantillonnage taxonomique grâce à un jackknife d'espèces permettra de vérifier son impact sur les résultats obtenus. Puis nous allons identifier les bipartitions possibles afin d'étudier les changements topologiques récurrents et identifier si certains sont dû à des artefacts. Enfin pour chaque réplica, nous testerons plusieurs effets : impact du modèle utilisé, retrait de sites…

Sources d'erreur systématique	Stratégies pour surmonter les difficultés	Références
Hétérogénéité compositionnelle	Supprimer les loci ou les taxons présentant une déviation extrême de la composition. Une autre solution, consistant à modéliser explicitement l'hétérogénéité de la composition, peut s'avérer plus satisfaisante	(Borowiec et al., 2019; Duchêne et al., 2017; Jeffroy et al., 2006; Roure & Philippe, 2011)
Données manquantes	Concevoir des ensembles de données plus petits (c'est-à-dire avec moins de loci) mais plus complets, par opposition à des ensembles de données plus grands (c'est-à-dire avec plus de loci) mais plus clairsemés. Analyser des sous-échantillons de données pour voir si les données manquantes sont source de conflit.	(Hosner et al., 2016; Kocot et al., 2017; Roure et al., 2013; Smith et al., 2020)
Hétérogénéité des longues branches	Effectuer l'analyse avec et sans les taxons et les gènes les plus susceptibles d'être sensibles à l'attraction des longues branches (par exemple en supprimant les sites ou les taxons qui évoluent rapidement).	(Kück & Wägele, 2016; Nosenko et al., 2013; Qu et al., 2017; Struck, 2014)
Paralogie	Filtrer et supprimer les loci paralogues à l'aide d'une méthode appropriée pour la prédiction de l'orthologie. Examiner soigneusement les topologies des gènes pour détecter les paralogues.	(Betancur-R. et al., 2014; Siu-Ting et al., 2019; Struck, 2013)
Hétérotachie, hétéropécillie	Supprimer des sites les plus variés de l'alignement. Utiliser des modèles évolutifs qui modélisent explicitement l'hétérotachie et/ou l'hétéropécillie	(Bouckaert & Lockhart, 2015; Crotty et al., 2020; Kocot et al., 2017; Zhong et al., 2011)
Hétérogénéité des arbres de gènes	Utiliser des approches de coalescence. Effectuer des analyses statistiques de binning, analyser la concordance des arbres.	(Betancur-R. et al., 2013; Richards et al., 2018)

Tableau 4. Principaux facteurs de confusion dans les analyses phylogénomiques et stratégies pour les atténuer (Young & Gillung, 2020).

Afin de contrer au mieux l'erreur stochastique, nous avons utilisé deux méthodes de reconstruction différentes : une méthode de super-matrice et une méthode de super-arbre (**Box 8**). Dans les deux cas, notre objectif a été d'obtenir une collection d'arbres comprenant un arbre LG4X, LG+C20+F+G et LG+C60+F+G et PMSF pour chacun des 5 réplicas d'espèces. En effet, les analyses de vraisemblance via les modèles LG+C20+F+G et LG+C60+G4+F améliorent le modèle LG4X en modélisant l'hétérogénéité de composition spécifique à chaque site à l'aide de 20 ou 60 catégories. Ces modèles ajustent beaucoup mieux les données que LG4X si l'on tient compte du BIC (*Bayesian Information Criterion*) (**Box 5**).

Box 8. Combattre l'erreur stochastique : super-matrice vs super-arbre

L'erreur stochastique (= erreur d'échantillonnage)

Même si l'évolution a eu lieu exactement telle que supposée par le modèle évolutif utilisé pour l'inférence phylogénétique, il est possible qu'un arbre incorrect soit obtenu à cause de la taille finie des alignements. Si le signal extrait est trop faible, on ne peut pas distinguer la meilleure topologie parmi plusieurs solutions, ce qui mène à des nœuds faiblement soutenus. Ce problème, très connu en statistique, se nomme l'erreur d'échantillonnage ou erreur stochastique, et dépend donc principalement de la quantité de données analysables. Ainsi, par définition, l'erreur stochastique disparaît dans des échantillons de taille infinie. Ce problème est particulièrement présent lorsque l'on travaille avec de grandes échelles de temps, car les sites deviennent rapidement saturés par de

multiples substitutions, effaçant alors le signal originel au profit d'un faux signal dû aux biais mutationnels. L'utilisation de modèles complexes décrivant au mieux les données peut significativement contribuer à diminuer ce problème. De même, plus une protéine évolue lentement, plus il est vraisemblable qu'elle conserve un signal phylogénétique ancien, mais aussi que le signal son évolution plus contrainte la rende plus susceptible aux convergences.

Pour inférer la phylogénie la plus représentative de « l'arbre vrai », une solution consiste à s'inspirer du concept méthodologique nommé « *Total Evidence* » en collectant plusieurs marqueurs phylogénétiques indépendants afin d'en extraire, dans un esprit de consensus, le signal majoritaire. C'est ainsi que s'est développée la phylogénomique, extension de la phylogénie qui vise à reconstruire l'histoire évolutive des espèces en combinant l'information phylogénétique de nombreux gènes.

Dans ce cas, deux approches sont possibles : la super-matrice ou le super-arbre.

Super-matrices

Une super-matrice est un alignement synthétique obtenu par concaténation d'un ensemble d'alignements de séquences. On applique alors ensuite sur ce « super-alignement » une méthode standard de reconstruction phylogénétique. Il a été montré que cette méthode permet d'obtenir des arbres fiables, notamment dû au fait que la concaténation de plusieurs alignements permet de diminuer l'erreur stochastique. Un biais à prendre en compte dans cette méthode est le nombre de gènes manquants chez certaines espèces, qui produisent des caractères manquants au sein des super-matrices. Toutefois, la proportion de données manquantes peut être relativement haute sans que ceci n'entraîne une baisse de résolution dans la reconstruction phylogénétique. En effet, la sélection d'un modèle adéquat d'évolution des séquences est plus bénéfique pour la précision phylogénétique que de réduire le taux de données manquantes. Un des inconvénients majeurs de cette méthode est le coût important en ressources informatiques. De plus, la super-matrice émet l'hypothèse que l'évolution de chacun des gènes sélectionnés puisse être expliquée par un modèle unique et donc que leur phylogénie soit identique à la phylogénie des espèces. Or, on sait que cela est faux, principalement pour deux raisons. La première est que l'histoire d'un gène ne reflète pas nécessairement l'histoire d'une espèce. C'est en particulier le cas lorsqu'il y a des transferts de gènes (gènes xénologues) ou en cas de polymorphisme ancestral (*incomplete lineage sorting* = ILS). Dans ce dernier cas, l'existence d'un polymorphisme ancestral disparu aujourd'hui conduit à la fixation d'un allèle qui ne reflète pas l'histoire de son espèce. Ce problème peut être géré par la réconciliation. La deuxième raison est quant à elle due à des problèmes d'hétérogénéité de substitution. Il y a donc ici un problème lié à la qualité de la modélisation de l'évolution. Pour combler cela, des modèles partitionnés supposent que l'évolution des séquences puissent être différente pour chaque gène.

Super-arbres

Un super-arbre est un arbre synthétique obtenu par l'assemblage d'un ensemble d'arbres phylogénétiques définissant des ensembles d'espèces recouvrant (souvent) incomplètement les espèces d'intérêt. Ce super-arbre se doit de refléter au mieux les différentes histoires évolutives modélisées par chaque arbre source (issus de gènes différents). Ce problème est une généralisation du consensus d'une forêt d'arbres, ce dernier consistant à rechercher l'arbre le plus représentatif d'une collection d'arbres sources (cf. **Box 10** sur le bootstrap et le jackknife). Parmi les nombreuses méthodes de super-arbres existantes, la plus populaire reste celle de la méthode de représentation

matricielle avec parcimonie (MRP pour *Matrix Representation with Parsimony*). Cette méthode permet d'obtenir des super-arbres proches des arbres issus des super-matrices, à condition que les arbres sources individuels soient congruents entre eux et que la représentation matricielle présente peu d'états de caractères manquants. Dans le cas contraire, le super-arbre sera peu fiable à cause de l'accumulation des erreurs stochastiques. La méthode du super-arbre ne résout pas directement le problème de l'erreur stochastique car chaque arbre de gène reste possiblement affecté. En revanche, un des avantages de cette méthode est qu'elle permet de combiner des arbres issus de données morphologiques et moléculaires et de détecter les problèmes de xénologie. De plus, c'est une méthode rapide en temps de calcul et qui facilite le jackknife de gènes. Dans le cadre de notre thèse, nous avons utilisé ASTRAL, qui est une amélioration de la méthode MRP. ASTRAL (*Accurate Species TRee ALgorithm*) est un outil d'estimation d'arbres d'espèces non racinés étant donné un ensemble d'arbres de gènes individuels non racinés. ASTRAL est statistiquement cohérent dans le modèle coalescent multi-espèces (et est donc utile pour gérer les polymorphisme ancestraux). Il utilise la programmation dynamique pour trouver une solution exacte, même avec de grands jeux de données. ASTRAL trouve l'arbre d'espèces qui a le nombre maximum d'arbres quatuor induits partagés avec l'ensemble d'arbres de gènes, sous la contrainte que l'ensemble de bipartitions dans l'arbre d'espèces provient d'un ensemble prédéfini de bipartitions. Cet ensemble prédéfini est décidé empiriquement par ASTRAL. ASTRAL utilise l'approche du *Maximum Quartet Support Species Tree* (MQSST). Il coupe l'arbre en $\binom{n}{4}$ quatuors d'espèces. Puis, il trouve l'arbre dominant pour chacun des quatuors. Enfin il combine ces derniers quatuors. Pour un quatuor, le MQSST tient compte de la fréquence relative des trois topologies et poids alternatifs et les pondère en conséquence. Ainsi, si la topologie dominante (c'est-à-dire la plus fréquente) d'un quatuor est beaucoup plus fréquente que les alternatives, les arbres qui n'induisent pas la topologie dominante sont pénalisés. En revanche, si les trois topologies du quatuor ont toutes des fréquences proches de 1/3, alors ce quatuor contribuera peu à l'optimisation du problème. Cela induit que plus un arbre de gène est fréquent, plus il contribuera à renforcer cette topologie dans l'arbre d'espèces.

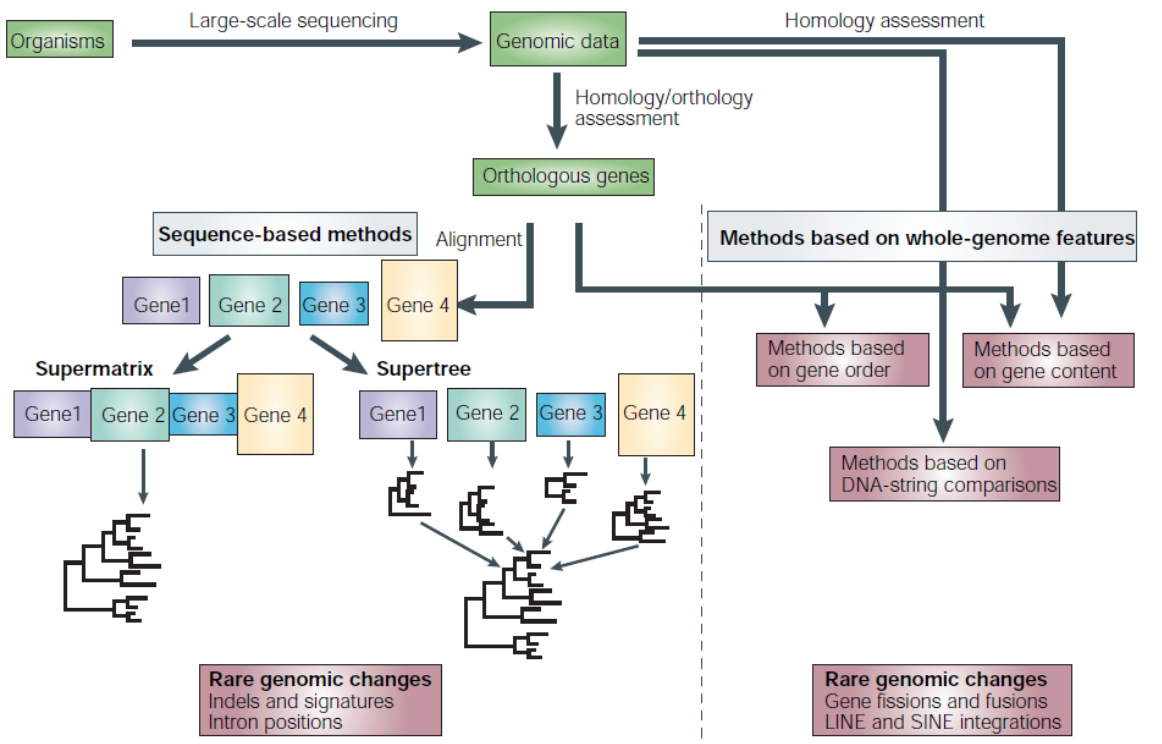


Figure 19. Méthodes d'inférence phylogénétique

Cet organigramme présente les étapes d'inférence menant à la construction d'arbres phylogénétiques à partir de données génomiques. Ces données sont obtenues à partir de séquençage à grande échelle d'ADN. Après évaluation de l'homologie et de l'orthologie de certains gènes, des jeux de données de gènes orthologues de notre sélection d'espèces sont ensuite assemblés. Les arbres sont alors calculés soit par la méthode des super-arbres soit par la méthode d'une super-matrice.

Références

Béatrice Roure, Denis Baurain, Hervé Philippe, Impact of Missing Data on Phylogenies Inferred from Empirical Phylogenomic Data Sets, Molecular Biology and Evolution, Volume 30, Issue 1, January 2013, Pages 197–214

Kapli P, Yang Z, Telford MJ. Phylogenetic tree building in the genomic age. Nat Rev Genet. 2020 Jul;21(7):428-444. Epub 2020 May 18. PMID: 32424311.

Carnap, R. Logical Foundations of Probability. (1950).

Swofford, D. L., G. J. Olsen, P. J. Waddell, et D. M. Hillis. 1996. Phylogeny inference. Pages 407-514 dans Molecular Systematics (2nd ed) (D. M. Hillis, C. Moritz, et B. K. Mable, eds.). Sinauer Associates, Sunderland, Massachusetts.

Ragan, M. A. (1992). Matrix representation in reconstructing phylogenetic relationships among the eukaryotes. Biosystems, 28(1-3), 47–55. doi:10.1016/0303-2647(92)90007-l

Lecointre, G. Total evidence requires exclusion of phylogenetically misleading data. Zool. Scr. 101–117 (2005).

Rieppel, O. The Philosophy of Total Evidence and its Relevance for Phylogenetic Inference. Pap. Avulsos Zool. 45, 77–89 (2005).

Rieppel, O. (2008). "Total evidence" in phylogenetic systematics. Biology & Philosophy, 24(5), 607–622. doi:10.1007/s10539-008-9122-1

Delsuc, F., Brinkmann, H. & Philippe, H. Phylogenomics and the reconstruction of the tree of life. Nat Rev Genet 6, 361–375 (2005).
Baurain, D., Philippe, H., 2010. Current approaches to phylogenomic reconstruction. In: Caetano-Anolles, G. (Ed), Evolutionary Genomics and Systems Biology. Wiley, Hoboken, NJ, pp. 17–41
Driskell, A. C., C. Ane, J. G. Burleigh, M. M. McMahon, C. O'Meara B, et M. J. Sanderson. 2004. Prospects for building the tree of life from large sequence databases. Science 306:1172-4.
Philippe, H., E. A. Snell, E. Bapteste, P. Lopez, P. W. Holland, et D. Casane. 2004. Phylogenomics of eukaryotes: impact of missing data on large alignments. Mol Biol Evol 21:1740-52.
Wiens, J. J. 2003. Missing data, incomplete taxa, and phylogenetic accuracy. Syst Biol 52:528-38.
Keeling PJ, Burger G, Durnford DG, Lang BF, Lee RW, Pearlman RE, Roger AJ, Gray MW. The tree of eukaryotes. Trends Ecol Evol. 2005 Dec;20(12):670-6. Epub 2005 Oct 10. PMID: 16701456.
Mirarab S, Reaz R, Bayzid MS, Zimmermann T, Swenson MS, Warnow T. ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics*. 2014;30(17):i541-i548.

2 SELECTION D'ESPECES

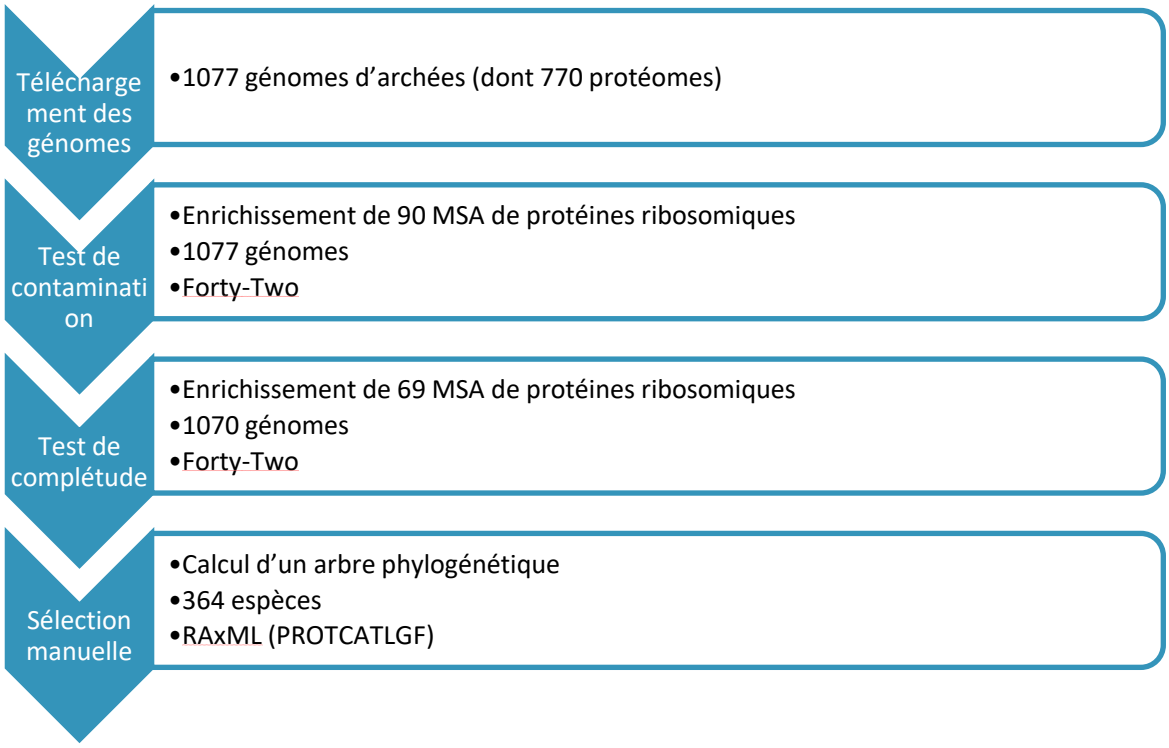


Figure 20. Protocole de sélection des espèces

2.1 COLLECTION DES GENOMES D'ARCHEES

1077 génomes d'Archaea ont été collectés à la fois sur GenBank et RefSeq via le portail du National Center for Biotechnology Information (NCBI). Sur ces 1077 génomes, on notera que 770 disposaient de leur prédiction en protéome, soit environ les trois quarts. La qualité de chaque génome a été évaluée afin de rassembler les génomes ayant la plus haute qualité possible. Nous avons ainsi évalué les paramètres suivants (cf. **Supp.Mat quast.csv**) :

- le nombre de contigs ;
- le nombre de contigs supérieur à 1000 nucléotides (**Figure 21**) ;
- la taille totale du génome ;

- la taille totale du génome basée sur les contigs supérieurs à 1000 nucléotides ;
- la taille du plus grand contig ;
- les proportions N50 et N75 ;
- les proportions L50 et L75 ;
- le nombre de N pour 100 kpb

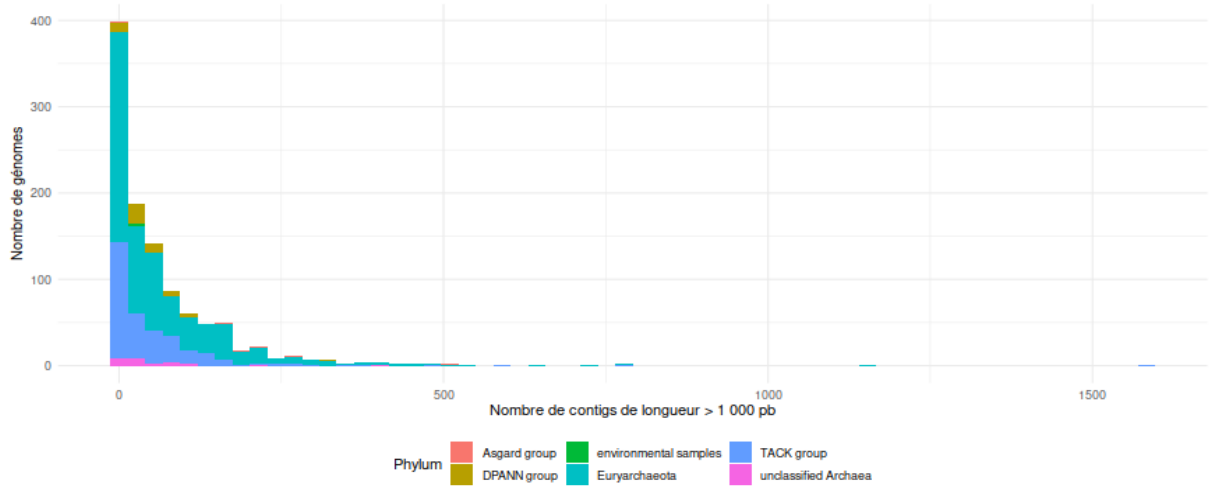


Figure 21. Distribution du nombre de contigs permettant d'estimer la qualité des assemblages de nos génomes d'archées.

Les Euryarchaeota (en particulier des Halobacteria) et des Crenarchaeota sont très sur-représentés au sein des génomes téléchargés (**Figure 22**). Les archées du groupe « *unclassified* » se verront par la suite attribuer un clade en accord avec leur phylogénie basée sur les protéines ribosomiques. On observe également une plus grande taille de génome chez les représentants du groupe Asgard (**Figure 23**). Sur le plan biologique, cela pourrait être interprété comme la mise en place d'une complexification vers la lignée eucaryote (ou au contraire comme une simplification des eucaryotes vers les archées) ou plus prosaïquement comme un problème technique lié à la reconstruction des MAG (ex. possibles contaminations dues à une mauvaise séparation des scaffolds de plusieurs organismes).

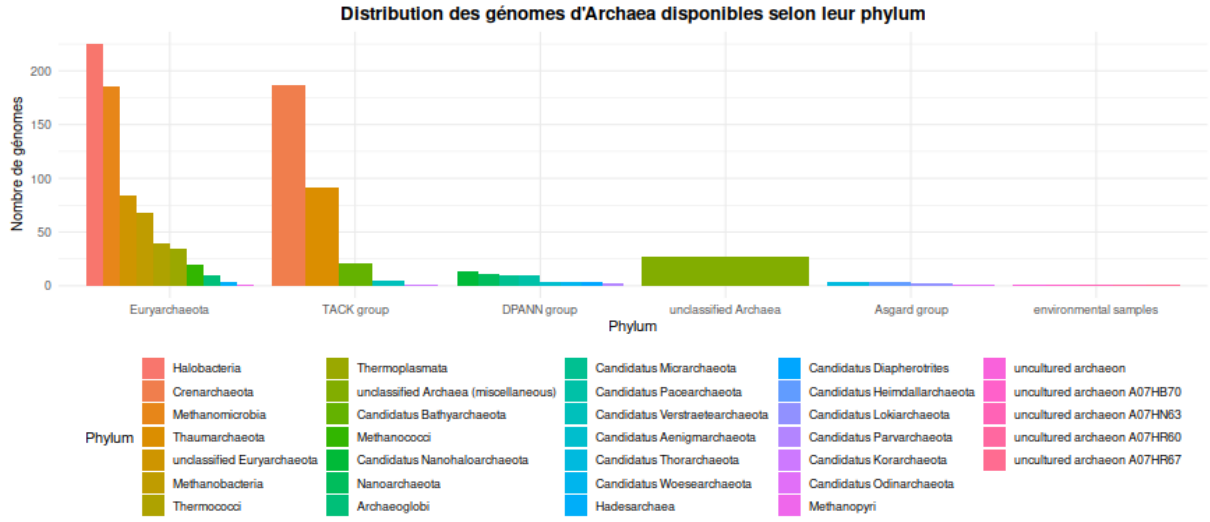


Figure 22. Distribution des génomes d'archées disponibles selon leur phylum.

Les couleurs correspondent à la taxonomie du NCBI. On observe une forte représentation des Euryarchaeota et des Crenarchaeota.

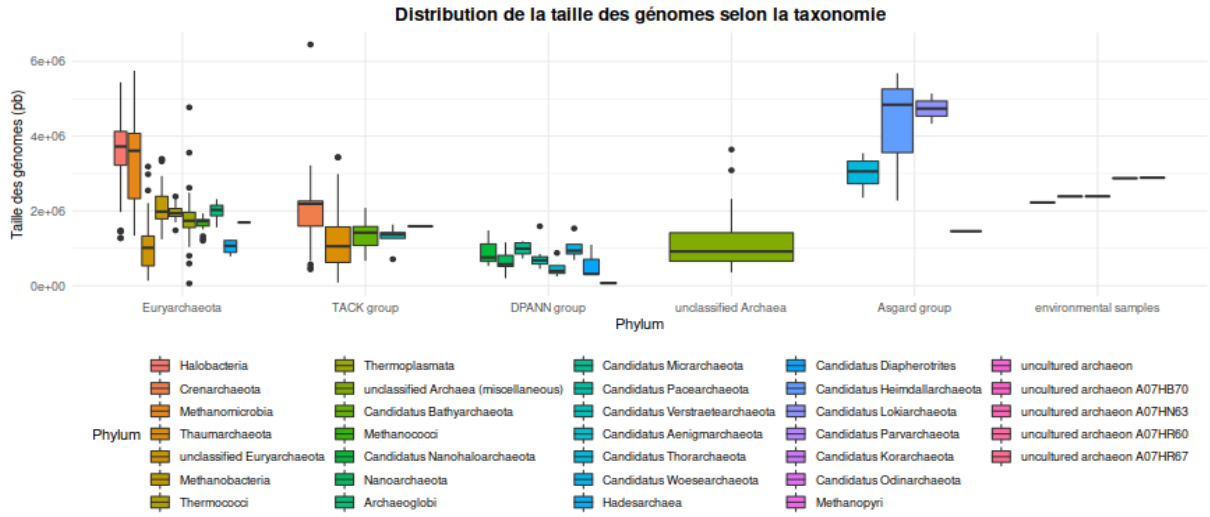


Figure 23. Distribution de la taille des génomes selon leur phylum.

Les couleurs correspondent à la taxonomie du NCBI. On observe une plus grande taille de génome chez les représentants du groupe Asgard qui pourrait être interprété soit comme la mise en place d'une complexification vers la lignée eucaryote (ou au contraire comme une simplification des eucaryotes vers les archées), soit comme un problème technique lié à la reconstruction des MAG.

2.2 ANALYSES DES PROTEINES RIBOSOMIQUES ET ECHANTILLONNAGE TAXONOMIQUE PRELIMINAIRE

Afin d'établir une phylogénie préliminaire des archées, un premier jeu de données a été assemblé en enrichissant 90 alignements multiples de séquences (MSA) de protéines ribosomiques d'archées et de bactéries. Certaines protéines ribosomiques sont universelles à l'ensemble des êtres vivants. En revanche, certaines sont spécifiques aux bactéries, d'autres aux eucaryotes, et certaines sont communes à l'ensemble archées + eucaryotes. Sur nos 1077 génomes, 1070 ont au moins une protéine ribosomique. Nous avons 7 « espèces » qui ont disparu car, très incomplètes, elles ne contenaient aucune protéine ribosomique (cf. **Supp.Mat 7-**

genomes-non-rajoutes.txt). Nous avons donc éliminé ces génomes de notre jeu de données. Certains génomes en revanche ne se sont pas ajoutés sur tous les alignements de protéines ribosomiques, laissant à penser qu'ils sont soit incomplets où que le BLAST n'a pas réussi à les aligner (propriété particulière de la protéine dû par exemple à une taille trop petite ?). L'étude de leur répartition au sein de nos génomes permet d'évaluer quels génomes sont contaminés. Pour chaque génome, nous avons recherché les séquences qui sont homologues aux séquences du MSA de l'alignement de référence puis trié les orthologues des paralogues possibles. On remarque que quelques archées se sont rajoutées sur les MSA de protéines ribosomiques en principe exclusivement bactériennes (en vert sur la **Figure 24**), ce qui laisse penser qu'elles pourraient être contaminées. En effet, les génomes correspondants sont censés être seulement archéens. Or, après enrichissement, des séquences bactériennes se rajoutent, interprétées par 42 comme des contaminations. Nous pouvons ainsi évaluer par la même occasion la complétude de nos génomes. En effet, on s'attendrait normalement à ce que chaque protéine ribosomique d'archées ou universelle soit retrouvée dans chacun de nos 1077 génomes.

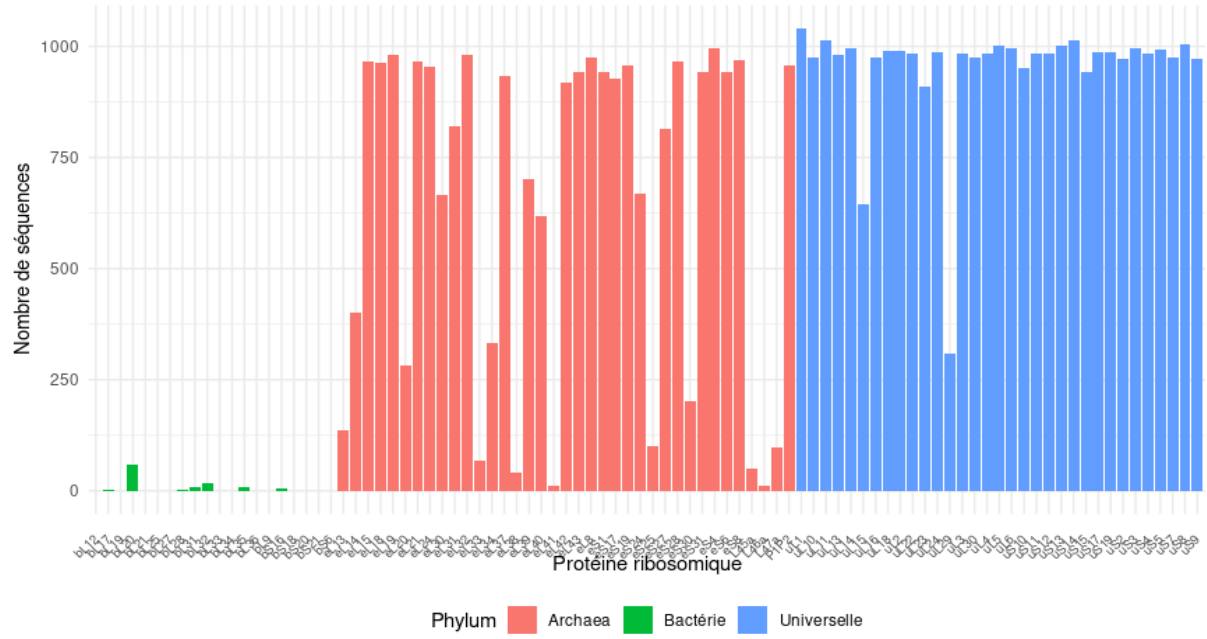


Figure 24. Nombre de séquences rajoutées par 42 pour chaque protéine ribosomique.
Quelques archées se sont rajoutées sur les MSA de protéines ribosomiques en principe exclusivement bactériennes (en vert)

Afin de vérifier si des génomes d'archées ont été contaminés (ou non) par des séquences bactériennes, nous avons effectué un test de contamination (Irisarri et al., 2017; Simion et al., 2017) avec "42", en utilisant cette fois des bactéries comme protéomes de référence lors du BRH (**Box 9**). En effet, ainsi, ce seront les séquences de protéines ribosomiques bactériennes qui seront spécifiquement recherchées et non celles des archées. Dans le cas de protéines ribosomiques universelles, on ne s'attend pas à observer de différences entre les deux analyses de « 42 » rapportées ici.

Box 9. Le BRH : *Best Reciprocal Hits*

L'acquisition d'un ensemble d'alignements de gènes orthologues sur base de données génomiques sert de point de départ à la construction d'un jeu de données phylogénomique. C'est à partir des

séquences constituant ces alignements que sont recherchés les orthologues correspondants chez d'autres espèces. La méthode du BRH s'appuie sur une meilleure conservation des séquences entre gènes orthologues mise en évidence par des comparaisons systématiques entre au moins deux génomes complets. Elle consiste à rechercher, lors de comparaisons inter-génomiques, les gènes de plusieurs espèces présentant les meilleures similitudes mutuelles. L'idée sous-jacente est que les "vrais orthologues" ont gardé les mêmes propriétés que leur ancêtre commun et, par conséquent, qu'ils présentent la meilleure conservation de séquence. Les paralogues, quant à eux, auront pu évoluer avec moins de contraintes (ou de nouvelles contraintes). Le BRH est probablement la définition la plus fonctionnelle de l'orthologie : deux gènes de deux génomes différents sont postulés orthologues si leurs protéines se retrouvent mutuellement comme *best hit* (meilleur score) dans les génomes comparés. Ainsi, si en partant d'une séquence A du protéome d'une espèce X, on obtient comme meilleur résultat de similarité (le BEST HIT) une séquence B dans le protéome de l'espèce Y, et si la séquence A est à son tour le BEST HIT dans le protéome de l'espèce X en partant de la séquence B de l'espèce Y, alors celles-ci seront considérées comme meilleurs hits réciproques et interprétées comme probablement orthologues.

On distingue deux principaux types d'algorithmes de recherche de similarité entre séquences : ceux basés sur l'alignement local entre séquences (exemple : BLAST), et ceux basés sur l'utilisation d'un modèle de Markov caché (i.e. HMM : *Hidden Markov Model*). Une fois les séquences homologues candidates récupérées, il est nécessaire de vérifier leur relation d'orthologie. Cette étape peut être réalisée par leur positionnement dans l'arbre du gène ou par leur comparaison à une base de données d'orthologues. Ce dernier point se rapproche du principe de réciprocité des meilleurs résultats de similarité. Cette condition de BRH est largement employée pour définir les relations d'orthologie. Il est à noter que les heuristiques de « 42 » sont plus complexes que ce qui est décrit ici, notamment pour gérer les paralogues récents (in-paralogues), même si le principe de base reste identique (cf**. Matériel & Méthodes**).

Références

Burki F, Shalchian-Tabrizi K, Pawlowski J. Phylogenomics reveals a new 'megagroup' including most photosynthetic eukaryotes. Biol Lett. 2008 Aug 23;4(4):366-9. doi: 10.1098/rsbl.2008.0224. PMID: 18522922; PMCID: PMC2610160.

Brown Matthew W., Sharpe Susan C., Silberman Jeffrey D., Heiss Aaron A., Lang B. Franz, Simpson Alastair G. B. and Roger Andrew J. 2013Phylogenomics demonstrates that breviate flagellates are related to opisthokonts and apusomonadsProc. R. Soc. B.2802013175520131755.

Ward N, Moreno-Hagelsieb G (2014) Quickly Finding Orthologs as Reciprocal Best Hits with BLAT, LAST, and UBLAST: How Much Do We Miss? PLoS ONE 9(7): e101850.

Tatusov RL, Koonin EV, Lipman DJ (1997) A genomic perspective on protein families. Science 278: 631–637.

Bork P, Dandekar T, Diaz-Lazcoz Y, Eisenhaber F, Huynen M, et al. (1998) Predicting function: from genes to genomes and back. J Mol Biol 283: 707–25.

Struck, T. H. (2013). The Impact of Paralogy on Phylogenomic Studies – A Case Study on Annelid Relationships. PLoS ONE, 8.

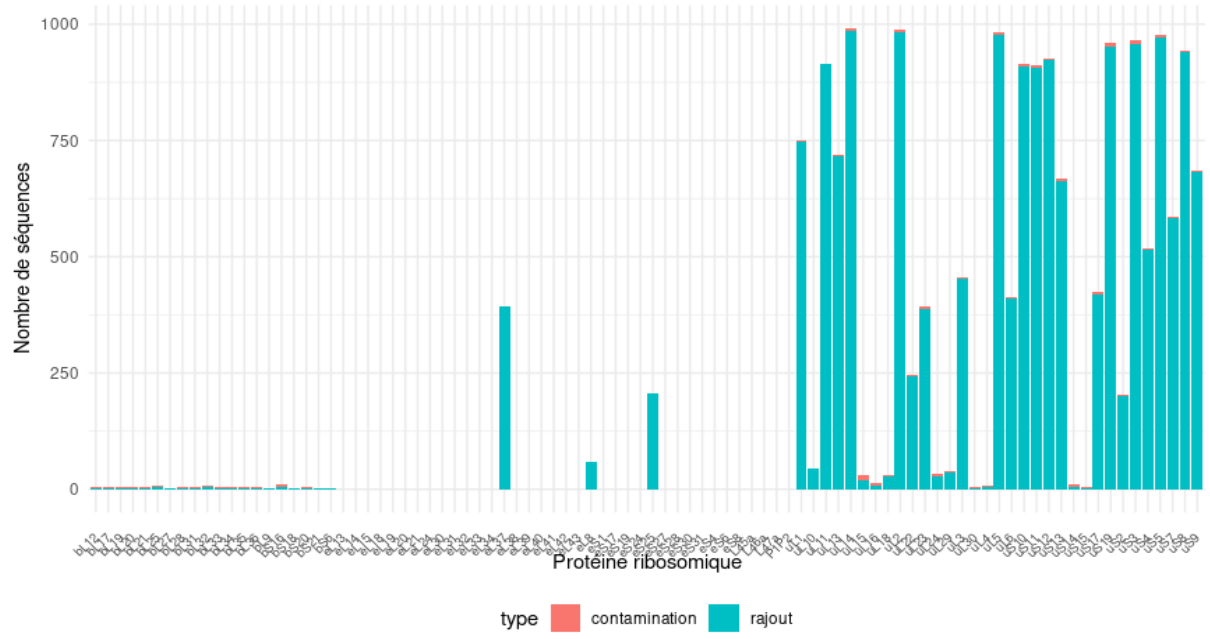


Figure 25. Nombre de séquences bactériennes rajoutées par 42 pour chaque protéine ribosomique. Notre protocole a permis de limiter la part de génomes contaminés.

On remarque une très faible part de contamination (en rouge sur la **Figure 25**). Nous obtenons ainsi une liste de 31 génomes ayant au moins une protéine ribosomique contaminée par une séquence non-archéenne (cf**. Supp.Mat 31-genomes-contamines.txt**).

Nous avons en parallèle concaténé 69 MSA de protéines ribosomiques en une super-matrice de 8 344 positions d'acides aminés bien conservés × 1070 espèces. Ces protéines ribosomiques nous ont servi d'indicateur afin d'évaluer la complétude des génomes (le nombre d'acides aminés concaténés reflétant le nombre de protéines ribosomiques retrouvées).

A partir de cet alignement, nous avons calculé un arbre phylogénétique avec RAxML avec une recherche par rapid-bootstrap (100 réplicas) en utilisant le modèle PROTCATLGF (cf. **Supp.Mat 1070sp-ribo.nex**). Nous avons alors sélectionné manuellement 364 espèces représentant au maximum la diversité des archées en tenant compte de nos informations de contamination et en favorisant à la fois la disponibilité de leur protéome ainsi que leur complétude (cf. **Supp.Mat archaea-364sp.lis**). Sur ces 364 archées, 305 avaient un protéome disponible (cf. **Supp.Mat selection-proteomes.txt**), tandis que les 59 espèces n'étaient disponibles que sous forme de génomes non-annotés (cf. **Supp.Mat selection-genomes.txt**). Un nouvel arbre ribosomique basé sur ce nouvel alignement de 7 817 positions et 364 espèces a alors été calculé, en utilisant la même méthode que précédemment (cf. **Figure 26** & **Supp.Mat FigS26**). Ainsi en réduisant le nombre d'espèces de 1070 à 364, nous avons perdu 527 colonnes. Cet arbre de 364 génomes est un condensé de notre arbre à 1070 espèces. Il représente bien la diversité des archées et lui est bien semblable, mais sans redondance excessive des groupes suréchantillonnés dans GenBank. Nous obtenons les groupes suivants : Archaeoglobi, Aenigmarchaeota, Bathyarchaeota, Diapherotrites, Heimdallarchaeota, Lokiarchaeota, Micrarchaeota, Nanohaloarchaeota, Pacearchaeota, Thorarchaeota, Verstraetearchaeota, Woesearchaeota, Crenarchaeota, Hadesarchaea, Halobacteria, Korarchaeota, Methanobacteria, Methanococci, Methanomicrobia, Methanopyrus, Nanoarchaeota, Odinarchaeota, Thaumarchaeota,

Thermococci, Thermoplasmata, Unclassified Euryarchaeota. Ce dernier groupe Unclassified Euryarchaeota pourrait être apparenté aux Hadesarchaeota.

On notera au sein de notre arbre ribosomique la présence de deux groupes appelés « Methanomicrobia » bien distincts. Nous remarquons que pour cet arbre ribosomique, le support statistique estimé par les proportions de bootstrap ne sont pas maximaux pour tous les nœuds, traduisant peut-être un manque de signal phylogénétique. De plus, il n'est pas à exclure que cet arbre soit affecté par divers artéfacts de reconstruction phylogénétique, en particulier l'artefact d'attraction des longues branches (cf. *archaeon CG2 30 31 98 GCA 001872885.1*). Ainsi, il est connu que les artéfacts génèrent un signal non-phylogénétique qui peut s'opposer au signal phylogénétique et affecter la résolution des arbres, que ceux-ci soient topologiquement corrects ou non (Baurain & Philippe, 2010). Afin de pallier ce problème, nous avons d'abord décidé d'augmenter notre nombre de gènes en étendant notre jeu de données aux gènes non ribosomiques, puis mis en place une série de stratégies pour limiter les artéfacts.

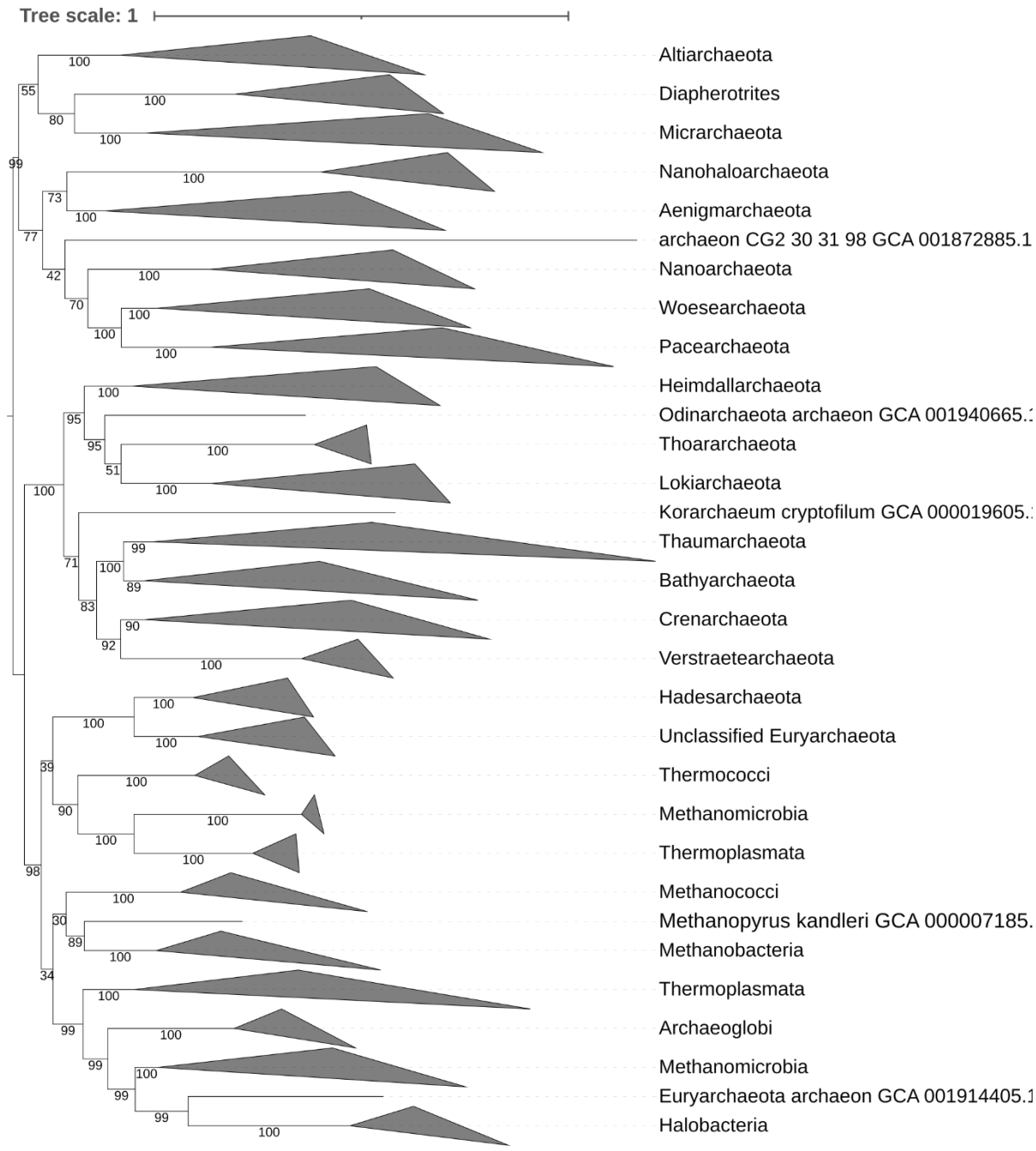


Figure 26. Arbre ribosomique de 364 archées calculé avec RAxML avec une recherche par rapid-bootstrap (100 réplicas) selon le modèle PROTCATLGF sur une super-matrice de 7 817 positions.
Les données sont fournies dans le **Supp.Mat FigS26**.

3 SELECTION DE GENES

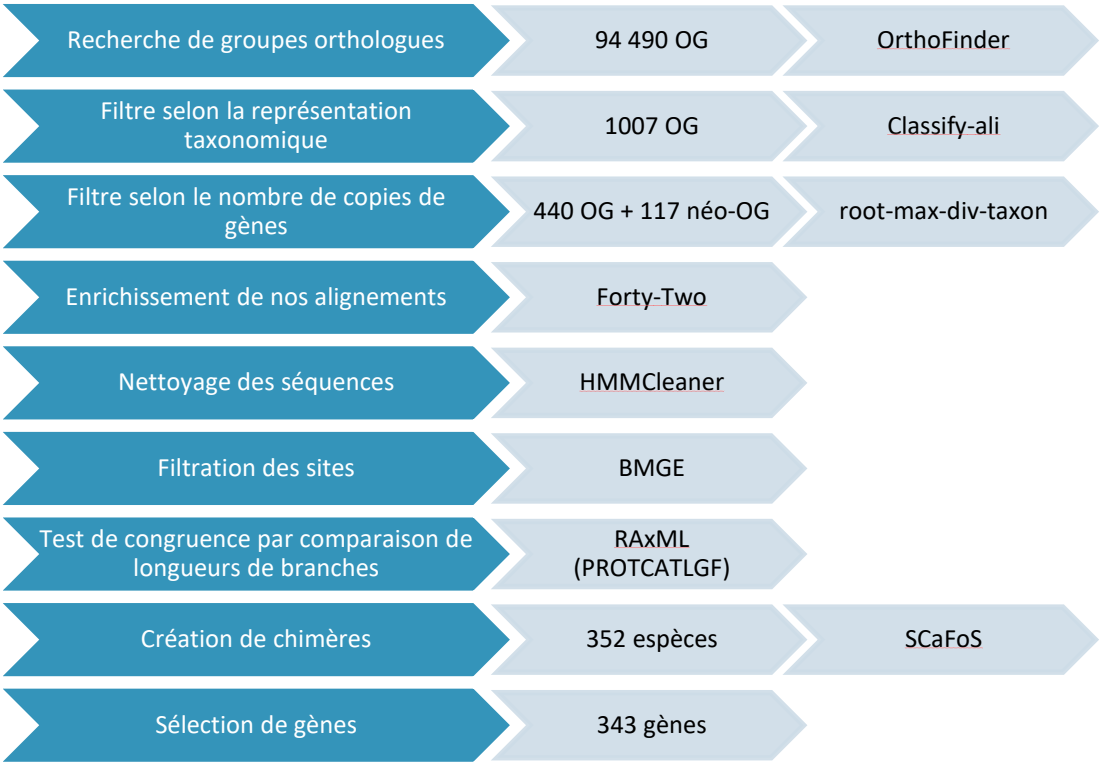


Figure 27. Protocole de sélection de gènes

3.1 CONSTRUCTION ET SELECTION DES GROUPES ORTHOLOGUES SELON LA REPRESENTATION TAXONOMIQUE

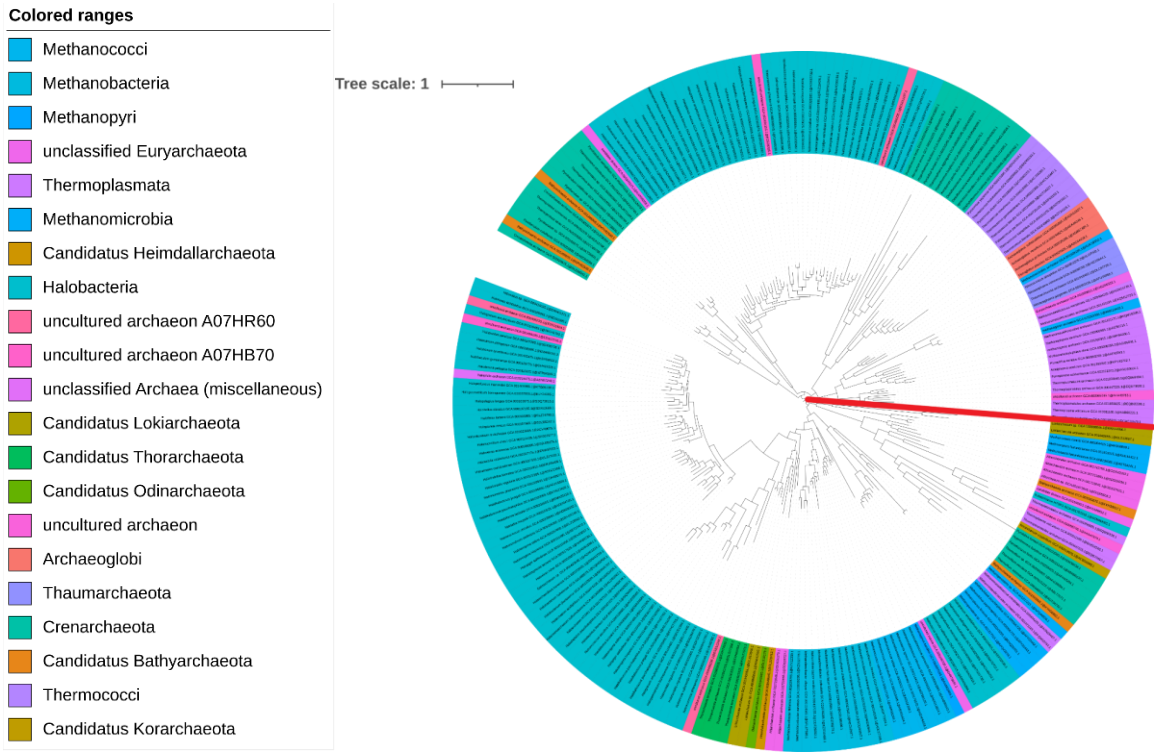
En vue d'élargir notre sélection de gènes au-delà des protéines ribosomiques, nous avons téléchargé depuis le NCBI les 305 protéomes correspondant à notre sélection d'espèces afin de générer des groupes de gènes orthologues. Nous avons ainsi généré 94 490 groupes orthologues (OG) (dont 66 798 singletons composés d'une seule séquence), sur lesquels nous avons appliqué successivement deux filtres, basés sur la taxonomie et le nombre d'espèces présentes. Le premier filtre, basé sur la taxonomie, nous a assuré d'avoir au moins une Euryarchaeota et une TACK (Thaumarchaeota, Aigarchaeota, Crenarchaeota, Korarchaeota) par gène. Nous avons ainsi retenu 4 843 OG. Quant au deuxième filtre, il nous a permis de conserver les gènes présents dans au moins ¼ de notre sélection de 364 espèces présentes (soit 91 espèces minimum), afin d'éviter les OGs trop faiblement échantillonnés. Après cette procédure, il nous est resté 1007 OG.

3.2 SELECTION DE GROUPES ORTHOLOGUES BASES SUR LE NOMBRE DE COPIES DES GENES ET GESTION DES FAMILLES MULTIGENIQUES

Nous avons eu recours à une méthode de découpe d'arbres phylogénétiques afin de trier et sélectionner nos gènes. Nous souhaitions créer deux jeux de données : l'un, de haute qualité, ne comprenant que des gènes en simple copie, l'autre, de qualité relativement moindre, ne comprenant que des gènes issus de duplications plus ou moins anciennes (out-paralogues).

Au total, la procédure de découpe d'arbres phylogénétiques a été répétée 4 fois. Un exemple de gène découpé par ce protocole est donné dans la **Figure 28**. Parmi les gènes découpés selon ces critères, nous avons supprimé, de la même façon qu'avec le jeu de données précédent, ceux tombant à moins de 91 espèces. Les résultats de chaque root-max-div-taxon sont donnés dans le **Tableau 5**. Au 1er tour, nous avons trouvé 21 gènes sans paralogue ni aucune espèce en

double et 731 gènes non coupés (ne répondant pas au critère de découpe ou qu'il n'y a pas assez de duplications), soit un total de 752 gènes. De ceux-là, nous récupérons ainsi après le 1er tour 440 gènes n'ayant aucune duplication. Les 312 gènes restants sont éliminés car jugés imparfaits (présences de duplications, mais impossibles de couper les arbres en deux sous-gènes). Ces gènes constituent notre premier jeu de données (cf. **Supp.Mat 440-genes.txt**). A l'issue de cette procédure, nous obtenons 117 gènes supplémentaires, provenant de la découpe de gènes dupliqués. Ces gènes seront traités à part et permettront de tester la robustesse des résultats obtenus sur les 440 gènes selon un principe de corroboration. Nous considérerons néanmoins ces 117 gènes « néo-orthologues » comme formant un jeu de données potentiellement moins fiable, en comparaison des 440 dont l'orthologie des séquences fait moins de doute.



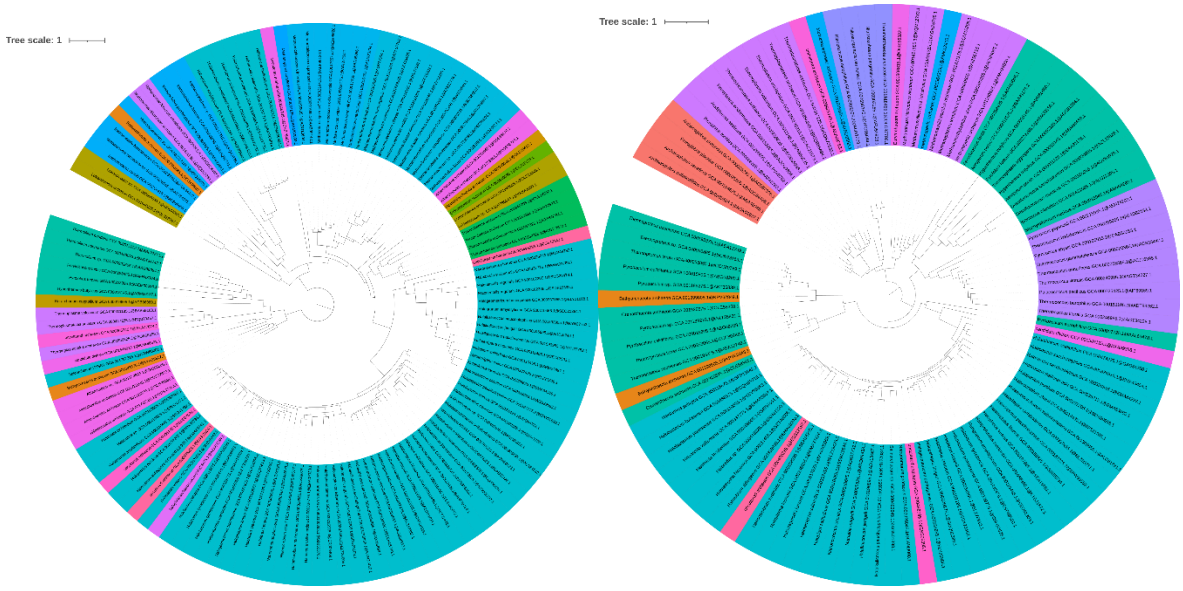


Figure 28. Exemple de gène dupliqué (OG0000527) découpé par le programme root-max-div-taxon avec RAxML selon le modèle PROTCATLGF.

Les données sont fournies **Supp.Mat FigS28**. Le trait rouge indique l'endroit de la découpe, donnant deux nouveaux gènes à partir desquels ont été calculés de nouveaux arbres. L'alignement initial contient 230 séquences, correspondant à 145 taxa. Parmi ces 145 taxa, 79 sont uniques tandis que 86 sont présents au moins deux fois. Après partitionnement, on obtient deux nouveaux arbres, l'un ayant 129 séquences et l'autre 101.

	ROUND 1	ROUND 2	ROUND 3	ROUND 4
Gènes en entrée	1007	417	144	65
Gènes sans paralogue	21 (21)	93	18	6
Gènes non coupés	731 (419)	228	85	59
Gènes coupés	255	96	41	X
Gènes coupés x2	510	192	82	X
> 90 espèces (passant au round n+1)	417	144	65	X

Tableau 5. Découpage des gènes paralogues afin de récupérer de nouveaux orthologues.

Les chiffres entre parenthèses correspondent aux gènes orthologues passant le filtre de 91 espèces et qui seront effectivement récupérés. Leur nombre au round 1 est bien égal à 440 (21 + 419). Les valeurs en surbrillance correspondent aux gènes « néo-orthologues » qui sont récupérés après découpe.

3.3 ÉCHANTILLONNAGE TAXONOMIQUE FINAL

Nous avons enrichi ces 440 + 117 MSA de protéines en rajoutant les 59 espèces qui n'étaient disponibles que sous forme de génomes, afin de faire correspondre leur échantillonnage taxonomique à notre sélection initiale de 364 espèces basée sur les MSA des protéines ribosomiques (cf. **Supp.Mat nb-genomes-rajoute-par-OG.txt**). Les MSA ont ensuite été filtrés avec HMMCleaner et BMGE, puis concaténés en une super-matrice avec SCaFoS.

Le premier jeu de données de 440 gènes totalise 89 135 positions d'acides aminés. Sur nos 61 913 séquences, 2771 présentent de la paralogie. Lors de la concaténation, nous avons utilisé un seuil maximal de divergence de 25 % par paire de séquences pour ne pas éliminer

systématiquement les deux paralogues et garder la séquence la moins divergente. Au-delà de ce seuil, les séquences ont été systématiquement éliminées. Parmi ces paralogues, 2017 sont des in-paralogues récents. Par conséquent, le choix du paralogue dans la concaténation n'a pas de conséquence sur la position phylogénétique de ce gène car, dans tous les cas, sa position phylogénétique est correcte. En revanche, 754 séquences ont été éliminées lors de la concaténation, dû à des séquences trop divergentes (out-paralogues ou xénologues) dont il a été impossible de déterminer laquelle était correcte. Six espèces ont des gènes ayant plus de 5 paralogues (cf. **Supp.Mat esp-paralog-sup-5.txt**). Le deuxième jeu de données de 117 gènes a pour sa part 47 154 positions.

Nous avons alors calculé un arbre des 440 gènes avec RAxML selon le modèle PROTCATLGF (**Figure 29** & **Supp.Mat FigS29**).

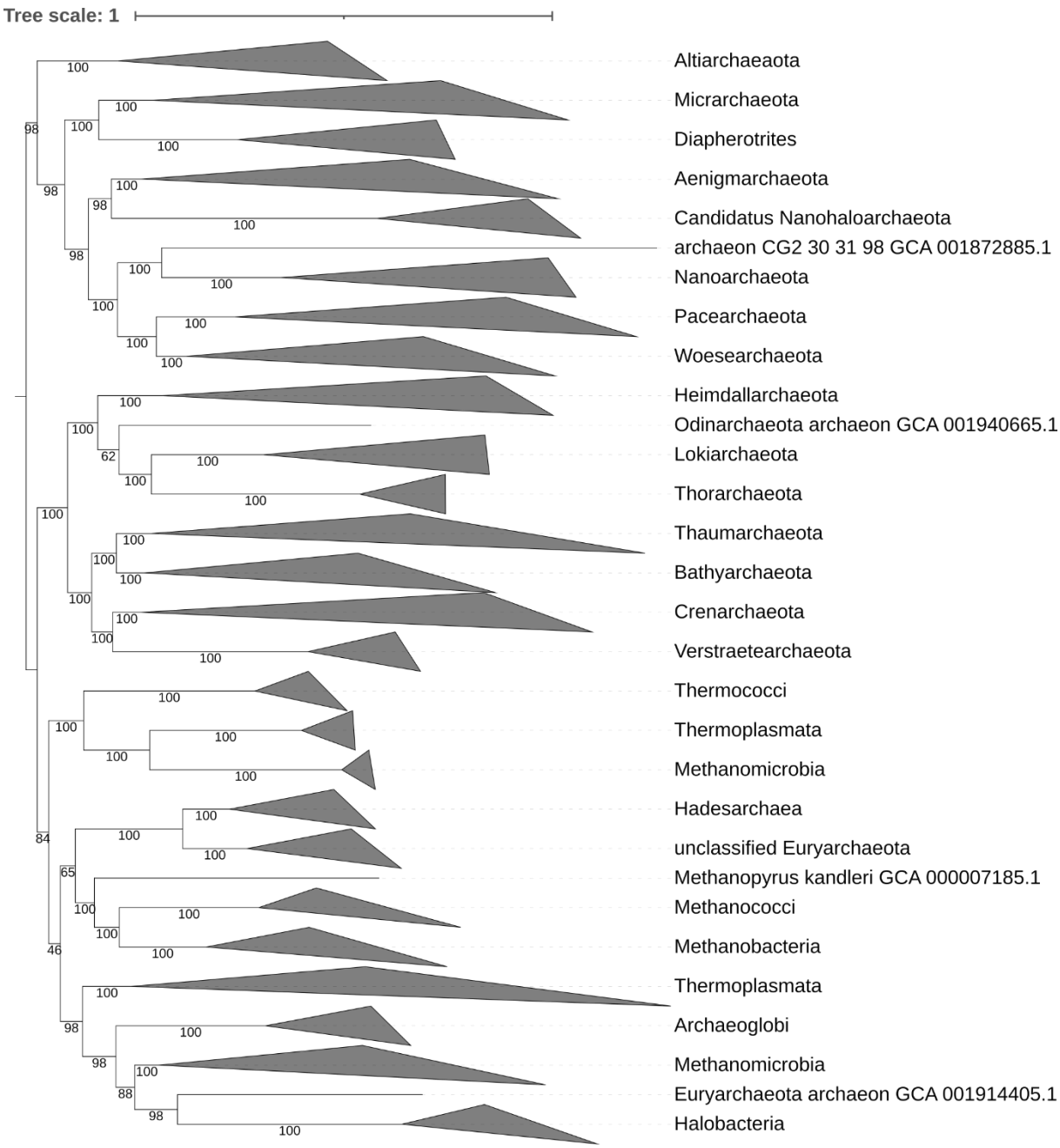


Figure 29. Arbre des 440 gènes calculé avec RAxML avec une recherche par rapid-bootstrap (100 réplicas) selon le modèle PROTCATLGF sur une super-matrices de 89 135 positions et 364 espèces.
Les données sont fournies **Supp.Mat FigS29.** L'utilisation de 440 gènes ne modifie pratiquement pas la topologie de notre arbre d'archées si on la compare à celle de notre arbre ribosomique. La majeure incertitude concerne la position des Hadesarchaea qui passe de groupe frère de (Thermococci + Methanomicrobia + Thermoplasmata = TTM) à groupe frère de (Methanopyrus + Methanococci + Methanobacteria). Toutefois, les faibles valeurs de bootstrap de respectivement 39% et 65% ne permet pas de trancher sur leur réelle affinité.

L'utilisation de 440 gènes modifie légèrement la topologie de l'arbre basé sur les protéines ribosomiques. En effet, dans l'arbre ribosomique, le groupe Hadesarchaea est le groupe frère de (Thermococci + Methanomicrobia + Thermoplasmata = TTM) avec un bootstrap de 39 % mais devient ici groupe frère du groupe (Methanopyrus + Methanococci + Methanobacteria) avec une

valeur de bootstrap de 65 %. Il semblerait que, dans les deux cas, leur position soit incertaine. Nous observons donc des différences entre nos deux jeux de données. On observe également deux groupes correspondant à des archées non déterminées (*Unclassified Euryarchaeota*). L'un de ces groupes comprend les Altiarchaeales, l'autre correspond à des archées ne répondant pas à une nomenclature définie. Toutefois, nous ne sommes pas convaincus des fortes valeurs de bootstrap. En effet, en phylogénomique, une branche doit avoir une valeur de bootstrap maximale pour être considérée comme fiable. Or on peut avoir de fortes valeurs qui pourraient être causées par du signal non phylogénétique (Baurain & Philippe, 2010). En effet, les méthodes de reconstruction d'arbres, en raison de leurs limitations, génèrent des signaux phylogénétiques parasites qui entrent en compétition avec le véritable signal authentique (Di Franco et al., 2022). Par exemple, lorsqu'un biais important favorise un ordre de branchement alternatif (par exemple, une attraction de branche longue entre deux entre deux taxons non apparentés à évolution rapide), le signal non phylogénétique peut prendre le pas sur le signal authentique, entraînant un signal phylogénétique en faveur d'une branche alternative, incorrecte. Ce signal non phylogénétique ne dépend pas seulement des propriétés de la radiation et de l'ensemble de données (par exemple, le taux global d'évolution ou l'échantillonnage taxonomique), mais également de la précision avec laquelle le modèle d'évolution déduit l'histoire de la substitution à chaque position. Dans les cas où tous les loci ne partagent pas la même histoire, le signal non phylogénétique peut être accru par d'autres violations du modèle, par exemple lors de l'utilisation d'un modèle concaténé en présence d'ILS (*Incomplete Lineage Sorting*).

3.4 TEST DE CONGRUENCE PAR COMPARAISON DES LONGUEURS DE BRANCHES (BLC)

Notre dernier contrôle de qualité a été un test de congruence (comparaison des longueurs de branches) basé sur le raisonnement selon lequel les séquences non orthologues (étant soit un contaminant, soit un xénologue, soit un paralogue) présentent généralement de très longues branches lorsqu'elles sont contraintes à être mal placées dans l'arbre d'espèces. Ce protocole ne sera effectué que sur la super-matrice de 440 gènes (cf. **Figure 29**) afin de ne pas réduire davantage la super-matrice issue des gènes dupliqués et mieux évaluer les potentiels bénéfices du traitement des données. Cet arbre a ensuite servi de référence pour imposer la topologie à la phylogénie de nos gènes individuels afin de révéler les espèces présentant de longues branches pour certains gènes.

Nous allons dans un premier temps éliminer des séquences individuelles jugées problématiques au sein d'un gène, puis dans un deuxième temps nous allons éliminer des gènes possédant trop de séquences dont les longueurs de branches ne satisfont pas le coefficient de corrélation de Pearson R^2. Avec cette méthode, nous sommes passés de 440 à 416 gènes.

3.5 CREATION DE CHIMERES

Nous avons ensuite ré-évalué la complétude de nos 416 gènes. Certains organismes n'étaient pas assez représentés dans nos MSA. Mais, parmi ceux-ci, plusieurs étaient étroitement apparentés. Nous avons donc créé des chimères par fusion de gènes in silico de façon que les séquences se complémentent, permettant d'avoir des organismes chimériques plus complets. Nous passons ainsi de 364 à 352 espèces par fusion des espèces proches les moins représentées dans nos 416 gènes.

Parmi nos 416 groupes orthologues à max 352 espèces, nous avons conservé ceux contenant au moins un représentant Asgard, un DPANN, un Euryarchaeota et un TACK. Il nous est ainsi resté 343 gènes.

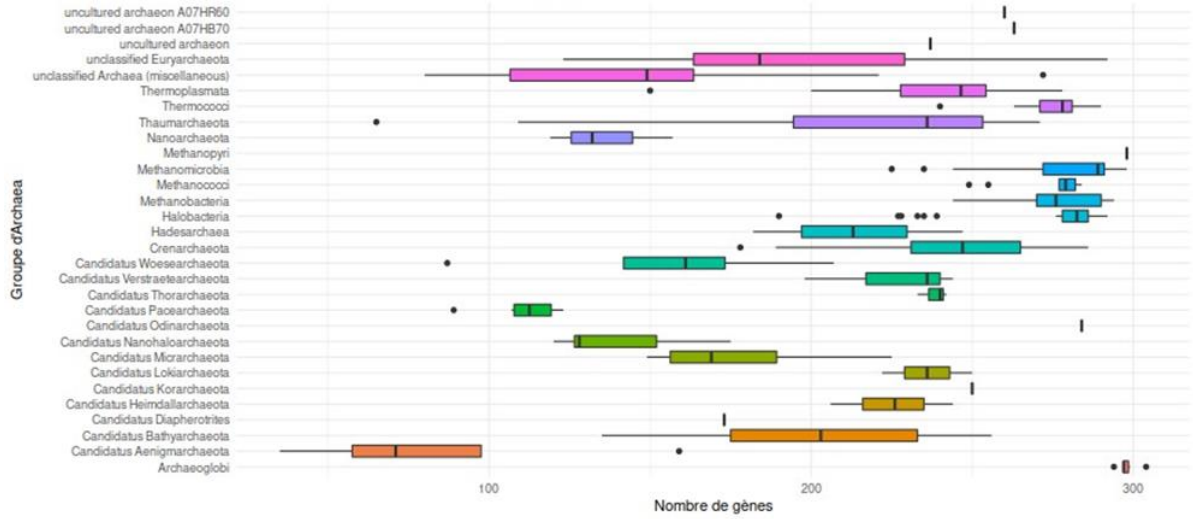


Figure 30. Distribution des groupes d'archées au sein de notre sélection de 343 MSA.
Les données sont fournies **Supp.Mat distr-sp-genes.csv**.

La majorité de nos groupes d'archées sont bien représentés dans nos 343 MSA (**Figure 30** & **Supp.Mat distr-sp-genes.csv**). On notera toutefois une sous-représentation des *Candidatus Aenigmarchaeota* (présents dans systématiquement moins de 100 gènes sauf pour un).

Nous disposons donc de 343 MSA de gènes orthologues de haute qualité et 117 MSA de gènes néo-orthologues de qualité plus brute. L'analyse de nos alignements révèle plusieurs points (**Figure 31, Figure 32, Figure 33, Supp.Mat stat-117-duplicates.csv & stat-343-genes.csv**). D'abord, nos MSA sont plus longs pour les gènes néo-orthologues (le mode est environ de 300 acides aminés après BMGE), contrairement aux vrais gènes orthologues, qui sont globalement deux fois plus courts (mode à environ 150 acides aminés). En revanche, ces derniers tendent à conserver plus d'espèces que pour les gènes néo-orthologues (ils ont été moins filtrés). Les gaps (indels et caractères manquants), quant à eux, ont une distribution similaire entre les gènes orthologues et néo-orthologues. La proportion de gaps tourne aux alentours de 10 % pour l'immense majorité des alignements (1er quartile = 7,7 % et 3ème quartile = 14,8 %).

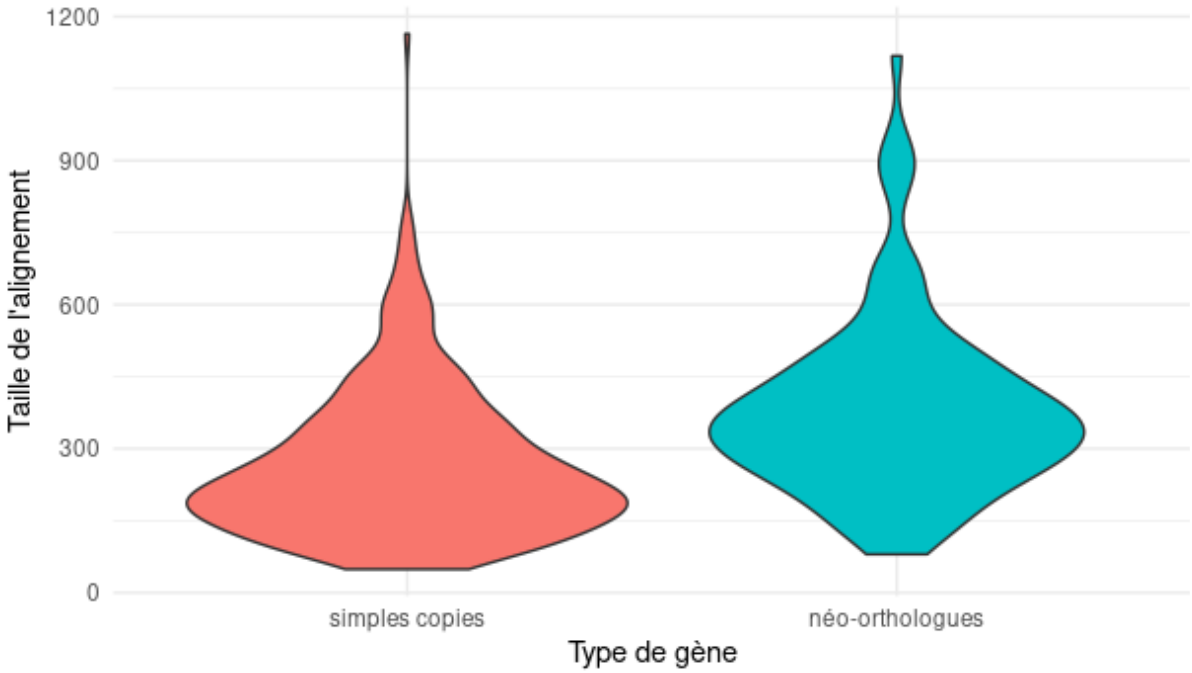


Figure 31. Graphe en violon de la distribution de la taille des alignements en fonction du type de gène.

Les données sont fournies **Supp.Mat stat-117-duplicates.csv & stat-343-genes.csv.** Nos MSA de gènes en simples copies tendent à être beaucoup plus courts (mode d'environ 150 acides aminés) que nos gènes néo-orthologues (mode supérieur à 300 acides aminés).

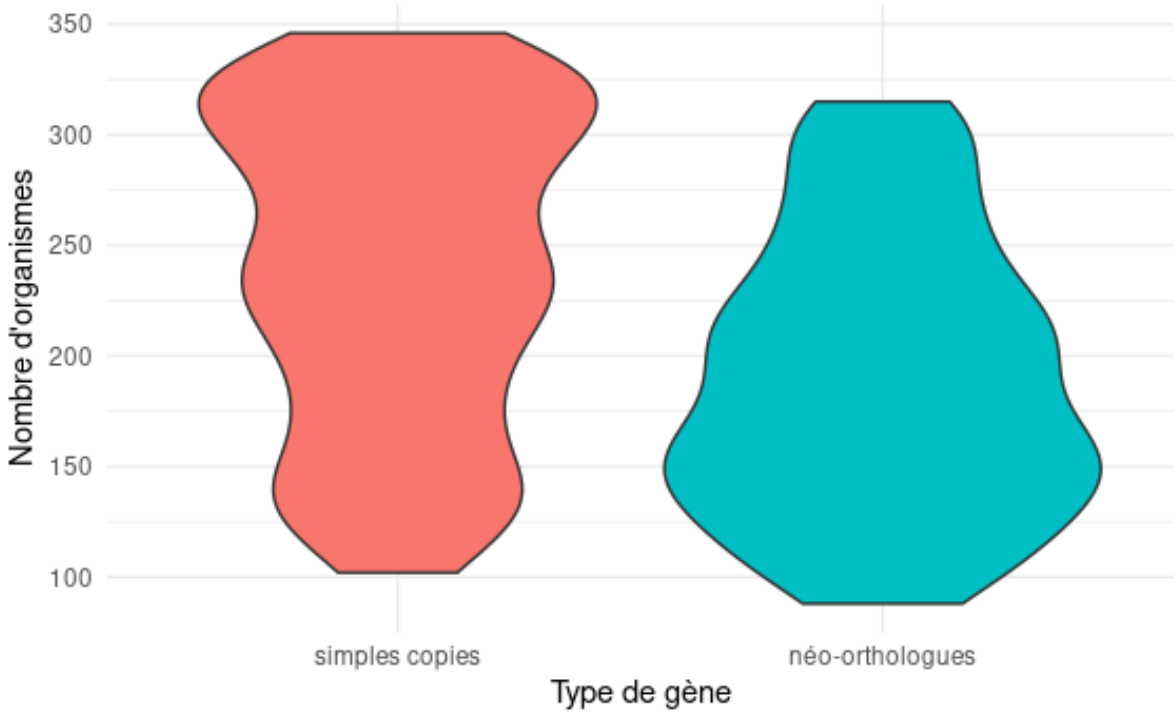


Figure 32. Graphe en violon de la distribution du nombre d'organismes conservés au sein des alignements.

Les données sont fournies **Supp.Mat stat-117-duplicates.csv & stat-343-genes.csv.** Nos MSA de gènes en simples copies derniers tendent à conserver plus d'espèces que pour les gènes néo-orthologues.

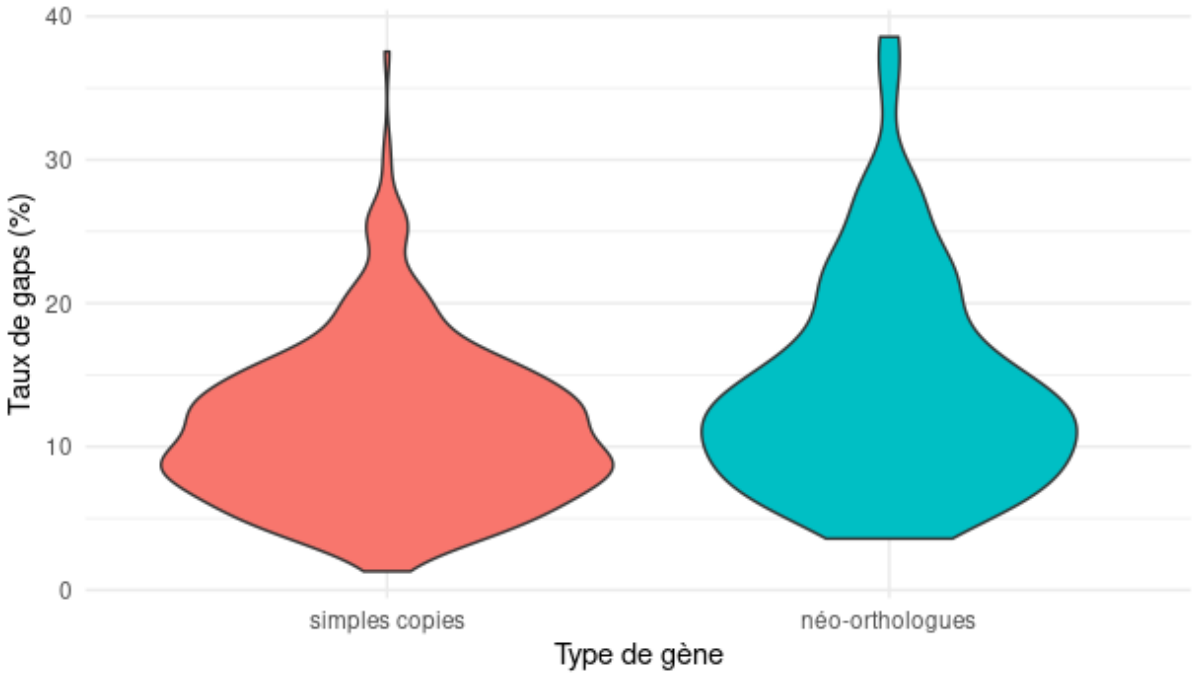


Figure 33. Graphe en violon de la distribution du taux de gaps au sein des alignements.

Les données sont fournies **Supp.Mat stat-117-duplicates.csv & stat-343-genes.csv.** La distribution des gaps est similaire entre nos MSA de gènes en simples copies et néo-orthologues. La proportion de gaps tourne aux alentours de 10 % pour l'immense majorité des alignements (1er quartile = 7,7 % et 3ème quartile = 14,8 %).

4 ANALYSE PRELIMINAIRE ET DISCUSSION PHYLOGENETIQUE DES ARBRES OBTENUS

Maintenant que nous avons créé nos jeux de données, nous allons calculer des arbres en utilisant des modèles sophistiqués. Les arbres ont été calculés avec IQ-TREE, dont nous avons une certaine expérience. Ce logiciel répond bien à nos attentes par sa rapidité (plus rapide que RAxML utilisé jusqu'à présent), son éventail de modèles utilisables plus conséquent et sa fiabilité. En effet, d'après des tests systématiques effectués au sein de notre laboratoire (M. Leleu), ses heuristiques en termes d'exploration des arbres sont presqu'aussi efficaces que celles de RAxML. Notre approche est une approche dite *data-driven*, c'est-à-dire que nous avons traité nos jeux de données sans aucun a priori sur les groupes supposés exister. Nous partons ainsi sans taxonomie prédéfinie et nous redéterminerons nous-mêmes les groupes. Notre but est d'analyser nos jeux de données de la façon la plus neutre et objective possible afin de les comparer ensuite à ce qui est connu dans la littérature. Nous avons calculé un arbre phylogénétique de 352 espèces avec IQ-TREE (Minh et al., 2020) selon le modèle LG4X avec Ultrafast Bootstrap × 1000 pour (1) notre super-matrice ribosomique de 7 819 positions (cf. **Supp.Mat arbre-ribo-352-sp**), (2) notre super-matrice de 94 485 positions représentant nos 343 gènes (**arbre-352-sp-343-genes**) et (3)

notre super-matrice de 47 154 positions représentant nos 117 gènes néo-orthologues issus de duplications.

4.1 ARBRE RIBOSOMIQUE

Afin de vérifier que la création d'organismes chimériques n'affecte pas la phylogénie des organismes, nous avons calculé un nouvel arbre ribosomique (**Figure 34** & **Supp.Mat FigS34**) avec IQ-TREE selon le modèle LG4X avec Ultrafast Bootstrap × 1000, sur une super-matrice de 7 819 positions de notre nouvelle sélection d'espèces (cf. **Supp.Mat FigS34**). Nous obtenons ainsi un même arbre actualisé avec les espèces chimériques. Aucune différence n'est toutefois à noter entre les deux arbres.

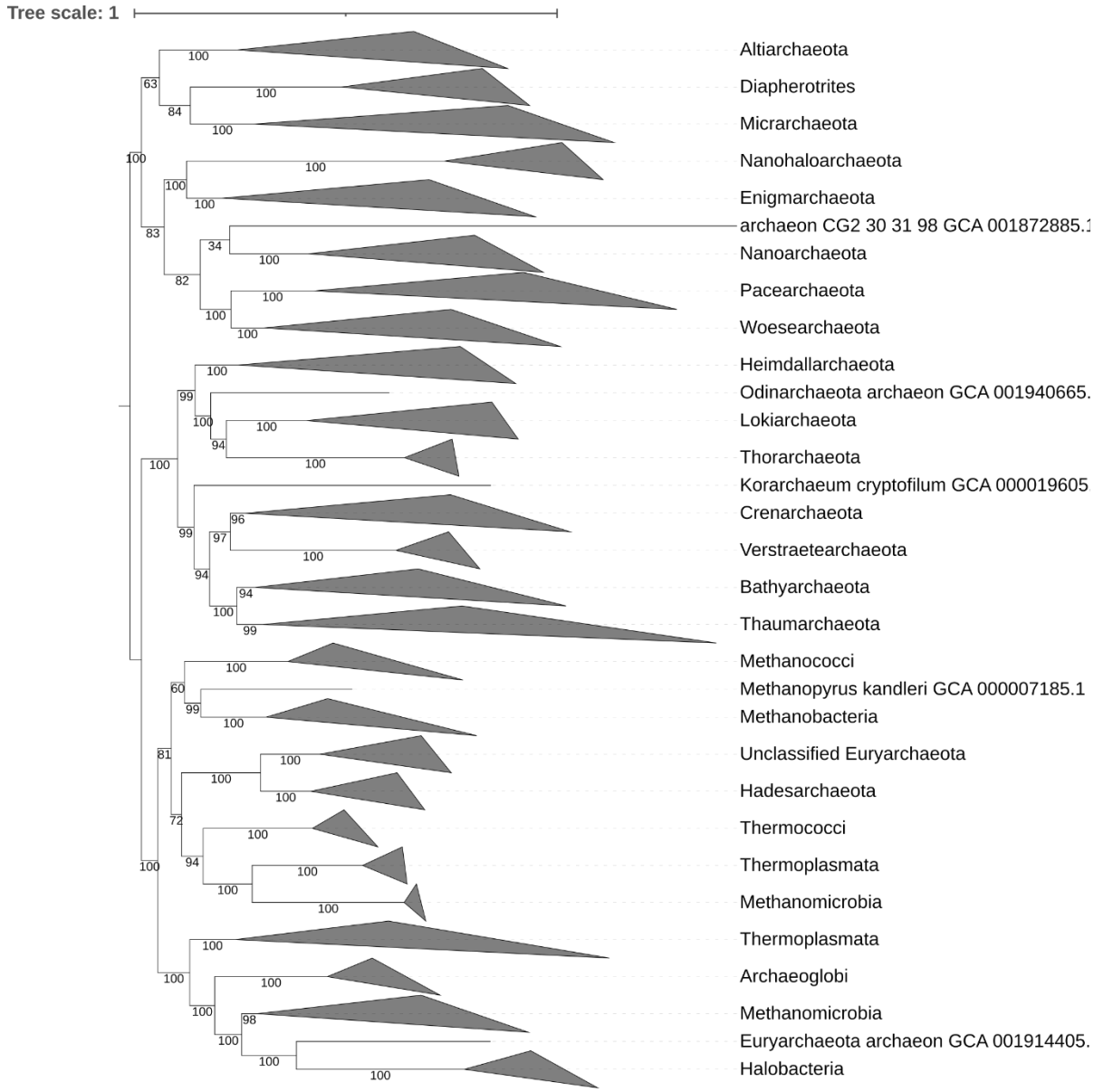


Figure 34. Arbre phylogénétique des protéines ribosomiques de 352 archées (dont 12 chimériques) calculé avec IQ-TREE selon le modèle LG4X et ultrafast bootstrap x 1000 sur une super-matrice de 7 819 positions.

Les données sont fournies **Supp.Mat FigS34.** La réduction de notre échantillonnage taxonomique à 352 archées de notre arbre ribosomique ne modifie pas notre topologie. On notera toutefois une augmentation du support statistique à 72% du groupe Hadesarchaeota avec les (Thermococci + Methanomicrobia + Thermoplasmata = TTM).

4.2 ARBRE 343 GENES SIMPLE COPIE

De cette sélection d'espèces, nous avons également calculé un nouvel arbre (**Figure 35** & **Supp.Mat FigS35**) comprenant l'ensemble de nos 352 espèces et les 343 gènes retenus après le test de congruence en utilisant IQ-TREE selon le modèle LG4X avec Ultrafast Bootstrap X 1000 sur une super-matrice de 94 485 positions (arbre-352-sp-343-genes).

On remarque des différences entre notre arbre ribosomique et notre arbre avec tous les gènes orthologues. Au sein des Euryarchaeota, on retrouve la même différence précédemment évoquée concernant la position des Hadesarchaea. Ceux-ci sont le groupe frère de Methanobacteria + Methanococci + Methanopyrus kandleri lorsqu'on utilise tous les gènes, alors qu'ils sont repoussés légèrement plus haut dans l'arbre lorsqu'on n'emploie que les gènes ribosomiques (**Figure 34**) et se retrouvent groupe frère de Thermococci + Methanomicrobia + Thermoplasmata.

De plus, nous pouvons également observer une différence notable entre notre arbre de 343 gènes et notre arbre préliminaire à 440 gènes. Dans notre arbre de 343 gènes (**Figure 35**), le groupe Thermococci + Thermoplasmata + Methanomicrobia se retrouve groupe frère de Hadesarchaeota + (Methanobacteria + Methanococci + Methanopyrus kandleri) avec une valeur d'ultrafast-bootstrap de 91 %. En revanche, dans l'arbre de 440 gènes (**Figure 29**), le groupe Hadesarchaeota + (Methanobacteria + Methanococci + Methanopyrus kandleri) est groupe frère du groupe (Thermoplasmata + Archaeoglobi + Methanomicrobia + Halobacteria) avec un ultrafast-bootstrap de 46%. Cela démontre un impact de notre sélection de gènes basée sur le test de congruence de comparaison des longues branches.

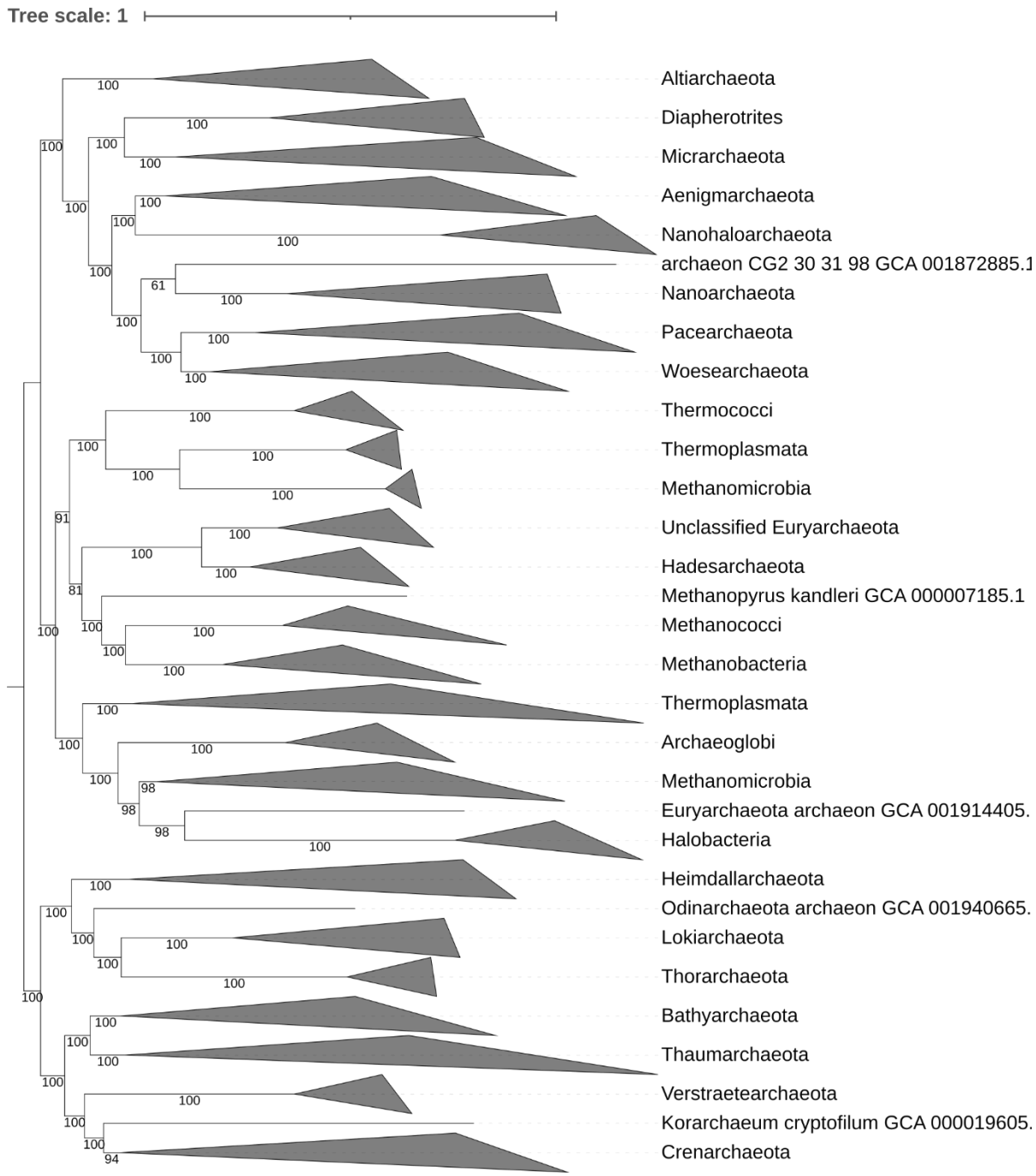


Figure 35. Arbre phylogénétique des 343 gènes de 352 archées calculé avec IQ-TREE selon le modèle LG4X et ultrafast bootstrap x 1000 sur une super-matrice de 94 485 positions.

Les données sont fournies **Supp.Mat FigS35**. La réduction de notre échantillonnage taxonomique à 352 archées de notre arbre à 343 gènes ne modifie pas notre topologie. On notera toutefois une augmentation du support statistique à 81% du groupe Hadesarchaea avec les (Methanopyrus + Methanococci + Methanobacteria).

4.3 ARBRE 117 GENES DUPLIQUES (NEO-ORTHOLOGUES)

Nous avons également calculé un arbre à partir des 117 gènes néo-orthologues issus de duplications (47 154 positions) avec cette même sélection de 352 espèces (les espèces chimérisées ont été ajoutées avec 42 sur les alignements de gènes individuels, puis les gènes ont

été concaténés en une super-matrice) (**Figure 36** & **Supp.Mat FigS36**). Le protocole est similaire aux 343 gènes, à l'exception des filtres qui n'ont pas été appliqués. Dans cet arbre, contrairement aux précédents et en particulier à l'arbre des 440 gènes initiaux dont il est l'équivalent pour les gènes néo-orthologues, les Hadesarchaea sont repoussés plus haut dans l'arbre à la base de tous les Euryarchaeota avec un ultrafast-bootstrap de 99%. De plus, les Thaumarchaeota apparaissent comme polyphylétiques, une partie venant s'insérer entre le groupe initial Crenarchaeota + Candidatus Verstraetearchaeota.

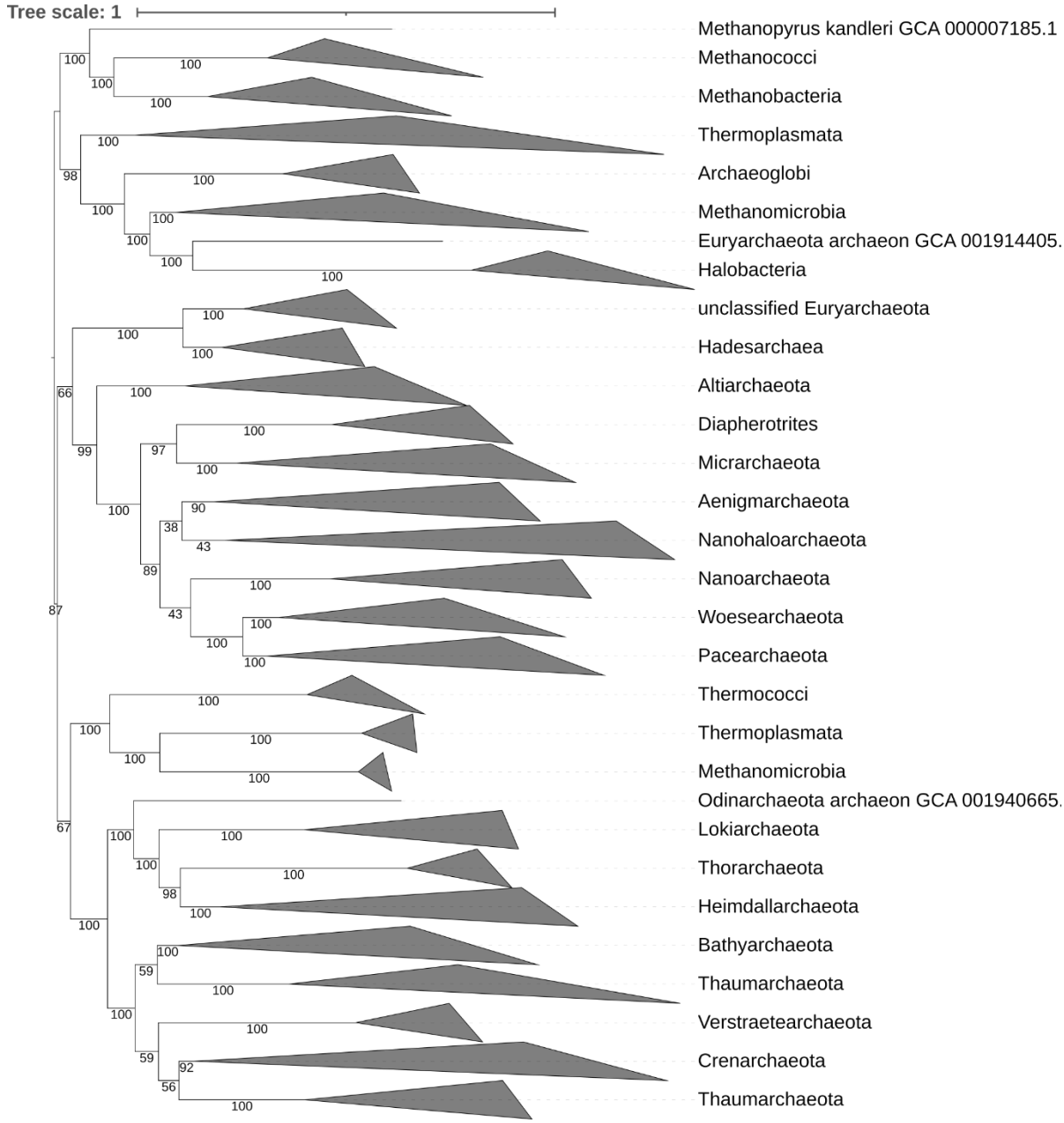


Figure 36. Arbre phylogénétique des 117 gènes néo-orthologues de 352 archées calculé avec IQ-TREE selon le modèle LG4X et ultrafast bootstrap x 1000 sur une super-matrice de 47 154 positions.

Les données sont fournies **Supp.Mat FigS36**. Les Hadesarchaea sont repoussés plus haut dans l'arbre à la base de tous les Euryarchaeota avec un très faut ultrafast-bootstrap de 99%. De plus, les Thaumarchaeota apparaissent comme polyphylétiques, une partie venant s'insérer entre le groupe initial Crenarchaeota + Candidatus Verstraetearchaeota.

Nous venons d'obtenir une première collection d'arbres dans lesquels nous pouvons donc repérer de premières incongruences. Notre but est de tenter de comprendre d'où viennent ces incongruences afin de déterminer, si possible, la bonne topologie. Nous allons donc dès à présent exploiter notre jeu de données en utilisant diverses méthodes. Nous allons faire différentes variations taxonomiques puis leur appliquer la même batterie de tests phylogénétiques : jackknife de gènes, recours à divers modèles évolutifs, le tout en comparant les méthodes de super-matrices et de super-arbres.

5 JACKKNIFE D'ESPECES / VARIATION D'ECHANTILLONNAGE TAXONOMIQUE

Nous voulons évaluer dans un premier temps l'impact de l'échantillonnage taxonomique sur la phylogénie des archées. Nous avons pour cela l'intention de procéder à un jackknife d'espèces (variation d'échantillonnage taxonomique) afin d'effectuer nos analyses (**Box 10**). Notre but est de remplacer des espèces pour détecter d'éventuels artefacts qui seraient dus à des espèces présentant des taux d'évolution différents, pouvant conduire en particulier à des phénomènes d'attractions des longues branches. Pour cela, nous allons utiliser notre arbre de protéines ribosomiques afin de définir des groupes monophylétiques représentant au maximum la diversité des archées et dont la granularité est suffisante pour découvrir les relations phylogénétiques d'ordre supérieur. Ces groupes doivent tous être supportés avec un bootstrap de 100 % dans notre arbre ribosomique. Dès lors, nous obtenons 70 groupes au sein desquels nous avons par la suite fait du jackknife d'espèces. Les espèces correspondant à nos 70 groupes sont données dans le fichier **70-groupes.xlsx**.

Box 10. Tester la solidité d'un arbre par ré-échantillonnage : le bootstrap et le jackknife

Le ré-échantillonnage est une technique statistique dans laquelle une procédure (telle que la construction d'un arbre phylogénétique) est répétée sur une série de jeux de données. Les résultats de l'analyse des jeux de données ré-échantillonnés sont ensuite combinés pour générer des informations récapitulatives sur l'ensemble de données d'origine.

Dans le contexte de la construction d'arbres phylogénétiques, le re-échantillonnage consiste à générer une série d'alignements de séquences en échantillonnant des colonnes à partir de l'alignement de séquences d'origine. Chacun de ces alignements (appelés pseudo-réplicas) est ensuite utilisé pour calculer un arbre phylogénétique individuel. Un arbre consensus peut finalement être construit en combinant les informations de l'ensemble des arbres générés. Les topologies produites peuvent également être triées par les fréquences d'apparition de leurs bipartitions constitutives.

Le principe général est le suivant :

1. Générer k nouveaux alignements en supprimant aléatoirement des gènes ou des colonnes

2. Recalculer un nouvel arbre pour chacun des k alignements
3. Calculer un arbre consensus avec les fréquences de bipartitions pour chacun des nœuds internes de l'arbre.

L'arbre consensus peut être calculé de différentes façons en utilisant :

- le consensus strict : les clades présents dans tous les arbres ;
- la loi de la majorité : l'arbre consensus contient seulement les branches présentes dans au moins la moitié des arbres individuels. Cet arbre consensus résume l'information de tous les arbres issus du jackknife sans arbre de référence en identifiant l'information partagée par tous les arbres et qui soit compatible avec une majorité des arbres. Toutefois, l'arbre obtenu peut encore contenir des multifurcations.
- Un procédé glouton de la loi de la majorité (*extended majority-rule*) qui ajoute ensuite graduellement les branches qui apparaissent dans moins de la moitié des arbres initiaux jusqu'à résoudre complètement l'arbre.

En général, l'arbre consensus n'a pas de longueur de branches. Il a été montré que cette méthode gloutonne donne de meilleurs résultats encore que la loi de la majorité.

La valeur de confiance de chaque nœud d'un arbre consensus (autrement dit la robustesse d'une bipartition) est alors définie comme le nombre de réplicas (le plus souvent en pourcentage) dans lequel ce nœud apparait. Généralement, pour les phylogénies simple gène, les branches avec un support supérieur à 75% peuvent être considérées comme fiables. En phylogénomique, en revanche, un support est fiable à partir de 99-100%.
Deux principales méthodes de ré-échantillonnage sont utilisées : le bootstrap et le jackknife.

Bootstrap
Le bootstrap est une méthode statistique de re-échantillonnage avec remise à partir d'un échantillon initial. Pour appliquer le bootstrap dans le contexte de la construction d'arbres, chaque pseudo-réplica est construit en échantillonnant aléatoirement des colonnes de l'alignement d'origine avec remise jusqu'à ce qu'un alignement de la taille désirée soit obtenu. La procédure de bootstrap commence généralement en créant de nouveaux alignements en sélectionnant aléatoirement des colonnes de l'alignement original et en recréant un arbre indépendant pour chaque nouvel alignement. Un arbre consensus est alors construit afin de résumer les résultats de tous les réplicas d'arbre. Cependant, une attention particulière sera donnée aux branches ayant des valeurs plus faibles.

Jackknife
L'un des problèmes lorsque l'on fait de la phylogénomique en utilisant des méthodes probabilistes et des modèles complexes est la puissance de calcul requise afin d'obtenir un support statistique fiable. Très souvent, pour de gros jeux de données, les temps de calculs sont longs voire impraticables. Pour de gros jeux de données, les calculs traditionnels de bootstrap sont presque irréalisables. Il est alors nécessaire de trouver un moyen de contourner ces problèmes d'ordre logistique, par exemple par des méthodes de sous-échantillonnage. Une méthode de re-échantillonnage de gènes appelée jackknife ("couteau Suisse") permet de réduire les alignements en plusieurs plus petits sous-échantillons de taille raisonnable afin d'évaluer la qualité des phylogénies. Le jackknife est une méthode de ré-échantillonnage numérique basée sur la suppression pour chaque pseudo-répliqua d'une portion des observations originales. Le jackknife

est très similaire au bootstrap : la seule différence repose dans le fait que les matrices pseudo-répliquées se génèrent en éliminant K positions de la matrice originale. Il a été montré qu'un taux de jackknife de 40% devrait idéalement être utilisé (autrement dit on garde 60% des données).

Références

Shi et al.: Using jackknife to assess the quality of gene order phylogenies. BMC Bioinformatics 2010 11:168.

Farris J, Albert V, Kallersjo M, Lipscomb D, Kluge A: Parsimony jackknifing outperforms neighbor-joining. Cladistics 1996, 12:99-124.

Felsenstein J: Confidence limits on phylogenies: An approach using the bootstrap. Evolution 1985, 39:783-791.

Irisarri, I., Baurain, D., Brinkmann, H. *et al.* Phylotranscriptomic consolidation of the jawed vertebrate timetree. *Nat Ecol Evol* **1**, 1370–1378 (2017).

Moret B, Warnow T: Advances in phylogeny reconstruction from gene order and content data. Methods in Enzymology 2005, 395:673-700.

Pattengale N, Alipour M, Bininda-Edmonds O, Moret B, Stamatakis A: How many bootstrap replicates are necessary? Proceedings of the 13th Int'l Conf on Research in Comput Molecular Biol (RECOMB'09) 2009, 184-200.

Belda E, Moya A, Silva F: Genome rearrangement distances and gene order phylogeny in g-Proteobacteria. Mol Biol Evol 2005, 22:1456-1467.

Luo H, Shi J, Arndt W, Tang J, Friedman R: Gene order phylogeny of the genus Prochlorococcus. PLoS ONE 2008, 3:e3837.

Luo H, Sun Z, Arndt W, Shi J, Friedman R, Tang J: Gene order phylogeny and the evolution of Methanogens. PLoS ONE 2009, 4:e6069.

Mueller, L. D., et F. J. Ayala. 1982. Estimation and interpretation of genetic distance in empirical studies. Genetical Research 40:127-137.

Penny, D., et M. D. Hendy. 1985. Testing methods of evolutionary tree constrution. Cladistics 1:266-278.

Penny, D., et M. D. Hendy. 1986. Estimating the reliability of evolutionary trees. Mol Biol Evol 3:403-417.

Mariadassou, M., Bar-Hen, A., & Kishino, H. (2018). Tree Evaluation and Robustness Testing. Reference Module in Life Sciences.

Nous avons procédé à 5 sélections d'espèces au sein de nos 70 groupes ribosomiques, avec des jackknives représentant ± 1/3 de nos 352 espèces, soit 5 × 121 espèces. Sur certains de nos groupes ayant trop peu de représentants, il a été impossible de faire du jackknife (ex. Korarchaeota représenté par une seule espèce, qui dès lors sera présente dans tous nos réplicas). A l'inverse, certains groupes sont sur-représentés (ex. les Halobacteria : 53 génomes) pour exploiter l'ensemble des espèces. Nous avons gardé 5 représentants au maximum par groupe. Dans notre cas, avec 5 jackknives de 5 espèces par groupe, on pourra utiliser au plus 25 espèces du groupe. Si un de nos groupe contient moins de 25 espèces, on réitérera le processus de tirage. Afin de déterminer quelles espèces garder, nous avons sélectionné les espèces selon le critère de complétude des protéomes, c'est-à-dire leur présence dans un plus grand nombre de gènes. Au final, 303 espèces sur 352 sont distribuées dans nos réplicas, les 49 espèces restantes appartenant à des groupes sur-représentés (ex. Halobacteria) et donc ne pouvant passer nos filtres pour être incluses dans nos réplicas.

6 JACKKNIFE DE GENES & CALCUL DES ARBRES

Nous avons soit concaténé nos gènes en super-matrices, soit calculé des arbres individuels pour chaque gène correspondant à nos 5 jackknives d'espèces afin de construire un super-arbre (**Figure 37**).

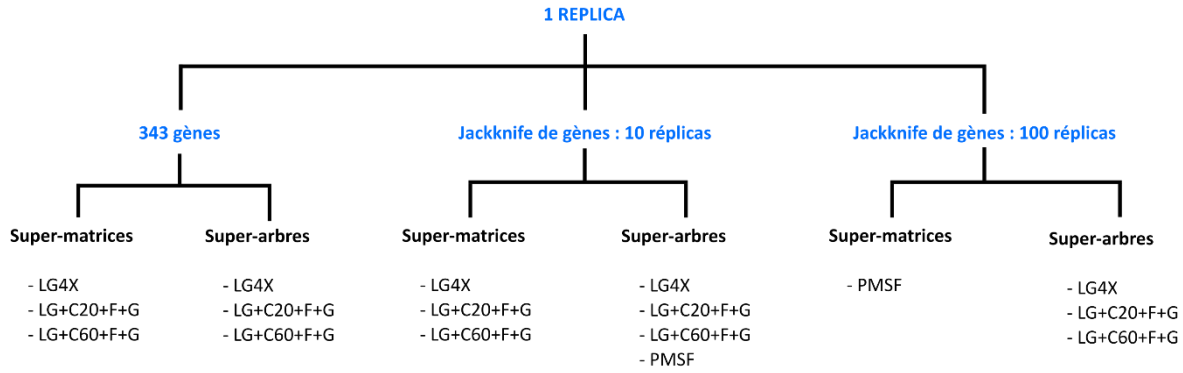


Figure 37. Vue d'ensemble de tous les arbres calculés pour un réplica de jackknife d'espèces. Pour un réplica, nous avons procédé par des approches de super-matrices et de super-arbres avec dans un premier temps l'ensemble des 343 gènes, un jackknife de gènes à 10 réplicas et un jackknife de gènes à 100 réplicas. Les modèles utilisés sont LG4X, LG+ C20+F+G et LG+C60+F+G et PMSF.

Nous avons ensuite cherché à nous mettre dans des conditions difficiles afin de vérifier si l'on retrouve les mêmes résultats qu'avec des phylogénies standards. Pour cela, nous avons couplé notre jackknife d'espèces à un sévère jackknife de gènes. Cette méthode tend à mettre à mal la solidité des arbres en révélant divers problèmes et les incongruences entre différents jeux de données. Nous allons ainsi tester la solidité de nos arbres (**Box 10**).

Pour chaque jeu de données, nous avons calculé des arbres selon les modèles LG4X, LG+ C20+F+G et LG+C60+F+G. Pour chacun des 3 modèles précédents, nous avons appliqué un premier jackknife de 10 réplicas de gènes, soit 5 réplicas d'espèces × 10 réplicas de gènes = 50 super-matrices × 3 modèles = 150 arbres (cf. **Figure 37**). Nous avons également employé en plus la méthode PMSF. Cette méthode est considérée comme celle combattant le mieux les artefacts de reconstruction phylogénétique (**Box 5**). Enfin, nous avons étendu le jackknife de gènes à 100 réplicas afin de calculer des arbres selon la méthode PMSF. Nous avons généré des tailles de super-matrices d'environ 35 000 positions, correspondant au tiers de la taille des super-matrices à 343 gènes et à la taille des super-matrices des 117 gènes néo-orthologues à laquelle on veut les comparer. Pour optimiser nos résultats, nous avons utilisé en arbre guide de référence celui issu de nos précédents IQ-TREE utilisant le modèle LG+C60+G4+F, soit un total de 5 × 100 = 500 super-matrices.

Enfin, nous avons analysé ces mêmes 100 réplicas de jackknife de gènes en utilisant la méthode des super-arbres. Les arbres consensus de chaque réplica (5 × 100 = 500) ont été calculés avec ASTRAL-III à partir des arbres simples gènes selon les différents modèles LG4X, LG+C20+F G et LG+C60+F+G.

Nous avons également comparé les résultats obtenus avec notre jeu de données issu des gènes néo-orthologues. Des super-matrices avec les mêmes réplicas d'espèces ont été calculées

comme précédemment en LG4X, LG+C20+F+G, LG+C60+F+G et PMSF (afin d'optimiser le résultat, l'arbre guide utilisé a été le même que celui utilisé avec les gènes orthologues initiaux).

7 ANALYSE DES DISTANCES DE ROBINSON-FOULDS

Nous avons effectué nos analyses des arbres en utilisant la méthode de Robinson-Foulds (**Box 11**). Cette méthode nous permet d'évaluer les facteurs qui affectent la topologie des arbres.

Box 11. Comparaison topologique : la distance de Robinson-Foulds (RF)

Un moyen de comparer un grand nombre d'arbres phylogénétiques est de mesurer la différence entre tous les arbres pris deux par deux. La distance de Robinson-Foulds (RF) est la mesure la plus largement utilisée à cette fin. La distance topologique de Robinson-Foulds entre deux arbres phylogénétiques est égale au nombre minimal d'opérations élémentaires de fusion et de séparation de nœuds nécessaires pour transformer un arbre en un autre. Autrement dit, elle compte le nombre de bipartitions différentes entre les deux arbres.

Une distance topologique d(T1, T2) entre deux arbres phylogénétiques non enracinés définis sur le même ensemble de n taxons est une mesure de dissemblance entre les topologies respectives de T1 et T2. Si d(T1, T2) = 0, alors les deux arbres sont identiques. Sachant que chaque branche interne d'un arbre phylogénétique induit une bipartition de l'ensemble de ses feuilles, la distance de bipartition dRF mesure le nombre de bipartitions induites par un arbre mais pas par l'autre. Elle est définie comme (A + B) où A est le nombre de partitions de données impliquées par le premier arbre mais pas le second arbre et B est le nombre de partitions de données impliquées par le second arbre mais pas par le premier arbre. Pour un arbre phylogénétique non enraciné dénombrant au plus n - 3 branches internes, la distance dRF est couramment normalisée par 2(n - 3) pour la situer dans l'intervalle [0, 1]. Une valeur de 0 signifie que les arbres sont strictement identiques, tandis qu'une valeur de 1 indique qu'ils sont totalement différents.

Références

J. B. MacQueen (1967): "Some Methods for classification and Analysis of Multivariate Observations, Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability", Berkeley, University of California Press, 1:281-297

Robinson D, Foulds L: Comparison of weighted labeled trees. Combinatorial Mathematics VI 1979, 748:119-126.

D.F. Robinson, L.R. Foulds, Comparison of phylogenetic trees, Mathematical Biosciences, Volume 53, Issues 1–2, 1981, Pages 131-147, ISSN 0025-5564.

7.1 DISTANCES DE ROBINSON-FOULDS DES SUPER-MATRICES

Dans un premier temps, nous avons cherché à évaluer l'impact du modèle sur nos arbres en comparant pour chaque réplica de jackknife les distances RF entre nos différents modèles pris deux à deux. Chaque boxplot (**Figure 38**) correspond donc à la comparaison de 10 paires d'arbres issus de 10 super-matrices. Alors que les modèles C20 et C60 tendent à donner des résultats similaires (nRF = 0.025-0.030) peu importe le réplica d'espèces, le modèle LG4X tend à donner des résultats plus différents (bien que cette différence reste faible : nRF = 0.050-0.090). De plus, on observe pour le réplica 2 une très forte variabilité lorsque l'on compare le modèle LG4X avec les modèles C20 et C60. On remarque aussi que la différence entre le réplica 1 et les autres réplicas est beaucoup plus marquée lorsque l'on compare le modèle LG4X avec les modèles C20 et C60. En

revanche, lorsque l'on compare les modèles C20 et C60 entre eux, cette variabilité est beaucoup plus faible, à l'exception du réplica 4, où elle est très importante, traduisant ainsi un nombre plus important de topologies retrouvées.

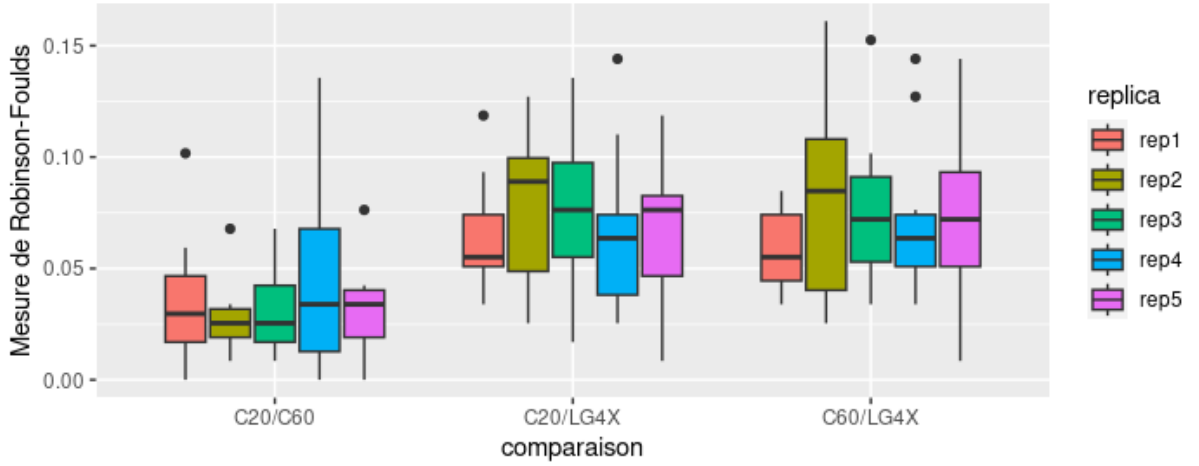


Figure 38. Comparaison entre modèles des distances topologiques de Robinson-Foulds des super-matrices pour chaque réplica.
Les données sont fournies **Supp.Mat Robinson-Foulds**. Alors que les modèles C20 et C60 tendent à donner des résultats similaires (nRF = 0.025-0.030) peu importe le réplica d'espèces, le modèle LG4X tend à donner des résultats plus différents (bien que cette différence reste faible : nRF = 0.050-0.090).

Nous avons ensuite évalué l'impact de notre sélection de gènes sur nos arbres (**Figure 39**). Pour un réplica et un modèle donnés, nous disposons de 10 réplicas de gènes, soit (10^2-10) / 2 = 45 comparaisons par boxplot. Si l'on classe classes les distances RF de nos 3 modèles d'après leur rang (1, 2, 3 ou 1, 2 en cas d'ex aequo) de leur médiane dans les 5 réplicas, on remarque qu'exception faite du réplica 3, le modèle LG4X tend à générer des distances RF plus élevées (de 0.15 à 0.20), suggérant qu'il est plus sensible à l'échantillonnage de gènes. Le modèle C20 semble en revanche être le plus stable.

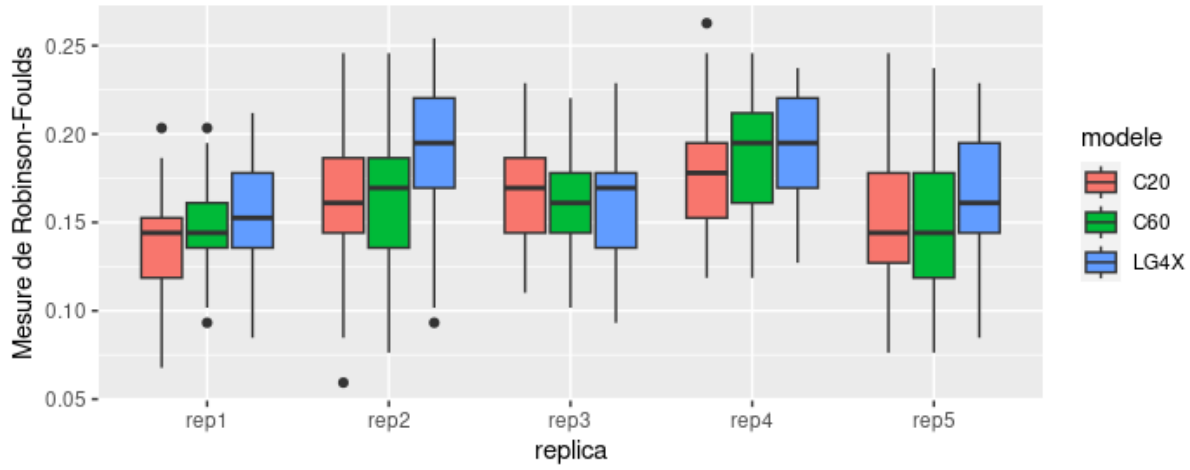


Figure 39. Distances topologiques de Robinson-Foulds des super-matrices selon les modèles LG4X, C20 et C60.

Les données sont fournies **Supp.Mat Robinson-Foulds**. Le modèle LG4X tend à générer des distances RF plus élevées (de 0.15 à 0.20), suggérant qu'il est plus sensible à l'échantillonnage de gènes.

Nous avons évalué l'impact de notre sélection de gènes sur les 5 × 100 arbres calculés selon la méthode PMSF (**Figure 40**). Chaque boxplot contient ainsi (100^2-100) / 2 = 4950 points. Le PMSF est encore moins sensible au réplica de gènes (on est à 0.1 de distance). Deux interprétations peuvent être envisagées : soit le fait de donner une topologie avec un arbre guide induit le résultat et donc diminue la variabilité des arbres obtenus, soit la méthode PMSF affine le résultat en combattant ce qui pourrait être considéré comme des artefacts (Wang et al., 2018).

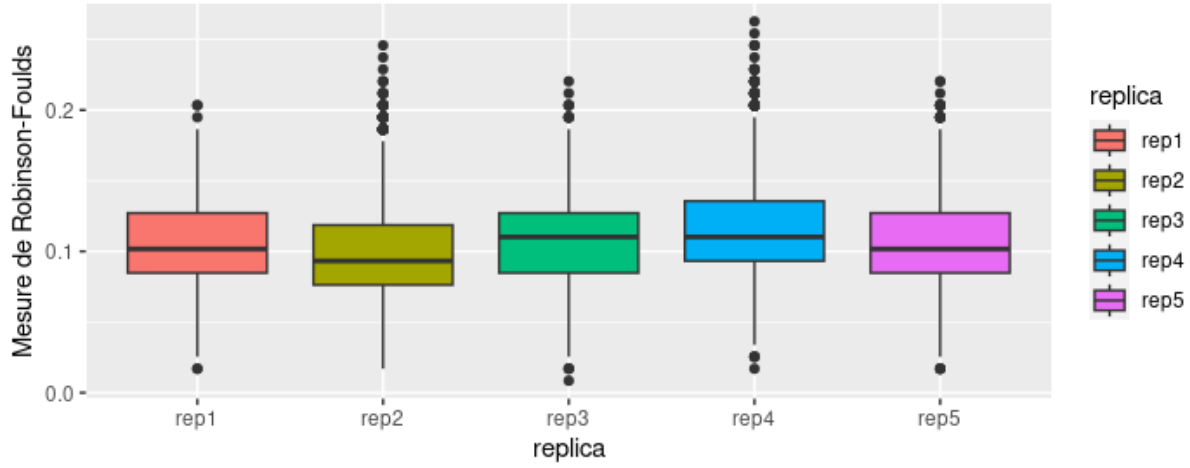


Figure 40. Distance topologique de Robinson-Foulds des super-matrices selon la méthode PMSF.

Les données sont fournies **Supp.Mat Robinson-Foulds**. Le PMSF est très peu sensible au réplica de gènes (0.1 de distance). Cela peut traduire que soit fournir un arbre guide induit le résultat, diminuant la variabilité des arbres obtenus, soit la méthode PMSF se montre très efficace pour combattre les artefacts et trouver le bon résultat.

7.2 DISTANCES DE ROBINSON-FOULDS DES SUPER-ARBRES

Pour comparer les modèles, nous disposons ici de 100 réplicas de gènes par boxplot (contre 10 pour les super-matrices) (**Figure 41**). Les résultats sont sensiblement les mêmes que

précédemment avec les super-matrices. Alors que les modèles C20 et C60 tendent à donner des résultats similaires (0.075-0.1) peu importe le réplica d'espèces, le modèle LG4X tend à donner des résultats plus différents (bien que cette différence reste faible : 0.10 à 0.14). Ces valeurs sont également légèrement plus élevées qu'avec les super-matrices. La variabilité entre modèles sont également moins marquées.

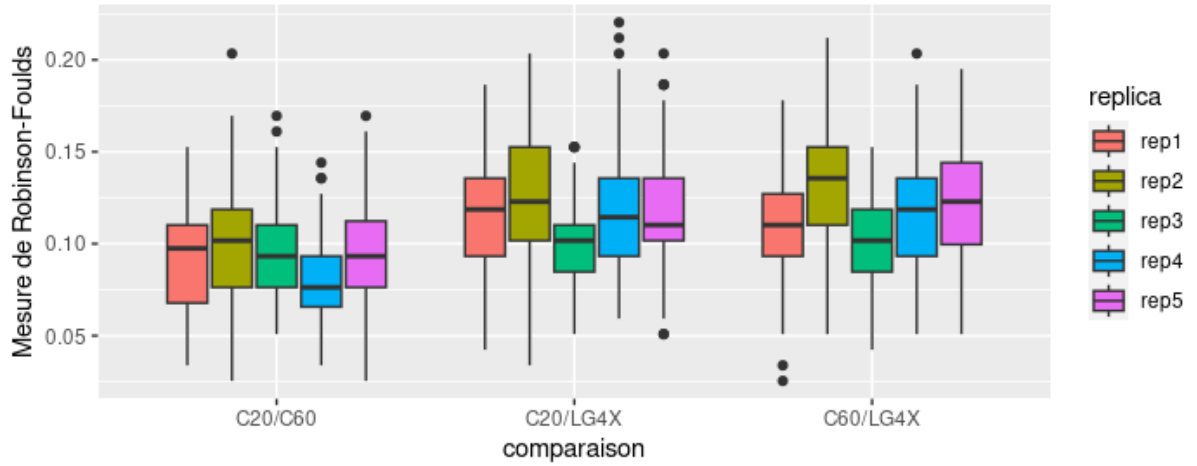


Figure 41. Comparaison entre modèles des distances topologiques de Robinsons-Foulds des super-arbres pour chaque réplica.

Les données sont fournies **Supp.Mat Robinson-Foulds**. Alors que les modèles C20 et C60 tendent à donner des résultats similaires (nRF = 0.075-0.1) peu importe le réplica d'espèces, le modèle LG4X tend à donner des résultats plus différents (bien que cette différence reste faible

En ce qui concerne l'influence de l'échantillonnage en gènes, nous disposons pour chaque réplica de 100 arbres (soit 4950 points par boxplot) (**Figure 42**). On observe une différence des moyennes plus importante entre le modèle LG4X et les autres modèles qu'entre les modèles C20 et C60. De plus, les réplicas 2 et 4 semblent plus instables selon le modèle employé avec des différences de 0.03 entre les médianes les plus hautes et les plus basses pour chaque réplica alors que les moyennes sont les mêmes pour chaque modèle dans les réplicas 1, 3 et 5.

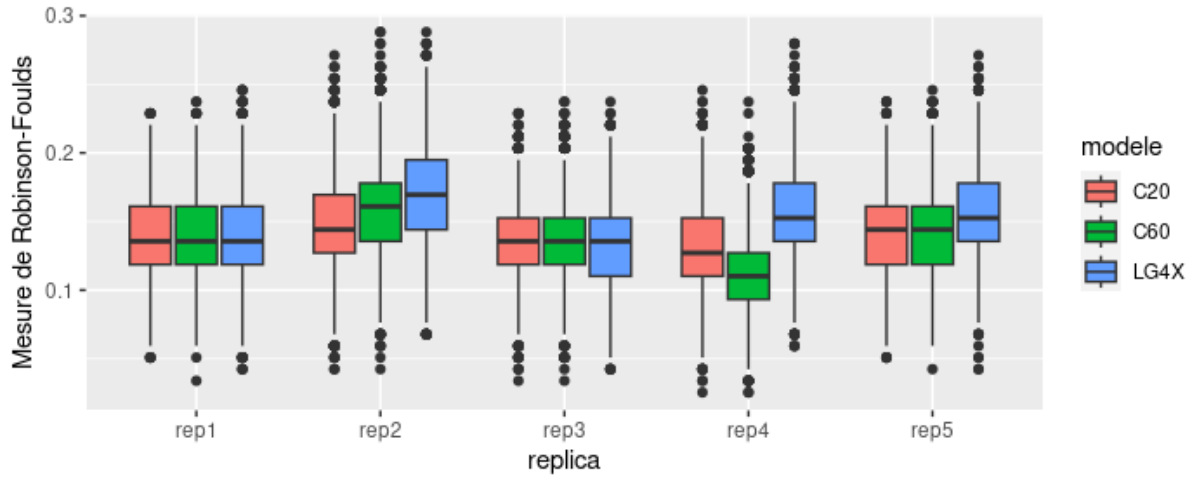


Figure 42. Distances topologiques de Robinson-Foulds des super-arbres selon les modèles LG4X, C20 et C60.

Les données sont fournies **Supp.Mat Robinson-Foulds**. Les moyennes diffèrent plus entre le modèle LG4X et les autres modèles qu'entre les modèles C20 et C60. Les réplicas 2 et 4 semblent plus instables selon le modèle employé.

Le bilan des analyses RF montre que le modèle LG4X est plus sensible aux variations d'échantillonnage de gènes que les autres modèles. De plus, les super-arbres sont moins stables que les super-matrices. Enfin, les réplicas d'espèces ne se comportent pas tous de la même façon, d'où l'intérêt d'en avoir fait 5 différents. Le choix du modèle semble donc le paramètre le plus important à considérer. Toutefois, les distances RF ne nous permettent pas d'évaluer qualitativement ces différences. Nous ne savons pas pour l'instant si ces différences proviennent de groupes terminaux mineurs ou au contraire de grands groupes qui se déplacent dans l'arbre ou d'une combinaison de ces deux phénomènes.

8 RECHERCHE DES BIPARTITIONS MAJORITAIRES

Nous avons jusqu'à présent mesuré les différences entre nos arbres sans toutefois regarder d'où elles provenaient. Pour ce faire, il faudrait en principe construire tous les arbres possibles. Malheureusement, la détection, parmi tous les arbres possibles, de l'arbre de plus haute vraisemblance est un problème dit « NP-dur », c'est-à-dire qu'aucun algorithme connu ne peut résoudre ce problème en un temps raisonnable. En effet, le nombre de topologies possibles pour un arbre phylogénétique non raciné de t espèces augmente de façon exponentielle lorsque t augmente, et peut être calculé grâce à la formule suivante :

$$n(t) = \prod_{i=3}^{t}(2i-5) = \frac{(2t-5)!}{(t-3)!\,2^{t-3}}$$

Ce qu'il faudrait encore multiplier par toutes les longueurs de branches possibles pour obtenir le nombre total d'arbres possibles. Par conséquent, il est impossible de tester chacun des arbres possibles.

Comme il nous est impossible d'envisager toutes les topologies possibles, nous avons recherché tous les nœuds rencontrés dans nos multiples arbres afin d'établir une liste de tous les clans (= clades non racinés) possibles. Ce travail a nécessité l'analyse individuelle de tous les arbres afin de relever tous les clans possibles (cf. **Supp.Mat archaea.clan**), qui ont été ensuite

comparés entre les arbres, les 5 réplicas n'ayant pas les mêmes espèces. A partir de cette collection de clans, nous avons cherché à identifier ceux systématiquement retrouvés et ceux plus instables. Pour ce faire, nous les avons exprimés en termes de regroupements successifs de nos 70 groupes ribosomiques initiaux, jusqu'à se rendre compte que des arbres semblent présenter des hypothèses alternatives (= combinaisons récurrentes de clans) (cf. **Supp.Mat 70groupe.xlsx**). Afin de comprendre d'où proviennent ces différences, nous avons focalisé notre attention sur les nœuds des arbres qui semblent les plus instables. Nous avons regardé les incongruences à partir d'une échelle de classification au minimum plus grande que nos 70 groupes, en délaissant volontairement les phylogénies à l'intérieur de ceux-ci. La liste des nœuds instables est donnée dans la table **Supp.Mat noeuds-à-problèmes.xlsx**.

Ces instabilités peuvent résulter de différentes causes. On distingue 5 cas :

- Impact du réplica d'espèces : est-ce que le choix d'espèces influe sur la topologie observée, peu importe la méthode employée ? Dans ce cas, les résultats obtenus sont sous l'influence de l'échantillonnage taxonomique et on peut soupçonner divers problèmes intrinsèques aux génomes utilisés (transfert de gènes, contamination etc.) ;
- Impact du modèle employé : tous les réplicas sont d'accord mais racontent des histoires différentes selon le modèle utilisé. Dans ce cas, il est fort possible que certains modèles soient sensibles à des artefacts de reconstruction phylogénétique impactant les (courtes) branches matérialisant les relations entre groupes de niveau plus élevé.
- Impact de la méthode employée : est-ce que les super-arbres donnent des résultats différents des super-matrices ? Cela pourrait pointer vers un manque de signal des arbres simples gènes à la base des super-arbres (erreur stochastique).
- Impact de certains gènes (ex : HGT au niveau d'un groupe) ?
- Impact combiné du réplica d'espèces et du modèle ou de la méthode employé(e).

8.1 REGROUPEMENT DES GROUPES RIBOSOMIQUES EN GROUPES DE NIVEAU SUPERIEUR COMMUNEMENT ACCEPTES PAR LA LITTERATURE

Nous avons jusqu'à présent procédé à une approche dite *data-driven*, c'est-à-dire que nous avons traité nos jeux de données sans aucun a priori. Nous allons maintenant comparer nos résultats par rapport à ce qui est connu de la littérature. Conformément à la littérature, le regroupement successif de nos 70 groupes ribosomiques initiaux nous permet d'attester de la monophylie de nombreux groupes déjà décrits. L'analyse de nos 70 groupes ribosomiques initiaux montre qu'ils restent systématiquement monophylétiques (bootstrap = 100) peu importe le réplica, le modèle (même dans des conditions difficiles de jackknife X 100 PMSF) ou la méthode employée. Nous pouvons alors définir nos hypothèses en fonction de nos 70 groupes ribosomiques, ce qui nous permet de réduire à 27 le nombre de groupes monophylétiques (souvent au-delà de nos groupes ribosomiques) systématiquement valides. Ces groupes sont donnés dans le **Tableau 6** et nommés d'après la littérature. Ce sont eux qui nous serviront à partir de maintenant d'OTU dans nos arbres phylogénétiques.

Groupes	Correspondance par rapport à nos 70 groupes
Groupe TACK	
Aigarchaeota	Aigarchaeota
Thaumarchaeota	Thaumarchaeota
Bathyarchaeota	Candidatus_Bathyarchaeota_1 Candidatus_Bathyarchaeota_2

	Candidatus_Bathyarchaeota_3 Candidatus_Bathyarchaeota_4 Candidatus_Bathyarchaeota_5 Candidatus_Bathyarchaeota_6 Candidatus_Bathyarchaeota_7 Candidatus_Bathyarchaeota_8
Korarchaeota	Candidatus_Korarchaeota
Verstratearchaeota	Candidatus_Verstraetearchaeota
SCGC	Crenarchaeota_SCGC
Crenarchaeota	
Thermoproteales	Crenarchaeota_Thermofilaceae Crenarchaeota_Thermoproteaceae
Desulfurococcales	Crenarchaeota_Desulfurococcales_Aeropyrum Crenarchaeota_Desulfurococcales_Desulfurococcaceae Crenarchaeota_Desulfurococcales_Ignicoccus Crenarchaeota_Desulfurococcales_Ignisphaera Crenarchaeota_Desulfurococcales_Pyrodictiaceae Crenarchaeota_Fervidicocccales
Sulfolobales	Crenarchaeota_Sulfolobales
Asgard	
Thorarchaeota	Candidatus_Thorarchaeota
Lokiarchaeota	Candidatus_Lokiarchaeota
Heimdallarchaeota	Candidatus_Heimdallarchaeota
Odinarchaeota	Candidatus_Odinarchaeota
Euryarchaeota	
Hadesarchaeota	Hadesarchaeota Unclassified Euryarchaeota
Theionarchaea	Theionarchaea
Thermococci	Thermococci
Altiarchaeales	Altiarchaeales
Methanobacteria	Methanobacteria
Methanopyri	Methanopyri
Thermoplasmatales	Thermoplasmata_Thermoplasmatales_1 Thermoplasmata_Thermoplasmatales_2 Thermoplasmata_Thermoplasmatales_3
Thermoplasmata_unknown_1	Thermoplasmata_unknown_1
Thermoplasmata_unknown_2	Thermoplasmata_unknown_2
Thermoplasmata_unknown_3	Thermoplasmata_unknown_3
Stenosarchaea	Halobacteria Methanomicrobia_Arc Methanomicrobia_Methanocellales Methanomicrobia_Methanomicrobiales Methanomicrobia_Methanoperedenaceae Methanomicrobia_Methanosaetaceae Methanomicrobia_Methanosarcinaceae Methanomicrobia_Methanosarcina_genus Methanomicrobia_Methanosarcinales Methanomicrobia_Syntrophoarchaeum
Archaeoglobi	Archaeoglobi

Natronarchaea	Methanonatronarchaeia
DPANN	
DPANN	Candidatus_Diapherotrites Candidatus_Micrarchaeota Candidatus_Aenigmarchaeota Candidatus_Nanohaloarchaeota Nanoarchaeota Candidatus_Pacearchaeota Pseudo_Woesearchaeota True_Woesearchaeota unclassified_Archaea_(miscellaneous)

Tableau 6. Regroupements de nos 70 groupes ribosomiques qui nous serviront d'OTU par la suite.

8.2 GROUPES PEU RETROUVES

Nous avons cherché au sein de nos bipartitions les nœuds présentant des instabilités. Nous avons considéré comme instables les nœuds présentant une valeur de BP inférieure à 75% (Shi et al., 2010) dans au moins une combinaison de réplica et de modèle (LG4X, C20 ou C60), que ce soit en méthode super-matrice ou super-arbre (cf. **Supp.Mat bilan-supermatrics.xlsx**, **bilan-superarbres.xlsx** & **bilan-duplicates.xlsx**). Les nœuds problématiques que nous avons relevés sont donnés dans le **Supp.Mat noeuds-à-problèmes.xlsx**. Nous allons regrouper ces problèmes selon qu'on les considère comme mineurs ou majeurs du point de vue de la problématique de cette thèse, c'est-à-dire dans le contexte d'une phylogénie globale des Archaea en relation avec l'origine des eucaryotes.

8.2.1 PROBLEMES MINEURS

Parmi nos 8 groupes initiaux de Bathyarchaeota, de nombreuses topologies alternatives ont été retrouvées. La relation entre les Bathyarchaeota et les Thaumarchaeota est bien supportée et a déjà été décrite (Adam et al., 2017; Brochier-Armanet et al., 2011). La phylogénie interne aux Bathyarchaeota a elle aussi déjà été traitée en profondeur dans la littérature et ne sera pas l'objet de notre étude ici.

Concernant les Euryarchaeota, la monophylie de nos deux groupes Thermoplasmata_unknown_2 et Thermoplasmata_unknown_3 est parfaitement supportée avec les super-matrices. En revanche, elle est moins systématique avec les super-arbres, où les valeurs d'ultrafast-bootstrap varient entre 6 et 63 %. Le groupe Thermoplasmata_unknown_1 vient alors casser cette monophylie au lieu d'être à leur racine.

Au sein des Stenosarchaea (cf. **Tableau 6**), la monophylie Methanomicrobia_Methanosarcinales + Methanomicrobia_Methanosaetaceae n'est jamais retrouvée lors de l'utilisation des super-arbres, alors qu'elle est toujours retrouvée avec l'emploi de super-matrices. Ce fait est dû au Methanonatronarchaeia Euryarchaeota archaeon GCA001914405.1, qui tend à se retrouver attiré par les Halobacteria dans nos réplicas 2, 4 et 5 (ultrafast-bootstrap < 72 %), empêchant la monophylie des Stenosarchaea. Nous nommerons par la suite SAN l'ensemble monophylétique Stenosarchaea + Archaeoglobi + Natronarchaea.

Le groupe Methanobacteria, toujours retrouvé en super-matrices, est moins supporté en super-arbres (support statistique compris entre 30 et 84 %, sans que le modèle ne semble exercer une influence sur cette valeur).

Concernant le groupe TACK, selon les réplicas, on trouve soit une Fervidicoccales soit une Desulfurococcales Pyrodictiaceae comme groupe frère des Acidilobales et des Aeropyrum.

La monophylie Thorarchaeota-Lokiarchaeota est parfaitement supportée en super-matrices, avec des valeurs d'ultrafast-bootstrap maximales (bien que légèrement mis à l'épreuve avec un jackknife de gènes x100, avec des valeurs comprises entre 76 et 83 % selon le réplica). En revanche, avec les super-arbres, les supports sont plus faibles et varient entre 42 et 72 %.

En PMSF, Heimdallarchaeota archaeon GCA 001940645.1 est problématique avec la méthode super-matrice. La monophylie des Heimdallarchaeota est mise à mal à cause de cette espèce qui vient s'insérer à la racine de tout le groupe Asgard. Ce problème est également très légèrement apparu avec la méthode des super-arbres.

8.2.2 PROBLEMES MAJEURS

La monophylie Euryarchaeota + Crenarchaeota + Bathyarchaeota est parfaitement supportée avec les super-matrices, en revanche elle peut paraître moins systématique avec l'emploi des super-arbres (support statistique compris entre 58 et 94 %, sans influence ni du modèle ni du réplica).

Chez les Euryarchaeota, la monophylie Thermoplasmata + SAN + Methanobacteria est bien supportée en super-matrices pour les réplicas 1 à 4 (ultrafast-bootstrap compris entre 62 et 100 %), sauf pour le réplica 5 à cause des Altiarchaeota qui viennent s'insérer comme groupe frère des Methanobacteria au lieu d'être à la base du groupe DPANN (ultrafast-bootstrap compris entre 1et 46 %). En revanche, en super-arbres, le support semble hésiter entre les deux positions alternatives des Altiarchaeales (ultrafast-bootstrap compris entre 31 et 76 % pour une monophylie du groupe). Il en va de même pour le groupe Hadesarchaea, qui oscille entre groupe frère des Euryarchaeota + DPANN et groupe frère des Theionarchaea + Thermococci (**Figure 44**). L'ultrafast-bootstrap est compris entre 34 et 75 % pour les super-arbres, sans influence du modèle ou du réplica. En revanche, concernant les super-matrices, les ultrafast-bootstraps varient entre 62 et 100 % pour les réplicas 1, 2 et 3. Concernant le réplica 4, le support est très faible, variant de 0 à 24 %. Les valeurs du réplica 5 varient entre 35 et 95 %, et entre 18 et 75 % pour le modèle PMSF. On notera que la valeur de bootstrap lorsqu'on inclut nos 354 espèces avec un modèle LG4X (cf. arbres préliminaires pré-jackknife) tombe à 19 % pour une hypothèse Hadesarchaea + Theionarchaea + Thermococci. La position des Hadesarchaea est donc source d'incertitude avec notre jeu de données.

Les Crenarchaeota forment un groupe parfaitement soutenu. En revanche, en son sein, la position de *Korarchaeota cryptofylum* varie selon le modèle et la méthode utilisés. En super-matrices, on tend à trouver préférentiellement les Korarchaeota comme groupe frère des Crenarchaeota. Pour le réplica 1, on observe une baisse significative des ultrafast-bootstrap lors de l'utilisation des modèles C60 et PMSF (respectivement 34 et 5 %) au profit du groupe SCGC comme groupe frère des Crenarchaeota (66 et 95 %). Par ailleurs, la super-matrice contenant les 354 espèces possède un faible support statistique (6 %) pour cette hypothèse. De plus, lors du JK de gènes × 100 avec nos 5 réplicas, cette hypothèse se retrouve presque à égalité avec l'hypothèse alternative d'une position des Korarchaeota à la base des Crenarchaeota (ultrafast-bootstrap entre 30 et 69 %). C'est cette dernière hypothèse qui est systématiquement retrouvée avec la méthode des super-arbres.

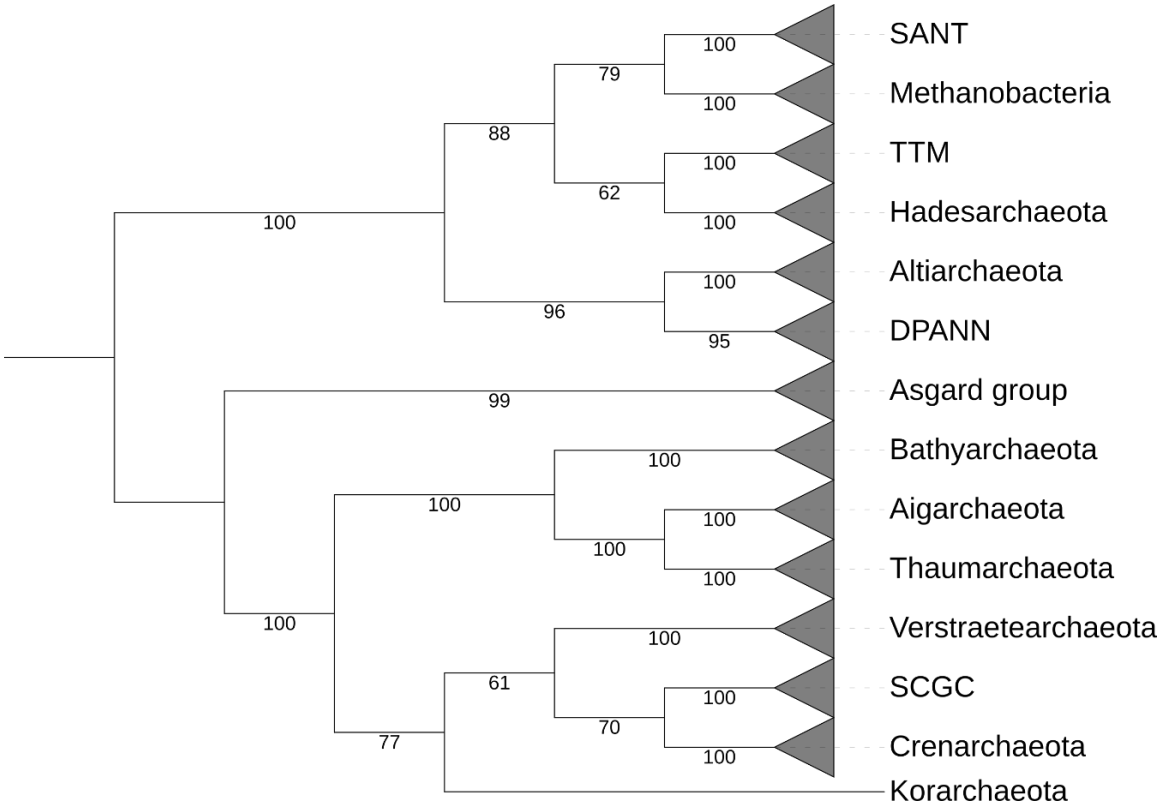
8.3 TOPOLOGIES POSSIBLES

Dans le cadre de notre étude, on s'intéresse à une sélection de problèmes d'un assez haut niveau. Nous venons d'explorer les bipartitions. Certaines sont stables, d'autres moins. Cependant, les trop faibles rangs taxonomiques ne font pas l'objet de notre étude. En effet, nous travaillons à grande échelle et notre but sera ensuite d'inclure les eucaryotes afin de vérifier leurs relations de parenté avec les archées. On aura donc pour cela besoin d'arbres les plus stables possibles avec des groupes dont nous sommes certains de la validité et de la position phylogénétique. C'est pourquoi nous procédons au regroupement de groupes de niveaux supérieurs dont nous pouvons certifier la monophylie pour observer les différences. La liste des groupes retenus que nous considérons comme robustes est la suivante : Aigarcheota, Altiarchaeales, Asgard, Bathyarchaeota, Crenarchaeota, DPANN, Hadesarchaeota, Korarchaeota, Methanobacteria, SANT (Stenosarchaea + Archaeoglobi + Natronoarchaeota + Thermoplasmatales), SCGC, Thaumarchaeota, TTM (Thermococci + Theionarchaea (regroupe des Thermoplasmatales différentes des "vraies thermoplasmatales") + Methanomicrobia_Arc (/!\ différentes des Methanomicrobia du groupe SANT)), Verstraetearchaeota. Les groupes SCGC, TTM et SANT ont été définis à partir de nos résultats. Les autres groupes sont déjà nommés ainsi dans la littérature.

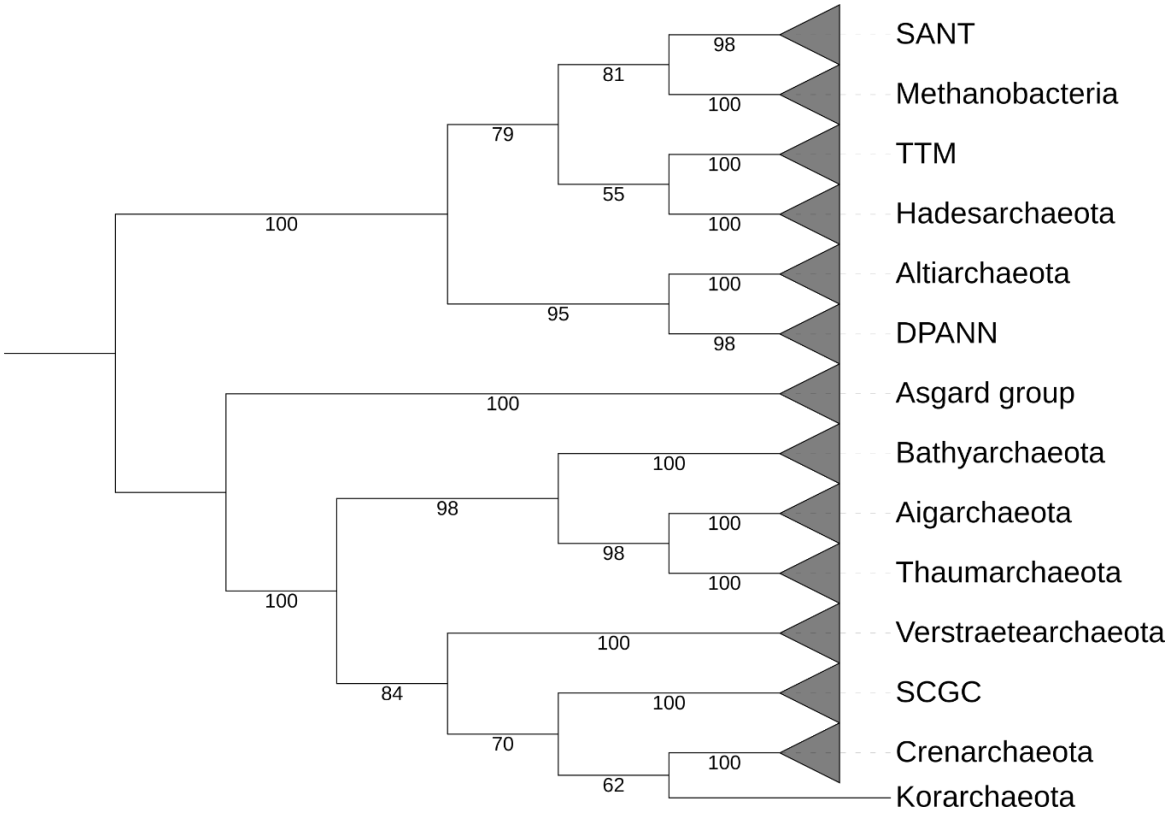
On notera la présence d'espèces appelées "thermoplasmatales" au sein de deux groupes : celui des "vraies" Thermoplasmata et celles formant le groupe des Theionarchaea, ce qui peut rendre ce terme ambigu. A ce jour, ces génomes ont été correctement renommés en Theionarchaea dans le NCBI. De même, les Methanomicrobia appelées Arc dans notre groupe TTM ne sont pas apparentées aux vraies Methanomicrobia, que l'on retrouve dans notre groupe SANT, mais à un groupe récemment nommé *Candidatus Methanofastidiosa*.

Basé sur ces groupes, les arbres consensus pour nos 5 réplicas d'espèces obtenus après un jackknife de gènes pour 100 réplicas par super-matrices calculés selon la méthode PMSF sont donnés **Figure 43**. L'enracinement de ses arbres par la méthode de midpoint laisse apparaître les deux grands groupes historiques que sont les Euryarchaeota et les TACK.

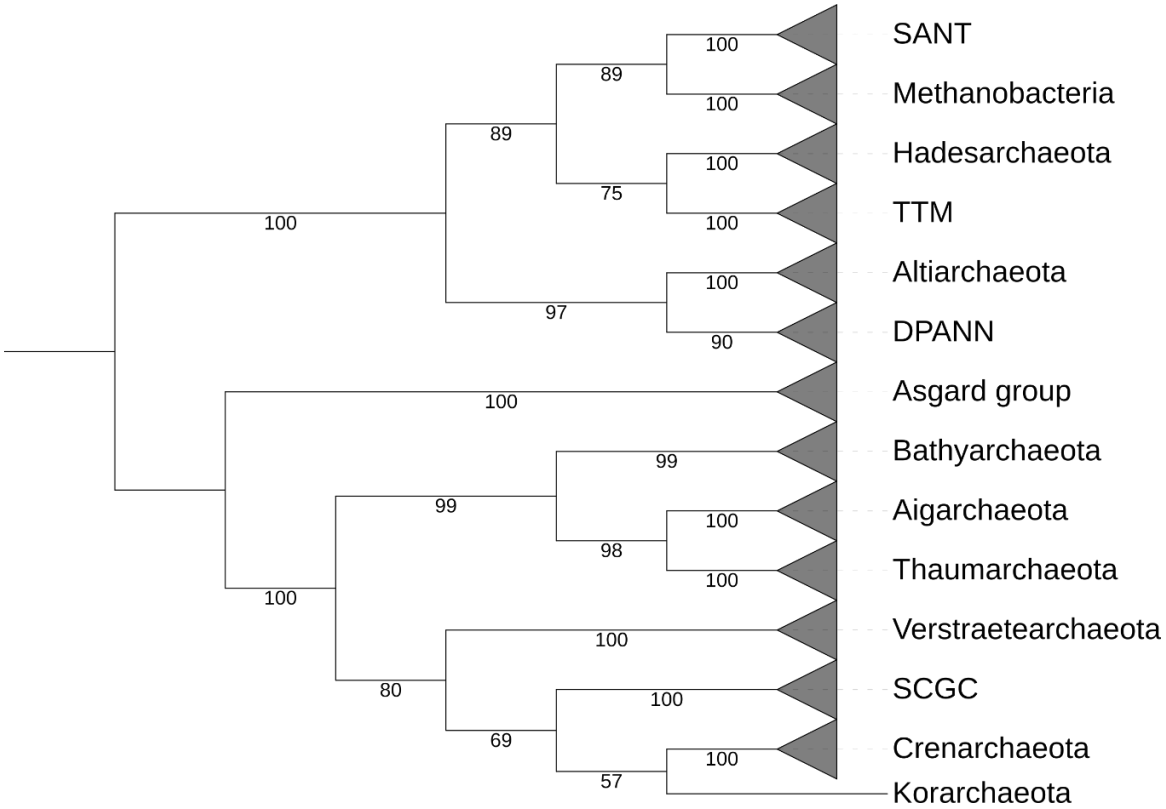
Réplica 1



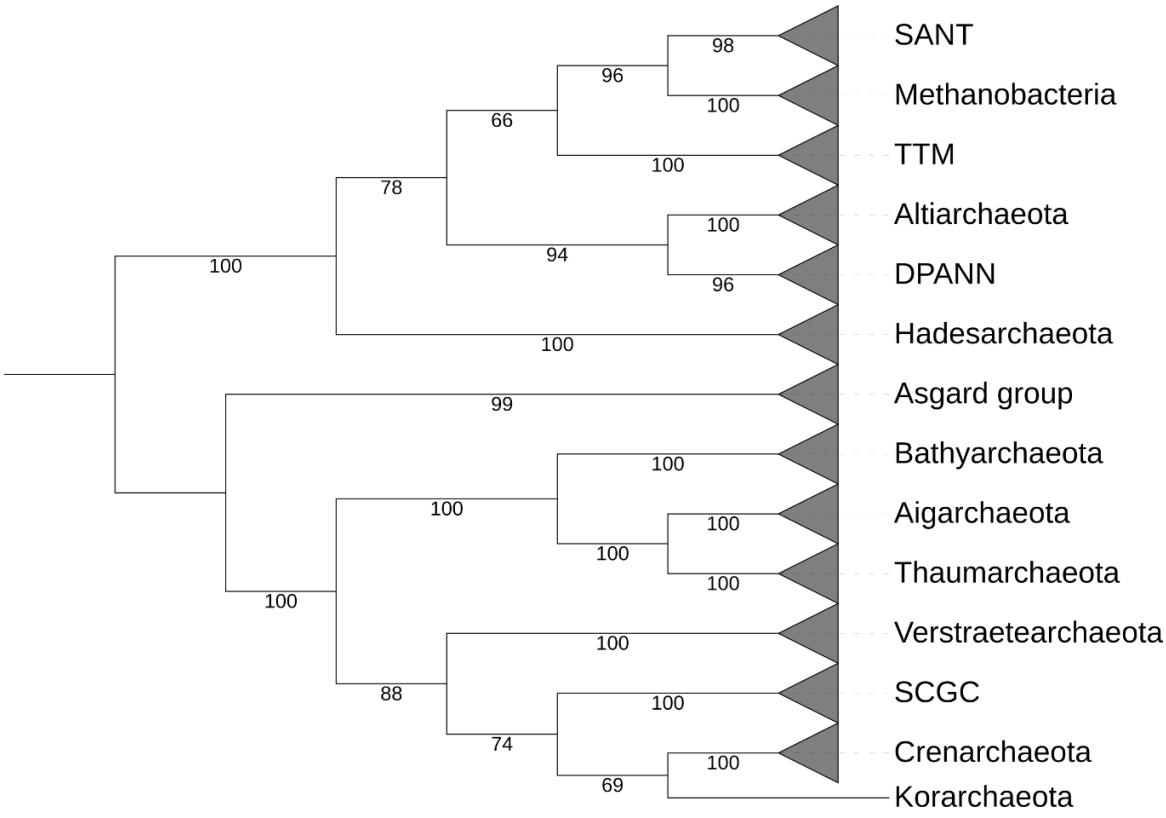
Réplica 2



Réplica 3



Réplica 4



Réplica 5

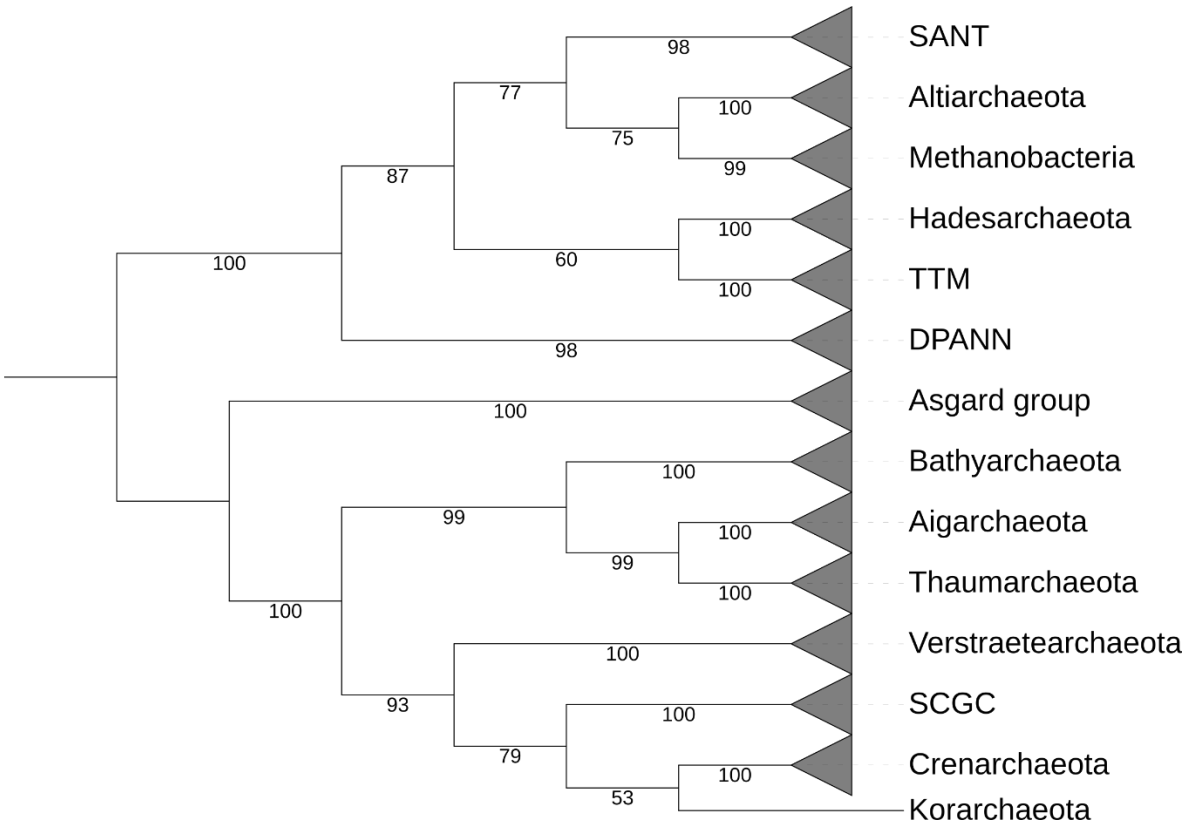


Figure 43. Arbres consensus obtenus pour nos 5 réplicas d'espèces (35 000 positions) obtenus après un jackknife de gènes pour 100 réplicas par super-matrices calculés selon la méthode PMSF.
Le détail des 100 arbres individuels sont donnés dans le **Supp.Mat FigS43**.
Les arbres ont été raciné selon la méthode du midpoint. Selon cette méthode, on retrouve les 2 groupes d'archées historiques que sont les Euryarchaeota et les TACK. Sur les 5 arbres, les réplicas 2 et 3 soutiennent la même topologie. Les Korarchaeota se placent soit comme groupe frère des Crenarchaeota, soit comme groupe frère des de l'ensemble Crenarchaeota + SCGC + Verstraearchaeota. Les Hadesarchaeota se placent soit comme groupe frère du groupe TTM, soit à la base de tous les Euryarchaeota et des DPANN. Les Altiarchaeales se placent soit comme groupe frère des DPANN soit comme groupe frère des Methanobacteria.

Nous distinguons en tout et pour tout pour l'ensemble de nos arbres 4 topologies possibles selon le réplica d'espèces utilisé ; 2 de nos 5 réplicas (les réplicas 2 et 3) tendent vers la même topologie). L'échantillonnage taxonomique joue donc un rôle important dans les études sur la phylogénie des archées. Ces 4 topologies sont données à la **Figure 44**. On peut observer ces 4 topologies en fixant la méthode complètement et en ne faisant varier que le réplica.

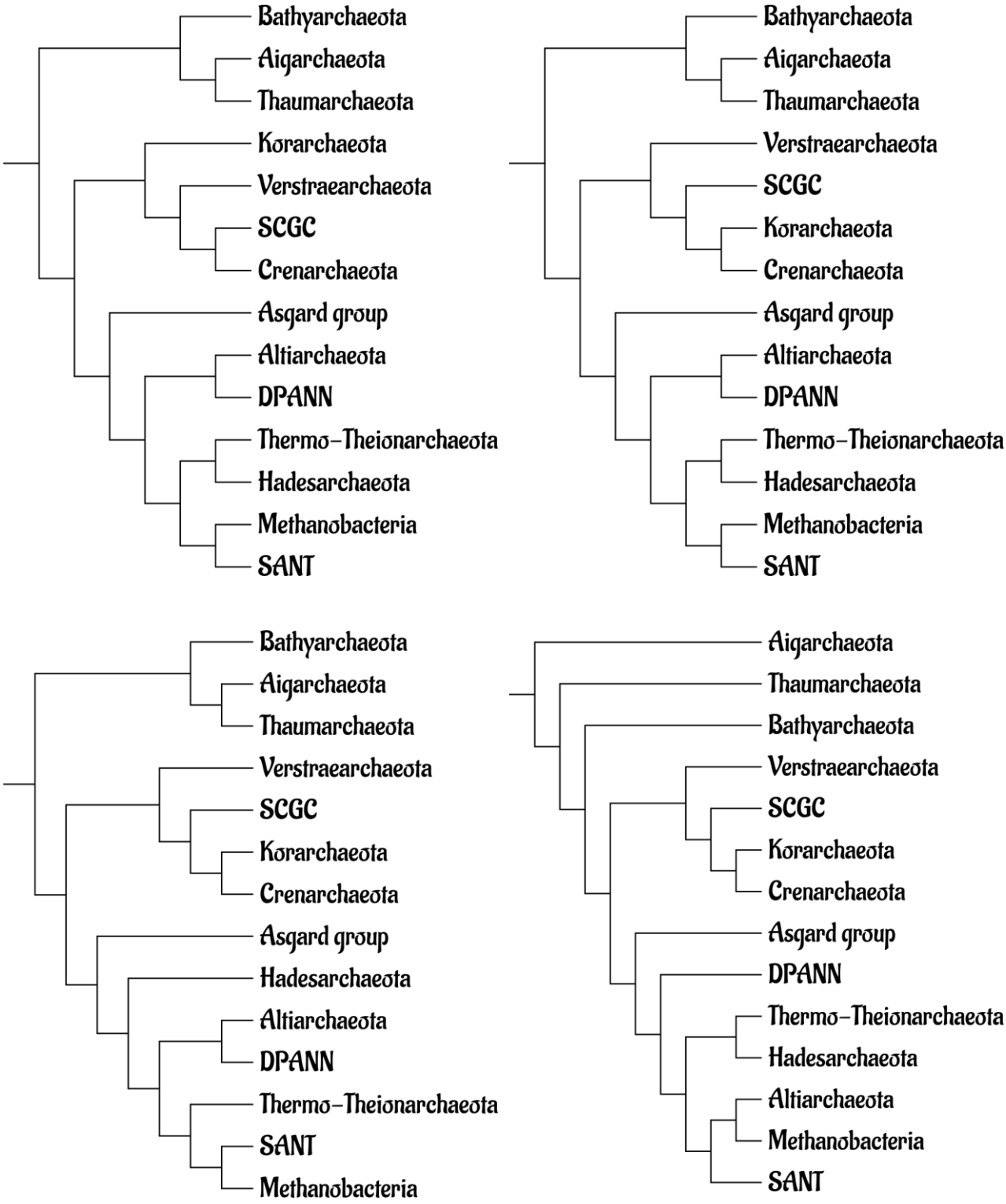


Figure 44. Différentes topologies possibles retrouvées avec nos jeux de données selon le réplica.
En fixant la topologie et en variant le réplica, 4 topologies sont retrouvées. Les positions variables concernent celles des Korarchaeota, des Hadesarchaeota et des Altiarchaeota.

Finalement, les groupes problématiques que nous retenons sont les suivants :

- les Korarchaeota, qui se placent soit comme groupe frère des Crenarchaeota, soit comme groupe frère des de l'ensemble Crenarchaeota + SCGC + Verstraearchaeota. Cette dernière hypothèse semble privilégiée par une étude se basant sur une phylogénie bayésienne sur une super-matrice de 41 gènes (36 gènes issus de la liste de gènes marqueurs Phylosift + les sous-unités A et B des ARN polymérase + 3 protéines ribosomiques universelles) (Adam et al., 2017; Raymann et al., 2015). En revanche, une autre étude basée sur une concaténation de 45 protéines ainsi que sur un super-arbre de 3 242 gènes insère les Korarchaeota à la base de tout le groupe TACK (T. A. Williams et al., 2017).
- Les Hadesarchaeota, qui se placent soit comme groupe frère du groupe Thermo-Theionarchaeota, soit à la base de tous les Euryarchaeota et des DPANN.
- Les Altiarchaeales, qui se placent soit comme groupe frère des DPANN soit comme groupe frère des Methanobacteria. Ce problème a déjà été étudié dans la littérature. Dans les premières analyses, les Altiarchaeales étaient associées aux Methanococcales (Probst et al., 2014). Il a été également suggéré (sur base de tentative d'enracinement des archées) qu'elles puissent représenter l'une des branches d'archées les plus profondes, peut-être au niveau de la classe ou même du phylum (Adam et al., 2017; Raymann et al., 2015). Cependant, en raison de leur taux d'évolution rapide, le placement des Altiarchaeales dans la phylogénie des archées doit être pris avec prudence. En effet, d'autres analyses ont également suggéré qu'elles pourraient être regroupées avec la lignée DPANN, elle aussi à évolution rapide (Aouad et al., 2022; Bird et al., 2016; Dombrowski et al., 2020; Hug et al., 2016), bien que cela puisse être causé par un artefact de reconstruction rendant leur position très incertaine (Martinez-Gutierrez & Aylward, 2021).

Une phylogénie présentant les positions alternatives des groupes problématiques est présentée **Figure 45**.

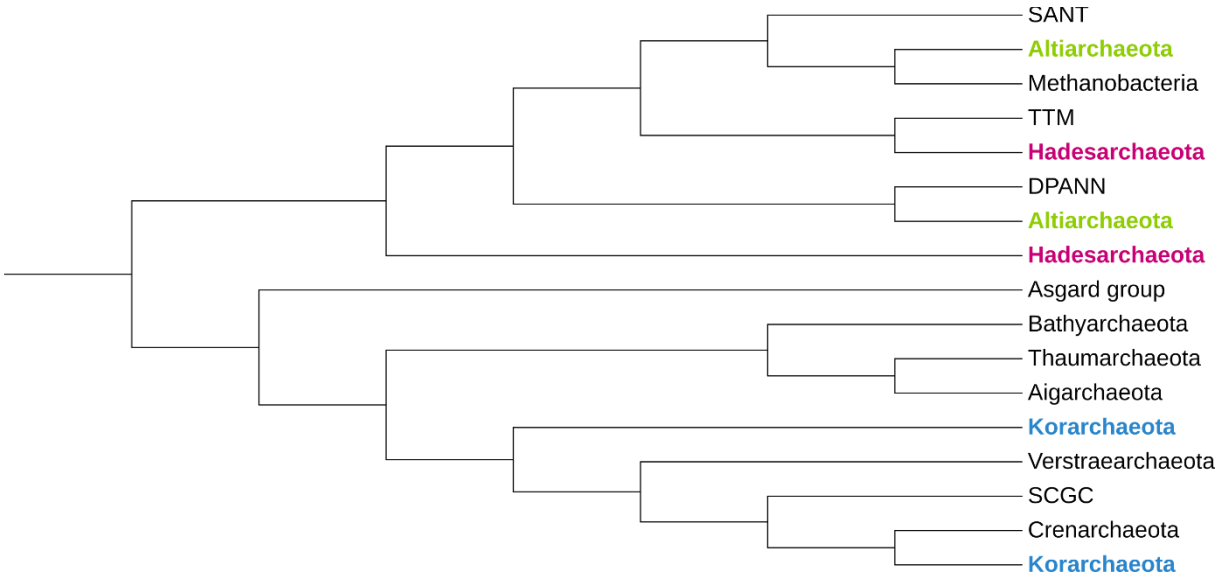


Figure 45. Cladogramme synthétique non raciné présentant les positions possibles des groupes instables (en couleur).

Nous avons trois groupes instables : les Korarchaeota (en bleu), les Hadesarchaeota (en rouge) et les Altiarchaeota (en vert), chacun ayant deux embranchements possibles dans l'arbre des archaea.

9 RETRAIT DE SITES (SLOW-FAST)

Nous soupçonnons que les variations dans les données et méthodes aient révélé des artefacts dû à du signal non-phylogénétique. En effet, une possibilité est que les positions à taux d'évolution rapide soient saturées par de multiples substitutions, introduisant alors un faux signal phylogénétique et une topologie d'arbre incorrecte rendant caduques les phylogénies : c'est l'erreur systématique (**Box 12**). Nous allons donc tenter d'évaluer l'importance de ce faux signal phylogénétique et de l'éliminer (si possible) de nos jeux de données ou de le neutraliser dans nos analyses. Afin de minimiser l'erreur systématique, nous avons choisi de tester notre jeu de données en éliminant les sites à taux d'évolution rapide via une approche dite slow-fast (Aouad et al., 2022; Brinkmann & Philippe, 1999; Delsuc et al., 2005; Roure & Philippe, 2011) afin de favoriser les sites plus lents, qui ont moins de chances de contredire les hypothèses d'homogénéité du processus de substitution (même si cela peut aussi être dû au fait qu'ils soient très contraints, pouvant provoquer des convergences). On combat ainsi la saturation mutationnelle pouvant mener à de l'homoplasie.

Le calcul des vitesses d'évolution par site et la construction des différentes matrices a été faite avec IQ-TREE en utilisant les 4 catégories de taux de la courbe gamma comme « proxy » pour 4 catégories de vitesses. La catégorie gamma d'une position est estimée par la somme des changements survenus au sein des séquences. Par définition, IQ-TREE classe les colonnes en 4 catégories, chacune correspondant à une vitesse d'évolution. Lors de l'inférence normale, chaque colonne appartient à chaque catégorie à 25% de probabilité. IQ-TREE estime les paramètres du modèle et applique alors une approche Bayésienne empirique afin d'attribuer les taux des sites comme la moyenne sur les catégories de taux, pondérée par la probabilité postérieure que le site tombe dans cette catégorie. Cette approche empirique Bayésienne est utilisée par IQ-Tree car elle est considérée comme la plus précise (Mayrose et al., 2004). Toutefois, à la demande, IQ-TREE peut effectuer une analyse bayésienne pour attribuer chaque colonne à l'une des catégories en particulier. Les sites sont ensuite assemblés en différentes matrices appelées « *bins* ». La première contient les sites ayant le taux d'évolution le plus lent, et les suivantes les sites ayant un taux d'évolution de plus en plus rapide. Nous partons ainsi de 4 bins de taille égale, que nous avons concaténés au fur et à mesure que l'on a incorporé des sites à taux d'évolution rapide, le 4ème alignement correspondant à notre alignement initial complet. Nous avons alors calculé avec IQ-TREE en LG4X et avec la méthode PMSF des arbres pour chacune de ces matrices, soit un total de 2 × 4 = 8 arbres. Les mêmes arbres ont également été calculés pour les 4 bins considérées individuellement.

Box 12. L'erreur systématique et le choix du modèle

L'erreur systématique

Un modèle d'évolution des séquences ne constitue qu'une représentation simplifiée du processus réel de substitution. Il se fonde sur des hypothèses qui doivent pouvoir expliquer les données analysées. Si les substitutions sont rares, il est facile de distinguer le signal phylogénétique. Quand les substitutions sont trop abondantes, elles se superposent (substitutions multiples) et font disparaitre le signal originel. La saturation du signal se produit lorsqu'il y a plus d'un changement par site en moyenne. Du fait de leur nombre d'états de caractères possibles réduit (seulement 4), les séquences nucléotidiques sont plus sensibles à la saturation que les séquences d'acides aminés (au nombre de 20). C'est pourquoi ce sont ces dernières qui sont préférentiellement utilisées lors de la résolution phylogénies anciennes, car elles sont moins susceptibles de saturer. L'erreur systématique survient alors si le processus évolutif accumulant les substitutions multiples viole en

plus les suppositions du modèle utilisé pour la reconstruction phylogénétique, c'est-à-dire par le fait que les particularités de l'évolution des séquences ne sont pas correctement prises en compte par les hypothèses sous-jacentes. Cette erreur est d'autant plus grande que le nombre de caractères utilisés est grand. L'erreur systématique produit un signal non-phylogénétique entrant en compétition avec le véritable signal phylogénétique. Contrairement à l'erreur stochastique, les analyses de bootstrap ou de jackknife n'indiquent pas forcément la présence d'erreur systématique dans un jeu de données, pouvant parfois conduire à des résultats à la fois faux et statistiquement supportés. Les matrices qui contiennent plus de positions rapides ont davantage tendance à induire des artefacts de reconstruction phylogénétique. Il est donc impératif de trouver un moyen de les détecter.

Le phénomène d'attraction des longues branches (LBA)

L'un des principaux artefacts de reconstruction phylogénétique lié à l'erreur systématique est le phénomène d'attraction des longues branches. L'attraction des longues branches est un artefact provoquant le regroupement de taxons (1) soit évoluant plus rapidement, (2) soit ayant divergé tôt (branchement ancien), sans refléter leurs véritables liens de parenté. En effet, lorsque les vitesses d'évolution entre les lignées étudiées sont très différentes, celles évoluant le plus vite ont plus de chances de partager des caractères par homoplasie plutôt que par homologie. Dans un contexte de parcimonie, le phénomène est principalement lié à l'accumulation de substitutions convergentes interprétées comme des synapomorphies (alors qu'il s'agit de convergences !), alors qu'avec les méthodes probabilistes, il est plutôt lié aux violations de modèles mentionnées ci-dessus. De ce fait, ces lignées qui ont de longues branches (c'est-à-dire un grand nombre d'événements évolutifs) sont regroupées ensemble dans l'arbre, et ce indépendamment de leur parenté. Par conséquent, si une phylogénie est établie à partir d'un gène évoluant plus vite dans une lignée que dans une autre, alors il est fort probable que celle-ci se retrouve placée à la base de l'arbre, attirée par le groupe externe servant à enraciner la phylogénie (lui-même pourvu d'une longue branche) et soit ainsi interprétée comme un groupe relativement ancestral.

Slow-Fast

Les positions à taux d'évolution rapide étant saturées par de multiples substitutions, elles ont perdu leur signal phylogénétique. Pour minimiser l'erreur systématique et ses effets dans la reconstruction phylogénétique, plusieurs stratégies ont été proposées. Celles-ci consistent à éliminer le signal non-phylogénétique des jeux de données, c'est-à-dire les substitutions mal interprétées par les méthodes de reconstruction et qui supportent une topologie incorrecte. En éliminant ainsi les sites à taux d'évolution rapide, on favorise ceux qui ont moins de chances de contredire les hypothèses d'homogénéité du processus de substitution. On combat ainsi la saturation mutationnelle pouvant mener à de l'homoplasie.

Ainsi, la méthode Slow/Fast consiste à identifier les positions qui n'ont subi aucune substitution à l'intérieur des groupes prédéfinis (positions les plus lentes) et à rajouter progressivement les positions qui ont subi au plus une, deux, trois, etc., substitutions. De là, on peut soit (1) créer un ensemble de super-matrices imbriquées qui contiennent de plus en plus de positions rapides, soit (2) créer des ensembles de super-matrices à taux d'évolution bien distinct. L'observation de l'évolution des supports pour les différentes hypothèses topologiques permet alors d'évaluer à partir de quel taux de mutation on perd le signal phylogénétique pour laisser place à de l'erreur systématique.

Références
Philippe H., de Vienne D.M., Ranwez V., Roure B., Baurain D. & Delsuc F. 2017. Pitfalls in supermatrix phylogenomics. European Journal of Taxonomy XXX: 1–25.
Philippe H, Brinkmann H, Lavrov DV, Littlewood DTJ, Manuel M, et al. (2011) Resolving Difficult Phylogenetic Questions: Why More Sequences Are Not Enough. PLoS Biol 9(3): e1000602.
Lartillot, N., Brinkmann, H., & Philippe, H. (2007). Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. BMC evolutionary biology, 7 Suppl 1(Suppl 1), S4.
Felsenstein, J. (1978). Cases in which Parsimony or Compatibility Methods Will be Positively Misleading. Systematic Zoology, 27(4), 401-410.
Phillips, M. J., F. Delsuc, et D. Penny. 2004. Genome-scale phylogeny and the detection of systematic biases. Mol Biol Evol 21:1455-8.
Rodríguez-Ezpeleta, Naiara, et al., 2007a, « Detecting and Overcoming Systematic Errors in Genome-Scale Phylogenies », sous la dir. de Frank Anderson, Systematic Biology 56, no 3 () : 389-399, issn : 1076-836X.
Brinkmann, H., et H. Philippe. 1999. Archaea sister group of Bacteria? Indications from tree reconstruction artifacts in ancient phylogenies. Mol Biol Evol 16:817-25.
Brochier, C., et H. Philippe. 2002. Phylogeny: a non-hyperthermophilic ancestor for bacteria. Nature 417:244.
Delsuc, F., H. Brinkmann, et H. Philippe. 2005. Phylogenomics and the reconstruction of the tree of life. Nat Rev Genet 6:361-75.
Keeling PJ, Burger G, Durnford DG, Lang BF, Lee RW, Pearlman RE, Roger AJ, Gray MW. The tree of eukaryotes. Trends Ecol Evol. 2005 Dec;20(12):670-6. Epub 2005 Oct 10. PMID: 16701456.
Philippe, H., P. Lopez, H. Brinkmann, K. Budin, A. Germot, J. Laurent, D. Moreira, M.
Muller, et H. Le Guyader. 2000. Early-branching or fast-evolving eukaryotes? An answer based on slowly evolving positions. Proc R Soc Lond B Biol Sci 267:1213-21.
Kapli P, Yang Z, Telford MJ. Phylogenetic tree building in the genomic age. Nat Rev Genet. 2020 Jul;21(7):428-444. Epub 2020 May 18. PMID: 32424311.

On notera plusieurs choses dans nos groupes ribosomiques :

Au sein du groupe SANT, il est difficile de trancher concernant la position des Halobacteria, des Archaeoglobi et des Methanonatronarchaeia, les trois pouvant potentiellement se positionner à la racine du groupe SANT. De plus, une autre possibilité est le rapprochement des Halobacteria avec les Methanomicrobiales. Toutefois, on note des variations entre bins mais également entre réplicas qui sont difficiles à interpréter (**Figure 46** & **Supp.Mat slow-fast**). Par exemple, concernant le rapprochement des Halobacteria avec les Methanomicrobiales, pour les bins non cumulés, pour le réplica 2, le premier bin est à 9 % (ultrafast-bootstrap ramené sur 100 car nous avons 10 000 réplicas de bootstrap), le deuxième à 57 %, le troisième à 2 % ; pour le réplica 4, le premier bin est à 3 %, le deuxième à 76 %, le troisième à 0. Les réplicas 1 et 3 privilégient largement un rapprochement des Methanomicrobiales avec les Halobacteria (ultrafast-bootstrap > 95 %) tandis que les réplicas 2 et 4 privilégient un rapprochement des Halobacteria avec Methanonatron. Cette dernière hypothèse est supportée pour le réplica 5 pour les deux premiers bins seulement. Enfin, la monophylie du groupe SANT n'est pas retrouvée avec le dernier bin dès lors que l'on utilise des positions où l'erreur systématique est maximale, alors que les super-matrices complètes et l'utilisation de l'ensemble des bins cumulés retrouvent bien ce groupe.

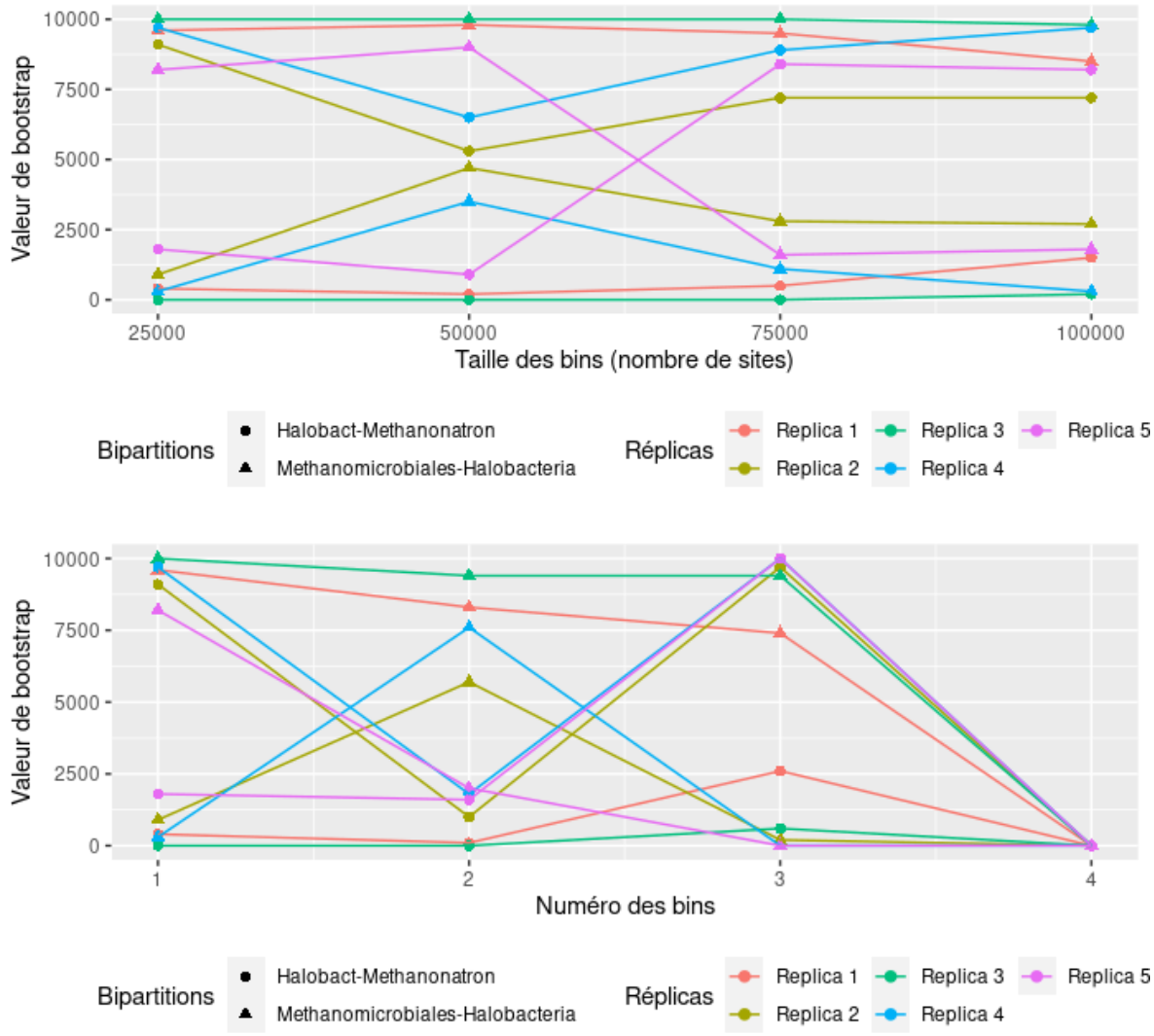


Figure 46. Proportion d'arbres indiquant les 2 positions possibles des Halobacteria (soit avec les Methanonatronarchaeia soit avec les Methanomicrobiales) selon la taille des bins cumulés et non cumulés pour chaque réplica selon le modèle LG4X.

Les données sont fournies **Supp.Mat slow-fast**. Il est difficile de se prononcer sur la position des Halobacteria, des Archaeoglobi et des Methanonatronarchaeia, les trois pouvant potentiellement se positionner à la racine du groupe SANT. Le résultat varie à la fois selon le réplica et le bin, laissant supposer que à la fois le choix d'espèces et l'alignement utilisé peut donner des résultats différents.

Les Altiarchaeota tendent à se rapprocher du groupe DPANN, sauf pour le réplica 5 (**Figure 47**). Pour les bins cumulés, l'ultrafast-bootstrap est systématiquement supérieur à 94 % pour les réplicas 1 à 4. En revanche, cette hypothèse n'est pas supportée pour le réplica 5, avec une valeur d'ultrafast-bootstrap inférieure à 11 %. Pour les bins non cumulés, on retrouve de faibles valeurs d'ultrafast-bootstrap pour l'hypothèse DPANN + Altiarchaeota pour le réplica 5 (ultrafast-bootstrap < 11 %). En revanche, pour les réplicas 1 à 4, les valeurs d'ultrafast-bootstrap favorisent cette hypothèse pour les trois premiers alignements (ultrafast-bootstrap > 92 %) tandis que le dernier alignement (contenant les sites les plus rapides) ne trouve jamais cette hypothèse. Il semblerait donc selon nos analyses que les Altiarchaeota sortent des Euryarchaeota, bien que la

conclusion soit dépendante du réplica. On notera que le réplica 5 ne supporte aucune de nos deux hypothèses.

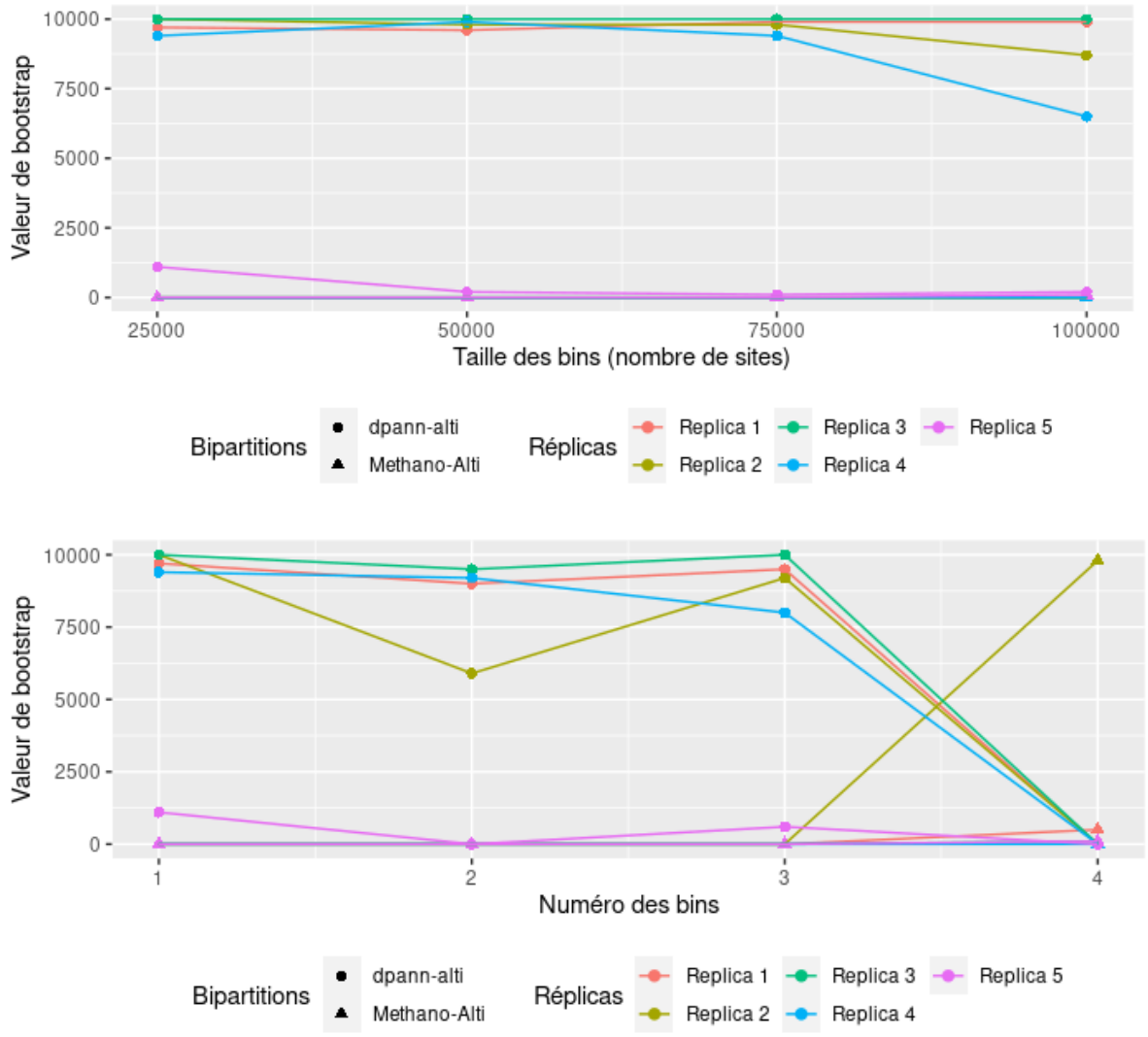


Figure 47. Proportion d'arbres indiquant les 2 positions possibles des Altiarchaeota (soit avec les DPANN soit avec les Methanomicrobiales) selon la taille des bins cumulés et non cumulés pour chaque réplica selon le modèle LG4X.

Les données sont fournies **Supp.Mat slow-fast**. Les Altiarchaeota tendent à se rapprocher du groupe DPANN, sauf pour le réplica 5.

Nous avons précédemment observé que la position des Korarchaeota était incertaine et variait selon le modèle utilisé, tantôt à la base des Crenarchaeota, tantôt à la base des Crenarchaeota + SCGC (**Figure 48**). Nos analyses *slow-fast* permettent ici de trancher quant à leur position en favorisant les Korarchaeota plutôt que le groupe SCGC comme groupe frère des Crenarchaeota. Cette solution est très bien supportée avec l'utilisation des bins cumulés (ultrafast-bootstrap > 71 %). En revanche, l'utilisation de bins individuels montrent bien que l'utilisation de sites rapides privilégie l'hypothèse alternative des SCGC à la base de tous les Crenarchaeota, signe d'un artefact de reconstruction phylogénétique. Les Korarchaeota et Crenarchaeota forment donc un groupe monophylétique qui s'enracine dans le groupe SCGC.

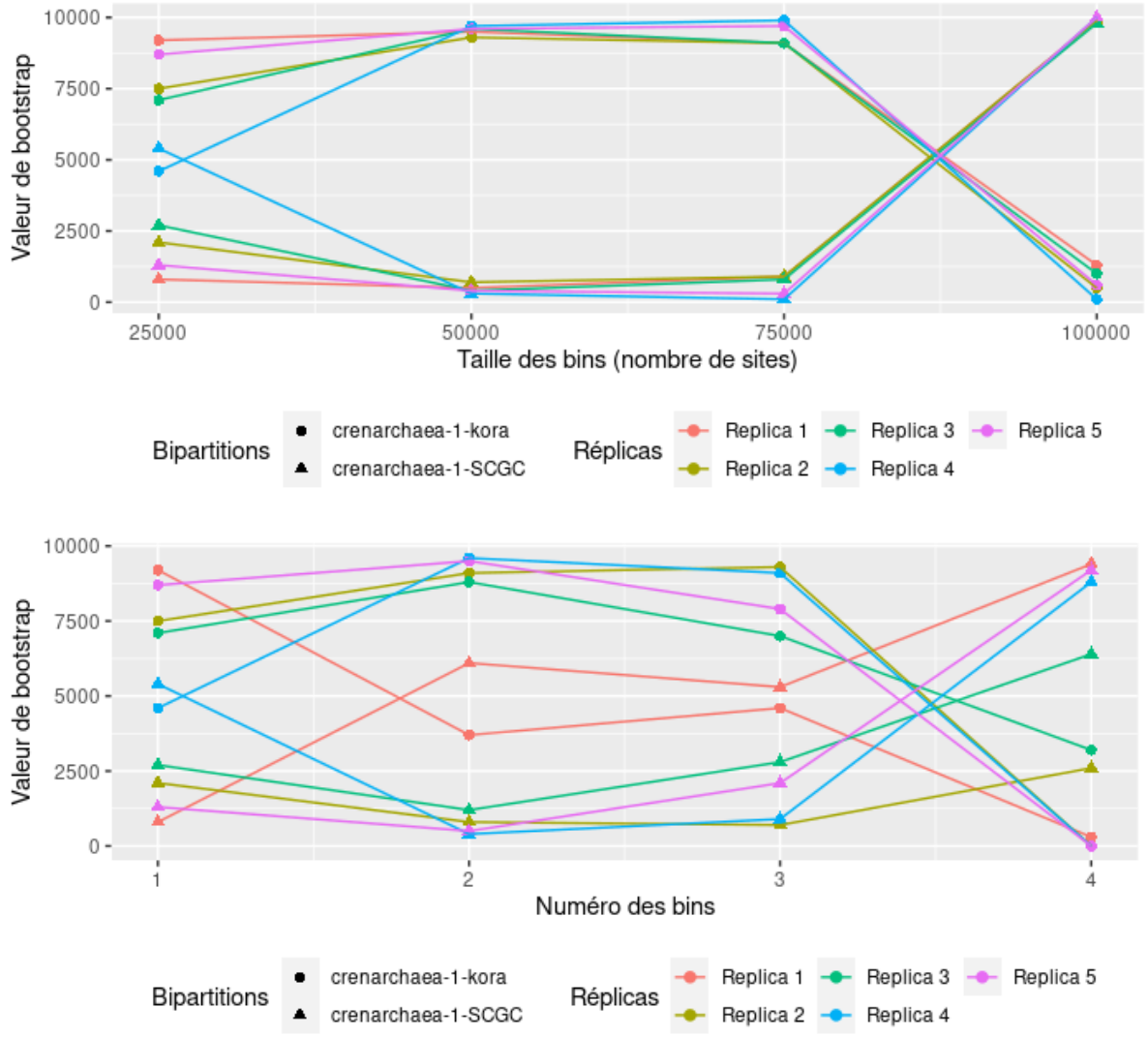


Figure 48. Proportion d'arbres indiquant les 2 positions possibles des Korarchaeota et des SCGC en fonction des Crenarchaeaota selon la taille des bins cumulés et non cumulés pour chaque réplica selon le modèle LG4X.

Les données sont fournies **Supp.Mat slow-fast**. Nos résultats montrent que les les Korarchaeota et Crenarchaeota forment un groupe monophylétique qui s'enracine dans le groupe SCGC.

Les Hadesarchaeota privilégient comme groupe frère des Thermococci, des Theionarchaea et des Methanomicrobia_Arc lors de l'utilisation de sites lents. Seul le réplica 4 peine à soutenir cette hypothèse, avec un ultrafast-bootstrap maximum de 7 % lors de l'utilisation de bins séparés, alors que l'utilisation de bins cumulés donne des ultrafast-bootstraps inférieurs à 17 % sauf pour le dernier qui atteint une forte valeur de 89 % (cf. **Supp.Mat slow-fast/SlowFast_bilan.xlsx**).

10 DISCUSSION

Plusieurs études ont montré qu'une augmentation de l'échantillonnage taxonomique peut améliorer la précision phylogénétique (Jeffroy et al., 2006; Pollock et al., 2002; Rokas & Carroll, 2005; Zwickl & Hillis, 2002). Cette stratégie est couramment utilisée comme solution pour résoudre les nœuds instables dans l'arbre du vivant (Young & Gillung, 2020). Dans certains cas, il

a été suggéré que les conflits entre les arbres rapportés peuvent résulter de différences dans la représentation taxonomique (Da Cunha et al., 2017; Nasir et al., 2016) et il n'est pas clair dans quelle mesure le sur-échantillonnage de certains taxons par rapport à d'autres peut affecter de manière délétère la reconstruction d'un arbre (Martinez-Gutierrez & Aylward, 2021).

Phylogénie des Bathyarchaeota

Bien que la phylogénie interne des Bathyarchaeota ne soit pas l'objet précis de notre étude, nous pouvons mentionner que la première phylogénie des Bathyarchaeota fut établie en 2012 sur la base de 4 720 séquences issues de la base de données SILVA (Kubo et al., 2012). Des séquences de 940 pb minimum ont été utilisées afin de construire le squelette de l'arbre, auquel on a ensuite rajouté des séquences sans changer la topologie générale de l'arbre. Que ce soit avec les méthodes de distance ou de maximum de vraisemblance, un total de 17 sous-groupes furent identifiés, avec 76 % de similarité partagée par les séquences les plus distantes. Cependant, 12 % des séquences sont restées non regroupées et isolées. Les sous-groupes 13 à 17 étaient instables selon les méthodes de reconstruction utilisées et présentaient des multifurcations. En 2016, une autre étude a rajouté deux nouveaux sous-groupes (18 et 19) avec de fortes valeurs d'ultrafast-bootstrap (respectivement 96 % et 86 %) (Fillol et al., 2016). Le sous-groupe 5 a de plus été divisé en 5a et 5b, chacun avec une similarité intra-groupe supérieure à 90 % selon les estimations de maximum de vraisemblance. Le groupe 5b fut lui-même subdivisé en 5b et 5bb au fur et à mesure que de nouvelles séquences étaient ajoutées. Les Bathyarchaeota sont caractérisés par une très forte diversité intra-groupe. La limite des séquences des ARNr 16S des Bathyarchaeota tombe dans l'étendue du minimum d'identité des séquences de ce qui est considéré comme un phylum (74.95–79.9%) et chaque sous-groupe tombe dans l'intervalle d'identité de séquence médian des séquences des familles et des ordres (respectivement 91.65-92.9 % et 88.25–90.1 %) (Yarza et al., 2014). Il a été proposé que cette haute diversité des Bathyarchaeota soit le reflet d'une haute diversité des métabolismes au sein des différents sous-groupes (Kubo et al., 2012). Actuellement, la classification des Bathyarchaeota basée sur les ARNr 16S supporte un total de 25 groupes (Zhou et al., 2018).

Monophylie du groupe TACK et apparentement avec le groupe Asgard

Pendant plus d'une décennie (1990-2002), les Euryarchaeota et les Crenarchaeota étaient les seuls deux groupes d'archées connus. Entre 2002 et 2011, grâce au développement des techniques de séquençage (*amplicon sequencing*, *shotgun metagenomics*…) plusieurs nouveaux embranchements ont été proposés sur la base d'analyses phylogénétiques et génomiques : Korarchaeota, Nanoarchaeota (cultivé en 2002) et Thaumarchaeota (qui oxyde l'ammoniac) et Aigarchaeota (Brochier-Armanet, Boussau, et al., 2008; Elkins et al., 2008; Huber et al., 2002; Nunoura et al., 2011). Ensemble, ils forment le superphylum TACK (ou Protoarchaeota) (Kozubal et al., 2013; Petitjean et al., 2014), auquel se sont récemment ajoutés les métagénomes Geoarchaeota (Guy et al., 2014; Kozubal et al., 2013), les Bathyarchaeota (Evans et al., 2015; He et al., 2016; Lazar et al., 2016) et les Verstraetearchaeota (Vanwonterghem et al., 2016). La monophylie des TACK est bien obtenue dans nos analyses. Nous retrouvons systématiquement le groupe Asgard à la base des TACK, indépendamment de la stratégie d'échantillonnage. Cette conclusion est soutenue par des études récentes (Adam et al., 2017; Spang et al., 2015; T. A. Williams et al., 2017; Zaremba-Niedzwiedzka et al., 2017). Certains ont néanmoins suggéré que le placement des archées d'Asgard près de TACK pourrait être dû en partie à un échantillonnage déséquilibré des taxons (Da Cunha et al., 2017; Nasir et al., 2016). Toutefois, dans notre cas, nous

avons bien fait en sorte que notre échantillonnage soit bien équilibré, ce qui indique que ce résultat est probablement correct.

Nomenclature et taxonomie générale des groupes d'archées

Une attention particulière doit également être portée à la nomenclature des archées. Par deux fois, nous avons relevé des cas d'homonymie pour des espèces appartenant à deux genres non apparentés. C'est le cas d'espèces nommées (1) « thermoplasmatales » au sein de deux groupes (celui des « vraies » Thermoplasmata et celles formant le groupe des Theionarchaea), et (2) de Methanomicrobia appelés Arc qui ne sont pas apparentées aux vraies Methanomicrobia que l'on retrouve dans notre groupe SANT. Il est difficile de dire si cette confusion est le fruit d'un descriptif originalement basé sur le métabolisme et la morphologie ou s'il s'agit d'une erreur phylogénétique. La taxonomie actuelle des groupes d'archées reconnus est incohérente en raison de l'absence de lignes directrices claires basées sur le génome pour la classification des micro-organismes (Gribaldo & Brochier-Armanet, 2012). Ces divergences peuvent s'expliquer en partie par le pouvoir de résolution limité de l'ARNr 16S, en particulier pour les nœuds les plus anciens, par des artefacts méthodologiques ou par la mauvaise qualité ou la petite taille des séquences peuvent expliquer ces problèmes de nomenclature ainsi que la multiplication du nombre de phylum décrits ces dernières années. La disponibilité croissante de données de séquences génomiques provenant de lignées non échantillonnées auparavant a conduit à une progression spectaculaire des propositions (au moins 3 super-phyla (TACK, DPANN, Asgard) et des dizaines de nouveaux phyla) qui sont venues s'ajouter à la dichotomie historique Crenarchaeota et Euryarchaeota définie par Woese en 1990. Compte tenu des progrès continus des technologies de séquençage et de l'exploration de la diversité microbienne, le risque existe d'une explosion anarchique de taxons de haut niveau. Il serait ainsi urgent d'établir des règles communes qui devraient être respectées avant une proposition formelle d'un nouvel embranchement d'archées. Ce n'est que depuis récemment que des efforts ont commencé à être entrepris afin d'utiliser des critères bien définis basés sur des cadres phylogénomiques et génomiques comparatifs (des distances génomiques en pourcentage) pour standardiser et aider à établir une taxonomie solide des archées (mais également des bactéries) (Parks et al., 2018; Rinke et al., 2021).

Importance de la sélection d'espèces, de gènes et du modèle utilisé

Lors de notre étude, les résultats sont fortement dépendant du réplica d'espèces plus que de la sélection de gènes. Pour chaque méthode employée (super-arbre vs super-matrice), les phylogénies obtenues ont des topologies congruentes dans l'ensemble, ce qui est la preuve d'un signal phylogénétique cohérent dans ces différents jeux de données. Un gros travail a été fourni afin de s'assurer d'utiliser des gènes fiables, et nos résultats semblent aller dans le sens d'une robustesse des topologies si l'on ne tient compte que des gènes. En revanche, l'incongruence s'observe surtout au niveau du réplica et du modèle utilisé. La variabilité entre modèles est moins marquée et se fait surtout sentir lors de l'utilisation du modèle LG4X, qui est plus sensible aux variations d'échantillonnage de gènes que les modèles catégoriels car moins sophistiqué. De plus, les super-arbres sont moins robustes que les super-matrices. Enfin, les réplicas d'espèces ne se comportent pas tous de la même façon, d'où l'intérêt d'en avoir étudié 5 différents. Le choix des espèces semble donc un critère primordial, avec l'utilisation de modèles sophistiqués.

Parmi nos propositions d'arbres phylogénétiques des archées, celui qui nous semble donc le plus crédible est donné à la **Figure 49** suivante, correspondant aux arbres de nos réplica 2 et 3 :

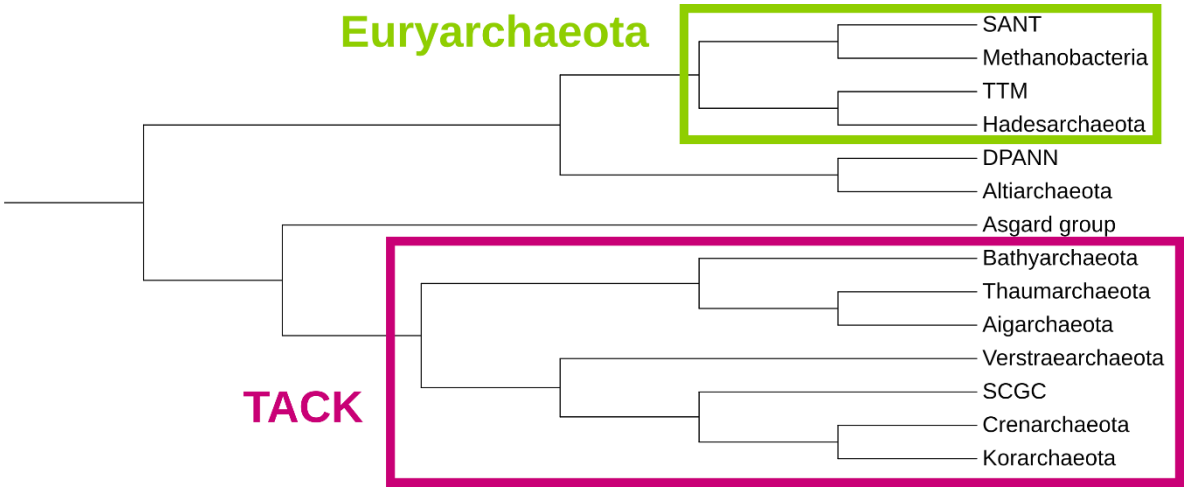


Figure 49. Arbres phylogénétiques non raciné des archées paraissant le plus crédible selon nos résultats.

Nous retrouvons les deux grands groupes que sont les TACK et les Euryarchaeota. Les DPANN et Altiarchaeota forment un troisième groupe en dehors des Euryarchaeota.

CHAPITRE 3

LES EUCARYOTES ET LEURS RELATIONS AVEC LES ARCHAEA

1 INTRODUCTION

Les dernières études suggèrent que les eucaryotes partagent un ancêtre commun avec le groupe Asgard. Toutefois, on peut discuter ce résultat qui pourrait résulter de biais méthodologiques, en particulier d'un artefact d'attraction des longues branches. Afin de tester cette hypothèse, nous allons sélectionner les gènes présents à la fois chez les archées et les eucaryotes et relancer les analyses phylogénétiques décrites précédemment afin de vérifier l'impact de l'ajout des eucaryotes sur nos arbres. Nous avons obtenu dans le chapitre précédent une phylogénie des archées pour 5 réplicas d'espèces. Notre objectif est donc maintenant d'insérer une sélection d'eucaryotes au sein de ces arbres.

2 SELECTION DES EUCARYOTES

Nous possédons au sein de notre laboratoire une expertise concernant les eucaryotes (Van Vlierberghe, Philippe, et al., 2021). Nous nous sommes focalisés sur les protéomes de qualité supérieure, en particulier ceux avec une grande complétude, ce qui est essentiel pour obtenir des groupes orthologues les plus complets et équilibrés possibles (Simão et al., 2015; Waterhouse et al., 2018), avec de faibles niveaux de contamination (Irisarri et al., 2017; Simion et al., 2017; Van Vlierberghe, Di Franco, et al., 2021). Nous possédons au sein de notre laboratoire une sélection de 73 protéomes eucaryotes jugés de qualité supérieur (Van Vlierberghe, Philippe, et al., 2021). Nous avons alors sélectionné parmi ces protéomes 11 espèces représentant au mieux la diversité des eucaryotes. Idéalement, cette sélection devrait nous permettre d'émettre des hypothèses sur les propriétés de LECA et d'éviter des synapomorphies qui ne concerneraient que certains sous-groupes d'eucaryotes. La sélection d'eucaryotes que nous souhaitons ajouter à nos archées est la suivante :

Espèces	Super-groupe	Phylum
Arabidopsis thaliana	Archaeplastida	Chloroplastida
Cyanidioschyzon merolae	Archaeplastida	Rhodophyta
Cyanophora paradoxa	Archaeplastida	Glaucophyta
Dictyostelium discoideum	Amorphea	Amoebozoa
Homo sapiens	Amorphea Obazoa	Opisthokonta
Ustilago maydis	Amorphea Obazoa	Opisthokonta
Emiliania huxleyi	Haptista	Haptophyta
Guillardia theta	Cryptista	Cryptophyta
Plasmodiophora brassicae	TSAR	Rhizaria
Vitrella brassicaformis	TSAR	Alveolata
Phytophthora infestans	TSAR	Stramenopiles

Nous avons ajouté avec 42 cette sélection de 11 eucaryotes sur nos 440 gènes (c'est-à-dire avant les tests de congruence et la création de chimères) de nos 5 réplicas d'archées (121 espèces), soit un total de 132 espèces par réplica. Ainsi, pour chaque réplica, 204, 199, 206, 202 et 202 gènes ont été enrichis d'au moins une séquence eucaryote.

3 SELECTION DES GENES

Maintenant que nous avons ajouté notre sélection d'eucaryotes, il nous faut sélectionner les gènes que nous allons garder. Plusieurs critères vont entrer en compte pour la sélection de nos gènes. Il nous faut des gènes où les eucaryotes sont suffisamment représentés, à la fois en nombre et en diversité. Nous avons également besoin de limiter les problèmes de paralogie susceptibles d'apparaitre lorsqu'un même eucaryote est ajouté plusieurs fois au sein d'un même groupe orthologue. En effet, certains gènes eucaryotes peuvent avoir subi de nombreuses duplications ayant leur propre évolution indépendante au cours de temps. Il convient alors de gérer ces paralogues et d'élucider lesquels correspondent à des homologues archéens.

Nous avons évalué si pour un MSA donné, certaines espèces étaient présentes plusieurs fois, traduisant de possibles paralogies. Pour ce faire, nous avons compté le nombre d'occurrences de chaque espèce eucaryote au sein de chaque MSA (**Figure 50**). Certains organismes ont été rajoutés en plusieurs copies au sein d'un même OG, traduisant de possibles phénomènes de paralogie. Citons en particulier les cas extrêmes d'*Arabidopsis thaliana*, présente 12 et 14 fois au sein de OG0000295 et OG0000825, et d'*Emiliania huxleyi*, présente 13 fois dans OG0000295. D'une façon générale, cette analyse montre qu'*Arabidopsis thaliana* et *Guillardia theta* sont plus susceptibles d'être ajoutées sous la forme de plusieurs paralogues dans nos groupes orthologues, résultant soit d'événements de duplications au sein de leur lignée, soit de transferts de gènes soit d'endosymbioses secondaires. Ainsi, *Guillardia theta* possède un nucléomorphe, vestige de 551 kpb du noyau de l'algue rouge qui aurait été englobée au cours de leur évolution, accentuant ce problème de paralogie. Les endosymbioses successives à l'origine de ces microorganismes confèrent à ces derniers une organisation génomique particulièrement complexe, avec quatre génomes différents (Sibbald & Archibald, 2020) :

- deux génomes procaryotiques, dans les mitochondries et les plastes des algues rouges et vertes ;
- deux génomes eucaryotiques, dans le noyau de la cellule hôte et dans le nucléomorphe.

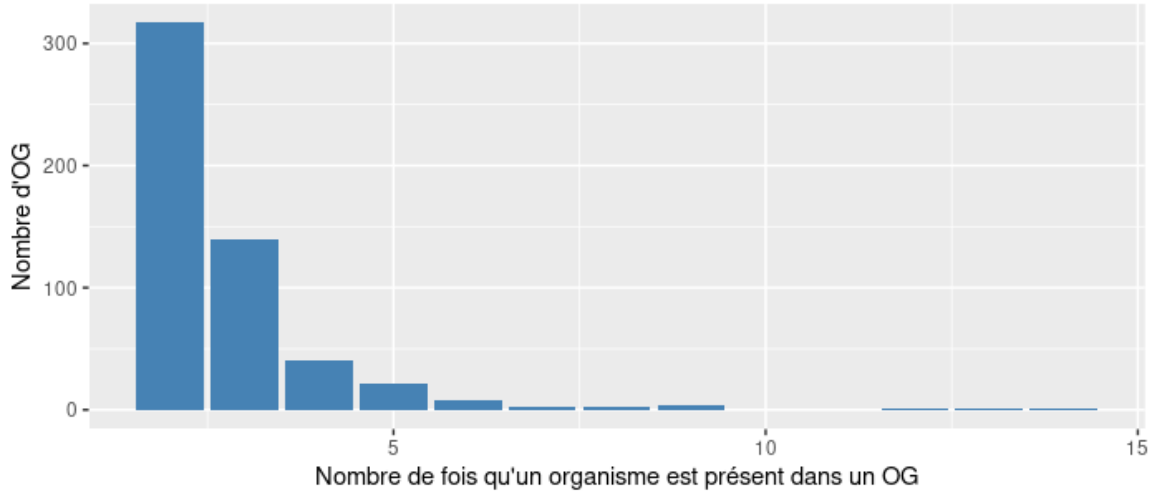


Figure 50. Distribution des doublons au sein des groupes orthologues (Attention qu'il s'agit de la distribution de doublons, donc par défaut l'axe des x commence à 2).
Les données sont fournies **Supp.Mat selection-genes**. Certains organismes ont été rajoutés en plusieurs copies au sein d'un même OG, traduisant de possibles phénomènes de paralogie. *Arabidopsis thaliana* et *Guillardia theta* sont plus susceptibles d'être ajoutées sous la forme de plusieurs paralogues dans nos groupes orthologues.

Puis, pour chaque eucaryote, nous avons compté dans combien de groupes orthologues il est représenté plusieurs fois (**Figure 51**).

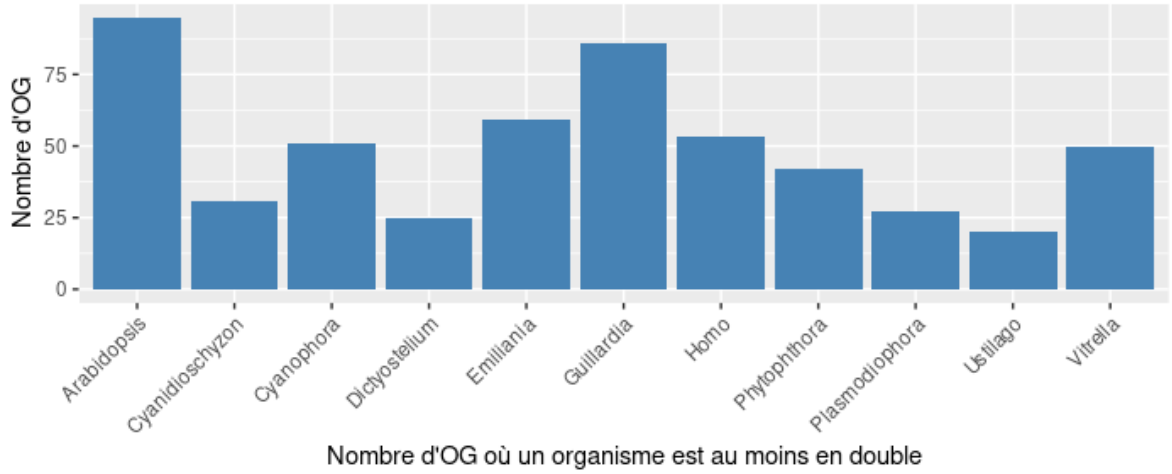


Figure 51. Nombre de groupes orthologues présentant plusieurs occurrences d'un eucaryote donné.
Les données sont fournies **Supp.Mat selection-genes**.

Nous avons également évalué pour chaque groupe orthologue le nombre d'espèces eucaryotes présentes au moins 2 fois (**Figure 52**). Ici, on cherche les gènes où les eucaryotes (en général) sont en double afin de gérer correctement les problèmes de duplications. En effet, selon les cas, il faudra soit sélectionner la bonne copie du gène eucaryote à garder, soit dans le cas où il serait impossible de décider quelle est la bonne, éliminer le gène.

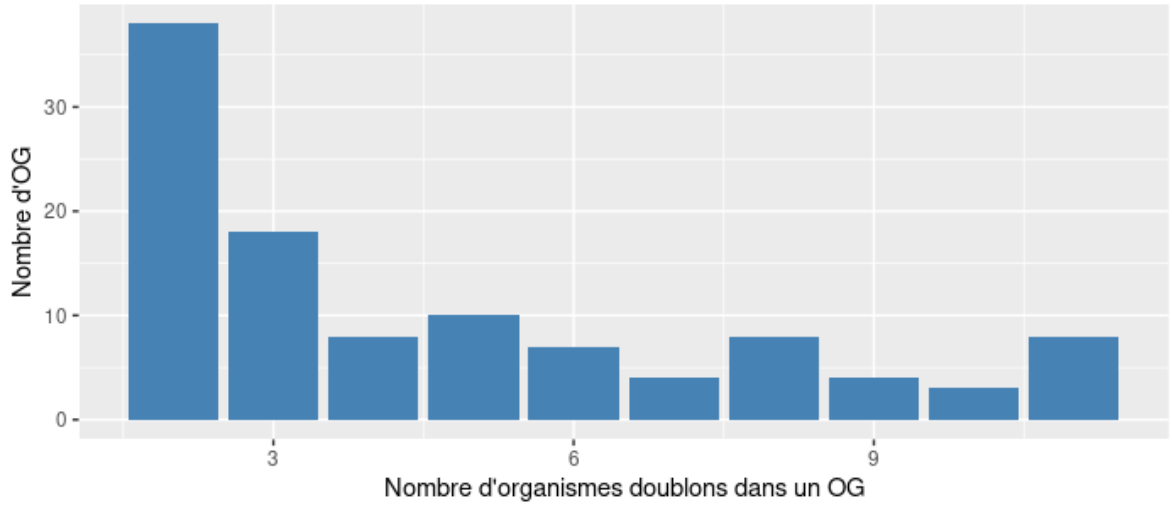


Figure 52. Distribution du nombre d'espèces eucaryotes doublons par groupe orthologue (ex. 10 OG ont 5 eucaryotes en double).
Les données sont fournies **Supp.Mat selection-genes**.

Nous avons ensuite évalué la distribution du nombre d'eucaryotes ajoutés au sein des groupes orthologues (**Figure 53**). Longtemps répartis en Unikontes et Bikontes, les analyses phylogénétiques des super-matrices de protéines montrent généralement un arbre des eucaryotes composé de cinq à huit « super-groupes » qui se répartissent en trois assemblages d'ordre encore plus élevé :

1. Amorphea (Amoebozoa plus Obazoa, ce dernier comprenant les animaux et les champignons) ;
2. Diaphoretickes (principalement Sar, Archaeplastida, Cryptista et Haptophyta) ;
3. Excavata (Discoba et Metanomonada)(Adl et al., 2012; M. W. Brown et al., 2017; Burki et al., 2020; Keeling & Burki, 2019).

La racine eucaryote serait située quelque part entre les Amorphea et le groupe Diaphoretickes + Excavata. Cette racine est surnommée « Opimoda-Diphoda » (Derelle et al., 2015). Nous disposons de 24 alignements où seuls ont été rajoutés des Diaphoretickes, 3 qui n'ont rajouté que des Amorphea et 163 qui ont à la fois des Amorphea et des Diaphoretickes.

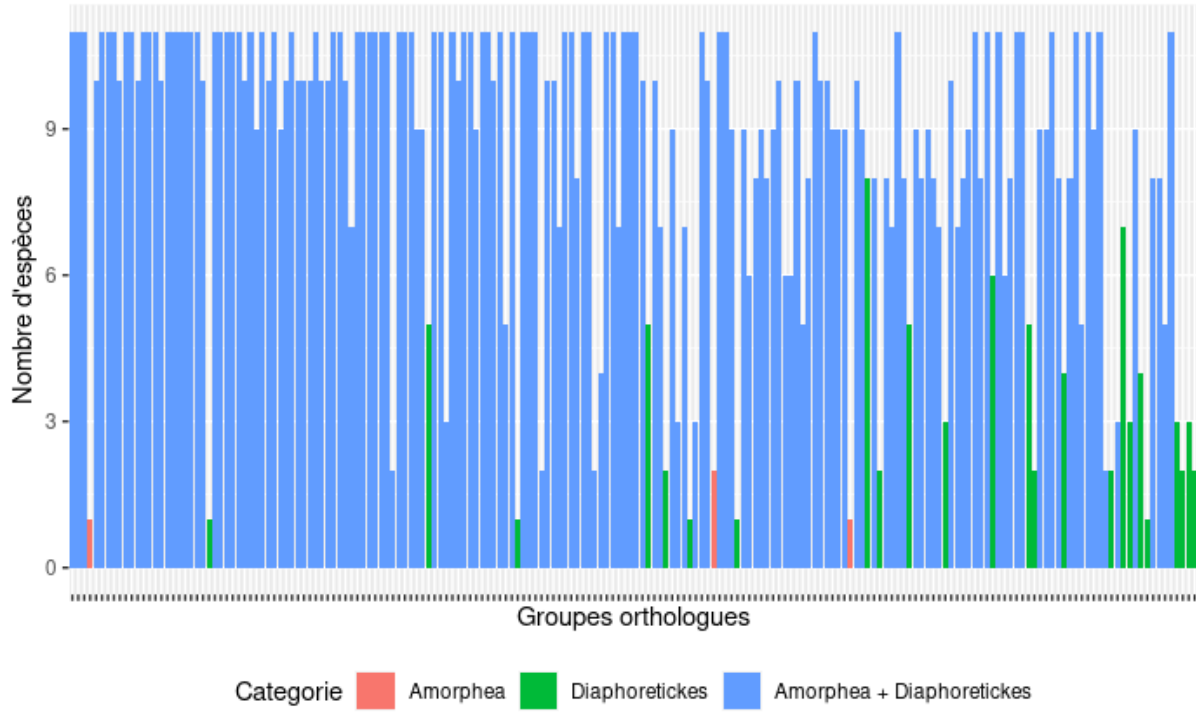


Figure 53. Distribution du nombre d'eucaryotes ajoutés au sein des groupes orthologues.
Les données sont fournies **Supp.Mat selection-genes**.

Nous avons alors cherché à savoir s'il y a des groupes orthologues plus susceptibles de se voir enrichis en eucaryotes. Nous avons pour cela regardé combien de groupes orthologues ont rajouté le même nombre d'eucaryotes (**Figure 54**).

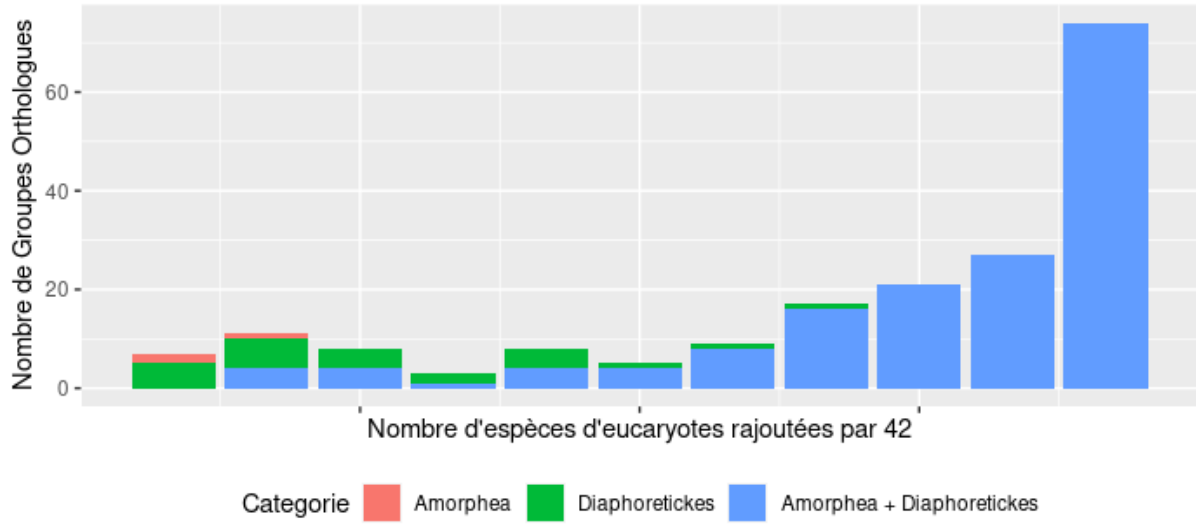


Figure 54. Nombre de groupes orthologues ayant rajouté le même nombre d'eucaryotes.
Les données sont fournies **Supp.Mat selection-genes**. La majorité de nos gènes ont ajouté au moins 9 de nos eucaryotes. Presque 74 gènes ont ajouté tous nos eucaryotes.

Enfin, nous avons évalué la proportion d'Amorphea et de Diaphoretickes pour chaque groupe orthologue, afin de privilégier les groupes orthologues ayant un représentant de chaque groupe (**Figure 55**). En effet, nous cherchons à nous rapprocher au maximum de l'état de LECA. Il

est donc important d'avoir des groupes orthologues dont on peut légitimement inférer qu'ils étaient déjà présents chez LECA.

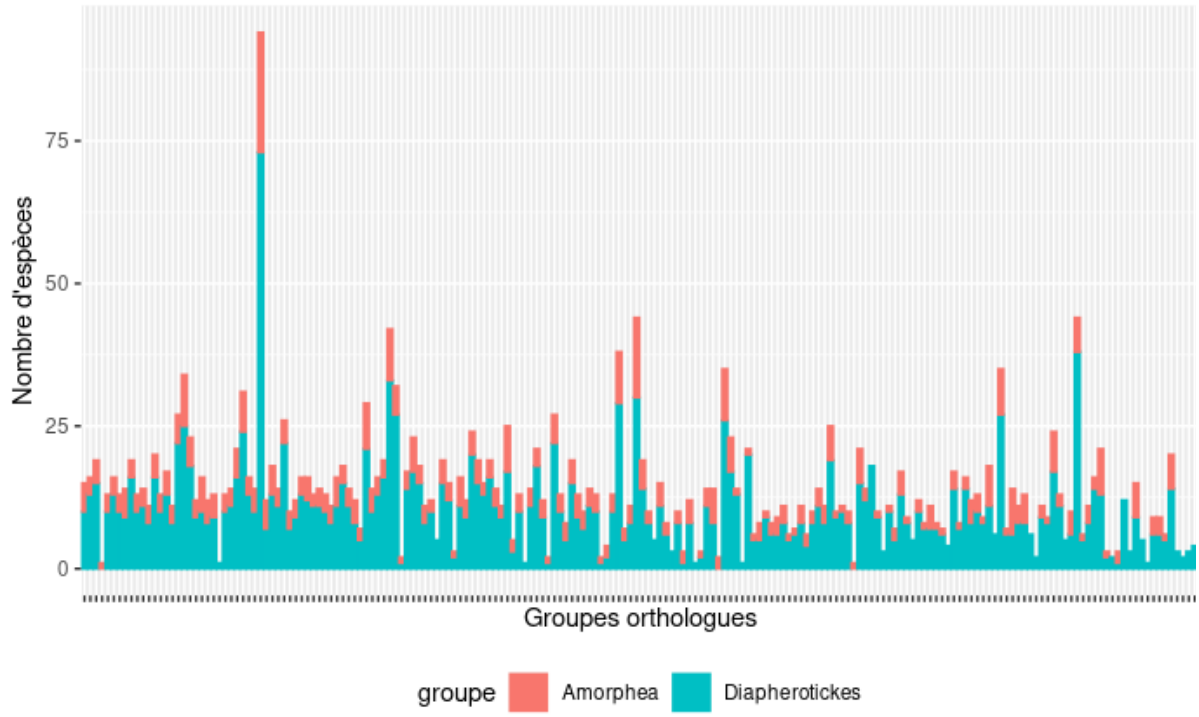


Figure 55. Proportion d'Amorphea et de Diaphoretickes ajoutés par 42 pour chaque groupe orthologue.
Les données sont fournies **Supp.Mat selection-genes**.

125 gènes présentent au moins 10 organismes eucaryotes dans 1 des 5 réplicas et 102 gènes ont au moins 10 organismes eucaryotes dans les 5 réplicas. Nous avons alors supprimé les séquences individuelles (sur base de leur accession) qui n'ont pas été ajoutées dans les 5 réplicas, de telle sorte que ces derniers aient rigoureusement le même jeu de séquences eucaryotes. Nous avons ainsi choisi de garder les groupes orthologues ayant au moins 8 organismes eucaryotes dans les 5 réplicas, soit 138 gènes sur 440.

Nous avons alors calculé avec IQTree et le modèle LG4X des arbres individuels de notre sélection de 138 gènes afin de vérifier la monophylie des eucaryotes et éviter qu'une des séquences rajoutées par 42 ne soit une contamination. Plusieurs cas se sont présentés :

1. les séquences eucaryotes sont monophylétiques en simple copie (50 gènes) (**Figure 56**) ;
2. les séquences eucaryotes sont dupliquées et polyphylétiques dans l'arbre des archées (85 gènes) (**Figure 57**); Dans ce cas, il s'agit soit de séquences isolées qui semblent être soumises à l'artefact d'attraction des longues branches, soit certaines espèces (en particulier *Arabidopsis thaliana*, *Cyanidioschyzon merolae*, *Cyanophora paradoxa*, *Emiliania huxleyi* et *Guillardia theta*.) qui possèdent de nombreux paralogues. L'identification et la suppression de ces séquences résout ce problème. Parmi ceux-ci, 9 arbres possèdent 2 sous-arbres eucaryotes monophylétiques dont il nous a été impossible de trancher objectivement lequel est le bon.

3. les séquences eucaryotes sont en simple copie mais polyphylétiques (3 gènes) (**Figure 58**). Ces gènes ont été systématiquement supprimés (les arbres ont été enracinés dans les eucaryotes par *mid-point root*).

Figure 56. Exemple de gène avec les eucaryotes monophylétiques (OG0000228).
En vert sont surlignés les eucaryotes.

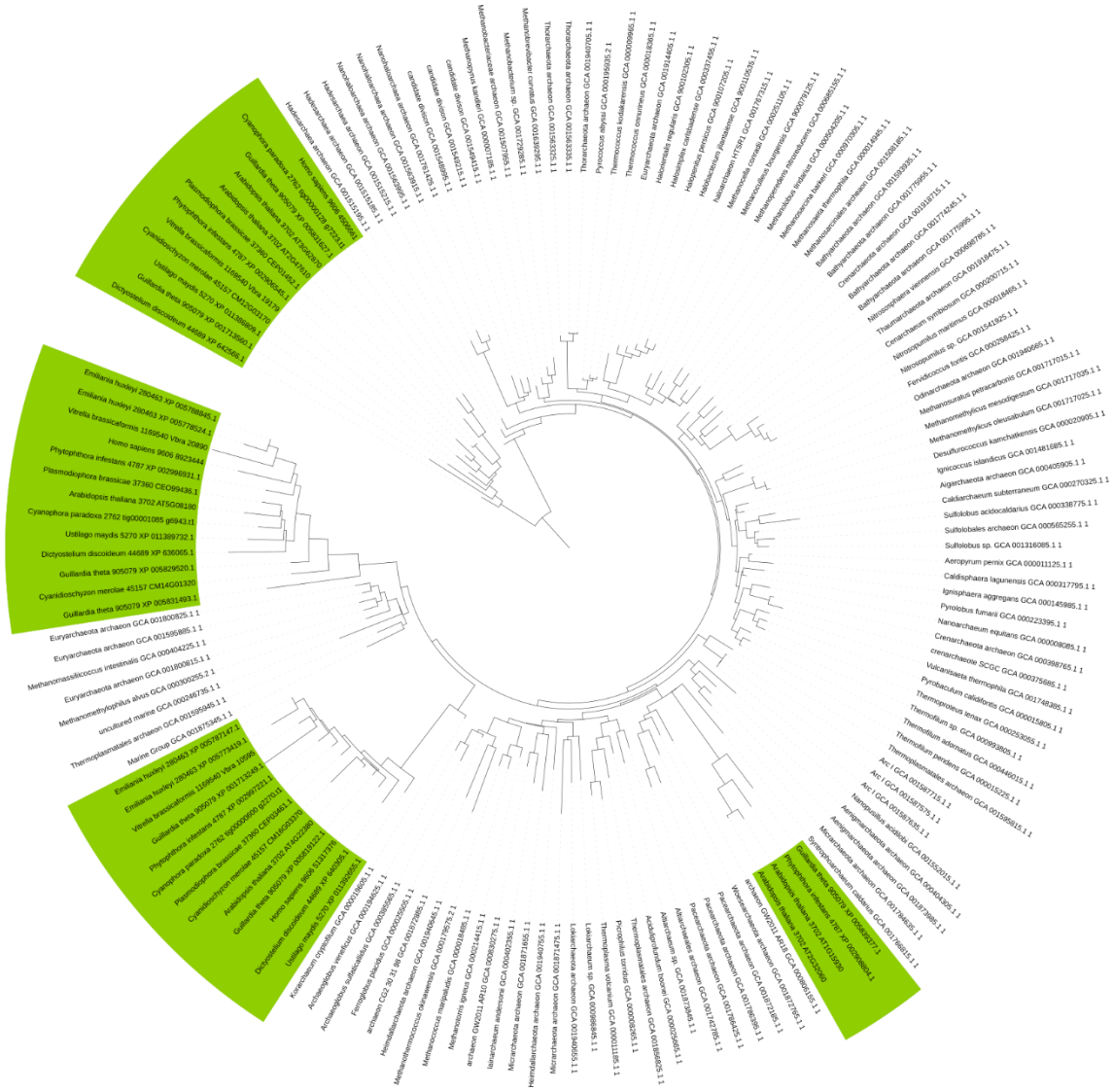


Figure 57. Exemple de gène avec les eucaryotes dupliqués et polyphylétique (OG0000343).
En vert sont surlignés les eucaryotes.

Figure 58. Exemple de gène où les eucaryotes sont polyphylétiques (OG0000773).
En vert sont surlignés les eucaryotes.

Nous avons également observé que l'archée Heimdallarchaeota archaeon GCA001940645.1 pourrait être un métagénome contaminé par un eucaryote. En effet, celui-ci vient s'insérer systématiquement au sein des eucaryotes rendant dès lors les Heimdallarchaeota polyphylétiques.

Après analyse individuelle de chaque arbre, les séquences ont été nettoyées afin de corriger ces problèmes (cf. **Supp.Mat genes-polyphyletiques-causes.txt**). Ainsi, afin de limiter ce qui pourrait être un problème de séquences incorrectes, nous avons éliminé les gènes où il nous était impossible d'obtenir la monophylie des eucaryotes (correspondant à 9 gènes du cas 2 et les 3 gènes du cas 3). Il nous est donc resté 126 gènes.

Nous avons ensuite ajouté avec 42 le génome d'Anaerobic archaeon_MK-D1, une nouvelle archée Asgard récemment séquencée (souche MK-D1 (Imachi et al., 2020)). Contrairement aux

séquences d'Asgard dont nous disposions jusqu'alors, celle-ci n'est pas un métagénome mais un vrai génome issu d'un organisme cultivé. Ce génome est intéressant car il s'agit du seul Asgard à avoir été cultivé. Son séquençage peut ainsi permettre d'évaluer la précision et la complétude des séquençages effectués à partir de métagénomes.

4 RETRAIT D'ESPECES

Nous avons observé depuis le début de notre étude une singularité de Heimdallarchaeota archaeon GCA001940645.1. Déjà lors de notre phylogénie des archées, celle-ci se distinguait de par son incertitude concernant sa position phylogénétique. Dans nos arbres incluant les eucaryotes, cette archée semble se rapprocher plus des eucaryotes que des autres Heimdallarchaeota. Nous soupçonnons que cette singularité soit la conséquence d'une contamination ou d'une nouvelle lignée. Pour évaluer l'impact de ce génome sur notre phylogénie, nous avons créé deux nouvelles super-matrices : une où nous avons retiré Heimdallarchaeota archaeon GCA001940645.1 et une où nous avons retiré tous les Heimdallarchaeota. Puis nous avons calculé des arbres selon la méthode PMSF pour nos 5 réplicas (cf. **Supp.Mat archaea+eukaryota_126_genes**).

Nous n'observons pas de différences concernant la position des eucaryotes dans nos arbres lors du retrait des Heimdallarchaeota. Nos résultats laissent à penser que Heimdallarchaeota archaeon GCA001940645 est un métagénome contaminé par un eucaryote. Ce retrait montre qu'indépendamment des Heimdallarchaeota, la position des eucaryotes dans cette région de l'arbre ne change pas bien que l'on rallonge la branche. Si leur position avait changé, on aurait pu se poser la question de savoir si des signaux opposés se combattent. Mais dans notre cas, l'ajout des Heimdallarchaeota tend à renforcer la position des eucaryotes.

5 EFFET DE NOTRE SELECTION DE GENES ET DE LA PRESENCE DES EUCARYOTES SUR LES TOPOLOGIES D'ARCHEES OBTENUES AU CHAPITRE 2

Nous nous sommes interrogés sur les effets que pourraient engendrer l'ajout d'eucaryotes au sein d'une super-matrice d'archées. Si sur un même réplica avec une même méthode, j'ai un changement topologique entre l'arbre des archées seul et l'arbre avec eucaryote, est-ce dû à la présence des eucaryotes ou au sampling de gène lié au fait qu'on ait mis les eucaryotes ? Pour le savoir, nous avons donc décidé de créer pour nos 5 réplicas une super-matrice de nos 126 gènes dépourvue de nos eucaryotes, puis de calculer l'arbre correspondant selon le PMSF afin de comparer sa topologie à celles obtenues précédemment avec 343 gènes. Nous avons alors calculé les différences entre toutes nos bipartitions des jeux de données avec 343 et 126 gènes afin d'évaluer l'impact de ce sous-échantillonnage de gènes (**Figure 59** & **FigS59**). Si ces différences en valeurs absolues tendent vers 0, alors les topologies coïncident et les différences observées après ajout des eucaryotes pourront être interprétées comme la conséquence de leur ajout. Dans le cas contraire, les différences observées pourront être interprétées comme le résultat d'un cas particulier de sous-échantillonnage de gène, voire un effet combiné avec l'ajout des eucaryotes. Il nous sera alors difficile de trancher. Comme nous l'avons déjà conclu au chapitre précédent, notre sous-échantillonnage de gènes n'affecte que très peu les bipartitions retrouvées, renforçant la stabilité de nos topologies trouvées au chapitre 2. Comme vu, au chapitre précédent, la topologie des archées est surtout dépendante du réplica.

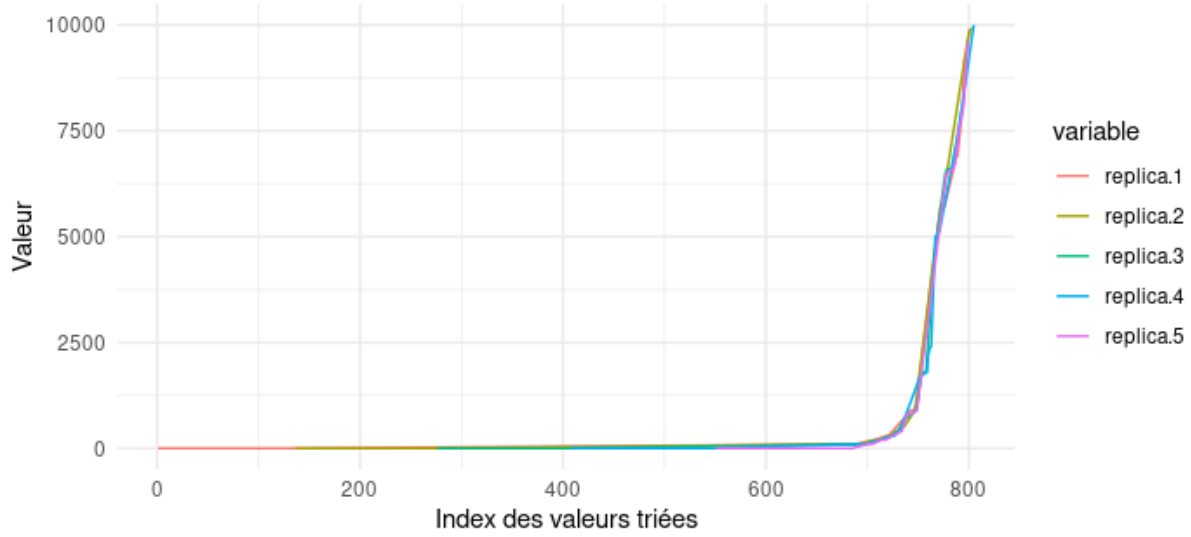


Figure 59. Différences entre les bipartitions de notre sélection de 343 gènes d'archées avec notre sélection de 126 gènes.

Les données sont fournies **Supp.Mat FigS59**. Les valeurs qui tendent vers 0 indiquent que les topologies entre la sélection de 343 gènes et les 126 gènes coïncident. Les différences observées sont interprétées soit comme le résultat d'un cas particulier de sous-échantillonnage de gènes.

Nous avons alors calculé une collection d'arbres à partir de nos super-matrices, cette fois en incluant les eucaryotes, comprenant un arbre LG4X, LG+C20+F+G et LG+C60+F+G et PMSF pour chacun des 5 réplicas. En suivant le même protocole que précédemment en évaluant les bipartitions des archées retrouvées, on se rend compte que dans notre cas, l'ajout des eucaryotes ne change en rien la topologie des archées trouvée au chapitre précédent. Nous retrouvons le même arbre, simplement avec les eucaryotes en plus comme groupe frère du groupe Asgard.

Toutefois, cette solution nous semble encore trop simplistes et nous soupçonnons que cette supposée parenté entre Asgards et eucaryotes soit le fruit de processus sous-jacents cachés et plus complexes. Pour vérifier, nous allons exploiter notre jeu de données en utilisant diverses stratégies.

Pour la suite du protocole, les arbres ont été systématiquement calculés selon la méthode PMSF, avec en arbre guide un arbre calculé selon le modèle LG+C60+F+G.

6 HETEROGENEITE DE SUBSTITUTION AU COURS DU TEMPS

L'un des problèmes pouvant intervenir dans la résolution de l'arbre est la portion de branches ayant des propriétés différentes. Ces branches aux propriétés si différentes du reste de l'arbre pourraient avoir un impact sur la topologie globale de l'arbre en attirant ou repoussant d'autres branches. Ainsi, le signal phylogénétique des substitutions affectant certaines colonnes de nos alignements est supplanté par des processus plus profonds et subtils liés aux hétérogénéités de substitution au cours du temps. Ceux-ci ont très bien pu se produire lors de la transition archée/eucaryote, faussant alors le signal phylogénétique contenu dans chaque colonne, par exemple dans les cas où les proportions de sites invariables d'espèces non apparentées convergent. Deux propriétés peuvent varier :

- le taux de substitution au cours du temps (hétérotachie ou covarion) ;

- le profil, c'est-à-dire la distribution des acides aminés acceptables à une position donnée (hétéropécillie).

6.1 HETEROTACHIE (COVARION) : HETEROGENEITE DE TAUX DE SUBSTITUTION D'UN SITE AU COURS DU TEMPS

Il se peut que, si certains groupes d'archées ou si les eucaryotes ont un taux de substitution très différent des autres espèces de notre arbre, cela conduise à des artefacts de reconstruction phylogénétique. Afin de vérifier cette hypothèse, nous allons inférer des taux d'évolution spécifiques à chaque site indépendamment pour les archées et les eucaryotes. Un calcul des deltas de vitesse nous permettra ensuite de filtrer progressivement les colonnes à éliminer afin de créer des super-matrices ayant des colonnes avec des vitesses d'évolution plus ou moins hétérogènes. Puis nous allons calculer des arbres en introduisant progressivement des sites au fur et à mesure que ceux-ci évoluent plus différemment.

Nous avons utilisé les alignements issus du jeu de données où nous avons retiré les gènes où les eucaryotes sont polyphylétiques. Les colonnes ayant au moins un représentant de chaque grand groupe taxonomique d'archées ont été retenues, à savoir : un eucaryote, un DPANN, un euryarchaeota, un TACK, un Asgard. Nous avons ainsi 126 gènes, correspondant à environ 33 000 positions pour chaque réplica. Le but est ainsi de maintenir par la suite des super-matrices d'archées et d'eucaryotes de taille identique tout en jouant sur le sous-ensemble d'archées utilisées pour calculer les taux. Nous avons alors divisé les super-matrices en deux, afin de séparer d'un côté les archées et de l'autre côté les eucaryotes. Les taux de chaque position ont été calculés séparément pour les super-matrices eucaryotes et archées avec IQ-Tree, puis les deltas entre chaque site ont été calculés entre ces deux groupes. On a utilisé 4 bins (une par catégorie gamma) car les taux étaient une version approximative de ces catégories.

Nous avons calculé pour chaque site de nos alignements la différence de taux de substitution entre nos eucaryotes et nos archées (pris soit dans leur ensemble, soit après le retrait de certains de nos sous-groupes mentionnés précédemment) (cf. **rates**). Nous avons par la suite également procédé à des manipulations de retrait d'espèces sur les super-matrices d'archées, en supprimant alternativement les Euryarchaeota, les DPANN, les TACK et Heimdallarchaeota archaeon GCA001940645 (qui pour rappel tend à s'insérer systématiquement au sein des eucaryotes). Puis ces différences ont été triées en allant des sites ayant la plus petite (homotachie) à la plus grande (hétérotachie). Cela nous permet de discriminer des sites ayant de forts écarts de taux de substitution entre archées et eucaryotes des sites ayant de plus faibles écarts. Les résultats sont donnés **Figure 60** & **Supp.Mat FigS60**. On n'observe pas de différences dans la distribution des différences de taux de substitutions selon l'ensemble d'archées utilisé. Par conséquent, il n'y a pas de variation de taux de substitution pour certains sous-groupes. L'allure de distribution globale des deltas ne semble pas affectée par le sous-échantillonnage d'archées, mais cela ne veut pas dire que la composition des bins est identique en termes de sites de la super-matrice originale. Donc même si on ne s'attend à priori pas à des variations, nous allons le vérifier en calculant les différents arbres avec toutes les espèces en alternant des fichiers de taux de substitution différents correspondant à nos différents ensembles d'archées. Nous allons donc pouvoir utiliser l'ensemble de nos archées comme référence afin de créer des sous-super-matrices, sans se soucier de variations qui auraient pu être intrinsèques à certains sous-groupes. De plus, il ne semble pas y avoir de différences selon nos réplicas.

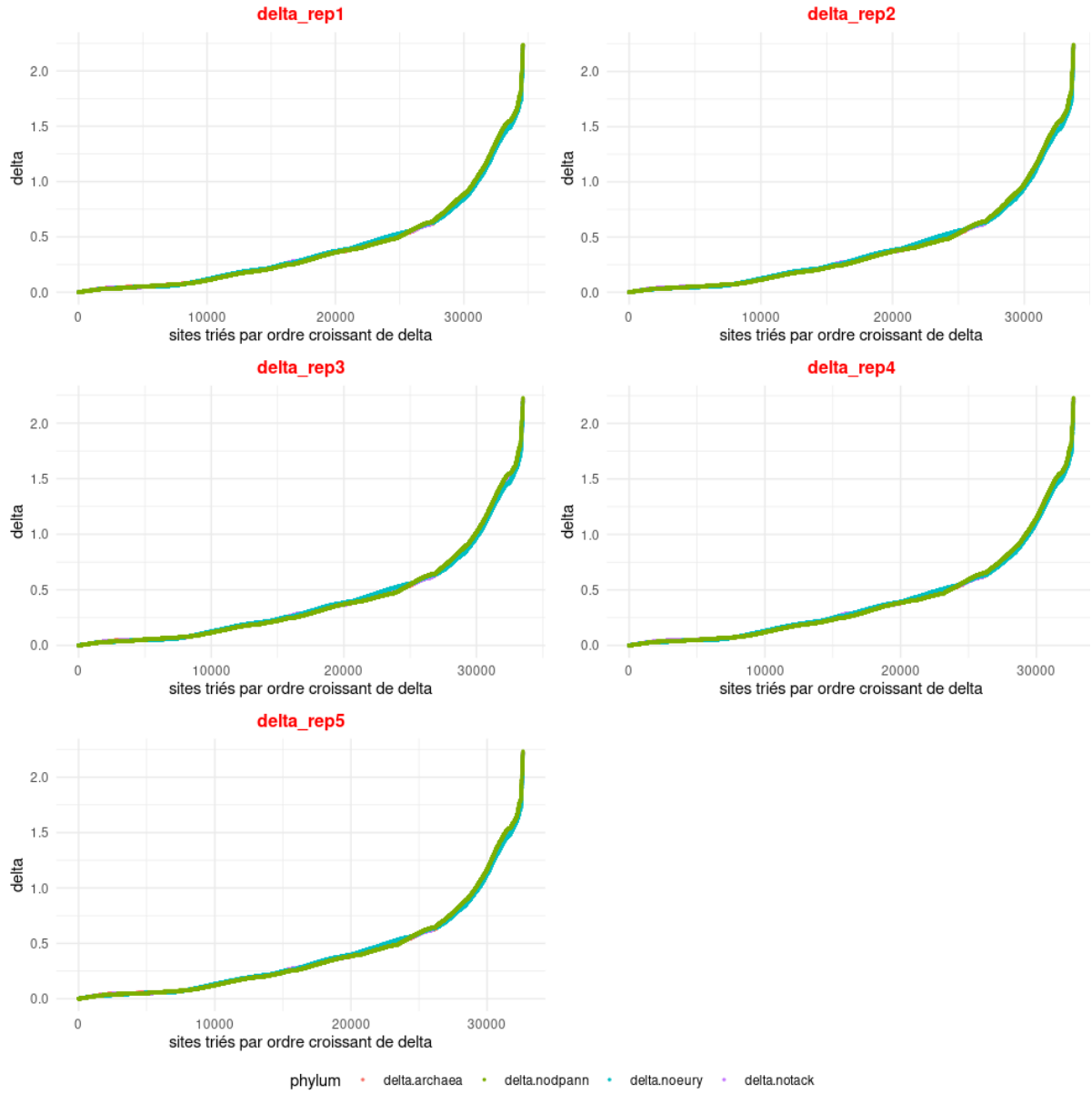


Figure 60. Variation du taux de substitution par site entre nos différents ensembles d'archées et nos eucaryotes.

Les données sont fournies **Supp.Mat FigS60**. L'allure de distribution globale des deltas ne semble pas affectée par le sous-échantillonnage d'archées, mais cela ne veut pas dire que la composition des bins est identique en termes de sites de la super-matrice originale. De plus, il ne semble pas y avoir de différences selon nos réplicas.

Nos super-matrices font entre 30 000 et 35000 positions. Nous avons choisi de créer des bins cumulatifs d'environ 6 000 positions basés sur ces taux de substitution, contenant toutes les espèces (à la fois les eucaryotes et les archées). Deux types de bins ont été créés. Dans un premier temps, nous avons utilisé des bins cumulatifs qui ajoutent progressivement des colonnes au fur et à mesure que les deltas entre taux de substitution augmentent. Dans un deuxième temps, nous n'avons pas cumulé les bins afin d'observer des arbres construits à partir d'alignements ayant des deltas de taux de substitution différents. Ces arbres ont été calculés en LG4X. Ce modèle a été choisi pour 2 raisons : (1) d'abord on veut utiliser un modèle qui sera sensible aux artefacts afin de les identifier plus facilement, et (2) nous sommes soumis à des limitations de temps de calcul.

On observe qu'on a toujours cette même Heimdallarchaeota archaeon GCA001940645.1 qui pose problème (cf. fichier **parse-consense**).

Les analyses de nos bins cumulés et non cumulés (**Figure 61, Figure 62** & **Supp.Mat heterotachie**) montrent que pour les sites les plus homogènes en termes de taux entre archées et eucaryotes, le regroupement des Asgards auprès des eucaryotes n'est que très peu représenté (moins de 2 500 arbres sur 10 000), au profit d'un regroupement des Asgards auprès des TACK (plus de 7 500 arbres sur 10 000). Il faut atteindre une taille de super-matrices de 18 000 positions avec l'ajout de sites à taux de substitution plus hétérogène pour que les Asgards soient systématiquement regroupés avec les eucaryotes. On peut dès lors supposer que le regroupement des Asgards avec les eucaryotes puisse être le fruit d'un biais de reconstruction phylogénétique lié à des sites présentant de l'hétérotachie mal modélisée. Autrement dit, ce regroupement fortuit serait la conséquence d'une erreur systématique produisant un signal non-phylogénétique à la fois faux et statistiquement supporté entrant en compétition avec le véritable signal phylogénétique.

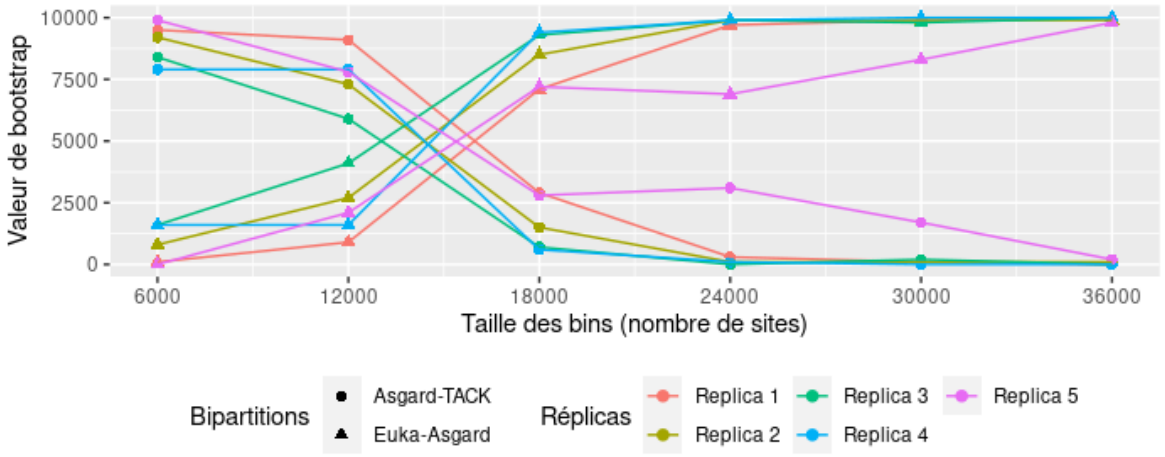


Figure 61. Proportion d'arbres indiquant la position des Asgards selon la taille des bins cumulés pour chaque réplica selon le modèle LG4X.

Les données sont fournies **Supp.Mat heterotachie**. Il faut atteindre une taille de super-matrices de 18 000 positions avec l'ajout de sites à taux de substitution plus hétérogène pour que les Asgards soient systématiquement regroupés avec les eucaryotes. On peut dès lors supposer que le regroupement des Asgards avec les eucaryotes puisse être le fruit d'un biais de reconstruction phylogénétique lié à des sites présentant de l'hétérotachie mal modélisée.

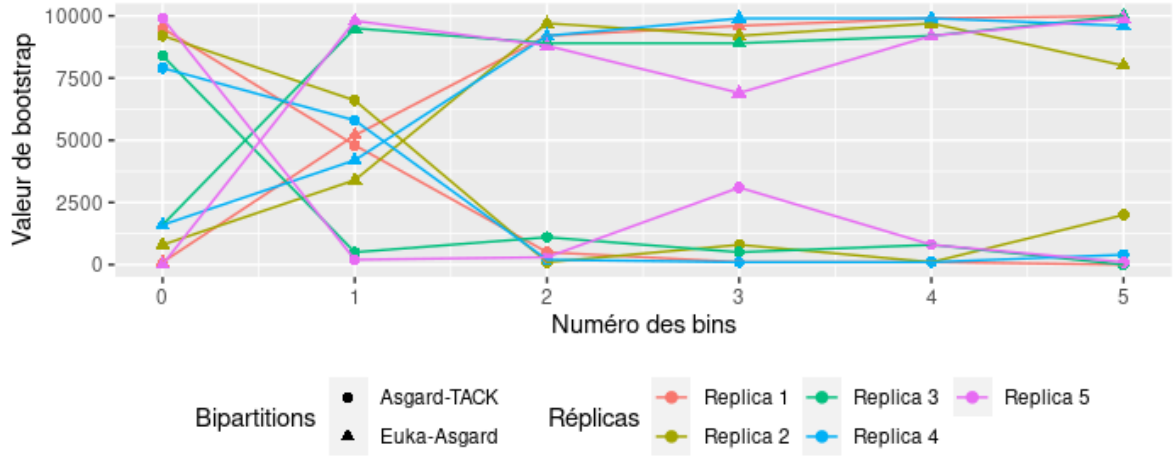


Figure 62. Proportion d'arbres indiquant la position des Asgards selon la taille des bins non cumulés pour chaque réplica selon le modèle LG4X.

Les données sont fournies **Supp.Mat heterotachie**. Les bins les plus homogènes tendent à favoriser les Asgards comme groupe frère des TACK, tandis que les sites plus hétérogènes soutiennent l'hypothèse des eucaryotes comme groupe frère des Asgards.

Deux exemples de représentation d'arbres phylogénétique correspondant aux deux topologies possibles entre eucaryotes, TACK et Asgard sont donnés **Figure 63** et **Figure 64**.

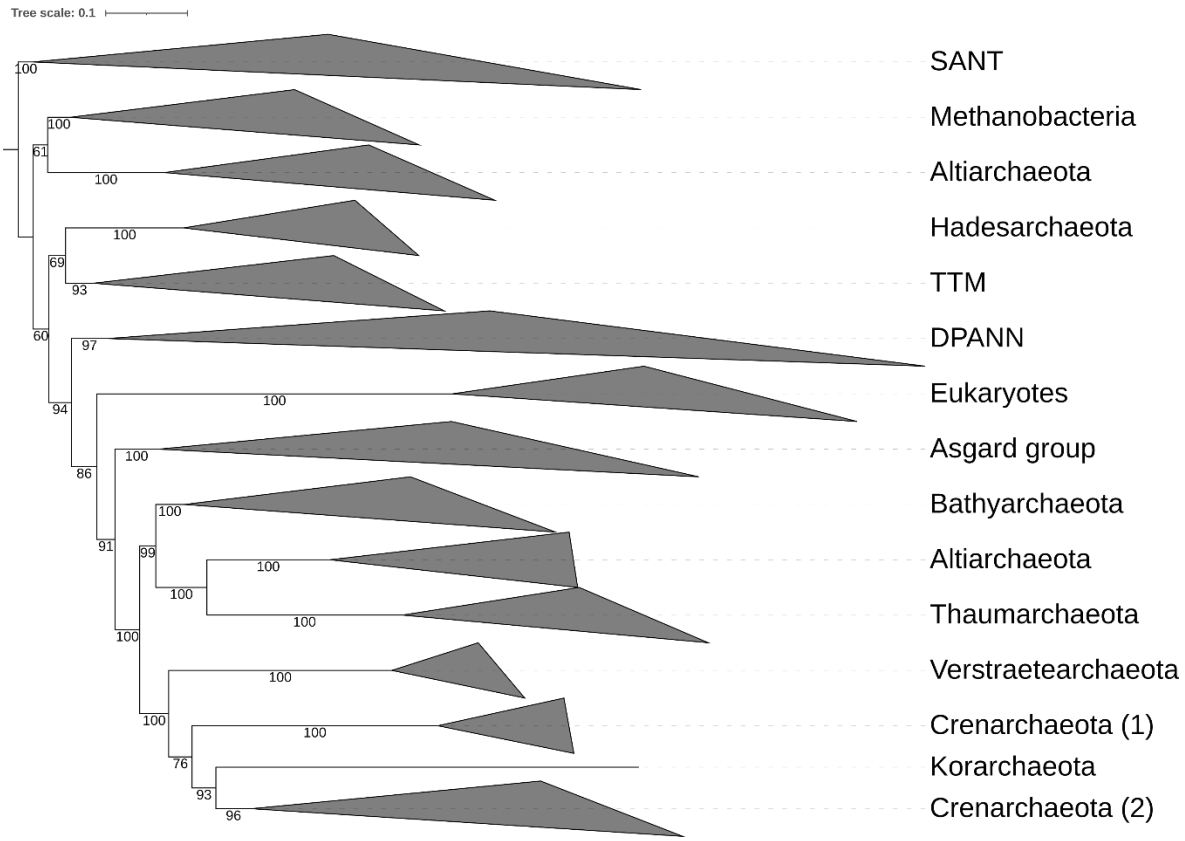


Figure 63. Arbre correspondant aux bins cumulés à 12 000 positions pour le réplica 1 (132 espèces) selon le modèle LG4X.

Les données sont fournies **Supp.Mat heterotachie**. Dans cet arbre, les eucaryotes sont à la base d'un groupe comprenant les Asgard et les TACK, qui se retrouvent groupes frères.

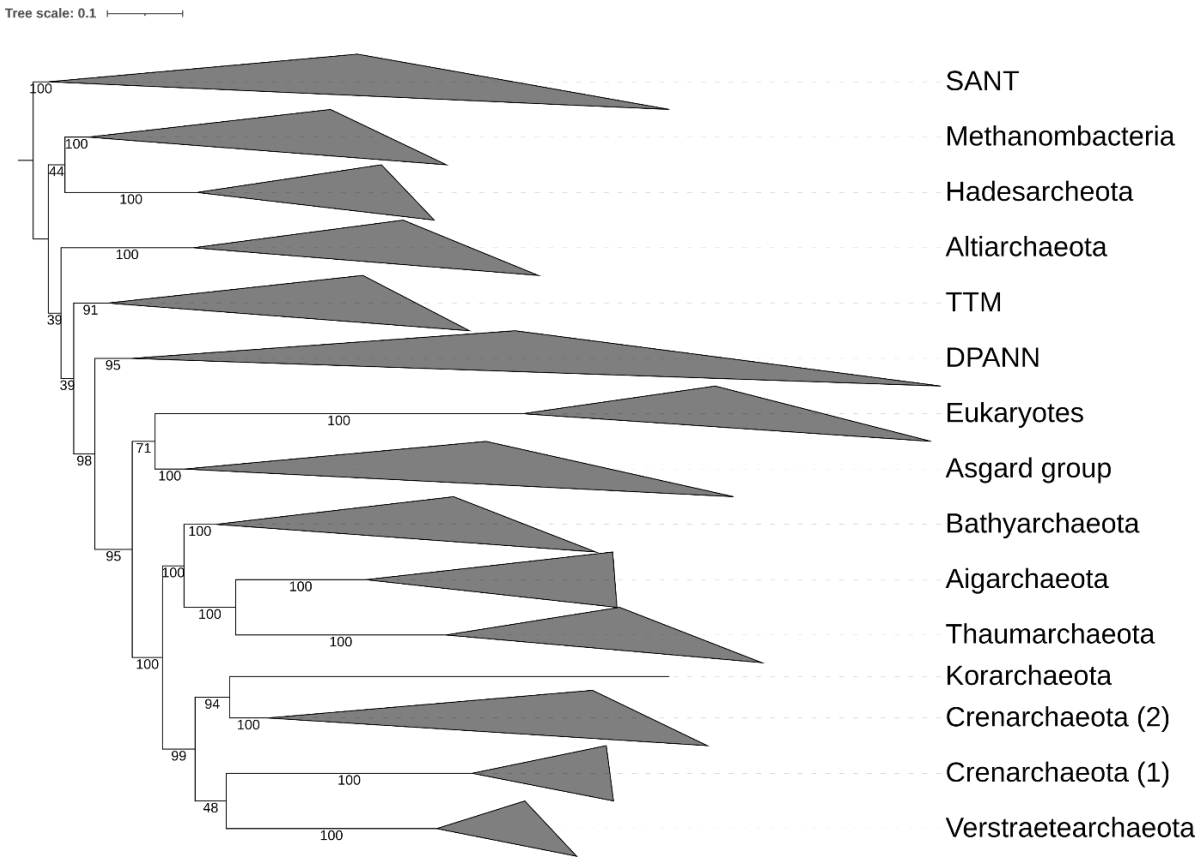


Figure 64. Arbre correspondant aux bins cumulés à 18 000 positions pour le réplica 1 (132 espèces) selon le modèle LG4X.

Les données sont fournies **Supp.Mat heterotachie**. Dans cet arbre, les eucaryotes sont le groupe frère du groupe Asgard.

Nous avons également construit de nouveaux arbres selon le modèle PMSF à partir de bins cumulés dont les vitesses ont été déterminées après le retrait alternativement des DPANN (**Figure 65**), puis des Euryarchaeota (**Figure 66**) et enfin des TACK (**Figure 67**). Dans tous les cas, on voit bien que le rapprochement des eucaryotes avec les Asgards résultent de l'utilisation de sites hétérotaches, favorisant notre interprétation que cette hypothèse est erronée et soit le fruit d'un artefact.

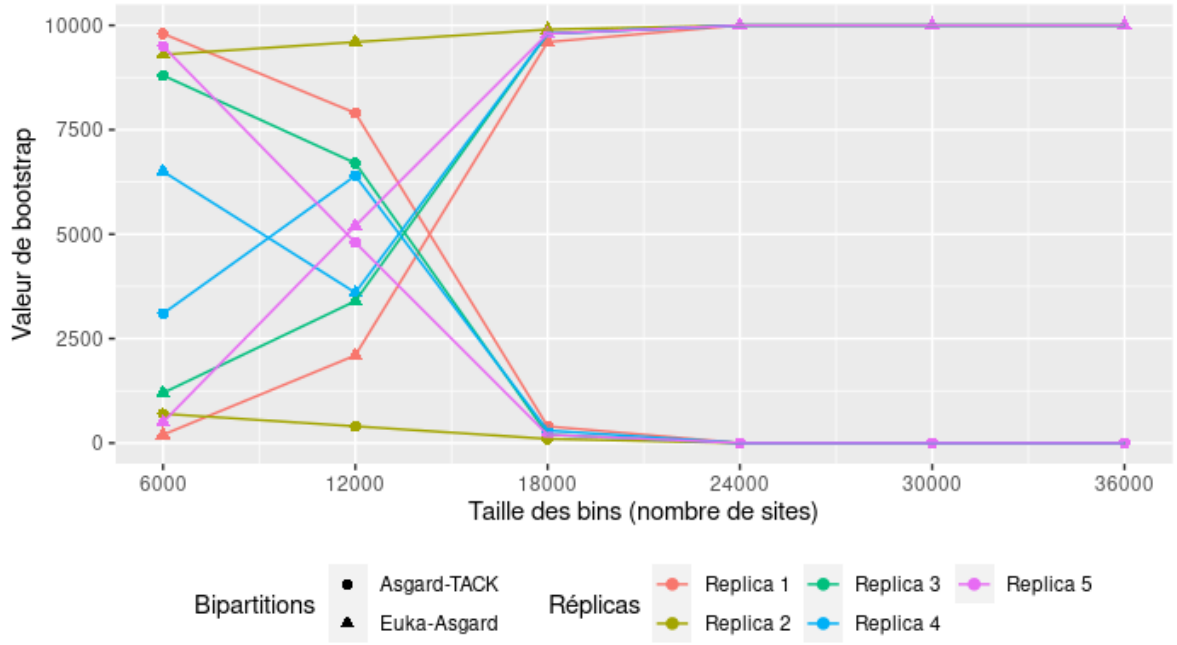


Figure 65. Proportion d'arbres selon le modèle PMSF à partir de bins cumulés dont les vitesses ont été déterminées après le retrait des DPANN.

Les données sont fournies **Supp.Mat heterotachie**. Avec le modèle PMSF, nous obtenons le même résultat que le modèle LG4X, à savoir qu'il faut atteindre une taille de super-matrices de 18 000 positions avec l'ajout de sites à taux de substitution plus hétérogène pour que les Asgards soient systématiquement regroupés avec les eucaryotes.

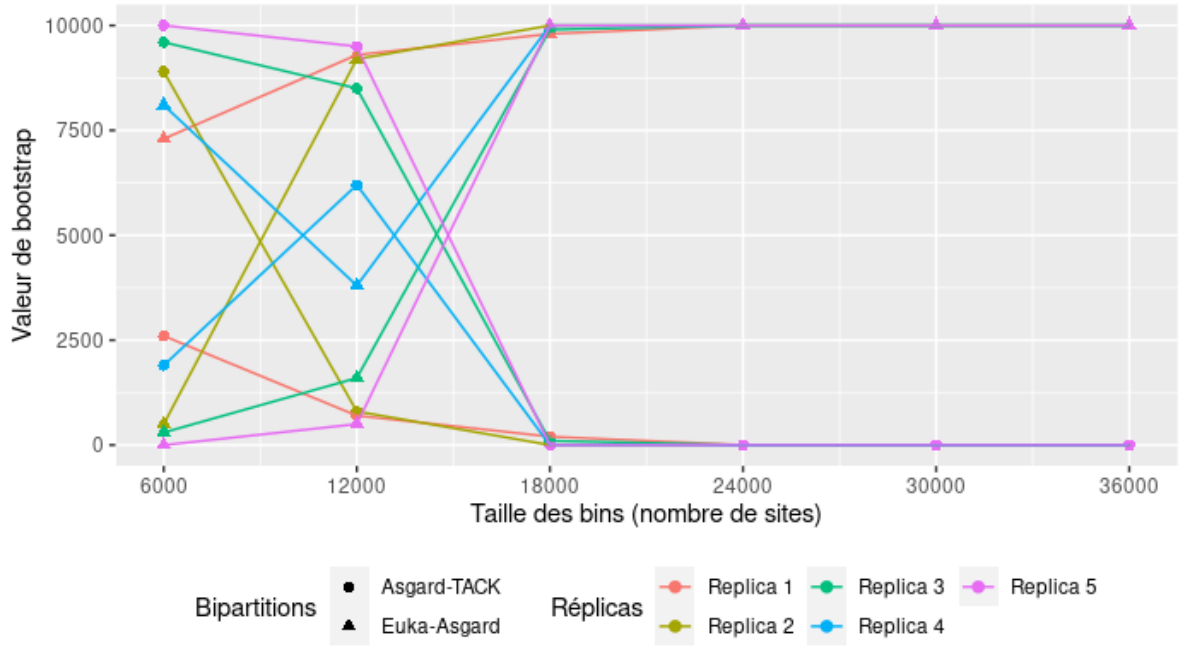


Figure 66. Proportion d'arbres selon le modèle PMSF à partir de bins cumulés dont les vitesses ont été déterminées après le retrait des Euryarchaeota.

Les données sont fournies **Supp.Mat heterotachie**. Encore une fois, les bins les plus homogènes tendent à favoriser les Asgards comme groupe frère des TACK, tandis que les sites plus hétérogènes soutiennent l'hypothèse des eucaryotes comme groupe frère des Asgards. Seul le réplica 1 soutient l'hypothèse des eucaryotes comme groupe frère des Asgards. Le réplica 4 est en revanche difficile à interpréter et semble dépendre du bin utilisé.

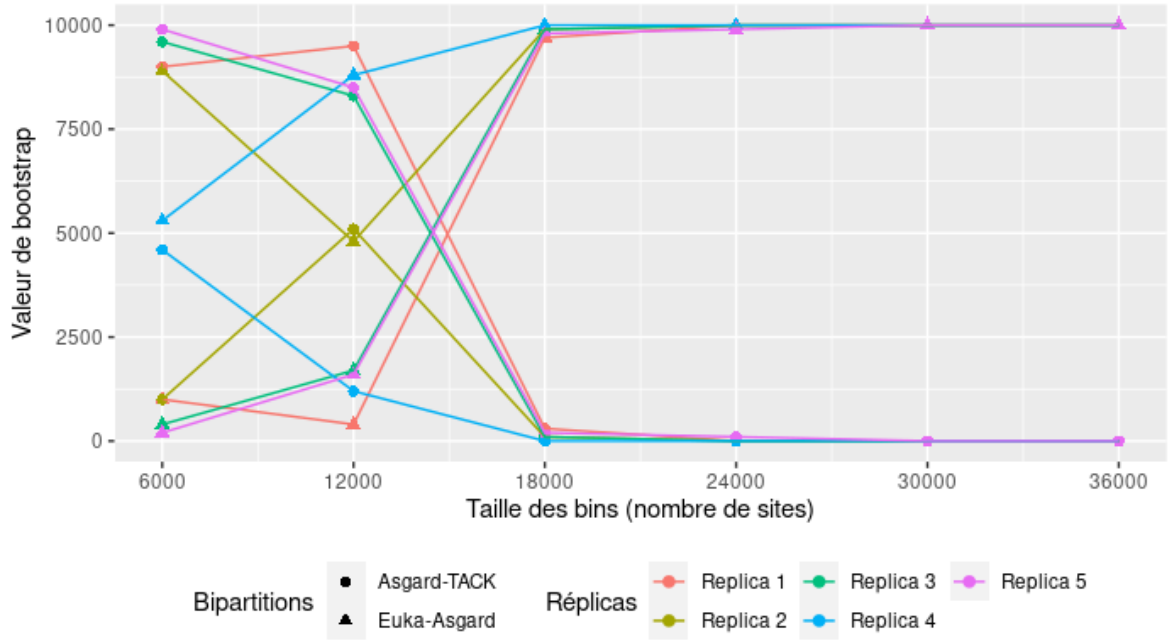


Figure 67. Proportion d'arbres selon le modèle PMSF à partir de bins cumulés dont les vitesses ont été déterminées après le retrait des TACK.

Les données sont fournies **Supp.Mat heterotachie**. Encore une fois, les bins les plus homogènes tendent à favoriser les Asgards comme groupe frère des TACK, tandis que les sites plus hétérogènes soutiennent l'hypothèse des eucaryotes comme groupe frère des Asgards. Seul le réplica 1 soutient l'hypothèse des eucaryotes comme groupe frère des Asgards. Le réplica 4 est en revanche difficile à interpréter et semble dépendre du bin utilisé.

6.2 HETEROPECILLIE : HETEROGENEITE DES PROCESSUS DE SUBSTITUTION D'UN SITE AU COURS DU TEMPS

Enfin, nous avons voulu vérifier si tous les acides aminés sont potentiellement possibles pour une colonne donnée. Pour nos 5 réplicas d'espèces, nous avons estimé séparément pour les archées et les eucaryotes les paramètres de modèles de mixtures, puis déduit le profil de fréquence spécifique à chaque site. Ces profils sont basés, pour chaque site, sur un Khi-carré comparant les distributions de fréquences des acides aminées entre les super-matrices. Nous avons ensuite généré 6 bins cumulatifs par réplica et calculé les arbres selon la méthode PMSF (**Figure 68** & **Supp.Mat heteropecillie**). Malgré tout, nos résultats de montrent pas de différence entre nos bins. Les eucaryotes sont systématiquement groupe frère des Asgards avec des valeurs d'ultrafast-bootstrap maximales quels que soit le réplica et le bin.

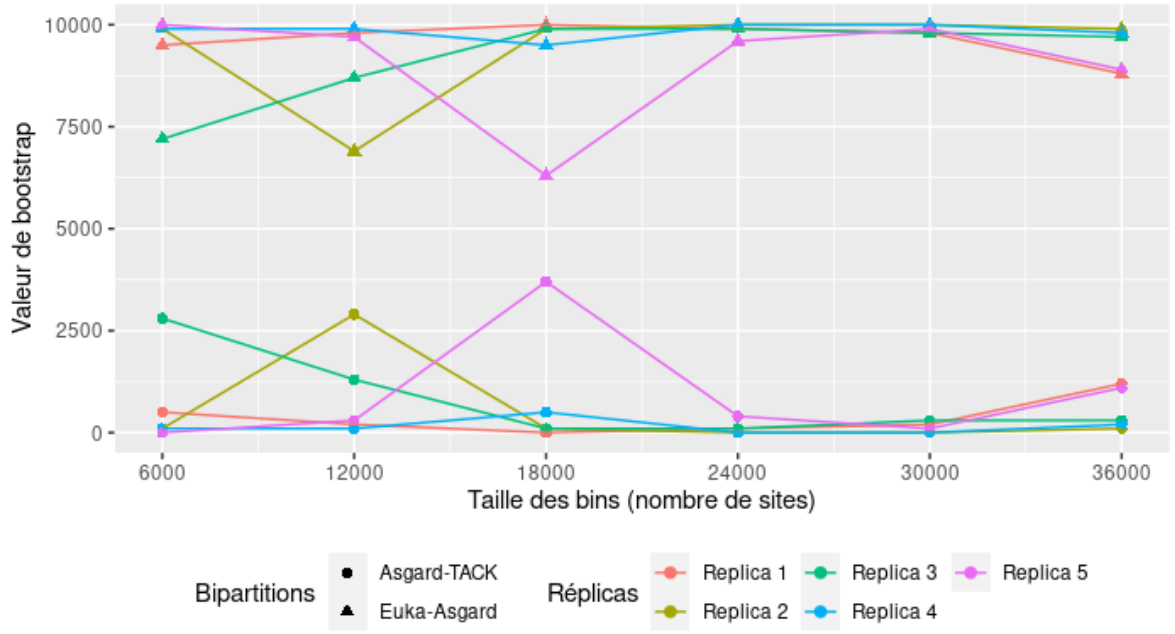


Figure 68. Proportion d'arbres indiquant la position des Asgards selon la taille des bins cumulés pour chaque réplica (132 espèces) selon le modèle PMSF.

Les données sont fournies **Supp.Mat heteropecillie**. Les eucaryotes sont systématiquement groupe frère des Asgards.

Nous avons alors voulu vérifier que tous les bins donnaient bien le même résultat pris séparément. Pour cela, nous avons décidé de générer des bins non cumulatifs et de calculer des arbres selon la méthode PMSF, ce dernier étant plus apte à combattre les artefacts mais théoriquement plus sensible à l'hétéropécillie (**Figure 69**). Cette fois, nos résultats montrent que le dernier bin ne retrouve pas la monophylie eucaryotes + Asgards mais favorise plutôt les Asgards comme groupe frère des TACK (les valeurs d'ultrafast-bootstrap du dernier bin pour ce regroupement sont supérieures à 90 %). Il apparait donc que le PMSF ne résiste pas au phénomène d'hétéropécillie. Le réplica (et donc la sélection d'espèces) et le modèle semblent aussi jouer un rôle non négligeable. Cela montre que l'hétéropécillie est un phénomène important à prendre en compte lorsque l'on décide d'insérer les eucaryotes au sein des archées.

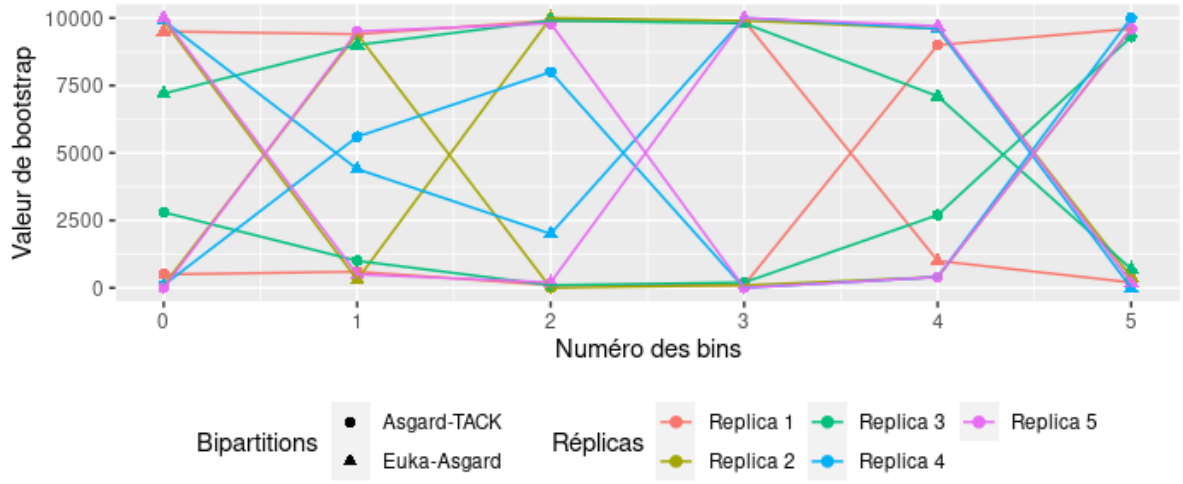


Figure 69. Proportion d'arbres indiquant la position des Asgards selon les bins non cumulés pour chaque réplica (132 espèces) avec la méthode PMSF.

Les données sont fournies **Supp.Mat heteropecillie**. Les eucaryotes sont regroupés avec les Asgards pour tous les bins sauf le dernier.

7 ENRACINEMENT DES ARCHEES

Suite à mes analyses phylogénétiques des arbres d'Archaea, il est crucial de procéder à l'enracinement de ces arbres pour mieux comprendre les relations évolutives et déterminer l'origine commune des différents groupes. Jusqu'à présent, mes phylogénies étaient non enracinées, ce qui limite l'interprétation des données en termes d'évolution temporelle et directionnelle. L'objectif est maintenant d'enraciner ces arbres en utilisant des méthodes robustes pour assurer la précision et la fiabilité des résultats.

Nous avons utilisé deux méthodes : l'AU-test et le rootstrap. L'AU-test (*Approximately Unbiased test*) est une méthode statistique permettant d'évaluer la confiance des hypothèses d'enracinement en comparant différents arbres enracinés hypothétiques. Cette approche utilise des simulations pour estimer la distribution des arbres possibles et calcule les probabilités de chaque hypothèse d'enracinement. Cela permet de sélectionner l'arbre le plus probable parmi plusieurs candidats, renforçant ainsi la validité de l'enracinement proposé. Le rootstrap est une méthode complémentaire qui combine le bootstrap traditionnel avec des techniques spécifiques d'enracinement pour estimer la stabilité de l'enracinement à travers des rééchantillonnages des données. Cette technique permet de tester la robustesse de l'enracinement en générant de multiples ensembles de données bootstrap, puis en observant la fréquence à laquelle un enracinement particulier est obtenu. Une haute fréquence d'enracinement dans les ensembles de données bootstrap indique une forte confiance dans la position de la racine. Nous avons appliqué ces méthodes à deux jeux de données distincts :

- un jeu de données composé uniquement de séquences d'archées. Cela permettra d'obtenir une première estimation de la racine sans influence externe.
- un jeu de données incluant les eucaryotes pour évaluer l'impact de ces séquences sur les résultats d'enracinement. Comparer les arbres enracinés obtenus à partir des deux jeux de données permettra de déterminer si l'inclusion des eucaryotes introduit des artefacts ou modifie significativement les résultats.

Nous avons choisi, pour des raisons de limitations de temps de calcul, de travailler exclusivement sur la super-matrice du réplica 2 (33 959 positions), qui, comme vu précédemment, nous semble être la topologie d'arbre la plus crédible. L'utilisation du AU-test sans eucaryotes avec la p-value la moins rejetante plus élevée (0.94) enracine l'arbre au sein des Heimdallarchaeota (**Figure 70**). Dans ce scénario, les Asgards sont paraphylétiques. L'évolution des archées se ferait alors par simplification secondaire. Ce scénario pourrait être en accord avec un scénario à 3 domaines indépendants. Archées et eucaryotes partageraient un ancêtre commun déjà complexe, et l'évolution des archées se serait faite par simplification secondaire. Cela expliquerait la forte proximité phylogénétique entre le groupe Asgard et les eucaryotes. Ce scénario pourrait également tout à fait être compatible avec le scénario 1D (D. P. Devos, 2021) qui propose lui aussi un ancêtre commun aux archées et eucaryotes déjà complexe. On retrouve ensuite deux sous-groupes correspondant aux Euryarchaeota et aux TACK. Les DPANN sont groupe frère des Altiarchaeota et s'insèrent donc au sein des Euryarchaeota.

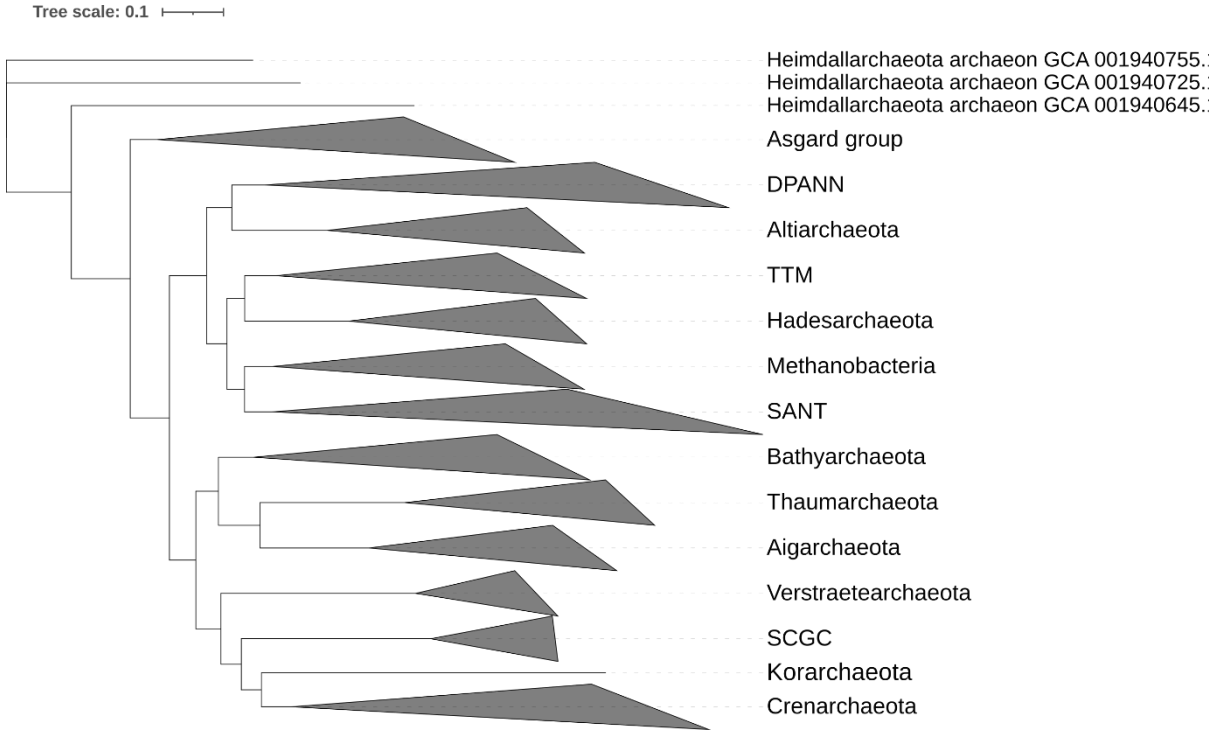


Figure 70. Arbre obtenu ayant la plus forte p-value (0.94) après un AU-test en absence des eucaryotes sur la super-matrice du réplica 2 (33 959 positions, 121 espèces).
Les données sont fournies **Supp.Mat enracinement**. Les Asgards sont paraphylétiques et à la base des autres archées. Dans un tel scénario, l'évolution des archées se seraient fait par simplification secondaire, l'ancêtre des archées étant un être déjà complexe.

Parmi les autres racines possibles de notre ensemble de confiance, l'AU-test sans les eucaryotes nous donne 26 arbres possibles dont 6 racines possibles :

- Heimdallarchaeota (2 cas, avec la p-value à 0.94)
- Altiarchaeota/DPANN (p-value = 0.47)
- SANT, pouvant être paraphylétiques ou inclure les Methanobacteria comme groupe frère (p-values valant respectivement 0.14, 0.70, 0.69, 0.09, 0.19)
- DPANN, ceux-ci pouvant ou non être paraphylétiques (p-values valant respectivement 0.22, 0.22, 0.17, 0.05, 0.06)

- TACK (3 cas, les p-values valant respectivement 0.21, 0.17, 0.13, 0.12)
- Crenarchaeaota + Korarchaeota, les TACK devenant paraphylétiques (4 cas, les p-values valant respectivement 0.16, 0.13, 0.08, 0.07)
- Bathyarchaeota (3 cas, les p-values valant respectivement 0.06, 0.16 et 0.16)
- enfin les archées pourraient être divisés en deux groupes frères correspondant aux Euryarchaeota et aux autres archées (2 cas, les p-values valant respectivement 0.77 et 0.49).

En revanche, l'ajout des eucaryotes lors du AU-test modifie drastiquement la position de la racine pour faire apparaître deux groupes : les Euryarchaeota et un grand ensemble DPANN + Asgard + Eucaryotes + TACK (p-value = 0.76) (**Figure 71**). Les DPANN ne sont donc plus ici membres des Euryarchaeota mais sont à la base du groupe Asgard + Eucaryotes + TACK. Nous avons également une paraphylie des Heimdallarchaeota, avec les eucaryotes qui viennent s'y insérer. Toutefois, on notera la longueur de la branche à la base des eucaryotes qui pourrait aussi laisser penser à un artefact d'attraction des longues branches (taille de branche : 0.38).

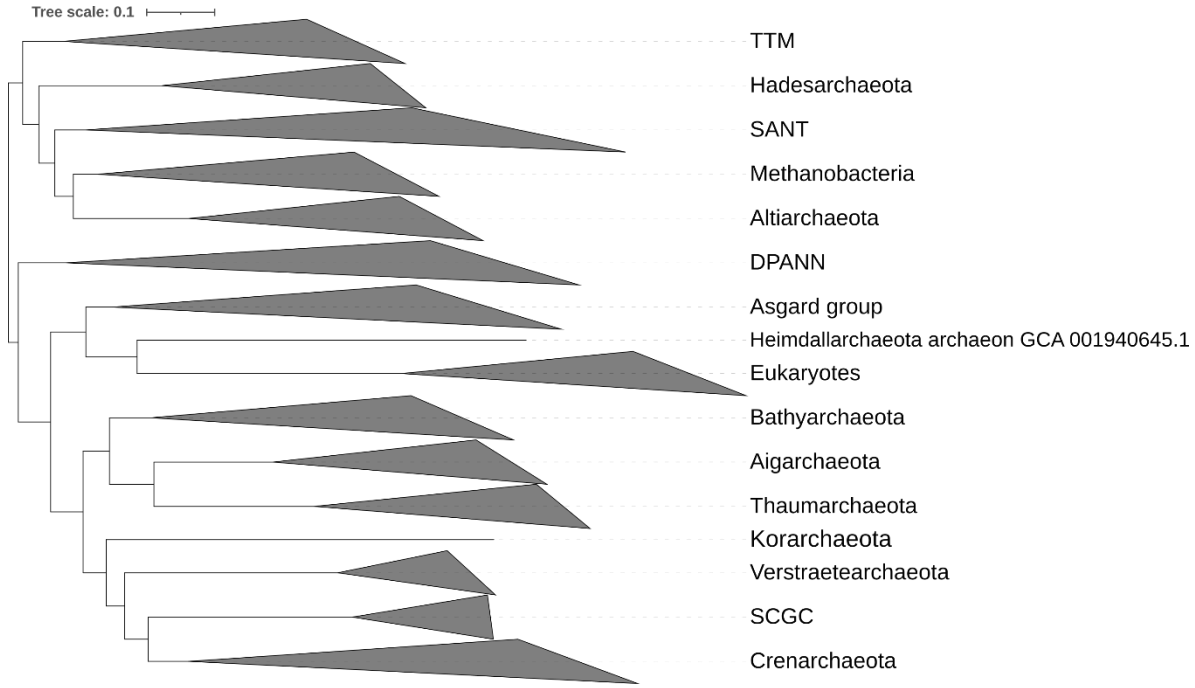


Figure 71. Arbre obtenu ayant la plus forte p-value (0.76) après un AU-test en présence des eucaryotes sur la super-matrice du réplica 2 (33 959 positions, 132 espèces).
Les données sont fournies **Supp.Mat enracinement**. L'ajout des eucaryotes favorise dans ce cas-là deux groupes d'archées distincts : d'un côté les Euryarchaeota, qui sont dès lors monophylétiques, et de l'autre l'ensemble des autres archées, eucaryotes inclus.

Parmi les autres racines possibles de notre ensemble de confiance, l'AU-test avec les eucaryotes nous donne 12 arbres possibles dont 6 racines possibles :

- les TTM (p-values = 0.05)
- les Thermoplasmatales (2 cas, les p-values valant respectivement 0.37 et 0.19)
- les DPANN (p-values = 0.37)
- les Altiarchaeota (3 cas, les p-values valant respectivement 0.41, 0.12, 0.08)
- les Stenosarchaea + Archaeoglobi, rendant notre groupe SANT paraphylétique (p-value = 0.14).

– enfin les archées pourraient être divisés en 2 groupes frères correspondant aux Euryarchaeota et aux autres archées (4 cas, les p-values valant respectivement 0.76, 0.68, 0.41, 0.19), avec les DPANN étant soit au sein des Euryarchaeota, soit à la base de toutes les autres archées. De plus, dans un cas, les Hadesarchaea et le groupe TTM sortent des Euryarchaeota, les rendant paraphylétiques, pour se retrouver à la base des DPANN et des autres archées.

Lorsqu'on utilise la méthode de rootstrap, que ce soit avec ou sans eucaryotes, la racine se situe au sein du groupe SANT (**Figure 72** & **Figure 73**). Les Euryarchaeota, devenant paraphylétiques, n'existent alors plus comme groupe valide. Les DPANN, comme avec le AU-test, s'insèrent à la base du groupe (Asgard + Eucaryotes + TACK). Les Verstraetearchaeota deviennent le groupe frère des SCGC. En revanche, la position des Altiarchaeota change avec l'ajout des eucaryotes, se rapprochant d'une partie du groupe TTM, rendant ce dernier polyphylétique. De même, les Asgard deviennent paraphylétiques, les Heimdallarchaeota étant alors le groupe frère des eucaryotes. De même qu'avec le AU-test, la longue branche des eucaryotes (taille de branche : 0.39) pourrait conduire à une position liée à un phénomène d'attraction des longues branches.

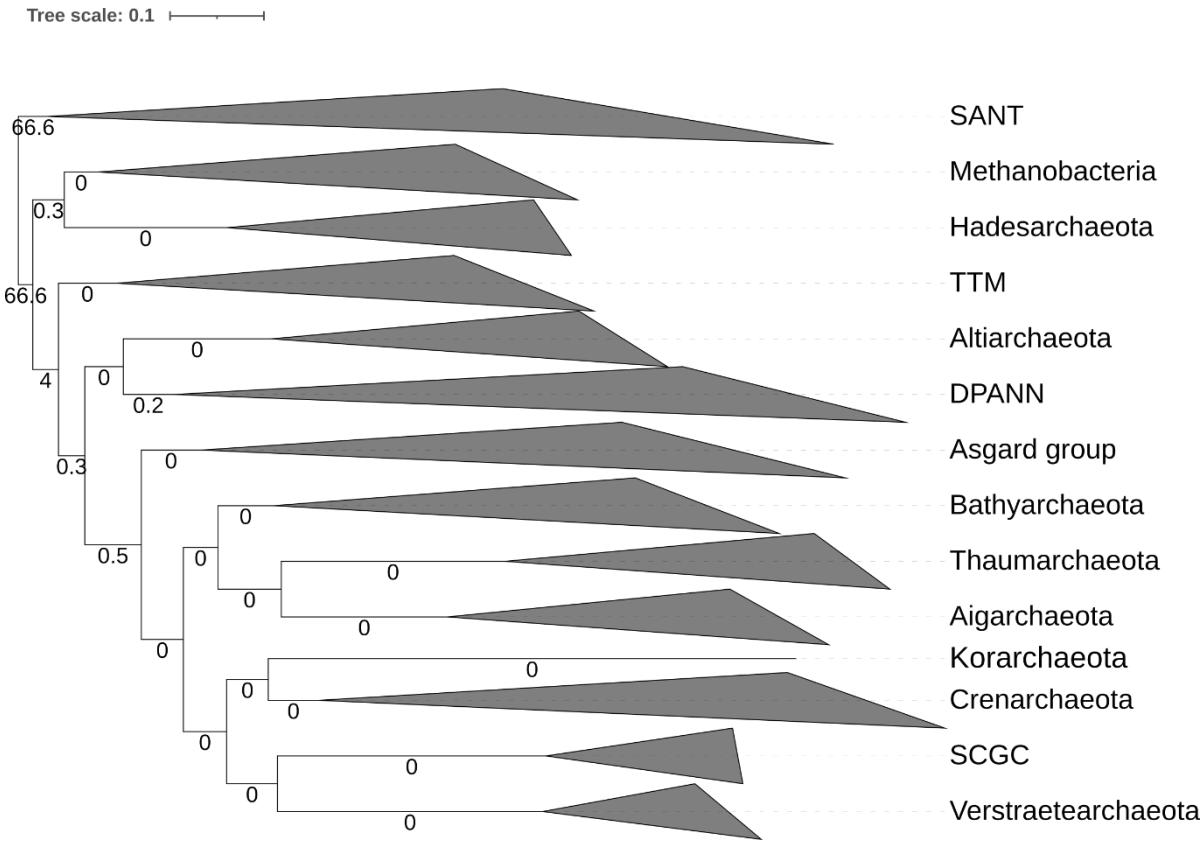


Figure 72. Arbre obtenu après un rootstrap en absence des eucaryotes sur la super-matrice du réplica 2 (33 959 positions, 121 espèces).

Les données sont fournies **Supp.Mat enracinement**. Cet arbre enracine les archées au sein du groupe SANT, rendant les Euryarchaeota paraphylétiques.

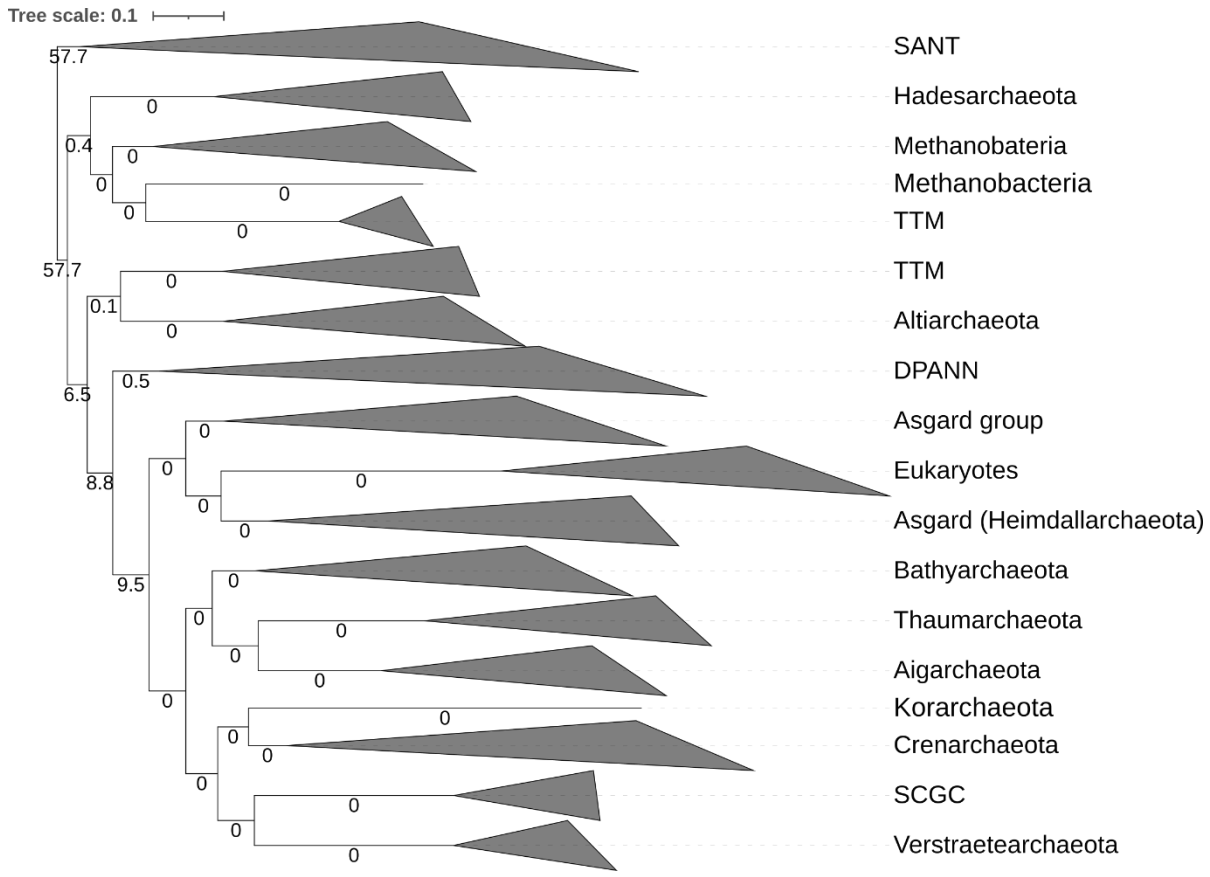


Figure 73. Arbre obtenu après un rootstrap en présence des eucaryotes sur la super-matrice du réplica 2 (33 959 positions, 132 espèces).

Les données sont fournies **Supp.Mat enracinement**. Cet arbre enracine les archées au sein du groupe SANT, rendant les Euryarchaeota paraphylétiques. Les Altiarchaeota se rapprochent de certains membres du groupes TTM, rendant ce dernier polyphylétique.

8 INFERENCE PHYLOGENETIQUE

Après avoir effectué des analyses phylogénétiques en utilisant IQ-TREE et la méthode PMSF sur un jeu de données contenant uniquement des Archaea et un autre jeu incluant des Archaea et des Eucaryotes, nous avons choisi de compléter ces analyses en utilisant le modèle CAT pour effectuer des inférences bayésiennes avec PhyloBayes. En effet, le modèle CAT (CATégories) est particulièrement bien adapté pour capturer l'hétérogénéité des fréquences d'équilibre des acides aminés à travers les sites d'une séquence. En refaisant les analyses avec le modèle CAT et PhyloBayes, nous pouvons comparer les résultats avec ceux obtenus précédemment via IQ-TREE et la méthode PMSF. Cette comparaison permettra de vérifier la consistance des résultats et d'identifier d'éventuels artefacts introduits par les méthodes ou modèles précédents. Ces analyses d'inférence bayésienne ont été réalisées à l'aide de PhyloBayes MPI 1.9a (Lartillot et al., 2013) selon les modèles CAT-G et CAT-GTR-G.

Conformément à nos résultats d'enracinement, nous avons choisi d'enraciner nos arbres au sein du groupe SANT. Avec nos deux modèles (CAT-G et CAT-GTR-G), nous retrouvons le même groupe Ouranosarchaea (TACK + Asgard + eucaryotes). Leur phylogénie est la même (**Figure 74** & **Figure 75**), et nous retrouvons également le même résultat avec notre analyse PMSF, à savoir les eucaryotes comme groupe frère du groupe Asgard. Dans nos deux modèles, deux chaînes sur 4 présentent une paraphylie des Asgard, avec soit toutes les Heimdallarchaeota soit

Heimdallarchaeota GCA001940645.1 qui devien(nen)t groupe frère des eucaryotes. Concernant la position des Korarchaeota, que nous avons discutée avec le modèle PMSF, nous retrouvons majoritairement une position des Korarchaeota comme groupe frère des Crenarchaeota. Cependant, avec le modèle CAT-G, nous retrouvons une chaîne (sur quatre) qui positionne les Korarchaeota à la base du groupe TACK. Avec le modèle CAT-GTR-G, sur quatre chaînes, une chaîne positionne les Korarchaeota à la base des TACK et une autre les positionne au sein des TACK à la base des SCGC, Crenarchaeota et Verstraetearchaeota. Concernant les Hadesarchaeota, ceux-ci sont systématiquement à la base des Ouranosarchaea avec le modèle CAT-G et trois chaînes sur quatre du modèle CAT-GTR-G. Ce résultat diffère des analyses effectuées avec IQ-TREE, qui privilégient les DPANN à cette position. Jusqu'à présent, nous avions les Hadesarchaeota comme groupe interne aux Euryarchaeota. Or nos résultats d'enracinement ne semblent pas valider les Euryarchaeota comme un groupe monophylétique valide. Ce résultat n'a été retrouvé que lors de l'AU-test en présence des eucaryotes. De plus, lors de nos analyses PMSF sur nos réplicas d'espèces, seul le réplica 4 obtenait les Hadesarchaeota à la base des Ouranosarchaea. L'hypothèse alternative positionne les Hadesarchaeota au sein des Euryarchaeota. On retrouve ensuite les DPANN, qui peuvent avoir comme groupe frère les Altiarchaeota. Si l'on enracine dans le groupe SANT nos arbres, les Methanobacteria sont systématiquement le groupe que l'on trouve juste après dans l'arbre, suivis du groupe TTM. Les incertitudes restent les positions des groupes DPANN, Altiarchaeota et Hadesarchaeota qui s'insèrent entre les Ouranosarchaea et les Methanobacteria.

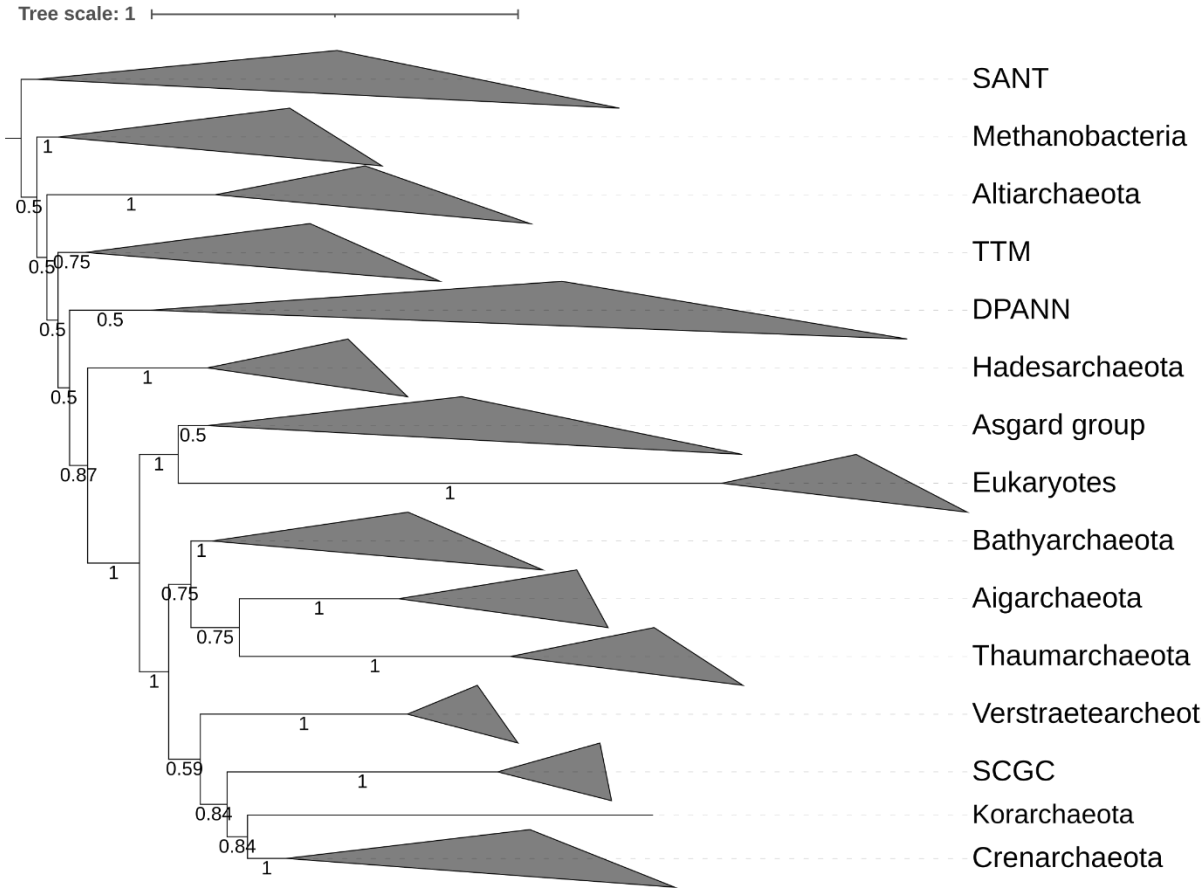


Figure 74. Arbre consensus calculé avec PhyloBayes selon le modèle CAT-G sur le réplica 2 (132 espèces).

Les données sont fournies **Supp.Mat phyloBayes**. Les eucaryotes sont groupe frère des Asgards. Les Hadesarchaeota sont systématiquement à la base des Ouranosarchaea, contrairement aux analyses effectuées en PMSF avec IQ-TREE qui favorisait les DPANN à cette position.

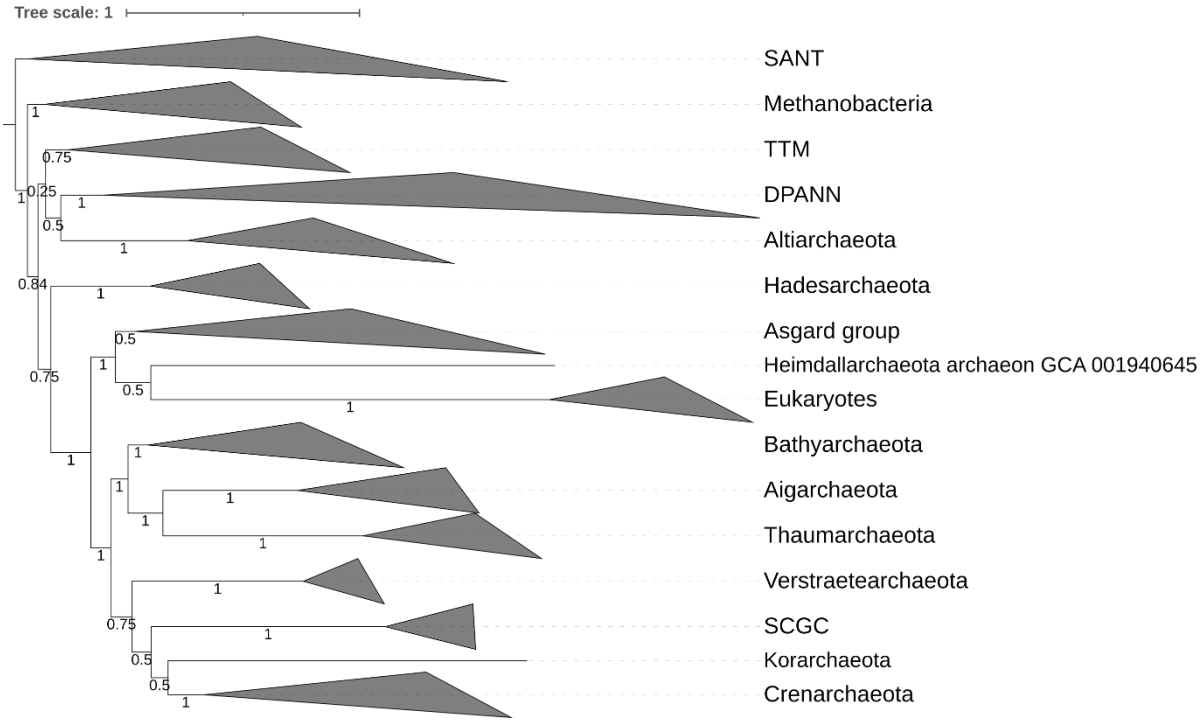


Figure 75. Arbre consensus calculé avec PhyloBayes selon le modèle CAT-GTR-G sur le réplica 2 (132 espèces).

Les données sont fournies **Supp.Mat phyloBayes**. Les eucaryotes sont groupe frère des Asgards. Les Hadesarchaeota sont à la base des Ouranosarchaea pour 3 de nos 4 chaînes, contrairement aux analyses effectuées en PMSF avec IQ-TREE qui favorisait les DPANN à cette position.

Plusieurs facteurs, notamment des variations dans les modèles (C60 vs CAT), les logiciels (PhyloBayes vs IQ-TREE) et les détails de mise en œuvre (par exemple, le nombre de catégories utilisées pour tenir compte de l'hétérogénéité des taux (= catégories gamma) pourraient expliquer la nouvelle variation des résultats observée ici parmi les modèles hétérogènes au niveau des sites. Le nombre de catégories déduites par CAT dans PhyloBayes peut être très élevé. Cela nécessite un très grand nombre de paramètres supplémentaires estimés. Enfin, on notera la longueur de branche importante des eucaryotes (CAT-G : 1.48 ; CAT-GTR-G : 1,71), dont la position peut donc faire l'objet d'un artefact d'attraction des longues branches.

9 DISCUSSION

La cellule eucaryote présente de nombreuses caractéristiques apparemment uniques, et il existe un écart considérable entre les cellules procaryotes et eucaryotes en termes de complexité et de développement. Malgré cet écart apparent, peu de caractéristiques sont véritablement eucaryotes uniquement, ce qui suggère une contribution mixte des deux autres domaines. En effet, la cellule eucaryote est une mosaïque évolutive composée de caractéristiques bactériennes et

archéennes, ainsi que d'innovations eucaryotes. La découverte des archées d'Asgard (Eme et al., 2017; Spang et al., 2015, 2017; Zaremba-Niedzwiedzka et al., 2017) a relancé un débat intense sur la topologie de l'arbre universel du vivant. Il a été suggéré que la ramification des eucaryotes au sein des archées dans certaines reconstitutions phylogénétiques soit le produit d'artefacts de reconstruction phylogénétiques (notamment l'attraction des longues branches) associés à une forte divergence ancienne (D. P. Devos, 2021). Les partisans d'un arbre à deux domaines primaires (2D, hypothèse Eocyte) proposent un scénario dans lequel les eucaryotes sont apparus au sein des archées, plus précisément en étant un sous-groupe du super-phylum Asgard, tandis que d'autres soutiennent un arbre dans lequel les trois domaines (3D) sont monophylétiques. Nous nous sommes placés dans notre étude dans un modèle 2D afin de tester la supposée parenté entre les Asgards et les eucaryotes. Malgré tout, un scénario à trois domaines pourrait rester envisageable (**Figure 6**).

De plus, il est possible que l'existence de transferts horizontaux de gènes entre les Asgards ancestraux et les proto-eucaryotes puisse être à l'origine de la répartition très éparse observée de certaines ESPs Asgards et de protéines marqueurs universels (Da Cunha, Gaïa, et al., 2022). En effet, il est possible que la position variable des Asgards dans les arbres simples gènes soient dues à un taux d'évolution rapide. Des reconstructions métaboliques ont suggéré que les Asgards dépendent d'interactions symbiotiques pour l'anabolisme et le catabolisme, ce qui pourrait expliquer pourquoi ils sont si difficiles à cultiver (2 espèces sont actuellement cultivées : *Prometheoarchaeum syntrophicum* (Imachi et al., 2020) et *Lokiarchaeum ossiferum* (Rodrigues-Oliveira et al., 2023). Il est donc possible que l'adaptation des Asgards à leurs partenaires ait augmenté le taux d'évolution de certaines de leurs protéines, au même titre que l'ancien groupe supposé des Archezoa. De plus, ces taux de transfert horizontaux élevés pourraient expliquer l'aspect disparate des protéines jusqu'alors supposées ESPs au sein des archées (Da Cunha, Gaïa, et al., 2022), ce que ne mentionnent pas de récentes études qui soutiennent une origine des eucaryotes au sein du groupe Asgard (Eme et al., 2023).

Le comportement anormal des protéines d'Asgard

De plus, dans les arbres basés sur une seule protéine, la position des eucaryotes est variable, se rattachant soit au sein des Asgards, soit comme groupe frère de tous les Asgards, soit comme groupe frère d'autres archées. Ce phénomène avait déjà été observé avec les trois premiers Asgards publiés (Loki 1, 2 et 3) dans 36 arbres à protéine unique (Da Cunha et al., 2017; Da Cunha, Gaïa, et al., 2022). Les Asgards étaient aussi souvent paraphylétiques dans ces arbres (les Heimdallarchaeota se retrouvant groupe frère des eucaryotes tandis que les autres groupes d'Asgards forment trois clades distincts non apparentés), alors que la monophylie d'autres clades majeurs d'archées était généralement retrouvée. Cela suggère que les positions dispersées des Asgards n'étaient probablement pas dues à un manque de résolution. Un comportement anormal des protéines Asgards a également été signalé pour les protéines ribosomiques (Garg et al., 2021). En effet, la dispersion des Asgards dans les arbres universels de protéines ribosomiques pourrait refléter des erreurs dans la reconstruction MAG (*Metagenome-Assembled Genome*), même si dans notre cas, l'utilisation d'un Asgard cultivé ne semble pas propice à ce type d'erreur et peut donc être tout à fait fiable. Cependant, ceci ne peut probablement pas expliquer toutes les situations. En effet, le même type de comportement anormal a été observé dans des arbres simples gènes basés sur des protéines universelles encodées par des archées DPANN, évoluant rapidement. Cela pourrait suggérer que la position variable des Asgards au sein de ces arbres d'archées résulterait d'un phénotype évoluant rapidement (Da Cunha, Gaïa, et al., 2022).

L'importance du choix des protéines

Un autre problème qui peut fausser la topologie de l'arbre du vivant est l'ensemble de données de marqueurs utilisés. Nous avons tenté de maximiser le nombre de gènes orthologues à utiliser, en ne nous limitant pas aux gènes ribosomiques. En effet, il a été montré que le passage d'un arbre 2D à 3D peut ne dépendre que d'une seule protéine. Ainsi, par exemple, une étude basée sur un jeu de données initial de 30 protéines montre que le retrait d'une seule protéine, l'ATPase YchF d'origine bactérienne, transformait un arbre 3D en un arbre 2D (Liu et al., 2021b). De même, le retrait du facteur d'élongation EF2 d'un ensemble de données de 36 protéines transformait un arbre 2D en un arbre 3D en cassant la monophylie Eucaryotes-Lokiarchaeota (Da Cunha et al., 2017). Le fait qu'une seule protéine puisse déterminer la topologie de l'arbre universel du vivant à partir de plusieurs dizaines de marqueurs souligne l'importance d'analyser soigneusement les arbres simples gènes avant d'effectuer la concaténation de leurs données.

Dans le choix des marqueurs, la longueur des gènes est également cruciale. Ainsi, les protéines de grande taille ont tendance à soutenir les arbres 3D, tandis que les protéines courtes ont tendance à soutenir les arbres 2D (Da Cunha et al., 2017). Il est possible que les protéines courtes n'abritent pas suffisamment de positions informatives pour détecter le signal correspondant à la monophylie des Archaea. Ainsi les jeux de données utilisés par Liu (Liu et al., 2021b), Xie (Xie et al., 2022) et leurs collègues sont tous deux enrichis en protéines courtes (80% et 40%, respectivement), ce qui favorise potentiellement la topologie 2D, et omettent plusieurs grandes protéines qui ont donné individuellement des arbres 3D par le passé (Da Cunha et al., 2017). En particulier, ils sont tous deux dépourvus de la grande sous-unité B de l'ARN polymérase. Or, les grandes sous-unités de l'ARN polymérase ont un signal phylogénétique très élevé (c'est-à-dire qu'elles prédisent avec plus de précision une véritable ligne d'ascendance verticale) (Martinez-Gutierrez & Aylward, 2021). En effet, sur un jeu de données de 41 marqueurs archéens et bactériens conservés, les grandes sous-unités de l'ARN polymérase obtiennent de meilleures performances malgré une longueur d'alignement globale plus courte. En revanche, les protéines ribosomiques ont tendance à présenter individuellement un faible signal phylogénétique. Le signal phylogénétique obtenu avec la concaténation des protéines ribosomiques était beaucoup plus élevé que celui obtenu avec les protéines ribosomiques individuelles, mais reste inférieur au signal obtenu avec la concaténation des deux grandes sous-unités de l'ARN polymérase (Martinez-Gutierrez & Aylward, 2021). Des résultats similaires donnant des arbres 3D ont également été obtenus (Da Cunha et al., 2017; T. A. Williams et al., 2020). Des arbres 2D n'ont été obtenus qu'après avoir recodé les acides aminés des alignements de séquences multiples. Étant donné que le recodage des acides aminés réduit fortement le signal dans l'analyse phylogénétique, il est possible que le passage d'un arbre du vivant 3D à un arbre 2D après recodage soit dû à un signal plus faible en faveur de la monophylie des archées (Da Cunha et al., 2017). C'est dans un souci d'avoir un maximum de signal phylogénétique que nous avons tout au long de notre étude tenté de récupérer un maximum de gènes en plus des protéines ribosomiques. Dans notre cas, le jackknife de gènes effectué sur les archées et les distances de Robinson-Foulds nous ont permis d'évaluer l'impact du choix de gènes, qui semblait faible contrairement au choix d'espèces. Ici, nous avons évalué l'impact de notre sélection de gènes en comparant nos résultats à ceux obtenus sur nos phylogénies d'archées. Notre sous-échantillonnage de gènes présents chez les eucaryotes n'affecte que très peu les bipartitions retrouvées, ne nous aidant pas à trancher entre nos topologies d'archées du chapitre précédent (mais n'en ajoutant pas non plus).

Transferts horizontaux de gènes et ESPs

La caractérisation de nouvelles lignées Asgards a conduit en leur sein à l'identification de plusieurs nouvelles ESPs. Toutefois, la distribution des ESPs est très inégale au sein des Asgards. De nombreuses ESPs sont spécifiques à une lignée Asgard tandis que les autres lignées en sont dépourvues (Liu et al., 2021b; Xie et al., 2022). C'est par exemple le cas de la tubuline, qui n'est présente que dans la lignée Odin (Xie et al., 2022; Zaremba-Niedzwiedzka et al., 2017). Dans les modèles 2D et 3D, la distribution phylogénétique extrêmement disparate de ces ESPs nécessite de nombreuses pertes et/ou des transferts entre les lignées d'archées. Dans les scenarii 2D on devrait en outre supposer que les eucaryotes ont émergé à partir d'une lignée Asgard éteinte qui possédait l'ensemble des ESPs actuellement répartis entre les différentes lignées d'Asgards actuelles (Eme et al., 2017). Dans les deux modèles, il est particulièrement difficile d'expliquer l'existence d'ESPs actuellement restreints à une seule ou à quelques lignées d'Asgards. Cependant, si l'on considère que les Eucaryotes sont à la base des Asgards + TACK, comme le laisse sous-entendre nos travaux d'hétérotachie, alors un ancêtre complexe à ce groupe est envisageable. Cet ancêtre serait issu d'une lignée aujourd'hui éteinte et qui aurait déjà une certaine partie des gènes et caractéristiques que peuvent partager les Eucaryotes, Asgard et TACK. Ainsi, jusqu'à cet ancêtre, l'évolution se serait faite par complexification. Puis, cette complexification se serait poursuivie chez les Eucaryotes, tandis que le groupe Asgard + TACK aurait évolué par simplification secondaire. Les caractéristiques communes pourraient très bien s'expliquer par héritage vertical de cet ancêtre commun, tandis que d'autres caractéristiques auraient été perdues ou modifiées, expliquant l'aspect disparate de certaines caractéristiques. D'ailleurs, cette situation n'est pas sans rappeler le piège de l'hypothèse Archézoa où l'on pensait que ces Eucaryotes supposés simples étaient à la base des Eucaryotes. Le fait que certains Asgards ou TACK vivent en symbiose ou en interaction avec d'autres organismes pourraient très bien expliquer ces phénomènes de simplification secondaire, comme ce fut le cas pour les Archézoa.

Une autre hypothèse qui pourrait plus facilement expliquer la distribution inégale des ESPs Asgards est que certains ESPs ont été recrutés par les Asgards parmi les proto-eucaryotes (c'est-à-dire les membres des lignées eucaryotes qui ont précédé le dernier ancêtre commun des eucaryotes). De tels transferts horizontaux de gènes ont pu se produire soit au début de l'évolution des Asgards, les ESPs correspondants étant présents chez la plupart ou la totalité des Asgards, soit plus tard, au cours de la diversification des lignées d'Asgard, expliquant ainsi la distribution restreinte des ESPs correspondants (Da Cunha et al., 2017). Cela pourrait également expliquer pourquoi une de nos trois Heimdallarchaeota se retrouve presque systématiquement rattachée aux eucaryotes au lieu des autres Heimdallarchaeota. Récemment, une étude a enraciné les eucaryotes au sein d'un nouveau groupe appelé Hodarchaeales (Eme et al., 2023). Nous avons cherché à quoi correspond cette Heimdallarchaeota dans leur jeu de données. L'analyse de notre Heimdallarchaeota GCA_001940645.1 correspond à LC-3 sp001940645 qui est considérée comme une Hodarchaeales, qui était jusqu'alors considérée comme une Heimdallarchaeota. Notre thèse mettrait donc aussi en évidence ce nouveau groupe.

Il est important de noter que cette hypothèse d'HGT proto-eucaryote expliquerait pourquoi certaines ESPs d'Asgard sont beaucoup plus semblables à leurs homologues eucaryotes qu'à ceux des archées. C'est le cas de la tubuline Odin, qui est beaucoup plus proche des tubulines eucaryotes que de la tubuline trouvée chez les Thaumarchaeota (Zaremba-Niedzwiedzka et al., 2017). De même, la plupart des actines d'Asgard sont beaucoup plus semblables aux actines eucaryotes et aux protéines liées à l'actine (ARP) qu'aux crénactines des archées (Da Cunha, Gaia, et al., 2022; Stairs & Ettema, 2020). Dans le cas de l'actine, l'hypothèse d'un HGT proto-eucaryote est fortement étayée par des analyses phylogénétiques montrant que les différentes formes

d'actines présentes chez les Asgards se ramifient entre les différents clades de paralogues d'actines eucaryotes (actine cytoplasmique et ARP) qui étaient déjà présents chez LECA (Da Cunha, Gaia, et al., 2022; Stairs & Ettema, 2020). Par ailleurs, corroborant cette hypothèse, des analyses de génomique comparative ont montré que les ESPs sont d'origine relativement récente et pourraient correspondre à des HGT tardifs entre archées et eucaryotes (Nasir et al., 2021). En effet, les ESPs pourraient être plus répandues que ce que l'on pensait, comme le laisse penser la découverte d'actine encodée par des virus (viractine) dans plusieurs génomes de virus (Cunha et al., 2020; Da Cunha, Gaia, et al., 2022) ou encore la présence de protéines d'actine chez les Bathyarchaeota (Dombrowski et al., 2018; Zhou et al., 2018). Ces découvertes suggèrent que la présence des ESPs dans d'autres lignées d'archées ou de virus est vraisemblablement sous-estimée (Nasir et al., 2021).

Il est également important de garder à l'esprit que les arbres phylogénétiques sont le résultat final d'étapes analytiques successives. Parmi celles-ci, les alignements de séquences multiples sont critiques, et malgré les améliorations indéniables de leurs méthodes, le risque de mauvais alignements augmente avec la taille de l'ensemble de données. Des défauts d'alignements de protéines universelles dus à l'inversion/substitution de domaines ou à des mésappariements ont été détectés dans 42% des marqueurs universels utilisés (36 arCOGS + protéines ribosomiques) dans le premier article d'Asgard qui ont ensuite servi aux études ultérieures (Nasir et al., 2021). Il serait donc important de vérifier les nouveaux alignements pour détecter d'éventuelles erreurs. Notamment, le risque de mauvais alignements est accru en cas de sur-échantillonnage des taxons et d'inclusion d'espèces à évolution rapide, deux facteurs fréquemment observés dans les études concernant l'arbre universel du vivant (Da Cunha, Gaïa, et al., 2022). De plus, nos résultats soutiennent une monophylie des Asgards avec le groupe TACK plus qu'avec les eucaryotes, en particulier lorsque l'on garde des sites à taux d'évolution lents. Les Asgards pourraient être très sensibles aux phénomènes d'hétérotachie et d'hétéropécillie. D'ailleurs, une de nos Heimdallarchaeota est systématiquement regroupée au sein des eucaryotes au lieu de se rapprocher des autres Heimdallarchaeota ; cela pourrait s'expliquer si on émet l'hypothèse d'un fort phénomène d'hétérotachie qui éclate les Heimdallarchaeota. En effet, si l'on tient en compte la possibilité d'artefact dus à des transferts horizontaux de gènes et à l'utilisation de sites à taux d'évolution rapide (en lien avec leur mode de vie symbiotique ou parasitaire d'eucaryotes), il peut être tout à fait légitime de douter de l'étroite parenté entre eucaryotes et Asgards.

Enfin, une possibilité est que ces ESP soient en fait des vestiges de ce qui était présent chez leur ancêtre commun LAECA. Cela suggérerait que ce dernier est déjà d'une certaine complexité et que l'évolution vers les archées se fait par simplification. Cette hypothèse est étayée par le fait que l'évolution des archées a probablement impliqué des pertes massives à partir d'un ancêtre complexe (Koonin, 2015). Dans ce scénario, les archées ont perdu la plupart des caractéristiques eucaryotes que possédait un LAECA complexe. Les Asgard, ayant moins divergé, conservent alors un plus grand nombre d'ESPs ainsi qu'une plus grande proximité phylogénétique avec les eucaryotes. Par conséquent, les scénarii 1D ne contredisent pas la proximité d'Asgard avec les eucaryotes, mais l'interprètent plutôt différemment (D. P. Devos, 2021). L'aspect éparse de certaines protéines au sein des archées peut alors s'interpréter comme le résultat de pertes et simplification secondaires, en particulier chez les TACK auxquels se rapprochent les Asgards dans nos analyses d'hétérotachie et d'hétéropécillie. La supposée grande quantité de transferts horizontaux de gènes des bactéries vers les archées (Nelson-Sathi et al., 2015) pourrait en réalité

représenter principalement les vestiges d'une lignée ayant divergé d'un ancêtre bactérien qui a été perdu dans certaines archées et maintenu divergent dans d'autres (D. P. Devos, 2021).

Par ailleurs, certains éléments, bien qu'encore limités, pourraient aller dans ce sens, avec notamment un placement des bactéries PVC à la base de la lignée conduisant à LAECA. Ainsi, un signal phylogénétique a été détecté dans les protéines ribosomales, soutenant un scénario 1D avec les bactéries PVC à la base du domaine eucaryotes-archées (Cavalier-Smith & Chao, 2020). Le signal est faible, mais on peut s'y attendre car il s'agit d'une des relations les plus anciennes de l'arbre de vie. En fait, on s'attend à ce que le signal entre les bactéries PVC et les eucaryotes soit beaucoup moins clair qu'entre les archées et les eucaryotes, car les premières se sont séparées beaucoup plus tôt que les secondes (D. P. Devos, 2021).

Racine des archées, paraphylie des Euryarchaeota, position des Altiarchaeota et relation avec le groupe DPANN

Depuis la découverte du premier représentant des DPANN (Rinke et al., 2013), le placement (T. A. Williams et al., 2015) de ce groupe au sein des archées est incertain en raison de la réduction extrême de leur génome (490 885 paires de bases pour environ 1000 gènes, le plus petit de tous les archées (Waters et al., 2003)) et de la longueur de leurs branches (autrement dit : leur taux de substitution rapide) (Dombrowski et al., 2019). Les DPANN pourraient à la place appartenir aux Euryarchaeota ou être polyphylétiques en occupant diverses positions au sein des Euryarchaeota.

Le placement basal des DPANN doit être traité avec une certaine prudence. Plusieurs études utilisant les bactéries en groupe externe avec des marqueurs protéiques conservés ont placé la racine des archées entre les DPANN et le reste des archées (Martinez-Gutierrez & Aylward, 2021; Petitjean et al., 2014; T. A. Williams et al., 2017). Des modèles bayésiens sophistiqués ont également placé les DPANN à la racine des archées (T. A. Williams et al., 2017). Ces modèles étaient basés sur des modèles d'évolution des génomes et prenaient en compte les duplications, les pertes et les transferts horizontaux de gènes. Cependant, l'ensemble des données analysées pour cette dernière étude était relativement limité et ne comprenait que des séquences d'ARN 16*S* et 23*S*, connues pour être biaisées au niveau compositionnel (Galtier & Lobry, 1997). Un super-arbre multigénique a également été enraciné entre les DPANN et l'ensemble de toutes les autres archées à l'aide d'un modèle d'évolution du génome récemment développé (T. A. Williams et al., 2017). D'après cette étude, les reconstructions métaboliques sur l'arbre enraciné suggèrent que les premières archées étaient des anaérobies qui avaient la capacité de réduire le CO_2 en acétate via la voie de Wood-Ljungdahl. Contrairement aux propositions suggérant que la réduction du génome a été le mode prédominant de l'évolution des archées (Csűrös & Miklós, 2009), cette étude propose un ancêtre archéen à génome relativement petit, qui aurait ensuite augmenté en complexité via la duplication et le transfert horizontal de gènes. La capacité génétique d'utiliser l'ancien système de fixation du carbone (Wood-Ljungdahl), la présence de caractères méthanogènes, ainsi que la capacité d'oxyder en anaérobiose le méthane et d'autres hydrocarbures courts ont été découverts dans diverses lignées d'archées vivant dans des environnements anaérobies (Spang et al., 2017). Par conséquent, ces résultats soutiennent les hypothèses qui suggèrent que toutes les archées actuelles ont évolué à partir d'un ancêtre anaérobie autotrophe qui utilisait la voie Wood-Ljungdahl et qui aurait pu obtenir de l'énergie par méthanogenèse. Leur position basale et leur monophylie présentent une grande incertitude, le faible échantillonnage de taxons disponible nous donnant des raisons de douter de ce résultat. Des études antérieures ont montré que la position phylogénétique des DPANN est sensible aux

taxons inclus (Dombrowski et al., 2019; T. A. Williams et al., 2017). Dans notre cas, nous arrivons à la même conclusion. En effet, l'ajout des eucaryotes tend à éloigner les DPANN des Altiarchaeota et les met à la base des Ouranosarchaea (TACK + Asgard (+ eucaryotes ?)). De plus, il est possible qu'il s'agisse d'un artefact causé par un LBA dû à leur haut taux de substitution et leur génome de très petite taille (Aouad et al., 2018, 2022; Petitjean et al., 2014; Raymann et al., 2015; Roure & Philippe, 2011). En effet, les DPANN sont des parasites d'archées (par ex. *Nanoarchaeum equitans* est un symbionte de *Ignicoccus hospitalis*, les *Micrarchaeota* spp. sont des parasites des Sulfolobales). Si les DPANN représentent le groupe ancestral des archées, alors l'ancêtre commun des archées actuelles aurait donc été une forme symbiotique ayant par la suite gagné en complexité. Cette conclusion irait à l'encontre d'une origine très ancienne des archées. Toutefois, ce raisonnement pourrait être circulaire puisque si le mode de vie ancestral des archées est symbiotique, l'hôte devrait être autre chose. Quoi qu'il en soit, nous ne retrouvons jamais, dans aucune de nos analyses, le groupe DPANN à la base des archées. Nous avons donc des raisons de douter de la véracité de ce résultat. Il est probable que le séquençage supplémentaire d'archées, et en particulier de DPANN elles-mêmes, permettra de clarifier leur place.

Les Euryarchaeota ont longtemps été définis comme un groupe monophylétique (Petitjean et al., 2014; T. A. Williams et al., 2017). Nous avons observé lors de nos analyses PMSF que ce résultat est instable, en fonction de nos sous-échantillonnages taxonomiques (jackknives d'espèces). Encore une fois ici, la monophylie des Euryarchaeota est mise à mal lors de l'utilisation du rootstrap. Les Euryarchaeota nous apparaissent comme un groupe paraphylétique. Le groupe Ouranosarchaea (TACK + Asgard (+ eucaryotes ?)) est bien retrouvé dans nos analyses, sauf lors de l'AU-test en absence d'eucaryotes où les Asgard se retrouvent à la racine. En revanche, la monophylie des Gaiarchaea (Euryarchaeota) reste difficile à retrouver. Il nous est, selon nos résultats, impossible de trancher pour l'une ou l'autre solution.

Les questions de la monophylie des Euryarchaeota, de la position du groupe DPANN et de ses liens avec les Altiarchaeales, et de la position basale des DPANN au sein des archées demeurent encore quelques-uns des plus grands mystères concernant la phylogénie des archées. Y répondre nécessiterait l'assemblage de jeux de de données spécifiques ainsi que la mise en œuvre de protocoles spécialement conçus pour répondre à cette question (Aouad et al., 2022) ainsi que la découverte de nouveaux membres de ce groupe. Les DPANN comme groupe basal des archées ne pourraient peut-être qu'être une version archée de ce que sont les Archezoa aux eucaryotes, à savoir des espèces symbiotiques à génomes très réduits subissant un artefact d'attraction des longues branches. Notre étude tendrait à privilégier une paraphylie des Euryarchaeota, un rapprochement des DPANN et des Altiarchaeota à la base des Ouranosarchaea + Eucaryotes et un enracinement des archées au sein du groupe SANT. Si l'on tient compte de nos travaux d'hétérotachie, les eucaryotes pourraient être à la base des Ouranosarchaea. Finalement, le cladogramme que nous privilégions dans le cadre de cette thèse est donné à la **Figure 76**.

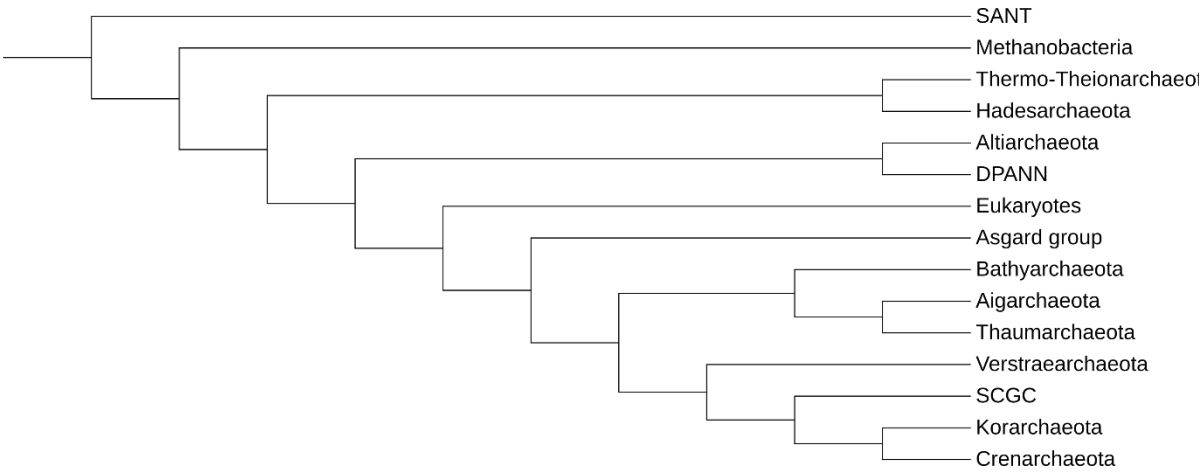


Figure 76. Cladogramme raciné des archées et eucaryotes basé sur l'ensemble de nos résultats.

La racine est située dans le groupe SANT. Compte tenu de nos résultats d'hétérotachie, il nous semblerait que les Eucaryotes soient à la base des Ouranosarchaea et que les Euryarchaeota soient polyphylétiques.

DISCUSSION GENERALE & CONCLUSION

Ce travail de thèse s'est axé autour des archées et de leurs relations avec les eucaryotes. Il avait pour objectif d'évaluer la pertinence de la tendance actuelle à inclure les eucaryotes au sein des archées. Nous avons ainsi profité de l'apport toujours plus important de données de séquençage afin d'améliorer la résolution de l'arbre évolutif des archées, avant d'y inclure les eucaryotes. Notre avons exploité cet apport massif de nouvelles séquences en assemblant un jeu de données le plus fiable et solide possible afin de le soumettre à différentes approches d'inférence phylogénétique et d'identifier de possibles biais méthodologiques.

Une grande partie du travail fut de construire un jeu de données qui soit le plus propre et complet possible afin de minimiser tout artefact susceptible d'altérer le résultat. Cela s'est fait par une évaluation de la qualité des gènes (tests de contamination, orthologie, évaluation des taux de substitution…) et par un choix d'espèces raisonné. Plusieurs jeux de données ont été créés et testés afin de comparer les résultats via des procédure de jackknife de gènes et d'espèces. Puis je me suis ensuite concentré sur l'évaluation de la qualité de l'inférence phylogénétique, en allant de l'alignement de séquences orthologues à l'inférence d'arbres d'espèces.

En général, pour résoudre des problématiques de phylogénie profonde, il est nécessaire d'analyser des ensembles de données à grande échelle génomique, car le signal phylogénétique exploitable dans des séquences évolutivement éloignées est faible. Toutefois, il faut prendre des précautions lors de l'utilisation de cette approche, car une résolution parfaite de l'arbre ne signifie pas nécessairement que l'évolution des séquences a été correctement reconstruite. En particulier, les valeurs de bootstrap ne peuvent pas être utiles pour détecter les erreurs systématiques (faux signal phylogénétique dû à la saturation du signal lorsque les substitutions sont trop abondantes) lors des analyses phylogénétiques, car le fait qu'une relation soit supportée avec une valeur de bootstrap de 100% ne signifie pas qu'elle soit correcte. Ma thèse confirme que les artefacts d'inférence phylogénétique sont plus importants lors de l'utilisation d'un grand nombre de positions et peuvent ainsi masquer le signal phylogénétique authentique (Da Cunha et al., 2017). Ce fut le cas lors de nos analyses de retrait de sites, à la fois pour le Slow-Fast des archées, ainsi que pour les analyses d'hétérotachie et d'hétéropécillie lorsque j'ai rajouté les eucaryotes. Les valeurs de bootstrap étaient élevées dans tous les cas malgré des résultats contradictoires selon la taille des alignements et le choix des sites conservés. Afin d'éviter les artefacts, il serait donc idéal de disposer de modèles d'évolution capables de prendre en compte la complexité et l'hétérogénéité des processus de substitution. Néanmoins, cette démarche est complexe. Il est plus simple de modéliser un seul processus à la fois que l'ensemble de tous les processus possibles en même temps. En attendant le modèle parfait, ce problème peut être contourné par des approches alternatives. Afin de détecter de possibles artefacts et de l'erreur systématique, l'approche utilisée dans le cadre de ma thèse fut de tester la robustesse des résultats à l'échantillonnage taxonomique, au retrait des sites rapides et à l'utilisation de différents modèles d'évolution des séquences.

Dans le premier chapitre, j'ai dressé le bilan des connaissances au début de ma thèse sur la problématique de l'enracinement de l'arbre du vivant. Ce chapitre a fait l'objet d'une publication dans la revue *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*. Cet article visait à mettre en lumière les difficultés méthodologiques rencontrées pour enraciner l'arbre de la vie et explorer les raisons (sociologiques) sous-jacentes au faible intérêt porté à cette question fondamentale. J'ai examiné dans 126 articles parus après l'an 2000 les difficultés qui affectent l'inférence phylogénétique ainsi que les moyens d'améliorer la modélisation du processus de substitution, qui est hautement hétérogène, tant d'un site à l'autre qu'au fil du temps.

De manière surprenante, moins de la moitié des études semblaient prêter un intérêt aux possibles artefacts pouvant affecter les phylogénies. J'ai également illustré qu'un meilleur échantillonnage taxonomique, de meilleures sélections de gènes et des stratégies raisonnées de suppression de données ont conduit à de nombreuses révisions de l'arbre du vivant, en déplaçant presque toujours les organismes simples plus haut dans l'arbre, à condition que les artefacts d'attraction des longues branches soient bien neutralisés. Ensuite, j'ai noté que, malgré la quantité de données génomiques disponibles depuis 2000 et jusqu'en 2015, il y a eu étonnamment peu d'intérêt à déterminer l'origine de l'arbre de la vie. De plus, les rares études traitant de cette question étaient presque toujours basées sur des méthodes datant des années 1990, dont on connait aujourd'hui très bien les limites. Cela m'a amené à affirmer que l'hypothèse d'une racine bactérienne pour l'arbre du vivant pouvait être attribué au préjugé de la *Grande Chaîne des Êtres* d'Aristote, selon laquelle les organismes simples sont les ancêtres de formes de vie plus complexes. Enfin, j'ai illustré via un enracinement de l'arbre du vivant en utilisant le facteur d'élongation, un gène paralogue anciennement dupliqué (Baldauf et al., 1996; Gouy et al., 2015), que même les meilleurs modèles actuels ne peuvent pas encore gérer totalement la complexité des processus évolutifs, en particulier concernant le processus de substitution. Dans nos analyses, les deux clades bactériens se ramifient à des positions différentes : en tant que groupe frère des archées + eucaryotes pour EF-Tu et comme groupe frère des eucaryotes pour EF-G. Dans les deux sous-arbres, les archées sont paraphylétiques, avec les Crenarcheota plus proches des eucaryotes, mais sans soutien statistique. De toute évidence, les erreurs stochastique et systématique affectent profondément cette phylogénie basée sur des facteurs d'élongation dupliqués (Gouy et al., 2015). J'en ai conclu que la racine bactérienne (communément acceptée au début de ma thèse en 2015) est encore non prouvée et que la racine de l'arbre du vivant devrait être revisitée en utilisant des super-matrices phylogénomiques si l'on veut interpréter correctement les processus évolutifs ayant favorisé l'émergence de la cellule eucaryote. En effet, si l'évaluation des caractères homologues (en particulier grâce à l'élimination des régions alignées de manière ambiguë) et des gènes orthologues est relativement précise et ne constitue pas le problème le plus important de la phylogénétique profonde, l'inférence basée sur les super-matrices semble être robuste à l'inclusion de séquences paralogues et xénologues (c'est-à-dire transférées horizontalement). La principale difficulté à résoudre est liée à la modélisation des processus de substitution, auxquelles les super-matrices sont sensibles (Philippe et al., 2005, 2017; Young & Gillung, 2020).

Étant donné que les erreurs stochastique et systématique ont plus d'impact sur l'enracinement de l'arbre du vivant que sur la résolution de l'une de ses parties, les stratégies d'enracinement devraient d'abord être validées sur des questions moins profondes et de difficulté similaire (nous avons dans notre cas utilisé l'exemple de la monophylie des Bivalves). À notre avis, il n'est pas judicieux d'appliquer directement de nouvelles approches, aussi intelligentes soient-elles, pour localiser la racine de l'arbre de la vie sans une validation préalable approfondie sur des questions difficiles dont les réponses sont connues. Les données de test nécessaires sont faciles à assembler étant donné la quantité de publications disponibles de nos jours. Compte tenu de ces conditions préalables, nous soutenons que l'approche par super-matrice reste la méthode de choix pour enraciner l'arbre du vivant car il s'agit de la stratégie la plus largement utilisée et validée (Boussau et al., 2013; T. A. Williams et al., 2017).

De plus, la découverte du groupe d'archées Asgard durant ma thèse a alors favorisé la communauté scientifique à envisager un modèle du vivant à 2 domaines, privilégiant un scénario de fusion pour expliquer l'origine de la cellule eucaryote actuelle. Cela m'a amené dans le deuxième chapitre à construire trois jeux de données de 352 OTU contenant (1) 90 protéines

ribosomiques, (2) 343 gènes orthologues et (3) 117 gènes « néo-orthologues » issus de la découpe phylogénétique de familles de gènes paralogues. Ces jeux de données constituent désormais un ensemble de gènes qui peuvent être utilisés dans des analyses ultérieures pour reconstruire la phylogénie des archées à la lumière de nouveaux génomes. Une des difficultés majeures de l'assemblage de jeux de données pour inférer la phylogénie des espèces a été la gestion de la paralogie et de la xénologie, les deux conduisant à observer plusieurs séquences par espèce pour un gène donné, certaines d'entre elles ne reflétant pas la vraie phylogénie des organismes. J'ai défini des groupes monophylétiques sur base de nos résultats. J'ai alors fait varier systématiquement l'échantillonnage taxonomique au travers de la génération de cinq réplicas, puis effectué des jackknives de gènes. Puis j'ai inféré des arbres selon deux stratégies différentes : super-matrices et super-arbres. Dans chaque cas, différents modèles d'évolution des séquences ont été employés afin d'évaluer au mieux la pertinence de chacun ainsi que les biais (hétérogénéité du processus évolutif) et artefacts susceptibles d'affecter nos phylogénies. J'ai observé que la sélection d'espèces joue un rôle important dans la topologie obtenue. En revanche, dans les manipulations de jackknife de gènes, les distances RF sont faibles car les arbres sont congruents. La sélection de gènes influe moins sur la topologie des arbres, dès lors que ceux-ci utilisent des modèles suffisamment sophistiqués (en particulier les modèles catégoriques et le PMSF). Enfin, les super-matrices tendent à donner des résultats plus robustes que les super-arbres. Après analyse des bipartitions majoritaires, j'ai relevé quatre topologies d'arbres que l'on pouvait retrouver, montrant alors qu'il y avait des incongruences selon la sélection d'espèces. Nous avons en particulier relevé les incongruences suivantes (**Figure 44**) :

- les Korarchaeota, qui se placent soit comme groupe frère des Crenarchaeota, soit comme groupe frère des de l'ensemble Crenarchaeota + SCGC + Verstraearchaeota.
- Les Hadesarchaeota, qui se placent soit comme groupe frère du groupe Thermo Theionarchaeota, soit à la base de tous les Euryarchaeota et des DPANN.
- Les Altiarchaeales, qui se placent soit comme groupe frère des DPANN soit comme groupe frère des Methanobacteria.

Afin de réduire l'erreur systématique, j'ai choisi de tester notre jeu de données en éliminant les sites à taux d'évolution rapide via une approche dite *slow-fast* (Brinkmann & Philippe, 1999; Roure & Philippe, 2011) afin de favoriser les sites plus lents, qui ont moins de chances de contredire les hypothèses d'homogénéité du processus de substitution. Mes analyses supportent les Korarchaeota plutôt que le groupe SCGC (*Single Cell Genomics Center*) comme groupe frère de Sulfolobales + Desulfurococcales + Thermofilaceae + Thermoproteaceae. En revanche, il est difficile de trancher parmi nos quatre topologies laquelle semble être la bonne. En effet, la sélection d'espèces semble être un facteur important dans les résultats obtenus. Mes résultats mettent en avant les difficultés rencontrées lors l'inférence des arbres. L'amélioration de la qualité de cette inférence vis à vis de l'élimination de données problématiques dépend d'un équilibre fragile entre bruit (erreur stochastique), signal non-phylogénétique (erreur systématique) et signal phylogénétique authentique (signal historique) (Baurain & Philippe, 2010) qui est difficile à prendre en compte par les méthodes automatiques actuelles, et requiert donc des développements méthodologiques supplémentaires. Notre étude tendrait à privilégier un rapprochement des DPANN et des Altiarchaeota à la base des Euryarchaeota, mais ce résultat semble dépendant de l'échantillonnage taxonomique. De plus la position des Hadesarchaeota reste incertaine, même s'il semble majoritairement se rapprocher de ce que nous avons nommé le groupe TTM (Thermococci + Theionarchaea + Methanomicrobia Arc).

Dans le troisième chapitre, j'ai décidé d'inclure une sélection d'eucaryotes au sein de nos alignements de sélections d'archées afin d'évaluer leur supposée relation avec les Asgards. Afin de faire face aux nombreuses séquences paralogues au sein des sous-arbres eucaryotes, j'ai réalisé des analyses phylogénétiques préliminaires pour chaque gène et sélectionné manuellement les séquences orthologues aux archées. Après de nombreuses analyses d'arbres de gènes individuels, j'ai construit des super-matrices de 126 gènes. Comme j'avais trouvé précédemment quatre topologies alternatives d'arbres d'archées selon nos sélections d'espèces, j'ai vérifié que l'ajout des eucaryotes n'impactait pas ces topologies. De même, j'ai vérifié et me suis assuré que le passage de 343 gènes à 126 gènes ne changeait pas les topologies de mes 5 réplicas d'espèces. Ces résultats m'ont conforté dans la confiance que je pouvais accorder à la sélection de gènes que j'ai pu faire, ainsi qu'au travail rigoureux effectué lié à la construction de mon jeu de données. J'ai alors voulu évaluer l'impact de l'hétérotachie (hétérogénéité du taux de substitution d'un site au cours du temps) et de l'hétéropécillie (hétérogénéité du processus de substitution d'un site au cours du temps) sur les arbres obtenus. Pour cela, après avoir inféré des taux d'évolution spécifiques à chaque site indépendamment pour les archées et les eucaryotes, j'ai filtré progressivement les colonnes à éliminer sur la base d'un calcul de delta de vitesse afin de créer des super-matrices ayant des colonnes avec des taux d'évolution plus ou moins hétérogènes. Concernant l'hétérotachie, contrairement à ce que laisse entendre la littérature récente (Aouad et al., 2022; Eme et al., 2017), le regroupement des Asgards auprès des eucaryotes n'est que très peu représenté lorsqu'on se concentre sur les sites homotaches entre archées et eucaryotes, au profit d'un regroupement des Asgards auprès des TACK. Il faut atteindre une grande taille de super-matrices (18 000 positions) avec l'ajout de sites à taux de substitution plus hétérogène pour que les Asgards soient systématiquement regroupés avec les eucaryotes, quel que soit notre échantillonnage taxonomique. Le regroupement des Asgards avec les eucaryotes pourrait donc être le fruit d'un biais de reconstruction phylogénétique lié à des sites présentant de l'hétérotachie mal modélisée. En revanche, les analyses d'hétéropécillie tendent à montrer le contraire. Ainsi lorsqu'on se concentre sur les sites homotaches entre archées et eucaryotes, ces derniers viennent s'insérer comme groupe frère des Asgards.

Ainsi, dans ma thèse, le retrait des sites les plus saturés s'avère être une méthode efficace afin d'améliorer l'extraction du signal phylogénétique, en ne générant pas de signal non-phylogénétique et en se focalisant uniquement sur le signal historique, soutenant ainsi une solution qui pourrait s'avérer plus pertinente. Toutefois, lorsqu'on regarde l'histoire de notre conception de l'évolution de la cellule eucaryote, on peut se poser la question de savoir si certaines solutions proposées ne sont pas le résultat de problèmes auxquels la science a déjà dû faire face dans le passé. Ainsi, la supposée racine archéenne des DPANN et l'étroite parenté entre les Asgards et les eucaryotes pourrait être une nouvelle version de ce qu'étaient les Archezoa aux eucaryotes, à savoir des espèces symbiotiques à génomes très réduits subissant un artefact d'attraction des longues branches. Ainsi, compte tenu de nos résultats, nous soupçonnons que les eucaryotes soient en réalité à la base des Ouranosarchaea, expliquant ainsi la distribution éparse des ESPs par des simplifications secondaires.

La racine de l'arbre du vivant est une problématique qui reste ouverte. Alors que l'on opposait historiquement deux structures cellulaires (procaryotes vs eucaryotes), les premières phylogénies moléculaires de l'arbre du vivant ont exacerbé la singularité des archées et leurs similitudes avec les eucaryotes (E. Chatton, 1925; É. P. L. Chatton, 1938; C. Woese et al., 1978; CarlR. Woese & Fox, 1977). Dès lors, il s'est posé la question de savoir si la structure de l'arbre du vivant comprenait deux ou trois domaines. De nombreuses hypothèses ont alors émergé

concernant l'origine de la cellule eucaryote, qui semble être une chimère entre une bactérie et une archée, bien qu'elle possède également des caractéristiques qui lui sont propres (T. M. Embley & Martin, 2006; T. M. Embley & Williams, 2015; Eme et al., 2017; Hartman & Fedorov, 2002; Lake, 2007). La découverte du groupe Asgard a donc été importante (Eme et al., 2017; Spang et al., 2015; T. A. Williams et al., 2017; Zaremba-Niedzwiedzka et al., 2017). Elle a confirmé la relation étroite entre eucaryotes et archées et apporté un soutien aux scenarii 2D (hypothèse Eocyte). Cependant, même si l'on a réduit le fossé phylogénétique entre archées et eucaryotes, l'origine proposée des eucaryotes au sein des Asgards soulève toujours les mêmes problèmes. Seuls deux Asgards ont été cultivés jusqu'à présent. La première cellule est *Candidatus Prometheoarchaeum syntrophicum* (Imachi et al., 2020), une Lokiarchaeota de très petite taille (0,5 mm) qui ne présente aucun signe de complexité cellulaire ou d'organisation membranaire. Elle est métaboliquement déficiente, vivant en symbiose avec une bactérie sulfato-réductrice du genre *Desulfovibrio* et une archée productrice du méthane du genre *Methanogenium*, ce qui est en contradiction à ce que l'on s'attendrait d'un proto-eucaryote complexe. En 2022, une nouvelle espèce d'Asgard récoltée dans la boue d'un estuaire en Slovénie, *Lokiarchaeum ossiferum*, est cultivée (Rodrigues-Oliveira et al., 2023) (après sept ans d'efforts) et étudiée en détail par microscopie électronique, qui dévoile plusieurs dizaines de fins tentacules comportant des épaississements et des excroissances ainsi qu'un cytosquelette s'étendant jusque dans les tentacules. Le génome de cette espèce et celui d'une autre espèce découverte au même endroit ont pu être séquencés entièrement. Ils comportent notamment quatre gènes codant des complexes protéiques qui chez les eucaryotes servent à plier, découper et assembler les membranes afin de relier les compartiments internes. L'hypothèse de transferts horizontaux de gènes proto-eucaryotes est pertinente à la fois dans les scenarii 2D et 3D. Cela implique que, indépendamment du scénario correct, l'étude des ESPs pourrait fournir des informations importantes sur l'eucaryogenèse en identifiant des étapes intermédiaires dans l'évolution de ces protéines. Dans la biosphère moderne, les HGT entre eucaryotes et bactéries sont particulièrement répandus entre les espèces vivant en étroite association. Des associations symbiotiques entre archées et eucaryotes ont également été décrites, comme les méthanogènes prospérant dans des protistes de différentes lignées eucaryotes (Husnik et al., 2021) ou *Cenarchaeum symbiosum* vivant dans des éponges marines (Preston et al., 1996). Il est donc possible que certains ancêtres du groupe Asgard aient été des symbiotes de proto-eucaryotes, échangeant des gènes avec leurs hôtes. Ainsi, si les ESPs d'Asgards sont issues de transferts horizontaux de gènes, alors elles n'expliquent en rien l'origine des eucaryotes. De plus, cette hypothèse implique que les anciens Asgards étaient déjà diversifiés avant le LECA et partageaient leurs biotopes avec des proto-eucaryotes (puisqu'ils sembleraient leur être symbiotiques). Par conséquent, l'étude de la distribution environnementale des Asgards modernes pourrait fournir des informations essentielles sur la nature des biotopes dans lesquels les proto-eucaryotes prospéraient. Il pourrait être intéressant d'étudier si certains Asgards vivent en symbiose avec des eucaryotes modernes en recherchant des signatures Asgards dans divers types de cellules eucaryotes. Nos travaux tendent à favoriser la racine des archées au sein du groupes SANT. Nous ne retrouvons pas systématiquement la monophylie des Euryarchaota.

Enfin, il se pose toujours les problèmes liés à l'aspect mosaïque des génomes et aux nombreuses différences biochimiques, métaboliques et cellulaires entre les archées et les eucaryotes, y compris la biocompatibilité des lipides (Martin & Muller, 1998; Peretó et al., 2004; Valentine, 2007). Les lipides des archées consistent en de longues chaînes d'alcool isoprénoïques attachées au glycérol par des liaisons éther, tandis que les eucaryotes et bactéries fabriquent les lipides de leurs membranes en assemblant deux chaînes d'acides gras avec une molécule de

glycérol par l'intermédiaire d'une liaison ester. Les chromosomes eucaryotes sont linéaires et ne possèdent pas de plasmides. Leur matériel génétique est entouré d'un noyau, et ils possèdent de nombreux organites. Les archées présentent une grande diversité métabolique (chimiolithotrophie, méthanogenèse, fixation de l'azote, fermentation, respiration anaérobie et aérobie…).

Près de 50 ans après leur découverte, la révolution génomique a révélé que les archées sont des organismes très divers. Les archées ont été découvertes dans presque tous les habitats (que ce soit via des métagénomes ou des souches mises en culture) et il est très probable que de nombreuses autres lignées d'archées soient encore inconnues. Les archées jouent des rôles clés dans des processus biogéochimiques importants et sont abondantes dans le microbiome des animaux, y compris le nôtre. Bien que plusieurs analyses métagénomiques ont contribué à en révéler plus sur leur diversité et leur prévalence, le fait qu'aucun pathogène archéen n'ait été découvert n'a pas grandement suscité d'intérêt dans leur étude (ce qui, signalons-le, peut être tout à fait curieux et mystérieux) et il est triste que la plupart du grand public ignore encore leur existence. Ils sont même présents dans nos nombrils sans que l'on ne sache toujours rien d'eux malgré leur découverte en… 2012 (Hulcr et al., 2012). Leur diversité au sein du microbiote est encore peu étudiée et méconnue. Ce n'est que récemment qu'une étude a révélé leur importance au sein d'un large spectre du règne animal, avec leur présence dans 175 espèces sur 250 échantillonnées (Thomas et al., 2022). Notre manque de connaissances sur la véritable diversité des archées entrave grandement notre capacité à comprendre la relation entre les archées et les eucaryotes. Maintenant, l'une des étapes les plus importantes serait d'explorer davantage la diversité de la vie. Avec les nouvelles technologies et techniques de séquençage, la quantité de génomes disponibles est impressionnante. La découverte et le séquençage des branches inexplorées de l'arbre du vivant pourraient permettre d'augmenter la résolution des arbres phylogénétiques en atténuant certains artefacts, en particulier en cassant les longues branches. Mais cet afflux de données massif auquel nous avons aujourd'hui accès doit pouvoir être exploité à sa juste valeur avec des méthodes qui puissent prendre en compte la complexité des processus biologiques si l'on veut espérer comprendre les relations entre bactéries, archées et eucaryotes.

BIBLIOGRAPHIE

Acehan, D., Santarella-Mellwig, R., & Devos, D. P. (2014). A bacterial tubulovesicular network. *Journal of Cell Science*, *127*(2), 277–280. https://doi.org/10.1242/jcs.137596

Adam, P. S., Borrel, G., Brochier-Armanet, C., & Gribaldo, S. (2017). The growing tree of Archaea: New perspectives on their diversity, evolution and ecology. In *ISME Journal* (Vol. 11, Issue 11, pp. 2407–2425). Nature Publishing Group. https://doi.org/10.1038/ismej.2017.122

Adl, S. M., Simpson, A. G. B., Lane, C. E., Lukes, J., Bass, D., Bowser, S. S., Brown, M. W., Burki, F., Dunthorn, M., Hampl, V., Heiss, A., Hoppenrath, M., Lara, E., Le Gall, L., Lynn, D. H., McManus, H., Mitchell, E. A., Mozley-Stanridge, S. E., Parfrey, L. W., … Spiegel, F. W. (2012). The revised classification of eukaryotes. *J Eukaryot Microbiol*, *59*, 429–514. https://doi.org/10.1111/j.1550-7408.2012.00644.x

Adoutte, A., Germot, A., Le Guyader, H., Philippe, H., Française De Génétique, S., Président, P. A. N., Jacob, F., Berger, V.-P. R., Pinon, H., Stoll, C., Bernheim, A., Bolotin-Fukuhara, M., Fellous, M., Génermont, J., Michel, B., Motta, R., Nicolas, A., Sommer, S., Thuriaux, P., … Prunier, M.-L. (1996). Que savons-nous de l'histoire évolutive des Eucaryotes ? 2. De la diversification des protistes à la radiation des multicellulaires* Comité de rédaction. *Medecines Sciences*, *12*(2), 1–17.

Akıl, C., Ali, S., Tran, L. T., Gaillard, J., Li, W., Hayashida, K., Hirose, M., Kato, T., Oshima, A., Fujishima, K., Blanchoin, L., Narita, A., & Robinson, R. C. (2022). Structure and dynamics of Odinarchaeota tubulin and the implications for eukaryotic microtubule evolution. *Science Advances*, *8*(12), eabm2225. https://doi.org/10.1126/sciadv.abm2225

Amiri, H., Karlberg, O., & Andersson, S. G. E. (2003). Deep origin of plastid/parasite ATP/ADP translocases. *Journal of Molecular Evolution*, *56*(2), 137–150. https://doi.org/10.1007/s00239-002-2387-0

Andersson, G. E., Karlberg, O., Canbäck, B., & Kurland, C. G. (2003). On the origin of mitochondria: a genomics perspective. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *358*(1429), 165–179. https://doi.org/10.1098/rstb.2002.1193

Aouad, M., Flandrois, J. P., Jauffrit, F., Gouy, M., Gribaldo, S., & Brochier-Armanet, C. (2022). A divide-and-conquer phylogenomic approach based on character supermatrices resolves early steps in the evolution of the Archaea. *BMC Ecology and Evolution*, *22*(1). https://doi.org/10.1186/s12862-021-01952-0

Aouad, M., Taïb, N., Oudart, A., Lecocq, M., Gouy, M., Taib, N., & Brochier-Armanet, C. (2018). Extreme halophilic archaea derive from two distinct methanogen Class II lineages. *Molecular Phylogenetics and Evolution*, *127*, 46–54. https://doi.org/10.1016/j.ympev.2018.04.011ï

Arcas, A., Cases, I., & Rojas, A. M. (2013). Serine/threonine kinases and E2-ubiquitin conjugating enzymes in Planctomycetes: unexpected findings. *Antonie van Leeuwenhoek*, *104*(4), 509–520. https://doi.org/10.1007/s10482-013-9993-2

Baldauf, S. L., Palmer, J. D., & Doolittle, W. F. (1996). The root of the universal tree and the origin of eukaryotes based on elongation factor phylogeny. *Proceedings of the National Academy of Sciences*, *93*(15), 7749–7754. http://www.pnas.org/content/93/15/7749.abstract

Bapteste, E., & Brochier, C. (2004). On the conceptual difficulties in rooting the tree of life. *Trends in Microbiology*, *12*(1), 9–13. https://doi.org/10.1016/j.tim.2003.11.002

Baurain, D., & Philippe, H. (2010). Current Approaches to Phylogenomic Reconstruction. In *Evolutionary Genomics and Systems Biology* (pp. 17–41). John Wiley and Sons. https://doi.org/10.1002/9780470570418.ch2

Bekker, A., Slack, J. F., Planavsky, N., Krapež, B., Hofmann, A., Konhauser, K. O., & Rouxel, O. J. (2010). Iron Formation: The Sedimentary Product of a Complex Interplay among Mantle, Tectonic, Oceanic, and Biospheric Processes*. *Economic Geology*, *105*(3), 467–508. https://doi.org/10.2113/gsecongeo.105.3.467

Bell, P. J. L. (2001). Viral Eukaryogenesis: Was the Ancestor of the Nucleus a Complex DNA Virus? *Journal of Molecular Evolution*, *53*(3), 251–256. https://doi.org/10.1007/s002390010215

Bell, P. J. L. (2005). The Viral Eukaryogenesis Theory. *Origins: Genesis, Evolution and Diversity of Life*, *6*, 347–367. https://doi.org/10.1007/1-4020-2522-X_22

Bell, P. J. L. (2006). Sex and the eukaryotic cell cycle is consistent with a viral ancestry for the eukaryotic nucleus. *Journal of Theoretical Biology*, *243*(1), 54–63. https://doi.org/http://dx.doi.org/10.1016/j.jtbi.2006.05.015

Bell, P. J. L. (2009). The Viral Eukaryogenesis Hypothesis. *Annals of the New York Academy of Sciences*, *1178*(1), 91–105. https://doi.org/10.1111/j.1749-6632.2009.04994.x

Betancur-R., R., Li, C., Munroe, T. A., Ballesteros, J. A., & Ortí, G. (2013). Addressing Gene Tree Discordance and Non-Stationarity to Resolve a Multi-Locus Phylogeny of the Flatfishes (Teleostei: Pleuronectiformes). *Systematic Biology*, *62*(5), 763–785. https://doi.org/10.1093/sysbio/syt039

Betancur-R., R., Naylor, G. J. P., & Ortí, G. (2014). Conserved Genes, Sampling Error, and Phylogenomic Inference. *Systematic Biology*, *63*(2), 257–262. https://doi.org/10.1093/sysbio/syt073

Bird, J. T., Baker, B. J., Probst, A. J., Podar, M., & Lloyd, K. G. (2016). Culture independent genomic comparisons reveal environmental adaptations for altiarchaeales. *Frontiers in Microbiology*, *7*(AUG). https://doi.org/10.3389/fmicb.2016.01221

Blobel, G., Walter, P., & Gilmore, R. (1986). Intracellular Protein Topogenesis. In G. Poste & S. T. Crooke (Eds.), *New Insights into Cell and Membrane Transport Processes* (pp. 277–283). Springer US. https://doi.org/10.1007/978-1-4684-5062-0_14

Boedeker, C., Schüler, M., Reintjes, G., Jeske, O., van Teeseling, M. C. F., Jogler, M., Rast, P., Borchert, D., Devos, D. P., Kucklick, M., Schaffer, M., Kolter, R., van Niftrik, L., Engelmann, S., Amann, R., Rohde, M., Engelhardt, H., & Jogler, C. (2017). Determining the bacterial cell biology of Planctomycetes. *Nature Communications*, *8*(1), 14853. https://doi.org/10.1038/ncomms14853

Borowiec, M. L., Rabeling, C., Brady, S. G., Fisher, B. L., Schultz, T. R., & Ward, P. S. (2019). Compositional heterogeneity and outgroup choice influence the internal phylogeny of the

ants. *Molecular Phylogenetics and Evolution*, *134*, 111–121. https://doi.org/https://doi.org/10.1016/j.ympev.2019.01.024

Bouckaert, R., & Lockhart, P. (2015). Capturing heterotachy through multi-gamma site models. *BioRxiv*, 018101. https://doi.org/10.1101/018101

Boussau, B., Szöllősi, G. J., Duret, L., Gouy, M., Tannier, E., & Daubin, V. (2013). Genome-scale coestimation of species and gene trees. *Genome Research*, *23*(2), 323–330. https://doi.org/10.1101/gr.141978.112

Brinkmann, H., & Philippe, H. (1999). Archaea sister group of Bacteria? Indications from tree reconstruction artifacts in ancient phylogenies. *Molecular Biology and Evolution*, *16*(6), 817–825. http://www.scopus.com/inward/record.url?eid=2-s2.0-0033015732&partnerID=40&md5=161b1246034ef74bd244336ad2381d67

Brinkmann, H., & Philippe, H. (2007). The diversity of eukaryotes and the root of the eukaryotic tree. *Advances in Experimental Medicine and Biology*, *607*, 20–37. https://doi.org/10.1007/978-0-387-74021-8_2

Brochier-Armanet, C., Boussau, B., Gribaldo, S., & Forterre, P. (2008). Mesophilic crenarchaeota: Proposal for a third archaeal phylum, the Thaumarchaeota. *Nature Reviews Microbiology*, *6*(3), 245–252. https://doi.org/10.1038/nrmicro1852

Brochier-Armanet, C., Forterre, P., & Gribaldo, S. (2011). Phylogeny and evolution of the Archaea: One hundred genomes later. In *Current Opinion in Microbiology* (Vol. 14, Issue 3, pp. 274–281). Elsevier Ltd. https://doi.org/10.1016/j.mib.2011.04.015

Brochier-Armanet, C., Gribaldo, S., & Forterre, P. (2008). A DNA topoisomerase IB in Thaumarchaeota testifies for the presence of this enzyme in the last common ancestor of Archaea and Eucarya. *Biology Direct*, *3*(1), 54. https://doi.org/10.1186/1745-6150-3-54

Brown, J., & Doolittle, W. (1995). Brown JR, Doolittle WF. Root of the universal tree of life based on ancient aminoacyl-tRNA synthetase gene duplications. Proc Natl Acad Sci USA 92: 2441-2445. *Proceedings of the National Academy of Sciences of the United States of America*, *92*, 2441–2445. https://doi.org/10.1073/pnas.92.7.2441

Brown, J. R., Robb, F. T., Weiss, R., & Doolittle, W. F. (1997). Evidence for the early divergence of tryptophanyl- and tyrosyl-tRNA synthetases. *Journal of Molecular Evolution*, *45*(1), 9–16. https://doi.org/10.1007/PL00006206

Brown, M. W., Heiss, A., Kamikawa, R., Inagaki, Y., Yabuki, A., Tice, A. K., Shiratori, T., Ishida, K., Hashimoto, T., Simpson, A. G. B., & Roger, A. J. (2017). Phylogenomics places orphan protistan lineages in a novel eukaryotic super-group. *BioRxiv*, 227884. https://doi.org/10.1101/227884

Bui, E. T., Bradley, P. J., & Johnson, P. J. (1996). A common evolutionary origin for mitochondria and hydrogenosomes. *Proceedings of the National Academy of Sciences of the United States of America*, *93*(18), 9651–9656. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC38483/

Burki, F., Roger, A. J., Brown, M. W., & Simpson, A. G. B. (2020). The New Tree of Eukaryotes. *Trends in Ecology & Evolution*, *35*(1), 43–55. https://doi.org/10.1016/j.tree.2019.08.008

Caetano-Anollés, G. (2002). Evolved RNA secondary structure and the rooting of the universal tree of life. *Journal of Molecular Evolution*, *54*(3), 333–345.

http://www.scopus.com/inward/record.url?eid=2-s2.0-0036182064&partnerID=40&md5=815d3db2853fefed260585f769c8dd1c

Castelle, C. J., & Banfield, J. F. (2018). Major New Microbial Groups Expand Diversity and Alter our Understanding of the Tree of Life. *Cell*, *172*(6), 1181–1197. https://doi.org/10.1016/j.cell.2018.02.016

Castelle, C. J., Wrighton, K. C., Thomas, B. C., Hug, L. A., Brown, C. T., Wilkins, M. J., Frischkorn, K. R., Tringe, S. G., Singh, A., Markillie, L. M., Taylor, R. C., Williams, K. H., & Banfield, J. F. (2015). Genomic Expansion of Domain Archaea Highlights Roles for Organisms from New Phyla in Anaerobic Carbon Cycling. *Current Biology*, *25*(6), 690–701. https://doi.org/10.1016/j.cub.2015.01.014

Cavalier-Smith, T. (1993). Kingdom protozoa and its 18 phyla. *Microbiological Reviews*, *57*(4), 953–994. http://mmbr.asm.org/content/57/4/953.abstract

Cavalier-Smith, T. (2002). The neomuran origin of archaebacteria, the negibacterial root of the universal tree and bacterial megaclassification. *International Journal of Systematic and Evolutionary Microbiology*, *52*(1), 7–76. http://ijs.sgmjournals.org/content/52/1/7.abstract

Cavalier-Smith, T. (2006). Rooting the tree of life by transition analyses. *Biology Direct*, *1*(1), 19.

Cavalier-Smith, T. (2009). Predation and eukaryote cell origins: A coevolutionary perspective. *The International Journal of Biochemistry & Cell Biology*, *41*(2), 307–322. https://doi.org/http://dx.doi.org/10.1016/j.biocel.2008.10.002

Cavalier-Smith, T., & Chao, E. E.-Y. (2020). Multidomain ribosomal protein trees and the planctobacterial origin of neomura (eukaryotes, archaebacteria). *Protoplasma*, *257*(3), 621–753. https://doi.org/10.1007/s00709-019-01442-7

Charlebois, R., Sensen, C. W., Doolittle, W., & Brown, J. (1997). Evolutionary analysis of the hisCGABdFDEHI gene cluster from the archaeon Sulfolobus solfataricus P2. *Journal of Bacteriology*, *179*, 4429–4432. https://doi.org/10.1128/jb.179.13.4429-4432.1997

Chatton, E. (1925). *Pansporella perplexa: amœbien à spores protégées parasite des daphnies: réflexions sur la biologie et la phylogénie des protozoaires*. Masson.

Chatton, É. P. L. (1938). *Titres et travaux scientifiques (1906-1937).* E. Sottano.

Cherlin, S., Heaps, S. E., Nye, T. M. W., Boys, R. J., Williams, T. A., & Embley, T. M. (2018). The Effect of Nonreversibility on Inferring Rooted Phylogenies. *Molecular Biology and Evolution*, *35*(4), 984–1002. https://doi.org/10.1093/molbev/msx294

Chistoserdova, L., Jenkins, C., Kalyuzhnaya, M. G., Marx, C. J., Lapidus, A., Vorholt, J. A., Staley, J. T., & Lidstrom, M. E. (2004). The Enigmatic Planctomycetes May Hold a Key to the Origins of Methanogenesis and Methylotrophy. *Molecular Biology and Evolution*, *21*(7), 1234–1241. https://doi.org/10.1093/molbev/msh113

Cleland, C. E., & Chyba, C. F. (2002). Defining 'Life.' *Origins of Life and Evolution of the Biosphere*, *32*(4), 387–393. https://doi.org/10.1023/A:1020503324273

Collins, L. J., Kurland, C. G., Biggs, P., & Penny, D. (2009). The Modern RNP World of Eukaryotes. *Journal of Heredity*, *100*(5), 597–604. https://doi.org/10.1093/jhered/esp064

Copeland, H. F. (1938). The Kingdoms of Organisms. *The Quarterly Review of Biology*, *13*(4). https://doi.org/10.1086/394568

Copeland, H. F. (1956). *The classification of lower organisms.*

Cox, C. J., Foster, P. G., Hirt, R. P., Harris, S. R., & Embley, T. M. (2008). The archaebacterial origin of eukaryotes. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(51), 20356–20361. https://doi.org/10.1073/pnas.0810647105

Crick, F. H. C. (1968). The origin of the genetic code. *Journal of Molecular Biology*, *38*(3), 367–379. https://doi.org/http://dx.doi.org/10.1016/0022-2836(68)90392-6

Criscuolo, A., & Gribaldo, S. (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*(1), 210. https://doi.org/10.1186/1471-2148-10-210

Crotty, S. M., Minh, B. Q., Bean, N. G., Holland, B. R., Tuke, J., Jermiin, L. S., & Haeseler, A. Von. (2020). GHOST: Recovering Historical Signal from Heterotachously Evolved Sequence Alignments. *Systematic Biology*, *69*(2), 249–264. https://doi.org/10.1093/sysbio/syz051

Csűrös, M., & Miklós, I. (2009). Streamlining and Large Ancestral Genomes in Archaea Inferred with a Phylogenetic Birth-and-Death Model. *Molecular Biology and Evolution*, *26*(9), 2087–2095. https://doi.org/10.1093/molbev/msp123

Čuboňová, L., Sandman, K., Hallam, S. J., DeLong, E. F., & Reeve, N. J. (2005). Histones in Crenarchaea. *Journal of Bacteriology*, *187*(15), 5482–5485. https://doi.org/10.1128/JB.187.15.5482-5485.2005

Cunha, V. Da, Gaia, M., Ogata, H., Jaillon, O., Delmont, T. O., & Forterre, P. (2020). Giant viruses encode novel types of actins possibly related to the origin of eukaryotic actin: the viractins. *BioRxiv*, 2020.06.16.150565. https://doi.org/10.1101/2020.06.16.150565

Da Cunha, V., Gaïa, M., & Forterre, P. (2022). The expanding Asgard archaea and their elusive relationships with Eukarya. *MLife*, *1*(1), 3–12. https://doi.org/10.1002/mlf2.12012

Da Cunha, V., Gaia, M., Gadelle, D., Nasir, A., & Forterre, P. (2017). Lokiarchaea are close relatives of Euryarchaeota, not bridging the gap between prokaryotes and eukaryotes. *PLoS Genetics*, *13*(6). https://doi.org/10.1371/journal.pgen.1006810

Da Cunha, V., Gaia, M., Ogata, H., Jaillon, O., Delmont, T. O., & Forterre, P. (2022). Giant Viruses Encode Actin-Related Proteins. *Molecular Biology and Evolution*, *39*(2), msac022. https://doi.org/10.1093/molbev/msac022

Davidov, Y., & Jurkevitch, E. (2007). How incompatibilities may have led to eukaryotic cell. *Nature*, *448*(7150), 130. http://dx.doi.org/10.1038/448130a

Davidov, Y., & Jurkevitch, E. (2009). Predation between prokaryotes and the origin of eukaryotes. *BioEssays*, *31*(7), 748–757. https://doi.org/10.1002/bies.200900018

de Duve, C. (2007). The origin of eukaryotes: a reappraisal. *Nature Reviews Genetics*, *8*(5), 395–403. https://doi.org/10.1038/nrg2071

DeLange, R. J., Green, G. R., & Searcy, D. G. (1981). A histone-like protein (HTa) from Thermoplasma acidophilum. I. Purification and properties. *Journal of Biological Chemistry*, *256*(2), 900–904. https://doi.org/https://doi.org/10.1016/S0021-9258(19)70064-7

Delsuc, F., Brinkmann, H., & Philippe, H. (2005). Phylogenomics and the reconstruction of the tree of life. *Nature Reviews Genetics*, *6*(5), 361–375. https://doi.org/10.1038/nrg1603

Derelle, R., Torruella, G., Klimeš, V., Brinkmann, H., Kim, E., Vlček, Č., Lang, B. F., & Eliáš, M. (2015). Bacterial proteins pinpoint a single eukaryotic root. *Proceedings of the National Academy of Sciences*, *112*(7), E693–E699. https://doi.org/10.1073/pnas.1420657112

Desmond, E., & Gribaldo, S. (2009). Phylogenomics of Sterol Synthesis: Insights into the Origin, Evolution, and Diversity of a Key Eukaryotic Feature. *Genome Biology and Evolution*, *1*, 364–381. https://doi.org/10.1093/gbe/evp036

Devos, D., Dokudovskaya, S., Alber, F., Williams, R., Chait, B. T., Sali, A., & Rout, M. P. (2004). Components of Coated Vesicles and Nuclear Pore Complexes Share a Common Molecular Architecture. *PLOS Biology*, *2*(12), e380-. https://doi.org/10.1371/journal.pbio.0020380

Devos, D. P. (2012). Regarding the presence of membrane coat proteins in bacteria: Confusion? What confusion? *BioEssays*, *34*(1), 38–39. https://doi.org/10.1002/bies.201100147

Devos, D. P. (2021). Reconciling Asgardarchaeota Phylogenetic Proximity to Eukaryotes and Planctomycetes Cellular Features in the Evolution of Life. In *Molecular Biology and Evolution* (Vol. 38, Issue 9, pp. 3531–3542). Oxford University Press. https://doi.org/10.1093/molbev/msab186

Devos, D. P., Gräf, R., & Field, M. C. (2014). Evolution of the nucleus. *Current Opinion in Cell Biology*, *28*(0), 8–15. https://doi.org/http://dx.doi.org/10.1016/j.ceb.2014.01.004

Devos, D. P., & Reynaud, E. G. (2010). Intermediate Steps. *Science* , *330*(6008), 1187–1188. https://doi.org/10.1126/science.1196720

Di Franco, A., Baurain, D., Glöckner, G., Melkonian, M., & Philippe, H. (2022). Lower Statistical Support with Larger Data Sets: Insights from the Ochrophyta Radiation. *Molecular Biology and Evolution*, *39*(1), msab300. https://doi.org/10.1093/molbev/msab300

Dombrowski, N., Lee, J. H., Williams, T. A., Offre, P., & Spang, A. (2019). Genomic diversity, lifestyles and evolutionary origins of DPANN archaea. *FEMS Microbiology Letters*, *366*(2). https://doi.org/10.1093/femsle/fnz008

Dombrowski, N., Teske, A. P., & Baker, B. J. (2018). Expansive microbial metabolic versatility and biodiversity in dynamic Guaymas Basin hydrothermal sediments. *Nature Communications*, *9*(1), 4999. https://doi.org/10.1038/s41467-018-07418-0

Dombrowski, N., Williams, T. A., Sun, J., Woodcroft, B. J., Lee, J.-H., Minh, B. Q., Rinke, C., & Spang, A. (2020). Undinarchaeota illuminate DPANN phylogeny and the impact of gene transfer on archaeal evolution. *Nature Communications*, *11*(1), 3939. https://doi.org/10.1038/s41467-020-17408-w

Dong, J.-H., Wen, J.-F., & Tian, H.-F. (2007). Homologs of eukaryotic Ras superfamily proteins in prokaryotes and their novel phylogenetic correlation with their eukaryotic analogs. *Gene*, *396*(1), 116–124. https://doi.org/https://doi.org/10.1016/j.gene.2007.03.001

Doolittle, W. F. (1998). You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. *Trends in Genetics*, *14*(8), 307–311. https://doi.org/http://dx.doi.org/10.1016/S0168-9525(98)01494-2

Doolittle, W. F. (1999). Phylogenetic Classification and the Universal Tree. *Science*, *284*(5423), 2124–2128. https://doi.org/10.1126/science.284.5423.2124

Doolittle, W. F., & Brown, J. R. (1994). Tempo, mode, the progenote, and the universal root. *Proceedings of the National Academy of Sciences*, *91*(15), 6721–6728. http://www.pnas.org/content/91/15/6721.abstract

Duchêne, D. A., Duchêne, S., & Ho, S. Y. W. (2017). New Statistical Criteria Detect Phylogenetic Bias Caused by Compositional Heterogeneity. *Molecular Biology and Evolution*, *34*(6), 1529–1534. https://doi.org/10.1093/molbev/msx092

Dufourc, E. J. (2008). Sterols and membrane dynamics. *Journal of Chemical Biology*, *1*(1–4), 63–77. https://doi.org/10.1007/s12154-008-0010-6

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, *26*(19), 2460–2461. https://doi.org/10.1093/bioinformatics/btq461

Elkins, J. G., Podar, M., Graham, D. E., Makarova, K. S., Wolf, Y., Randau, L., Hedlund, B. P., Brochier-Armanet, C., Kunin, V., Anderson, I., Lapidus, A., Goltsman, E., Barry, K., Koonin, E. V, Hugenholtz, P., Kyrpides, N., Wanner, G., Richardson, P., Keller, M., & Stetter, K. O. (2008). A korarchaeal genome reveals insights into the evolution of the Archaea. *Proceedings of the National Academy of Sciences*, *105*(23), 8102–8107. https://doi.org/10.1073/pnas.0801980105

Embley, M., van der Giezen, M., Horner, D. S., Dyal, P. L., & Foster, P. (2003). Mitochondria and hydrogenosomes are two forms of the same fundamental organelle. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *358*(1429), 191–203. https://doi.org/10.1098/rstb.2002.1190

Embley, T. M., & Hirt, R. P. (1998). Early branching eukaryotes? *Current Opinion in Genetics & Development*, *8*(6), 624–629. https://doi.org/http://dx.doi.org/10.1016/S0959-437X(98)80029-4

Embley, T. M., & Martin, W. (2006). Eukaryotic evolution, changes and challenges. *Nature*, *440*(7084), 623–630. http://dx.doi.org/10.1038/nature04546

Embley, T. M., & Williams, T. A. (2015). Evolution: Steps on the road to eukaryotes. *Nature*, *521*(7551), 169–170. http://dx.doi.org/10.1038/nature14522

Eme, L., Spang, A., Lombard, J., Stairs, C. W., & Ettema, T. J. G. (2017). Archaea and the origin of eukaryotes. *Nature Reviews Microbiology*, *15*(12), 711–723. https://doi.org/10.1038/nrmicro.2017.133

Eme, L., Tamarit, D., Caceres, E. F., Stairs, C. W., De Anda, V., Schön, M. E., Seitz, K. W., Dombrowski, N., Lewis, W. H., Homa, F., Saw, J. H., Lombard, J., Nunoura, T., Li, W.-J., Hua, Z.-S., Chen, L.-X., Banfield, J. F., John, E. S., Reysenbach, A.-L., … Ettema, T. J. G. (2023). Inference and reconstruction of the heimdallarchaeial ancestry of eukaryotes. *Nature*, *618*(7967), 992–999. https://doi.org/10.1038/s41586-023-06186-2

Emms, D. M., & Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology*, *16*(1), 157. https://doi.org/10.1186/s13059-015-0721-2

Emms, D. M., & Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology*, *20*(1), 238. https://doi.org/10.1186/s13059-019-1832-y

Evans, P. N., Parks, D. H., Chadwick, G. L., Robbins, S. J., Orphan, V. J., Golding, S. D., & Tyson, G. W. (2015). Methane metabolism in the archaeal phylum Bathyarchaeota revealed by genome-centric metagenomics. *Science*, *350*(6259), 434–438. https://doi.org/10.1126/science.aac7745

F, F. I., Rui, Z., & F, B. J. (2021). "Sifarchaeota," a Novel Asgard Phylum from Costa Rican Sediment Capable of Polysaccharide Degradation and Anaerobic Methylotrophy. *Applied and Environmental Microbiology*, *87*(9), e02584-20. https://doi.org/10.1128/AEM.02584-20

Fenchel, T. M., & Finlay, B. J. (1995). *Ecology and evolution in anoxic worlds*. https://api.semanticscholar.org/CorpusID:82577097

Filée, J., Forterre, P., Sen-Lin, T., & Laurent, J. (2002). Evolution of DNA polymerase families: evidences for multiple gene exchange between cellular and viral proteins. *Journal of Molecular Evolution*, *54*(6), 763–773. https://doi.org/10.1007/s00239-001-0078-x

Fillol, M., Auguet, J. C., Casamayor, E. O., & Borrego, C. M. (2016). Insights in the ecology and evolutionary history of the Miscellaneous Crenarchaeotic Group lineage. *ISME Journal*, *10*(3), 665–677. https://doi.org/10.1038/ismej.2015.143

Forterre, P. (2001). Genomics and early cellular evolution. The origin of the DNA world. *Comptes Rendus de l'Académie Des Sciences - Series III - Sciences de La Vie*, *324*(12), 1067–1076. https://doi.org/http://dx.doi.org/10.1016/S0764-4469(01)01403-2

Forterre, P. (2007). Quand les évolutionnistes découvrent l'importance des virus. *Virologie*, *11*(1), 5–12.

Forterre, P. (2011). A new fusion hypothesis for the origin of Eukarya: better than previous ones, but probably also wrong. *Research in Microbiology*, *162*(1), 77–91. https://doi.org/http://dx.doi.org/10.1016/j.resmic.2010.10.005

Forterre, P., & Philippe, H. (1998). La préhistoire du vivant. In *Curr Opin Genet Dev* (Vol. 396, Issue 2). Academic Press. www.nceas.ucsb.edu/

Forterre, P., & Philippe, H. (1999a). The Last Universal Common Ancestor (LUCA),Simple or Complex? *Biol Bull*, *196*(June), 373–377.

Forterre, P., & Philippe, H. (1999b). Where is the root of the universal tree of life? *BioEssays*, *21*(10), 871–879. https://doi.org/10.1002/(SICI)1521-1878(199910)21:10<871::AID-BIES10>3.0.CO;2-Q

Fox, G. E., Magrum, L. J., Balch, W. E., Wolfe, R. S., & Woese, C. R. (1977). Classification of methanogenic bacteria by 16S ribosomal RNA characterization. *Proceedings of the National Academy of Sciences*, *74*(10), 4537–4541. https://doi.org/10.1073/pnas.74.10.4537

Fox, G., Pechman, K. R., & Woese, C. R. (1977). Comparative cataloging of 16S ribosomal ribonucleic acid: Molecular approach to procaryotic systematics. *International Journal of Systematic Bacteriology*, *27*, 44–57. https://doi.org/10.1099/00207713-27-1-44

Francis, W. R. (2021). *The eukaryotic last common ancestor was bifunctional for hopanoid and sterol production*. www.preprints.org

Franklin, R. E., & Gosling, R. G. (1953). Molecular Configuration in Sodium Thymonucleate. *Nature*, *171*(4356), 740–741. https://doi.org/10.1038/171740a0

Fuerst, J. A., & Nisbet, E. G. (2004). Buds from the tree of life: linking compartmentalized prokaryotes and eukaryotes by a non-hyperthermophile common ancestor and implications for understanding Archaean microbial communities. *International Journal of Astrobiology*, *3*(3), 183–187. https://doi.org/DOI: 10.1017/S1473550404002150

Fuerst, JohnA. (2013). The PVC superphylum: exceptions to the bacterial definition? *Antonie van Leeuwenhoek*, *104*(4), 451–466. https://doi.org/10.1007/s10482-013-9986-1

Galtier, N., & Lobry, J. R. (1997). Relationships Between Genomic G+C Content, RNA Secondary Structures, and Optimal Growth Temperature in Prokaryotes. *Journal of Molecular Evolution*, *44*(6), 632–636. https://doi.org/10.1007/PL00006186

Ganti Tibor. (2003a). *Chemoton Theory Volume 1: Theoretical Foundations of Fluid Machineries*.

Ganti Tibor. (2003b). *Chemoton Theory Volume 2: Theory of Living Systems*.

Ganti Tibor. (2003c). *The Principles of Life*.

Garg, S. G., Kapust, N., Lin, W., Knopp, M., Tria, F. D. K., Nelson-Sathi, S., Gould, S. B., Fan, L., Zhu, R., Zhang, C., & Martin, W. F. (2021). Anomalous Phylogenetic Behavior of Ribosomal Proteins in Metagenome-Assembled Asgard Archaea. *Genome Biology and Evolution*, *13*(1). https://doi.org/10.1093/gbe/evaa238

Germot, A., Philippe, H., & Le Guyader, H. (1997). Evidence for loss of mitochondria in Microsporidia from a mitochondrial-type HSP70 in Nosema locustae1. *Molecular and Biochemical Parasitology*, *87*(2), 159–168. https://doi.org/http://dx.doi.org/10.1016/S0166-6851(97)00064-9

Germot, A., Philippe, H., & Le Guyader, H. (1996). Presence of a mitochondrial-type 70-kDa heat shock protein in Trichomonas vaginalis suggests a very early mitochondrial endosymbiosis in eukaryotes. *Proceedings of the National Academy of Sciences of the United States of America*, *93*(25), 14614–14617. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC26182/

Gest, H. (2004). The discovery of microorganisms by Robert Hooke and Antoni van Leeuwenhoek, fellows of the Royal Society. In *Notes and Records of the Royal Society* (Vol. 58, Issue 2). https://doi.org/10.1098/rsnr.2004.0055

Gilbert, W. (1978). Why genes in pieces? *Nature*, *271*(5645), 501. https://doi.org/10.1038/271501a0

Gilbert, W., Marchionni, M., & McKnight, G. (1986). On the antiquity of introns. *Cell*, *46*(2), 151–153. https://doi.org/https://doi.org/10.1016/0092-8674(86)90730-0

Gogarten, J. P., Kibak, H., Dittrich, P., Taiz, L., Bowman, E. J., Bowman, B. J., Manolson, M. F., Poole, R. J., Date, T., & Oshima, T. (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), 6661–6665. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC297905/

Golding, G. B., & Gupta, R. S. (1995). Protein-based phylogenies support a chimeric origin for the eukaryotic genome. *Molecular Biology and Evolution*, *12*(1), 1–6. https://doi.org/10.1093/oxfordjournals.molbev.a040178

Gouy, R., Baurain, D., & Philippe, H. (2015). Rooting the tree of life: The phylogenetic jury is still out. In *Philosophical Transactions of the Royal Society B: Biological Sciences* (Vol. 370, Issue 1678). Royal Society of London. https://doi.org/10.1098/rstb.2014.0329

Gray, M. W. (2015). Mosaic nature of the mitochondrial proteome: Implications for the origin and evolution of mitochondria. *Proceedings of the National Academy of Sciences*, *112*(33), 10133–10138. https://doi.org/10.1073/pnas.1421379112

Gray, M. W., Burger, G., & Lang, B. F. (1999). Mitochondrial Evolution. *Science*, *283*(5407), 1476–1481. https://doi.org/10.1126/science.283.5407.1476

Gray, M. W., Burger, G., & Lang, B. F. (2001). The origin and early evolution of mitochondria. *Genome Biology*, *2*(6), reviews1018.1. https://doi.org/10.1186/gb-2001-2-6-reviews1018

Gray, M. W., & Doolittle, W. F. (1982). Has the endosymbiont hypothesis been proven? *Microbiological Reviews*, *46*(1), 1–42. https://doi.org/10.1128/mr.46.1.1-42.1982

Gribaldo, S., & Brochier-Armanet, C. (2012). Time for order in microbial systematics. *Trends in Microbiology*, *20*(5), 209–210. https://doi.org/10.1016/j.tim.2012.02.006

Gribaldo, S., & Cammarano, P. (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*, 508–516. https://doi.org/10.1007/PL00006407

Gribaldo, S., Poole, A. M., Daubin, V., Forterre, P., & Brochier-Armanet, C. (2010). The origin of eukaryotes and their relationship with the Archaea: are we at a phylogenomic impasse? *Nature Reviews Microbiology*, *8*(10), 743–752. http://dx.doi.org/10.1038/nrmicro2426

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, *29*(8), 1072–1075. https://doi.org/10.1093/bioinformatics/btt086

Guy, L., & Ettema, T. J. G. (2011). The archaeal 'TACK' superphylum and the origin of eukaryotes. *Trends in Microbiology*, *19*(12), 580–587. https://doi.org/http://dx.doi.org/10.1016/j.tim.2011.09.002

Guy, L., Spang, A., Saw, J. H., & Ettema, T. J. G. (2014). 'Geoarchaeote NAG1' is a deeply rooting lineage of the archaeal order Thermoproteales rather than a new phylum. *The ISME Journal*, *8*(7), 1353–1357. https://doi.org/10.1038/ismej.2014.6

Haeckel, E. (1866). Generelle Morphologie der Organismen. In *Generelle Morphologie der Organismen*. https://doi.org/10.1515/9783110848281

Hannah, J. L., Bekker, A., Stein, H. J., Markey, R. J., & Holland, H. D. (2004). Primitive Os and 2316 Ma age for marine shale: implications for Paleoproterozoic glacial events and the rise of atmospheric oxygen. *Earth and Planetary Science Letters*, *225*(1), 43–52. https://doi.org/https://doi.org/10.1016/j.epsl.2004.06.013

Hartman, H., & Fedorov, A. (2002). The origin of the eukaryotic cell: A genomic investigation. *Proceedings of the National Academy of Sciences*, *99*(3), 1420–1425. https://doi.org/10.1073/pnas.032658599

He, Y., Li, M., Perumal, V., Feng, X., Fang, J., Xie, J., Sievert, S. M., & Wang, F. (2016). Genomic and enzymatic evidence for acetogenesis among multiple lineages of the archaeal phylum Bathyarchaeota widespread in marine sediments. *Nature Microbiology*, *1*(6), 16035. https://doi.org/10.1038/nmicrobiol.2016.35

Hirt, R. P., Healy, B., Vossbrinck, C. R., Canning, E. U., & Embley, T. M. (1997). A mitochondrial Hsp70 orthologue in Vairimorpha necatrix: molecular evidence that microsporidia once contained mitochondria. *Current Biology*, *7*(12), 995–998. https://doi.org/http://dx.doi.org/10.1016/S0960-9822(06)00420-9

Hirt, R. P., Logsdon, J. M., Healy, B., Dorey, M. W., Doolittle, W. F., & Embley, T. M. (1999). Microsporidia are related to Fungi: Evidence from the largest subunit of RNA polymerase II and other proteins. *Proceedings of the National Academy of Sciences* , *96*(2), 580–585. https://doi.org/10.1073/pnas.96.2.580

Hjort, K., Goldberg, A. V, Tsaousis, A. D., Hirt, R. P., & Embley, T. M. (2010). Diversity and reductive evolution of mitochondria among microbial eukaryotes. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *365*(1541), 713–727. https://doi.org/10.1098/rstb.2009.0224

Horner, D. S., Hirt, R. P., Kilvington, S., Lloyd, D., & Embley, T. M. (1996). Molecular Data Suggest an Early Acquisition of the Mitochondrion Endosymbiont. *Proceedings of the Royal Society of London B: Biological Sciences*, *263*(1373), 1053–1059. http://rspb.royalsocietypublishing.org/content/263/1373/1053.abstract

Hoshino, Y., & Gaucher, E. A. (2018). On the Origin of Isoprenoid Biosynthesis. *Molecular Biology and Evolution*, *35*(9), 2185–2197. https://doi.org/10.1093/molbev/msy120

Hosner, P. A., Faircloth, B. C., Glenn, T. C., Braun, E. L., & Kimball, R. T. (2016). Avoiding Missing Data Biases in Phylogenomic Inference: An Empirical Study in the Landfowl (Aves: Galliformes). *Molecular Biology and Evolution*, *33*(4), 1110–1125. https://doi.org/10.1093/molbev/msv347

Huber, H., Hohn, M. J., Rachel, R., Fuchs, T., Wimmer, V. C., & Stetter, K. O. (2002). A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont. *Nature*, *417*(6884), 63–67. https://doi.org/10.1038/417063a

Huelsenbeck, J. P., Bollback, J. P., & Levine, A. M. (2002). Inferring the Root of a Phylogenetic Tree. *Systematic Biology*, *51*(1), 32–43. https://doi.org/10.1080/106351502753475862

Hug, L. A., Baker, B. J., Anantharaman, K., Brown, C. T., Probst, A. J., Castelle, C. J., Butterfield, C. N., Hernsdorf, A. W., Amano, Y., Kotaro, I., Suzuki, Y., Dudek, N., Relman, D. A., Finstad, K. M., Amundson, R., Thomas, B. C., & Banfield, J. F. (2016). A new view of the tree and life's diversity. *Nature Microbiology*, *1*(April), 16048. https://doi.org/10.1038/nmicrobiol.2016.48

Hulcr, J., Latimer, A. M., Henley, J. B., Rountree, N. R., Fierer, N., Lucky, A., Lowman, M. D., & Dunn, R. R. (2012). A Jungle in There: Bacteria in Belly Buttons are Highly Diverse, but Predictable. *PLOS ONE*, *7*(11), e47712-. https://doi.org/10.1371/journal.pone.0047712

Husnik, F., Tashyreva, D., Boscaro, V., George, E. E., Lukeš, J., & Keeling, P. J. (2021). Bacterial and archaeal symbioses with protists. *Current Biology*, *31*(13), R862–R877. https://doi.org/https://doi.org/10.1016/j.cub.2021.05.049

Imachi, H., Nobu, M. K., Nakahara, N., Morono, Y., Ogawara, M., Takaki, Y., Takano, Y., Uematsu, K., Ikuta, T., Ito, M., Matsui, Y., Miyazaki, M., Murata, K., Saito, Y., Sakai, S., Song, C., Tasumi, E., Yamanaka, Y., Yamaguchi, T., … Takai, K. (2020). Isolation of an archaeon at the prokaryote–eukaryote interface. *Nature*, *577*(7791), 519–525. https://doi.org/10.1038/s41586-019-1916-6

Irisarri, I., Baurain, D., Brinkmann, H., Delsuc, F., Sire, J. Y., Kupfer, A., Petersen, J., Jarek, M., Meyer, A., Vences, M., & Philippe, H. (2017). Phylotranscriptomic consolidation of the jawed vertebrate timetree. *Nature Ecology and Evolution*, *1*(9), 1370–1378. https://doi.org/10.1038/s41559-017-0240-5

Iwabe, N., Kuma, K., Hasegawa, M., Osawa, S., & Miyata, T. (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), 9355–9359. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC298494/

James, T. Y., Pelin, A., Bonen, L., Ahrendt, S., Sain, D., Corradi, N., & Stajich, J. E. (2013). Shared Signatures of Parasitism and Phylogenomics Unite Cryptomycota and Microsporidia. *Current Biology*, *23*(16), 1548–1553. https://doi.org/http://dx.doi.org/10.1016/j.cub.2013.06.057

Javaux, E. J. (2019). Challenges in evidencing the earliest traces of life. *Nature*, *572*(7770), 451–460. https://doi.org/10.1038/s41586-019-1436-4

Jeffroy, O., Brinkmann, H., Delsuc, F., & Philippe, H. (2006). Phylogenomics: the beginning of incongruence? *Trends in Genetics*, *22*(4), 225–231. https://doi.org/10.1016/j.tig.2006.02.003

Jékely, G. (2003). Small GTPases and the evolution of the eukaryotic cell. *BioEssays*, *25*(11), 1129–1138. https://doi.org/10.1002/bies.10353

Joyce, G. F. (2002). The antiquity of RNA-based evolution. *Nature*, *418*(6894), 214–221. http://dx.doi.org/10.1038/418214a

Karnkowska, A., Vacek, V., Zubáčová, Z., Treitli, S. C., Petrželková, R., Eme, L., Novák, L., Žárský, V., Barlow, L. D., Herman, E. K., Soukal, P., Hroudová, M., Doležal, P., Stairs, C. W., Roger, A. J., Eliáš, M., Dacks, J. B., Vlček, Č., & Hampl, V. (2016). A Eukaryote without a Mitochondrial Organelle. *Current Biology*, *26*(10), 1274–1284. https://doi.org/10.1016/j.cub.2016.03.053

Katoh, K., Misawa, K., Kuma, K., & Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, *30*(14), 3059–3066. https://doi.org/10.1093/nar/gkf436

Katoh, K., & Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*, *30*(4), 772–780. https://doi.org/10.1093/molbev/mst010

Keeling, P. J. (1998). A kingdom's progress: Archezoa and the origin of eukaryotes. *BioEssays*, *20*(1), 87–95. https://doi.org/10.1002/(SICI)1521-1878(199801)20:1<87::AID-BIES12>3.0.CO;2-4

Keeling, P. J., & Burki, F. (2019). Progress towards the Tree of Eukaryotes. *Current Biology*, *29*(16), R808–R817. https://doi.org/https://doi.org/10.1016/j.cub.2019.07.031

Keeling, P. J., & Doolittle, W. F. (1996). Alpha-tubulin from early-diverging eukaryotic lineages and the evolution of the tubulin family. *Molecular Biology and Evolution* , *13*(10), 1297–1305. http://mbe.oxfordjournals.org/content/13/10/1297.abstract

Keeling, P. J., & Doolittle, W. F. (1997). Evidence that eukaryotic triosephosphate isomerase is of alpha-proteobacterial origin. *Proceedings of the National Academy of Sciences of the United States of America*, *94*(4), 1270–1275. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC19780/

Kelly, S., Wickstead, B., & Gull, K. (2011). Archaeal phylogenomics provides evidence in support of a methanogenic origin of the Archaea and a thaumarchaeal origin for the eukaryotes. *Proceedings of the Royal Society B: Biological Sciences*, *278*(1708), 1009–1018. https://doi.org/10.1098/rspb.2010.1427

Kocot, K. M., Struck, T. H., Merkel, J., Waits, D. S., Todt, C., Brannock, P. M., Weese, D. A., Cannon, J. T., Moroz, L. L., Lieb, B., & Halanych, K. M. (2017). Phylogenomics of Lophotrochozoa with Consideration of Systematic Error. *Systematic Biology*, *66*(2), 256–282. https://doi.org/10.1093/sysbio/syw079

Koonin, E. V. (2009). Intron-dominated genomes of early ancestors of eukaryotes. *The Journal of Heredity*, *100*(5), 618–623. http://europepmc.org/abstract/MED/19617525

Koonin, E. V. (2015). Origin of eukaryotes from within archaea, archaeal eukaryome and bursts of gene gain: eukaryogenesis just made easier? *Philosophical Transactions of the Royal Society B: Biological Sciences*, *370*(1678), 20140333. https://doi.org/10.1098/rstb.2014.0333

Koonin, E. V, & Martin, W. (2005). On the origin of genomes and cells within inorganic compartments. *Trends in Genetics*, *21*(12), 647–654. https://doi.org/10.1016/j.tig.2005.09.006

Kozubal, M. A., Romine, M., Jennings, R. deM, Jay, Z. J., Tringe, S. G., Rusch, D. B., Beam, J. P., McCue, L. A., & Inskeep, W. P. (2013). Geoarchaeota: a new candidate phylum in the Archaea from high-temperature acidic iron mats in Yellowstone National Park. *The ISME Journal*, *7*(3), 622–634. https://doi.org/10.1038/ismej.2012.132

Kubo, K., Lloyd, K. G., F Biddle, J., Amann, R., Teske, A., & Knittel, K. (2012). Archaea of the Miscellaneous Crenarchaeotal Group are abundant, diverse and widespread in marine sediments. *ISME Journal*, *6*(10), 1949–1965. https://doi.org/10.1038/ismej.2012.37

Kück, P., & Wägele, J. W. (2016). Plesiomorphic character states cause systematic errors in molecular phylogenetic analyses: a simulation study. *Cladistics*, *32*(4), 461–478. https://doi.org/https://doi.org/10.1111/cla.12132

Kump, L. R. (2008). The rise of atmospheric oxygen. *Nature*, *451*(7176), 277–278. https://doi.org/10.1038/nature06587

Kurland, C. G., & Andersson, S. G. E. (2000). Origin and Evolution of the Mitochondrial Proteome. *Microbiology and Molecular Biology Reviews*, *64*(4), 786–820. https://doi.org/10.1128/MMBR.64.4.786-820.2000

Kutschera, U., Levit, G. S., & Hossfeld, U. (2019). Ernst Haeckel (1834–1919): The German Darwin and his impact on modern biology. *Theory in Biosciences*. https://doi.org/10.1007/S12064-019-00276-4

Labedan, B., Boyen, A., Baetens, M., Charlier, D., Chen, P., Cunin, R., Durbeco, V., Glansdorff, N., Herve, G., Legrain, C., Liang, Z., Purcarea, C., Roovers, M., Sanchez, R., Toong, T.-L., de Casteele, M., van Vliet, F., Xu, Y., & Zhang, Y.-F. (1999). The Evolutionary History of Carbamoyltransferases: A Complex Set of Paralogous Genes Was Already Present in the Last Universal Common Ancestor. *Journal of Molecular Evolution*, *49*, 461–473. https://doi.org/10.1007/PL00006569

Lake, J. A. (2007). Disappearing act. *Nature*, *446*(7139), 983. http://dx.doi.org/10.1038/446983a

Lake, J. A., Henderson, E., Oakes, M., & Clark, M. W. (1984). Eocytes: a new ribosome structure indicates a kingdom with a close relationship to eukaryotes. *Proceedings of the National Academy of Sciences*, *81*(12), 3786–3790. http://www.pnas.org/content/81/12/3786.abstract

Lane, N., & Martin, W. (2010). The energetics of genome complexity. *Nature*, *467*(7318), 929–934. https://doi.org/10.1038/nature09486

Lartillot, N., & Philippe, H. (2004). A Bayesian Mixture Model for Across-Site Heterogeneities in the Amino-Acid Replacement Process. *Molecular Biology and Evolution*, *21*(6), 1095–1109. https://doi.org/10.1093/molbev/msh112

Lartillot, N., & Philippe, H. (2006). Computing Bayes factors using thermodynamic integration. *Systematic Biology*, *55*(2), 195–207.

Lartillot, N., Rodrigue, N., Stubbs, D., & Richer, J. (2013). PhyloBayes MPI: Phylogenetic Reconstruction with Infinite Mixtures of Profiles in a Parallel Environment. *Systematic Biology* , *62*(4), 611–615. https://doi.org/10.1093/sysbio/syt022

Lawson, F. S., Charlebois, R. L., & Dillon, J.-A. R. (1996). Phylogenetic analysis of carbamoylphosphate synthetase genes: complex evolutionary history includes an internal duplication within a gene which can root the tree of life. *Molecular Biology and Evolution*, *13 7*, 970–977.

Lazar, C. S., Baker, B. J., Seitz, K., Hyde, A. S., Dick, G. J., Hinrichs, K.-U., & Teske, A. P. (2016). Genomic evidence for distinct carbon substrate preferences and ecological niches of Bathyarchaeota in estuarine sediments. *Environmental Microbiology*, *18*(4), 1200–1211. https://doi.org/https://doi.org/10.1111/1462-2920.13142

Le, S. Q., Dang, C. C., & Gascuel, O. (2012). Modeling Protein Evolution with Several Amino Acid Replacement Matrices Depending on Site Rates. *Molecular Biology and Evolution*, *29*(10), 2921–2936. https://doi.org/10.1093/molbev/mss112

Le, S. Q., & Gascuel, O. (2008). An Improved General Amino Acid Replacement Matrix. *Molecular Biology and Evolution* , *25*(7), 1307–1320. https://doi.org/10.1093/molbev/msn067

Letunic, I., & Bork, P. (2021). Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Research*, *49*(W1), W293–W296. https://doi.org/10.1093/nar/gkab301

Levit, G. S., & Hossfeld, U. (2019). Ernst Haeckel in the history of biology. *Current Biology*, *29*(24), R1276–R1284. https://doi.org/10.1016/J.CUB.2019.10.064

Liu, Y., Makarova, K. S., Huang, W.-C., Wolf, Y. I., Nikolskaya, A. N., Zhang, X., Cai, M., Zhang, C.-J., Xu, W., Luo, Z., Cheng, L., Koonin, E. V, & Li, M. (2021a). Expanded diversity of Asgard archaea

and their relationships with eukaryotes. *Nature*, *593*(7860), 553–557. https://doi.org/10.1038/s41586-021-03494-3

Liu, Y., Makarova, K. S., Huang, W.-C., Wolf, Y. I., Nikolskaya, A. N., Zhang, X., Cai, M., Zhang, C.-J., Xu, W., Luo, Z., Cheng, L., Koonin, E. V, & Li, M. (2021b). Expanded diversity of Asgard archaea and their relationships with eukaryotes. *Nature*, *593*(7860), 553–557. https://doi.org/10.1038/s41586-021-03494-3

Lombard, J., López-García, P., & Moreira, D. (2012). The early evolution of lipid membranes and the three domains of life. *Nature Reviews Microbiology*, *10*(7), 507–515. http://dx.doi.org/10.1038/nrmicro2815

Lonhienne, T. G. A., Sagulenko, E., Webb, R. I., Lee, K.-C., Franke, J., Devos, D. P., Nouwens, A., Carroll, B. J., & Fuerst, J. A. (2010). Endocytosis-like protein uptake in the bacterium Gemmata obscuriglobus. *Proceedings of the National Academy of Sciences*, *107*(29), 12883–12888. https://doi.org/10.1073/pnas.1001085107

Lopez, P., Forterre, P., & Philippe, H. (1999). The root of the tree of life in the light of the covarion model. *Journal of Molecular Evolution*, *49*(4), 496–508. https://doi.org/10.1007/PL00006572

López-García, P., & Moreira, D. (2006). Selective forces for the origin of the eukaryotic nucleus. *BioEssays*, *28*(5), 525–533. https://doi.org/10.1002/bies.20413

López-García, P., & Moreira, D. (1999). Metabolic symbiosis at the origin of eukaryotes. *Trends in Biochemical Sciences*, *24*(3), 88–93. https://doi.org/http://dx.doi.org/10.1016/S0968-0004(98)01342-5

Lwoff, A. (1957). The concept of virus. *Journal of General Microbiology*, *17*(2). https://doi.org/10.1099/00221287-17-2-239

Lynn, M., Michael, C., Ricardo, G., & John, H. (2006). The last eukaryotic common ancestor (LECA): Acquisition of cytoskeletal motility from aerotolerant spirochetes in the Proterozoic Eon. *Proceedings of the National Academy of Sciences*, *103*(35), 13080–13085. https://doi.org/10.1073/pnas.0604985103

Lyons, T. W., Reinhard, C. T., & Planavsky, N. J. (2014). The rise of oxygen in Earth's early ocean and atmosphere. *Nature*, *506*(7488), 307–315. https://doi.org/10.1038/nature13068

Macleod, F., Kindler, G. S., Wong, H. L., Chen, R., & Burns, B. P. (2019). Asgard archaea: Diversity, function, and evolutionary implications in a range of microbiomes. In *AIMS Microbiology* (Vol. 5, Issue 1, pp. 48–61). AIMS Press. https://doi.org/10.3934/microbiol.2019.1.48

Magrum, L. J., Luehrsen, K. R., & Woese, C. R. (1978). Are extreme halophiles actually "bacteria"? *Journal of Molecular Evolution*, *11*(1), 1–8. https://doi.org/10.1007/BF01768019

Makarova, K. S., & Koonin, E. V. (2010). Two new families of the FtsZ-tubulin protein superfamily implicated in membrane remodeling in diverse bacteria and archaea. *Biology Direct*, *5*(1), 33. https://doi.org/10.1186/1745-6150-5-33

Margulis, L. (1970). *Origin of eukaryotic cells: Evidence and research implications for a theory of the origin and evolution of microbial, plant and animal cells on the precambrian Earth*. Yale University Press.

Margulis, L., Dolan, M. F., & Guerrero, R. (2000). The chimeric eukaryote: Origin of the nucleus from the karyomastigont in amitochondriate protists. *Proceedings of the National Academy of Sciences*, *97*(13), 6954–6959. https://doi.org/10.1073/pnas.97.13.6954

Martijn, J., & Ettema, T. J. (2013). From archaeon to eukaryote: the evolutionary dark ages of the eukaryotic cell. *Biochemical Society Transactions*, *41*(1), 451–457. http://europepmc.org/abstract/MED/23356327

Martin Embley, T. (2006). Multiple secondary origins of the anaerobic lifestyle in eukaryotes. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *361*(1470), 1055–1067. https://doi.org/10.1098/rstb.2006.1844

Martin, W. (1999). A briefly argued case that mitochondria and plastids are descendants of endosymbionts, but that the nuclear compartment is not. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *266*(1426), 1387–1395. https://doi.org/10.1098/rspb.1999.0792

Martin, W. (2005). Archaebacteria (Archaea) and the origin of the eukaryotic nucleus. *Current Opinion in Microbiology*, *8*(6), 630–637. https://doi.org/http://dx.doi.org/10.1016/j.mib.2005.10.004

Martin, W., & Koonin, E. V. (2006). Introns and the origin of nucleus–cytosol compartmentalization. *Nature*, *440*(7080), 41–45. http://dx.doi.org/10.1038/nature04531

Martin, W., & Muller, M. (1998). The hydrogen hypothesis for the first eukaryote. *Nature*, *392*(6671), 37–41. http://dx.doi.org/10.1038/32096

Martinez-Gutierrez, C. A., & Aylward, F. O. (2021). Phylogenetic Signal, Congruence, and Uncertainty across Bacteria and Archaea. *Molecular Biology and Evolution*, *38*(12), 5514–5527. https://doi.org/10.1093/molbev/msab254

Mayrose, I., Graur, D., Ben-Tal, N., & Pupko, T. (2004). Comparison of site-specific rate-inference methods for protein sequences: Empirical Bayesian methods are superior. *Molecular Biology and Evolution*, *21*(9), 1781–1791. https://doi.org/10.1093/molbev/msh194

McInerney, J. O., Martin, W. F., Koonin, E. V, Allen, J. F., Galperin, M. Y., Lane, N., Archibald, J. M., & Embley, T. M. (2011). Planctomycetes and eukaryotes: A case of analogy not homology. *BioEssays*, *33*(11), 810–817. https://doi.org/10.1002/bies.201100045

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., & Lanfear, R. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution*, *37*(5), 1530–1534. https://doi.org/10.1093/molbev/msaa015

Mirarab, S., & Warnow, T. (2015). ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics*, *31*(12), i44–i52. https://doi.org/10.1093/bioinformatics/btv234

Moreira, D., & López-García, P. (1998). Symbiosis Between Methanogenic Archaea and δ-Proteobacteria as the Origin of Eukaryotes: The Syntrophic Hypothesis. *Journal of Molecular Evolution*, *47*(5), 517–530. https://doi.org/10.1007/PL00006408

Muñoz-Gómez, S. A., Susko, E., Williamson, K., Eme, L., Slamovits, C. H., Moreira, D., López-García, P., & Roger, A. J. (2022). Site-and-branch-heterogeneous analyses of an expanded dataset

favour mitochondria as sister to known Alphaproteobacteria. *Nature Ecology & Evolution*, *6*(3), 253–262. https://doi.org/10.1038/s41559-021-01638-2

Nasir, A., Kim, K. M., Da Cunha, V., & Caetano-Anollés, G. (2016). Arguments Reinforcing the Three-Domain View of Diversified Cellular Life. In *Archaea* (Vol. 2016). Hindawi Publishing Corporation. https://doi.org/10.1155/2016/1851865

Nasir, A., Mughal, F., & Caetano-Anollés, G. (2021). The tree of life describes a tripartite cellular world. *BioEssays*, *43*(6). https://doi.org/10.1002/bies.202000343

Nelson-Sathi, S., Sousa, F. L., Roettger, M., Lozada-Chávez, N., Thiergart, T., Janssen, A., Bryant, D., Landan, G., Schönheit, P., Siebers, B., McInerney, J. O., & Martin, W. F. (2015). Origins of major archaeal clades correspond to gene acquisitions from bacteria. *Nature*, *517*(7532), 77–80. https://doi.org/10.1038/nature13805

Nosenko, T., Schreiber, F., Adamska, M., Adamski, M., Eitel, M., Hammel, J., Maldonado, M., Müller, W. E. G., Nickel, M., Schierwater, B., Vacelet, J., Wiens, M., & Wörheide, G. (2013). Deep metazoan phylogeny: When different genes tell different stories. *Molecular Phylogenetics and Evolution*, *67*(1), 223–233. https://doi.org/http://dx.doi.org/10.1016/j.ympev.2013.01.010

Nunoura, T., Takaki, Y., Kakuta, J., Nishi, S., Sugahara, J., Kazama, H., Chee, G.-J., Hattori, M., Kanai, A., Atomi, H., Takai, K., & Takami, H. (2011). Insights into the evolution of Archaea and eukaryotic protein modifier systems revealed by the genome of a novel archaeal group. *Nucleic Acids Research*, *39*(8), 3204–3223. https://doi.org/10.1093/nar/gkq1228

O'Malley, M. A., Leger, M. M., Wideman, J. G., & Ruiz-Trillo, I. (2019). Concepts of the last eukaryotic common ancestor. *Nature Ecology & Evolution*, *3*(3), 338–344. https://doi.org/10.1038/s41559-019-0796-3

Parks, D. H., Chuvochina, M., Waite, D. W., Rinke, C., Skarshewski, A., Chaumeil, P.-A., & Hugenholtz, P. (2018). A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nature Biotechnology*, *36*(10), 996–1004. https://doi.org/10.1038/nbt.4229

Penny, D., Hoeppner, M. P., Poole, A. M., & Jeffares, D. C. (2009). An Overview of the Introns-First Theory. *Journal of Molecular Evolution*, *69*(5), 527–540. https://doi.org/10.1007/s00239-009-9279-5

Peretó, J., López-García, P., & Moreira, D. (2004). Ancestral lipid biosynthesis and early membrane evolution. *Trends in Biochemical Sciences*, *29*(9), 469–477. https://doi.org/10.1016/j.tibs.2004.07.002

Petitjean, C., Deschamps, P., López-Garciá, P., & Moreira, D. (2014). Rooting the domain archaea by phylogenomic analysis supports the foundation of the new kingdom Proteoarchaeota. *Genome Biology and Evolution*, *7*(1), 191–204. https://doi.org/10.1093/gbe/evu274

Peyretaillade, E., Broussolle, V., Peyret, P., Méténier, G., Gouy, M., & Vivarès, C. P. (1998). Microsporidia, amitochondrial protists, possess a 70-kDa heat shock protein gene of mitochondrial evolutionary origin. *Molecular Biology and Evolution* , *15*(6), 683–689. http://mbe.oxfordjournals.org/content/15/6/683.abstract

Philippe, H., Delsuc, F., Brinkmann, H., & Lartillot, N. (2005). Phylogenomics. In *Annual Review of Ecology, Evolution, and Systematics* (Vol. 36, pp. 541–562). https://doi.org/10.1146/annurev.ecolsys.35.112202.130205

Philippe, H., & Forterre, P. (1999). The rooting of the universal tree of life is not reliable. *Journal of Molecular Evolution*, *49*(4), 509–523. https://doi.org/10.1007/PL00006573

Philippe, H., Germot, A., Le Guyader, H., Adoutte, A., Française, S., Président, G., Président, · A Nicolas, Jacob, F., Berger, R., Rinon, H., Stoll Secrétaire, C., Sohgnac Trésorier, M., & Sinet, P.-M. (1995). Que savons-nous de l'histoire évolutive des eucaryotes ? 1. L'arbre universel du vivant et les difficultés de la reconstruction phylogénétique Vice-présidents. *Medecines Sciences*, *11*(8), 1–13.

Philippe, H., Germot, A., & Moreira, D. (2000). The new phylogeny of eukaryotes. *Current Opinion in Genetics & Development*, *10*(6), 596–601. https://doi.org/http://dx.doi.org/10.1016/S0959-437X(00)00137-4

Philippe, H., Vienne, D. M. de, Ranwez, V., Roure, B., Baurain, D., & Delsuc, F. (2017). Pitfalls in supermatrix phylogenomics. *European Journal of Taxonomy*, *0*(283). https://doi.org/10.5852/ejt.2017.283

Pilhofer, M., Ladinsky, M. S., McDowall, A. W., Petroni, G., & Jensen, G. J. (2011). Microtubules in Bacteria: Ancient Tubulins Build a Five-Protofilament Homolog of the Eukaryotic Cytoskeleton. *PLOS Biology*, *9*(12), e1001213-. https://doi.org/10.1371/journal.pbio.1001213

Podolsky, S. H., & Tauber, A. I. (1994). Origins of Life: The Central Concepts. David W. Deamer , Gail R. Fleischaker. *The Quarterly Review of Biology*, *69*(2), 253–254. https://doi.org/10.1086/418549

Pollock, D. D., Zwickl, D. J., McGuire, J. A., & Hillis, D. M. (2002). Increased taxon sampling is advantageous for phylogenetic inference. *Systematic Biology*, *51*(4), 664–671. https://doi.org/10.1080/10635150290102357

Poole, A., Jeffares, D., & Penny, D. (1999). Early evolution: Prokaryotes, the new kids on the block. *BioEssays*, *21*(10), 880–889. https://doi.org/10.1002/(SICI)1521-1878(199910)21:10<880::AID-BIES11>3.0.CO;2-P

Poole, A. M., & Penny, D. (2007). Evaluating hypotheses for the origin of eukaryotes. *BioEssays*, *29*(1), 74–84. https://doi.org/10.1002/bies.20516

Poole, A., & Penny, D. (2001). Does endosymbiosis explain the origin of the nucleus? *Nature Cell Biology*, *3*(8), E173--E173. http://dx.doi.org/10.1038/35087102

Poole, A., & Penny, D. (2007). Eukaryote evolution: Engulfed by speculation. *Nature*, *447*(7147), 913. http://dx.doi.org/10.1038/447913a

Poulton, S. W., Fralick, P. W., & Canfield, D. E. (2004). The transition to a sulphidic ocean ~ 1.84 billion years ago. *Nature*, *431*(7005), 173–177. https://doi.org/10.1038/nature02912

Preston, C. M., Wu, K. Y., Molinski, T. F., & DeLong, E. F. (1996). A psychrophilic crenarchaeon inhabits a marine sponge: Cenarchaeum symbiosum gen. nov., sp. nov. *Proceedings of the National Academy of Sciences*, *93*(13), 6241–6246. https://doi.org/10.1073/pnas.93.13.6241

Probst, A. J., Weinmaier, T., Raymann, K., Perras, A., Emerson, J. B., Rattei, T., Wanner, G., Klingl, A., Berg, I. A., Yoshinaga, M., Viehweger, B., Hinrichs, K. U., Thomas, B. C., Meck, S., Auerbach, A. K., Heise, M., Schintlmeister, A., Schmid, M., Wagner, M., … Moissl-Eichinger, C. (2014).

Biology of a widespread uncultivated archaeon that contributes to carbon fixation in the subsurface. *Nature Communications*, *5*. https://doi.org/10.1038/ncomms6497

Qu, X.-J., Jin, J.-J., Chaw, S.-M., Li, D.-Z., & Yi, T.-S. (2017). Multiple measures could alleviate long-branch attraction in phylogenomic reconstruction of Cupressoideae (Cupressaceae). *Scientific Reports*, *7*(1), 41005. https://doi.org/10.1038/srep41005

Ragan, M. A. (2009). Trees and networks before and after Darwin. *Biology Direct*, *4*(1), 43. https://doi.org/10.1186/1745-6150-4-43

Raymann, K., Brochier-Armanet, C., & Gribaldo, S. (2015). The two-domain tree of life is linked to a new root for the Archaea. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(21), 6670–6675. https://doi.org/10.1073/pnas.1420858112

Reynaud, E. G., & Devos, D. P. (2011). Transitional forms between the three domains of life and evolutionary implications. *Proceedings of the Royal Society B: Biological Sciences* , *278*(1723), 3321–3328. https://doi.org/10.1098/rspb.2011.1581

Richards, E. J., Brown, J. M., Barley, A. J., Chong, R. A., & Thomson, R. C. (2018). Variation Across Mitochondrial Gene Trees Provides Evidence for Systematic Error: How Much Gene Tree Variation Is Biological? *Systematic Biology*, *67*(5), 847–860. https://doi.org/10.1093/sysbio/syy013

Rinke, C., Chuvochina, M., Mussig, A. J., Chaumeil, P.-A., Davín, A. A., Waite, D. W., Whitman, W. B., Parks, D. H., & Hugenholtz, P. (2021). A standardized archaeal taxonomy for the Genome Taxonomy Database. *Nature Microbiology*, *6*(7), 946–959. https://doi.org/10.1038/s41564-021-00918-8

Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N. N., Anderson, I. J., Cheng, J. F., Darling, A., Malfatti, S., Swan, B. K., Gies, E. A., Dodsworth, J. A., Hedlund, B. P., Tsiamis, G., Sievert, S. M., Liu, W. T., Eisen, J. A., Hallam, S. J., Kyrpides, N. C., Stepanauskas, R., … Woyke, T. (2013). Insights into the phylogeny and coding potential of microbial dark matter. *Nature*, *499*(7459), 431–437. https://doi.org/10.1038/nature12352

Rivas-Marín, E., & Devos, D. P. (2018). The Paradigms They Are a-Changin': past, present and future of PVC bacteria research. *Antonie van Leeuwenhoek*, *111*(6), 785–799. https://doi.org/10.1007/s10482-017-0962-z

Rivera, M. C. (2007). Genomic analyses and the origin of the eukaryotes. *Chemistry and Biodiversity*, *4*(11), 2631–2638. https://doi.org/10.1002/cbdv.200790215

Rivera, M. C., Jain, R., Moore, J. E., & Lake, J. A. (1998). Genomic evidence for two functionally distinct gene classes. *Proceedings of the National Academy of Sciences*, *95*(11), 6239–6244. http://www.pnas.org/content/95/11/6239.abstract

Rivera, M. C., & Lake, J. A. (1992). Evidence that eukaryotes and eocyte prokaryotes are immediate relatives. *Science (New York, N.Y.)*, *257*(5066), 74–76. http://europepmc.org/abstract/MED/1621096

Rivera, M. C., & Lake, J. A. (2004). The ring of life provides evidence for a genome fusion origin of eukaryotes. *Nature*, *431*(7005), 152–155. http://dx.doi.org/10.1038/nature02848

Rodrigues-Oliveira, T., Wollweber, F., Ponce-Toledo, R. I., Xu, J., Rittmann, S. K.-M. R., Klingl, A., Pilhofer, M., & Schleper, C. (2023). Actin cytoskeleton and complex cell architecture in an

Asgard archaeon. *Nature*, *613*(7943), 332–339. https://doi.org/10.1038/s41586-022-05550-y

Roger, A. J. (1999). Reconstructing early events in eukaryotic evolution. *American Naturalist*, *154*(4 SUPPL.), S146–S163. http://www.scopus.com/inward/record.url?eid=2-s2.0-0032702602&partnerID=40&md5=3653006777c96da395d1670f5c42894b

Roger, A. J., Clark, C. G., & Doolittle, W. F. (1996). A possible mitochondrial gene in the early-branching amitochondriate protist Trichomonas vaginalis. *Proceedings of the National Academy of Sciences of the United States of America*, *93*(25), 14618–14622. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC26183/

Roger, A. J., Muñoz-Gómez, S. A., & Kamikawa, R. (2017). The Origin and Diversification of Mitochondria. *Current Biology*, *27*(21), R1177–R1192. https://doi.org/https://doi.org/10.1016/j.cub.2017.09.015

Rokas, A., & Carroll, S. B. (2005). More genes or more taxa? The relative contribution of gene number and taxon number to phylogenetic accuracy. *Molecular Biology and Evolution*, *22*(5), 1337–1344. https://doi.org/10.1093/molbev/msi121

Rotte, C., & Martin, W. (2001). Does endosymbiosis explain the origin of the nucleus? *Nature Cell Biology*, *3*(8), E173--E173. http://dx.doi.org/10.1038/35087104

Roure, B., Baurain, D., & Philippe, H. (2013). Impact of Missing Data on Phylogenies Inferred from Empirical Phylogenomic Data Sets. *Molecular Biology and Evolution*, *30*(1), 197–214. https://doi.org/10.1093/molbev/mss208

Roure, B., & Philippe, H. (2011). Site-specific time heterogeneity of the substitution process and its impact on phylogenetic inference. *BMC Evolutionary Biology*, *11*(1). https://doi.org/10.1186/1471-2148-11-17

Roure, B., Rodriguez-Ezpeleta, N., & Philippe, H. (2007). SCaFoS: a tool for selection, concatenation and fusion of sequences for phylogenomics. *BMC Evolutionary Biology*, *7*(Suppl 1), S2.

Sanger, F. (1959). Chemistry of Insulin. *Science*, *129*(3359), 1340–1344. https://doi.org/10.1126/science.129.3359.1340

Santana-Molina, C., Rivas-Marin, E., Rojas, A. M., & Devos, D. P. (2020). Origin and Evolution of Polycyclic Triterpene Synthesis. *Molecular Biology and Evolution*, *37*(7), 1925–1941. https://doi.org/10.1093/molbev/msaa054

Santarella-Mellwig, R., Franke, J., Jaedicke, A., Gorjanacz, M., Bauer, U., Budd, A., Mattaj, I. W., & Devos, D. P. (2010). The compartmentalized bacteria of the planctomycetes-verrucomicrobia- chlamydiae superphylum have membrane coat-like proteins. *PLoS Biology*, *8*(1). https://doi.org/10.1371/journal.pbio.1000281

Santarella-Mellwig, R., Pruggnaller, S., Roos, N., Mattaj, I. W., & Devos, D. P. (2013). Three-Dimensional Reconstruction of Bacteria with a Complex Endomembrane System. *PLOS Biology*, *11*(5), e1001565-. https://doi.org/10.1371/journal.pbio.1001565

Sapp, J. (2005). The Prokaryote-Eukaryote Dichotomy: Meanings and Mythology. *Microbiology and Molecular Biology Reviews*, *69*(2), 292–305. https://doi.org/10.1128/MMBR.69.2.292-305.2005

Sapp, J. (2006). Two faces of the prokaryote concept. *International Microbiology*, *9*(3), 163–172.

Sato, N. (2021). Are Cyanobacteria an Ancestor of Chloroplasts or Just One of the Gene Donors for Plants and Algae? *Genes*, *12*(6). https://doi.org/10.3390/genes12060823

Scamardella, J. M. (1999). Not plants or animals: a brief history of the origin of Kingdoms Protozoa, Protista and Protoctista. *International Microbiology*, *2*(4), 207–216.

Schmitt, S. (2009). Haeckel, un darwinien allemand ? *Comptes Rendus Biologies*, *332*(2–3), 110–118. https://doi.org/10.1016/J.CRVI.2008.07.006

Schrodinger, E. (2012). What is Life?: With Mind and Matter and Autobiographical Sketches. In *Canto Classics*. Cambridge University Press. https://doi.org/DOI: 10.1017/CBO9781107295629

Schwartz, R. M., & Dayhoff, M. O. (1978). Origins of prokaryotes, eukaryotes, mitochondria, and chloroplasts. *Science* , *199*(4327), 395–403. https://doi.org/10.1126/science.202030

Shen, Y., Knoll, A. H., & Walter, M. R. (2003). Evidence for low sulphate and anoxia in a mid-Proterozoic marine basin. *Nature*, *423*(6940), 632–635. https://doi.org/10.1038/nature01651

Shi, J., Zhang, Y., Luo, H., & Tang, J. (2010). Using jackknife to assess the quality of gene order phylogenies. *BMC Bioinformatics*, *11*(1), 168. https://doi.org/10.1186/1471-2105-11-168

Shimodaira, H. (2002). An Approximately Unbiased Test of Phylogenetic Tree Selection. *Systematic Biology*, *51*(3), 492–508. https://doi.org/10.1080/10635150290069913

Shiratori, T., Suzuki, S., Kakizawa, Y., & Ishida, K. (2019). Phagocytosis-like cell engulfment by a planctomycete bacterium. *Nature Communications*, *10*(1), 5529. https://doi.org/10.1038/s41467-019-13499-2

Si Quang, L., Gascuel, O., & Lartillot, N. (2008). Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics* , *24*(20), 2317–2323. https://doi.org/10.1093/bioinformatics/btn445

Sibbald, S. J., & Archibald, J. M. (2020). Genomic Insights into Plastid Evolution. *Genome Biology and Evolution*, *12*(7), 978–990. https://doi.org/10.1093/gbe/evaa096

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, *31*(19), 3210–3212. https://doi.org/10.1093/bioinformatics/btv351

Simion, P., Philippe, H., Baurain, D., Jager, M., Richter, D. J., Di Franco, A., Roure, B., Satoh, N., Quéinnec, É., Ereskovsky, A., Lapébie, P., Corre, E., Delsuc, F., King, N., Wörheide, G., & Manuel, M. (2017). A Large and Consistent Phylogenomic Dataset Supports Sponges as the Sister Group to All Other Animals. *Current Biology*, *27*(7), 958–967. https://doi.org/10.1016/j.cub.2017.02.031

Siu-Ting, K., Torres-Sánchez, M., San Mauro, D., Wilcockson, D., Wilkinson, M., Pisani, D., O'Connell, M. J., & Creevey, C. J. (2019). Inadvertent Paralog Inclusion Drives Artifactual Topologies and Timetree Estimates in Phylogenomics. *Molecular Biology and Evolution*, *36*(6), 1344–1356. https://doi.org/10.1093/molbev/msz067

Smith, B. T., Mauck III, W. M., Benz, B. W., & Andersen, M. J. (2020). Uneven Missing Data Skew Phylogenomic Relationships within the Lories and Lorikeets. *Genome Biology and Evolution*, *12*(7), 1131–1147. https://doi.org/10.1093/gbe/evaa113

Sogin, M. L. (1997). Organelle origins: Energy-producing symbionts in early eukaryotes? *Current Biology*, *7*(5), R315--R317. https://doi.org/http://dx.doi.org/10.1016/S0960-9822(06)00147-3

Spang, A., Caceres, E. F., & Ettema, T. J. G. (2017). Genomic exploration of the diversity, ecology, and evolution of the archaeal domain of life. *Science*, *357*(6351), eaaf3883. https://doi.org/10.1126/science.aaf3883

Spang, A., Hatzenpichler, R., Brochier-Armanet, C., Rattei, T., Tischler, P., Spieck, E., Streit, W., Stahl, D. A., Wagner, M., & Schleper, C. (2010). Distinct gene set in two different lineages of ammonia-oxidizing archaea supports the phylum Thaumarchaeota. *Trends in Microbiology*, *18*(8), 331–340. https://doi.org/10.1016/j.tim.2010.06.003

Spang, A., Saw, J. H., Jorgensen, S. L., Zaremba-Niedzwiedzka, K., Martijn, J., Lind, A. E., van Eijk, R., Schleper, C., Guy, L., & Ettema, T. J. G. (2015). Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature*, *521*(7551), 173–179. http://dx.doi.org/10.1038/nature14447

Stairs, C. W., & Ettema, T. J. G. (2020). The Archaeal Roots of the Eukaryotic Dynamic Actin Cytoskeleton. *Current Biology*, *30*(10), R521–R526. https://doi.org/10.1016/j.cub.2020.02.074

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312–1313. https://doi.org/10.1093/bioinformatics/btu033

Stanier, R. Y., & Van Niel, C. B. (1962). The Concept of a Bacterium. *Archiv Für Mikrobiologie*, *42*, 17–35.

Struck, T. H. (2013). The Impact of Paralogy on Phylogenomic Studies – A Case Study on Annelid Relationships. *PLoS ONE*, *8*.

Struck, T. H. (2014). Trespex-detection of misleading signal in phylogenetic reconstructions based on tree information. *Evolutionary Bioinformatics*, *10*, 51–67. https://doi.org/10.4137/EBo.s14239

Sun, J., Evans, P. N., Gagen, E. J., Woodcroft, B. J., Hedlund, B. P., Woyke, T., Hugenholtz, P., & Rinke, C. (2021). Recoding enhances the metabolic capabilities of two novel methylotrophic Asgardarchaeota lineages. *BioRxiv*, 2021.02.19.431964. https://doi.org/10.1101/2021.02.19.431964

Takemura, M. (2001). Poxviruses and the Origin of the Eukaryotic Nucleus. *Journal of Molecular Evolution*, *52*(5), 419–425. https://doi.org/10.1007/s002390010171

Thomas, C. M., Desmond-Le Quéméner, E., Gribaldo, S., & Borrel, G. (2022). Factors shaping the abundance and diversity of the gut archaeome across the animal kingdom. *Nature Communications*, *13*(1), 3358. https://doi.org/10.1038/s41467-022-31038-4

Tyagi, N., Anamika, K., & Srinivasan, N. (2010). A Framework for Classification of Prokaryotic Protein Kinases. *PLOS ONE*, *5*(5), e10608-. https://doi.org/10.1371/journal.pone.0010608

Vacek, V., Novák, L. V. F., Treitli, S. C., Táborský, P., Čepička, I., Kolísko, M., Keeling, P. J., & Hampl, V. (2018). Fe–S Cluster Assembly in Oxymonads and Related Protists. *Molecular Biology and Evolution*, *35*(11), 2712–2718. https://doi.org/10.1093/molbev/msy168

Valentine, D. L. (2007). Adaptations to energy stress dictate the ecology and evolution of the Archaea. *Nature Reviews Microbiology*, *5*(4), 316–323. https://doi.org/10.1038/nrmicro1619

Van Vlierberghe, M., Di Franco, A., Philippe, H., & Baurain, D. (2021). Decontamination, pooling and dereplication of the 678 samples of the Marine Microbial Eukaryote Transcriptome Sequencing Project. *BMC Research Notes*, *14*(1), 306. https://doi.org/10.1186/s13104-021-05717-2

Van Vlierberghe, M., Philippe, H., & Baurain, D. (2021). Broadly sampled orthologous groups of eukaryotic proteins for the phylogenetic study of plastid-bearing lineages. *BMC Research Notes*, *14*(1). https://doi.org/10.1186/s13104-021-05553-4

Vanwonterghem, I., Evans, P. N., Parks, D. H., Jensen, P. D., Woodcroft, B. J., Hugenholtz, P., & Tyson, G. W. (2016). Methylotrophic methanogenesis discovered in the archaeal phylum Verstraetearchaeota. *Nature Microbiology*, *1*(12), 16170. https://doi.org/10.1038/nmicrobiol.2016.170

Vellai, T., Takács, K., & Vida, G. (1998). A New Aspect to the Origin and Evolution of Eukaryotes. *Journal of Molecular Evolution*, *46*(5), 499–507. https://doi.org/10.1007/PL00006331

Vellai, T., & Vida, G. (1999). The origin of eukaryotes: The difference between prokaryotic and eukaryotic cells. *Proceedings. Biological Sciences / The Royal Society*, *266*, 1571–1577. https://doi.org/10.1098/rspb.1999.0817

Villanueva, L., von Meijenfeldt, F. A. B., Westbye, A. B., Yadav, S., Hopmans, E. C., Dutilh, B. E., & Damsté, J. S. S. (2021). Bridging the membrane lipid divide: bacteria of the FCB group superphylum have the potential to synthesize archaeal ether lipids. *The ISME Journal*, *15*(1), 168–182. https://doi.org/10.1038/s41396-020-00772-2

Villarreal, L. P., & DeFilippis, V. R. (2000). A Hypothesis for DNA Viruses as the Origin of Eukaryotic Replication Proteins. *Journal of Virology*, *74*(15), 7079–7084. https://doi.org/10.1128/JVI.74.15.7079-7084.2000

Wagner, M., & Horn, M. (2006). The Planctomycetes, Verrucomicrobia, Chlamydiae and sister phyla comprise a superphylum with biotechnological and medical relevance. *Current Opinion in Biotechnology*, *17*(3), 241–249. https://doi.org/http://dx.doi.org/10.1016/j.copbio.2006.05.005

Wallace, D. C. (1982). Structure and evolution of organelle genomes. *Microbiological Reviews*, *46*(2), 208–240. https://doi.org/10.1128/mr.46.2.208-240.1982

Wang, H. C., Minh, B. Q., Susko, E., & Roger, A. J. (2018). Modeling Site Heterogeneity with Posterior Mean Site Frequency Profiles Accelerates Accurate Phylogenomic Estimation. *Systematic Biology*, *67*(2), 216–235. https://doi.org/10.1093/sysbio/syx068

Waterhouse, R. M., Seppey, M., Simao, F. A., Manni, M., Ioannidis, P., Klioutchnikov, G., Kriventseva, E. V., & Zdobnov, E. M. (2018). BUSCO applications from quality assessments to gene prediction and phylogenomics. *Molecular Biology and Evolution*, *35*(3), 543–548. https://doi.org/10.1093/molbev/msx319

Waters, E., Hohn, M. J., Ahel, I., Graham, D. E., Adams, M. D., Barnstead, M., Beeson, K. Y., Bibbs, L., Bolanos, R., Keller, M., Kretz, K., Lin, X., Mathur, E., Ni, J., Podar, M., Richardson, T., Sutton, G. G., Simon, M., Söll, D., … Noordewier, M. (2003). The genome of Nanoarchaeum equitans:

Insights into early archaeal evolution and derived parasitism. *Proceedings of the National Academy of Sciences*, *100*(22), 12984–12988. https://doi.org/10.1073/pnas.1735403100

Watson, J. D., & Crick F H C. (1953). Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature*, *171*(4356), 737–738. https://doi.org/10.1038/171737a0

Whittaker, R. H. (1969). New concepts of kingdoms of organisms. *Science*, *163*(3863). https://doi.org/10.1126/science.163.3863.150

Wiegand, S., Jogler, M., & Jogler, C. (2018). On the maverick Planctomycetes. *FEMS Microbiology Reviews*, *42*(6), 739–760. https://doi.org/10.1093/femsre/fuy029

Williams, B. A. P., Hirt, R. P., Lucocq, J. M., & Embley, T. M. (2002). A mitochondrial remnant in the microsporidian Trachipleistophora hominis. *Nature*, *418*(6900), 865–869. http://dx.doi.org/10.1038/nature00949

Williams, T. A., Cox, C. J., Foster, P. G., Szöllősi, G. J., & Embley, T. M. (2020). Phylogenomics provides robust support for a two-domains tree of life. *Nature Ecology & Evolution*, *4*(1), 138–147. https://doi.org/10.1038/s41559-019-1040-x

Williams, T. A., Foster, P. G., Nye, T. M. W., Cox, C. J., & Embley, T. M. (2012). A congruent phylogenomic signal places eukaryotes within the Archaea. *Proceedings of the Royal Society B: Biological Sciences*, *279*, 4870–4879. https://doi.org/10.1098/rspb.2012.1795

Williams, T. A., Heaps, S. E., Cherlin, S., Nye, T. M. W., Boys, R. J., & Embley, T. M. (2015). New substitution models for rooting phylogenetic trees. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *370*(1678), 20140336. https://doi.org/10.1098/rstb.2014.0336

Williams, T. A., Szöllosi, G. J., Spang, A., Foster, P. G., Heaps, S. E., Boussau, B., Ettema, T. J. G., & Martin Embley, T. (2017). Integrative modeling of gene and genome evolution roots the archaeal tree of life. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(23), E4602–E4611. https://doi.org/10.1073/pnas.1618463114

Woese, C., Magrum, L. J., & Fox, G. (1978). Archaebacteria. *Journal of Molecular Evolution*, *11*, 245–251. https://doi.org/10.1007/BF01734485

Woese, C. R., & Fox, G. E. (1977). Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proceedings of the National Academy of Sciences*, *74*(11), 5088–5090. https://doi.org/10.1073/pnas.74.11.5088

Woese, C. R., Kandler, O., & Wheelis, M. L. (1990). Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proceedings of the National Academy of Sciences*, *87*(12), 4576–4579. https://doi.org/10.1073/pnas.87.12.4576

Woese, CarlR., & Fox, GeorgeE. (1977). The concept of cellular evolution. *Journal of Molecular Evolution*, *10*(1), 1–6. https://doi.org/10.1007/BF01796132

Wollman, A. J. M., Nudd, R., Hedlund, E. G., & Leake, M. C. (2015). From animaculum to single molecules: 300 years of the light microscope. In *Open Biology* (Vol. 5, Issue 4). https://doi.org/10.1098/rsob.150019

Xie, R., Wang, Y., Huang, D., Hou, J., Li, L., Hu, H., Zhao, X., & Wang, F. (2022). Expanding Asgard members in the domain of Archaea sheds new light on the origin of eukaryotes. *Science China Life Sciences*, *65*(4), 818–829. https://doi.org/10.1007/s11427-021-1969-6

Yarza, P., Yilmaz, P., Pruesse, E., Glöckner, F. O., Ludwig, W., Schleifer, K. H., Whitman, W. B., Euzéby, J., Amann, R., & Rosselló-Móra, R. (2014). Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nature Reviews Microbiology*, *12*(9), 635–645. https://doi.org/10.1038/nrmicro3330

Young, A. D., & Gillung, J. P. (2020). Phylogenomics — principles, opportunities and pitfalls of big-data phylogenetics. In *Systematic Entomology* (Vol. 45, Issue 2, pp. 225–247). Blackwell Publishing Ltd. https://doi.org/10.1111/syen.12406

Yutin, N., Makarova, K. S., Mekhedov, S. L., Wolf, Y. I., & Koonin, E. V. (2008). The Deep Archaeal Roots of Eukaryotes. *Molecular Biology and Evolution*, *25*(8), 1619–1630. https://doi.org/10.1093/molbev/msn108

Yutin, N., Wolf, M., Wolf, Y., & Koonin, E. (2009). The origins of phagocytosis and eukaryogenesis. *Biology Direct*, *4*(1), 1–26. https://doi.org/10.1186/1745-6150-4-9

Zaremba-Niedzwiedzka, K., Caceres, E. F., Saw, J. H., Bäckström, Di., Juzokaite, L., Vancaester, E., Seitz, K. W., Anantharaman, K., Starnawski, P., Kjeldsen, K. U., Stott, M. B., Nunoura, T., Banfield, J. F., Schramm, A., Baker, B. J., Spang, A., & Ettema, T. J. G. (2017). Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature*, *541*(7637), 353–358. https://doi.org/10.1038/nature21031

Zhaxybayeva, O., Lapierre, P., & Gogarten, J. P. (2005). Ancient gene duplications and the root(s) of the tree of life. *Protoplasma*, *227*(1), 53–64. https://doi.org/10.1007/s00709-005-0135-1

Zhong, B., Deusch, O., Goremykin, V. V, Penny, D., Biggs, P. J., Atherton, R. A., Nikiforova, S. V, & Lockhart, P. J. (2011). Systematic Error in Seed Plant Phylogenomics. *Genome Biology and Evolution*, *3*, 1340–1348. https://doi.org/10.1093/gbe/evr105

Zhou, Z., Pan, J., Wang, F., Gu, J. D., & Li, M. (2018). Bathyarchaeota: Globally distributed metabolic generalists in anoxic environments. In *FEMS Microbiology Reviews* (Vol. 42, Issue 5, pp. 639–655). Oxford University Press. https://doi.org/10.1093/femsre/fuy023

Zhu, Q., Mai, U., Pfeiffer, W., Janssen, S., Asnicar, F., Sanders, J. G., Belda-Ferre, P., Al-Ghalith, G. A., Kopylova, E., McDonald, D., Kosciolek, T., Yin, J. B., Huang, S., Salam, N., Jiao, J.-Y., Wu, Z., Xu, Z. Z., Cantrell, K., Yang, Y., … Knight, R. (2019). Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea. *Nature Communications*, *10*(1), 5477. https://doi.org/10.1038/s41467-019-13443-4

Zillig, W. (1991). Comparative biochemistry of Archaea and Bacteria. *Current Opinion in Genetics & Development*, *1*(4), 544–551. https://doi.org/http://dx.doi.org/10.1016/S0959-437X(05)80206-0

Zillig, W., Klenk, H., Palm, P., Leffers, H., Pühler, G., Gropp, F., Garrett, R. A., Biochemie, M., & Martinsried, D.-. (1989). Did eukaryotes originate by a fusion event ? *Endocytobiosis & Cell Research*, *6*, 1–25.

Zuckerkandl, E., & Pauling, L. (1965). *Evolutionary Divergence and Convergence, in Proteins*.

Zwickl, D. J., & Hillis, D. M. (2002). Increased taxon sampling greatly reduces phylogenetic error. *Systematic Biology*, *51*(4), 588–598. https://doi.org/10.1080/10635150290102339