UNIVERSITY OF LIEGE (BELGIUM)

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

PHYLOGENOMICS OF ARCHAEA & RELATIONSHIPS WITH EUKARYOTES

DOCTORAL THESIS
Submitted for the degree of Doctor of Science
Defended by

Richard Thierry GOUY

Funding: F.R.S.-FNRS FRIA

Thesis defended on November 21, 2024 before a jury composed of:

Thesis supervisor
Prof. Denis BAURAIN, InBioS, ULiège
Jury chairman
Prof. Patrick MEYER, InBioS, ULiège
Reviewers
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

To my family, who saw me complete this thesis,
who have supported me
throughout my dream of becoming a researcher

ACKNOWLEDGMENTS

I am deeply grateful to my PhD supervisor, Professor Denis Baurain, for his invaluable support and mentorship throughout this intellectually challenging yet rewarding journey. His guidance and encouragement have been instrumental in helping me navigate the complexities of this project.

I would also like to express my gratitude to Hervé Philippe for enabling me to publish my first article and for his advice in constructing the dataset. My thanks extend to the members of my thesis committee, who have consistently believed in this project and allowed me to see it through to completion.

I am equally indebted to my colleagues in the lab: Amandine, Raphaël, Mick, Luc, Loïc, Catherine, and Valérian. The moments we shared are among the most memorable and cherished of my journey. My sincere thanks also go to Rosa and Damien for their technical and logistical support.

Finally, I would like to thank my family—my parents and my brother—for their unwavering love and support, which have been my foundation throughout this endeavor.

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.

TABLE OF CONTENTS

INTRODUCTION

1 DEFINING LIFE

« If no one asks me, I know; but if I want to explain it to someone, I do not know. »
Saint Augustin d'Hippone, 4th century.

This thought by Saint Augustine concerned the definition of time. However, today, it could just as easily be paraphrased for life, as defining it remains an elusive endeavor. The transition from inanimate to living matter is unclear. A fundamental question about the origin of life is understanding the evolutionary steps that allowed the transition from complex prebiotic chemistry to simple biology.

Currently, three major molecules encode and transmit genetic information: two families of nucleic acids (ribonucleic acid (RNA) and deoxyribonucleic acid (DNA)) and proteins composed of around twenty different amino acids. The central dogma of molecular biology, introduced in the 1950s by Francis Crick (co-discoverer of DNA's structure, which earned him the Nobel Prize in 1962), established the link between genetic material (DNA and RNA) within the cell and the proteins it synthesizes. This dogma states that in all living beings (at least current ones), genetic information flows in a single direction: from DNA to proteins via RNA, a transient structure enabling the information to be conveyed to a translation machinery that produces proteins—the basic building blocks that enable the cell, tissues, organs, and entire organisms to function (Watson & Crick F H C, 1953). This transfer of information is based on the genetic code, universal to all living beings.

Ultimately, what is life? Definitions of life can generally be grouped into three broad categories:

- Self-replication and evolution: This definition, adopted officially by NASA, states: “Life is a self-sustaining chemical system capable of Darwinian evolution” (Joyce, 2002; Podolsky & Tauber, 1994). A potential criticism of this definition is that it excludes entities incapable of reproducing (e.g., sterile hybrids like mules, worker ants, etc.).
- Functional characteristics: Based on the presence of necessary and sufficient characteristics, this view was formalized in the "chemoton"(Ganti Tibor, 2003c, 2003a, 2003b), model by Tibor Ganti, which posits that life relies on three properties: metabolism, self-catalysis/self-replication, and a membrane made of a lipid bilayer. In this model, a confined structure with an envelope capable of developing metabolism is essential. According to P.L. Luisi: “The minimal form of life is a system enclosed by a semi-permeable compartment, capable of self-organizing and self-maintaining by producing its own components through the transformation of energy and external nutrients with its own production mechanisms” (Cleland & Chyba, 2002). This is the focus of prebiotic chemistry experiments that explore pathways leading from inorganic to organic substances or biomolecules.
- Physicochemical perspective: This idea, proposed by Schrödinger (Schrodinger, 2012), characterizes living beings as "negative entropy points" or "dissipative entropy systems." A living being is capable of extracting energy from its environment to organize matter, thereby temporarily counteracting the second law of thermodynamics. Living beings are highly localized negative entropy zones in space and time. This definition, which detaches itself from strictly terrestrial conditions, could be a good reference for exobiologists seeking signs of life elsewhere in the universe using non-Earth-like chemistry.

Interestingly, systems generally considered non-living can partially fit into these definitions: viruses (obligate parasites), certain proteins like prions, liposomes, artificial protocells, self-replicating ribozymes, some crystals, and even phenomena like fire.

2 DESCRIBING LIFE

2.1 FROM SYSTEMATICS…

Systematics aims to name, describe, and classify all living beings, facilitating the study of life's diversity. This reflects the inherent order in nature. Attempts to classify life date back to antiquity, where efforts were made to organize the living world systematically.

Aristotle (384–322 BC), considered the father of zoology, was the first to propose a zoological system (in *History of Animals and Parts of Animals*), describing and classifying the animal kingdom (Sapp, 2005, 2006). He distinguished animals with blood (vertebrates) and those without (invertebrates) and described humans as "beings endowed with reason." His classification was based on an ascending complexity of organisms, from the simplest to the most complex, placing humans at the top. This hierarchical view, organizing life linearly, was later termed the scala naturae (the "ladder of nature"). This hierarchical view of life influenced thinkers like Theophrastus (372–287 BC), a student of Aristotle, who is considered the father of botany. In his works (*History of Plants* and *Research on Plants*), he attempted to classify plants systematically. Later, Pliny the Elder (23–79 AD) in Rome expanded on these ideas in his *Natural History*, which described the animal and plant kingdoms, integrating utilitarian criteria (e.g., agricultural, medicinal uses) but without a true classification spirit.

The Middle Ages brought little innovation in the classification of life. Ancient texts acquired a mystical aura and were heavily influenced by creationist and fixist views supported by monotheistic religions. Life was perceived as immutable, with classification systems reflecting a static vision of the living world.

In the late 17th century, Joseph Pitton de Tournefort introduced a more precise system of classification: “First name, then classify,” aiming to standardize plant names. Around the same period, in England, John Ray addressed the problem of species definition, basing it on the necessity of resemblance between parents and offspring. He also suggested that individuals from different species could not produce viable, fertile offspring. Ray pioneered fossil studies, challenging biblical accounts of the Great Flood by asserting that fossils were remnants of once-living organisms.

In the 18th century, Carl Linnaeus (1707–1778) revolutionized classification by publishing *Systema Naturae* in 1735. He introduced a hierarchical taxonomy (class, order, genus, and species) based on similarities among species. Linnaeus emphasized that many organisms shared common traits and controversially placed humans alongside apes within the primates. In 1758, Linnaeus standardized the binomial nomenclature for species, with a genus and a species name in Latin.

2.2 … TO THE THEORY OF EVOLUTION

For Linnaeus, species were considered fixed and immutable. His classification was based on the fixity of traits. Linnaeus, whose father was a pastor and who had studied theology, was therefore a firm believer in fixism, believing in a "sovereign order of nature”. He classified humans and apes in the same order, *Anthropomorpha*, against religious authorities who saw man as the ultimate result of divine creation. However, for him, the goal was not to question the intellectual and moral superiority of man over the ape, as both were the result of Creation and thus as perfect and immutable as each other. At the same time, the idea was developing that species were not fixed and immutable, but rather that they transform and evolve. Georges-Louis Leclerc, Comte de Buffon, used the criteria of interbreeding and sterility to describe species. He considered individuals that could produce fertile offspring as belonging to the same species, and those that were sterile as not belonging to the same species. Current species were seen as the continuation of previous species. By developing this idea, Buffon was one of the precursors of transformism. In his *Natural History*, in 44 volumes (the first volume published in 1744), he attempted to write an encyclopedia describing all the biology and geology of his time. He proposed that African and Asian elephants were descendants of the mammoths, whose fossils had recently been discovered in Siberia. He also discussed the formation of the Earth and the origin of life, estimating the age of the Earth to be 74,000 years. Although we now know this age to be far from the truth, his work not only allowed for the emancipation of ideas from the Bible (6000 years) but also introduced a new factor to be considered in the transformation of species: time.

In 1789, Antoine-Laurent de Jussieu published a botanical classification, developing the principle of hierarchical categorization of traits. To identify a given taxon, for example, a class, the ideal is to have one (or several) constant trait(s) within this class, and different traits in all other classes. Thus, a type of trait is useful at a specific level of classification, some at the order level, others at the genus level, and so on. Traits are therefore "subordinated."

In 1809, Jean-Baptiste Pierre Antoine de Monet, Chevalier de Lamarck, in his *Zoological Philosophy*, introduced the concept of gradual evolution of species: this is transformism. He is notably, with the German Treviranus, credited with coining the term "biology." His transformist theory of the inheritance of acquired traits is based on two principles:

1. The increasing complexity of the organization of living beings, from the simplest to the most complex;
2. Their diversification, or speciation, into species, following an adaptation of their behavior or organs to their environment.

To explain the second principle, Lamarck proposed two laws:

1. The "law of use and disuse," often summarized as "function creates the organ," states that after more frequent and sustained use of an organ, it gradually develops according to its use, while, conversely, it deteriorates when not used.
2. The inheritance of acquired traits, which refers to the possibility of transmitting the morphological or organic changes acquired during life to the offspring, related to the first law.

This expert in the classification of mollusks, worms, and insects observed a set of changes (evolution) between fossils of mollusks and living mollusks he had grouped in his classification. He proposed that new species form when animals and plants adapt to a changing environment, altering their organs and appearance from generation to generation through strengthening. His vision of life was based on the assumption of the inheritance of acquired traits. He believed to see a sense in this evolution, which would have occurred from the simplest forms to the most complex, but without explaining the mechanism that allowed living beings to adapt to their environment, other than by postulating that an organ used intensively would develop while another, rarely used, would regress and disappear. Lamarck was heavily criticized by his contemporaries, but he was the first to clearly publish the idea, supported by observations of living beings and fossils, that species are subject to evolution over time.

In 1796, Étienne Geoffroy Saint-Hilaire proposed that nature had formed all living beings according to a single organizational plan, with numerous variations, suggesting a common origin for all species. Cuvier, a student of Geoffroy, opposed Lamarck and Geoffroy Saint-Hilaire's theory of transformism. For him, the appearance and disappearance of species were the result of sudden catastrophes that wiped out species, leaving space for new ones to emerge. This theory of catastrophism could explain the appearance and disappearance of species without affecting their immutability. He also adopted Jussieu's method, applying it to animals with his law of the subordination of organs. A founder of comparative anatomy, Cuvier postulated that the organs of an animal are functionally dependent on each other and made this a principle of classification, which he applied to the study of fossils.

In 1859, in *On the Origin of Species*, Charles Darwin popularized the depiction of life as a phylogenetic tree (**Figure 1**), thereby challenging Aristotle's *scala naturae*, and developing the idea of a branching biological evolution based on natural selection (Ragan, 2009). At the same time, Alfred R. Wallace (father of biogeography) reached the same conclusions as Darwin, highlighting sexual selection as an evolutionary process. It was a letter from Wallace that prompted Darwin to "hasten" the publication of his theory of evolution. Thus, for Darwin and Wallace, the similarities or differences in traits should be considered the result of evolutionary phenomena.

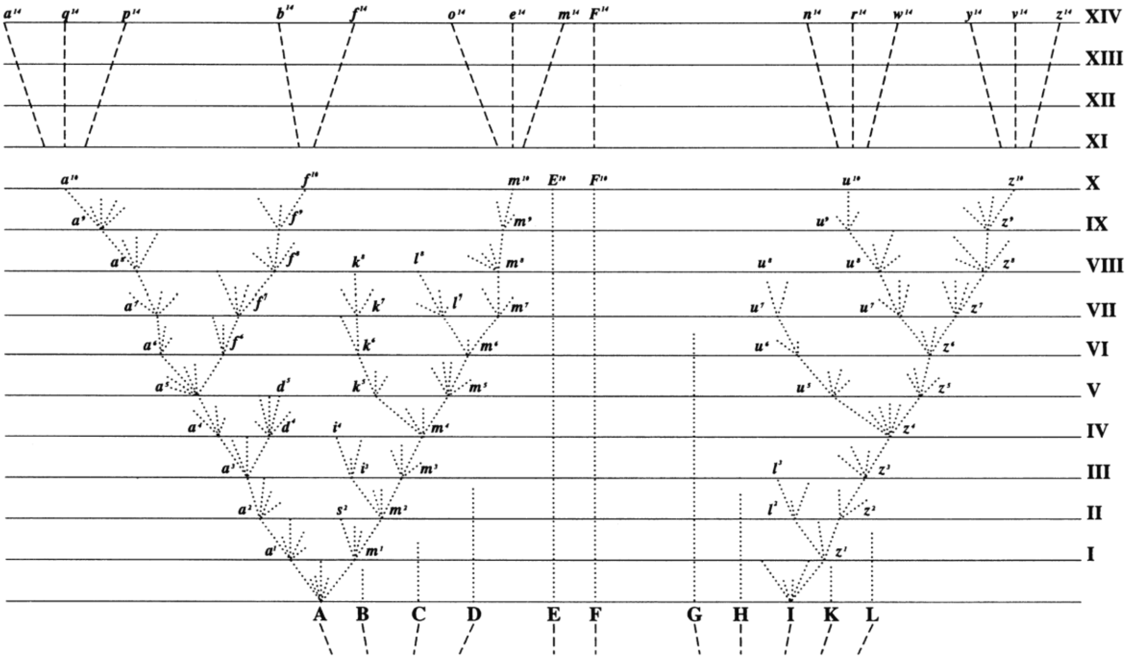


Figure 1. The unique figure from Darwin's book *On the Origin of Species by Means of Natural Selection* (1859).
This illustration highlights the process of descent with modification. This process establishes genealogical links between living beings. Ancestral species and forms of life are placed at the base of this representation. Under the effect of natural selection, the members of a species diverge over generations by accumulating variations, giving rise to new lineages. Only variations that proved advantageous were preserved by natural selection. The intervals between the horizontal lines represent significant numbers of generations. This is why contemporary species (at the top of the diagram) are connected to their ancestors through a long series of now-extinct intermediate lineages. The whole forms an evolutionary tree.

In 1866, in *Generelle Morphologie*, Ernst Haeckel, a staunch supporter of Darwin's ideas, also addressed a topic that Darwin had ignored in On the Origin of Species, but which is essential for a complete picture of evolution: the origin of life. Haeckel postulated that life on Earth originated from an "archegonium," meaning the spontaneous generation of the most primitive organisms without structure (monera) from inorganic matter. Thus, according to Haeckel, the initial appearance of all life was polyphyletic, and living matter appeared directly from inorganic chemical substances, not from previously generated organic substances (Kutschera et al., 2019; Levit & Hossfeld, 2019; Schmitt, 2009). He thus published the first phylogenetic trees (**Figure 2**), although he still placed man at the top of his trees. As with his contemporaries, the legacy of Aristotle's thinking was still present. His trees were drawn with the simplest organisms at the root, considered primitive, up to the most complex beings at the terminal branches, in line with his "biogenetic law," according to which ontogeny recapitulates phylogeny. Man is thus the "terminal product" of evolution. He proposed that the Monera were at the lowest stage of a third kingdom, which he called Protista (Haeckel, 1866). Haeckel thus divided life into three kingdoms: plants, animals, and protists, the latter corresponding to unicellular forms.

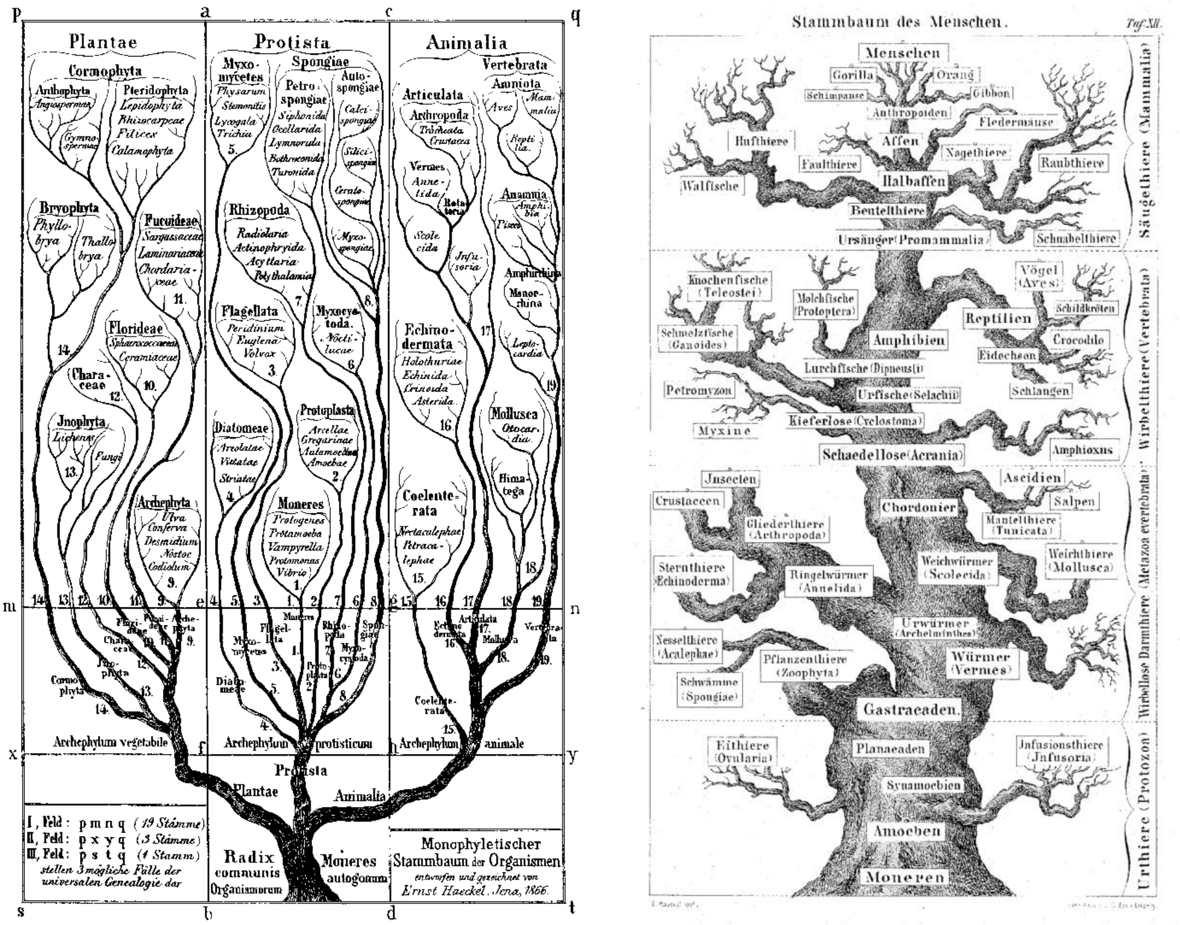


Figure 2. Haeckel's Trees (a) 1866 version showing plants, protists, and animals; (b) 1874 version leading to the human species.

In 1866, the German naturalist Ernst Haeckel created a unique evolutionary tree that grouped all known living organisms into three major kingdoms: animals, plants, and also protists. Under this latter term, he grouped unicellular organisms such as flagellates, amoebae, diatoms, sponges, and bacteria (then called monera). This tree, considered one of the first modern phylogenetic trees, illustrates the diversity of life and the evolutionary relationships between different forms of life. In 1874, Haeckel continued to explore these ideas in his work *Anthropogenie*, where he sketched a similar tree to illustrate the evolution of humans and their connection to other forms of life. These trees helped popularize the theory of evolution and paved the way for future research on the diversity and history of life on Earth.

2.3 FROM MICROBIOLOGY…

Since antiquity, it has been explained that life could appear almost anywhere, not just through organisms of their own species (reproduction), but also in mud, plants, or decomposing matter: this is the theory of spontaneous generation of living organisms. This idea, inherited from the Greeks, holds that life is a property of matter and can arise whenever the conditions are favorable, thus responding to a divine creative will. Until the 17th century, the microscopic world (invisible) was not yet known. Scientists did not offer an explanation for the diversity of living organisms, mostly because they held a fixist view. In 1668, Francesco Redi demonstrated that maggots did not form spontaneously in decaying meat, contrary to what was commonly believed up until then; rather, they were fly larvae that developed in a favorable environment, not the result of an imagined case of spontaneous generation. At the same time, the Dutch draper and amateur naturalist Antoni van Leeuwenhoek began to refine microscope lenses to check the purity of fabrics (Gest, 2004; Wollman et al., 2015). He was the first to discover bacteria and various protozoa and made the first descriptions of red blood cells and sperm. Redi and van Leeuwenhoek were among the first to challenge the theory of spontaneous generation. Robert Hooke, for his part, also built his own microscope and named the "cells" he observed. He was one of the first to propose the idea that the Earth's surface had undergone transformations over time and that fossils were the remains of species that had existed in earlier eras. This led to numerous controversies within the scientific community of the time, known as the "boiling battle." While belief in spontaneous generation for visible living organisms began to fade, controversies between scientists, philosophers, and religious figures persisted and intensified until the 19th century. It was only with Louis Pasteur, who developed a rigorous experimental protocol for sterilization, that the view of the living world was dramatically altered and the final blow was dealt to the theory of spontaneous generation in favor of the cellular theory. This theory states that the cell is the structural unit, the functional unit, and the reproductive unit of life. Today, the main points of consensus in cellular theory are as follows:

- All living organisms are composed of one or more cells (excluding viruses from this category).
- Every cell comes from a pre-existing cell through cellular division.
- The cell is a living unit and the basic unit of life, meaning a cell is a more or less autonomous entity capable of performing certain necessary and sufficient functions for its life.
- Cellular individuality is maintained by the plasma membrane, which regulates exchanges between the cell and its environment.
- The cell contains DNA that provides the information necessary for its functioning and reproduction.

Although the discovery of unicellular organisms dates back to the 17th century (Adoutte et al., 1996; Philippe et al., 1995), and the first description of the microbial world dates to the 19th century, the distinction between prokaryotes and eukaryotes is relatively recent in biology and closely linked to the advances in microscopy, initially photon-based and later primarily electron microscopy. It was Edouard Chatton, in 1925, who was the first to differentiate two major types of cells (cells with or without a nucleus), which he named "eukaryotes" and "prokaryotes" respectively (E. Chatton, 1925; É. P. L. Chatton, 1938; Sapp, 2005; Scamardella, 1999). However, his idea was initially rejected (Sapp, 2005).

The following year, in 1938, Herbert Copeland proposed that Haeckel's Monera have their own kingdom, based on the idea that they are “relatively little modified descendants of any single form of life that appeared on Earth, and that they are distinctly different from protists due to the absence of nuclei” (Copeland, 1938, 1956). He introduced a four-kingdom classification system: Monera, Protista, Plantae, and Animalia.

In 1957, André Lwoff, a student of Edouard Chatton, distinguished between viruses and bacteria by showing that viruses contain only one type of nucleic acid (RNA or DNA) and that they do not reproduce by division like a cell (Lwoff, 1957).

In 1962, Roger Yate Stanier and Cornelius Bernardus Van Niel revisited Chatton's idea in "The Concept of a Bacterium" and showed that they could only define bacteria in a negative manner relative to eukaryotes, their main characteristic being the absence of internal membranes and genetic material surrounded by a nuclear membrane (Stanier & Van Niel, 1962). The only positive definition of a bacterium is the presence of a cell wall made of peptidoglycan (bacteria without a wall, such as Mollicutes, are the result of secondary simplification).

In 1969, Robert Harding Whittaker published a new classification of life into five kingdoms (Whittaker, 1969) : Monera (bacteria + blue-green algae (= Cyanobacteria)), Fungi (= fungi), Protista, Plantae, and Animalia (**Figure 3**). His classification took into account the type and level of organization, cellular complexity (prokaryote, eukaryote uni- or multicellular), and type of nutrition (photosynthesis, absorption, and ingestion). He considered bacteria to be the lowest life forms from which more complex life forms gradually developed, leading to multicellular organisms. He was also the first to consider Fungi as a separate group, due to their absorption-based nutrition. Whittaker's schema contains uncertainties and ambiguities because the author chose not to form strictly monophyletic groups but instead to group living beings into five major "kingdoms," defined by ecological characteristics or organizational grades (Adoutte et al., 1996). The groups are represented in a vertical stack, as if they were derived from one another, although certain lineages do extend into another group. While his classification created serious ambiguities by forming a non-natural (i.e., non-phylogenetic) classification, it nonetheless gained significant success.

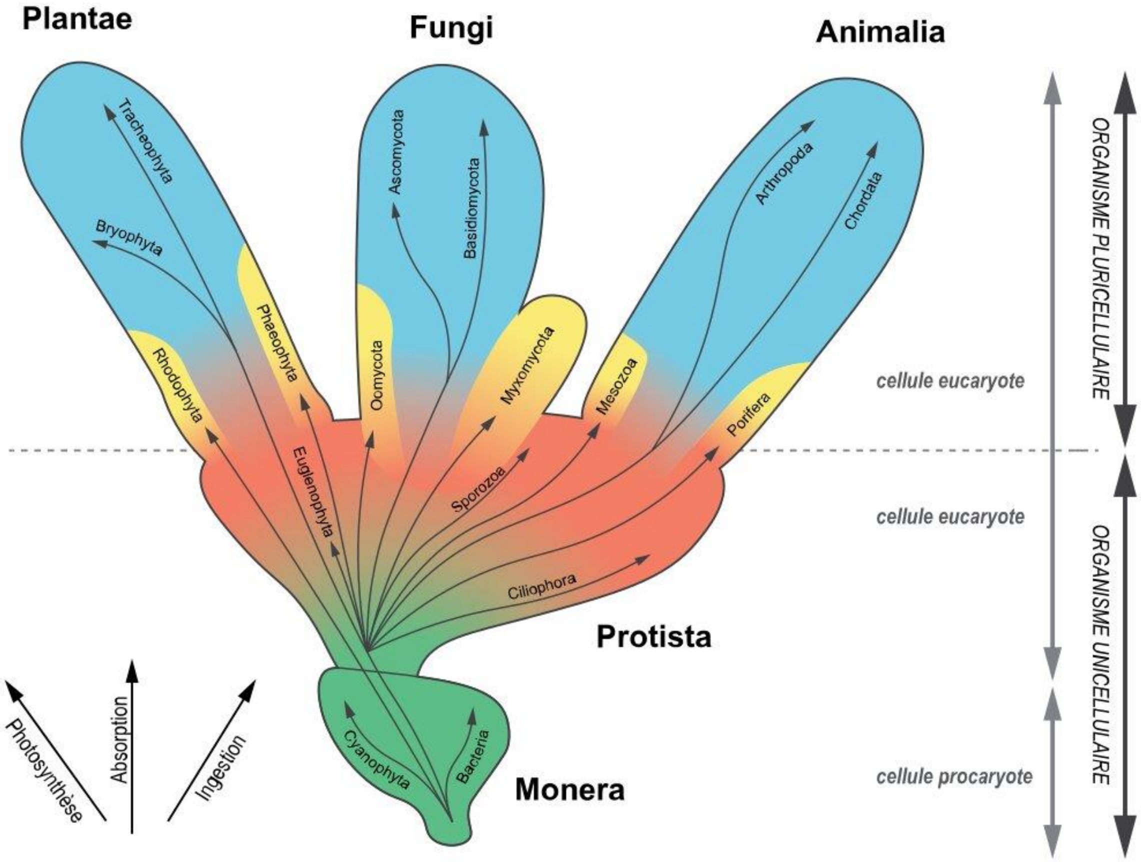


Figure 3. Five-Kingdom System According to Robert Whittaker (Whittaker, 1969).
In 1969, Robert Harding Whittaker, a prominent American ecologist, proposed an innovative classification of life into five kingdoms: Monera, Protista, Fungi, Plantae, and Animalia. His classification takes into account the type and level of organization, cellular complexity, as well as the type of nutrition. He considered Monera to be the most primitive life forms from which progressively more complex life forms developed, including Protists, and then multicellular organisms (Fungi, Plantae, and Animalia).

2.4 … TO THE MOLECULAR REVOLUTION

The molecular revolution began in 1953 with the discovery of the double helix structure of DNA by Rosalind Franklin (Franklin & Gosling, 1953), later confirmed by James Dewey Watson and Francis Crick (which earned them the Nobel Prize in Physiology or Medicine in 1962) (Watson & Crick F H C, 1953). Crick introduced the central dogma of molecular biology (**Box 1**) by establishing the link between the genetic material (DNA and RNA) contained within the cell and the proteins that the cell synthesizes (Crick, 1968). Thus, genetic information is transmitted in a one-way direction: DNA → RNA → protein. He did, however, mention an exception to the dogma: retroviruses, which can reverse the flow from RNA to DNA with the help of an enzyme, reverse transcriptase. At the time, he was unable to support this statement with evidence, but it was later proven to be true.

Box 1. The Central Dogma of Molecular Biology

Cells contain many components, the main ones being water, inorganic ions, small molecules and their precursors (sugars, amino acids, nucleotides), fatty acids and their precursors, and macromolecules (polysaccharides, proteins, nucleic acids…). Proteins and nucleic acids are long linear chains of small molecules (polymers): amino acids for proteins, nucleotides for nucleic acids. Among nucleic acids, there are two categories: deoxyribonucleic acid (DNA) and ribonucleic acids (RNA). DNA and RNA are made up of a chain of β-D-nucleoside 5'-phosphates. In other words, they consist of four molecules called monophosphate nucleotides linked together by 3'-5' phosphodiester bonds. Nucleotides are formed by the association of a sugar (β-D-ribofuranose for RNA, β-D-2'-deoxyribofuranose for DNA), a phosphoric acid, and a heterocyclic nitrogenous base. The nitrogenous bases are heterocycles derived from purine (adenine (A) and guanine (G)) or pyrimidine (cytosine (C), thymine (T), and uracil (U)). Three of these bases are common to both DNA and RNA: adenine, guanine, and cytosine. The fourth differs depending on the type of nucleic acid: uracil for RNA and thymine for DNA (which is actually a methylated form of uracil).

Some of these bases have the particularity of pairing with each other. Thus, guanine pairs with cytosine through three hydrogen bonds, while adenine pairs with thymine (in DNA) and uracil (in RNA) through two hydrogen bonds. However, RNA molecules have a very high diversity of nucleotides, with over a hundred modified nucleotides currently known. Moreover, while DNA is present as a double-stranded helix, RNA is (with rare exceptions, including some viruses) single-stranded. On the same strand of RNA, bases can pair with each other and form secondary structures, which play a central role in the function of all RNAs, imposing a three-dimensional structure dictated by the base sequence.

It is the order of these nucleotides that determines, in the form of a code, all the information necessary to confer species their properties and individuals their unique characteristics. The information contained in DNA is organized into units called genes. A gene can be considered as a sequence of DNA that specifies the synthesis of a polypeptide chain (protein) or a functional ribonucleic acid (RNA). More specifically, a gene is defined as a functional nucleotide segment of DNA (or RNA in certain viruses) that contains genetic information determining the structure of a polypeptide chain (or several isomers) or the structure of an RNA molecule. A gene encodes functional molecules (proteins or RNA) that contain signals necessary for regulating its expression depending on cellular conditions. Thus, each gene corresponds to a nucleotide sequence that determines the order of amino acids in a particular protein, and thus its function.

The genetic code establishes the rules for translating this information to produce proteins. The information carried by a gene is first copied in the form of RNA, called messenger RNA (mRNA): this is transcription. This mRNA, which has a temporary lifespan, carries out either structural functions, enzymatic functions, or functions of transporting genetic information. In the latter case, it directs protein synthesis via a ribonucleoprotein complex (i.e., composed of RNA and proteins) called the ribosome: this is the translation of the information, where each triplet of nucleotides in the mRNA induces the incorporation of an amino acid via a complementary transfer RNA (tRNA). RNA molecules are therefore functional intermediaries that facilitate the expression of DNA as a protein. This is known as the flow of genetic information, or the central dogma of molecular biology, as presented by Francis Crick in 1958. DNA is considered the stable and transmissible carrier of genetic information, which defines the biological functions of an organism. Its replication also ensures genetic continuity from one generation to the next.

However, the central dogma does not predict that one can reverse the flow of information from RNA to DNA. Yet, an enzyme dependent on RNA, reverse transcriptase, has been found in retroviruses and retrotransposons. Reverse transcriptase is capable of using an RNA strand as a template and catalyzing its reverse transcription into complementary DNA (cDNA). Moreover, some RNA molecules also have the ability to self-replicate. Indeed, certain RNA viruses replicate through an RNA intermediate. To do this, they encode an RNA-dependent RNA polymerase. Without completely overturning the central dogma of molecular biology, this discovery has provided a new perspective on the possibilities of life and opened new research avenues regarding the origins of the current genetic system.

References

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

In 1959, Frederick Sanger determined that the double polypeptide chain of insulin is made up of a precise sequence of amino acids, leading him to formulate the idea that all proteins are composed of a unique amino acid sequence (Sanger, 1959). He won his first Nobel Prize in 1958 for his work on the structure of proteins. On the other hand, a DNA sequencing method bears his name: the Sanger method. This earned him a second Nobel Prize in 1980. Today, his technique is less commonly used for genomic studies in favor of High-Throughput Sequencing (HTS) and Third-generation Sequencing (TGS). However, it remains widely used in molecular biology approaches.

In 1964, Emile Zuckerkandl and Linus Pauling demonstrated that the sequence of proteins contains a wealth of information about ancient evolutionary history (Zuckerkandl & Pauling, 1965). It is, in fact, modified by mutations that lead from an ancestral organism to its descendants. Therefore, the sequence of the same protein differs more from one organism to another the less closely related they are. This discovery laid the foundation for phylogeny, which is the study of the evolutionary relationships between living organisms (both current and extinct). They were pioneers in comparing amino acids by studying the phylogeny of primates based on hemoglobin sequences, which also led them to introduce the concept of the "molecular clock." These authors also rightly suggested that the same would likely apply to the sequences of nucleic acids (DNA and RNA), which were technically impossible to "read" at the time. Molecular phylogeny was born (**Box 2**). This new discipline, which reconstructed the relationships between organisms based on the comparison of sequences, revolutionized our understanding of the origins of current phyla. Thus, using genes, we attempt to trace the history of organisms.

Box 2. Molecular Evolution & phylogeny

Evolutionary Forces & molecular variation

The Austrian biologist Emile Zuckerkandl and the American biochemist Linus Pauling were the first to focus on the mechanisms of molecular variation. Although very stable from one generation to the next, the genetic heritage does vary, as evidenced by the differences observed between the genomes of various species. Comparing the genetic sequences of two species allows us to establish the relationships between them. The study of variations in molecular evolution rates is a powerful tool for measuring the evolution of genomes and species. Historically, the first phylogenies representing the evolution of species were based on morphological traits. But today, advances in genetics allow us to transcend this approach by comparing their protein or nucleic acid sequences. The traits of individuals (and thus of species) are linked to genes. Variations in traits, which are the raw material of evolution, are linked to variations in genes. Molecular variations can be observed within individuals of the same species (polymorphism) or between distinct species (divergence). The analysis of polymorphism data is part of population genetics, while the study of divergences falls under phylogeny.

Various evolutionary forces generate and influence molecular variations. The first of these variations can occur through sudden changes in one or more genes: mutations. In each generation, within a population, an individual may present a different state at a given position in its genome compared to its parents. This is when a new allele appears, creating a polymorphism at the corresponding locus. Various types of mutations are recognizable: nucleotide changes, insertions or deletions, elongations/shortenings of repeated motifs, gene or genomic segment duplications, transpositions, and inversions. These mutations can be non-coding, synonymous (the coding sequence is modified, but the protein remains unchanged), or non-synonymous (the protein differs). Mutations between purines (A <-> G) or pyrimidines (C <-> T) are called transitions, while mutations that change a purine to a pyrimidine (or vice versa) are called transversions. Mutations that provide individuals with a reproductive advantage are passed on and spread, while others disappear; this is our second evolutionary force: natural selection. Our third evolutionary force is genetic drift, which corresponds to random variations in allele frequencies, driven by the randomness of reproduction. Selective drift is the only force that applies to neutral loci, which are those whose variations do not influence the reproductive success of individuals. Furthermore, genes can be more or less physically linked on chromosomes. This physical linkage results in genetic linkage, meaning that certain alleles at different loci are preferentially associated, forming a combination of alleles called a haplotype. Thus, selective events occurring on neighboring genes will interfere with each other. Recombination, and more specifically the crossovers it generates, is the process that "breaks" these associations between genes, making their evolution less dependent on each other. Finally, the demography and structure of populations also play a role in molecular variations at the intraspecific level. Demographic effects (bottlenecks, selective sweeps, migration, panmixia, etc.) apply to the entire genome, while selective effects remain localized.

Molecular phylogeny has several advantages over comparing morphological traits: (i) first, the number of character states is fixed (four nucleotides and twenty amino acids), and their universality means that they can be compared across all living organisms; (ii) furthermore, the evolution of proteins and nucleic acid sequences follows a more or less regular pattern, which allows the use of mathematical models to formalize their changes; (iii) finally, the genomes of all organisms consist of very long nucleotide sequences from which we can deduce protein sequences, providing an immense amount of information that surpasses that provided by morphological traits.

Search for Homologous Traits

To assess molecular similarities (and thus determine the relationships between species), phylogenetic analysis should focus on sequences called homologous, meaning inherited from a common ancestor. After aligning sequences to account for events of insertion and deletion of nucleotides, it is possible to evaluate their similarity. The more two sequences resemble each other, the more likely it is that they share a common ancestor. The search for homology with the BLAST program also gives an E-value, which is the mathematical expectation. This E-value represents the number of sequences in a database of the same size and composition that would align by chance with the reference sequence (query) with a similarity score equal to or greater than the one obtained.

Three major types of homologous genes can be distinguished:

- Orthologous genes (from the Greek *ortho*, “straight”), present in different species and whose divergence dates back to the speciation events that produced them. Thus, α genes derived from the α copy in the common ancestor, and similarly for β genes derived from the β copy. Orthologous genes reflect the history of speciation events between species. These are the genes used in molecular phylogenies;
- Paralogous genes (from the Greek *para*, “parallel”), resulting from gene duplications within the same species. Therefore, multiple copies of these genes, diverging from one another, can exist within this species. Thus, the α and β copies of the same gene within the same organism are paralogous. There are also three types of paralogy:
    - In-paralogy: Two paralogous sequences within a species are in-paralogous if the duplication event occurred after their speciation;
    - Out-paralogy: Two paralogous sequences within a species are out-paralogous if the duplication event occurred before their speciation;
    - Ohnology: Two paralogous sequences are ohnologues if they result from a complete genome duplication event. The term “ohnologue” was proposed by Ken Wolfe in honor of Susumu Ohno.
- Xenologous genes (from the Greek *xenos*, “foreign”), representing transfers of genetic material from another species (known as horizontal gene transfer), where the phylogenetic link is not direct. Such genes cannot be used to trace the phylogeny of species.

The inference of species phylogeny, in this context inherited from the morphology of homologous traits, requires the analysis of strictly orthologous genes, which are the only witnesses of speciation events. The inclusion of paralogous and/or xenologous genes in such analysis must be avoided at all costs, as they distort the results of the inference.

References

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.

During the 1980s, with the technological advancements in sequencing, a universal tree of life was gradually developed, based on the sequence of a particular nucleic acid, the ribosomal small subunit RNA, responsible for protein synthesis. This molecule was chosen for several reasons. First, because the ribosome is universal: it is found in thousands of copies in all cells, which allows the establishment of phylogenetic links between all living beings (excluding viruses, which lack ribosomes). Second, because the size of this RNA is sufficient to provide information while remaining "manageable." Since the 2000s, our knowledge of microbial diversity has been revolutionized by the development of sequencing techniques and the possibility of sequencing environmental DNA (metagenomics). The ability to massively sequence all DNA present in a given environment at a low cost, using high-throughput sequencing technology, has allowed the reconstruction, through computational methods, of genomes of organisms (called MAGs for metagenome-assembled genomes) that had never been cultured before.

Some have gone further, rejecting the tree-of-life model and questioning phylogenetic classifications. Thus, in 1998, Ford Doolittle suggested that bacterial genes present in eukaryotes were acquired via ingestion (the "you are what you eat" theory) (Doolittle, 1998). His work demonstrated the importance of horizontal gene transfers between living beings. For him, no hierarchical classification of life is valid. Phylogenetic classifications do not reflect the history of organisms but rather that of genes (Doolittle, 1999; Hirt et al., 1999).

3 THE WOESE PARADIGM

3.1 FIRST MOLECULAR PHYLOGENIES AND THE DISCOVERY OF A NEW DOMAIN : THE « ARCHÉOBACTÉRIA »

When Carl Woese began his research in 1969, he decided to compare the sequence of the ribosomal small subunit RNA (SSU rRNA, known as 16S rRNA) of the bacteria known at the time (G. Fox et al., 1977). However, at that time, DNA or RNA could not yet be sequenced quickly. To overcome this problem, Woese used a method developed by Frederick Sanger, which involved creating a "genetic fingerprint" of SSU rRNA molecules by reproducibly cutting them into about a hundred small fragments (or oligonucleotides) and quantifying the divergence between two species A and B using a similarity coefficient, S_{AB} (G. E. Fox et al., 1977). Specifically, he cultured the organisms he wanted to analyze in the presence of a radioactive isotope of phosphorus to isolate their SSU rRNA. Then, after digesting the RNA with ribonuclease T1, which cleaves the molecule after G residues, he generated oligonucleotide catalogs consisting of a specific collection of U, C, and A with a single G residue (in the terminal position). The resulting oligonucleotides were then separated on paper according to their chemical properties using two-dimensional chromatography. The radioactivity revealed the various RNA fragments as spots: this was the genetic fingerprint unique to each species. Each oligonucleotide on the paper could then be eluted, and its sequence identified. Many small motifs were present in all bacteria and were thus irrelevant for analysis. Woese decided to consider only oligonucleotides of at least six nucleotides as relevant. Consequently, the more two species shared a significant number of large oligonucleotides, the more closely related they were. After sequencing the resulting oligonucleotides, Woese compared their sequences and frequencies to define a phylogenetic distance, the PD value (Phylogenetic Distance value), between pairs of organisms.

Woese calculated the similarity coefficient S_{AB} for all species in pairs using the formula: $S_{AB} = 2\ N_{AB}/(N_A + N_B)$, where N_{AB} is the number of oligonucleotides common to both organisms, and N_A and N_B are the numbers of oligonucleotides in the sequences of size 6 or greater. The S_{AB} value equals 1 for identical species and approaches 0 for species that are highly divergent (G. Fox et al., 1977; G. E. Fox et al., 1977). This phylogenetic distance is expressed as a percentage corresponding to the number of base changes needed to transition from one species' oligonucleotide catalog to another's. It serves as an approximation of the number of substitutions between the sequences of two organisms (**Tableau 1**).

When George Fox (C. R. Woese & Fox, 1977), a student of Woese, analyzed the similarity coefficients between pairs of bacteria, he discovered that a group of bacteria was no more closely related to other bacteria than it was to eukaryotes: these were the so-called "methanogenic" bacteria. These results revealed the existence of not two but three types of SSU rRNA, and thus three types of ribosomes in the living world. This led to the emergence, under Woese's vision, of life being divided into three domains: Bacteria, Eukaryotes, and Archaea. Subsequent studies showed that at least two other known groups could be assigned to this new domain: halophilic bacteria (Magrum et al., 1978) and thermo-acidophilic bacteria (C. Woese et al., 1978).

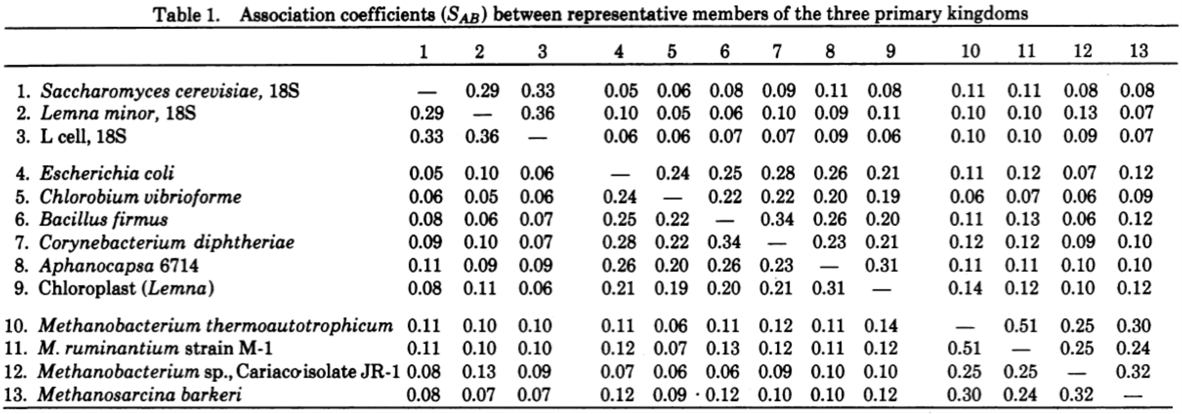


Table 1. Association coefficients (S_{AB}) between representative members of the three primary kingdoms

	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 showing the association coefficients (S_{AB}) between the SSU rRNA sequences of 13 representatives of the three primary domains (C. R. Woese & Fox, 1977).

This classification holds particular historical significance because it was the first time scientists had access to a classification based on objective and measurable phylogenetic criteria, in line with Darwin's theory of evolution. It demonstrates two key points: first, that classification must follow a naturally logical system, the most evident being the evolutionary history of life; and second, that molecular phylogeny is the most effective method for achieving this task.

3.2 SEARCHING FOR THE ROOT OF THE UNIVERSAL TREE OF LIFE

A potential solution to understanding the evolution of the eukaryotic cell and recounting its history is to establish its relationships with the other domains of life. These relationships among the three domains of life are highly debated. Which is the sister group to the other two? Are the three domains monophyletic, paraphyletic, or polyphyletic? Where is the root of the tree of life? These questions remain unanswered due to the difficulty of polarizing molecular characters. As a result, it is challenging to determine whether a given character state is ancestral or derived.

The LUCA (Last Universal Common Ancestor) represents the most recent common ancestor of all current organisms. The "root" of a phylogenetic tree refers to the oldest node in a given phylogeny. By definition, the root also represents the ancestor of all the taxa studied, thereby providing a direction to all the transformational processes used to reconstruct the phylogeny. The most common method for rooting the tree of life is to use universal paralogous genes, meaning genes that underwent an ancestral duplication in LUCA, such that they are present across all three domains of life in the form of at least two copies (**Figure 4**). The concept is as follows: let us consider an ancestral gene X. Over time, it undergoes a duplication, resulting in two copies: copy α and copy β. This provides two records of evolutionary processes over time. One needs only to establish the phylogeny of one of the copies and root it with the phylogeny of the other copy. This produces a mirrored tree, which can then be folded onto itself to identify the outgroup. Typically, the selection of an outgroup for rooting a tree is based on a group of taxa external to the studied group. However, in this case, rooting within an external group is based on one of the paralogous copies.

This technique was first employed using ferredoxin genes. Each ferredoxin gene is composed of two paralogous sequences linked together, such that the phylogeny of one half of the gene can be used to root the phylogeny of the other half (Schwartz & Dayhoff, 1978). However, ferredoxin is a very short protein, and only about fifteen conserved positions were used, which is insufficient to construct a reliable phylogeny of organisms.

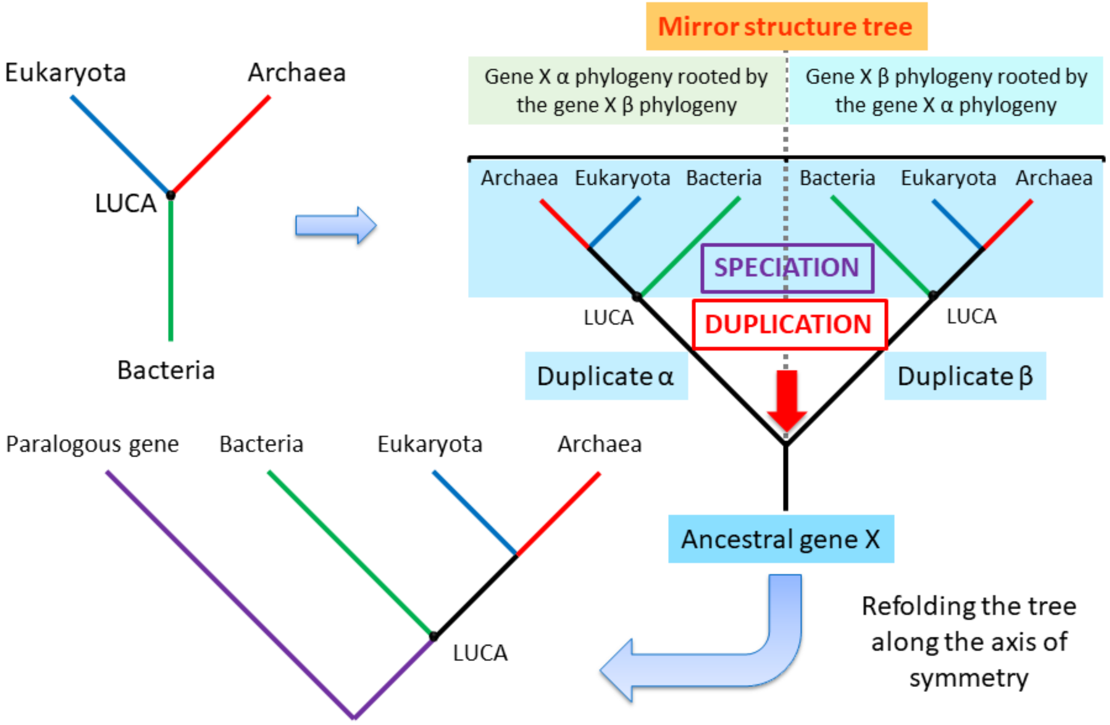


Figure 4. Method for Rooting a Phylogenetic Tree Using Paralogous Genes.
The paralogous gene method is used to root a phylogenetic tree based on homologous genes that underwent duplication during evolution. Paralogous gene sequences (resulting from duplication) are selected from different species. After alignment and construction of a phylogenetic tree, the ancestral paralogous gene from which the duplications originated can be identified by rooting the phylogeny of one copy with the phylogeny of the other copy.

It was through this method that, at the turn of the 1990s, the first phylogenetic trees of life were rooted in Bacteria (Gogarten et al., 1989; Iwabe et al., 1989; C. R. Woese et al., 1990) using ATPases and elongation factors. Unfortunately, we now know that these trees are incorrect (Philippe & Forterre, 1999).

Subsequently, this method was applied to other universal paralogous genes (**Tableau 2**). The problem is that only a few families of ancestrally duplicated genes are known as potential phylogenetic markers, and their analyses have led to varied and controversial conclusions.

Localization of the root	Phylogenetic Marker Used	Reference(s)
Before the original trifurcation	Various characteristics	(C. Woese et al., 1978)
On the branch leading to Bacteria	Regulatory and catalytic subunits of type V and F ATPases	(Gogarten et al., 1989)
On the branch leading to Bacteria	Translation elongation factor proteins, EF-Tu/1 and EF-G/2	(Iwabe et al., 1989)
On the branch leading to Bacteria	Val/lle amino-acyl tRNA synthetases	(J. Brown & Doolittle, 1995)
On the branch leading to Bacteria	Internal duplication in CPS (carbamoyl phosphate synthetase)	(Lawson et al., 1996)
On the branch leading to Bacteria	Tyr/Trp amino-acyl tRNA synthetases	(J. R. Brown et al., 1997)
Inconclusive results: too weak a phylogenetic signal, incongruence of individual his genes	Histidine biosynthesis genes hisA/hisF	(Charlebois et al., 1997)
On the branch leading to Bacteria	Signal recognition particle SRP54 and its receptor SRα	(Gribaldo & Cammarano, 1998)
On the branch leading to Bacteria	Aspartate et Ornithine transcarbamoylases	(Labedan et al., 1999)
On the branch leading to Eukaryotes	Translation elongation factors EF-1α and EF-2	(Lopez et al., 1999; Philippe & Forterre, 1999)
On the branch leading to Bacteria, but likely due to a long-branch attraction artifact	Elongation factors, ATPases, tRNA synthetases, CPS, SRP proteins	(Philippe & Forterre, 1999)
Monophyly of prokaryotes with slow-evolving sites vs Archaea as sister group to Eukaryotes with fast-evolving sites	Signal recognition particle SRP54 and its receptor SRα	(Brinkmann & Philippe, 1999)
Within Gram-negative Bacteria	Bacterial cell wall architecture	(Cavalier-Smith, 2002)
Inconclusive results	RNA secondary structure	(Caetano-Anollés, 2002)
Conceptual difficulties	Not applicable	(Bapteste & Brochier, 2004)

Tableau 2. Different perspectives on the localization of the root of the universal tree of life (adapted from (Zhaxybayeva et al., 2005)**).**

Moreover, the history of a gene does not necessarily reflect the history of a species. The discovery of horizontal gene transfers has shown that some genes are unusable for reconstructing phylogenies, as different genes can tell different stories.

3.3 THE 3 DOMAINS OF LIFE : BACTÉRIA, EUKARYOTA AND ARCHAEA

Today, the classification of living organisms is based on heritable traits, whether visible (anatomical, embryological, morphological criteria) or non-visible (DNA, RNA, and protein sequences). This classification is the subject of a specific discipline, phylogeny, which studies the relationships of kinship among living beings to understand their evolutionary history (phylogenesis) (Haeckel, 1866). This discipline can also incorporate data from paleontology.

The tree published by Woese in 1990, classifying life into three domains (Archaea, Bacteria, and Eukaryotes) (C. R. Woese et al., 1990) (**Figure 5**) revolutionized our understanding of life. His tree was rooted in the bacterial branch based on the study of paralogous genes (elongation factors and ATPases). Archaea and Eukaryotes were shown to be related and to share a common ancestor. The basal branching point represents the position of the last common ancestor, which Woese called the "Progenote" (CarlR. Woese & Fox, 1977). He defined it as "an entity with a rudimentary, imprecise link between its genotype and phenotype" (Doolittle & Brown, 1994; CarlR. Woese & Fox, 1977). However, rooting the tree of life remains an open problem. Placing Bacteria as the outgroup may simply be the result of a phylogenetic artifact caused by long-branch attraction (Philippe et al., 2000) (**Box 7**). For this reason, given the uncertainties surrounding the phylogenetic relationships among the three domains of life (Philippe & Forterre, 1999), it is preferable to work with unrooted universal trees.

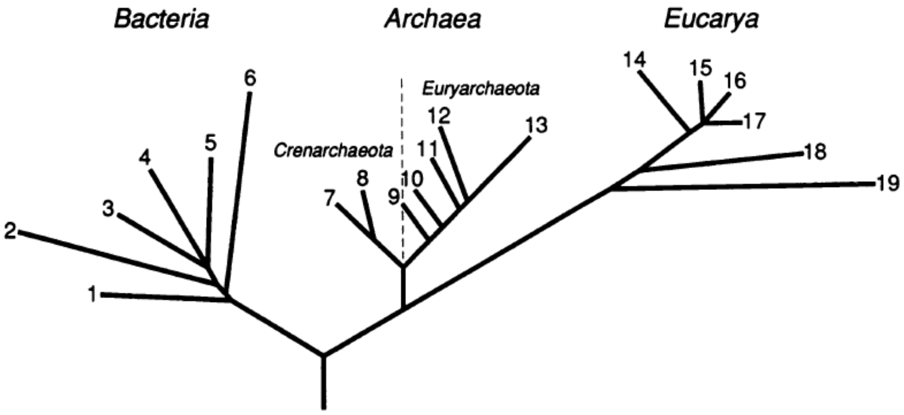


Figure 5. Woese's original rooted tree published in 1990 (C. R. Woese et al., 1990).
This tree is based on the analysis of SSU rRNA genes. The position of the root was determined by comparing the sequences of paralogous gene pairs of elongation factors and ATPases that diverged before the emergence of LUCA (cf. **Tableau 2**). This classification into three domains (Bacteria, Archaea, Eukaryotes) revolutionized our understanding of the diversity of life.

3.4 THE ARCHEZOA HYPOTHESIS

Within eukaryotic trees based on rRNA and elongation factors, a group of simple unicellular (often parasitic) organisms lacking organelles appears at the base. This group includes Metamonads (*e.g., Giardia intestinalis*), Microsporidia (*e.g., Encephalitozoon cuniculi*), Parabasalids (*e.g., Trichomonas vaginalis*), and Archamoebae (*e.g., Entamoeba histolytica*). Thomas Cavalier-Smith grouped them into a separate kingdom called *Archezoa* (Cavalier-Smith, 1993), considering them organisms that diverged from other eukaryotes before the capture of a proteobacterium later transformed into a mitochondrion: this is the Archezoa hypothesis.

The theory of eukaryotic evolution at the time postulated that (1) early eukaryotes were amitochondriate and lived in anaerobic environments, and later, (2) the emergence of aerobic conditions promoted the appearance and diversification of mitochondrial lineages capable of utilizing oxygen for respiration(T. M. Embley & Martin, 2006; Martin Embley, 2006). Indeed, oxygen is highly toxic to anaerobic organisms and exterminates most that cannot neutralize it. Mitochondria are, in a sense, the oxygen resistance organ of a cell, converting oxygen into water via respiration. The Archezoa were thus thought to have differentiated earlier in eukaryotic evolution and were considered relics of an ancient oxygen-free world, living in anaerobic conditions before the Great Oxidation Event (2.4 Ga) (Bekker et al., 2010; Hannah et al., 2004; Kump, 2008; Lyons et al., 2014). They represented a "primitive" stage of eukaryotic evolution, corresponding to ancestral organizational levels along the path to the progressive complexity of eukaryotic cells. The origin of mitochondria was interpreted as a consequence of rising oxygen levels, triggering a crisis for mitochondria-lacking cells and ensuring the survival of mitochondria-bearing cells capable of detoxifying oxygen through natural selection (Kurland & Andersson, 2000; Vellai et al., 1998).

However, several observations contradict this scenario:

1. Phylogenetic reconstructions based on genes other than rRNA (*e.g., α-tubulin* (Keeling & Doolittle, 1996), the large subunit of RNA polymerase II (RBP1) (Hirt et al., 1999) place Archezoa within various eukaryotic lineages, making them polyphyletic. In other words, different genes tell different stories, a phenomenon called *phylogenetic incongruence*. Some Archezoa are thus more recent than previously thought, particularly Microsporidia, now grouped with Fungi (Hirt et al., 1999; James et al., 2013; Keeling, 1998; Keeling & Doolittle, 1996).

2. Genetic analyses have shown that many Archezoa possess mitochondrial-derived genes (*e.g., 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), or even vestigial mitochondria lacking DNA (hydrogenosomes and mitosomes), proving that they once had mitochondria during evolution. Today, it is known that the apparent absence of typical mitochondria in these eukaryotes is due to significant secondary reduction (and rarely loss), not indicative of an ancestral state or a direct relationship among them (Vacek et al., 2018). Two types of mitochondrial reduction are distinguished: hydrogenosomes (performing fermentation that produces H_2, as in some Ciliates) and mitosomes, whose function long remained cryptic (M. Embley et al., 2003; Hjort et al., 2010; B. A. P. Williams et al., 2002) but which play a role in Fe-S cluster assembly (Karnkowska et al., 2016). Mitosomes lack proteins involved in other major mitochondrial functions (aerobic respiration, heme biosynthesis) but have proteins necessary for Fe-S cluster biosynthesis (*e.g., frataxin, cysteine desulfurase, Isu1, and mitochondrial Hsp70*) (Vacek et al., 2018). Thus, it seems mitochondrial symbiosis predates all Archezoa and, therefore, all known eukaryotes. Consequently, the adaptation of eukaryotes to anaerobic environments occurred multiple times independently, and there is no evidence they originated in such conditions (Martin Embley, 2006). This simply shows that the common ancestor of current eukaryotes already had mitochondria (Sogin, 1997) and was likely aerobic.

3. Long-branch attraction is responsible for the artifactual placement of Archezoa within the eukaryotic tree. Homoplasy (accumulation of convergent substitutions) can affect only certain divergent sequences, causing long-branch attraction, particularly in ancient or rapidly evolving lineages. Species with faster-evolving genes appear more distantly

related to their close relatives than they actually are. Undetected, these multiple substitutions create a non-phylogenetic signal that prevents accurate reconstruction of relationships (Forterre & Philippe, 1998). Saturation of mutations at specific positions erases the phylogenetic signal, replacing it with convergent biases, such as base composition, leading to systematic errors.

4. Isotopic studies indicate that anoxic environments have persisted locally and globally over the last 2 billion years. While it is generally accepted that oxygen first appeared in the atmosphere 2 billion years ago (Kump, 2008; Lyons et al., 2014), it is thought that until about 600 million years ago, the oceans were in an intermediate oxidation state, with oxygenated surface waters (where photosynthesis occurred) and anoxic, sulfide-rich deep waters (Poulton et al., 2004; Shen et al., 2003). Oxygen did not appear suddenly in geological time; its abundance gradually increased, linked to the evolution or adaptation of microorganisms (Fenchel & Finlay, 1995). Early "aerobes" were microaerophiles that thrived under very low oxygen pressures, in an atmosphere with oxygen levels between 0.1% and 1%. Consequently, the Great Oxidation Event of the atmosphere must be decoupled from anoxic marine environments, where anaerobic eukaryotes living on the fringes of an oxic world could have thrived, as they still do today. Mitochondria do not even increase respiratory rates: gram for gram, many prokaryotes respire faster than eukaryotes (Lane & Martin, 2010).

The phylogeny of eukaryotes appears to be less straightforward than was believed during the era of the Archezoa hypothesis. Numerous convergences and/or losses have occurred, making conjectures about the nature of their common ancestor speculative. This view of eukaryotic cell evolution was influenced by the assumption that life evolved from "simpler" or "primitive" forms to more "complex" or "advanced" forms, suggesting a false notion of continuous progress in evolution. This teleological view of evolution (cf. *scala naturae*) has been abandoned since it was demonstrated that many life forms (notably parasites) resulted from secondary simplifications. Today, it is widely accepted that all known modern eukaryotes evolved from a relatively complex ancestor equipped with mitochondria (A. M. Poole & Penny, 2007; A. Poole & Penny, 2007) and that this ancestor lived in an aerobic environment. This eukaryotic ancestor is commonly referred to as *LECA* (Last Eukaryotic Common Ancestor). Some lineages subsequently underwent secondary simplification (*e.g.*, the loss of introns and parts of the spliceosome, or the "mitosomization" of mitochondria). But if LECA was already a complex organism, was the same true for the earliest eukaryotes? Does the presence of mitochondria mark the first step in eukaryotic evolution, or did eukaryotes without mitochondria (like the proposed Archezoa) already exist beforehand? What influence did environmental factors have on their emergence? Did they appear in aerobic or anaerobic conditions?

The data we have today do not allow us to look back further than LECA. It is highly plausible that proto-eukaryotes existed without possessing all the properties of modern eukaryotes. The mystery remains unresolved...

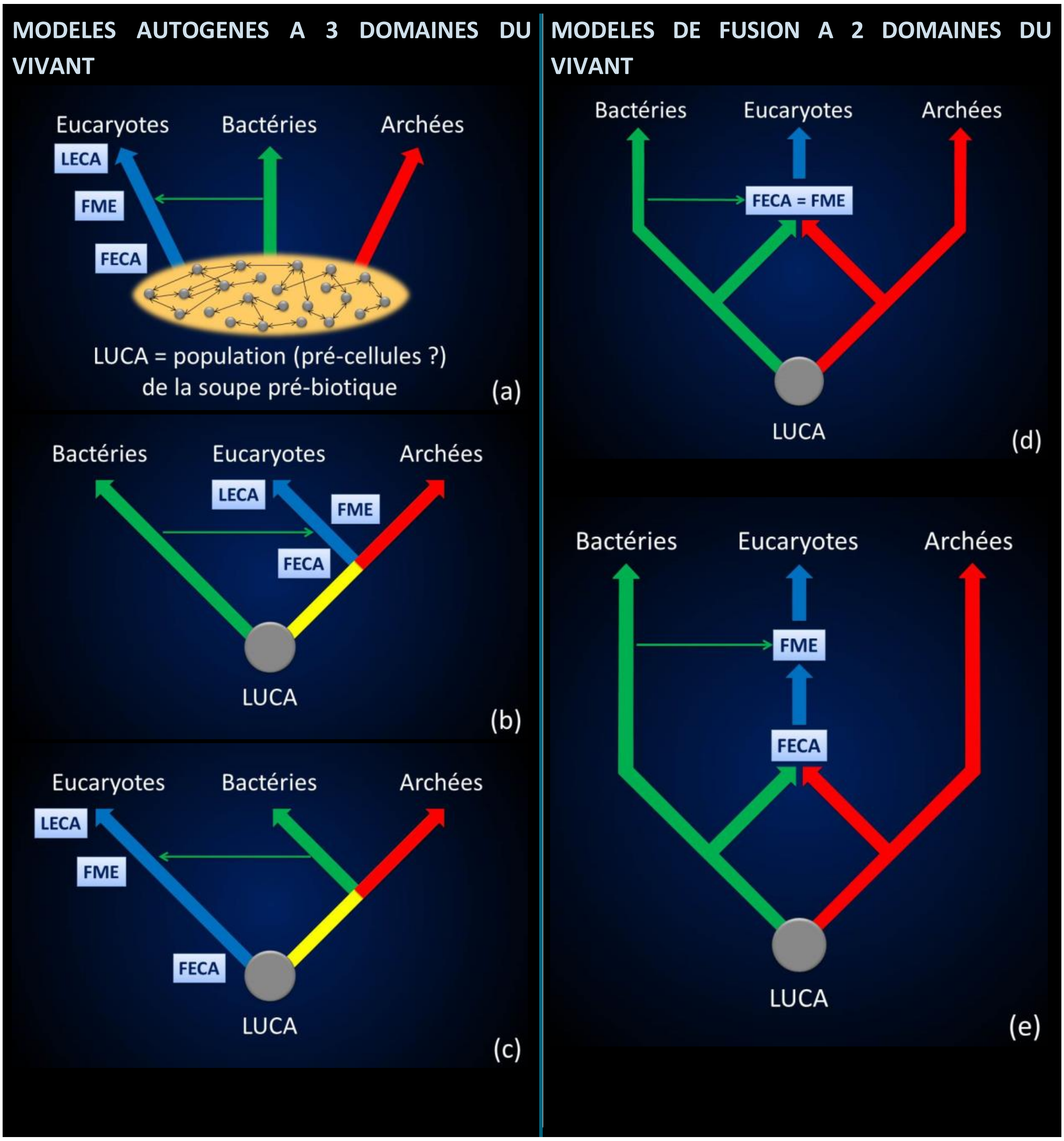
4 THE ORIGIN OF THE EUKARYOTIC CELL

Several scenarios attempt to explain the origin of the eukaryotic cell and its evolutionary relationships with the two prokaryotic domains, but they do not agree on the nature of the host or the number of partners involved in the emergence of the eukaryotic cell. Without being exhaustive, the models considered here aim to generalize the main characteristics of a large number of proposals (A. M. Poole & Penny, 2007; A. Poole & Penny, 2007). These scenarios can be grouped into several categories based on factors such as the nature of LUCA (DNA or RNA) (Forterre, 2007), the number of primary domains of life (two vs. three domains) (Forterre, 2011; Gribaldo et al., 2010) or the nature of the first eukaryote (with or without mitochondria) (T. M. Embley & Martin, 2006) (**Figure 6**).

If we consider a three-domain scenario, the ancestral eukaryotic lineage (stem group) does not correspond to modern eukaryotes but rather to an ancient lineage leading to LECA. This lineage may be as old, or even older, than bacteria and archaea. It is therefore not excluded that amitochondriate eukaryotes (stem group) existed before the endosymbiosis of the bacterium. Applying coalescence theory to current eukaryotes leads us back to this point as well. However, it is possible that other eukaryotic lineages posterior to LECA existed but left no modern descendants (**Box 2**). Consequently, this lack of information makes it difficult to infer the identity of LECA. Furthermore, LECA might be just one of several representatives of a group of lineages that emerged from FECA. In this context, we consider eukaryotes as a whole, including lineages that have since gone extinct after FECA.

On the other hand, proponents of fusion scenarios for the origin of eukaryotes often blur the distinction between eukaryotes in a broad sense (including extinct lineages that lived before acquiring mitochondria) and modern eukaryotes belonging to extant lineages (Forterre, 2011). In these scenarios, eukaryotes are recent because they arose from the fusion of an archaeon and a bacterium. Accordingly, LUCA is a prokaryote, and FECA only appears after the diversification of eubacteria and archaea.

The fusion hypothesis is based on the observation that eukaryotes and archaea share a similar cellular machinery (translation, transcription, replication), which is very different (sometimes non-homologous) from that of bacteria (Golding & Gupta, 1995) (cf. Janus Paradox of Lake (Lake, 2007)). At the same time, metabolic genes are of bacterial origin. Fusion scenarios involve various types of associations (symbiosis, syntrophy...), often justified by contemporary metabolic considerations to explain the association of two partners (López-García & Moreira, 1999).



MODELES AUTOGENES A 3 DOMAINES DU VIVANT

MODELES DE FUSION A 2 DOMAINES DU VIVANT

Figure 6. Proposed Models for Explaining the Evolution of the Three Domains of Life.
(a) The three domains of life have independent origins and emerged directly from the prebiotic soup. In this case, LUCA does not exist. (b) The three domains derive from a simple common ancestor, with evolution proceeding through increasing complexity in eukaryotes. LUCA would then be prokaryotic. (c) The three domains derive from a complex common ancestor, with evolution proceeding through simplification in prokaryotes. LUCA could be a prokaryote, a eukaryote, or something intermediate. (d) Eukaryotes emerged from the fusion of a bacterium and an archaeon, appearing with the mitochondrion (*mitochondrion-early model*) or with a proto-mitochondrion (*mitochondrion-intermediate model*), with their other characteristics evolving after this event. (e) Eukaryotes emerged from the fusion of a bacterium and an archaeon, appearing before the mitochondrion (*mitochondrion-late model*).

We will now review various hypotheses and models that have been formulated to date (in a non-exhaustive way) and that attempt to explain the origin of the eukaryotic cell as best we can.

4.1 THREE-DOMAIN SCEARIOS

In these scenarios, eukaryotes are ancient, possibly as old as bacteria or archaea, or even older. These models constitute the autogenous model of eukaryotes. In this case, the eukaryotic lineage is independent of prokaryotes, creating a three-domain scenario. There would have been a "proto-eukaryote," amitochondrial, referred to as FECA (First Eukaryote Common Ancestor). Over time, these early eukaryotes would have acquired mitochondria, leading to the FME (First Mitochondriate Eukaryote). These then evolved into LECA (Last Eukaryotic Common Ancestor), the ancestor of all modern eukaryotic lineages. Three hypotheses are conceivable:

1. The three domains of life have independent origins and emerged directly from the prebiotic soup. In this case, LUCA does not exist (**Figure 6 (a)**).
2. The three domains derive from a simple common ancestor, with evolution proceeding through increasing complexity in eukaryotes. Here, we have a three-domain model originating from a prokaryotic LUCA (**Figure 6 (b)**).
3. The three domains derive from a complex common ancestor, with evolution occurring through secondary simplification in prokaryotes. This results in a three-domain model originating from a LUCA that could be either a prokaryote or a proto-eukaryote (**Figure 6 (c)**).

In these hypotheses, the prokaryote-to-eukaryote transition is explained by the gradual complexity of ancestral structures within a prokaryotic lineage. The acquisition of new structures (*e.g.*, cytoskeleton, endomembrane systems) enabled the emergence of phagocytosis, which in turn facilitated the incorporation of an α-proteobacterium that became the mitochondrion (Muñoz-Gómez et al., 2022). Metabolic genes would have originated from horizontal transfer from the mitochondrial ancestor, while the remaining proto-eukaryotic genes would either be closely related to archaea through vertical inheritance (if archaea and eukaryotes share a common ancestor) or through horizontal transfer (if archaea and eukaryotes are not related). The first eukaryote would therefore lack mitochondria (FECA before FME). All these models consider mitochondria as a late addition, derived from an α-proteobacterium of the order *Rickettsiales*. However, this point is now debated due to the mosaic nature of the mitochondrial proteome, with only 10–20% being of α-proteobacterial origin (Gray, 2015). Mitochondrial genomes exhibit great diversity in terms of gene content and organization, complicating the reconstruction of the common ancestor of mitochondria and the precise identification of its bacterial ancestor. Phylogenetic analyses based on different mitochondrial proteins or entire genomes yield conflicting results, with some suggesting a relationship with α-proteobacteria and others indicating affinities with other bacterial groups (Roger et al., 2017). Certain mitochondrial genes exhibit similarities to genes from other bacterial groups, not just α-proteobacteria, suggesting horizontal gene transfers or complex recombination events.

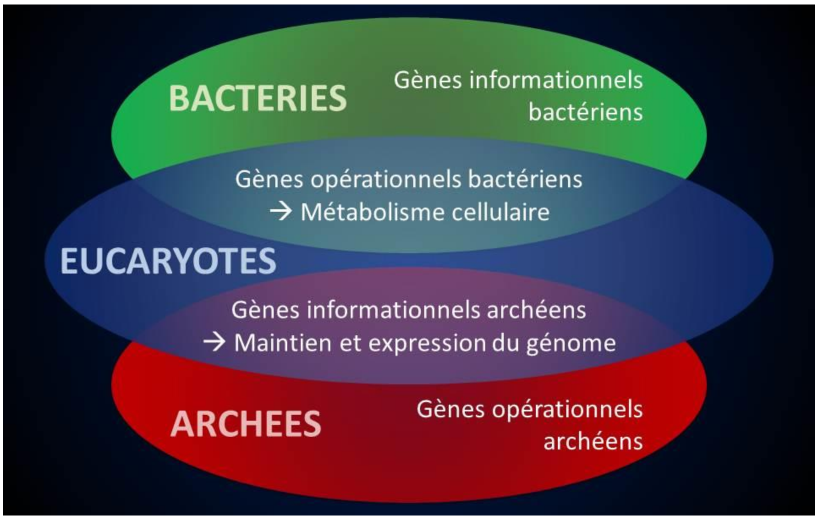
In 1991, Mitchell Sogin proposed that eukaryotes are very ancient and belong to a third lineage, independent of archaea and bacteria. He suggested that a proto-cell from the RNA world, already complex, engulfed an archaeon, giving rise to eukaryotes. This hypothesis followed a series of proposals from the late 1970s by researchers such as James Darnell, Hyman Hartman, and Ford Doolittle, who posited that introns were relics of pre-cellular gene assemblies. Indeed, it seems that LECA was already complex and rich in introns (*Introns-First* theory) (Koonin, 2009; Penny et al., 2009). In 2002, Hyman Hartman and Alexis Fedorov compared fully sequenced genomes from five eukaryotes (*Saccharomyces cerevisiae, Drosophila melanogaster, Caenorhabditis elegans, Arabidopsis thaliana,* and *Giardia lamblia*) and 44 complete archaeal and eubacterial genomes from GenBank (Hartman & Fedorov, 2002). They identified 347 eukaryote-specific proteins, which they termed *eukaryotic signature proteins* (ESPs). To explain these findings, they proposed the involvement of a third type of RNA-genomed cell (referred to as a "chronocyte"), distinct from archaea and eubacteria, to explain the formation of the eukaryotic cell. They suggested that the endosymbiont being hosted in an RNA-genomed cell could explain why transcription occurs in the nucleus and translation in the cytoplasm. This task division would be the result of a communication system established between the endosymbiont (a DNA-genomed cell) and the host cell (an RNA-genomed cell). This chronocyte would have possessed, among other features, a cytoskeleton enabling phagocytosis, an endomembrane system, and a complex system of intracellular signaling pathways.

In 2000, Luis Villarreal and Victor DePhilippis hypothesized a viral origin for eukaryotic replication proteins, based on studies of DNA polymerases from phycodnaviruses (viruses infecting microalgae) (Villarreal & DeFilippis, 2000). These polymerases show significant similarity to those of eukaryotes, with phylogenetic reconstructions placing them near the root of all eukaryotic delta DNA polymerases. In 2001, Philip Bell and Masaharu Takemura independently proposed a viral origin for eukaryotes (Bell, 2001, 2005, 2006, 2009; Takemura, 2001). According to them, the eukaryotic nucleus resulted from the infection of a cell by a complex DNA virus—this is the *viral eukaryogenesis* theory. It posits that eukaryotes evolved from an ancestor of modern archaea, into which a DNA virus integrated to form the nucleus. Their hypothesis is based on numerous similarities (mRNA capping, linear chromosomes, separation of transcription and translation...) between the intracellular cycle of certain cytoplasmic viruses (such as poxviruses) and the intracellular cycle of the nucleus. The strongest similarity is the recruitment of the endoplasmic reticulum to form, in one case, the viral factory membrane and, in the other, the nuclear membrane. Given the complexity of the eukaryotic cell, Forterre speculated that multiple viruses contributed to the formation of the nucleus (Forterre, 2007). This would explain the presence of several DNA and RNA polymerases in eukaryotes. Phylogenetic analyses show that different versions of these enzymes did not arise from duplications in a primordial eukaryotic lineage. Instead, eukaryotic DNA polymerases α, δ, and ε, as well as RNA polymerases I, II, and III, do not form monophyletic groups; each version is separated in phylogenetic trees by archaeal and viral enzymes (Filée et al., 2002; Forterre, 2007). The common ancestor of these eukaryotic enzymes would also have given rise to archaeal and viral enzymes, suggesting a very ancient FECA. DNA and its replication mechanisms may have first appeared and evolved in the viral world before being transferred to the cellular world (Forterre, 2001).

4.2 TWO-DOMAIN SCENARIOS: THE JANUS PARADOX, TWO ORGANISMS IN ONE?

To test Woese's model regarding the common origin of Archaea and Eukaryotes, Brian Golding and Radhey Gupta, in 1995, compared the genome sequences of bacteria, archaea, and eukaryotes. They showed that the eukaryotic genome was part archaean and part bacterial (Gram-negative type) (Golding & Gupta, 1995). Schematically, eukaryotic genes can be divided into three groups (Andersson et al., 2003; Hartman & Fedorov, 2002; A. M. Poole & Penny, 2007; Rivera et al., 1998; Rivera & Lake, 2004; Yutin et al., 2008) : those that are eukaryote-specific, those that are similar to bacterial genes, and those that are similar to archaeal genes. Furthermore, when considering the function of these genes, it becomes apparent that bacterial genes compose the operational genes responsible for cellular maintenance processes and code for nucleotide and amino acid biosynthesis in eukaryotes, while archaeal genes make up the informational genes of the cellular machinery (translation, transcription, replication) (**Figure 7**). Thus, unlike simple gene trees based on ribosomal RNA that place eukaryotes within the archaea, genome analysis shows that the eukaryotic cell likely resulted from the fusion of archaeal and bacterial genomes, transforming the tree of life into a "ring of life" (Rivera & Lake, 2004).

A fundamental question then arises: why do the informational genes of eukaryotes resemble archaeal genes, while the operational genes resemble bacterial genes? This strange correlation was named the "Janus Paradox" by James Lake in 2007 (Lake, 2007), referring to the Roman god with one head but two opposing faces (**Figure 7**). This paradox underpins two questions that need to be explained.


BACTERIES
Gènes informationnels bactériens
Gènes opérationnels bactériens → Métabolisme cellulaire
EUCARYOTES
Gènes informationnels archéens → Maintien et expression du génome
ARCHEES
Gènes opérationnels archéens

Bust of Janus, Vatican Museul

Figure 7. Duality of the Eukaryotic Genome, Between Bacterial and Archaeal, Illustrated by the Roman Bust of Janus.

Bacterial genes make up the operational genes responsible for cellular maintenance processes and code for nucleotide and amino acid biosynthesis in eukaryotes, while archaeal genes compose the informational genes of the cellular machinery (translation, transcription, replication).

1. **Why are bacterial informational genes absent from eukaryotic genomes?**

Informational genes are less prone to lateral gene transfers than operational genes. Based on the observation that two types of ribosomal genes cannot coexist within the same nucleus, Lake suggested that the archaeal ribosome was, by chance, the only survivor, while bacterial ribosomal genes were inactivated by various proteins, leading to their elimination (Lake, 2007). This would have resulted in a centralization of information in the host nucleus, thus facilitating coordination actions with the multiple copies of mitochondrial genes. Furthermore, the separation of energy production (mitochondria) and information (nucleus) would allow for a balanced cooperation between the two components (rather than a form of parasitism), with each benefiting from the arrangement.

2. **Why are archaeal operational genes absent from eukaryotic genomes?**

One of the reasons proposed is related to the structure of their membrane, which is much less permeable in archaea than in bacteria. The composition and biosynthesis pathways of phospholipids in archaea are very different from those observed in bacteria and eukaryotes (Peretó et al., 2004). The membranes of bacteria and eukaryotes are composed of linear fatty acids attached to glycerol-3-phosphate by ester bonds. In contrast, archaeal membranes are made of long chains of isoprenoid alcohols attached to glycerol-1-phosphate by ether bonds, making them much less permeable to ions. These membranes outperform bacterial membranes in environments exposed to chronic ecological stress (Valentine, 2007). It seems that the ecology and evolution of archaea are driven by their adaptations to chronic energy stress (Valentine, 2007). Thus, the biochemical mechanisms that enable archaea to cope with these stresses include the low permeability of their membranes and specific metabolic pathways adapted to these ecological niches.

If the membranes of archaea provide a more effective barrier against ions, why do eukaryotes have membranes similar to those of bacteria? In fact, it is highly likely that the low permeability of archaeal membranes hindered lateral diffusion of respiration products and bacterial signal transduction. The use of more fluid and permeable membranes may represent an adaptation aimed at enhancing energy production at the cost of energy conservation, with unsaturated membranes in the most extreme cases. Finally, chemical modifications, such as the inclusion of specific sterols (e.g., cholesterol), likely played a role in this fluidity (Dufourc, 2008). As the proto-mitochondrion became the energy organelle (powerhouse) of the proto-eukaryotic cell, the adaptation of anaerobic archaea to chronic energy stress would have become negligible and surpassed by the much more energy-efficient bacterial aerobic metabolism (Davidov & Jurkevitch, 2007). The mitochondrion would thus have ended the chronic energy stress faced by the archaea by becoming a more efficient energy conversion organelle. The adaptations of archaea to chronic energy stress would no longer have been advantageous and would have gradually disappeared, as they were incompatible with an energy-rich environment, overtaken by more energy-efficient bacterial metabolism.

This *Janus Paradox* thus serves as a metaphor for two types of incompatibilities between archaea and bacteria: structural incompatibility between informational systems and environmental/ecological incompatibility between metabolic systems (Davidov & Jurkevitch, 2007). This would explain (1) why eukaryotic membranes are of bacterial type while the host is of archaeal type, and (2) the absence of archaeal operational genes in eukaryotes (Davidov & Jurkevitch, 2009). The similarity between the cellular machinery of eukaryotes and archaea can then be interpreted as a vertical inheritance from a common ancestor (synapomorphy) or, conversely, as an ancestral state (symplesiomorphy), with bacteria being less similar due to an accelerated evolutionary rate since their divergence from LUCA. In contrast, bacteria and eukaryotes share many metabolic proteins, the origin of which is debated. However, it remains difficult to explain the tempo and modalities of these transfers. These observations do not clarify whether these changes were rapid and sudden or slow and gradual. In particular, the transition from an archaeal membrane to a bacterial membrane remains difficult to explain, especially in the absence of a reliable model of eukaryogenesis. Did the membranes coexist for a period of time, or was the archaeal membrane's elimination mechanical and instantaneous? The question remains open, and current observations do not explain the transition from one membrane type to another. We would then need to consider scenarios that could explain this gene mix. What are they?

4.2.1 CASE WHERE FECA IS AMITOCHONDRIAL

In these scenarios, the characteristics of eukaryotes appear before the mitochondrion (*mitochondrion-late model*). In this case, FECA results either from the physical fusion of two cells (the membranes fuse), or from the endosymbiosis of an archaeon, which then gives rise to the nucleus via another cell (which can never be a eukaryote) (**Figure 6 (e)**). This type of model requires three partners. For fusion, two cells are necessary, followed by an endosymbiosis for the mitochondrion. For endosymbioses, two endosymbionts are required (one for the nucleus and one for the mitochondrion).

Among fusion models, many partners have been proposed. The first chimera hypothesis was that of serial endosymbiosis by Lynn Margulis, proposed in 1970 (Margulis, 1970). This theory of the endosymbiotic origin of mitochondria and plastids had already been proposed by Mereschkowsky in 1905, before being rejected and forgotten (Mereschkowsky even suggested that the nucleus derived from bacteria). In her model, Margulis suggests that mitochondria derive from ancient endosymbiotic bacteria. After facing heavy criticism, her ideas were partially accepted thanks to molecular phylogenies (Gray et al., 1999; Gray & Doolittle, 1982; Wallace, 1982) and the accumulation of data that continued into the early 2000s, when genomic sequences of key model organisms became available (Gray et al., 2001; Sato, 2021). Molecular analyses based on rRNA showed that plastids derive from cyanobacteria and mitochondria from α-proteobacteria. Later, based on morphological criteria, Margulis proposed that the host that incorporated the ancestor of the mitochondrion (α-proteobacterium) itself resulted from a symbiosis event between a wall-less archaeon, similar to current *Thermoplasma acidophilum*, and a spirochete (bacterium) (Lynn et al., 2006). However, no genetic data have confirmed the involvement of a spirochete, as Margulis proposed this partner based solely on morphological grounds. Indeed, *T. acidophilum* is a euryarchaeon without a cell wall and possesses histone proteins biochemically similar but analogous to those in eukaryotes, allowing it to stabilize DNA by forming structures resembling nucleosomes. As for spirochetes, they possess structures resembling the microtubules of the eukaryotic cytoskeleton. During their symbiosis, the bacterium and the archaeon would have formed a proto-eukaryote without a nucleus or mitochondrion, which Margulis called the "Thiodendron stage" (Margulis et al., 2000). However, sequencing their genomes showed that spirochetes and *T. acidophilum* have no direct phylogenetic relationship to eukaryotes.

In 1984, James Lake proposed a model counter to Woese’s vision, in which he made archaea paraphyletic by creating a fourth domain more closely related to eukaryotes: the *Crenarchaeota* (which he called "eocyte") (Lake et al., 1984). Initially, this hypothesis was based on the structural study of ribosomes from both groups, which showed similar forms. Subsequent studies reinforced this hypothesis (Cox et al., 2008), based on sequences of the small subunit rRNA or the insertion of 11 amino acids in the GTPase domain of the elongation factor 1-α (EF-1α, also called EF-Tu) protein in eocytes and eukaryotes (Rivera & Lake, 1992) as well as the phylogeny of this gene, which places *Crenarchaeota* as the sister group of eukaryotes (Baldauf et al., 1996). More recently, in 2010, Bayesian approaches and maximum likelihood methods identified, within *Crenarchaeota*, the group *Thaumarchaeota* as the sister group or even the root group of eukaryotes (Kelly et al., 2011). Lake and Rivera proposed a vision of life based on a "ring of life" scenario, where the genome of eukaryotes results from the fusion of a bacterial genome (probably a proteobacterium or a member of an ancient photosynthetic clade like cyanobacteria) and a crenarchaeon (Rivera, 2007; Rivera & Lake, 2004) followed by numerous horizontal gene transfers.

In 1989, Wolfram Zillig, based on a comparison of 26 biochemical traits as well as the comparison of DNA polymerases, developed a fusion scenario between an archaeon and a bacterium (Zillig, 1991; Zillig et al., 1989). This model assumes the physical fusion of two cells into one new cell with a single genome. This process is independent of the origin of the mitochondrion, implying that modern eukaryotes emerged from three cells: two fused into an amitochondrial proto-eukaryote, and a third introduced through endosymbiosis. However, in this scenario, the term "fusion" remains imprecise.

However, not all models necessarily explain the emergence of a fundamental characteristic of eukaryotes: the nucleus. To justify the origin of the nucleus, Purificación López-García and David Moreira proposed in 1998 a syntrophic hypothesis based on hydrogen transfer in an anaerobic environment between an ancient sulfate-reducing myxobacterium (δ-proteobacterium) and a methanogenic euryarchaeon (future nucleus) (Moreira & López-García, 1998). A second endosymbiosis involving an α-proteobacterium would then be responsible for the mitochondrion. The appearance of the nucleus would have occurred in two stages, under the influence of two selective forces (López-García & Moreira, 2006). (1) Initially, there would have been metabolic compartmentalization to avoid the deleterious coexistence of anabolic pathways (autotrophic synthesis by the methanogenic archaeon) and catabolic pathways (fermentation by the myxobacterium). Indeed, bacterial fermentation releases hydrogen, carbon dioxide, and acetate, while the methanogenic archaeon derives its energy by reducing carbon dioxide with hydrogen to produce methane. This first phase of compartmentalization would thus prevent the constant recycling of molecules by these two opposing metabolic pathways. (2) Later, the acquisition of oxygenic respiration linked to the arrival of mitochondria would have led to the abandonment of methanogenesis (which has a much lower energy yield). Once the ability to methanogenize was lost, the archaeal membrane would have been lost as well. Simultaneously, a secretory endomembrane system would have formed by invagination of the bacterial membrane, which would have replaced the original archaeal membrane, thus forming the nuclear membrane (Lombard et al., 2012). The appearance of the nucleus would then ensure protection against the aberrant protein synthesis caused by the invasion of introns after the mitochondrion’s arrival, by decoupling transcription and translation.

4.2.2 CASE WHERE FECA IS MITOCHONDRIAL (= FME)

In these scenarios, eukaryotes appear with the mitochondrion (*mitochondrion-early model*), with their other characteristics evolving after this event, or with a proto-mitochondrion (*mitochondrion-intermediate model*) (Roger et al., 2017). In the latter intermediate model, the proto-mitochondrion is described as an ancestor of modern mitochondria, an alphaproteobacterium that was integrated into a host cell. Unlike the *mitochondrion-early* model, where this endosymbiosis is seen as the triggering event for the formation of eukaryotes, the intermediate model suggests that the host already had some eukaryotic characteristics. Thus, the symbiosis with the proto-mitochondrion took place at an intermediate stage of cellular evolution. This cell was not yet a complete eukaryote but had started evolving in that direction. This symbiotic acquisition at an intermediate stage of cellular evolution favored the complexity and diversification of eukaryotes. Therefore, eukaryotes directly arise from the endosymbiosis of a bacterium (which gave rise to the mitochondrion) by an archaeon. Here, only two partners are needed (a host cell and the ancestor of the mitochondrion). In these scenarios, FECA and FME are the same (**Figure 6 (d)**).

Many metabolic endosymbiosis scenarios have been developed between an archaeon and a bacterium, which later evolved into the mitochondrion. In 1992, Dennis Searcy hypothesized that eukaryotes derive from a symbiosis based on sulfur transfer between a wall-less euryarchaeon like *Thermoplasma* and a proteobacterium, which will become the mitochondrion. While the partners are the same as in Margulis's scenario previously mentioned, the difference lies in the nature of the symbiosis (based on sulfur), the number of partners involved, and the nature of the first eukaryote. For Searcy, eukaryotes appear directly after the acquisition of the mitochondrion via endosymbiosis, involving two partners.

In 1998, William Martin and Miklos Müller made hydrogen transfer the key to the symbiosis between the α-proteobacterium (providing hydrogen and carbon dioxide) which would become the mitochondrion and an anaerobic autotrophic methanogenic euryarchaeon which consumes it (Martin & Muller, 1998). In their model, the archaeal membrane is gradually replaced by bacterial phospholipids.

In 2006, Martin and Koonin proposed that the appearance of the nuclear membrane was linked to the diffusion of group II introns from the mitochondrial genome into the nuclear genome and to the formation of spliceosomes (Martijn & Ettema, 2013; Martin & Koonin, 2006). The appearance of the nucleus would have then been selected because it decouples transcription (nuclear) from translation (cytosolic), thus preventing the coexistence of deleterious proteins after the diffusion of mitochondrial-originated introns into the nuclear genome (Koonin & Martin, 2005; López-García & Moreira, 2006). Indeed, because ribosomal translation is faster than the splicing process, the host would have been unable to express its genes without their introns, with transcription and translation happening simultaneously. Without this, aberrant proteins would have been synthesized from genes still containing introns. Translation would have been too fast, translating the introns alongside the genes, making the proteins defective. Thus, separation between splicing and translation became necessary. One of the primary functions of the nuclear membrane would have been to allow the splicing of mRNA to remove its introns so that translation occurs on a continuous reading frame of the mRNA (Koonin & Martin, 2005). The nuclear membrane would then have been formed de novo by vesicles of bacterial phospholipids, which would have formed vesicles encapsulating the archaeal genome.

In 2013, Joran Martijn and Thijs Ettema developed a new model, the "PhAT" (phagocytosing archaeon theory), suggesting the possibility of the existence of archaea capable of phagocytosis, to explain the complexity of the eukaryotic cell (Martijn & Ettema, 2013). Among fusion models, many studies tend to strengthen an archaeal origin for eukaryotes, probably from an organism belonging to the TACK superphylum (Thaumarchaeota, Aigarchaeota, Crenarchaeota, Korarchaeota) (Guy & Ettema, 2011; Kelly et al., 2011; Martin, 2005; T. A. Williams et al., 2012). Moreover, an increasing number of ESPs (eukaryotic signature proteins) are regularly discovered among these groups (Hartman & Fedorov, 2002) suggesting that the archaeal ancestor of the eukaryotic cell may have been more eukaryotic than previously imagined. It might even have already had the ability to phagocytose. In their five-step scenario (Martijn & Ettema, 2013), (1) an archaeon (probably from the TACK group) with an actin cytoskeleton and ESPs (2) loses its wall and thus gains flexibility. (3) The cytoskeleton would then have evolved into a primitive phagocytosis machinery, allowing it to ingest other prokaryotes. This would result in an increase in horizontal gene transfer rates, destabilizing the genome of the archaeal host. This would then lead to an acceleration of the genome's evolutionary rate. (4) A protective membrane would then have formed around the genome by invagination movements to stabilize it, giving rise to a proto-nucleus. An ancient α-proteobacterium would then have been ingested but not digested, allowing it to maintain a symbiotic relationship with its host. (5) Finally, this α-proteobacterium would have evolved by reduction to give the mitochondrion. This mitochondrion, providing the host cell with an energy surplus in the form of ATP, would have facilitated the maturation of the nucleus and endomembrane systems.

What if the secret to the origin of eukaryotes lay not in symbiosis but in predation? This is the idea developed by Yaacov Davidov and Edouard Jurkevitch in 2009 (Davidov & Jurkevitch, 2009). If prokaryotes cannot phagocytose, could we assume that a small bacterial parasite entered a large archaeon? Indeed, until then, it seemed that phagocytosis appeared late in eukaryotes (Jékely, 2003). The small GTPases of the Ras superfamily, specific to eukaryotes, play a crucial role in cytoskeletal dynamics and vesicular trafficking. However, their phylogeny shows that the initial function of endomembranes was secretory and that phagocytosis appeared later (Jékely, 2003). Furthermore, an analysis of the proteins involved in phagocytosis in different eukaryotic groups reveals that phagocytosis has evolved independently at least three times among current eukaryotes (Yutin et al., 2009). A late origin for phagocytosis thus contradicts scenarios that postulate phagotrophy as a prerequisite for endosymbiosis (Cavalier-Smith, 2009; A. Poole & Penny, 2007). Finally, we know of predatory interactions between unicellular organisms that are not based on phagocytosis. For example, BALOs (Bdellovibrio and similar organisms, δ-proteobacteria) parasitize other Gram-negative bacteria, Rickettsia-like organisms (α-proteobacteria) parasitize eukaryotes, and *Nanoarchaeota* live in symbiosis or parasitism with *Ignicoccus hospitalis* (Davidov & Jurkevitch, 2009). Davidov and Jurkevitch thus propose that predation or some form of parasitism may have played a role in the evolution of the eukaryotic cell. The ancestor of the mitochondrion would thus have been capable of crossing the cellular barrier of its prey while maintaining the host's cellular integrity and joining it to develop predatory/parasitic relationships while bypassing its defense mechanisms.

4.3 ONE-DOMAIN SCENARIOS

For Thomas Cavalier-Smith, archaea and eukaryotes are two sister groups, which he groups under the term *Neomura* (New Walls) (Cavalier-Smith, 2002). According to him, the tree of life is rooted in bacteria, Gram-negative, and *Neomura* would be derived from Gram-positive bacteria capable of replacing rigid murine (peptidoglycan) with flexible glycoproteins that allow phagocytosis (Cavalier-Smith, 2006). This would explain the presence of a single plasma membrane in both archaea, eukaryotes, and Gram-positive bacteria, while Gram-negative bacteria have two. The conversion of a bacterium into a eukaryote would then have required around sixty innovations (Cavalier-Smith, 2009). Cavalier-Smith believes that eukaryotes are recent (800 to 850 million years ago). However, this hypothesis contradicts the fossil record, which suggests a much older origin for eukaryotes (**Box 3**).

Some authors go even further, considering eukaryotes as ancestral and prokaryotes as derived through reduction (i.e., secondary simplification). For David Penny and Anthony Poole, the RNA world (if it did indeed exist) is the origin of all cellular life (Collins et al., 2009; A. Poole et al., 1999). According to them, if the RNA used in splicing introns and stabilizing RNA in modern eukaryotes are relics from the RNA world, then these mechanisms must have been present in LUCA. Thus, this common ancestor would have been a eukaryote or proto-eukaryote itself, and prokaryotes would have been derived from it, long before the appearance of the mitochondrion (A. M. Poole & Penny, 2007; A. Poole & Penny, 2007).

In the same vein, Patrick Forterre and Hervé Philippe also presented, in 1999, a model in which LUCA would already have been a complex organism, with a genome rich in genes (Forterre & Philippe, 1999a). Indeed, suppose a phylogeny is established based on a gene evolving faster in one lineage than another. In this case, the faster-evolving lineage would be placed at the base of the tree near the external group: this is the phenomenon of long-branch attraction. By taking this bias into account and eliminating genes with an excessively fast evolution rate, Forterre and Philippe demonstrated that LUCA could have been a complex cell, like a eukaryote (Forterre & Philippe, 1999a, 1999b). The trees would then be rooted in eukaryotes, in which case evolution toward archaea and bacteria would have occurred by simplification and genome reduction. Forterre had already formulated this idea in 1995, assuming that this simplification occurred when transitioning to a thermophilic ecological niche (≈ 50–70°C). This is the hypothesis of thermo-reduction. Indeed, in such a context, organisms that reproduce the fastest and possess the most stable DNA in hot environments are favored. And this is precisely the case for prokaryotes, which reproduce very quickly and whose circular DNA is more stable in hot environments than the linear DNA of eukaryotes. However, at present, fossil data contradict this hypothesis, with the oldest prokaryote traces dating back at least 3.4 billion years (Javaux, 2019) (**Box 3**). That being said, the absence of proof is not proof of absence! Moreover, since this simplification is genetic, there is nothing to say that this ancestor was morphologically a eukaryote. It could very well have been a eukaryote or proto-eukaryote with a different structure and membranes unlike those of current eukaryotes. It has been suggested that this membrane was bifunctional for the production of hopanoids and sterols (Desmond & Gribaldo, 2009; Francis, 2021; Lombard et al., 2012; Peretó et al., 2004). Furthermore, molecular fossils actually allow us to date when the eukaryotic membrane appeared (**Box 3**), but there is nothing to say that these membranes did not exist earlier with a different type of membrane (Francis, 2021). However, this does not explain the origin of the eukaryotic cell itself, which could be older than the membrane traces found, which actually correspond to the membranes observed today.

Box 3. Microfossils support ancient origin of eukaryotes

Molecular, biogeochemical, and paleontological data suggest that eukaryotes appeared early in Earth's history. The oldest fossil eukaryote proposed is *Grypania spiralis*, discovered in Michigan (USA), and dated to 2.1 billion years ago. Its classification as a eukaryote is based solely on its large size. However, size is not a reliable criterion for distinguishing prokaryotes from eukaryotes. Moreover, its morphology could correspond to filaments of cyanobacteria.

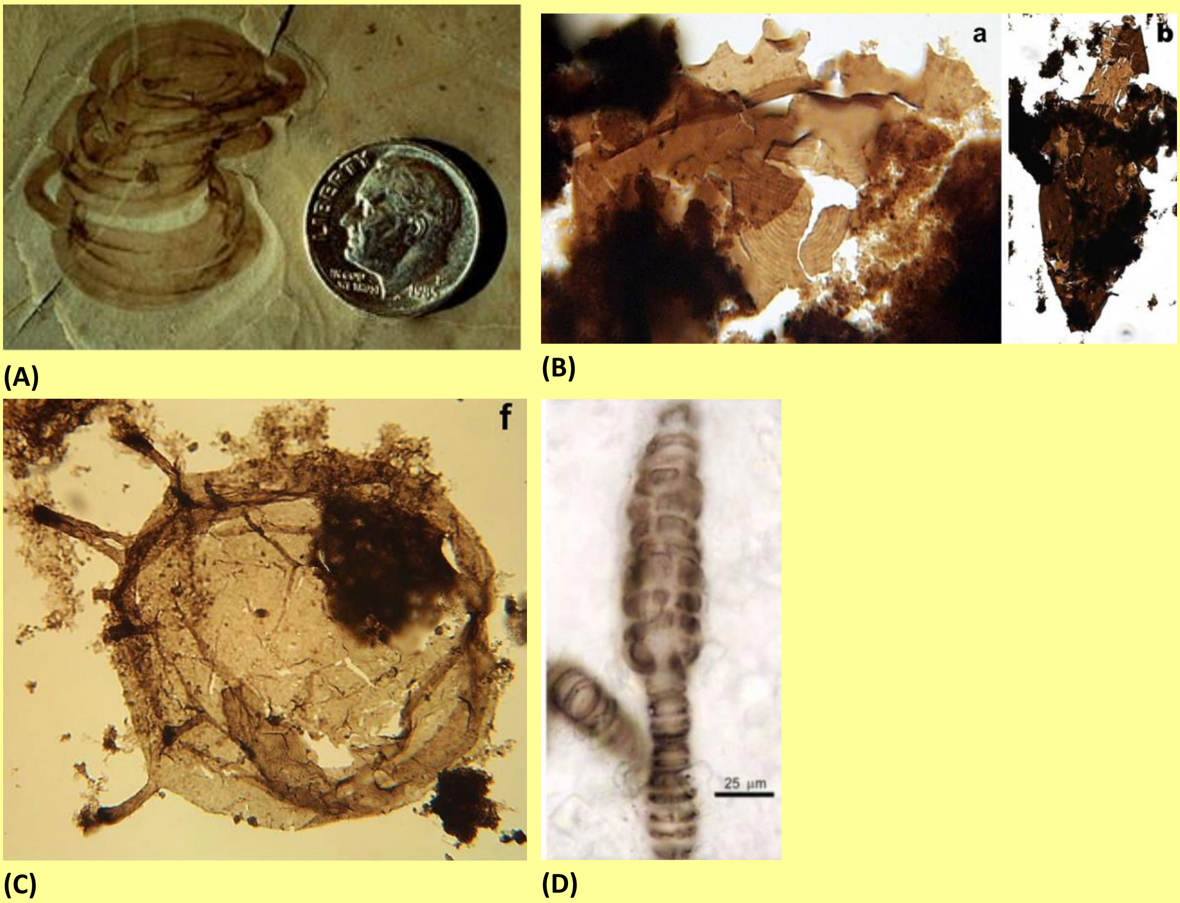


Figure 8. (A) *Grypania spiralis* - A possible early eukaryote fossil, dated to 2.1 billion years ago. (B) *Valeria lophostriata*, > 1.65 billion years ago, Mallapunyah Formation, Australia. This early eukaryote has a wall ornamented with concentric striations. (C) *Tappania plana*, a protist from the Roper Group, Australia, dated to 1.5 billion years ago. (D) *Bangiomorpha pubescens*, a possible red algae, which could make it the oldest eukaryote belonging to a current group.

A more reliable morphological feature distinguishing eukaryotes from prokaryotes is the ornamentation of cellular surfaces, which is specific to eukaryotes. The oldest ornamented microfossils are acritarchs (from the Greek akritos, meaning "uncertain" or "confused," and arche, meaning "origin," or "the first"). These are microfossils of uncertain biological affinity, polyphyletic in origin.

Fossils of mycelial structures dated to 2.4 billion years ago, discovered in marine basalt from the Ongeluk Formation in South Africa and initially attributed to fungi, remain controversial and require further evidence for definitive identification. These may actually be biomorphs, structures resembling living organisms but not necessarily of biological origin, or fossils of primitive microorganisms distinct from classical eukaryotes. These microfossils found in marine basalt resemble fungal mycelia due to their filamentous appearance and branched, anastomosing structures. However, their great age raises questions because fungi, as eukaryotes, were not expected to exist at that time, well before the widely accepted appearance of this group around 1 billion years ago. It is possible that these represent primitive forms of life that evolved in the deep biosphere of the ocean floor in volcanic environments. It is also possible that these fossils represent unicellular life forms close to prokaryotes, capable of developing in these extreme habitats.

Ancient microorganisms have also been discovered in the Paleoproterozoic Hutuo Group, in the Wutai Mountains of Shanxi Province, North China. The Dongye sub-group within the Hutuo group is particularly interesting. This includes *Dongyesphaera tenuispina* and *Dongyesphaera gen. nov*., characterized by spiny ornamentation, and *Dictyosphaera*, which has a network of ornamentations. Recent zircon dating places the Hutuo Group deposits around 2150-1950 million years ago. These microfossils provide crucial evidence of eukaryotic metabolism at a time when Earth was undergoing significant geological changes.

The oldest confirmed fossil eukaryotes are organic-walled microfossils (OWM) from the lower Changcheng Group (1673-1638 million years ago, Changzhougou and Chuanlinggou formations) in the Yanshan chain, North China, consisting of 15 species. The fossil assemblage is dominated by spheroidal forms, with fewer vesicles bearing processes, as well as colonial and filamentous forms. Among these, six morphologically complex taxa (*Dictyosphaera delicata*, two species of *Germinosphaera*, *Pterospermopsimorpha*, *Simia*, and *Valeria lophostriata*) are unambiguously identified as unicellular eukaryotes. Four species (*Cucumiforma*, *Navifusa*, *Schizofusa*, and large *Leiosphaeridia*), with relatively simple morphology but large size and thick walls, some of which exhibit exocytosis structures with a median cleft, are likely of eukaryotic affinity. However, various colonial microfossils may represent either eukaryotes or prokaryotes. The new OWM fossil record, morphologically diverse, represents one of the earliest occurrences of eukaryotes in China and worldwide, indicating that eukaryotic life was well-established by the late Paleoproterozoic and exhibited moderate diversity similar to that of the Mesoproterozoic.

Another group of Paleoproterozoic eukaryotic microfossils (1642 million years ago) was found in the Limbunya Group in the Birrindudu Basin, Northern Australia. These microfossils, including new and well-preserved species, show a rich diversity of eukaryotes, challenging the idea that eukaryotic diversification occurred only about 800 million years later. Twenty-six taxa were identified, including 12 considered eukaryotic, with four newly named species. These eukaryotes thrived in marginal marine environments such as intertidal zones and lagoons, particularly in the Blue Hole formation, which shows a highly diverse assemblage. Another major discovery is that these fossils display signs of morphological and cellular complexity, suggesting that eukaryotes already possessed sophisticated characteristics such as cytoskeletons and internal structures. This morphological diversity and abundance of species in these ancient ecosystems indicate that the ancestors of modern eukaryotes evolved much earlier than expected, raising new questions about the evolution of these organisms and their adaptation to various environments.

Valeria lophostriata is also found in the Mallapunyah Formation (McArthur Supergroup, Australia, 1.65 billion years ago) and in many younger Proterozoic siliciclastic successions. This spherical acritarch is easily distinguished by its concentric striations. These striations are wall ornaments, as shown by scanning electron microscopy (SEM) images. The ridges are spaced 1µm apart and are located on the inner face of the vesicle. *Valeria* has a wide stratigraphic distribution, extending from the late Paleoproterozoic to the Mesoproterozoic and Neoproterozoic.

Other fossils, such as *Tappania plana*, aged 1.5 billion years and discovered in the Roper Group shales in northern Australia, show acanthomorphic structure (= presence of spiny processes), indicating that the cytoskeleton and ecological prerequisites for eukaryotic diversification were already established at this time. These acritarchs could represent a basal group of eukaryotes.

The earliest fossils that might correspond to a present-day group (crown group) are those of *Bangiomorpha pubesens*, found in the Hunting Formation in Canada and dated to 1.047 billion years ago. These have been interpreted as red bangiaceae algae, based on cellular division figures (multicellular appearance) and sexual reproduction, placing a lower limit on the age of the appearance of "primary" chloroplast-bearing plants.

Regarding biomarkers for eukaryotes (molecular fossils of steranes), they are abundant and unambiguous from 1.7 billion years ago, supporting the idea that eukaryotes are ancient. In the Barney Creek formation, in northern Australia near Borroloola, the protosterol biota consists of aquatic bacteria producing protosterols and ancient basal eukaryotes, dating from 1.6 to 0.8 billion years ago (Tonian period). These organisms were more complex than current bacteria and preceded the last common ancestors of all modern eukaryotes. They were predators feeding on bacteria and possibly other eukaryotes. They were abundant in aquatic environments of the seas and significantly impacted the terrestrial ecosystem of the time. These microorganisms adapted to the much lower oxygen levels of the time and likely also produced protosteroids. Chemical signals suggest that these molecules may originate from an ancestor of the last common eukaryotic ancestor from which fungi, plants, and animals all evolved. Modern eukaryotes likely appeared during this "Tonian transformation", followed by the proliferation of red algae about 800 million years ago. This phase represents one of the most profound ecological turning points in Earth's history.

References

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.

However, a particular monophyletic group of organisms tends to support a bacterial origin for eukaryotes and archaea: the PVC group bacteria. This group includes Chlamydiae, Lentisphaerae, the candidate phylum Omnitrophica, Planctomycetes, the candidate phylum Poribacteria, and Verrucomicrobia. PVC bacteria exhibit unusual genetic and cellular characteristics for bacteria but are specific to archaea and/or eukaryotes (D. P. Devos & Reynaud, 2010; Reynaud & Devos, 2011; Rivas-Marín & Devos, 2018; Santarella-Mellwig et al., 2010; Wiegand et al., 2018). Three major steps lead to the emergence of eukaryotes and archaea: (1) the loss of the peptidoglycan-based cell wall in bacteria, (2) the replacement of the bacterial cytoskeleton protein FtsZ by eukaryotic tubulin, and (3) the development of an endomembrane system in eukaryotes. Members of the PVC group appear to be "intermediate" for these two events.

Based on the homology between many components of the eukaryotic endomembrane system and those of the bacterial periplasm, it has been suggested that the eukaryotic endomembrane system is the result of the internalization of the bacterial periplasm (Blobel et al., 1986; de Duve, 2007). Moreover, the emergence and development of the eukaryotic endomembrane system have been based on MCPs (membrane coat proteins) (D. Devos et al., 2004). Planctomycetes have always been of interest for eukaryogenesis scenarios because of their developed endomembrane system (Forterre, 2011; J. A. Fuerst & Nisbet, 2004), which is undoubtedly one of the most developed among prokaryotes (Acehan et al., 2014; Boedeker et al., 2017; D. P. Devos et al., 2014; Santarella-Mellwig et al., 2013). In these bacteria, the cytoplasmic membrane sends invaginations into the cytoplasm, forming a complex organization and representing a "true" internalization of the periplasm (D. P. Devos et al., 2014; Santarella-Mellwig et al., 2013). However, without molecular links between eukaryotic and bacterial endomembrane systems, Planctomycetes remained a curiosity in the prokaryote world. The detection of proteins with the same structural signature as eukaryotic MCPs in the proteomes of various Planctomycetes (and related species), and the demonstration that these proteins support their endomembranes, have significantly changed this situation (Santarella-Mellwig et al., 2010). Although there is no sequence signal between the MCPs of Planctomycetes and those of eukaryotes, the presence of these proteins in prokaryotes is unique and represents the first molecular link between eukaryotic and prokaryotic endomembrane systems. If their structural and functional similarities suggest an evolutionary relationship, the exact nature of this relationship, convergent or divergent, remains to be determined (D. P. Devos, 2012).

The development of phagocytosis was a crucial step in the process of eukaryogenesis. Once again, Planctomycetes display related phenotypes. Indeed, it has been demonstrated that the planctomycete *Gemmata obscuriglobus* is capable of internalizing proteins before degrading them internally in a process reminiscent of eukaryotic endocytosis (Lonhienne et al., 2010). This observation has since been extended to other molecules, such as dextran, and to another species of Planctomycetes, *Planctopirus limnophila*, suggesting that this ability is more widespread

(Boedeker et al., 2017). Furthermore, it has been shown that another Planctomycete, *Candidatus Uab amorphum*, is capable of phagocytosing other bacteria in a manner similar to eukaryotes (Shiratori et al., 2019).

Other characteristics that are generally not found in bacteria but are more often associated with eukaryotes or archaea, or both, are found within the PVC group. These characteristics include:

- Sterol synthesis, which was previously thought to be linked to eukaryogenesis, turns out to be of bacterial origin (Santana-Molina et al., 2020). Thus, some bacteria, such as the planctomycete *Gemmata obscuriglobus*, are capable of producing sterols.
- Tubulin-related proteins have been described in *Verrucomicrobia* (Pilhofer et al., 2011) and proteins containing a tubulin-like domain have been detected in Planctomycetes (Makarova & Koonin, 2010). Similarly, the genome of *Candidatus Uab amorphum*, the phagocytic planctomycete, encodes a protein related to actin (Shiratori et al., 2019). However, the phylogenetic relationships of these tubulin and actin proteins relative to their homologues in archaea and eukaryotes are still undefined.
- Anammox Planctomycetes possess hydrocarbon chains that are linked by either an ether or an ester to glycerol, known as ladderane lipids, providing a possible solution to the question of lipid transition (Villanueva et al., 2021).
- Various metabolic pathways, previously considered to be eukaryotic, have been discovered within the PVC group. These include genes related to C1 transfer metabolism (Planctomycetes), which may be linked to the origin of methanogenesis (Chistoserdova et al., 2004), and enzymes from the mevalonate pathway of isoprenoid biosynthesis in the *Verrucomicrobia* and *Lentisphaerae* phyla, which are generally associated with archaea and eukaryotes (Hoshino & Gaucher, 2018). Similarly, homologues of eukaryotic serine/threonine kinases and ubiquitin E2 have been detected in Planctomycetes (Arcas et al., 2013).
- Finally, Verrucomicrobia divide by binary fission using the FtsZ protein, demonstrating that the last common ancestor of the PVC group had the ftsZ gene and divided by binary fission, like most bacteria. However, all members of *Chlamydiae* and *Planctomycetes* have lost the ftsZ gene, as well as other genes related to division and cell wall (dcw), and now divide by asymmetric division (Rivas-Marín & Devos, 2018; Santana-Molina et al., 2020). Their division mechanisms are currently unknown, but it is known that the loss of ftsZ and the development of a new mode of cell division occurred during eukaryogenesis. The common ancestor of the PVC group and LAECA (Last Archaea-Eukaryote Common Ancestor) may therefore have been a bacterium that began accumulating characteristics that would later be recognized as specific to eukaryotes and/or archaea (Makarova & Koonin, 2010; Reynaud & Devos, 2011).

Despite the presence of some of these characteristics in other bacteria (McInerney et al., 2011), members of the PVC group are the only ones to exhibit so many of these traits in a single group of related organisms. This blend of phenotypes in current PVC members raises the hypothesis that they might be linked to intermediate stages in the transition from bacteria to an ancestor of LAECA (D. P. Devos & Reynaud, 2010). In this scenario, where PVC bacteria and LAECA share a common ancestor, there exists a lineage, or community of organisms (O'Malley et al., 2019), linked to PVC bacteria and intermediate between the ancestor of archaea and eukaryotes on the path to LAECA, which had already developed some of its shared characteristics (those shared between the ancestor of eukaryotes and archaea). Thus, in this scenario, PVC bacteria are the sister taxa of the LAECA lineage. The root of archaea and eukaryotes may, but not necessarily, be situated between the archaea and eukaryote lineages. The question of the complexity of the ancestor of each domain is also addressed in this 1D scenario, where the LAECA diverges from bacteria toward greater complexity. There is thus a continuous increase in complexity from bacteria to eukaryotes, while it decreases in archaea (secondary simplification) (D. P. Devos, 2021).

4.4 DISCUSSION OF THE PROPOSED MODELS

A multitude of models aimed at explaining the origin of the eukaryotic cell have emerged in recent years (**Figure 9**). However, none seem to fully address all the questions simultaneously. Moreover, the nature and properties of FECA, as well as its phylogenetic relationship, remain a source of ongoing questions (**Box 4**).

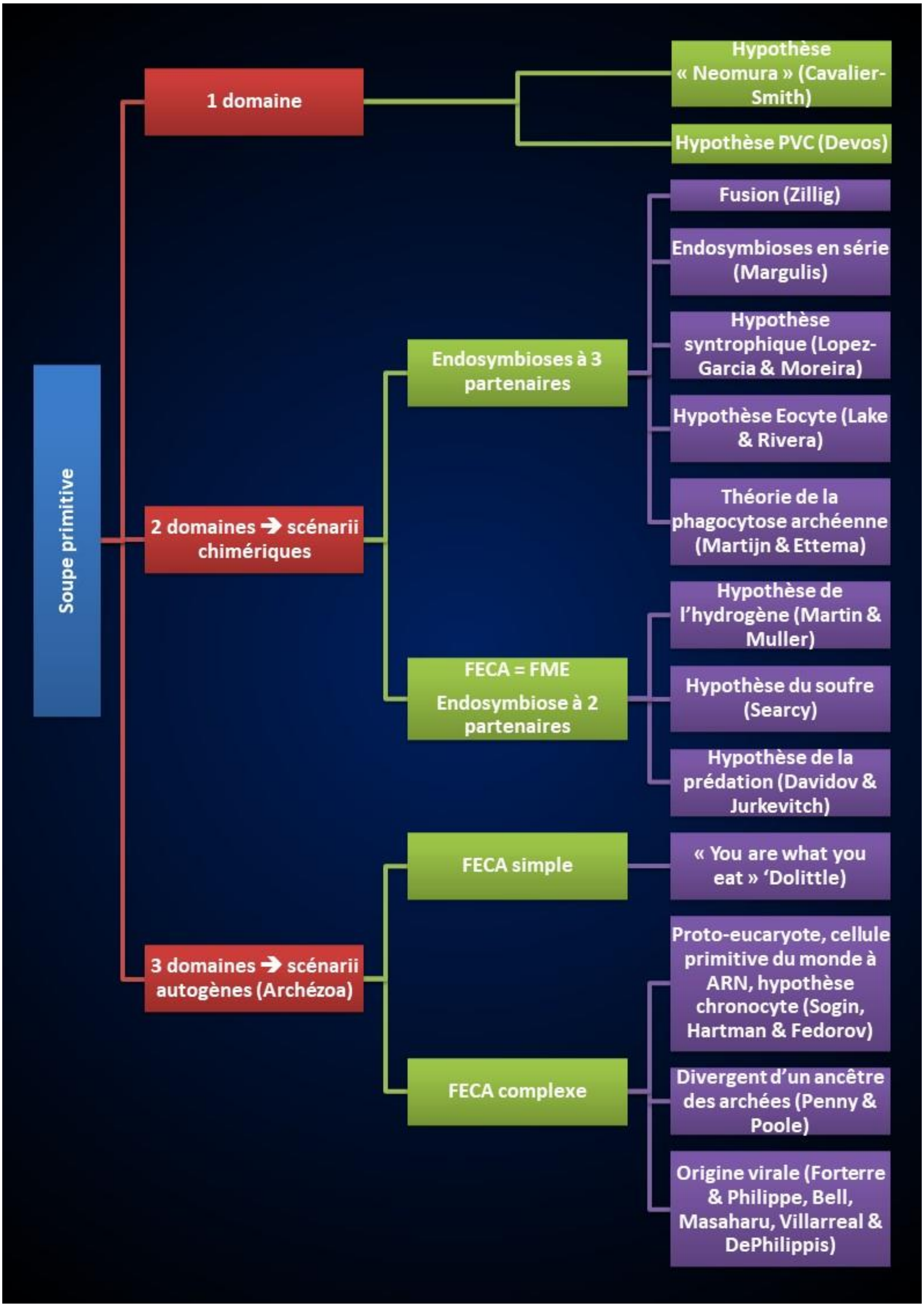
In light of recent discoveries in the physiology, molecular biology, and genomics of archaea and bacteria, many fusion models have been challenged (Forterre, 2011). For instance, the models of Margulis and Searcy involving Thermoplasma-like archaea as the origin of the eukaryotic nucleus have since been refuted. Molecular analyses have shown that the "histones" of Thermoplasma are not homologous to those of eukaryotes but to bacterial HU proteins (DeLange et al., 1981). Likewise, the idea that eukaryotes might have derived from methanogenic archaea (Martin & Müller's hydrogen hypothesis, Moreira & Lopez-Garcia's syntrophy hypothesis) is outdated since the discovery of these same histones, analogous to those of eukaryotes, in mesophilic Thaumarchaea (Brochier-Armanet, Gribaldo, et al., 2008; Čuboňová et al., 2005). Moreover, the involvement of a δ-proteobacterium or a spirochete, as suggested by the syntrophy hypothesis, has also been dismissed after genome sequencing failed to reveal specific affinities with eukaryotes and following the discovery of kinases and G proteins in various groups of archaea and bacteria (Dong et al., 2007; Tyagi et al., 2010).

The models in which the nucleus is an endosymbiont (J A Lake and Rivera, 1994) have been firmly rejected (Martin, 1999; A. Poole & Penny, 2001; Rotte & Martin, 2001). The discovery of proteins once thought to be specific to eukaryotes (ESPs) was used as an argument to support a three-partner fusion model (Hartman & Fedorov, 2002). However, a simpler explanation is that the eukaryotic domain had already separated from the archaeal domain before the acquisition of the mitochondrion (Amiri et al., 2003).

In fact, the choice of partners in the proposed fusion models has always been based on the state of knowledge at the time. In 2011, Forterre proposed a new fusion model, the best possible at that time, and also pointed out what was still incorrect in his model (Forterre, 2011). This "exercise in style" aimed to show that many fusion partners could actually be proposed. He then proposed a fusion model in which FECA resulted from the phagocytosis of a thaumarchaeon by a PVC bacterium (Planctomycetes, Verrucomicrobia, Chlamydiae), followed by the massive invasion of its descendants by various viral lineages (such as NCLDV = Nucleo-Cytoplasmic Large DNA Viruses) and retroviruses. Indeed, eukaryotic traits have been identified in PVC bacteria (Santarella-Mellwig et al., 2010; Wagner & Horn, 2006) and Thaumarchaeota (Brochier-Armanet, Boussau, et al., 2008; Brochier-Armanet, Gribaldo, et al., 2008; Spang et al., 2010). Thaumarchaeota would have provided informational and operational proteins, while PVC bacteria would have provided phospholipids, tubulin, and membrane proteins necessary for the formation of the nucleus. Viruses would have then contributed to eukaryotic-specific proteins, leading to the complexity of the proto-eukaryote. However, his model does not explain the origin of the nucleus, the distribution of archaeal and bacterial traits within eukaryotes, the origin of eukaryotic viruses, or the existence of three distinct ribosome lineages. Thus, he concludes that eukaryotes and their viruses most likely evolved from a specific lineage, according to the three-domain scenario proposed by Woese.

Box 4. What did LECA, the common ancestor of today's eukaryotes, look like?

It is not certain that the characteristics of eukaryotes (nuclear envelope, endomembrane system, intron splicing, and linear chromosomes) preceded the appearance of the mitochondrion. However, by comparing the properties of different groups of extant eukaryotes, it is possible to construct a profile of the common ancestor of modern eukaryotes (LECA) and infer its minimal characteristics. Thus, LECA was probably heterotrophic, lacked a cell wall, and engaged in extracellular digestion. The first step leading to modern eukaryotes must have been the development of phagocytosis. The formation of membrane folds inside the cell would have enabled intracellular digestion, giving rise to lysosomes. This ability to form small intramembrane vesicles may have been the origin of the nuclear envelope and the endomembrane system. Its chromosomes were likely already linear, thus requiring a continuous process of end-repair via the formation of telomeres. Introns were already present and had to be excised from RNA transcripts during the excision-splicing maturation process by a spliceosome. Only after this step would the RNA transcripts be transported out of the nucleus to the cytoplasm to serve as a template for protein synthesis. Unlike bacteria and archaea, transcription and translation were therefore no longer synchronous. The appearance of a cytoskeleton was likely a response to the loss of the cell wall, in order to maintain the cell membrane and structure the cytoplasm. LECA was thus already a complex organism with all eukaryotic characteristics and was therefore a typical eukaryotic cell. As a result, all key innovations of eukaryogenesis must have occurred before the appearance of LECA, in the root groups (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. Summary of the Different Scenarios for Explaining the Origin of the Eukaryotic Cell. These scenarios can be grouped based on the number of domains considered, the mechanism of eukaryote formation (symbiosis vs. autogenous), and the nature of the first eukaryote (with or without mitochondria).

Modern eukaryotes differ from prokaryotes through various features (Vellai & Vida, 1999) :

- The cell is compartmentalized by an internal membrane network enclosing the nucleus and organelles. The nucleus houses the genetic material within a nuclear envelope, while organelles include the endoplasmic reticulum, the Golgi apparatus, and lysosomes.
- The cytoplasm is structured by a cytoskeleton made up of actin filaments and tubulin microtubules.
- Cell division occurs via mitosis, during which DNA is compacted into chromosomes before being divided.
- Reproduction is genuinely sexual in many eukaryotes, with each sexual type contributing an equal share of genetic material to the next generation.
- The genome contains repeated DNA sequences (microsatellites).
- Introns often undergo splicing, making the eukaryotic genome a "genes in pieces" structure(Gilbert, 1978; Gilbert et al., 1986).
- Originally, eukaryotes possess organelles bounded by membranes that perform the cell's energy functions: mitochondria (responsible for respiration, among other roles) and, in photosynthetic organisms, chloroplasts (the site of photosynthesis).

Alongside eukaryotes, prokaryotes (bacteria and archaea) have long been defined negatively by the absence of certain features (Philippe et al., 1995):

- They lack internal membrane networks, particularly around their DNA, meaning they do not have a well-defined nucleus. A notable exception is the nucleoid in bacteria of the PVC group (JohnA. Fuerst, 2013; McInerney et al., 2011; Wagner & Horn, 2006).
- They lack individualized cytoplasmic organelles.
- Genetic material division occurs through a segregation process where bacterial DNA remains attached to the cell envelope rather than undergoing mitosis.
- In bacteria, a mucopeptide cell wall can provide an external structural framework.

This paradigm of prokaryotes aligns with the traditional idea of gradual evolution from simple to more complex life forms (Brinkmann & Philippe, 2007). Prokaryotes encompass cells from two distinct domains: bacteria and archaea. Classifying prokaryotes phylogenetically is challenging. Unlike animals and plants, where comparative anatomy and physiology can be applied, bacteria lack complex morphological traits and display high physiological and biochemical diversity that is difficult to interpret (Philippe et al., 1995). Consequently, molecular phylogenetic techniques, based on genetic markers, are now preferred. Phylogenetic classification initially relied on individual genes (notably ribosomal RNA) but has increasingly incorporated more genes as genomics has advanced. Recent breakthroughs in molecular techniques, particularly high-throughput sequencing and metagenomics, have significantly improved our ability to discover new organisms, filling gaps in phylogenetic classifications.

GOALS

Our thesis aims to provide new insights into the ongoing discussions on the domains of life. We will endeavor to answer three key questions:

1. Why are the number of domains and the rooting of the tree of life still unresolved?
2. What is the phylogeny of archaea?
3. What is the relationship between eukaryotes and archaea?

These three questions are central to the field of evolution and significantly impact our overall understanding of evolutionary processes. Each question will be addressed in a dedicated chapter of this thesis.

In the **first chapter**, we will discuss the biases that hinder the quest for answers, as well as the importance of techniques and cognitive biases. The appropriateness of the methods used will be examined, given that these are only approximations. From a methodological standpoint, this chapter will address six major issues inherent to biological data and phenomena: (1) the heterogeneity of substitution among characters, (2) the heterogeneity of substitution rates across sites, (3) the heterogeneity of substitution processes across sites, (4) compositional heterogeneity over time, (5) the heterogeneity of substitution processes at a site over time, and (6) the heterogeneity of substitution rates at a site over time within a lineage.

We will also discuss methodological biases, the methods used to create alignments (alignment, position selection, etc.), and the models employed in phylogenetic reconstructions, which struggle to address these issues caused by the heterogeneity of biological phenomena. Furthermore, we will explore our own perspective on evolution and the inherently anthropocentric nature of our perception.

This conceptual first chapter will establish a solid foundation for the arguments concerning the evolution of archaea and eukaryotes defended in this thesis. It will be crucial for understanding evolutionary questions, particularly those concerning deep evolutionary events, such as the origin of the domains of life or the emergence of the eukaryotic cell. This chapter will serve as the basis for the rest of the thesis.

In the **second chapter**, the focus will shift to the evolution of archaea and the relationships among the various groups that comprise them. The primary goals will be to (1) create a robust dataset to establish the phylogenetic relationships between different archaea, and (2) test these relationships by detecting potential biases that could lead to misinterpretations.

We will begin by selecting proteomes of interest, then generate and select orthologous groups (OGs) based on several criteria (taxonomy, number of species, etc.). After enriching these groups, we will create two datasets: one containing strictly orthologous OGs and another containing paralogous OGs derived using a phylogenetic slicing method. The first dataset will represent a highly reliable resource, while the second will be more prone to various artifacts.

The OGs will undergo additional filtering based on branch lengths, taxonomic diversity (including at least one Asgard, one DPANN, one Euryarchaeota, and one TACK). These datasets will then be used for phylogenetic reconstruction. Using robustness tests (gene jackknife, five species replicas) with different models, we will compare the resulting trees using Robinson-Foulds distances. From this, we will deduce the majority topologies based on the datasets to identify groups with unstable positions. Finally, slow-fast tests will be performed using positions less prone to evolutionary changes and, therefore, less affected by phylogenetic biases such as long-branch attraction (LBA).

In the **third chapter**, the results from Chapter 2 will be utilized to determine the placement of eukaryotes. To this end, data from 11 eukaryotic species will be added to the previously generated datasets. The formed orthologous groups will be evaluated to remove duplicates, particularly those arising from organellar genomes (mitochondria, plastids).

Using the various replicas generated in Chapter 2, we will conduct analyses to identify potential biases discussed in Chapter 1 (especially phenomena like heterotachy and heteropecilly). Following this, to infer evolutionary scenarios, we will attempt rooting using AU-tests and rootstrap methods on the most reliable species replica. Finally, Bayesian methods will be employed to calculate a final tree, which will be rooted based on our previous findings.

In conclusion, the general discussion will integrate all results to draw appropriate conclusions regarding the phylogenetic relationships between Asgard archaea and eukaryotes.

MATERIALS & METHODS

1 GENOME DOWNLOAD

Genomes and proteomes of archaea were retrieved from both RefSeq and GenBank (RefSeq genomes are a subset of those in GenBank) via the National Center for Biotechnology Information (NCBI) portal. A total of 584 conceptual proteomes corresponding to an equal number of genomes were downloaded from RefSeq, while 1,077 genomes (of which only 770 proteomes were available) came from GenBank. The downloads were performed on February 8, 2017.

2 GENOME QUALITY EVALUATION WITH QUAST

QUAST (v2.4) (Gurevich et al., 2013) was used to estimate the quality of genome assemblies in order to retain the highest-quality genomes possible. For each genome, the following metrics were collected: the number of scaffolds, the number of scaffolds > 1,000 nt, total genome size, total genome size based on scaffolds > 1,000 nt, the largest scaffold size, GC content, N50/N75/L50/L75 values, and the number of Ns per 100 kbp (Ns being unassigned positions). Assemblies were assessed based on three key dimensions: contiguity, completeness, and purity (including contamination presence). Contiguity is often measured using the contig N50 metric. Given a minimal set of contigs sorted in decreasing length, the N50/N75 values represent the length of the contig at 50%/75% of the genome’s total length, while L50/L75 denotes the rank of that specific contig. A contig N50 exceeding 1 Mbp is generally considered excellent. QUAST results are provided in the **quast.csv** file. Completeness was assessed by searching for ribosomal proteins in each genome. Accuracy was evaluated by contamination checks. High-quality genomes were defined as those with good assembly contiguity, maximal completeness, and minimal contamination.

3 RIBOSOMAL PROTEIN ANALYSIS AND PRELIMINARY TAXONOMIC SAMPLING

To establish a preliminary phylogeny of archaea, an initial dataset was assembled by enriching 90 multiple sequence alignments (MSAs) from the RiboDB 1.4.0131 ribosomal protein database (available at https://bitbucket.org/phylogeno/42-ribo-msas/src/master/MSAs/prokaryotes/). Orthologous groups were generated from our selection of archaea and enriched using these MSAs.

The *Forty-Two* (42) tool (Irisarri et al., 2017; Simion et al., 2017), was employed to add and align sequences to pre-existing MSAs while ensuring orthology and screening for potential contaminants. In practice, 42 identifies homologous sequences in each genome, refines orthologs from potential paralogs via advanced heuristics, and integrates them into the MSAs. It was run on the 1,077 GenBank genomes (including the 584 from RefSeq). The number of genomes from which ribosomal protein orthologs were successfully retrieved is provided in TABLEAU 2 (**stats-completude.txt**).

To verify whether the archaeal genomes were contaminated by bacterial sequences, a contamination test was performed using *42*, with bacterial proteomes serving as the reference for the Best Reciprocal Hit (BRH) method (**Box 9**). This method relies on a heuristic known as multiple BRH. The user selects a list of organisms present in the multiple sequence alignments (MSAs) and a set of proteomes that will serve as the reference for validating orthology (**Figure 10**). The sequences from the first list (the "queries") are used to search for corresponding homologs using BLAST in two sets: (1) the search database, which can consist of genomes, proteomes, or transcriptomes, and (2) the reference proteome set. Following this, orthology is validated between the best hits obtained from the different reference proteomes based on the BRH principle. This step consolidates a list of reference proteomes (and sequences) that best confirm orthologous relationships for the given MSA. Additionally, the homologous sequences recovered in step (1) are compared with the remaining reference proteomes. If the best hits from a homolog belong to the consolidated list of sequences, it is considered orthologous. It is important to note that the definition of orthology is not restricted by the initial "query." A sequence identified as homologous by one query can be validated as orthologous by reference sequences identified through another query. In addition to this orthology verification heuristic, *42* also implements a taxonomic check of the added sequences to identify or avoid contamination (**Figure 10**). After determining the orthologous sequences, each is compared to the alignment sequences to find the most similar one. This closest sequence serves as a template for aligning the new sequence and inserting it into the MSA, while also assigning a taxonomy. The taxonomy is automatically retrieved from the organism's name corresponding to the most similar sequence in the initial alignment, according to the NCBI database. For each new organism to be added, a taxonomic filter can be applied to its orthologous sequences. The affiliated taxonomy is then compared to this filter, which can be inclusive (the sequence must belong to the mentioned clade), exclusive (it must not belong to the mentioned clade), or both. This final step allows for the removal or marking of sequences that do not meet the requested taxonomic filter. Its purpose is to avoid contamination from known or potentially unknown sources, as well as to limit the addition of recently transferred sequences (LGT).

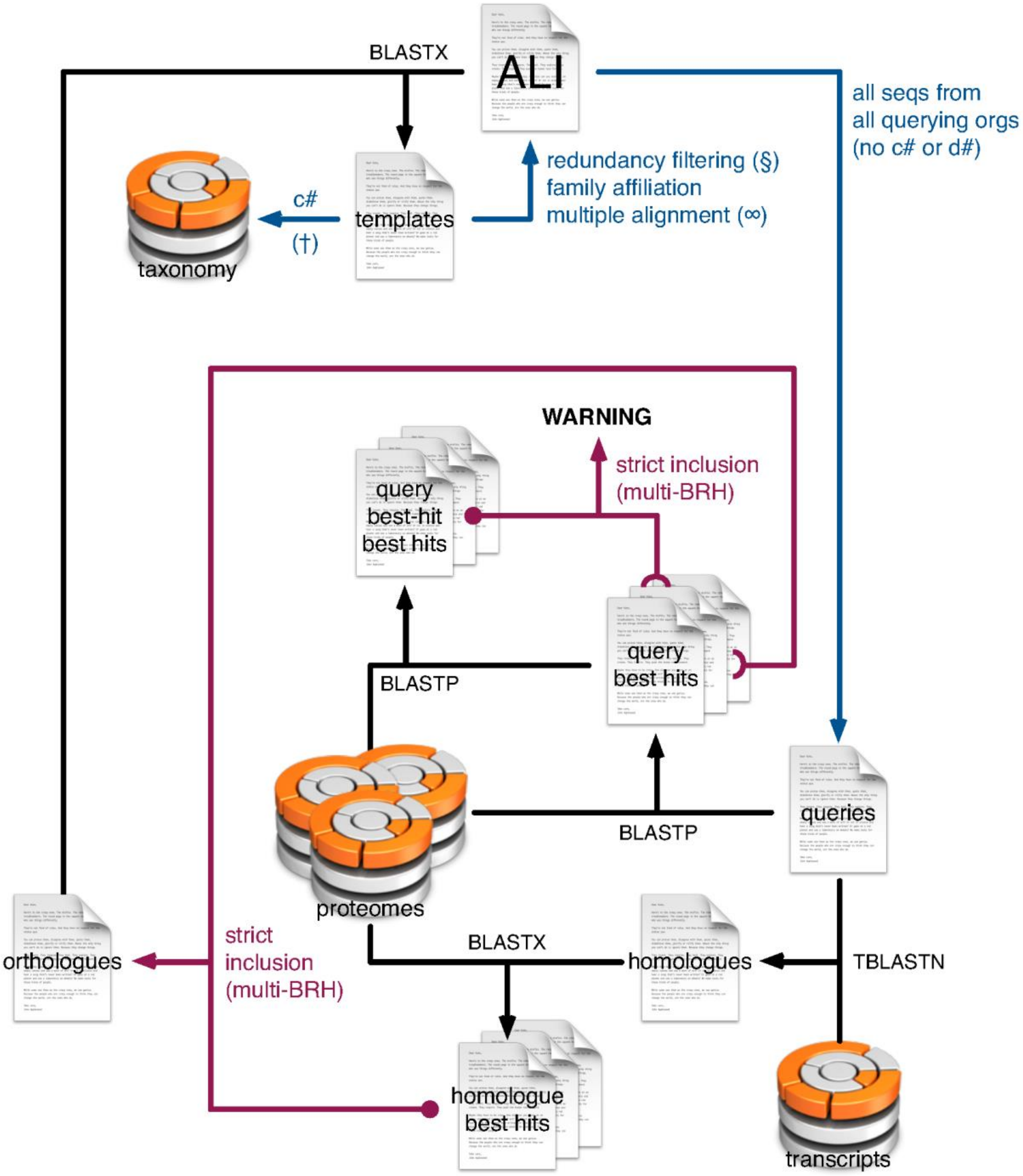

BLASTX
ALI
all seqs from
all querying orgs
(no c# or d#)
taxonomy
c#
(†)
templates
redundancy filtering (§)
family affiliation
multiple alignment (∞)
WARNING
strict inclusion
(multi-BRH)
query
best-hit
best hits
query
best hits
BLASTP
BLASTP
queries
proteomes
strict
inclusion
(multi-BRH)
orthologues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

Figure 10. Diagram of how 42 works (Denis Baurain, User Guide for 42)
42 uses the Best Reciprocal Hit (BRH) principle to identify orthologous sequences within given genomes and add them to proteomes. A BRH occurs when two proteins encoded by genes in two different genomes are identified as each other's best match in the other genome. In the case of 42, the addition of orthologous sequences to alignments is performed via a triangular BRH, involving a set of intermediate proteomes as reference points to validate orthology criteria and thus prevent the inclusion of paralogous sequences. The user starts with a set of alignments of orthologous groups they wish to enrich with translated genomes. They extract the list of all organisms present in these alignments and select a subset that best represents the taxonomic diversity of their sampling. These are referred to as the *query_orgs*. It is also important that these *query_orgs* appear in as many alignments as possible. First, 42 extracts the sequences of all the *query_orgs* to obtain a set of *query_seqs*. Then, it performs a BLASTP of these *query_seqs* against a set of reference proteomes. Ideally, the reference proteomes should correspond to the *query_orgs* to optimize BLAST scores. If a *query_org* does not have an available proteome, a phylogenetically close species can be used as a substitute. In the second step, a TBLASTN is conducted using the *query_seqs* on the genomes in the target database to identify homologous sequences, and thus potential orthologs. A reverse check is then performed on these potential orthologs by running a BLASTX of these sequences against the reference proteomes. This strategy eliminates paralogous sequences, retaining only orthologous ones. For a candidate sequence to be considered orthologous, its best score in the BLASTX against the reference proteomes must correspond to the same best score obtained by the *query_orgs* in the initial BLASTP against these same proteomes.

42 not only assesses the quality of genomic and proteomic data—by estimating contamination or evaluating data completeness—but also enriches orthologous groups with sequences from additional organisms. In summary, each orthologous sequence is classified by calculating the Last Common Ancestor (LCA) of the two sequences from the multiple sequence alignment (MSA) that are most similar to the newly added ortholog. This classification requires a minimum identity threshold of 90%. The acquisition of a set of aligned orthologous gene sequences from genomic data serves as the starting point for constructing a phylogenomic dataset. Orthologous sequences from other species are then identified based on these alignments. By default, the inferred LCA is expected to align with the taxonomy of the genome from which the ortholog originates, typically of archaeal origin. If this is not the case, the ortholog is flagged as a contaminant of the genome assembly—or remains unclassified if it does not match anything sufficiently close in the RiboDB reference database. Our goal is to construct a ribosomal tree to guide our species selection process. To achieve this, we used SCaFoS v1.30k (Roure et al., 2007) to retain 1,070 genomes while excluding 7 contaminated genomes and the original genomes (all duplicates) from RiboDB. We then concatenated our ribosomal protein MSAs into a supermatrix of 8,344 amino acid positions x 1,070 species, again using SCaFoS. Before phylogenetic inference, conducted with RAxML v8.1.17 (Stamatakis, 2014) using a rapid-bootstrap search (100 replicates) under the PROTCATLGF substitution model, we removed sites with >90% missing character states. This was done using ali2phylip.pl (option --max=0.1) from the Bio-MUST-Core distribution (D. Baurain; available at https://metacpan.org/pod/Bio::MUST::Core). From the inferred tree (**Tree-ribo-1070-sp**), we manually selected 364 species collectively representing the diversity of archaea (**archaea-364sp.lis**). Of these 364 species, 305 had available proteomes (cf. **Supp. Mat. selection-proteomes.txt**), while the remaining 59 species were only available as unannotated genome assemblies (cf. **Supp. Mat. selection-genomes.txt**). A smaller ribosomal

protein tree was subsequently computed based on this species selection, using the same method and model as described previously (cf. **Supp.Mat arbre-ribo-364-sp.pdf** & **arbre-racine.tre**).

4 CONSTRUCTION AND SELECTION OF ORTHOLOGOUS GROUPS BASED ON TAXONOMIC REPRESENTATION

To expand our selection of genes beyond ribosomal proteins, we used USEARCH v8 (Edgar, 2010) and OrthoFinder v1.12 (Emms & Kelly, 2015, 2019) on our 305 selected proteomes to generate orthologous gene groups. The process involved the following command: usearch64.8 -quiet -ublast ./Species0.fa -db ./Species0.fa.udb -evalue 1e-5 -accel 1 -threads 1 -blast6out Blast0_0.txt. This resulted in a total of **94,490 orthologous groups (OGs)**, including **66,798 singletons**. To refine this dataset, we applied two successive filters using **classify-ali.pl** from the **Bio::MUST::Core** distribution.

5 IDENTIFICATION AND GENERATION OF ORTHOLOGOUS GROUPS BASED ON GENE COPY NUMBER USING A PHYLOGENETIC PRUNING METHOD

To sort and select genes, we applied a phylogenetic tree pruning method. This approach allowed us to identify single-copy orthologous groups (OGs) suitable for concatenation. The process involved the following steps:

1. Alignment of OGs: We aligned 1,007 OGs using MAFFT (Katoh et al., 2002; Katoh & Standley, 2013)
2. Phylogenetic Inference: Resulting MSAs were analyzed using RAxML v8.1.17 with rapid-bootstrap searches (100 replicates) under the PROTCATLGF substitution model (Stamatakis, 2014) sur les MSAs résultants.
3. Tree Pruning: We used the program root-max-div-taxon (H. Philippe, CNRS) on the 1,007 inferred trees. This program splits phylogenetic trees according to four user-defined parameters:
    - Minimum number of species in clade 1,
    - Minimum number of species in clade 2,
    - Percentage of internal branches longer than the one used for splitting the tree,
    - Minimum number of species shared between the two subtrees.

   We tested three parameter sets for this process:
    - Set 1: 40, 90, 100, 40
    - Set 2: 90, 40, 100, 40
    - Set 3: 90, 90, 100, 50
4. Iterative Pruning: The procedure was repeated as needed, splitting trees to extract additional paralogs. Genes containing multiple paralogs were iteratively divided until all sequences within a group were orthologous.

This approach identified 440 directly orthologous MSAs and 117 newly orthologous MSAs (neo-orthologs).

6 FINAL TAXONOMIC SAMPLING AND ORTHOLOGOUS GROUP SELECTION

To refine our initial taxonomic sampling, we used SCaFoS (Roure et al., 2007) to construct a supermatrix of our 440 genes, yielding a total of 89,342 aligned amino acid (AA) positions (cf. Supplementary Material: supermatrix-440genes.ali). On this raw supermatrix, we applied ali2phylip.pl with the --max=0.1 option, resulting in a supermatrix of 89,342 unambiguously aligned amino acid positions. We then calculated a phylogenetic tree using RAxML (Stamatakis, 2014)with a rapid-bootstrap search (100 replicates) under the PROTCATLGF model.

7 CONGRUENCE TEST, SEQUENCE REMOVAL, AND MSA CONSOLIDATION

After the initial sequence additions, we applied a method to eliminate both poor sequences (paralogs, xenologs) and poor alignments (those where the orthology of numerous sequences was ambiguous). To do this, we compared branch lengths between gene trees and the species tree. This requires the inference of an intermediate species tree to serve as a reference. This tree is obtained by inferring a phylogeny from a supermatrix concatenating the different alignments of the dataset (**Figure 11**). Next, the phylogeny of each alignment used for the concatenation (one sequence per OTU) is inferred by fixing the tree topology to that of the supermatrix and using the same model. Finally, the branches of each gene tree are compared to those of the species tree reduced to the same sampling. With a fixed topology, if a sequence is forced into an incorrect position in the tree (i.e., the expected taxonomic position, which is incorrect due to paralogy or xenology), the sequence evolution model can only explain its position by a significant number of substitution events. As a result, the branch length of the sequence in question is disproportionately increased by the model. Consequently, this type of sequence can be identified by comparing its branch length to that expected from the reference species tree. Moreover, considering all branch lengths simultaneously, the presence of multiple unexpected branch lengths in a gene tree affects the correlation value of branch lengths (Pearson correlation coefficient R^2) between the reference and the gene. A low R^2 value thus also identifies alignments possibly containing many paralogs or displaying a phylogenetic signal contrary to that of the majority of genes (xenology or severe incompatibility of the gene phylogeny with the species phylogeny).

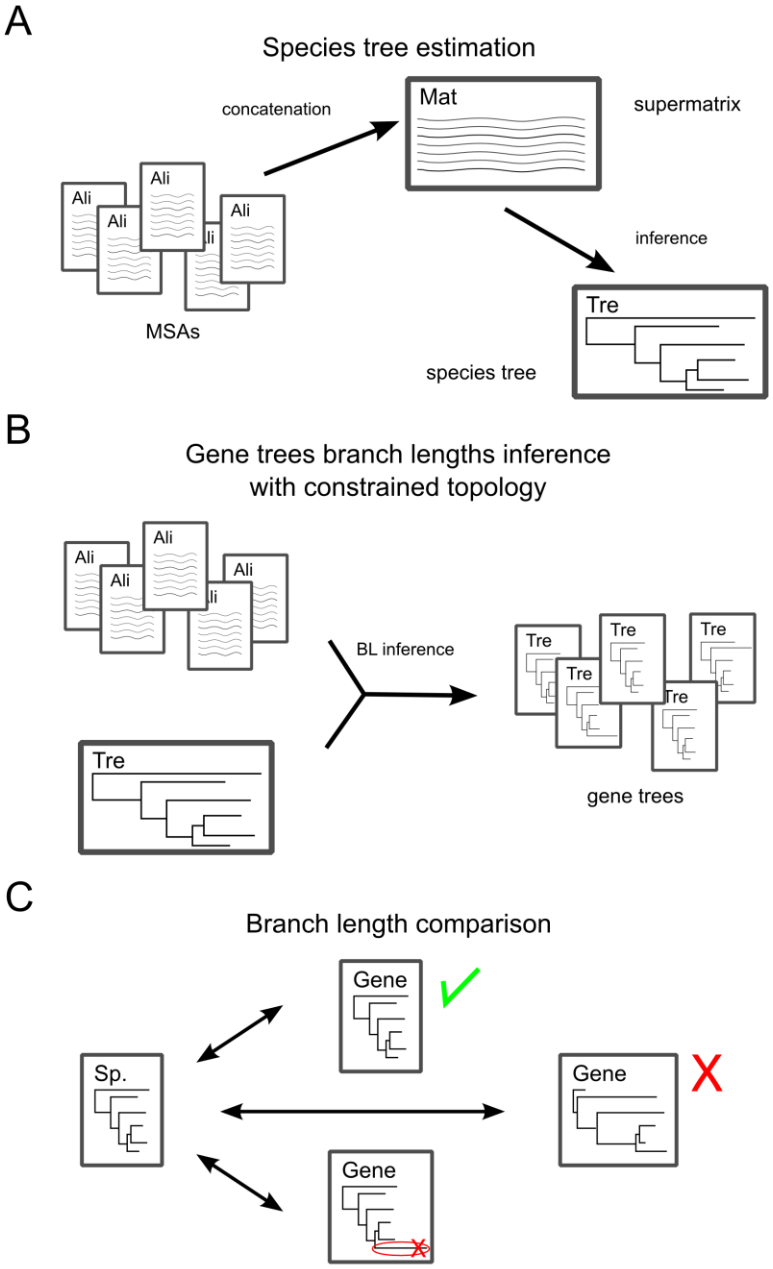


Figure 11. Decontamination protocol by branch length comparison and correlation.
A. Creation of a reference species tree. B. Estimation of branch lengths for gene trees by enforcing the reference topology. C. Comparison of branch lengths between gene trees and the species tree.

Specifically, we masked non-homologous sequence segments using HmmCleaner and removed high-entropy positions with BMGE (Criscuolo & Gribaldo, 2010) before concatenating them with SCaFoS (Roure et al., 2007), which, by default, selects the longest sequence per OTU when residual paralogues are present. We used the minimum evolutionary distance as the selection criterion among paralogous sequences of the same OTU (with a complete elimination threshold at 25% distance between the two paralogues). The maximum percentage of missing sites for a complete sequence was set at 10, and the maximum number of missing OTUs was set at 25. Trees were inferred using the PROTCATLGF model with RAxML (Stamatakis, 2014) and rapid-bootstrap searches (100 replicates). This allowed us to compare the terminal branch lengths observed in these individual gene trees to those in the corresponding reference supermatrix tree, leading to the removal of sequences with branch length ratios exceeding 5 times those in the reference tree. A total of 227 sequences were removed (cf. **Supp.Mat remove_seq_branchlength_9**). We then applied the same protocol by recalculating trees, this time using the Pearson correlation coefficient R^2 (mean – 1.96 x standard deviation = 0.626322348) of branch lengths between individual gene trees and the corresponding supermatrix (cf**. Supp.Mat congruence.csv**). With this method, we reduced the dataset from 440 to 416 genes.

8 CREATION OF CHIMERIC ORGANISMS

We next reassessed the completeness of our 416 genes. Some organisms were insufficiently represented in our MSAs, but among these, several were closely related. We thus created chimeras by in silico merging of complementary gene sequences, resulting in more complete chimeric organisms. This step reduced the number of species from 364 to 352 by merging underrepresented closely related species within our 416 genes.

After completing these MSA cleaning steps, the final supermatrix was constructed by realigning the sequences. We retained alignments containing at least one representative from Asgard, DPANN, Euryarchaeota, and TACK before concatenating them with SCaFoS. This process yielded a final dataset of 343 genes and 352 OTUs with 94,485 positions. This dataset was then used to evaluate the impact of taxonomic sampling and evolutionary models on the inference of the archaeal tree.

9 JACKKNIFE

We performed jackknife tests on both species and genes. To do so, we used our ribosomal tree to define monophyletic groups best representing archaeal diversity (groups-uniq.txt). These groups were defined according to three principles:

- A group must have 100% statistical support;
- A group can have a maximum of 5 species;
- If a group has 5 or more species, it is reduced to 3 species.

To optimize our selection, we used SCaFoS to prioritize species present in the largest number of our OGs (in other words, we retained the most organismally complete genes). We chose to work with 5 species replicates.

10 CONSTRUCTION OF SUPERMATRICES & PHYLOGENETIC INFERENCES

From this point onward, all subsequent steps were conducted in the same manner. The datasets were built as follows:

We systematically aligned all initial alignments using ali2phylip.pl, implementing the BMGE filter with the option --bmge-mask = loose. For this purpose, we tested the optimal value for the --max option within ali2phylip.pl, varying between 0.1, 0.2, 0.3, 0.4, and 0.5 (cf. Supp.Mat a2p-test-max10-50.csv). We observed that the maximum number of columns was removed at --max = 0.1, and beyond that, few additional columns were eliminated. Consequently, we retained --max = 0.1 and set --min = 0.1 to exclude sequences shorter than 10% of the longest sequence. Additionally, we applied a filter for 40 species using Classify-ali.pl to ensure sufficient species were retained per alignment in cases where some species might be lost. If this condition was not met, the alignment was discarded.

We also applied HmmCleaner. From the sequences of target species present in subtrees of interest, we performed two main types of analysis: a congruence analysis of the signal by concatenating MSAs into a supermatrix with SCaFoS (Roure et al., 2007) and a supertree analysis.

For each replicate, several analyses were performed (Box 5) :

- An analysis of 343 genes via supermatrix and supertree methods under the models LG4X, LG+C20+F+G, and LG+C60+F+G (Le et al., 2012; Le & Gascuel, 2008; Si Quang et al., 2008) ;
- A jackknife analysis of genes across 10 replicates for supermatrices under the models LG4X, LG+C20+F+G, and LG+C60+F+G, as well as by the PMSF method using the tree derived from the LG+C60+F+G model as the reference tree;
- A jackknife analysis of genes across 100 replicates for supermatrices calculated via the PMSF method, using the same reference tree as above (LG+C60+F+G), and for supertrees under the models LG4X, LG+C20+F+G, and LG+C60+F+G.

The supermatrix trees were computed with IQ-TREE 1.6.9 (Minh et al., 2020) using Ultrafast-Bootstrap × 1000. In parallel, congruence signal analyses based on supertree creation were conducted with ASTRAL-III (v5.7.5) (Mirarab & Warnow, 2015), which generated a supertree for each pipeline based on single-gene phylogenetic trees produced by IQ-TREE (LG4X, Ultrafast bootstrap) (Simion et al., 2017). Normalized Robinson-Foulds distances were also calculated using IQ-TREE with the -rf_all option, allowing pairwise distance computation for all trees.

The determination of majority partitions across all resulting trees was facilitated by the program parse_consense_out.pl (Supp.Mat bilan-parse_consens_out.csv). For this, we identified species groups expected to be recovered based on our ribosomal results and the literature (Supp.Mat groupes.otu). The selected trees were automatically formatted using format-tree.pl (found in Bio::MUST::Core) and visualized in iTOL (Letunic & Bork, 2021).

Box 5. Phylogenetic models

In order to represent the evolutionary processes of molecules as faithfully as possible, numerous models, relaxing certain hypotheses of a base model, have been proposed. The relevance of these models is typically evaluated through the likelihood gain they provide, penalized by their degrees of freedom, i.e., the number of additional parameters they introduce. This is the underlying principle of likelihood ratio tests (LRTs; Felsenstein, 1981). Among these sophisticated models, we can cite:

- Site-specific models that take into account the fact that the rate of evolution, or even the nature of the bases or amino acids used, varies from one position to another within the molecule;
- Non-stationary models that account for variations in base composition between species and estimate ancestral base compositions;
- DNA codon models which are particularly useful for detecting genes/sites/lineages under positive selection, characterized by a higher non-synonymous substitution rate than synonymous substitutions;
- Models accounting for differences in evolutionary rates between lineages, particularly used to estimate divergence dates based on fossil calibrations – referred to as deviation from the molecular clock hypothesis;
- Models representing variations in the evolutionary process followed by a site over time between lineages (or "covarion"), which allow detection of changes in functional constraint over the course of a molecule’s history.

Homogeneous Models: The GTR Model

Homogeneous models of sequence evolution differ based on two main components: the exchangeability matrix R and the equilibrium amino acid frequency vector Π. The exchangeability matrix R describes the relative rates at which one amino acid changes to another. In homogeneous models, it is assumed that substitution rates are uniform across all sites in a sequence. This means that every pair of nucleotides (in the case of DNA or RNA sequences) or every pair of amino acids (in the case of protein sequences) has the same substitution rate. In heterogeneous models, it is assumed that substitution rates can vary across different sites of a sequence. This means that some sites may evolve faster or slower than others. Heterogeneous models better capture the complexity of molecular evolution by accounting for variations in substitution rates between sites. They are often used when it is known or suspected that evolution does not occur uniformly across the sequence.

The GTR (General Time Reversible) model, the most general model under the assumption of homogeneous substitution rates between sites, corresponds to a matrix where all parameters, relative exchange rates, and stationary probabilities are considered unknown and are thus inferred from the data (Lartillot and Philippe, 2006). It is a general substitution model that allows substitution rates specific to each nucleotide or amino acid pair. The GTR model uses a single set of equilibrium frequencies (π) for all sites in the sequence. These equilibrium frequencies represent the relative proportions of amino acids or nucleotides in the entire set of sequences studied and are constant across all sites in the base model. GTR is a flexible model because it allows different substitution rates for each nucleotide (or amino acid) pair. For example, the substitution rates from A to C, A to G, etc., are all modeled separately.

Difference between Partition Model and Mixture Model

Mixture models, like partition models, allow more than one substitution model along sequences. However, whereas a partition model assigns a specific model to each site in the alignment, mixture models do not require this information. A mixture model calculates the probability (or weight) of each site belonging to each class of the mixture (also called categories or components). Since the site-class assignment is unknown, the likelihood of the site in mixture models is the weighted sum of the likelihoods of the site by mixture class.

For instance, the discrete rate heterogeneity model Gamma is a simple mixture model. It contains several rate categories with equal weight. IQ-TREE also supports a number of predefined protein mixture models such as the C10 to C60 profile mixture models (ML variants of the Bayesian CAT models). Among these, the following models are used in this thesis:

- (C20, C60): Mixture models with 10, 20, 30, 40, 50, 60 profiles [Le et al., 2008a] as variants of the CAT model [Lartillot and Philippe, 2004] for ML. These models assume a Poisson AA replacement and implicitly include Gamma rate heterogeneity between sites.
- LG4X: Model with four merged matrices with FreeRate heterogeneity [Le et al., 2012].
- LG+C20: Application of the LG matrix instead of Poisson for the 20 AA profile classes and Gamma rate heterogeneity.
- LG+C20+F: Application of the LG matrix for the 20 classes plus the 21st empirical AA profile class (evaluated from actual data) and Gamma rate heterogeneity.
- +F: Empirical amino acid frequencies evaluated from the data.
- +G: Discrete Gamma model ([Yang, 1994]) with 4 rate categories by default. The number of categories can be changed, e.g., +G8.

Bayesian Approaches: PhyloBayes vs IQ-TREE

The ML approach seeks to estimate the model parameters that maximize the likelihood of the observed data. In this approach, parameters are considered fixed values, and the best estimate is the one that makes the observed data most probable. It provides a point estimate of the model parameters but does not directly give information about the uncertainty associated with those estimates. The ML approach does not use priors on the model parameters. Parameter estimates are based solely on the probability of the observed data. In contrast, the Bayesian approach treats the model parameters as random variables and estimates them by calculating their posterior distribution conditioned on the observed data. This approach incorporates both prior information on the parameters (priors) and the observed data to estimate the distribution of the parameters. It provides a full estimate of the uncertainty of the model parameters in the form of posterior distributions. This allows uncertainty to be taken into account in predictions and decisions based on parameter estimates.

Among the Bayesian approaches using heterogeneous models, the CAT model (for Categories of Substitution) (Lartillot and Philippe, 2004), implemented in the PhyloBayes software, is designed to capture the heterogeneity of equilibrium amino acid (or nucleotide) frequencies across sites (Lartillot et al., 2013). Each site in the alignment is modeled using a set of equilibrium frequency profiles (π). A profile π is a vector describing the equilibrium frequencies of the different amino acids (or nucleotides) at that site. There are several profiles (or π vectors), and each site in the alignment is probabilistically assigned to one of these profiles. This allows modeling variations in selective pressures across different sites in the sequence. One of the main differences between amino acid substitution models (e.g., C60) and heterogeneous CAT models at the site level is the number of equilibrium frequency categories. The standard CAT model uses a Dirichlet process before inferring the number of equilibrium frequency categories, so the number of categories is variable (Lartillot and Philippe, 2004, 2006). This model operates like a mixture model, where the number of classes, rather than being fixed a priori, is one of the parameters of the problem. Through Bayesian treatment, the number of classes, as well as all other parameters of the model, are estimated by sampling, using the Markov Chain Monte Carlo (MCMC) principle. The underlying idea behind MCMC is that a Markov chain, taking the form of a guided walk through the multidimensional parameter space, can be used to estimate a probability distribution by periodically sampling the values of these parameters. The approximation of the distribution will be more accurate the higher the number of steps taken by the Markov chain. Conversely, IQ-TREE implements heterogeneous amino acid substitution models (Si Quang et al., 2008) with a fixed number of categories, which can range from 10 (C10) to 60 (C60). When running IQ-TREE with a category model, the program initializes several substitution profiles. IQ-TREE then estimates which profile is most appropriate for each site based on the alignment data. Each site in the sequence alignment is assigned to a particular profile based on its probability of matching that profile. This assignment is done iteratively and probabilistically, using algorithms like Markov Chain Monte Carlo (MCMC) or maximum likelihood techniques. By using multiple profiles, IQ-TREE can more accurately model the heterogeneity of substitution rates across sites. This allows for more robust and realistic phylogenetic estimates, particularly for protein sequences where substitution rates can vary considerably from one site to another.

PMSF

IQ-TREE provides a new "Posterior Mean Site Frequency" (PMSF) model as a rapid approximation of profile mixture models such as C10 to C60, which are computationally intensive and require significant memory (Le et al., 2008a). PMSF can be seen as a variant of the CAT model in PhyloBayes. PMSF represents amino acid profiles for each site of the alignment, calculated using an input mixture model and a guide tree. The PMSF model is much faster and requires significantly less RAM than the C10 to C60 models (cf. table below), regardless of the number of mixture classes. Moreover, extensive simulations and empirical analyses of phylogenomic data demonstrate that PMSF models can effectively mitigate long-branch attraction artifacts.

For novel phylogenomic datasets lacking well-accepted phylogenies, an ML tree can first be estimated under LG+F+Γ, which is quick to compute. This tree can then be used as a guide tree to fit a model such as LG+C20+F+Γ or (even better, LG+C60+F+Γ) to derive PMSF profiles. Subsequently, tree searching can be completed under LG+PMSF+Γ to obtain an ML tree, which is relatively faster than estimating a tree under the full LG+C20+F+Γ mixture. Guide trees can also be iteratively updated. As mentioned, PMSF profiles derived from the LG+F+Γ guide tree can yield an initial LG+PMSF+Γ tree. This LG+PMSF+Γ tree can then serve as a new guide tree, leading to updated PMSF profiles and, consequently, a new LG+PMSF+Γ tree. The process can be repeated a predetermined number of times or until no further topological changes occur. As the guide tree becomes more accurate with each iteration, the PMSF profiles are also expected to approach the true profiles, resulting in more accurate phylogenies.

PMSF refers to both an estimation method and a model that allows amino acid frequencies to vary by site, not necessarily in a manner consistent with a finite mixture. While mixtures with a finite number of classes are computationally necessary, the unique structural and functional constraints of protein sites suggest that frequency vectors are better modeled as a continuous variation. Posterior means tend to exhibit continuous variation between sites, a behavior demonstrated for posterior mean rate estimates calculated within a finite mixture by Susko et al. (2003).

The PMSF method performs as well as, if not better than, empirical mixture models like C20+F and C60+F for estimating phylogenies in the presence of site heterogeneity, in both simulations and empirical studies. Furthermore, it is much more computationally efficient. Indeed, one advantage of the PMSF method is that it allows for continuous variation of frequency vectors across sites.

References

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.

We also analyzed our supermatrices previously examined with IQ-TREE using Bayesian models (Lartillot & Philippe, 2004, 2006). Bayesian inference (BI) analyses were performed using PhyloBayes MPI 1.9a (Lartillot et al., 2013). The CAT-G and CAT-GTR-G models were employed to account for both site-specific substitution rate heterogeneity (modeled by the gamma distribution, which partitions sites into different substitution rate categories, thereby reflecting that some sites evolve faster or slower than others) and amino acid equilibrium frequency heterogeneity (modeled by the multiple equilibrium frequency vectors, or profiles, of the CAT model). Due to their computational intensity (using 40 or 50 CPU cores per chain, depending on the model), BI analyses were conducted on the supermatrices corresponding to replica 2 analyzed with the PMSF method, both before and after the addition of eukaryotes. All BI analyses were performed with four parallel chains, each run for 1000 cycles (approximately 12 or 36 days of computation per chain, depending on the model). A manual examination of the ".trace" files produced by PhyloBayes allowed us to exclude the first 200 or 500 trees of each chain as "burnin," depending on the model. For each analysis, a consensus tree was calculated per chain and across the four chains, each time using a 25% majority rule and sampling 100 of the remaining trees from each chain. Regardless of the model, the maxdiff values calculated by PhyloBayes (bpcomp) across the four chains were equal to 1, indicating that the chains had not converged according to this strict measure. Nevertheless, the meandiff values were 0.06 and 0.05, respectively. Additionally, the parameter samples from each chain individually were very similar, albeit with a few key topological differences.

11 SLOW-FAST

The stepwise elimination of the fastest-evolving sites was performed using the Slow-Fast method (Brinkmann & Philippe, 1999). At each step, a phylogenetic tree was computed with IQ-TREE (Minh et al., 2020) using the LG4X model, allowing us to monitor the impact of the removed positions on key branches. This approach requires the definition of reliable groups to estimate the evolutionary rate of each site. Groups corresponding to indisputable monophyletic classes or phyla were considered. For the Archaea, we defined four groups: DPANN, Euryarchaeota, Asgard, and TACK. When including eukaryotes, the Archaea and eukaryotes were separated into two groups.

12 ROOTING ARCHAEA

Box 6. Rooting methods

When estimating phylogenetic trees, unrooted trees are often used to calculate the likelihood of the data. Once the optimal unrooted tree is determined, additional methods can be employed to identify where the root should be placed on the tree: rooting via molecular clocks, branch length distribution (midpoint rooting, minimal ancestor deviation, minimum variance rooting), gene duplications, indels (insertions-deletions), unrooted gene tree distributions, probabilistic co-estimation of gene and species trees. However, the most widely used method for rooting phylogenetic trees is the outgroup method. Although using an outgroup to root an unrooted phylogeny is generally more accurate than other rooting methods, the primary challenge lies in identifying an appropriate outgroup. Outgroups that are too distant from the ingroup of interest may exhibit molecular evolution significantly different from that of the ingroup, potentially compromising the method's accuracy.

Felsenstein's pulley principle is a key concept in phylogenetic tree reconstruction, particularly concerning the probabilities of rooted and unrooted trees. This principle states that the root position in a phylogenetic tree does not affect the probabilities of branch length distributions when considering relationships among terminal species. In other words, the substitution path from one species to another remains the same regardless of the root position, allowing analyses to begin with unrooted trees. Using unrooted trees simplifies the exploration of the possible tree space and accelerates calculations, as there is no need to initially determine the root's exact placement. However, this also means that ultrametric constraints (as in a dated chronogram) cannot be imposed because the direction of time is undefined in an unrooted tree. This principle is especially relevant to reversible models, where transition probabilities are symmetrical.

The only existing approach to include root placement within probabilistic inference is the application of non-reversible models. Using non-reversible substitution models relaxes the fundamental assumption of temporal reversibility present in most widely used phylogenetic inference models. This assumption posits that the evolutionary process of DNA sequences is reversible over time, meaning that the mutation probabilities between nucleotides are identical in both directions. This reversibility simplifies substitution models and facilitates phylogenetic tree inference. However, it is important to recognize that this assumption is a simplification and may not always align with biological reality. Factors such as selective pressure, mutational biases, and error correction mechanisms can lead to asymmetric substitution rates. The assumption of temporal reversibility similarly applies to protein sequence-based phylogenies, where amino acid substitution models account for replacement rates between different amino acids instead of nucleotide bases.

References

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).

We conducted our rooting analyses (**Box 6**) within a maximum likelihood (ML) framework using IQ-TREE (Minh et al., 2020). One advantage of ML compared to Bayesian analysis is that it does not require prior assumptions about parameter distributions, which can sometimes affect tree inference (Cherlin et al., 2018; Huelsenbeck et al., 2002). While estimating parameters of the non-reversible model through likelihood maximization may seem computationally more intensive than calculating posterior probabilities (Huelsenbeck et al., 2002), IQ-TREE's algorithm is fast enough to estimate root placements for very large datasets. Thus, we assessed the rooting of archaea using rootstrap support values and the AU-test, which estimate the extent to which the data support each branch as a possible root of a phylogenetic tree. The rootstrap analysis uses resampling techniques to evaluate the robustness of different root positions, while the AU-test statistically compares the likelihoods of different rooting hypotheses to identify the root position most compatible with the data. These analyses were performed on our species replica 2, with and without eukaryotes.

We calculated the confidence set for all branches likely to contain the root of our IQ-TREE PMSF tree using an AU-test (Approximately Unbiased test) (Shimodaira, 2002). The AU-test is a statistical method for comparing phylogenetic trees and evaluating their support by the data. In the context of tree rooting, the AU-test can compare rooted trees with different root positions. The unrooted trees obtained during the PMSF analyses conducted previously served as guide topologies to steer the root search. IQ-TREE tests all possible positions on the provided topology. For each tree, its likelihood (the probability of the data given the tree) is calculated, which involves measuring the extent to which the data support each tree. The AU-test compares the likelihoods of different trees, accounting for variations due to data resampling. It uses resampling methods such as bootstrapping to evaluate the distribution of likelihoods and calculate a p-value for each tree. The p-values indicate whether a topology is significantly rejected by the data. If the p-value is less than 0.05, the root is not on the tested branch. To do this, we reset our ML tree with all possible root positions (one position for each branch) and calculated the likelihood of each tree. Using the AU-test, we then determined which root placements could be rejected, using an alpha threshold of 5%. We defined the root branch confidence set as the set of branches that are not rejected in favor of the ML root placement.

We also performed analyses using rootstrap. The rootstrap value measures the robustness of root placement, given the model and the data. It evaluates the stability and confidence in different hypotheses about the root position using bootstrap resampling of the data. Rootstrap analysis is a method specific to IQ-TREE for testing root positions on a phylogenetic tree. IQ-TREE generates a large number of bootstrap trees (resampled) from the data. Each bootstrap tree is a phylogenetic reconstruction based on a subset of the original data, capturing variability in tree inference. This allows us to see how the data support each possible root position. The results show the frequency with which each root position is supported by the bootstrap trees. This provides a measure of robustness and confidence in the various root positions. To calculate rootstrap supports, we performed a bootstrap analysis, resampling sites from the supermatrix with replacement to obtain a certain number of bootstrap replicate trees. The rootstrap support for each branch of the ML tree is defined as the proportion of bootstrap trees where the root is located on that branch. Since the root can be on any branch of an unrooted tree, root support values are calculated for all branches, including external branches. The sum of root support values along the tree is always less than or equal to 1. A sum less than 1 can occur when one or more bootstrap trees are rooted on a branch that does not appear in the tree. Indeed, some bootstrap replicate trees may have a topology different from that of the reference tree. If one or more bootstrap replicates have their root on a branch that does not exist in the reference tree, then these replicates do not contribute to the rootstrap support values for the branches of the reference tree, which explains why the sum may be less than 1.

An important difference between the AU-test and rootstrap support is that the AU-test is conditioned on a single maximum likelihood (ML) tree topology, whereas rootstrap support is not. Because of this, they provide very different information about root position. The AU-test assumes the ML tree topology is correct and then seeks to determine the confidence set of root positions conditioned on this topology. Rootstrap, on the other hand, does not assume a particular topology and evaluates how often a particular root position appears in a set of bootstrap replicates.

CHAPTER 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

- 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.

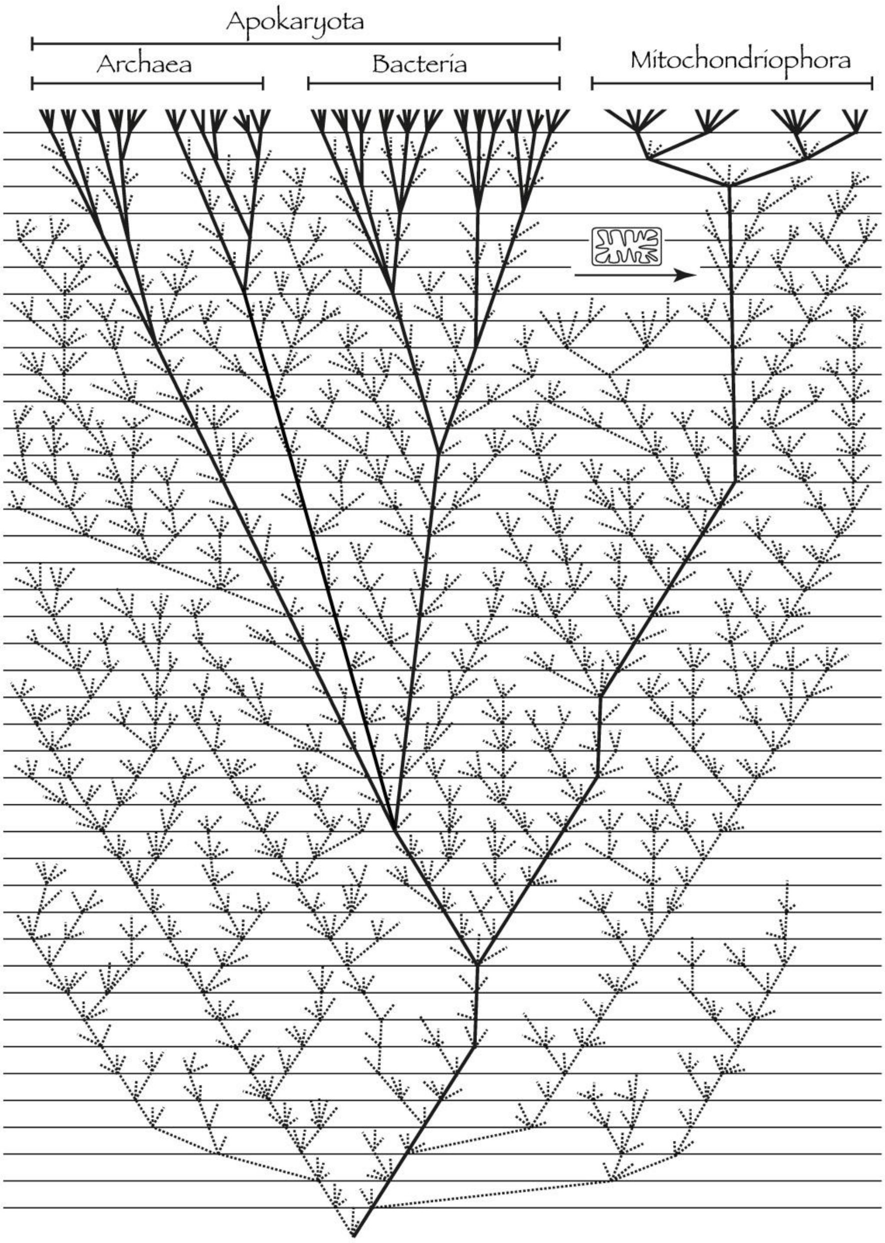

Apokaryota
Archaea
Bacteria
Mitochondriophora

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.

	model	LG	GTR	CAT	CATGTR
outgroup	topology				
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		62	54	100	100
Bivalvia					
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

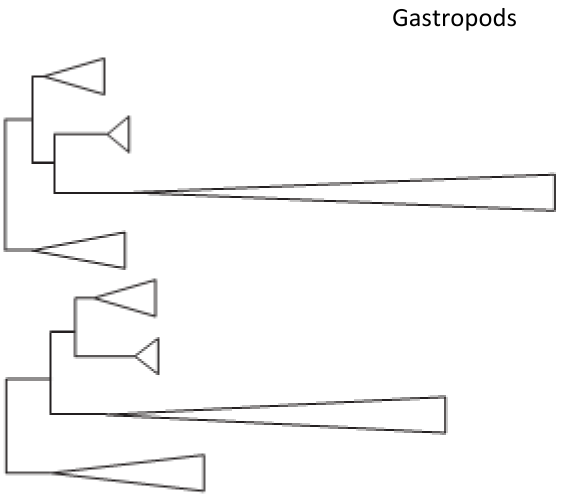


Figure 13.

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!

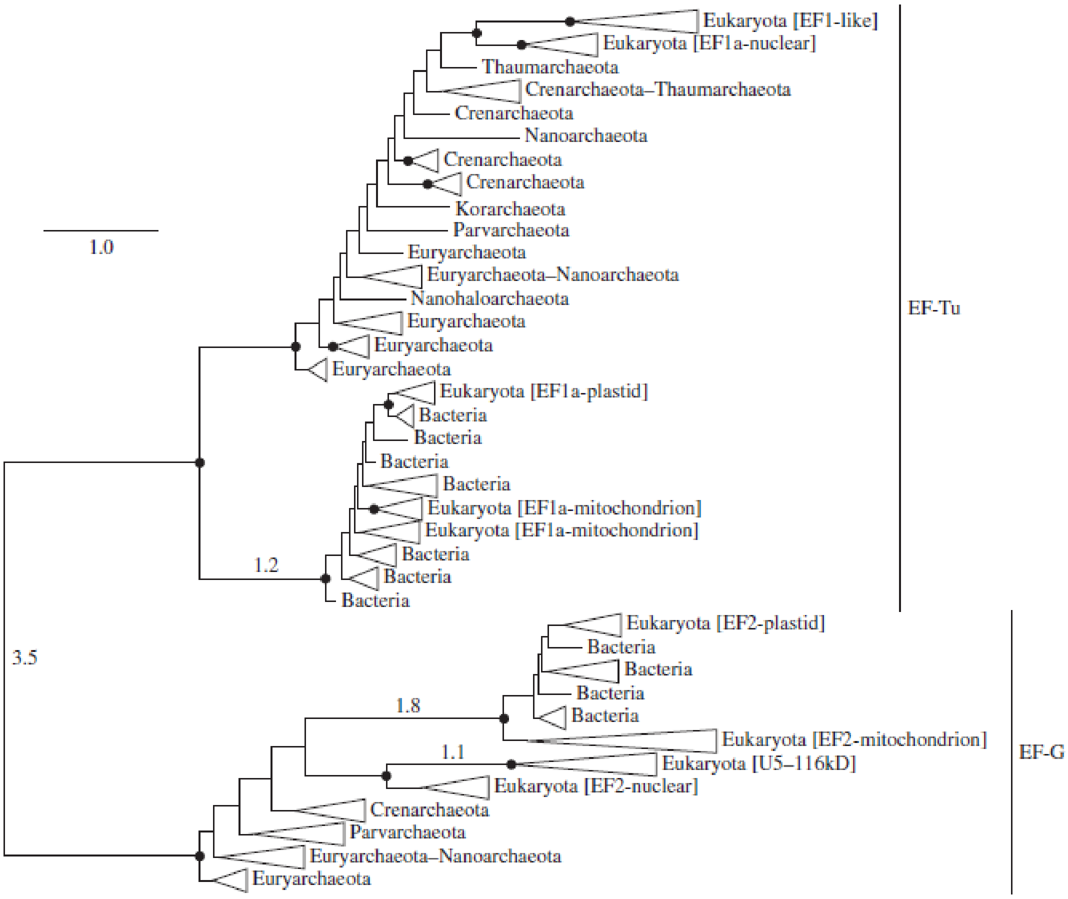


Figure 14.

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 16S 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)

UPDATE

11 COMPLEX RELATIONSHIPS BETWEEN ARCHAEA AND EUKARYOTES

11.1 DISCOVERY OF THE ASGARD GROUP

In 2015, Thijs Ettema’s team from Uppsala University in Sweden reconstructed a genome from the δ clade of the Deep Sea Archaeal Group (DSAG), which they claimed bridges the gap between prokaryotes and eukaryotes (Spang et al., 2015). Metagenomic studies indicate that they emerge within the TACK superphylum and are closely related to eukaryotes.

Named "Lokiarchaeota", they were collected at a depth of 3,283 meters in marine sediments of the Arctic Ocean between Greenland and Norway, near the hydrothermal vent zone of the Gakkel Ridge, close to the site known as Loki's Castle (hence their name), discovered in July 2008. They are represented by one complete genome (5.1 Mbp encoding 5,381 protein-coding genes) and two nearly complete genomes named Loki2 and Loki3.

In 2017, the same authors published a second article describing the MAGs of several new archaeal lineages related to the Lokiarchaeota, which they also named after Scandinavian figures: Thorarchaeota, Odinarchaeota, and Heimdallarchaeota (Eme et al., 2017; Spang et al., 2017). These new lineages have been found in a wide variety of environments: hot springs, deep oxygen-poor sediments, coastal sediments, and oceanic plankton (Macleod et al., 2019).

As these new genomes are named after gods from Scandinavian mythology, they were grouped into a superphylum named Asgard, after the dwelling of these gods. Since then, new samples from various regions around the globe (Loki's Castle, Yellowstone National Park, Aarhus Bay in Denmark, an aquifer near the Colorado River, Radiata Pool in New Zealand, hydrothermal vents near the Taketomi Islands in Japan, and the White Oak River estuary) have revealed various strains of this superphylum, such as Helarchaeota and Gerdarchaeota.

It has been suggested that the close relationship between Asgards and eukaryotes might result from possible contamination during metagenome assembly, combined with phylogenetic reconstruction artifacts due to gene selection and the inclusion of rapidly evolving archaeal groups (Da Cunha et al., 2017). However, since the discovery of new Asgard metagenomes, the low likelihood that contamination and/or homologous recombination with eukaryotic sequences occurred in the same way across all Asgard archaea—whose genomes were assembled from diverse sources and using different methodologies—undermines the hypothesis that the close relationship between eukaryotes and Asgards is the result of contamination with eukaryotic sequences.

Additionally, in 2019, an organism named *Candidatus Prometheoarchaeum syntrophicum* (strain MK-D1), related to the Lokiarchaeota, became the only member of this group to be successfully cultivated and fully sequenced (Imachi et al., 2020). Four more years of effort were required to characterize this Loki archaeon, as it divides very slowly: it takes 60 days to initiate growth, with a doubling time of approximately 14–25 days.

The researchers were unable to isolate it in pure culture but managed to co-culture it with a sulfate-reducing bacterium from the *Desulfovibrio* genus and a methane-producing archaeon from the *Methanogenium* genus. The Loki archaeon produces hydrogen, which is used by the bacterium to produce hydrogen sulfide and by *Methanogenium* to produce methane. Although the bacterium was eventually eliminated from the culture system, it appears that Lokiarchaeota cannot survive on their own and depend on a close association with the *Methanogenium* archaeon. In return, *Methanogenium* provides Loki with numerous components, including amino acids and vitamins, which MK-D1 is incapable of producing on its own.

According to the study's authors, all Asgard archaea might similarly rely on symbiotic associations for their development. This could explain the difficulty in cultivating them and the absence in their genomes of several critical metabolic pathways, particularly those involved in amino acid biosynthesis.

A second Asgard species, harvested from the mud of an estuary in Slovenia, was also cultivated after 7 years of effort (Rodrigues-Oliveira et al., 2023). Named *Lokiarchaeum ossiferum*, this organism features several dozen fine tentacles with thickening and growths, along with a cytoskeleton extending into the tentacles. It is an anaerobic organism found in deep marine environments with low oxygen levels, such as hydrothermal sediments. It appears capable of chemiosynthesis, obtaining energy from the reduction of compounds like sulfur or methane, which is typical of archaea in these extreme environments. Its genome has been fully sequenced. It contains genes encoding proteins similar to those found in eukaryotes, including cytoskeletal components, proteins involved in membrane trafficking, and other complex processes previously thought to be exclusive to eukaryotic cells. This suggests that this common ancestor of eukaryotes may have had a greater cellular complexity than previously imagined.

The genome of *Lokiarchaeum ossiferum* reveals the ability to perform metabolic processes typical of archaea, but with a notable increase in the repertoire of genes associated with eukaryotic functions. For example, it contains genes for E2 ubiquitin domain proteins, proteins involved in membrane remodeling, and GTPases associated with the cytoskeleton. These characteristics suggest an enhanced capacity for interacting with intracellular membranes and forming compartments, a key step in the evolution toward more complex cells.

Lokiarchaeota is a small (~550 nm), anaerobic organism that degrades amino acids and peptides through syntrophy with the archaeon *Methanogenium* and/or the bacterium *Halodesulfovibrio*. The culture of this strain confirms the presence of 80 eukaryotic signature proteins (ESPs), also found in other Asgards. Additionally, it produces "arms" or "tentacles" that allow it to attach to other microorganisms and potentially "consume" another.

Based on the characteristics identified in the new MK-D1 archaeon and assuming it could be a possible host for the endosymbiotic theory, the authors of the publication have established a new model called E3 for Entanglement – Engulfment – Endogenization (= internalization).

The authors then proposed a narrative of events that may have led an archaeon related to MK-D1 to "swallow" another organism, potentially explaining the origin of eukaryotes. According to them, several steps would have been necessary (**Figure 15**) :

- The first step is a transition from an anaerobic lifestyle to one that tolerates the presence of oxygen. A dependency relationship would have developed between the two organisms: the archaeon, the future host, and the bacterium, the future endosymbiont (**Figure 15 (a)**).
- Next, the host's external structures ("arms or tentacles") would have interacted with the bacterium, improving the physical interaction and allowing it to "engulf" the bacterium. One could say the host "swallowed" the bacterium without "digesting" it (**Figure 15 (b)**).
- After this ingestion, the host would have shared molecules as an energy source with the symbiotic bacterium. In return, the bacterium would have consumed the surrounding oxygen and provided energy back to the host in the form of ATP (**Figure 15 (c)**).
- The symbiosis between the two organisms would have evolved over time. Eventually, the host would have delegated ATP production to the bacterium. A channel, called AAC, which allows ATP to pass between the two cells, would have formed (**Figure 15 (d)**). This channel is now essential for the mitochondria in modern eukaryotic cells.
- Finally, over the course of evolution, internal structural modifications would have occurred, giving rise to present-day mitochondria. It is also possible that other endosymbioses took place in a series to lead to eukaryotes as we know them today.

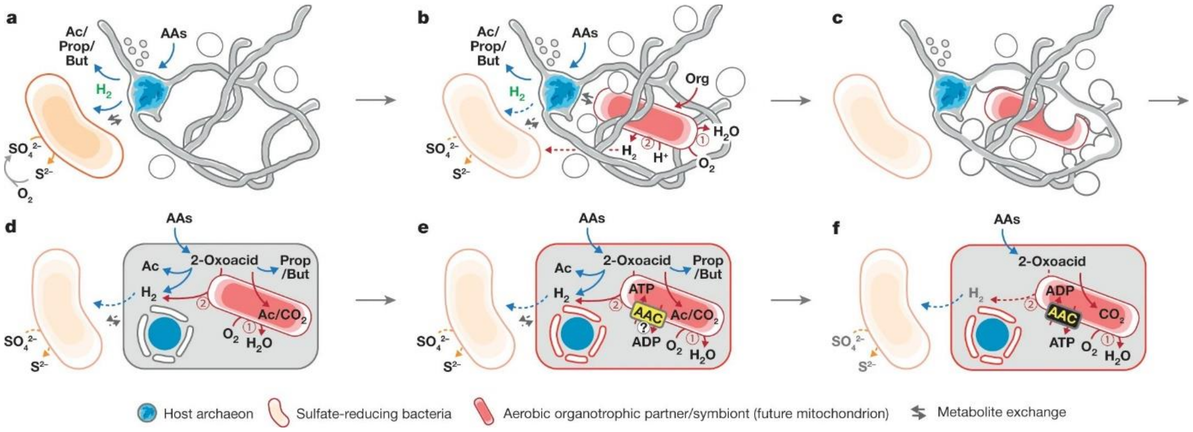


Figure 15. Hypothetical model explaining the emergence of eukaryotes (Imachi et al., 2020).
a) Interaction between the host archaeon and the symbiotic bacterium. The symbiotic bacterium uses the oxygen (O2) from the environment in exchange for nutrients produced by the host. **b)** The host's "arms" interact with the symbiotic bacterium, improving the physical interaction between the two organisms and leading to the engulfment of the bacterium. **c)** The interaction continues after engulfment with the development of a channel (AAC) allowing the transfer of energy molecules (ATP). **d)** Internalization and delegation by the host of energy molecule production to the symbiont.

The number of representatives from the Asgard phylum has been steadily growing since 2015, with a total of 18 lineages to date (Da Cunha, Gaïa, et al., 2022) (**Figure 16**).

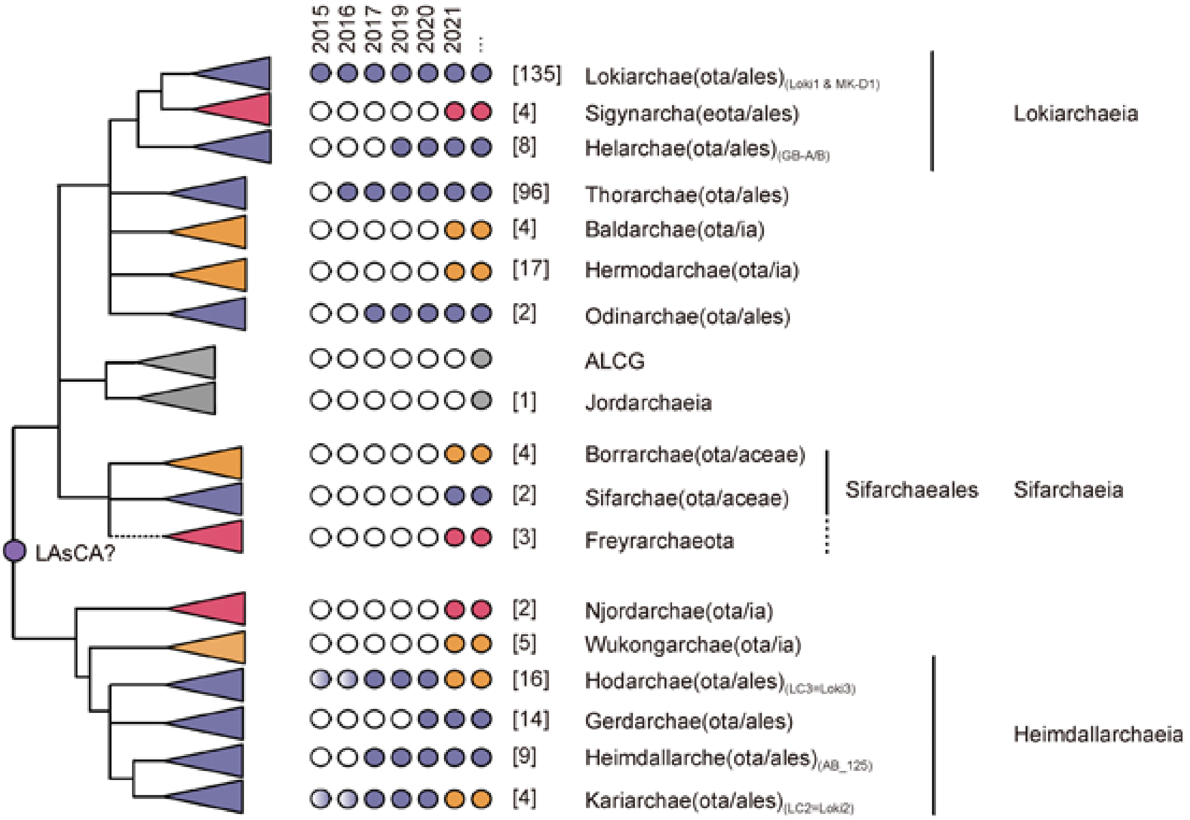

2015
2016
2017
2019
2020
2021
...
[135] Lokiarchae(ota/ales)(Loki1 & MK-D1)
[4] Sigynarcha(eota/ales)
[8] Helarchae(ota/ales)(GB-A/B)
Lokiarchaeia
[96] Thorarchae(ota/ales)
[4] Baldarchae(ota/ia)
[17] Hermodarchae(ota/ia)
[2] Odinarchae(ota/ales)
ALCG
[1] Jordarchaeia
[4] Borrarchae(ota/aceae)
[2] Sifarchae(ota/aceae)
[3] Freyrarchaeota
Sifarchaeales
Sifarchaeia
LAsCA?
[2] Njordarchae(ota/ia)
[5] Wukongarchae(ota/ia)
[16] Hodarchae(ota/ales)(LC3=Loki3)
[14] Gerdarchae(ota/ales)
[9] Heimdallarche(ota/ales)(AB_125)
[4] Kariarchae(ota/ales)(LC2=Loki2)
Heimdallarchaeia

Figure 16. A schematic representation of the known diversity of Asgard archaea, including the recently published lineages. (Da Cunha, Gaïa, et al., 2022).

In this diagram, orange and red colors represent newly discovered lineages introduced in the publications by (Liu et al., 2021a) and (Xie et al., 2022) respectively. Grey represents the new lineages not included in these publications, such as Sifarchaea (F et al., 2021), Jordarchaeia (Sun et al., 2021) and Asgard Lake Cootharaba (Sun et al., 2021). The schematic was designed by combining the phylogenetic trees from all these publications. The suffix in parentheses indicates whether the lineage is considered a Phylum (ota), a Family (aceae), or an Order (ales), according to the authors. The position of the Last Asgard Archaeal Common Ancestor (LAsCA) is shown based on the root observed in most phylogenetic trees, as discussed by (Liu et al., 2021a) which relates to the distribution of Eukaryotic Signature Proteins (ESPs). The year of publication of the first genome for each Asgard lineage is indicated, and changes in taxonomy are reflected by color changes in the points. Light purple points show that Asgards from the Karia and Hodar lineages were described before 2017 as Loki 2 and Loki 3, respectively (Spang et al., 2015). Additionally, the number of genomes available for each lineage is indicated in square brackets.

11.2 CHARACTERISTICS OF THE ASGARD GROUP

These new metagenomes have greatly challenged the three-domain tree of life model, in favor of a two-domain model (the Eocyte hypothesis). Indeed, phylogenetic analyses based on 36 universal proteins tend to place eukaryotes within the Asgard archaea (Spang et al., 2015). Moreover, these new genomes share numerous features previously thought to be exclusive to eukaryotes, suggesting that eukaryotes could have originated from them. For instance, a number of genes coding for eukaryotic-like proteins that had never been identified in archaea have been discovered in Asgard genomes. Asgard archaea possess Eukaryotic Signature Proteins (ESPs), including proteins involved in membrane trafficking mechanisms, eukaryotic structural proteins (cytoskeleton), proteins involved in the ubiquitination modification system, and eukaryotic ribosomal proteins. In particular, Asgards possess actin proteins related to those found in eukaryotes. Additionally, in Odinarchaeota, two genes are homologous to FtsZ, and one gene (OdinTubulin) shows greater homology to eukaryotic tubulin than to any other archaeal tubulin (Akıl et al., 2022). These two proteins are major components of the eukaryotic cytoskeleton. The cytoskeleton allows cells to move and deform to engulf other cells (phagocytosis), which could have been a key advantage in the emergence of the eukaryotic cell. Several authors have suggested that such a cytoskeleton could have appeared in Asgard archaea, enabling one of them to engulf the bacterium that gave rise to mitochondria, thus initiating the process that led to the appearance of eukaryotes. While Asgard archaea are not our direct ancestors, the presence of many eukaryotic-like proteins in these organisms suggests that they shared the same environment as proto-eukaryotes for a long time. Although Asgard archaea prefer to live in association with other organisms, it is even possible that they once lived in symbiosis with our ancestors.

The presence of eukaryote-like systems within the Asgard phylum suggests that eukaryotes inherited simple variations of the cellular machinery from an archaeal ancestor. However, several of these putatively eukaryotic systems coded by Asgards are incomplete or still need to be functionally characterized. Since they have functions that are still unknown in archaea, it is difficult to deduce the mechanism by which eukaryotic systems developed within archaea.

It has been suggested that horizontal gene transfers between ancestral Asgards and proto-eukaryotes may have resulted in the highly diversified topologies of phylogenetic trees for certain Asgard ESPs and universal marker proteins. This hypothesis is relevant regardless of the scenario considered for eukaryogenesis. It implies that Asgards were already diversified before the last common ancestor of eukaryotes and that they shared the same biotopes with proto-eukaryotes. The ESPs found in Asgards and universal proteins may have simply been recruited by a proto-eukaryote, which would explain the scattered distribution of ESPs and the atypical placement of certain Asgard lineages in universal trees based on specific genes. It is possible that some Asgards could still live in symbiosis with modern eukaryotes today.

These genomes contain many genes encoding eukaryotic signature proteins, which strongly link them to eukaryotes (**Figure 17** and **Figure 18**). According to the endosymbiotic model, the ancestral host already possessed some key components that governed the emergence of eukaryotic cellular complexity after endosymbiosis. Among these genes, we can mention (Eme et al., 2017; Spang et al., 2017) :

- The gene encoding the eL22 protein of the large ribosomal subunit. However, the gene encoding the eL41 protein is absent.
- Boxes coding for domains homologous to gelsolin, a protein that dismantles and caps actin filaments in the presence of calcium ions, enabling the formation of exocytosis vesicles.
- Several genes encoding the ESCRT (Endosomal Sorting Complexes Required for Transport) complex and several proteins homologous to components of the eukaryotic

multivesicular body endosome pathway. It is noteworthy that the ESCRT complex is already present in archaea in various forms (Frohn Béla P. AND Härtel 2022).

- Various subfamilies of the Ras superfamily of small GTPases.
- Proteins from the BAR/IMD superfamily involved in cellular trafficking and membrane remodeling.
- MK-D1 simultaneously expresses three systems capable of participating in cell division: FtsZ, actin, and the ESCRT-II/III complex.

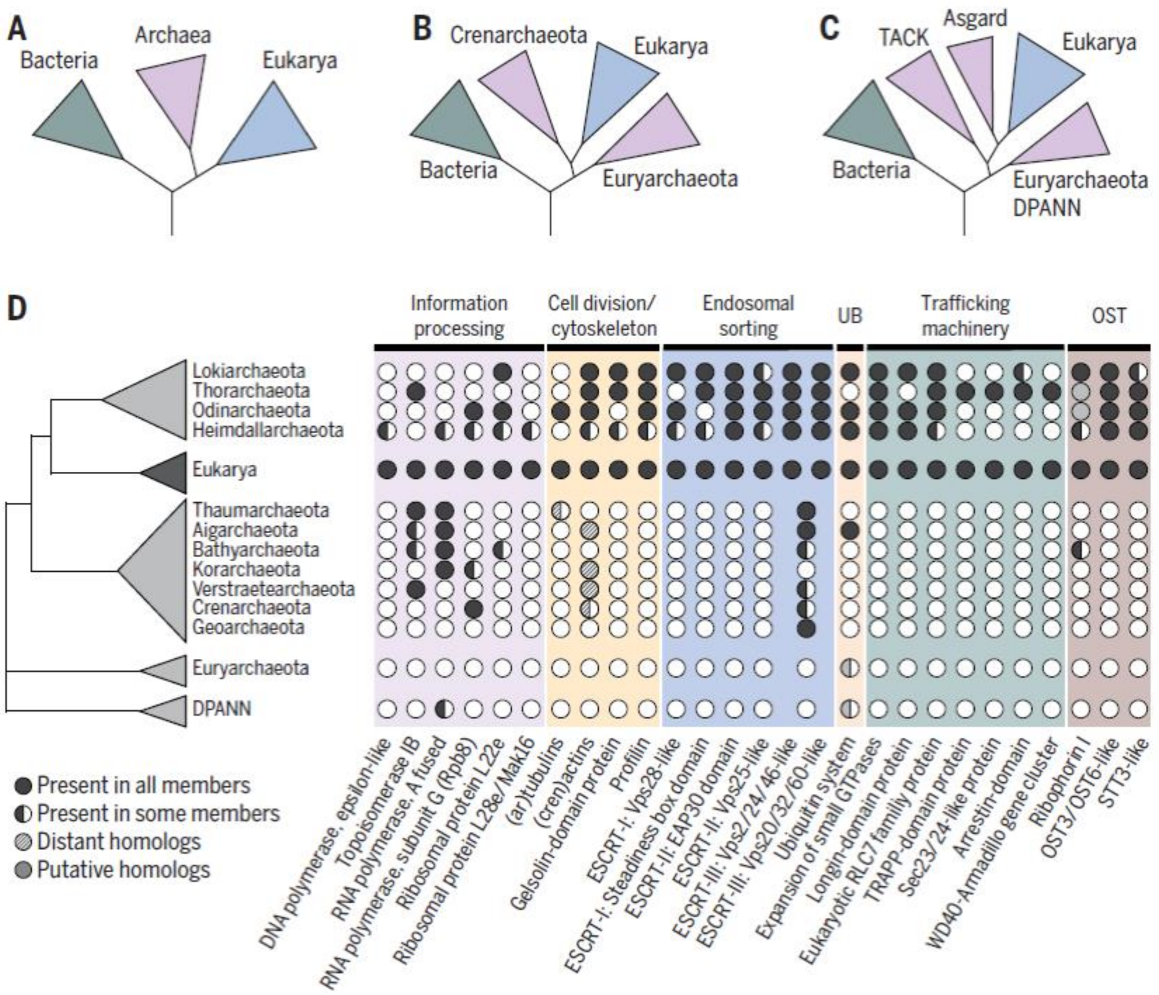


Figure 17. Archaea and the Evolution of the Tree of Life (Spang et al., 2017).
(A to C) Schematic representation of the relationship between archaea, bacteria, and eukaryotes according to the three-domain topology (A) and the two-domain topology (B), updated with the inclusion of Asgard archaea (C). The three-domain tree suggests that archaea, bacteria, and eukaryotes constitute primary domains. In contrast, a two-domain topology [(B) and (C)] aligns better with the idea that eukaryotes represent a secondary domain of life, emerging from the fusion of two prokaryotes: an archaeal host and an alphaproteobacterial symbiont. (D) A schematic tree of archaea and their relationship with eukaryotes, highlighting the placement of eukaryotes within Asgard and showing the distribution of eukaryotic signature proteins (ESPs) across the various lineages.

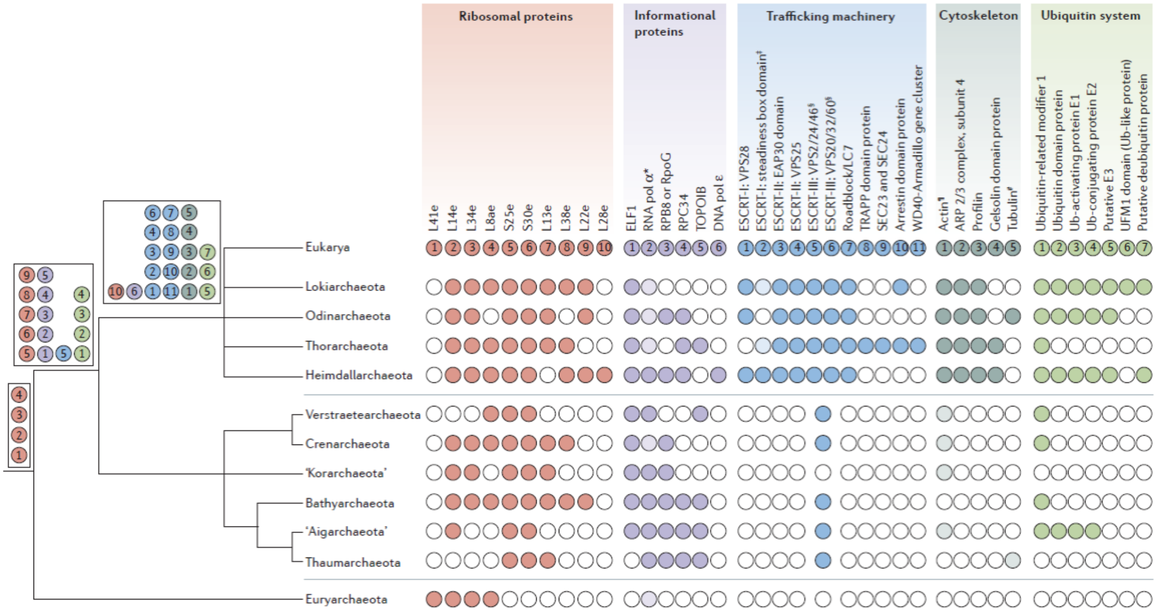


Figure 18. Origins of Eukaryotic Signature Proteins (ESPs) Found in FECA (Eme et al., 2017).
This figure illustrates the presence of homologs of eukaryotic signature proteins (ESPs) across various archaeal lineages (filled circles) and their presumed emergence along the schematic archaeal tree. The origin of each ESP is indicated on a schematic tree of life (left side). Notably, this figure assumes a known phylogeny of archaea.

Typically, these genes are irregularly distributed within the TACK group. However, they appear to be consistently present in the Asgard group, supporting the hypothesis that eukaryotes may have arisen from a fusion between one of these archaea and a bacterium (Da Cunha, Gaïa, et al., 2022) This fusion might have facilitated subsequent endosymbiotic events, enabled by the presence of a dynamic cytoskeleton (e.g., mitochondria, plastids) and the development of membrane systems (e.g., nuclear envelope, organelles). Through genes involved in membrane remodeling and cellular trafficking, this proto-eukaryotic cell could have achieved the internal complexity characteristic of modern eukaryotes.

However, these evolutionary relationships remain contentious. The inclusion of eukaryotes in current datasets is highly limited due to challenges in identifying detectable homology between the genomes of different domains (Zhu et al., 2019).

Analysis of individual phylogenetic trees for each of the 36 universal proteins reveals inconsistencies underlying the global analysis of these same proteins. Specifically, the placement of Asgard archaea varies significantly in individual trees depending on the protein analyzed. In many cases, they do not cluster together: some are closer to eukaryotes, while others are grouped with other archaea. Their proximity to eukaryotes is primarily observed in small proteins, which favor two-domain models (Eocyte hypothesis). Larger proteins, which provide stronger signals for accurate organism placement by allowing the comparison of more amino acid positions, frequently position Asgard archaea among other archaea, supporting a three-domain scenario. In concatenated analyses, the numerous small universal proteins may obscure the signal provided by larger proteins (Da Cunha, Gaïa, et al., 2022).

Currently, the domain Archaea includes four well-recognized supergroups (Adam et al., 2017; Castelle et al., 2015; Castelle & Banfield, 2018; Spang et al., 2017; T. A. Williams et al., 2017) :

- Euryarchaeota, comprising about 15 well-differentiated subgroups.
- TACK group, including Thaumarchaeota, Aigarchaeota, Crenarchaeota, and Korarchaeota.
- DPANN group, encompassing Diapherotrites, Parvarchaeota, Aenigmarchaeota, Nanoarchaeota, and Nanohaloarchaeota.
- Asgards, the most recently defined group.

CHAPTER 2

THE SUPERMATRIX AND PHYLOGENY OF ARCHAEA

1 INTRODUCTION

Recent studies tend to reinforce the Eocyte hypothesis, considering eukaryotes as a subgroup of archaea, likely originating from the Asgard group (Eme et al., 2017; Spang et al., 2017; T. A. Williams et al., 2017). However, the datasets used and the methods employed are regularly questioned, as they may result from phylogenetic reconstruction artifacts (Brinkmann & Philippe, 2007; Da Cunha, Gaïa, et al., 2022; Forterre, 2011; Philippe & Forterre, 1999) (**Box 7**). Our aim is therefore to build the most reliable dataset possible to avoid any bias and to conduct original phylogenomic analyses to test and verify the robustness of the results obtained. To achieve this, we will compile a dataset that is highly representative of the diversity of archaea, ensuring completeness, removing contamination (eliminating horizontal gene transfers), and addressing paralogy issues. Once this dataset is constructed, we will subject it to various stress tests.

Box 7. Phylogenetic Reconstruction Artifacts: Substitution Heterogeneities

When comparing homologous sequences, if they diverged from a recent common ancestor (*shallow phylogeny*), the sequences likely differ only slightly, by a few bases. However, the further back in time we go with species sharing an ancient common ancestor (*deep phylogeny*), the more challenging it becomes to find homologies. Homologous DNA sequences often differ significantly in terms of both the number of bases and their lengths. These differences arise from the accumulation of mutations over time (substitutions, insertions, deletions, etc.). Mutations are alterations to the nucleotide sequence of a genome, such as one of the four bases (A, T, C, G) being replaced by another, deleted, or a new base being inserted. These modifications generally occur through two processes: copying errors or DNA damage.

The simplest evolutionary model, called Jukes-Cantor (JC), assumes that substitutions accumulate at the same rate across sequences, regardless of base composition or mutation type (transition vs. transversion). This model presumes homogeneous and stationary evolution along the tree of life. However, it is now evident that these assumptions are far from reality in actual sequences. It is therefore necessary to adjust distances (pairwise sequence similarities) by accounting for specific substitution patterns in nucleotide sequences.

One of the main challenges in building phylogenies is accounting for substitutions to extract the correct signal. Although Darwin proposed evolution as a continuous process, there is no indication that this evolution proceeds uniformly across lineages. Simpson introduced the concept of "tempo and mode of evolution" to describe how the rate (*tempo*) and type (*mode*) of evolutionary change can vary between lineages, over time, and according to biogeography. To compare modes and tempos across lineages and time, it is necessary to estimate evolutionary rates. Unfortunately, the substitution process is highly heterogeneous, both by site and over time, complicating the creation of comprehensive and effective models. Several types of heterogeneities have been identified:

- Heterogeneity of substitutions among character states. Some substitutions are easier and therefore more frequent. For example, transitions (between purines or pyrimidines) occur more often than transversions (between purines and pyrimidines). The Kimura 2-parameter model (K2P) estimates transition and transversion rates. Additionally, one might expect equal proportions of each base (25%), but certain bases are often more prevalent. The Felsenstein-81 (F81) model accounts for these base proportions. The HKY model incorporates both transition/transversion rates and base proportions.

- Heterogeneity of the substitution rate across sites. Early models assumed that mutations occurred with equal probability across all sites in a sequence. This is incorrect, as some sites are more prone to substitutions, generating artifacts. For example, some sites may experience strong selection and high constraints, while others (e.g., the third base in codons) have fewer constraints. Additionally, functionally significant sites are generally less tolerant to mutations. A "fast" site accumulates more substitutions than a "slow" site over the same time. Ideally, separating slow and fast sites into different datasets would yield the same tree topologies, but branch lengths would be much longer with fast sites. To address this, substitution rates can be modeled as variable across sites, often using a gamma distribution.
- Heterogeneity of the substitution process across sites. Early phylogenetic models assumed any amino acid could occur at any site. However, biochemical evidence shows that only certain amino acids are acceptable at specific positions based on their properties (e.g., charge, hydrophobicity). Similar amino acid sequences from different species often retain functional similarity; sequence homology reflects similar biological functions. These "conservative" substitutions preserve functionality. The number of parameters required to describe amino acid preferences at each site is vast. Thus, grouping sites with similar properties into categories is a practical solution. CAT models use Dirichlet processes to assign each site to a category based on character state frequencies. These models are less prone to homoplasy and long-branch attraction artifacts.
- Composition Heterogeneity Over Time: This refers to changes in the stationary frequencies of nucleotides/amino acids throughout the tree.
- Substitution Process Heterogeneity Over Time (Heteropecilly): This describes changes in acceptable amino acid profiles at individual sites over time within the tree.
- Heterogeneity of rates within positions over time (Heterotachy and covarion): Due to changes in the functional constraints of protein sites and/or epistasis (interaction between genes where one gene’s expression determines another’s), the probability of accepting a mutation at a given position varies along tree branches. For instance, in cytochrome b (a mitochondrial protein involved in cellular respiration), the evolutionary rate of all variable positions changes over vertebrate evolution, independently of changes in other positions. Furthermore, within the same protein, critical functional positions may shift over time. In the 1970s, Fitch and Markowitz (1970) demonstrated that, at any given time, only a small fraction of positions are susceptible to variation, allowing mutations to become fixed. Biologically, this is unsurprising since the functional constraints of protein sites—and thus their evolutionary rates—change independently across lineages over time. Consequently, this can lead to phylogenetic artifacts when the proportions of invariable sites in unrelated species converge.

References

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.

The question of the validity of a tree is crucial in phylogenetics. Should we trust and accept the tree obtained as definitive? Proper model fitting is necessary both for reliable phylogeny estimation and for result interpretation, especially in the presence of confounding factors—methodological and biological phenomena that violate the model (Young & Gillung, 2020) (**Tableau 4**). Assessing the impact of these phenomena is a key objective of this thesis. Our goal is to evaluate the variability of phylogenetic trees by varying genes, species, models, and phylogenetic reconstruction methods. Various approaches are commonly used to measure the robustness of each branch in a tree. In this thesis, evaluation will be performed using a jackknife resampling method (without replacement) and by calculating normalized Robinson-Foulds distances between the obtained trees. Taxonomic sampling variation through species jackknifing will be used to assess its impact on the results. Then, we will identify possible bipartitions to study recurrent topological changes and determine whether some are due to artifacts. Finally, for each replicate, we will test several effects: the impact of the model used, site removal, and more.

Sources of systematic error	Strategies to overcome	Main references
Compositional heterogeneity	Remove the loci or taxa exhibiting extreme deviation in composition. Alternatively, explicitly modelling compositional heterogeneity may be more satisfactory.	(Borowiec et al., 2019; Duchêne et al., 2017; Jeffroy et al., 2006; Roure & Philippe, 2011)
Missing data	Design smaller (i.e. with fewer loci) but more complete datasets, as opposed to larger (i.e. with more loci) but sparser datasets. Analyse subsamples of the data to see if missing data cause conflict.	(Hosner et al., 2016; Kocot et al., 2017; Roure et al., 2013; Smith et al., 2020)
Branch-length heterogeneity	Perform analysis with and without taxa and genes most likely to be susceptible to long-branch attraction (e.g. remove rapidly evolving sites or taxa).	(Kück & Wägele, 2016; Nosenko et al., 2013; Qu et al., 2017; Struck, 2014)
Paralogy	Filter and remove paralogous loci using an appropriate method for orthology prediction. Carefully examine gene topologies to detect paralogues.	(Betancur-R. et al., 2014; Siu-Ting et al., 2019; Struck, 2013)
Hétérotachy, hétéropécilly	Removal of most varied sites from alignment. Use evolutionary models that explicitly model heterotachy (e.g. ghost, implemented in iqtree).	(Bouckaert & Lockhart, 2015; Crotty et al., 2020; Kocot et al., 2017; Zhong et al., 2011)
Gene tree heterogeneity	Use coalescence approaches. Perform statistical binning, tree concordance analysis.	(Betancur-R. et al., 2013; Richards et al., 2018)

Tableau 4. Major confounding factors in phylogenomic analyses and the strategies to mitigate them (Young & Gillung, 2020).

To best counter stochastic error, we used two different reconstruction methods: a supermatrix method and a supertree method (**Box 8**). In both cases, our goal was to obtain a collection of trees, including one LG4X tree, LG+C20+F+G, LG+C60+F+G, and PMSF for each of the five species replicates. Indeed, likelihood analyses using the LG+C20+F+G and LG+C60+F+G4 models improve upon the LG4X model by accounting for site-specific compositional heterogeneity using 20 or 60 categories. These models fit the data much better than LG4X when evaluated using the Bayesian Information Criterion (BIC) (**Box 5**).

Box 8. Combating Stochastic Error: Supermatrix vs. Supertree

Stochastic Error (= Sampling Error)

Even if evolution occurred exactly as assumed by the evolutionary model used for phylogenetic inference, an incorrect tree might be obtained due to the finite size of alignments. If the extracted signal is too weak, it becomes impossible to distinguish the best topology among multiple solutions, leading to poorly supported nodes. This issue, well-known in statistics, is called sampling error or stochastic error, and it primarily depends on the amount of analyzable data. By definition, stochastic error disappears with infinitely large samples. This issue is particularly significant when working with large timescales because sites quickly become saturated with multiple substitutions, erasing the original signal and replacing it with a false signal caused by mutational biases. Using complex models that best describe the data can significantly help mitigate this issue. Additionally, the more slowly a protein evolves, the more likely it is to retain an ancient phylogenetic signal, though its constrained evolution might also make it more prone to convergences.

To infer the phylogeny most representative of the "true tree," one solution is to adopt the methodological concept known as "Total Evidence" by collecting several independent phylogenetic markers to extract the majority signal in a consensus approach. This idea led to the development of phylogenomics, an extension of phylogeny aiming to reconstruct the evolutionary history of species by combining the phylogenetic information of numerous genes.

In this case, two approaches are possible: the supermatrix or the supertree.

Supermatrices

A supermatrix is a synthetic alignment obtained by concatenating a set of sequence alignments. A standard phylogenetic reconstruction method is then applied to this "superalignment." It has been shown that this method yields reliable trees, notably because concatenating multiple alignments reduces stochastic error. One bias to consider in this method is the number of missing genes in certain species, which leads to missing characters within the supermatrices. However, the proportion of missing data can be relatively high without reducing the resolution of the phylogenetic reconstruction. In fact, selecting an appropriate sequence evolution model benefits phylogenetic accuracy more than reducing the proportion of missing data. A significant drawback of this method is its high computational resource cost. Moreover, the supermatrix assumes that the evolution of each selected gene can be explained by a single model and that their phylogeny matches the species phylogeny. However, this is known to be incorrect for two main reasons.

1. The history of a gene does not necessarily reflect the history of a species. This is particularly the case with gene transfers (xenologs) or ancestral polymorphism (Incomplete Lineage Sorting = ILS). In the latter case, the existence of an ancestral polymorphism that has since disappeared results in the fixation of an allele that does not reflect the species' history. This issue can be managed through reconciliation.
2. The second reason relates to substitution heterogeneity issues, tied to the quality of the evolutionary modeling. Partitioned models, which assume that sequence evolution can vary for each gene, address this limitation.

Supertrees

A supertree is a synthetic tree obtained by assembling a set of phylogenetic trees defining sets of species that often partially overlap the species of interest. This supertree must best reflect the various evolutionary histories modeled by each source tree (from different genes). This issue generalizes the consensus of a tree forest, which consists of finding the tree most representative of a collection of source trees (cf. **Box 10** on bootstrap and jackknife). Among the many existing supertree methods, the most popular is the Matrix Representation with Parsimony (MRP) method. This method can produce supertrees close to those derived from supermatrices, provided that the individual source trees are congruent and the matrix representation contains few missing character states. Otherwise, the supertree's reliability diminishes due to the accumulation of stochastic errors. The supertree method does not directly solve the stochastic error problem because each gene tree may still be affected. However, one advantage of this method is its ability to combine trees derived from both morphological and molecular data and detect xenology issues. Additionally, it is computationally efficient and facilitates gene jackknifing. In our thesis, we used ASTRAL, an improvement on the MRP method. ASTRAL (Accurate Species Tree Algorithm) is a tool for estimating unrooted species trees given a set of unrooted gene trees. ASTRAL is statistically consistent under the multispecies coalescent model (and is thus useful for handling ancestral polymorphisms). It uses dynamic programming to find an exact solution, even with large datasets. ASTRAL identifies the species tree with the maximum number of shared induced quartet trees with the set of gene trees, constrained by a predefined set of bipartitions empirically determined by ASTRAL. ASTRAL employs the Maximum Quartet Support Species Tree (MQSST) approach. It divides the tree into $\binom{n}{4}$ species quartets, identifies the dominant tree for each quartet, and then combines these quartets. For a quartet, MQSST considers the relative frequency of the three alternative topologies and their weights, adjusting accordingly. Thus, if the dominant topology (i.e., the most frequent) of a quartet is much more frequent than the alternatives, trees that do not induce the dominant topology are penalized. Conversely, if all three topologies of the quartet have frequencies close to 1/3, that quartet contributes little to problem optimization. This implies that the more frequent a gene tree is, the more it reinforces its topology in the species tree.

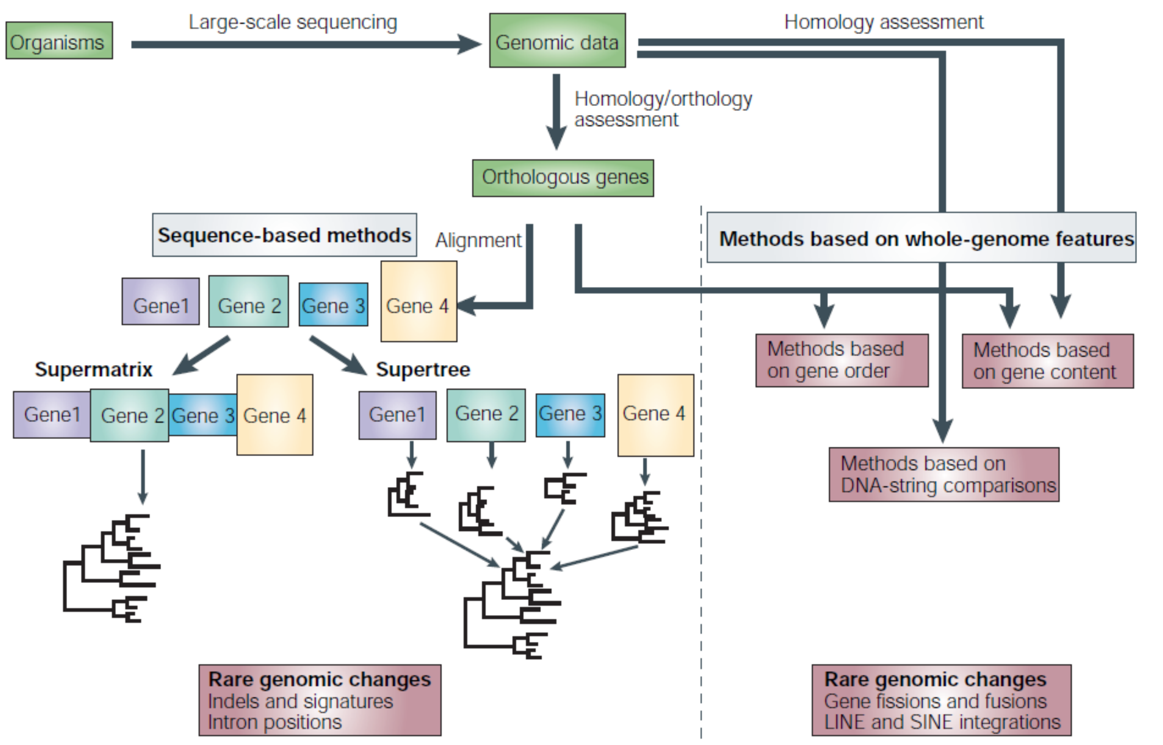


Figure 19. Phylogenetic Inference Methods.

This flowchart illustrates the inference steps leading to the construction of phylogenetic trees from genomic data. These data are obtained through large-scale DNA sequencing. After assessing the homology and orthology of certain genes, datasets of orthologous genes from the selected species are assembled. The trees are then computed using either the supertree method or the supermatrix method.

References

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 SPECIES SELECTION

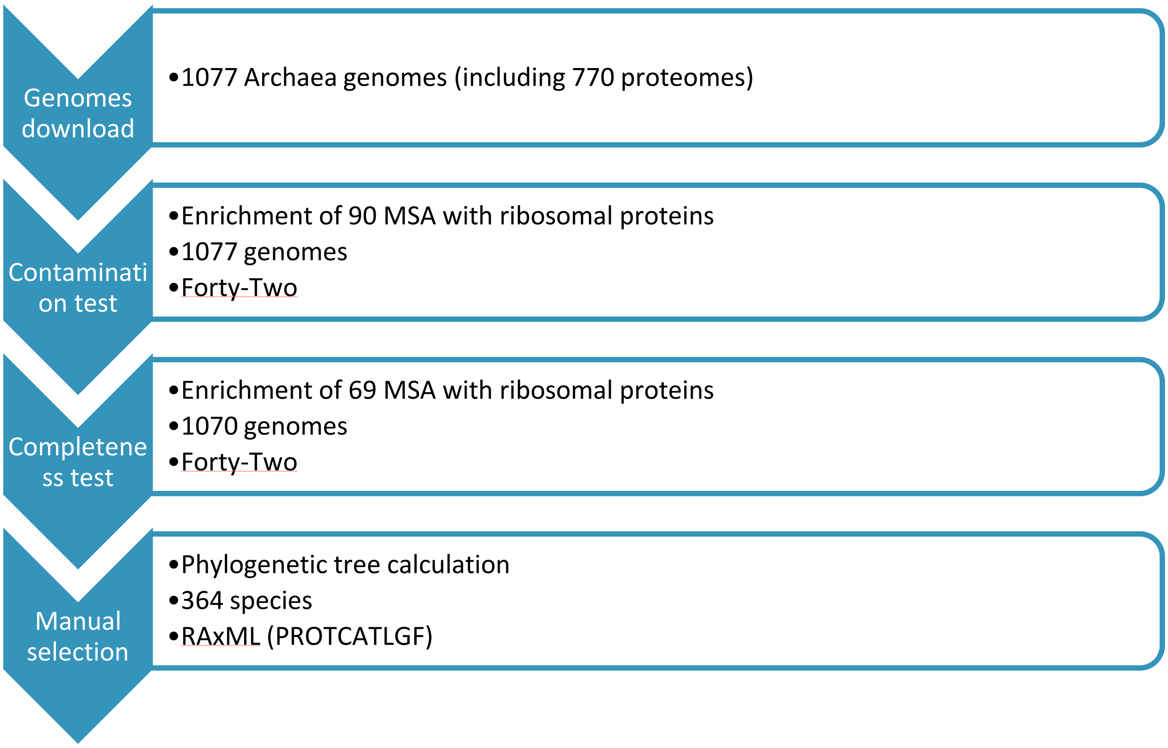

Genomes download
•1077 Archaea genomes (including 770 proteomes)
Contamination test
•Enrichment of 90 MSA with ribosomal proteins
•1077 genomes
•Forty-Two
Completeness test
•Enrichment of 69 MSA with ribosomal proteins
•1070 genomes
•Forty-Two
Manual selection
•Phylogenetic tree calculation
•364 species
•RAxML (PROTCATLGF)

Figure 20. Species selection protocol.

2.1 COLLECTION OF ARCHAEA GENOMES

A total of 1,077 Archaea genomes were collected from both GenBank and RefSeq via the National Center for Biotechnology Information (NCBI) portal. Among these genomes, 770 had proteome predictions available, representing approximately three-quarters of the dataset. The quality of each genome was assessed to gather the highest-quality genomes possible. The following parameters were evaluated (cf. **Supp.Mat quast.csv**):

- The number of contigs;
- The number of contigs exceeding 1,000 nucleotides (**Figure 21**);
- The total genome size;
- The total genome size based on contigs exceeding 1,000 nucleotides;
- The size of the largest contig;
- The N50 and N75 proportions;
- The L50 and L75 proportions;
- The number of N's per 100 kbp.

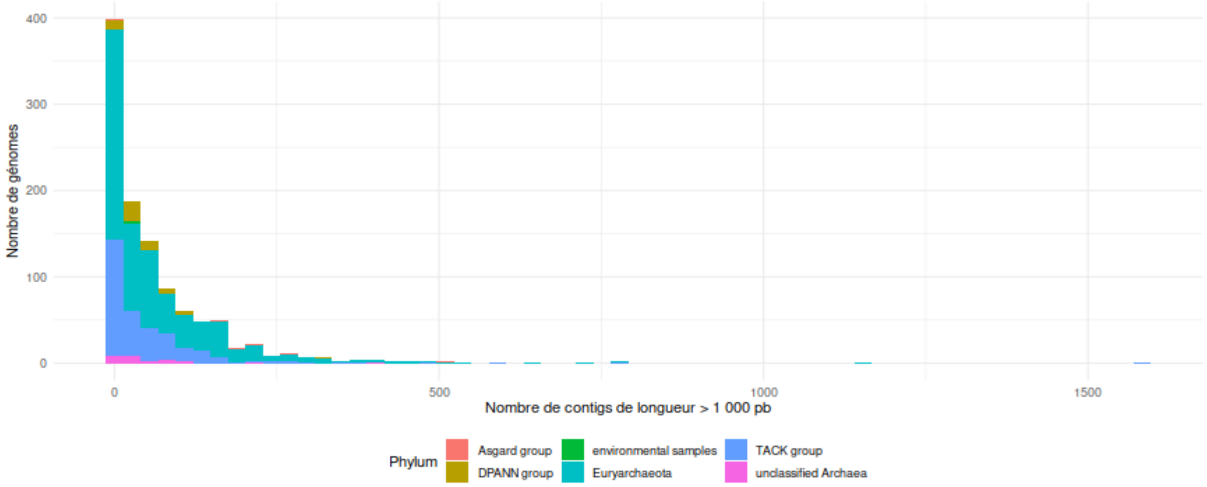


Figure 21. Distribution of the number of contigs used to estimate the quality of our archaeal genome assemblies.

The *Euryarchaeota* (particularly *Halobacteria*) and *Crenarchaeota* are highly overrepresented among the downloaded genomes (**Figure 22**). Archaea from the "unclassified" group will later be assigned to a clade based on their phylogeny inferred from ribosomal proteins. A larger genome size is also observed among representatives of the *Asgard* group (**Figure 23**). Biologically, this could be interpreted as evidence of a trend toward increasing complexity along the eukaryotic lineage (or conversely as a simplification of eukaryotes toward Archaea). Alternatively, it may more pragmatically reflect a technical issue related to MAG reconstruction (e.g., potential contamination due to improper separation of scaffolds from multiple organisms).

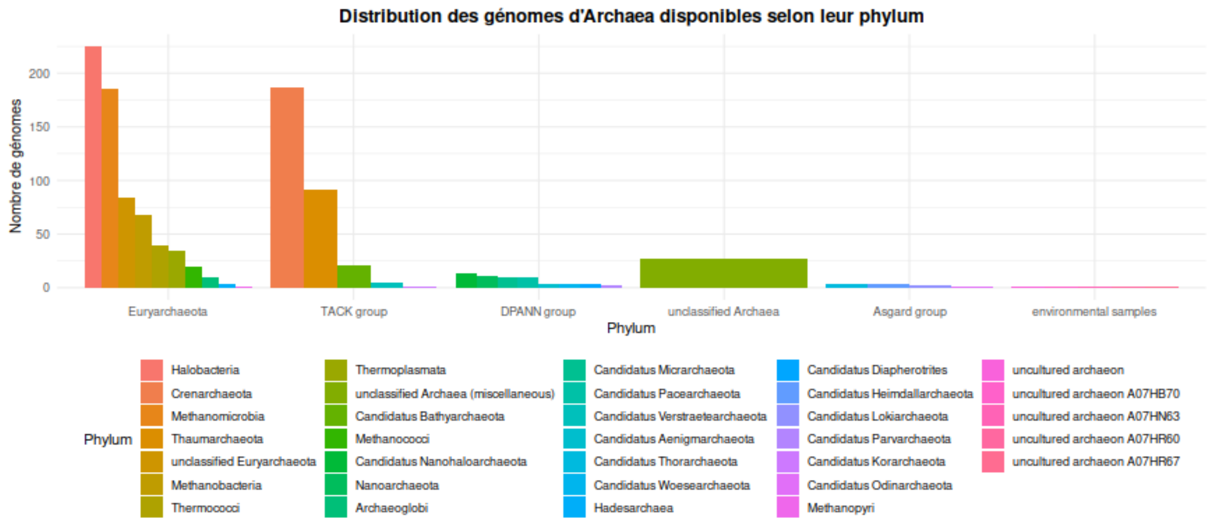


Figure 22. Distribution of available archaea genomes by phylum.
Colors correspond to NCBI taxonomy. Euryarchaeota and Crenarchaeota are strongly represented.

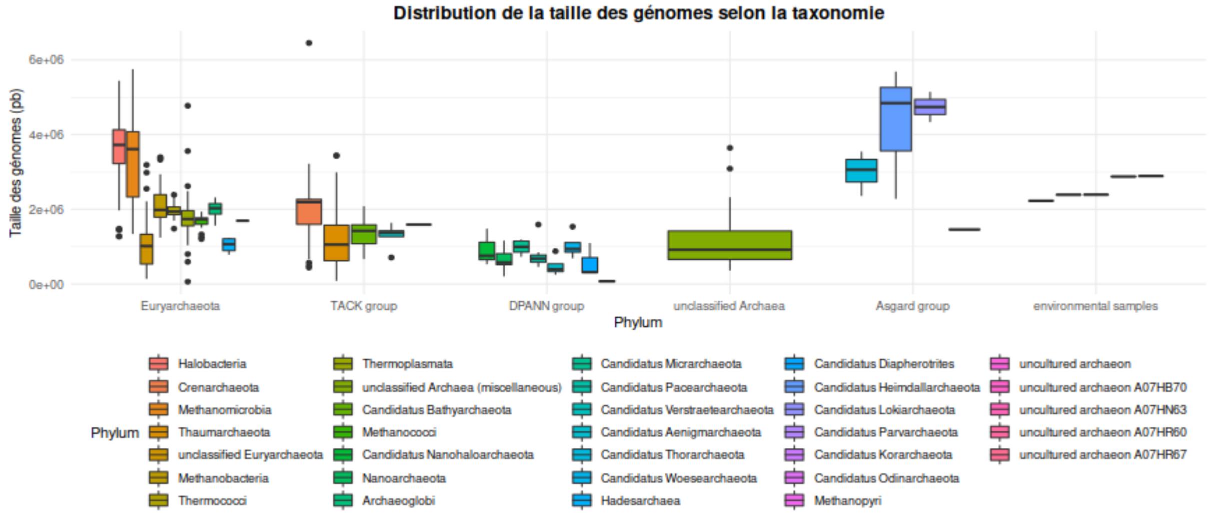


Figure 23. Genome size distribution by phylum.
Colors correspond to NCBI taxonomy. A larger genome size is observed in the Asgard group, which could be interpreted either as the implementation of a complexification towards the eukaryotic lineage (or, on the contrary, as a simplification from eukaryotes to archaea), or as a technical problem linked to MAG reconstruction.

2.2 ANALYSES OF RIBOSOMAL PROTEINS AND PRELIMINARY TAXONOMIC SAMPLING

To establish a preliminary phylogeny of archaea, an initial dataset was assembled by supplementing 90 multiple sequence alignments (MSAs) of archaeal and bacterial ribosomal proteins. Some ribosomal proteins are universal to all living organisms, while others are specific to bacteria, eukaryotes, or shared exclusively by archaea and eukaryotes. Among the 1077 genomes, 1070 contained at least one ribosomal protein. Seven "species" were excluded as they were highly incomplete and lacked any ribosomal proteins (cf. **Supp.Mat 7-genomes-non-rajoutes.txt**). Consequently, these genomes were removed from our dataset. However, certain genomes were not included in all ribosomal protein alignments, suggesting they are either incomplete or that BLAST failed to align them (perhaps due to specific protein properties, such as a particularly small size). Analyzing their distribution across our genomes helps identify potentially contaminated genomes. For each genome, we searched for sequences homologous to those in the reference alignment MSAs, then differentiated orthologs from potential paralogs. Interestingly, a few archaea were added to MSAs of ribosomal proteins that are typically bacterial (highlighted in green in **Figure 24**). This suggests potential contamination, as these genomes should be strictly archaeal. Upon enrichment, bacterial sequences were added, interpreted as contamination in 42 cases. This process simultaneously allowed us to assess the completeness of our genomes. Ideally, each archaeal or universal ribosomal protein should be present in all 1077 genomes.

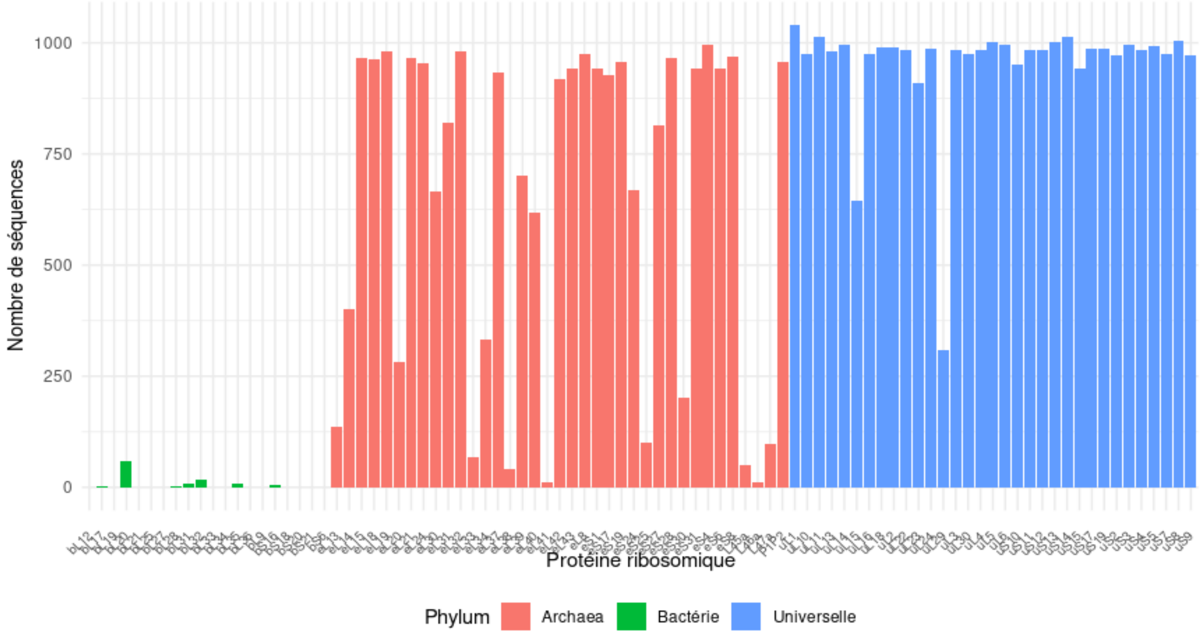


Figure 24. Number of sequences added per 42 for each ribosomal protein.

A few archaea have been added to the MSAs of ribosomal proteins that are in principle exclusively bacterial (green).

To verify whether any archaeal genomes have been contaminated (or not) by bacterial sequences, we performed a contamination test (Irisarri et al., 2017; Simion et al., 2017) using "42," this time using bacteria as reference proteomes during the BRH (**Box 9**). By doing so, we specifically searched for bacterial ribosomal protein sequences rather than archaeal ones. For universal ribosomal proteins, we would not expect to observe any differences between the two analyses of "42" reported here.

Box 9. The BRH : *Best Reciprocal Hits*

The acquisition of a set of orthologous gene alignments from genomic databases serves as the starting point for constructing a phylogenomics dataset. It is from the sequences that constitute these alignments that the corresponding orthologs in other species are searched. The BRH method relies on the better conservation of sequences between orthologous genes, as demonstrated by systematic comparisons of at least two complete genomes. It involves searching, during inter-genomic comparisons, for genes across multiple species that exhibit the best mutual similarities. The underlying idea is that "true orthologs" have retained the same properties as their common ancestor and, therefore, show the highest sequence conservation. Paralogues, on the other hand, may have evolved with fewer constraints (or new constraints). BRH is probably the most functional definition of orthology: two genes from two different genomes are considered orthologous if their proteins are mutually identified as the best hit (best score) in the compared genomes. Thus, if starting from a sequence A in the proteome of species X, the best similarity result (the BEST HIT) is sequence B in the proteome of species Y, and if sequence A is also the BEST HIT in the proteome of species X when starting from sequence B of species Y, then these are considered reciprocal best hits and are interpreted as likely orthologs.

Two main types of sequence similarity search algorithms are distinguished: those based on local sequence alignment (e.g., BLAST) and those using a Hidden Markov Model (HMM). Once the homologous candidate sequences are retrieved, it is necessary to verify their orthology relationship. This step can be done by placing them in the gene tree or by comparing them to a database of orthologs. The latter approach is akin to the principle of reciprocity in best-hit results. This BRH condition is widely used to define orthology relationships. It should be noted that the heuristics of "42" are more complex than described here, particularly for handling recent paralogs (in-paralogs), although the basic principle remains the same (cf**. Material & Methods**).

References

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.

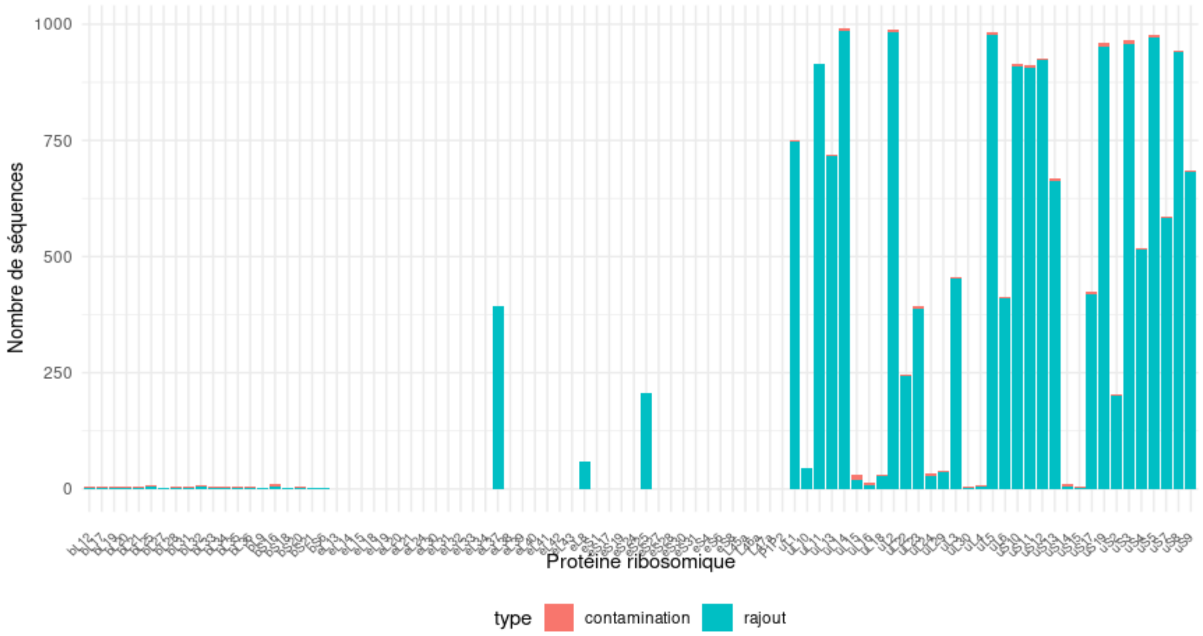


Figure 25. Number of bacterial sequences added per 42 for each ribosomal protein.
Our protocol enabled us to limit the proportion of contaminated genomes.

A very small proportion of contamination is observed (in red on **Figure 25**). We then obtained a list of 31 genomes that contain at least one ribosomal protein contaminated by a non-archaeal sequence (cf**. Supp.Mat 31-genomes-contamines.txt**).

Simultaneously, we concatenated 69 MSA of ribosomal proteins into a supermatrix of 8,344 well-conserved amino acid positions × 1,070 species. These ribosomal proteins were used as indicators to assess genome completeness (the number of concatenated amino acids reflects the number of ribosomal proteins found). From this alignment, we calculated a phylogenetic tree using RAxML with a rapid bootstrap search (100 replicas) and the PROTCATLGF model (cf. **Supp.Mat 1070sp-ribo.nex**). We then manually selected 364 species representing the maximum diversity of archaea, considering contamination information and favoring species with available proteomes and completeness (cf. **Supp.Mat archaea-364sp.lis**). Of these 364 archaea, 305 had a proteome available (cf. **Supp.Mat selection-proteomes.txt**), while 59 species were only available as unannotated genomes (cf. **Supp.Mat selection-genomes.txt**). A new ribosomal tree was then calculated based on this new alignment of 7,817 positions and 364 species, using the same method as before (cf. **Figure 26** & **Supp.Mat FigS26**). By reducing the number of species from 1,070 to 364, we lost 527 columns. This tree of 364 genomes is a condensed version of our tree with 1,070 species. It effectively represents the diversity of archaea and closely resembles it, but without excessive redundancy from over-sampled groups in GenBank.

We obtained the following groups: 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, and Unclassified Euryarchaeota. The last group, Unclassified Euryarchaeota, could be related to Hadesarchaeota.

Within our ribosomal tree, we noted the presence of two distinct groups called "Methanomicrobia." We observed that, for this ribosomal tree, the statistical support estimated by bootstrap proportions was not maximal for all nodes, possibly reflecting a lack of phylogenetic signal. Additionally, it cannot be excluded that this tree is affected by various phylogenetic reconstruction artifacts, particularly the long-branch attraction artifact (cf. archaeon CG2 30 31 98 GCA 001872885.1). It is known that artifacts generate non-phylogenetic signals that may oppose phylogenetic signals and affect tree resolution, whether or not the trees are topologically correct (Baurain & Philippe, 2010). To address this issue, we first decided to increase the number of genes by extending our dataset to include non-ribosomal genes, then implemented a series of strategies to minimize artifacts.

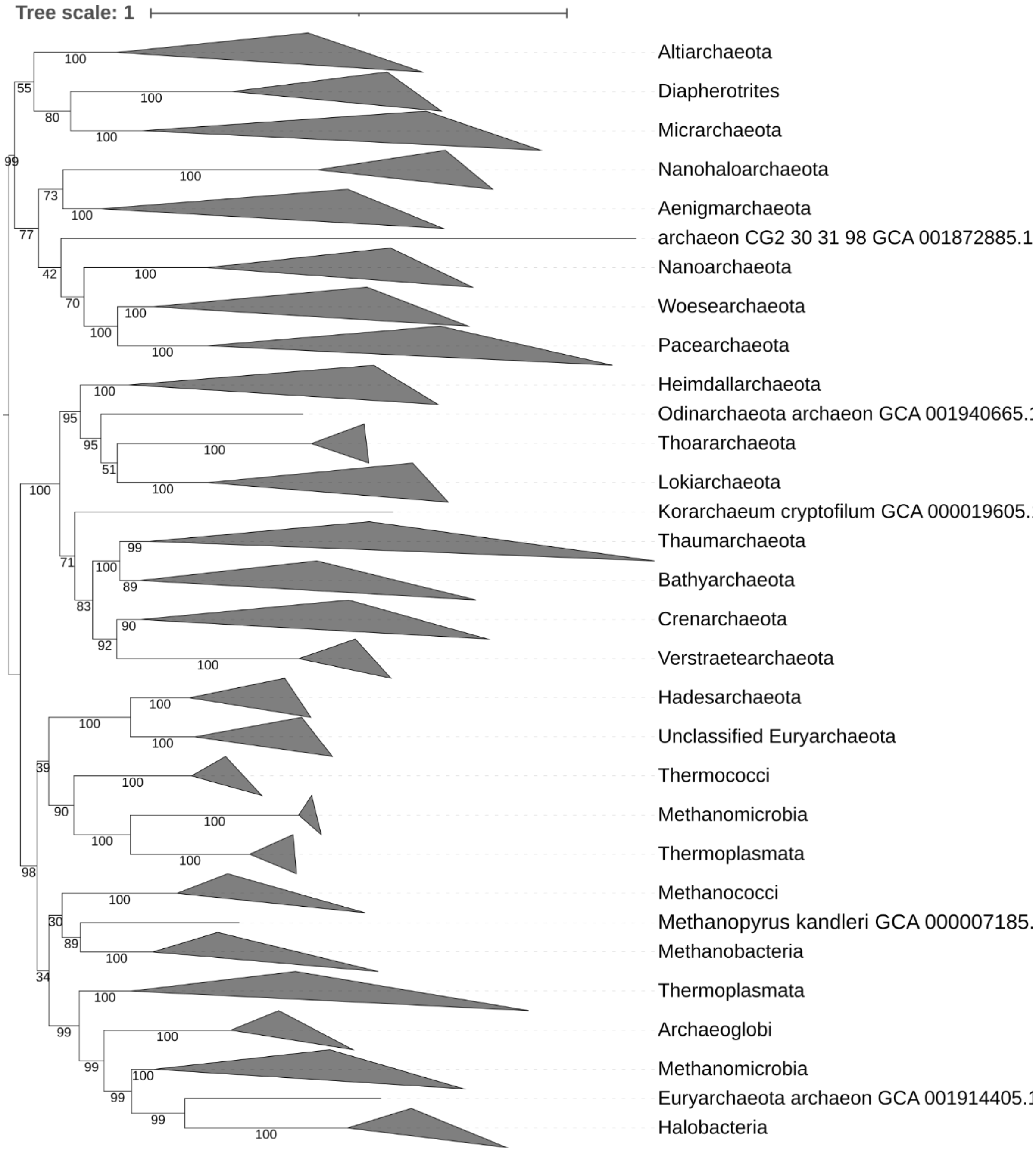
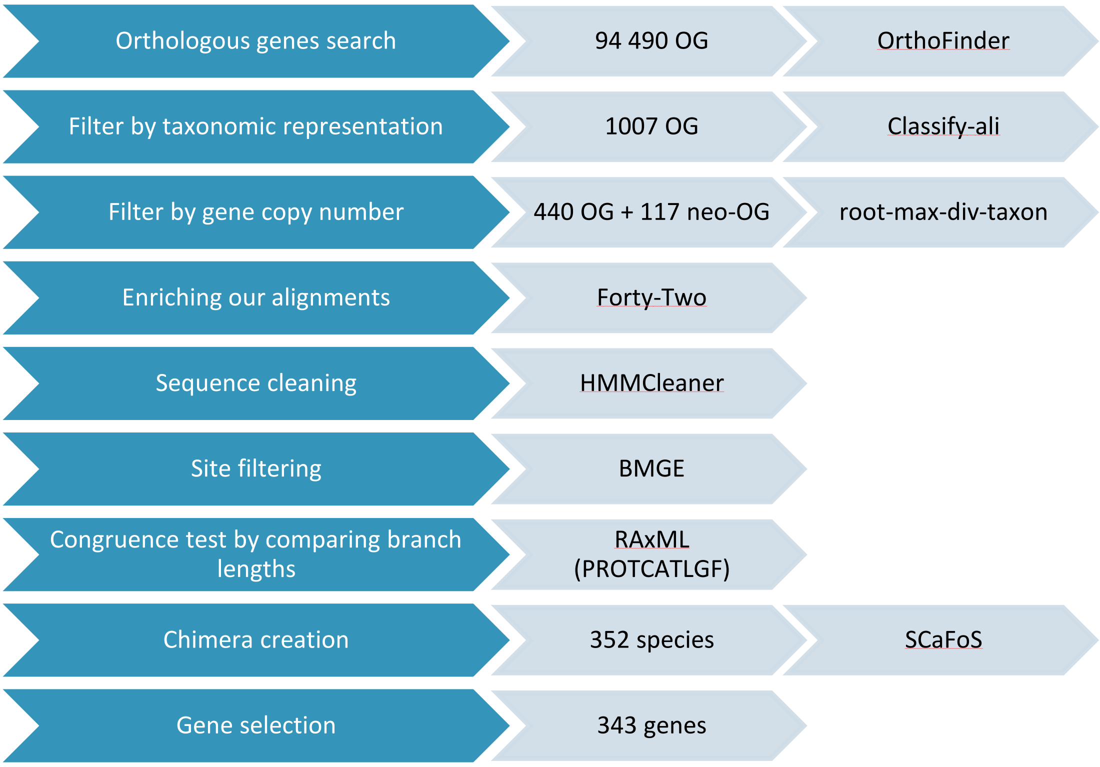


Figure 26. Ribosome tree of 364 archaea calculated with RAxML using rapid-bootstrap searching (100 replicates) according to the PROTCATLGF model on a super-matrix of 7,817 positions. Data are provided in **Supp.Mat FigS26**.

3 GENES SELECTION



Orthologous genes search	94 490 OG	OrthoFinder
Filter by taxonomic representation	1007 OG	Classify-ali
Filter by gene copy number	440 OG + 117 neo-OG	root-max-div-taxon
Enriching our alignments	Forty-Two	
Sequence cleaning	HMMCleaner	
Site filtering	BMGE	
Congruence test by comparing branch lengths	RAxML (PROTCATLGF)	
Chimera creation	352 species	SCaFoS
Gene selection	343 genes	

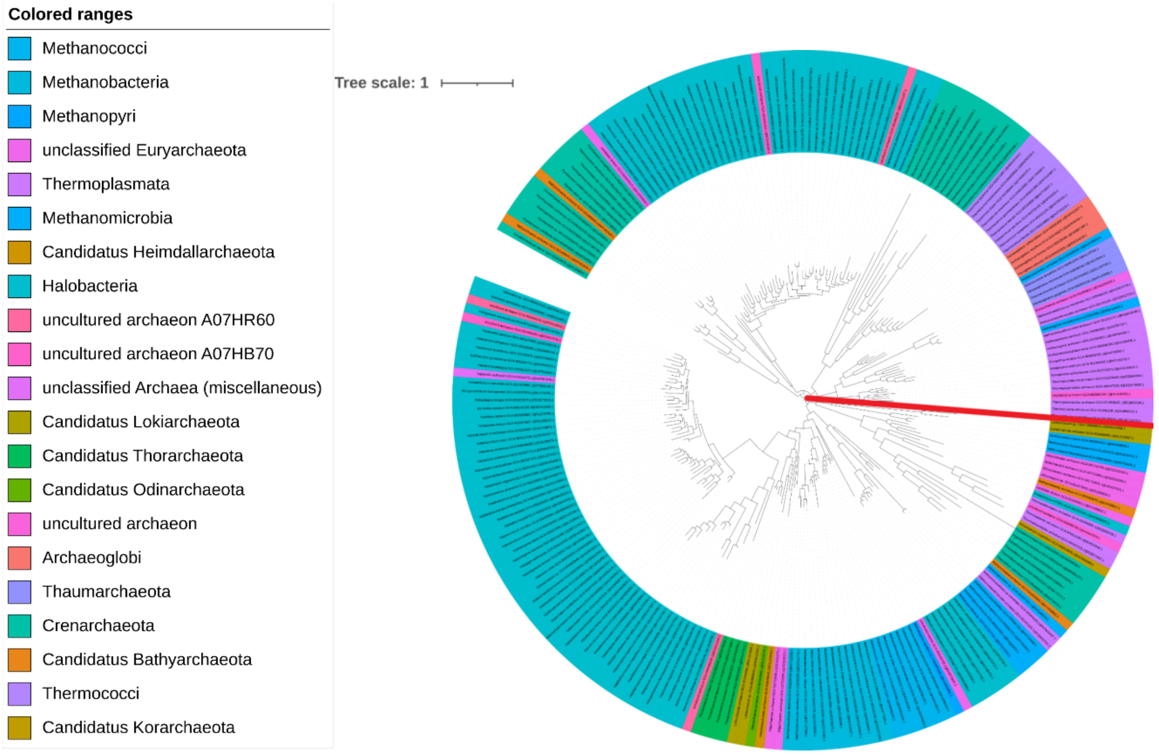
Figure 27. Gene selection protocol.

3.1 CONSTRUCTION AND SELECTION OF ORTHOLOGOUS GROUPS ACCORDING TO TAXONOMIC REPRESENTATION

To expand our gene selection beyond ribosomal proteins, we downloaded from NCBI the 305 proteomes corresponding to our selected species in order to generate orthologous gene groups. We generated 94,490 orthologous groups (OGs), including 66,798 singletons consisting of a single sequence. We then applied two successive filters based on taxonomy and the number of species present. The first filter, based on taxonomy, ensured that each gene included at least one Euryarchaeota and one TACK (Thaumarchaeota, Aigarchaeota, Crenarchaeota, Korarchaeota) species. This resulted in 4,843 OGs. The second filter allowed us to retain only those genes present in at least one-quarter of our selected 364 species (i.e., at least 91 species) to avoid including OGs that were too sparsely sampled. After applying this filter, 1,007 OGs remained.

3.2 SELECTION OF ORTHOLOGOUS GROUPS BASED ON GENE COPY NUMBER AND MANAGEMENT OF MULTIGENIC FAMILIES

We applied a tree-cutting method to sort and select our genes. Our goal was to create two datasets: one of high quality, containing only single-copy genes, and the other, of relatively lower quality, containing genes from older or more recent duplications (out-paralogues). The tree-cutting procedure was repeated four times. An example of a gene cut by this protocol is shown in **Figure 28**. Among the genes cut according to these criteria, we removed those that were present in fewer than 91 species, in the same way as in the previous dataset. The results of each root-max-div-taxon are provided in **Tableau 5**. In the first round, we identified 21 genes without paralogues and no species duplication, and 731 genes that were not cut (either because they didn't meet the cutting criteria or because there were not enough duplications), totaling 752 genes. After the first round, we retained 440 genes without duplications. The remaining 312 genes were discarded as imperfect (due to duplications that could not be split into two sub-genes). These genes constitute our first dataset (cf. **Supp.Mat 440-genes.txt**). At the end of this process, we obtained 117 additional genes from the cutting of duplicated genes. These genes will be treated separately and will be used to test the robustness of the results obtained from the 440 genes using a corroboration principle. However, we will consider these 117 "neo-orthologous" as forming a potentially less reliable dataset compared to the 440, for which the orthology of the sequences is less doubtful.



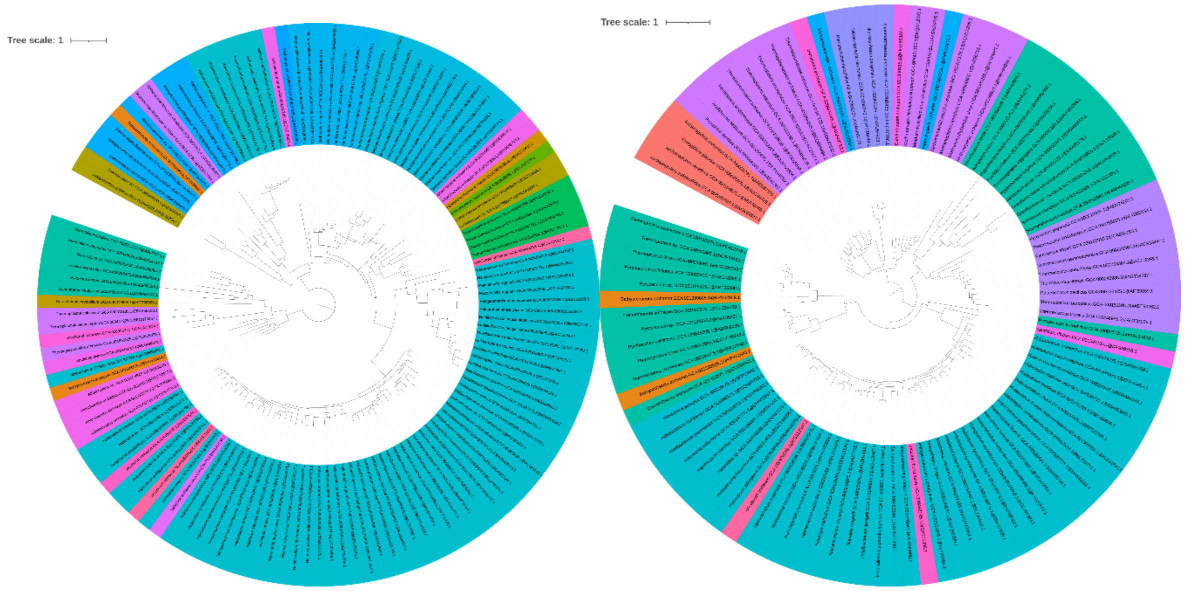


Figure 28. Example of a duplicated gene (OG0000527) cut by the root-max-div-taxon program with RAxML according to the PROTCATLGF model.

Data provided **Supp.Mat FigS28**. The red line indicates the cutting location, giving two new genes from which new trees were calculated. The initial alignment contains 230 sequences, corresponding to 145 taxa. Of these 145 taxa, 79 are unique, while 86 occur at least twice. After partitioning, we obtain two new trees, one with 129 sequences and the other with 101.

	ROUND 1	ROUND 2	ROUND 3	ROUND 4
Input genes	1007	417	144	65
Genes without paralog	21 (21)	93	18	6
Uncut genes	731 (419)	228	85	59
Cut genes	255	96	41	X
Cut genes x2	510	192	82	X
> 90 species (moving to round n+1)	417	144	65	X

Tableau 5. Cutting out paralogous genes to recover new orthologs.

The numbers in brackets correspond to the orthologous genes that pass the 91 species filter and will actually be recovered. Their number in round 1 is equal to 440 (21 + 419). Highlighted values correspond to “neo-orthologous” genes that are recovered after cutting.

3.3 FINAL TAXONOMIC SAMPLING

We enriched the 440 + 117 protein MSAs by adding the 59 species that were only available as genomes, in order to match their taxonomic sampling with our initial selection of 364 species based on the ribosomal protein MSAs (cf. **Supp.Mat nb-genomes-rajoute-par-OG.txt**). The MSAs were then filtered using HMMCleaner and BMGE, and concatenated into a super-matrix using SCaFoS. The first dataset of 440 genes consists of 89,135 amino acid positions. Out of the 61,913 sequences, 2,771 exhibit paralogy. During the concatenation process, we applied a maximum divergence threshold of 25% per sequence pair to avoid systematically eliminating both paralogues and to retain the least divergent sequence. Sequences with divergence beyond this threshold were systematically removed. Among these paralogues, 2,017 are recent in-paralogues.

Therefore, the choice of paralogue during concatenation did not impact the phylogenetic position of these genes, as in all cases, their phylogenetic placement remains correct. However, 754 sequences were removed during concatenation due to excessive divergence (out-paralogues or xenologues), for which it was impossible to determine the correct one. Six species had genes with more than five paralogues (cf. **Supp.Mat esp-paralog-sup-5.txt**). The second dataset of 117 genes consists of 47,154 positions.

We then calculated a phylogenetic tree for the 440 genes using RAxML with the PROTCATLGF model (**Figure 29** & **Supp.Mat FigS29**).

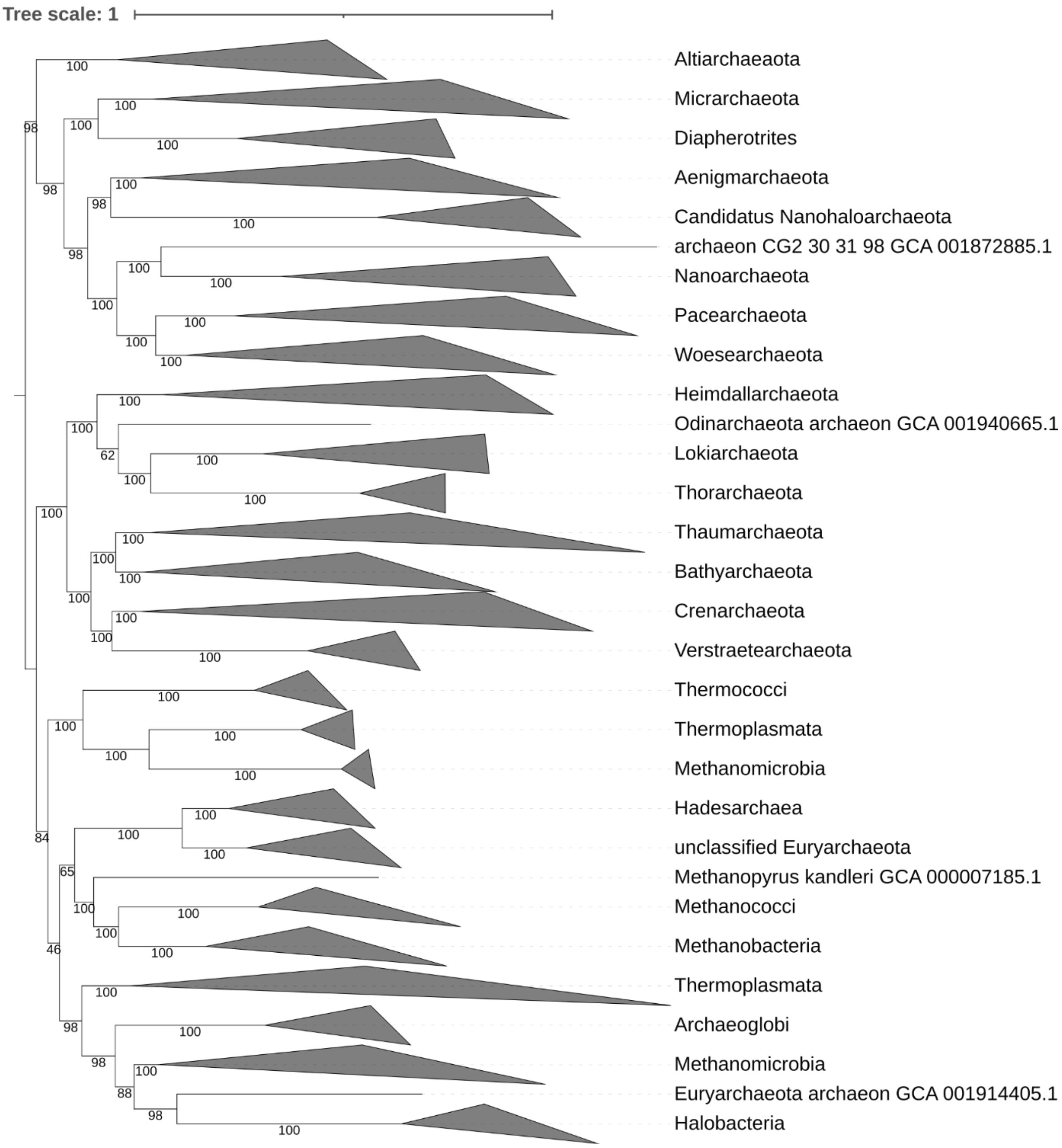


Figure 29. 440-gene tree calculated with RAxML using a rapid-bootstrap search (100 replicas) according to the PROTCATLGF model on a super-array of 89,135 positions and 364 species.

The data are provided in **Supp.Mat FigS29.** The use of 440 genes has little effect on the topology of our archaeal tree when compared to the ribosomal tree. The main uncertainty concerns the position of the Hadesarchaea, which shifts from being the sister group to (Thermococci + Methanomicrobia + Thermoplasmata = TTM) to being the sister group to (Methanopyrus + Methanococci + Methanobacteria). However, the low bootstrap values of 39% and 65% do not allow us to make a definitive conclusion about their true affinity.

The use of 440 genes slightly alters the topology of the tree based on ribosomal proteins. Indeed, in the ribosomal tree, the group Hadesarchaea is the sister group of (Thermococci + Methanomicrobia + Thermoplasmata = TTM) with a bootstrap value of 39%, but here it becomes the sister group of the group (Methanopyrus + Methanococci + Methanobacteria) with a bootstrap value of 65%. It seems that, in both cases, their position is uncertain. We therefore observe differences between our two datasets. We also observe two groups corresponding to undetermined archaea (*Unclassified Euryarchaeota*). One of these groups includes the Altiarchaeales, the other corresponds to archaea that do not fit a defined nomenclature. However, we are not convinced by the high bootstrap values. Indeed, in phylogenomics, a branch must have a maximum bootstrap value to be considered reliable. However, high values can result from non-phylogenetic signal (Baurain & Philippe, 2010). Indeed, tree reconstruction methods, due to their limitations, generate parasitic phylogenetic signals that compete with the true, authentic signal (Di Franco et al., 2022). For example, when a significant bias favors an alternative branching order (e.g., a long-branch attraction between two unrelated taxa with rapid evolution), the non-phylogenetic signal may dominate over the authentic signal, resulting in a phylogenetic signal supporting an incorrect alternative branch. This non-phylogenetic signal depends not only on the properties of the radiation and the dataset (e.g., overall rate of evolution or taxonomic sampling), but also on the accuracy with which the evolutionary model infers substitution history at each position. In cases where not all loci share the same history, the non-phylogenetic signal can be amplified by other model violations, such as when using a concatenated model in the presence of ILS (*Incomplete Lineage Sorting*).

3.4 BRANCH LENGTH COMPARISON CONGRUENCE TEST (BLC)

Our last quality control step was a congruence test (branch length comparison) based on the reasoning that non-orthologous sequences (whether contaminants, xenologs, or paralogs) generally exhibit very long branches when they are constrained to be mispositioned in the species tree. This protocol will only be performed on the 440-gene supermatrix (cf. **Figure 29**) in order to avoid further reducing the supermatrix derived from duplicated genes and to better assess the potential benefits of data treatment. This tree was then used as a reference to impose the topology onto the phylogeny of our individual genes to reveal species with long branches for certain genes. First, we will eliminate individual sequences considered problematic within a gene, and then, in the second step, we will eliminate genes with too many sequences whose branch lengths do not satisfy the Pearson correlation coefficient R^2. With this method, we reduced the number of genes from 440 to 416.

3.5 CHIMERA CREATION

We then reevaluated the completeness of our 416 genes. Some organisms were underrepresented in our MSAs. However, among these, several were closely related. We therefore created chimeras by in silico gene fusion so that the sequences complement each other, allowing for more complete chimeric organisms. This resulted in a reduction from 364 to 352 species by fusing the less represented closely related species in our 416 genes.
Among our 416 orthologous groups to a maximum of 352 species, we retained those containing at least one representative of Asgard, DPANN, Euryarchaeota, and TACK. We were left with 343 genes.

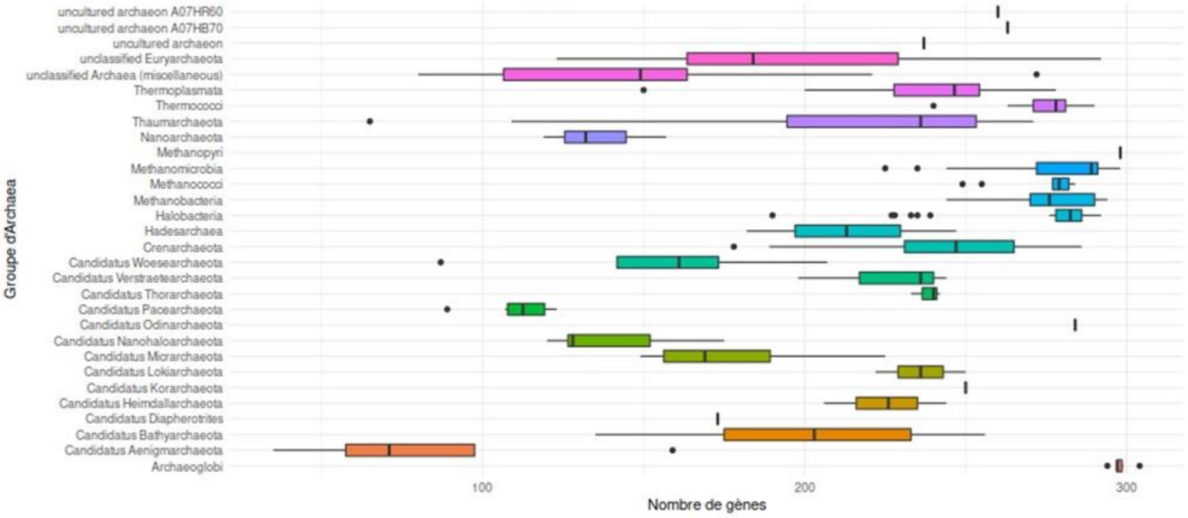


Figure 30. Distribution of archaeal groups within our selection of 343 MSAs.
Data are provided in **Supp.Mat distr-sp-genes.csv**.

The majority of our archaeal groups are well represented in our 343 MSAs (**Figure 30** & **Supp.Mat distr-sp-genes.csv**). However, there is an underrepresentation of *Candidatus Aenigmarchaeota* (present in consistently fewer than 100 genes, except for one). We therefore have 343 high-quality MSA of orthologous genes and 117 MSA of neo-orthologous genes of rougher quality. The analysis of our alignments reveals several points (**Figure 31, Figure 32, Figure 33, Supp.Mat stat-117-duplicates.csv & stat-343-genes.csv**). First, our MSAs are longer for neo-orthologous genes (the mode is around 300 amino acids after BMGE), unlike true orthologous genes, which are generally about twice as short (mode around 150 amino acids). However, these latter genes tend to retain more species than neo-orthologous genes (they have been less filtered). The gaps (indels and missing characters) have a similar distribution between orthologous and neo-orthologous genes. The gap proportion is around 10% for the vast majority of alignments (1st quartile = 7.7% and 3rd quartile = 14.8%).

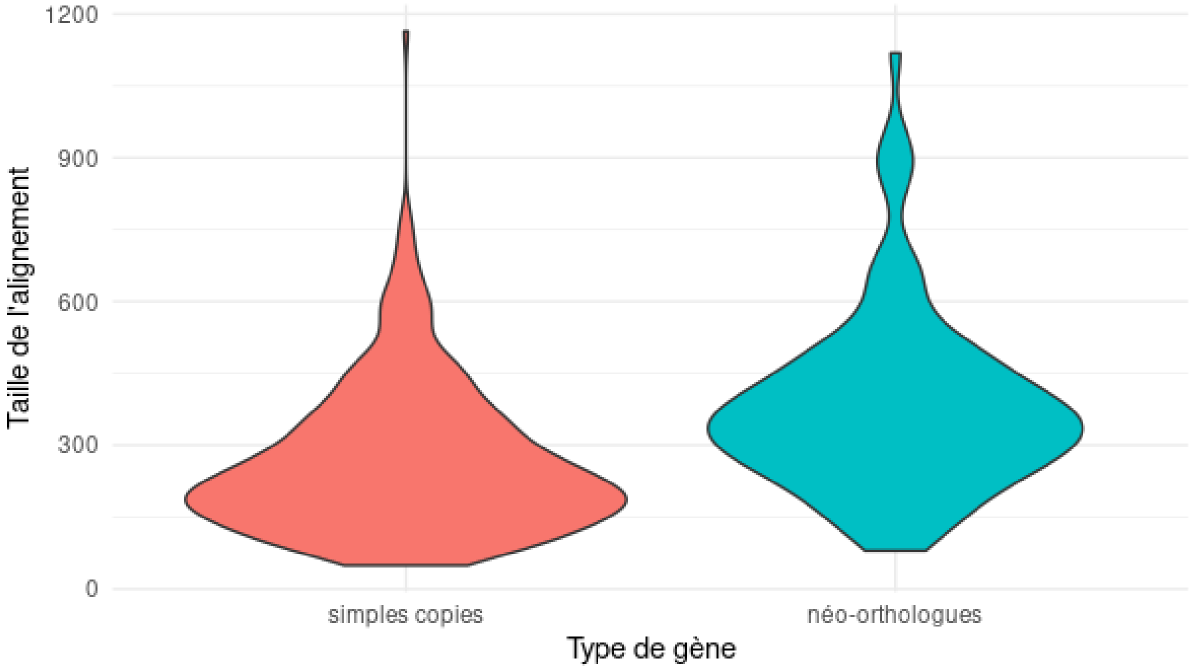


Figure 31. Violin plot of alignment size distribution as a function of gene type.
Data are provided in **Supp.Mat stat-117-duplicates.csv & stat-343-genes.csv.** Our single-copy gene MSAs tend to be much shorter (mode around 150 amino acids) than our neo-orthologous genes (mode above 300 amino acids).

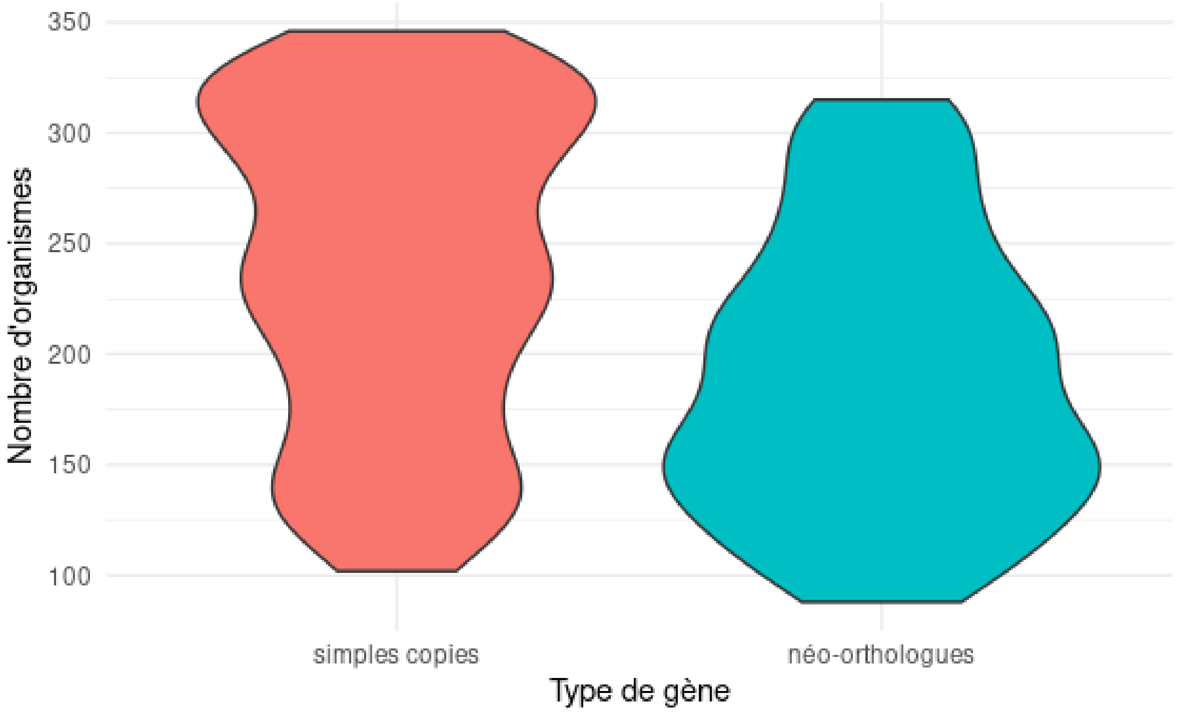


Figure 32. Violin plot of the distribution of the number of organisms retained within alignments.
Data are provided in **Supp.Mat stat-117-duplicates.csv & stat-343-genes.csv.** Our MSAs of single-last-copy genes tend to conserve more species than for neo-orthologous genes.

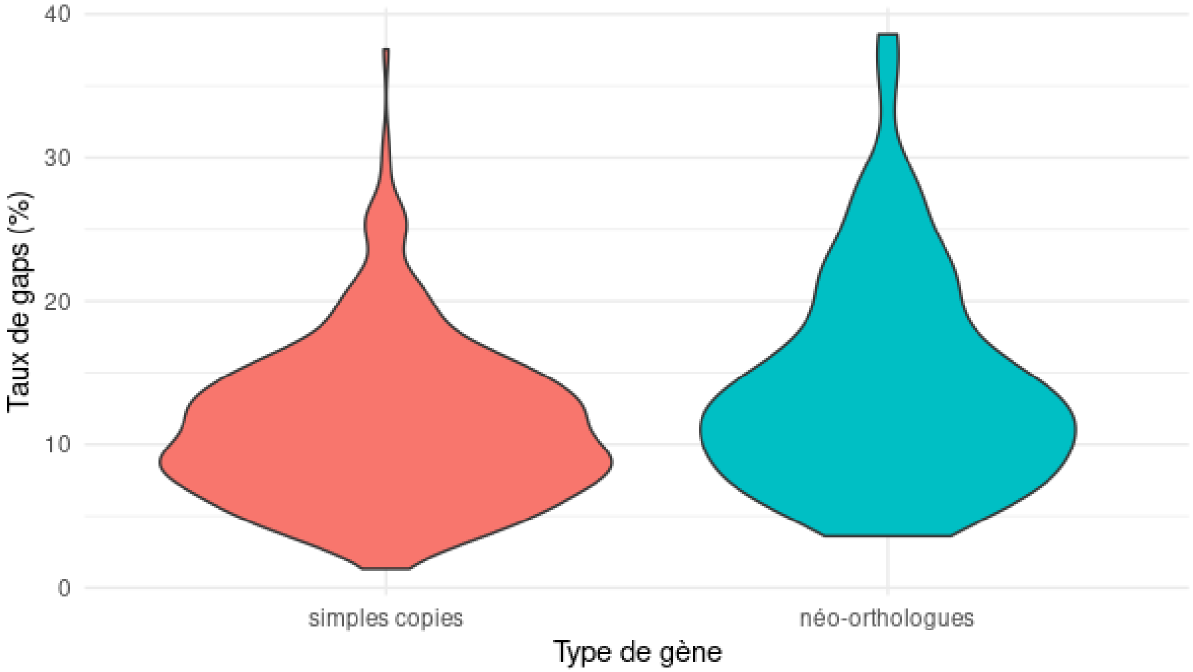


Figure 33. Violin plot of gap rate distribution within alignments.

Data are provided **Supp.Mat stat-117-duplicates.csv & stat-343-genes.csv.** The distribution of gaps is similar between our MSAs of single-copy and neo-orthologous genes. The proportion of gaps is around 10% for the vast majority of alignments (1st quartile = 7.7% and 3rd quartile = 14.8%).

4 PRELIMINARY ANALYSIS AND PHYLOGENETIC DISCUSSION OF THE OBTAINED TREES

Now that we have created our datasets, we will compute trees using sophisticated models. The trees were computed with IQ-TREE, with which we have some experience. This software meets our expectations in terms of speed (faster than RAxML used previously), a broader range of usable models, and reliability. Indeed, according to systematic tests conducted in our laboratory (M. Leleu), its heuristics for tree exploration are nearly as effective as those of RAxML. Our approach is a data-driven approach, meaning that we treated our datasets without any prior assumptions about the groups that were supposed to exist. We thus start without predefined taxonomy and will redetermine the groups ourselves. Our goal is to analyze our datasets as neutrally and objectively as possible, in order to compare them later to what is known in the literature. We computed a phylogenetic tree of 352 species with IQ-TREE (Minh et al., 2020) using the LG4X model with Ultrafast Bootstrap × 1000 for (1) our ribosomal supermatrix of 7,819 positions (cf. **Supp.Mat arbre-ribo-352-sp**), (2) our supermatrix of 94,485 positions representing our 343 genes (**arbre-352-sp-343-genes**) and (3) our supermatrix of 47,154 positions representing our 117 neo-orthologous genes from duplications.

4.1 RIBOSOMAL TREE

To verify that the creation of chimeric organisms does not affect the phylogeny of the organisms, we computed a new ribosomal tree (**Figure 34** & **Supp.Mat FigS34**) with IQ-TREE using the LG4X model with Ultrafast Bootstrap × 1000, on a supermatrix of 7,819 positions from our new selection of species (cf. **Supp.Mat FigS34**). We thus obtain the same tree updated with the chimeric species. No differences, however, are observed between the two trees.

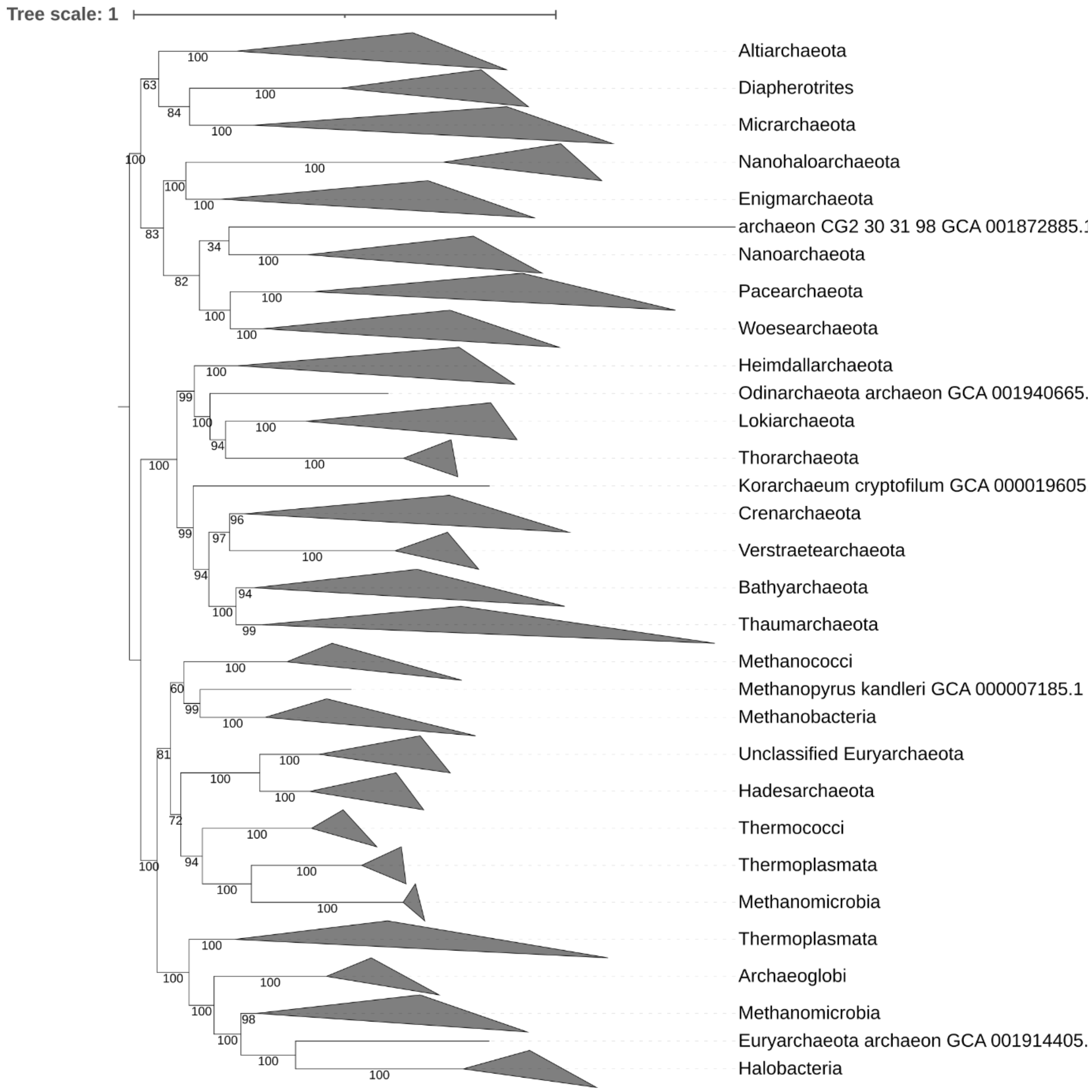


Figure 34. Phylogenetic tree of ribosomal proteins from 352 archaea (including 12 chimeras) calculated with IQ-TREE using the LG4X model and ultrafast bootstrap x 1000 on a super-matrix of 7,819 positions.

Data are provided in **Supp.Mat FigS34.** Reducing our taxonomic sampling to 352 archaea from our ribosomal tree does not alter our topology. However, we note an increase in statistical support to 72% for the Hadesarchaeota group with the (Thermococci + Methanomicrobia + Thermoplasmata = TTM).

4.2 SINGLE COPY 343-GENE TREE

From this selection of species, we also computed a new tree (**Figure 35** & **Supp.Mat FigS35**) including all of our 352 species and the 343 genes retained after the congruence test, using IQ-TREE with the LG4X model and Ultrafast Bootstrap × 1000 on a supermatrix of 94,485 positions (**arbre-352-sp-343-genes**).

We observe differences between our ribosomal tree and our tree with all orthologous genes. Within the *Euryarchaeota*, we find the same difference previously mentioned regarding the position of the *Hadesarchaea*. These are the sister group of *Methanobacteria* + *Methanococci* + *Methanopyrus kandleri* when using all genes, whereas they are pushed slightly higher in the tree when using only the ribosomal genes (**Figure 34**) and are placed as the sister group of *Thermococci* + *Methanomicrobia* + *Thermoplasmata*. Furthermore, we can also observe a notable difference between our 343-gene tree and our preliminary 440-gene tree. In our 343-gene tree (**Figure 35**), the group *Thermococci* + *Thermoplasmata* + *Methanomicrobia* is the sister group of *Hadesarchaeota* + (*Methanobacteria* + *Methanococci* + *Methanopyrus kandleri*) with an Ultrafast Bootstrap value of 91%. In contrast, in the 440-gene tree (**Figure 29**), the group *Hadesarchaeota* + (*Methanobacteria* + *Methanococci* + *Methanopyrus kandleri*) is the sister group of the group (*Thermoplasmata* + *Archaeoglobi* + *Methanomicrobia* + *Halobacteria*) with an Ultrafast Bootstrap of 46%. This demonstrates the impact of our gene selection based on the branch length comparison congruence test.

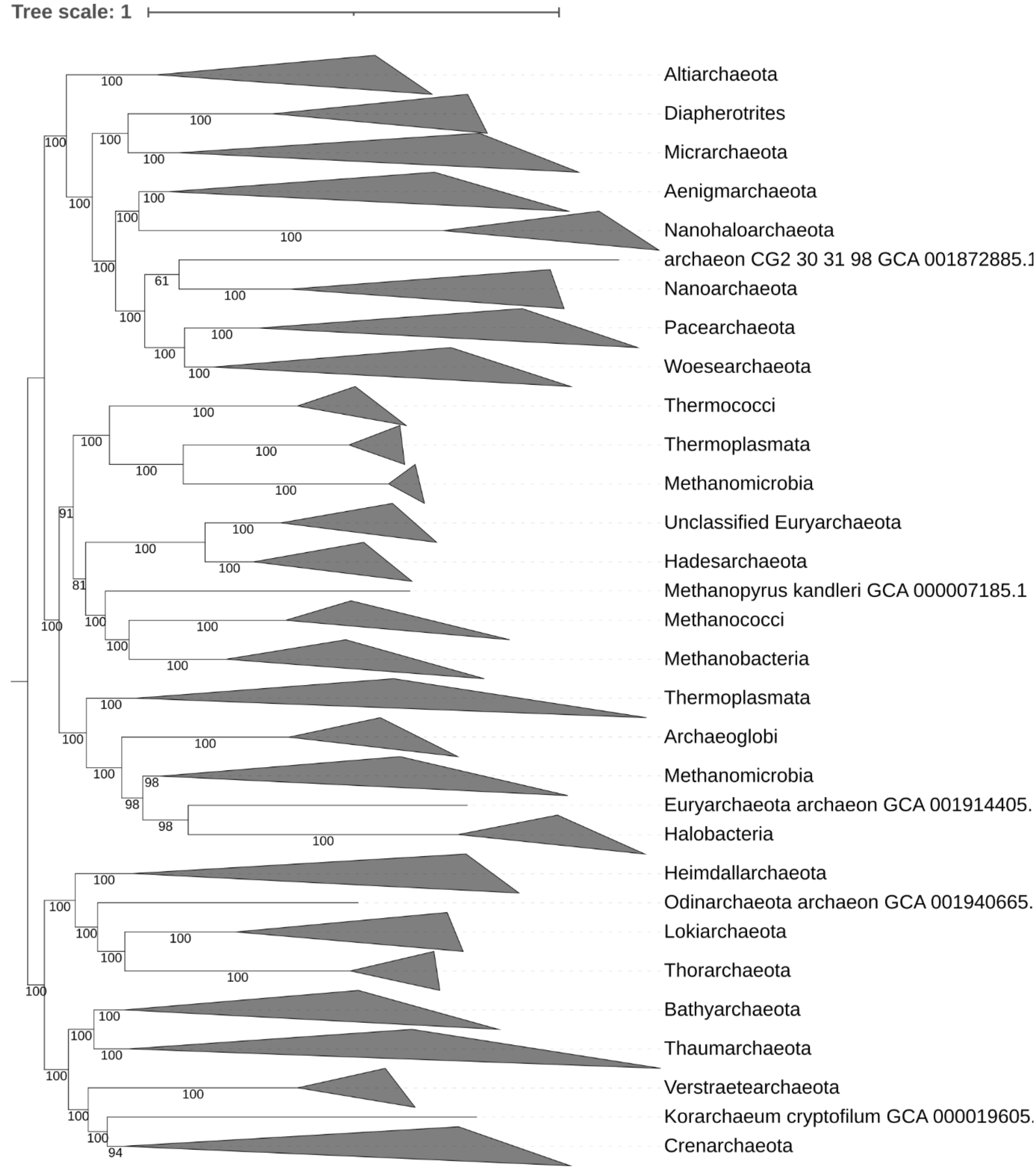


Figure 35. Phylogenetic tree of 343 genes from 352 archaea calculated with IQ-TREE according to the LG4X model and ultrafast bootstrap x 1000 on a super-matrix of 94,485 positions.
Data are provided in **Supp.Mat FigS35**. Reducing our taxonomic sampling to 352 archaea from our 343-gene tree does not alter our topology. However, we note an increase in statistical support to 81% for the Hadesarchaea group with the (Methanopyrus + Methanococci + Methanobacteria).

4.3 117 DUPLICATED GENES TREE (NEO-ORTHOLOGUES)

We also computed a tree from the 117 neo-orthologous genes resulting from duplications (47,154 positions) with the same selection of 352 species (the chimeric species were added with 42 on the individual gene alignments, then the genes were concatenated into a supermatrix) (**Figure 36** & **Supp.Mat FigS36**). The protocol is similar to the 343-gene tree, except that no filters were applied. In this tree, unlike the previous ones, and particularly the tree with the initial 440 genes, which it corresponds to for the neo-orthologous genes, the *Hadesarchaea* are pushed higher in the tree to the base of all *Euryarchaeota* with an Ultrafast Bootstrap of 99%. Additionally, the *Thaumarchaeota* appear as polyphyletic, with part of them inserted between the initial group *Crenarchaeota* + *Candidatus Verstraetearchaeota*.

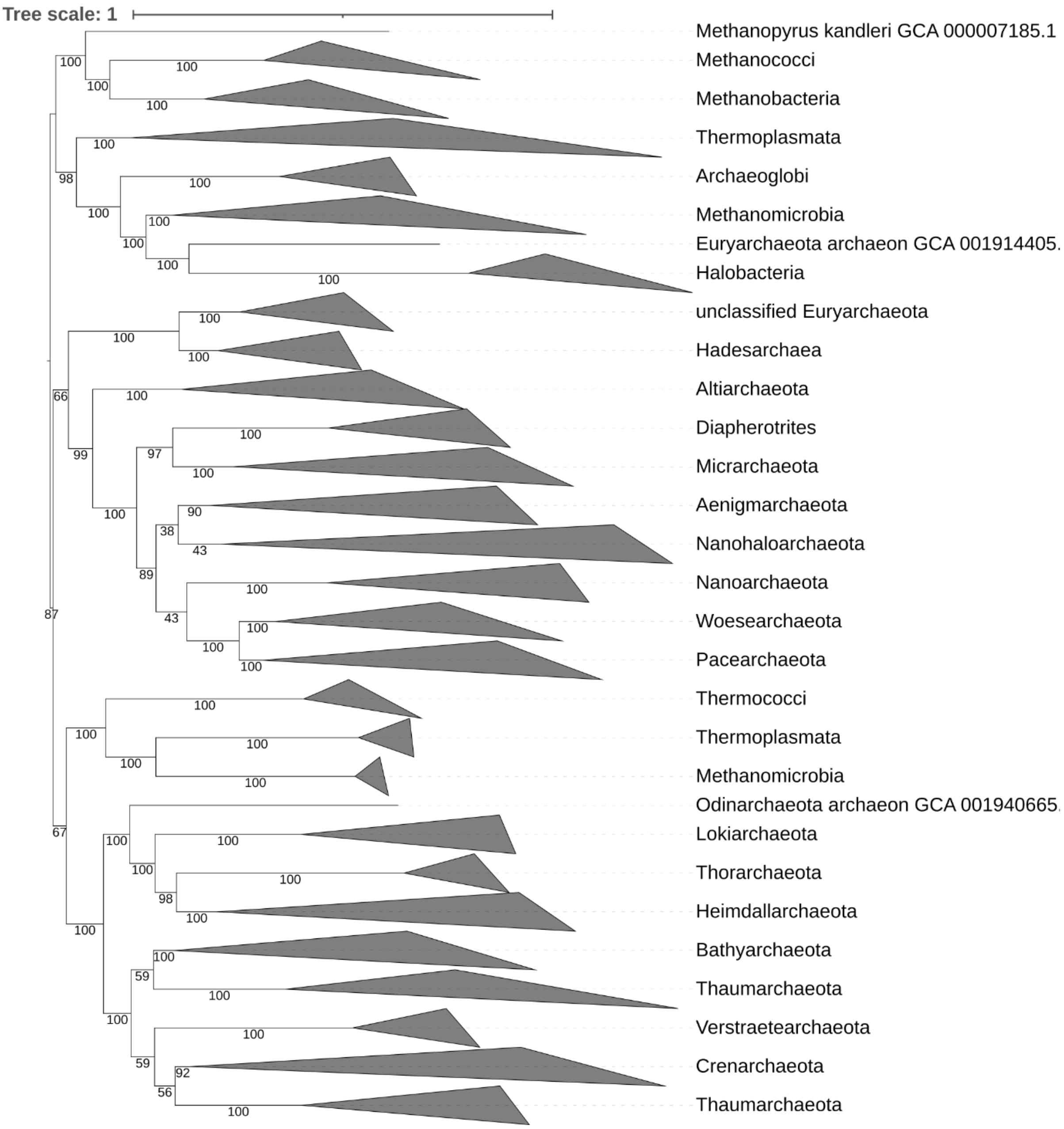


Figure 36. Phylogenetic tree of 117 neo-orthologous genes from 352 archaea calculated with IQ-TREE according to the LG4X model and ultrafast bootstrap x 1000 on a super-matrix of 47,154 positions.

Data are provided in **Supp.Mat FigS36**. Hadesarchaea are pushed higher up the tree at the base of all Euryarchaeota with a very faut ultrafast-bootstrap of 99%. In addition, Thaumarchaeota appear polyphyletic, with one part fitting between the initial Crenarchaeota + Candidatus Verstraetearchaeota group.

We have just obtained an initial collection of trees in which we can identify some initial incongruences. Our goal is to try to understand where these incongruences come from in order to determine, if possible, the correct topology. We will now exploit our dataset using various methods. We will apply different taxonomic variations and then subject them to the same battery of phylogenetic tests: gene jackknife, use of various evolutionary models, all while comparing the supermatrix and supertree methods.

5 SPECIES JACKKNIFE / TAXONOMIC VARIATION SAMPLING

First, we want to assess the impact of taxonomic sampling on the phylogeny of archaea. For this, we plan to perform a species jackknife (taxonomic sampling variation) to carry out our analyses (**Box 10**). Our goal is to replace species to detect potential artifacts that could be due to species with different evolutionary rates, which may lead, in particular, to long-branch attraction phenomena. To do this, we will use our ribosomal protein tree to define monophyletic groups representing the maximum diversity of archaea, with sufficient granularity to uncover higher-order phylogenetic relationships. These groups must all be supported with a bootstrap of 100% in our ribosomal tree. Thus, we obtain 70 groups within which we subsequently performed the species jackknife. The species corresponding to our 70 groups are provided in the file **70-groupes.xlsx**.

Box 10. Testing the Robustness of a Tree through Resampling: Bootstrap and Jackknife

Resampling is a statistical technique in which a procedure (such as constructing a phylogenetic tree) is repeated on a series of datasets. The results of the analysis of the resampled datasets are then combined to generate summary information about the original dataset.

In the context of phylogenetic tree construction, resampling involves generating a series of sequence alignments by sampling columns from the original sequence alignment. Each of these alignments (called pseudo-replicas) is then used to calculate an individual phylogenetic tree. A consensus tree can finally be constructed by combining the information from all the generated trees. The topologies produced can also be sorted by the frequency of occurrence of their constituent bipartitions.

The general principle is as follows:

1. Generate k new alignments by randomly removing genes or columns.
2. Recalculate a new tree for each of the k alignments.
3. Compute a consensus tree with the frequencies of bipartitions for each of the internal nodes of the tree.

The consensus tree can be calculated in different ways:

- Strict consensus: The clades that appear in all trees.
- Majority rule: The consensus tree contains only the branches that are present in at least half of the individual trees. This tree summarizes the information from all the trees resulting from the jackknife without a reference tree, identifying shared information compatible with the majority of trees. However, the resulting tree can still contain multifurcations.
- Greedy majority-rule procedure (extended majority-rule): This method gradually adds branches that appear in less than half of the initial trees until the tree is fully resolved.

In general, the consensus tree has no branch lengths. It has been shown that this greedy method provides better results than the majority rule alone. The confidence value of each node in a consensus tree (i.e., the robustness of a bipartition) is then defined as the number of replicas (usually expressed as a percentage) in which the node appears. Typically, for single-gene phylogenies, branches with a support greater than 75% are considered reliable. In phylogenomics, however, support is considered reliable from 99-100%.

Bootstrap

The bootstrap is a resampling method with replacement from an initial sample. To apply bootstrap in tree construction, each pseudo-replica is built by randomly sampling columns from the original alignment with replacement until an alignment of the desired size is obtained. The bootstrap procedure generally starts by creating new alignments by randomly selecting columns from the original alignment and reconstructing an independent tree for each new alignment. A consensus tree is then constructed to summarize the results of all the tree replicas. Special attention will be given to branches with lower values.

Jackknife

One problem when performing phylogenomics using probabilistic methods and complex models is the computational power required to obtain reliable statistical support. For large datasets, computation times can be long or even impractical. In such cases, traditional bootstrap calculations are almost unfeasible. It is necessary to find ways to overcome these logistical challenges, such as using subsampling methods. A resampling method for genes called *jackknife* (or "Swiss knife") reduces the alignments into several smaller subsamples of manageable size to assess the quality of the phylogenies. The jackknife is very similar to bootstrap: the only difference is that pseudo-replicated matrices are generated by eliminating K positions from the original matrix. It has been shown that a jackknife rate of 40% should ideally be used (i.e., keeping 60% of the data).

References

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.

We performed 5 species selections within our 70 ribosomal groups, with jackknives representing ± 1/3 of our 352 species, i.e., 5 × 121 species. For some of our groups with too few representatives, it was impossible to perform jackknife (e.g., Korarchaeota represented by a single species, which will therefore be present in all our replicates). Conversely, some groups are over-represented (e.g., Halobacteria: 53 genomes), so to exploit all the species, we kept a maximum of 5 representatives per group. In our case, with 5 jackknives of 5 species per group, we can use a maximum of 25 species from the group. If any of our groups contains fewer than 25 species, we will repeat the sampling process. To determine which species to keep, we selected species based on the criterion of proteome completeness, meaning their presence in the largest number of genes. Ultimately, 303 species out of 352 are distributed across our replicates, with the remaining 49 species belonging to over-represented groups (e.g., Halobacteria) and therefore unable to pass our filters for inclusion in the replicates.

6 GENE JACKKNIFE & TREE CALCULATION

We either concatenated our genes into super-matrices or calculated individual trees for each gene corresponding to our 5 species jackknives to construct a super-tree (**Figure 37**).

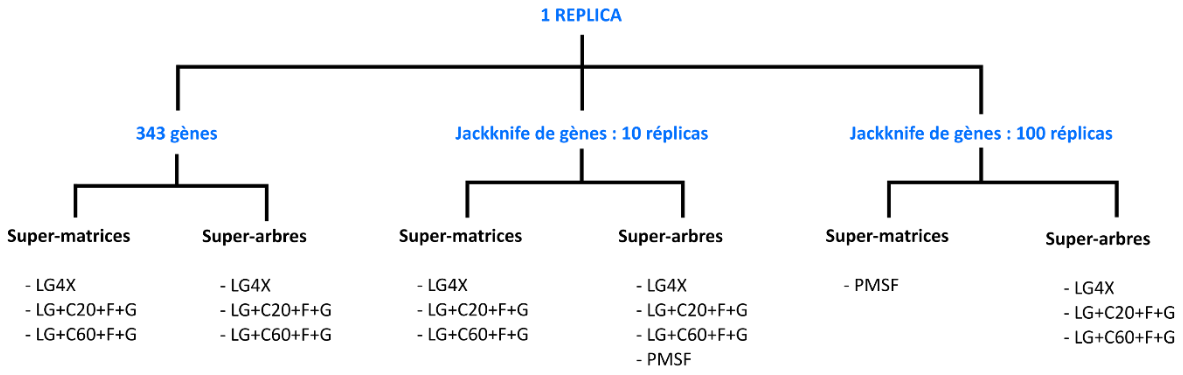


Figure 37. Overview of all trees calculated for a species jackknife replica.

For one replica, we proceeded with super-matrix and super-tree approaches, initially with all 343 genes, a 10-replicate gene jackknife and a 100-replicate gene jackknife. The models used are LG4X, LG+ C20+F+G and LG+C60+F+G and PMSF.

We then sought to put ourselves in difficult conditions to verify whether we obtain the same results as with standard phylogenies. To do this, we combined our species jackknife with a stringent gene jackknife. This method tends to challenge the robustness of trees by revealing various problems and incongruences between different datasets. We will thus test the robustness of our trees (**Box 10**).

For each dataset, we calculated trees using the LG4X, LG+ C20+F+G, and LG+C60+F+G models. For each of the 3 previous models, we applied an initial 10-replica gene jackknife, i.e., 5 species replicates × 10 gene replicates = 50 super-matrices × 3 models = 150 trees (cf. **Figure 37**). We also employed the PMSF method. This method is considered the best for combating phylogenetic reconstruction artifacts (**Box 5**). Finally, we extended the gene jackknife to 100 replicates in order to calculate trees using the PMSF method. We generated super-matrix sizes of approximately 35,000 positions, which corresponds to one-third of the size of the super-matrices from the 343 genes and the size of the super-matrices of the 117 neo-orthologous genes, with which we wish to compare them. To optimize our results, we used as a reference guide tree the one from our previous IQ-TREE using the LG+C60+G4+F model, for a total of 5 × 100 = 500 super-matrices.

Finally, we analyzed these same 100 gene jackknife replicates using the super-tree method. The consensus trees of each replicate (5 × 100 = 500) were calculated with ASTRAL-III from the individual gene trees according to the different LG4X, LG+C20+F+G, and LG+C60+F+G models.

We also compared the results obtained with our dataset from the neo-orthologous genes. Super-matrices with the same species replicates were calculated as before in LG4X, LG+C20+F+G, LG+C60+F+G, and PMSF (to optimize the result, the guide tree used was the same as the one used with the initial orthologous genes).

7 ANALYSIS OF ROBINSON-FOULDS DISTANCES

We performed our tree analyses using the Robinson-Foulds method (**Box 11**). This method allows us to evaluate the factors affecting tree topology.

Box 11. Topological Comparison: The Robinson-Foulds (RF) Distance

One way to compare a large number of phylogenetic trees is to measure the difference between each pair of trees. The Robinson-Foulds (RF) distance is the most widely used measure for this purpose. The topological Robinson-Foulds distance between two phylogenetic trees is equal to the minimum number of elementary node fusion and separation operations required to transform one tree into the other. In other words, it counts the number of different bipartitions between the two trees.

A topological distance d(T1, T2) between two unrooted phylogenetic trees defined on the same set of n taxa is a measure of dissimilarity between the respective topologies of T1 et T2. If d(T1, T2) = 0, then the two trees are identical. Since each internal branch of a phylogenetic tree induces a bipartition of its leaf set, the bipartition distance dRF measures the number of bipartitions induced by one tree but not the other. It is defined as (A + B) where A is the number of partitions of data involved by the first tree but not the second, and B is the number of partitions of data involved by the second tree but not the first.

For an unrooted phylogenetic tree with at most n - 3 internal branches, the dRF distance is commonly normalized by 2(n - 3) to bring it into the range [0, 1]. A value of 0 indicates that the trees are strictly identical, while a value of 1 indicates that they are completely different.

References

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 ROBINSON-FOULDS DISTANCES OF SUPERMATRICES

Initially, we aimed to assess the impact of the model on our trees by comparing, for each jackknife replica, the RF distances between our different models taken pairwise. Each boxplot (**Figure 38**) represents the comparison of 10 pairs of trees from 10 super-matrices. While the C20 and C60 models tend to yield similar results (nRF = 0.025-0.030) regardless of the species replica, the LG4X model tends to give more divergent results (although this difference remains small: nRF = 0.050-0.090). Additionally, we observe for replica 2 a very high variability when comparing the LG4X model with the C20 and C60 models. It is also noticeable that the difference between replica 1 and the other replicas is much more pronounced when comparing the LG4X model with the C20 and C60 models. However, when comparing the C20 and C60 models to each other, this variability is much lower, except for replica 4, where it is significant, indicating a greater number of topologies identified.

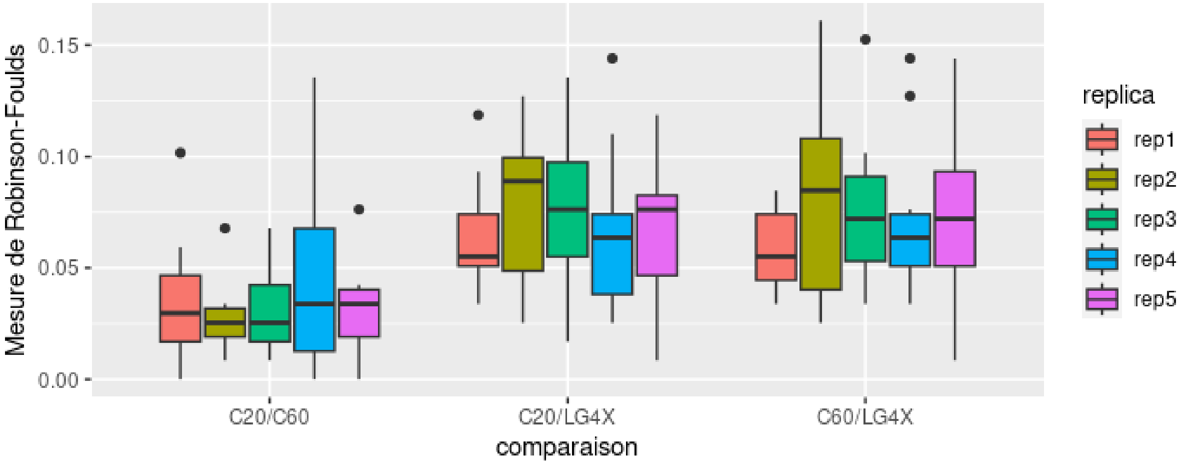


Figure 38. Model-to-model comparison of Robinson-Foulds topological distances of super-arrays for each replica.

Data are provided in **Supp.Mat Robinson-Foulds**. While the C20 and C60 models tend to give similar results (nRF = 0.025-0.030) regardless of species replica, the LG4X model tends to give more different results (although this difference remains small: nRF = 0.050-0.090).

We then assessed the impact of our gene selection on our trees (**Figure 39**). For a given replica and model, we have 10 gene replicas, meaning (10^2-10) / 2 = 45 comparisons per boxplot. When ranking the RF distances of our three models based on their median rank (1, 2, 3, or 1, 2 in case of ties) across the 5 replicas, we observe that, with the exception of replica 3, the LG4X model tends to generate higher RF distances (from 0.15 to 0.20), suggesting that it is more sensitive to gene sampling. On the other hand, the C20 model appears to be the most stable.

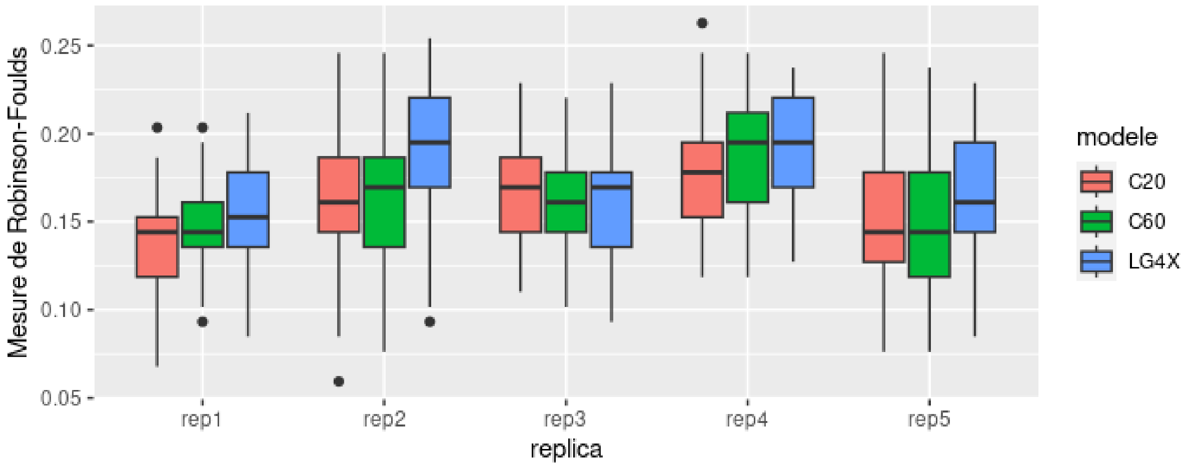


Figure 39. Robinson-Foulds topological distances of supermatrices according to LG4X, C20 and C60 models.

Data are provided in **Supp.Mat Robinson-Foulds**. The LG4X model tends to generate higher RF distances (from 0.15 to 0.20), suggesting that it is more sensitive to gene sampling.

We assessed the impact of our gene selection on the 5 × 100 trees calculated using the PMSF method (**Figure 40**). Each boxplot thus contains (100^2-100) / 2 = 4950 points. PMSF is even less sensitive to the gene replica (with a distance of 0.1). Two interpretations can be considered: either the fact that providing a topology with a guide tree induces the result and thus reduces the variability of the obtained trees, or the PMSF method refines the result by combating what could be considered as artifacts (Wang et al., 2018).

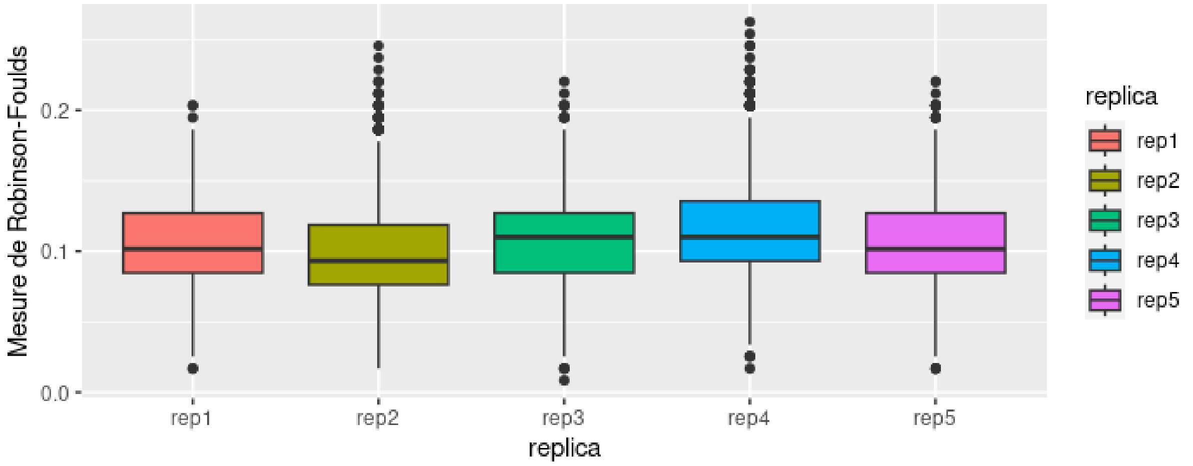


Figure 40. Robinson-Foulds topological distance of super-matrices using the PMSF method.

Data are provided in **Supp.Mat Robinson-Foulds**. PMSF is very insensitive to gene replication (0.1 distance). This may mean that either providing a guide tree induces the result, reducing the variability of the trees obtained, or that the PMSF method proves very effective in combating artifacts and finding the right result.

7.2 ROBINSON-FOULDS DISTANCES OF SUPERTREES

To compare the models, we have 100 gene replicas per boxplot here (compared to 10 for the super-matrices) (**Figure 41**). The results are essentially the same as with the super-matrices. While the C20 and C60 models tend to give similar results (0.075-0.1) regardless of the species replica, the LG4X model tends to give more divergent results (although this difference remains small: 0.10 to 0.14). These values are also slightly higher than those with the super-matrices. The variability between models is also less pronounced.

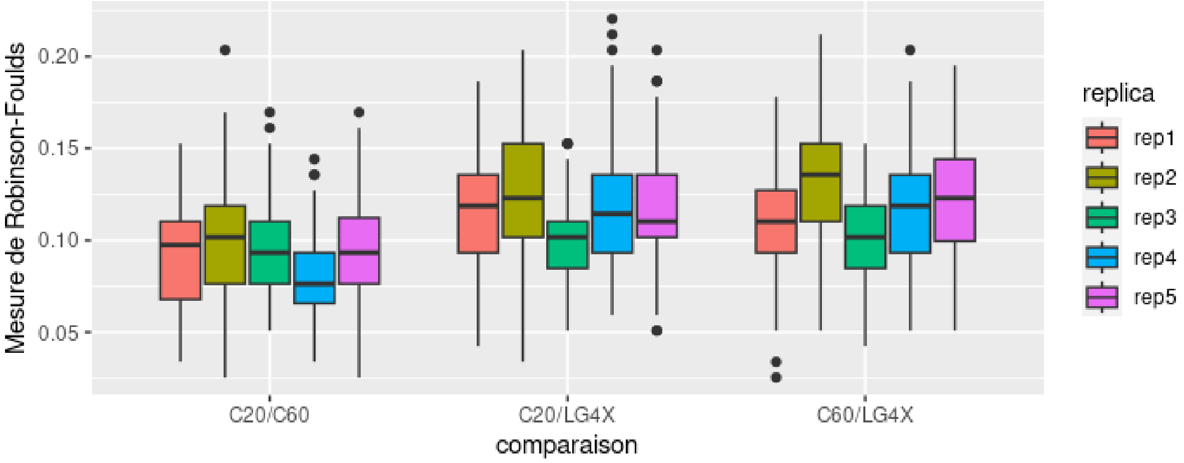


Figure 41. Model-to-model comparison of topological Robinsons-Foulds distances of supertrees for each replica.

Data are provided in **Supp.Mat Robinson-Foulds**. While the C20 and C60 models tend to give similar results (nRF = 0.075-0.1) regardless of species replica, the LG4X model tends to give more different results (although this difference remains small.

Regarding the influence of gene sampling, we have 100 trees per replica (i.e., 4950 points per boxplot) (**Figure 42**). A more significant difference in means is observed between the LG4X model and the other models than between the C20 and C60 models. Additionally, replicas 2 and 4 appear to be more unstable depending on the model used, with differences of 0.03 between the highest and lowest medians for each replica, while the means are the same for each model in replicas 1, 3, and 5.

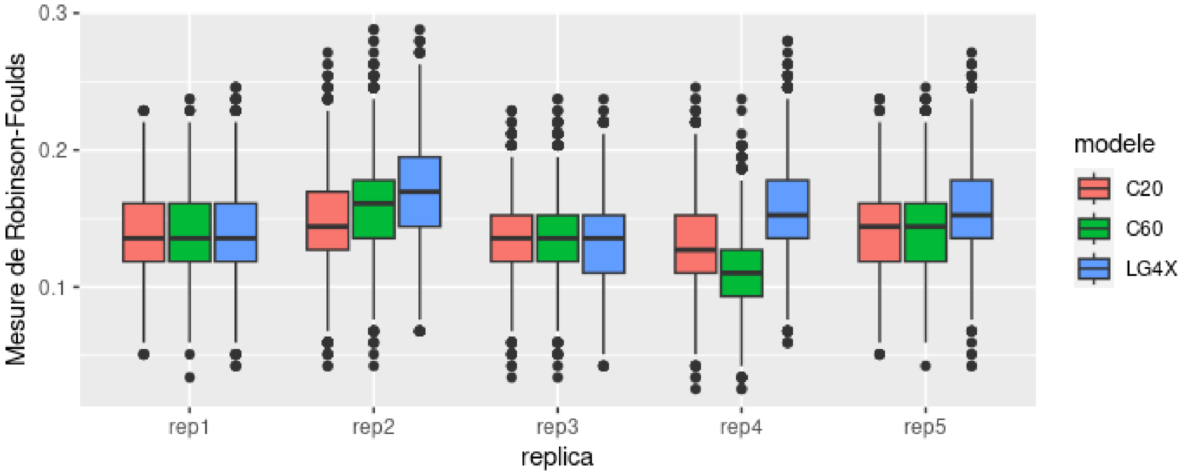


Figure 42. Robinson-Foulds topological distances of supertrees according to LG4X, C20 and C60 models.

Data are provided in **Supp.Mat Robinson-Foulds**. Means differ more between LG4X and the other models than between C20 and C60. Replicas 2 and 4 appear to be more unstable, depending on the model used.

The results from the RF analyses show that the LG4X model is more sensitive to variations in gene sampling than the other models. Additionally, super-trees are less stable than super-matrices. Finally, the species replicas do not all behave the same way, which underscores the importance of having used 5 different replicas. Thus, the choice of model seems to be the most important parameter to consider. However, the RF distances do not allow us to qualitatively assess these differences. We are currently unable to determine whether these differences stem from minor terminal groups, larger groups that move within the tree, or a combination of both phenomena.

8 SEARCH FOR MAJORITY BIPARTITIONS

Until now, we have measured the differences between our trees without examining where they came from. To do this, one would ideally need to construct all possible trees. Unfortunately, detecting the most likely tree among all possible trees is an NP-hard problem, meaning that no known algorithm can solve this problem in a reasonable amount of time. Indeed, the number of possible topologies for an unrooted phylogenetic tree with t species increases exponentially as t increases, and it can be calculated using the following formula:

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

The total number of possible trees would then need to be multiplied by all possible branch lengths to obtain the total number of possible trees. As a result, it is impossible to test each possible tree.

Since it is not feasible to consider all possible topologies, we focused on identifying all the nodes encountered in our multiple trees in order to establish a list of all possible clades (unrooted clades). This work involved the individual analysis of all the trees to identify all the possible clades (cf. **Supp.Mat archaea.clan**), which were then compared across trees, with the 5 replicas not containing the same species. From this collection of clades, we sought to identify those that were consistently found and those that were more unstable. To do this, we expressed them in terms of successive groupings of our initial 70 ribosomal groups, only to realize that some trees appear to present alternative hypotheses (recurrent combinations of clades) (cf. **Supp.Mat 70groupe.xlsx**). To understand the source of these differences, we focused our attention on the nodes of the trees that appeared to be the most unstable. We examined the incongruences starting from a classification scale at least larger than our 70 groups, deliberately excluding phylogenies within these groups. A list of the unstable nodes is provided in the table **Supp.Mat noeuds-à-problèmes.xlsx**.

These instabilities can arise from different causes. We distinguish five cases:

- Impact of the species replica: Does the choice of species affect the observed topology, regardless of the method used? In this case, the results obtained are influenced by taxonomic sampling, and various intrinsic problems with the genomes used (gene transfer, contamination, etc.) may be suspected.
- Impact of the model used: Are all replicas in agreement but tell different stories depending on the model employed? In this case, it is very likely that some models are sensitive to phylogenetic reconstruction artifacts affecting the (short) branches representing relationships between higher-level groups.
- Impact of the method used: Do super-trees give different results from super-matrices? This could point to a lack of signal from the gene simple trees that form the basis of the super-trees (stochastic error).
- Impact of certain genes (e.g., HGT at the level of a group)?
- Combined impact of the species replica and the model or method employed.

8.1 GROUPING OF RIBOSOMAL GROUPS INTO HIGHER-LEVEL GROUPS COMMONLY ACCEPTED IN THE LITERATURE

Until now, we have used a data-driven approach, meaning that we have analyzed our datasets without any prior assumptions. We will now compare our results with what is known in the literature. In line with the literature, the successive grouping of our 70 initial ribosomal groups allows us to confirm the monophyly of many groups already described. The analysis of our 70 initial ribosomal groups shows that they remain systematically monophyletic (bootstrap = 100) regardless of the replica, model used (even under difficult conditions such as jackknife X 100 PMSF), or method employed. This allows us to define our hypotheses based on our 70 ribosomal groups, which reduces the number of monophyletic groups to 27, which are consistently valid, often beyond our ribosomal groups. These groups are listed in **Tableau 6** and are named according to the literature. From now on, these will serve as OTUs (Operational Taxonomic Units) in our phylogenetic trees.

Groups	Correspondence with our 70 groups
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. Grouping of our 70 ribosomal groups, which will later serve as our OTUs.

8.2 GROUPS RARELY FOUND

We searched within our bipartitions for nodes exhibiting instabilities. We considered nodes as unstable if their BP value was lower than 75% (Shi et al., 2010) in at least one combination of replica and model (LG4X, C20, or C60), whether using the super-matrix or super-tree method (cf. **Supp.Mat bilan-supermatrics.xlsx**, **bilan-superarbres.xlsx** & **bilan-duplicates.xlsx**). The problematic nodes we identified are provided in **Supp.Mat noeuds-à-problèmes.xlsx**. We will categorize these issues as minor or major, based on the context of this thesis, specifically in the context of a global phylogeny of the Archaea in relation to the origin of eukaryotes.

8.2.1 MINOR PROBLEMS

Among our 8 initial Bathyarchaeota groups, many alternative topologies were found. The relationship between the Bathyarchaeota and Thaumarchaeota is well supported and has already been described (Adam et al., 2017; Brochier-Armanet et al., 2011). The phylogeny within the Bathyarchaeota has also been thoroughly studied in the literature and will not be the focus of this study.

Regarding the Euryarchaeota, the monophyly of our two groups *Thermoplasmata_unknown_2* and *Thermoplasmata_unknown_3* is well supported with the super-matrices. However, it is less consistent with the super-trees, where ultrafast-bootstrap values vary between 6 and 63%. The group *Thermoplasmata_unknown_1* then disrupts this monophyly instead of being at their root.

Within the *Stenosarchaea* (cf. **Tableau 6**), the monophyly *Methanomicrobia Methanosarcinales* + *Methanomicrobia_Methanosaetaceae* is never found when using super-trees, while it is always found with super-matrices. This issue is due to *Methanonatronarchaeia Euryarchaeota archaeon GCA001914405.1*, which tends to be attracted to the *Halobacteria* in our replicas 2, 4, and 5 (ultrafast-bootstrap < 72%), preventing the monophyly of *Stenosarchaea*. We will later refer to the monophyletic group *Stenosarchaea* + *Archaeoglobi* + *Natronarchaea* as SAN.

The *Methanobacteria* group, always found in super-matrices, is less supported in super-trees (statistical support between 30 and 84%, with no model appearing to influence this value).

Regarding the *TACK* group, depending on the replicas, either *Fervidicoccales* or *Desulfurococcales Pyrodictiaceae* is found as the sister group of *Acidilobales* and *Aeropyrum*.

The monophyly *Thorarchaeota-Lokiarchaeota* is well supported in super-matrices, with maximum ultrafast-bootstrap values (although slightly challenged with a jackknife of genes x100, with values between 76 and 83% depending on the replica). However, with super-trees, the supports are weaker, varying between 42 and 72%.

In PMSF, *Heimdallarchaeota archaeon GCA 001940645.1* is problematic with the super-matrix method. The monophyly of *Heimdallarchaeota* is undermined due to this species, which inserts itself at the root of the entire *Asgard* group. This problem also very slightly appeared with the super-tree method.

8.2.2 MAJOR PROBLEMS

The monophyly of Euryarchaeota + Crenarchaeota + Bathyarchaeota is well supported with super-matrices; however, it may appear less systematic with the use of super-trees (statistical support ranging from 58 to 94%, with no influence from the model or the replica).

Among the Euryarchaeota, the monophyly of *Thermoplasmata + SAN + Methanobacteria* is well supported in super-matrices for replicas 1 to 4 (ultrafast-bootstrap ranging from 62 to 100%), except for replica 5 due to the *Altiarchaeota*, which insert themselves as the sister group to the *Methanobacteria* rather than being at the base of the *DPANN* group (ultrafast-bootstrap ranging from 1 to 46%). However, in super-trees, the support seems to fluctuate between the two alternative positions for *Altiarchaeales* (ultrafast-bootstrap ranging from 31 to 76% for a monophyly of the group). The same is true for the *Hadesarchaea* group, which oscillates between being the sister group of *Euryarchaeota + DPANN* and being the sister group of *Theionarchaea + Thermococci* (**Figure 44**). The ultrafast-bootstrap ranges from 34 to 75% for the super-trees, with no influence from the model or the replica. However, for super-matrices, the ultrafast-bootstrap ranges from 62 to 100% for replicas 1, 2, and 3. For replica 4, the support is very weak, ranging from 0 to 24%. The values for replica 5 range from 35 to 95%, and from 18 to 75% for the PMSF model. It is noteworthy that the bootstrap value when including our 354 species with an LG4X model (cf. preliminary pre-jackknife trees) drops to 19% for the hypothesis of *Hadesarchaea + Theionarchaea + Thermococci*. The position of the *Hadesarchaea* is thus a source of uncertainty with our dataset.

The *Crenarchaeota* form a perfectly supported group. However, within it, the position of *Korarchaeota cryptofylum* varies depending on the model and method used. In super-matrices, *Korarchaeota* tend to be found preferentially as the sister group to the *Crenarchaeota*. For replica 1, there is a significant decrease in the ultrafast-bootstrap values when using the C60 and PMSF models (respectively 34 and 5%), in favor of the *SCGC* group as the sister group to the *Crenarchaeota* (66 and 95%). Furthermore, the super-matrix containing the 354 species has a low statistical support (6%) for this hypothesis. In addition, during the jackknife of genes × 100 with our 5 replicas, this hypothesis is nearly equal to the alternative hypothesis of a position for the *Korarchaeota* at the base of the *Crenarchaeota* (ultrafast-bootstrap ranging from 30 to 69%). It is this latter hypothesis that is systematically found with the super-tree method.

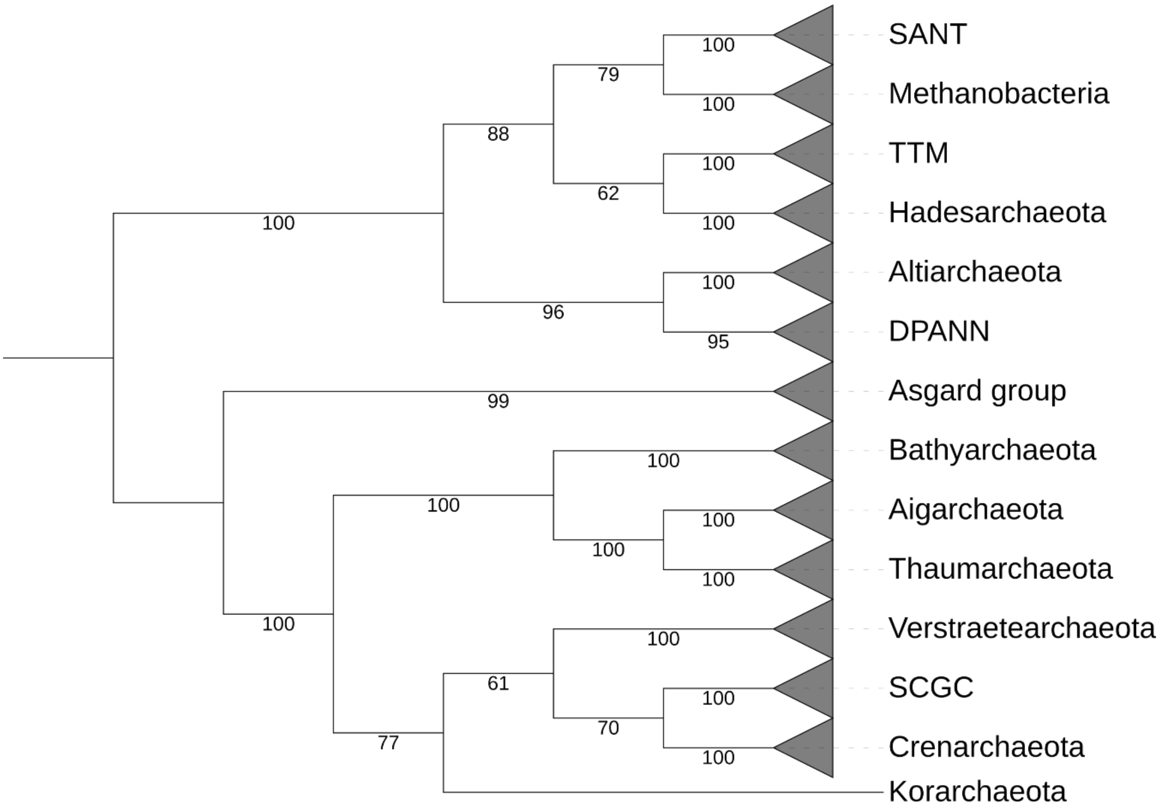
8.3 TOPOLOGIES POSSIBLES

In the context of our study, we focus on a selection of relatively high-level problems. We have just explored the bipartitions. Some are stable, others less so. However, very low taxonomic ranks are not the subject of our study. Indeed, we are working on a large scale, and our goal will be to include eukaryotes in order to verify their phylogenetic relationships with the archaea. For this, we will need the most stable trees possible, with groups whose validity and phylogenetic position we are certain of. This is why we proceed by grouping higher-level groups for which we can certify monophyly, to observe the differences. The list of the groups selected, which we consider robust, is as follows: Aigarcheota, Altiarchaeales, Asgard, Bathyarchaeota, Crenarchaeota, DPANN, Hadesarchaeota, Korarchaeota, Methanobacteria, SANT (Stenosarchaea + Archaeoglobi + Natronoarchaeota + Thermoplasmatales), SCGC, Thaumarchaeota, TTM (Thermococci + Theionarchaea (grouping Thermoplasmatales different from "true Thermoplasmatales") + Methanomicrobia_Arc (/!\ different from the Methanomicrobia of the SANT group)), Verstraetearchaeota. The SCGC, TTM, and SANT groups were defined based on our results. The other groups are already named as such in the literature.

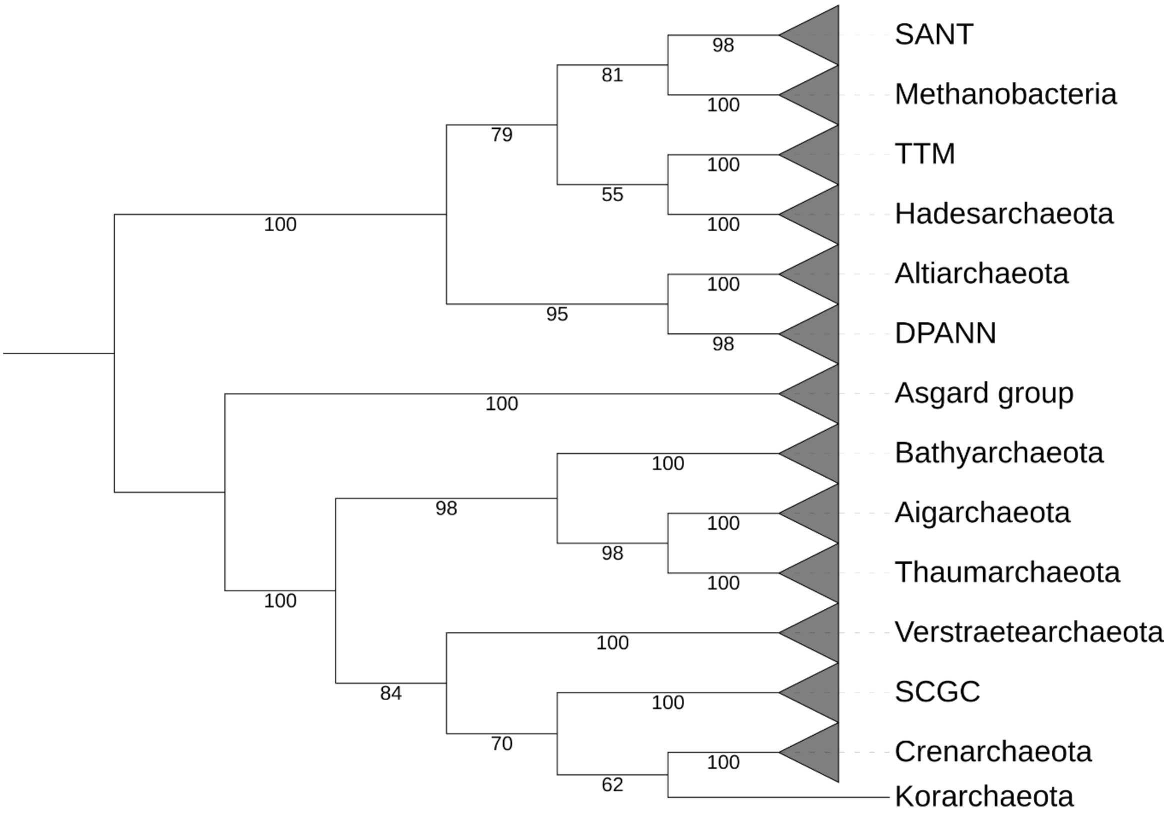
It should be noted that species called "thermoplasmatales" are present in two groups: the "true" Thermoplasmata and those forming the Theionarchaea group, which can make this term ambiguous. To date, these genomes have been properly renamed Theionarchaea in NCBI. Similarly, the Methanomicrobia called Arc in our TTM group are not related to the true Methanomicrobia found in our SANT group, but to a group recently named *Candidatus Methanofastidiosa*.

Based on these groups, the consensus trees for our 5 species replicas obtained after a gene jackknife for 100 replicas by super-matrices calculated using the PMSF method are given in **Figure 43**. The rooting of these trees by the midpoint method reveals the two main historical groups, Euryarchaeota and TACK.

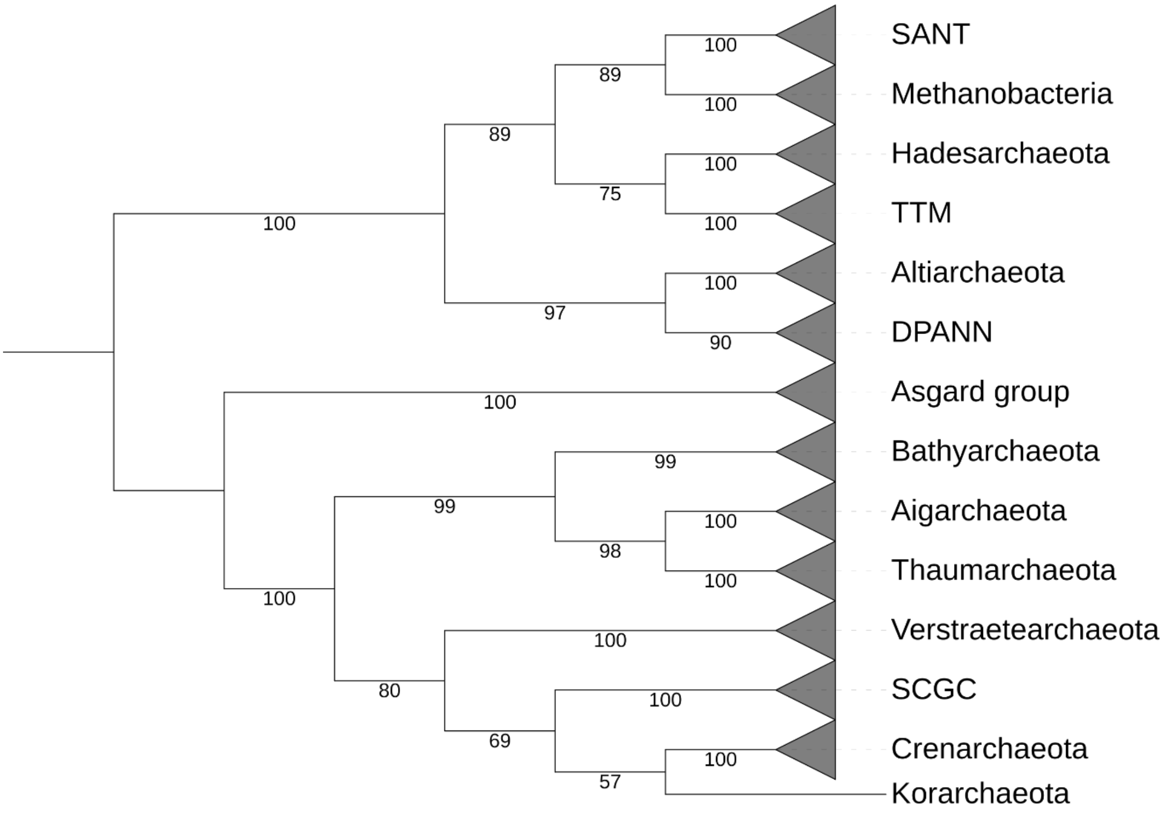
Replica 1



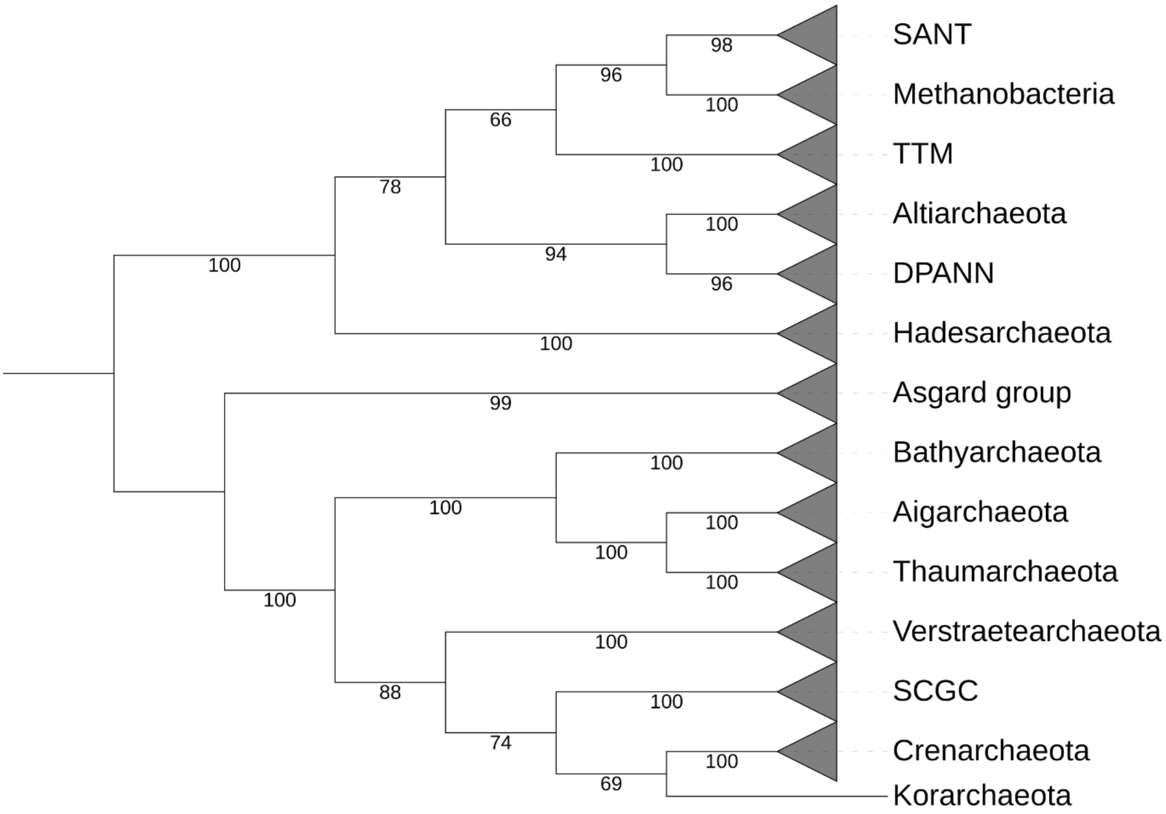
Replica 2



Replica 3



Replica 4



Replica 5

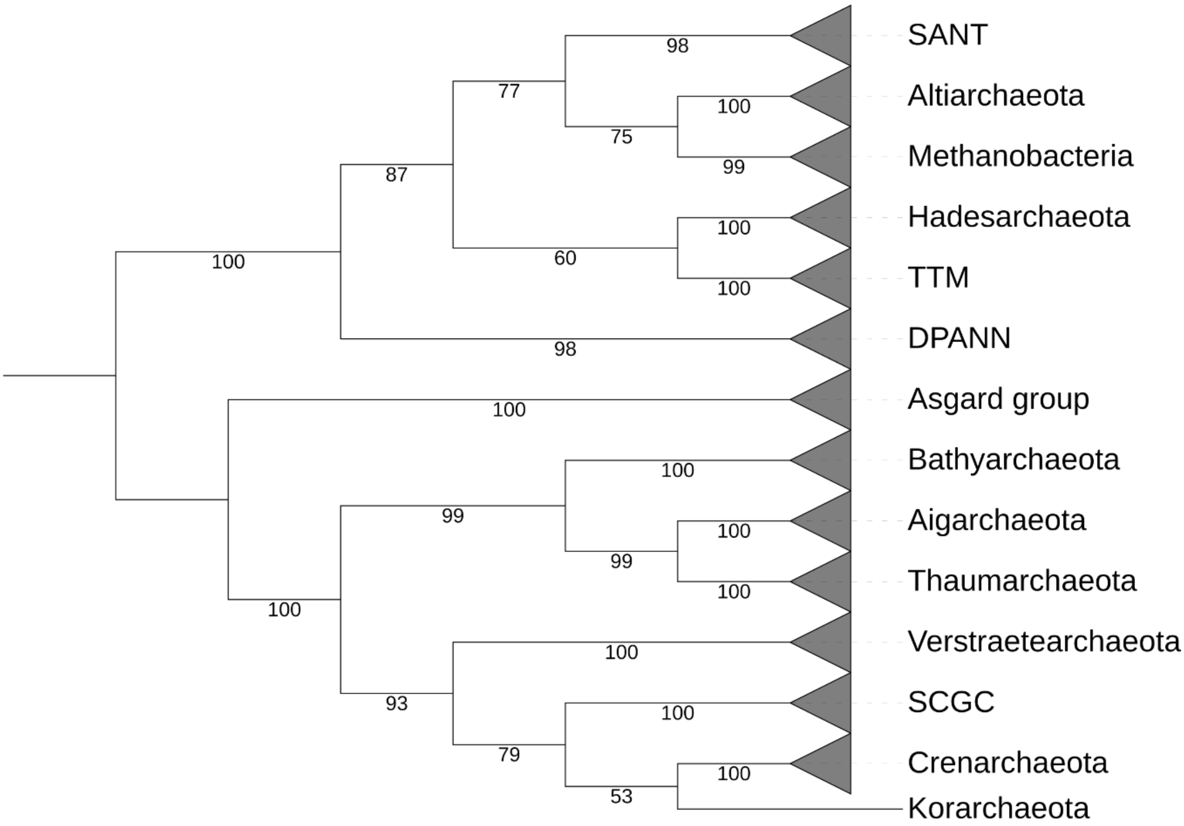


Figure 43. Consensus trees obtained for our 5 species replicas (35,000 positions) after a gene jackknife for 100 replicas per super-matrix calculated using the PMSF method.

Details of the 100 individual trees are shown in **Supp.Mat FigS43**.

The trees were rooted using the midpoint method. According to this method, the 2 historical archaea groups are Euryarchaeota and TACK. Of the 5 trees, replicas 2 and 3 support the same topology. Korarchaeota are positioned either as a sister group to Crenarchaeota, or as a sister group to Crenarchaeota + SCGC + Verstraearchaeota. The Hadesarchaeota are placed either as a sister group to the TTM group, or at the base of all Euryarchaeota and DPANN. The Altiarchaeales are either sister groups to DPANN or to Methanobacteria.

We distinguish a total of 4 possible topologies for all of our trees, depending on the species replica used; 2 of our 5 replicas (replicas 2 and 3) tend towards the same topology. Taxonomic sampling thus plays an important role in studies on the phylogeny of archaea. These 4 topologies are shown in **Figure 44**. One can observe these 4 topologies by fixing the method completely and varying only the replica.

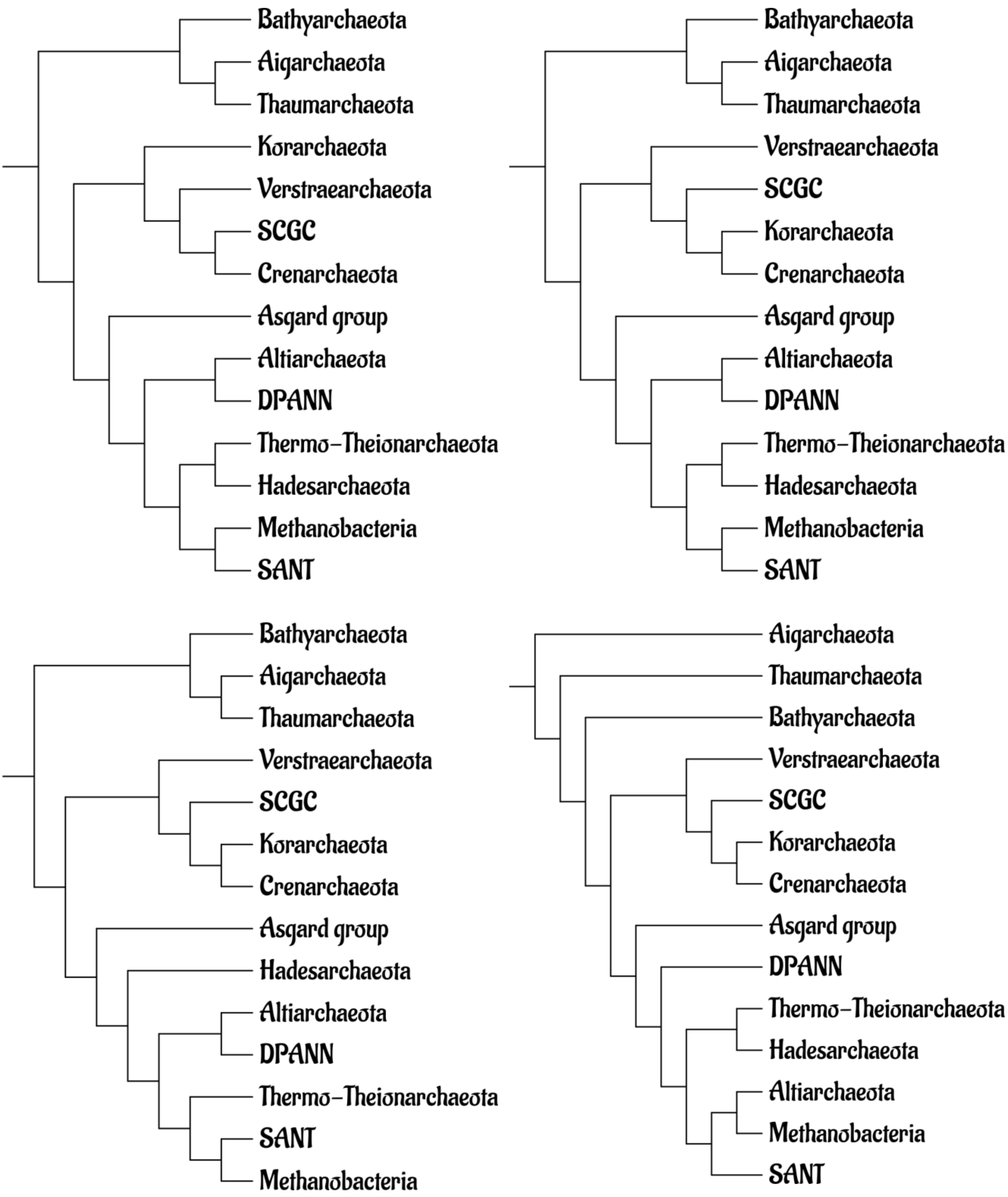


Figure 44. Different possible topologies found with our datasets depending on the replica.
By fixing the topology and varying the replica, 4 topologies are found. The variable positions concern Korarchaeota, Hadesarchaeota and Altiarchaeota.

Finally, the problematic groups we retain are as follows:

- The Korarchaeota, which are placed either as a sister group to the Crenarchaeota or as a sister group to the Crenarchaeota + SCGC + Verstraearchaeota clade. This latter hypothesis seems to be favored by a study based on a Bayesian phylogeny of a super-matrix of 41 genes (36 genes from the Phylosift marker gene list + the A and B subunits of RNA polymerase + 3 universal ribosomal proteins) (Adam et al., 2017; Raymann et al., 2015). On the other hand, another study based on a concatenation of 45 proteins as well as a super-tree of 3,242 genes places the Korarchaeota at the base of the entire TACK group (T. A. Williams et al., 2017).

- The Hadesarchaeota, which are placed either as a sister group to the Thermo-Theionarchaeota group or at the base of all Euryarchaeota and DPANN.
- The Altiarchaeales, which are placed either as a sister group to the DPANN or as a sister group to the Methanobacteria. This issue has already been studied in the literature. In early analyses, the Altiarchaeales were associated with the Methanococcales (Probst et al., 2014). It has also been suggested (based on attempts to root archaea) that they may represent one of the deepest branches of archaea, perhaps at the class or even phylum level (Adam et al., 2017; Raymann et al., 2015). However, due to their rapid evolutionary rate, the placement of the Altiarchaeales in the phylogeny of archaea should be approached with caution. Indeed, other analyses have also suggested that they could be grouped with the rapidly evolving DPANN lineage (Aouad et al., 2022; Bird et al., 2016; Dombrowski et al., 2020; Hug et al., 2016), although this may be caused by a reconstruction artifact rendering their position very uncertain (Martinez-Gutierrez & Aylward, 2021).

A phylogeny presenting the alternative positions of the problematic groups is shown in **Figure 45**.

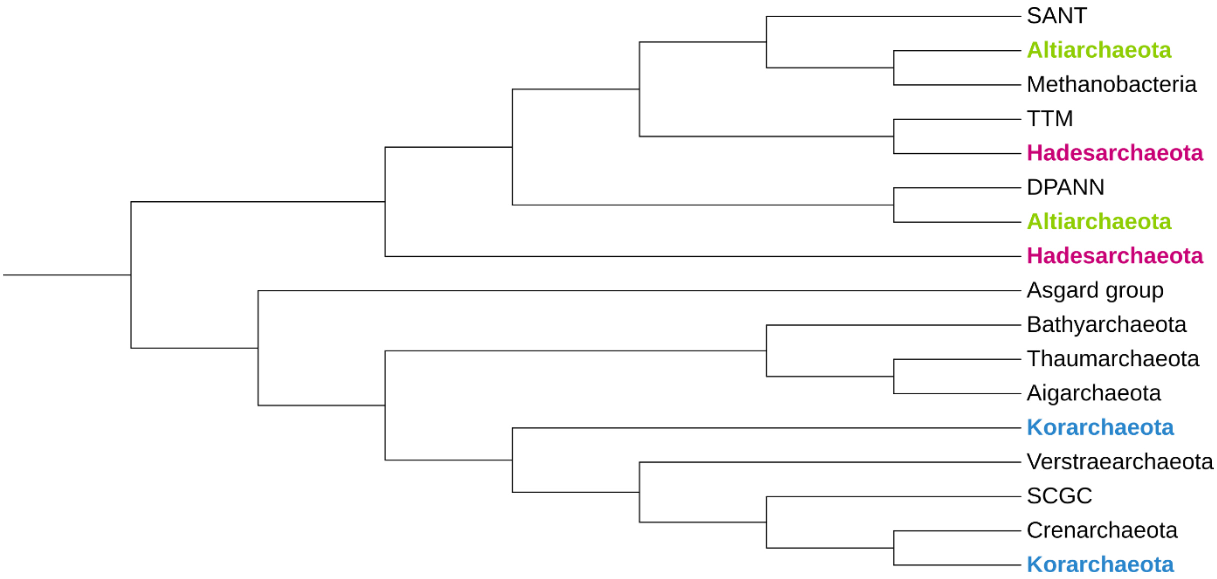


Figure 45. Synthetic unrooted cladogram showing possible positions of unstable groups (in color). We have three unstable groups: Korarchaeota (blue), Hadesarchaeota (red) and Altiarchaeota (green), each with two possible branches in the archaea tree.

9 REMOVAL OF SITES (SLOW-FAST)

We suspect that variations in the data and methods have revealed artifacts due to non-phylogenetic signal. Indeed, one possibility is that positions with rapid evolutionary rates are saturated by multiple substitutions, introducing a false phylogenetic signal and an incorrect tree topology, thereby invalidating the phylogenies: this is systematic error (**Box 12**). We will therefore attempt to assess the importance of this false phylogenetic signal and remove it (if possible) from our datasets or neutralize it in our analyses. To minimize systematic error, we have chosen to test our dataset by removing sites with rapid evolutionary rates using a so-called slow-fast approach (Aouad et al., 2022; Brinkmann & Philippe, 1999; Delsuc et al., 2005; Roure & Philippe, 2011) in order to favor slower sites, which are less likely to contradict the assumptions of substitution process homogeneity (although this could also be due to their high constraints, potentially causing convergences). This approach combats mutation saturation, which could lead to homoplasy.

The calculation of evolutionary rates per site and the construction of the different matrices was done with IQ-TREE using the 4 categories of rate from the gamma curve as a "proxy" for 4 rate categories. The gamma category of a position is estimated by the sum of changes that occurred within the sequences. By definition, IQ-TREE classifies columns into 4 categories, each corresponding to a different evolutionary rate. In the normal inference, each column belongs to each category with a 25% probability. IQ-TREE estimates the model parameters and then applies an empirical Bayesian approach to assign the site rates as the average across the rate categories, weighted by the posterior probability that the site belongs to that category. This empirical Bayesian approach is used by IQ-TREE as it is considered the most accurate (Mayrose et al., 2004). However, upon request, IQ-TREE can perform a Bayesian analysis to assign each column to a particular category. The sites are then assembled into different matrices called "bins." The first contains the slowest evolving sites, and the following bins contain sites with progressively faster evolutionary rates. We start with 4 equal-sized bins, which we concatenated as we incorporated sites with faster evolutionary rates, with the 4th alignment corresponding to our full initial alignment. We then calculated trees for each of these matrices with IQ-TREE using LG4X and the PMSF method, totaling 2 × 4 = 8 trees. The same trees were also calculated for the 4 bins considered individually.

Box 12. The Systematic Error and Model Choice

The systematic error

A sequence evolution model is only a simplified representation of the actual substitution process. It is based on assumptions that must be able to explain the data being analyzed. When substitutions are rare, it is easy to distinguish the phylogenetic signal. When substitutions are too abundant, they overlap (multiple substitutions) and obscure the original signal. Signal saturation occurs when there is more than one change per site on average. Due to their reduced number of possible character states (only 4), nucleotide sequences are more sensitive to saturation than amino acid sequences (which have 20 states). This is why amino acids are preferentially used when resolving ancient phylogenies, as they are less likely to saturate. The systematic error arises if the evolutionary process accumulating multiple substitutions also violates the assumptions of the model used for phylogenetic reconstruction, i.e., when the particularities of sequence evolution are not correctly accounted for by the underlying assumptions. This error increases as the number of characters used increases. The systematic error produces a non-phylogenetic signal that competes with the true phylogenetic signal. Unlike stochastic error, bootstrap or jackknife analyses do not necessarily indicate the presence of systematic error in a dataset, sometimes leading to results that are both incorrect and statistically supported. Matrices containing more fast-evolving positions are more likely to induce phylogenetic reconstruction artifacts. It is therefore imperative to find a way to detect them.

The Long Branch Attraction (LBA) artifact

One of the main phylogenetic reconstruction artifacts related to systematic error is the long branch attraction (LBA) phenomenon. LBA is an artifact that causes the grouping of taxa (1) either evolving more quickly, or (2) having diverged early (ancient branching), without reflecting their true evolutionary relationships. Indeed, when the evolutionary rates between the lineages studied are very different, those evolving faster are more likely to share characters through homoplasy rather than homology. In a parsimony context, the phenomenon is primarily linked to the accumulation of convergent substitutions interpreted as synapomorphies (while they are actually convergences!), whereas in probabilistic methods, it is rather linked to violations of the models mentioned above. As a result, these lineages with long branches (i.e., a large number of evolutionary events) are grouped together in the tree, regardless of their true relationship. Consequently, if a phylogeny is established from a gene evolving faster in one lineage than in another, it is highly likely that the faster-evolving lineage will be placed at the base of the tree, attracted by the external group used to root the phylogeny (which itself has a long branch), and thus interpreted as a relatively ancestral group.

Slow-Fast

Positions with rapid evolutionary rates are saturated by multiple substitutions and have lost their phylogenetic signal. To minimize the systematic error and its effects on phylogenetic reconstruction, several strategies have been proposed. These strategies involve removing non-phylogenetic signal from datasets, that is, substitutions incorrectly interpreted by reconstruction methods and supporting an incorrect topology. By removing fast-evolving sites, we favor those that are less likely to contradict the assumptions of substitution process homogeneity. This approach combats mutation saturation, which could lead to homoplasy.

Thus, the Slow/Fast method involves identifying positions that have not undergone any substitutions within predefined groups (the slowest positions) and gradually adding positions that have undergone one, two, three, etc., substitutions. From there, one can either (1) create a set of nested super-matrices containing increasingly faster positions, or (2) create sets of super-matrices with clearly distinct evolutionary rates. Observing the evolution of support for different topological hypotheses allows the evaluation of at which mutation rate the phylogenetic signal is lost and replaced by systematic error.

References

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. <i>Systematic Zoology,</i> <i>27</i>(4), 401-410.

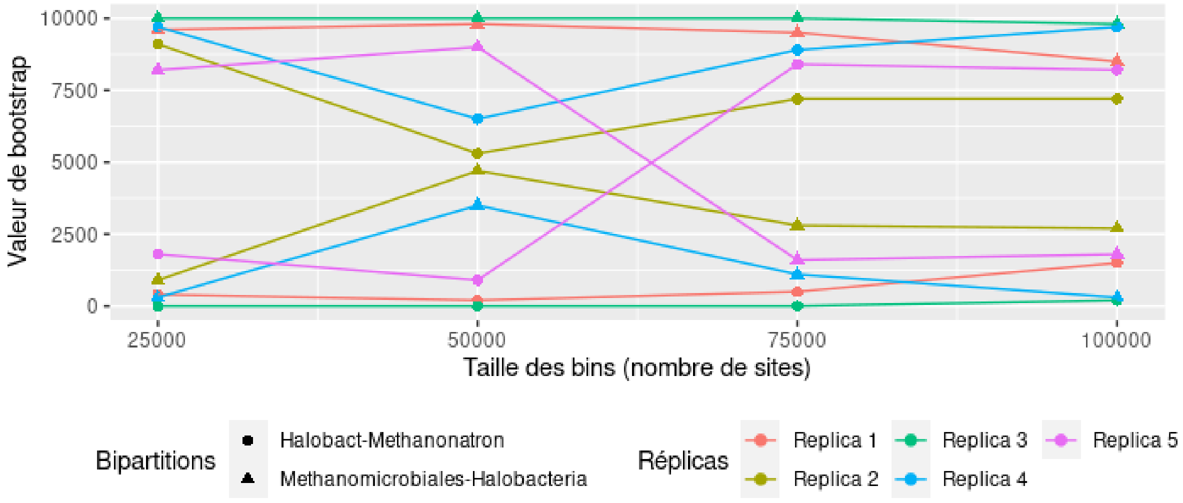
Phillips, M. J., F. Delsuc, et D. Penny. 2004. Genome-scale phylogeny and the detection of systematic biases. Mol Biol Evol 21:1455-8.

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.

There are several things to note about our ribosomal groups:

Within the SANT group, it is difficult to determine the position of the *Halobacteria*, *Archaeoglobi*, and *Methanonatronarchaeia*, as all three could potentially position themselves at the root of the SANT group. Additionally, another possibility is the proximity of *Halobacteria* with *Methanomicrobiales*. However, variations are noted between bins as well as between replicates, which are difficult to interpret (**Figure 46** & **Supp.Mat slow-fast**). For instance, regarding the proximity of *Halobacteria* with *Methanomicrobiales*, for the non-cumulative bins, for replicate 2, the first bin is at 9% (ultrafast-bootstrap reduced to 100 since we have 10,000 bootstrap replicates), the second is at 57%, and the third is at 2%; for replicate 4, the first bin is at 3%, the second at 76%, and the third at 0%. Replicates 1 and 3 strongly favor a proximity of *Methanomicrobiales* with *Halobacteria* (ultrafast-bootstrap > 95%), whereas replicates 2 and 4 favor a proximity of *Halobacteria* with *Methanonatron*. The latter hypothesis is supported for replicate 5 for the first two bins only. Finally, the monophyly of the SANT group is not recovered with the last bin when using positions with maximal systematic error, while the complete super-matrices and the use of all the cumulative bins recover this group.



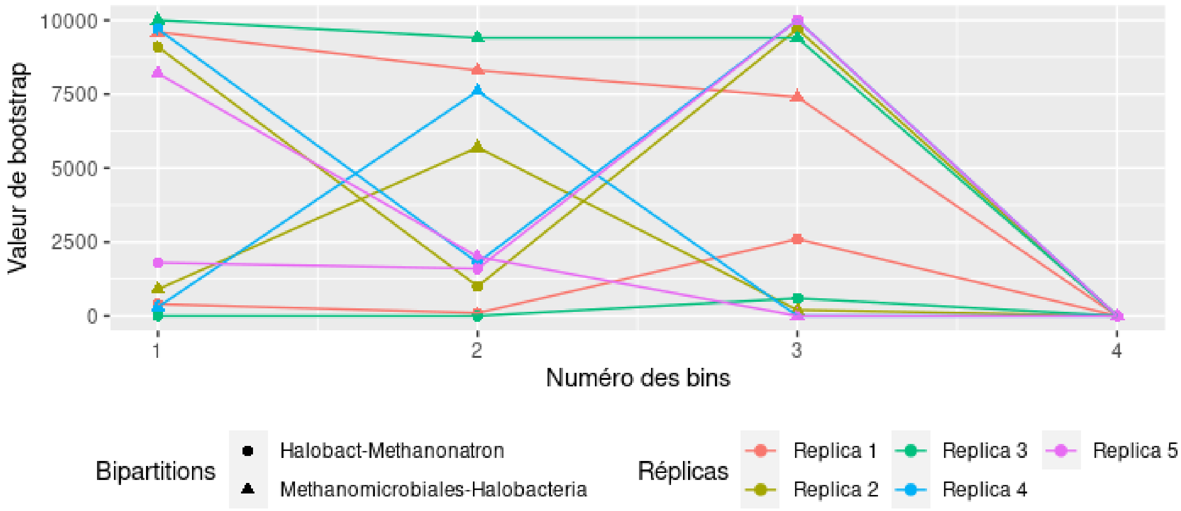
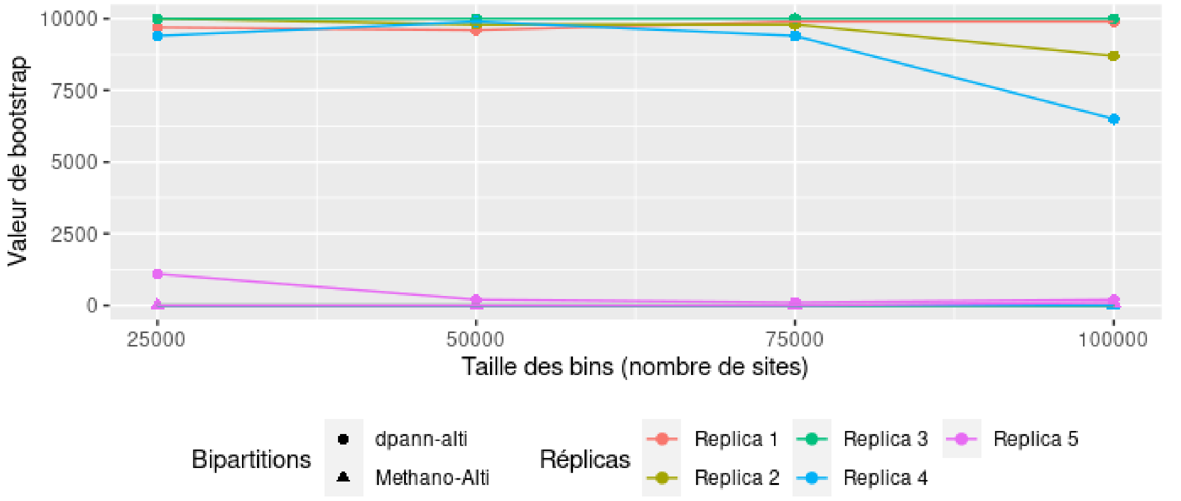


Figure 46. Proportion of trees indicating the 2 possible positions of Halobacteria (either with Methanonatronarchaeia or Methanomicrobiales) according to cumulative and non-cumulative bins size for each replica according to the LG4X model.

Data are provided in **Supp.Mat slow-fast**. The position of Halobacteria, Archaeoglobi and Methanonatronarchaeia is difficult to judge, as all three could potentially be positioned at the root of the SANT group. The result varies according to both replica and bin, suggesting that both the choice of species and the alignment used may give different results.

The Altiarchaeota tend to cluster with the DPANN group, except for replicate 5 (**Figure 47**). For the cumulative bins, the ultrafast-bootstrap is consistently above 94% for replicates 1 to 4. However, this hypothesis is not supported for replicate 5, with an ultrafast-bootstrap value lower than 11%. For the non-cumulative bins, low ultrafast-bootstrap values are found for the DPANN + Altiarchaeota hypothesis in replicate 5 (ultrafast-bootstrap < 11%). In contrast, for replicates 1 to 4, the ultrafast-bootstrap values favor this hypothesis for the first three alignments (ultrafast-bootstrap > 92%), while the last alignment (containing the fastest sites) never recovers this hypothesis. Therefore, according to our analyses, it appears that the Altiarchaeota are emerging from the Euryarchaeota, although this conclusion is dependent on the replicate. It should be noted that replicate 5 does not support either of our two hypotheses.



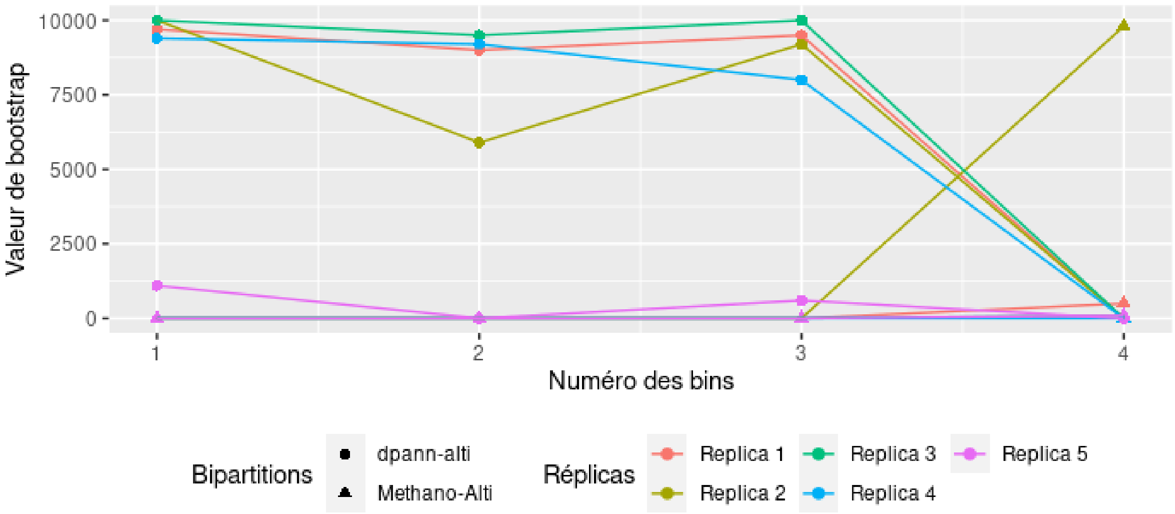
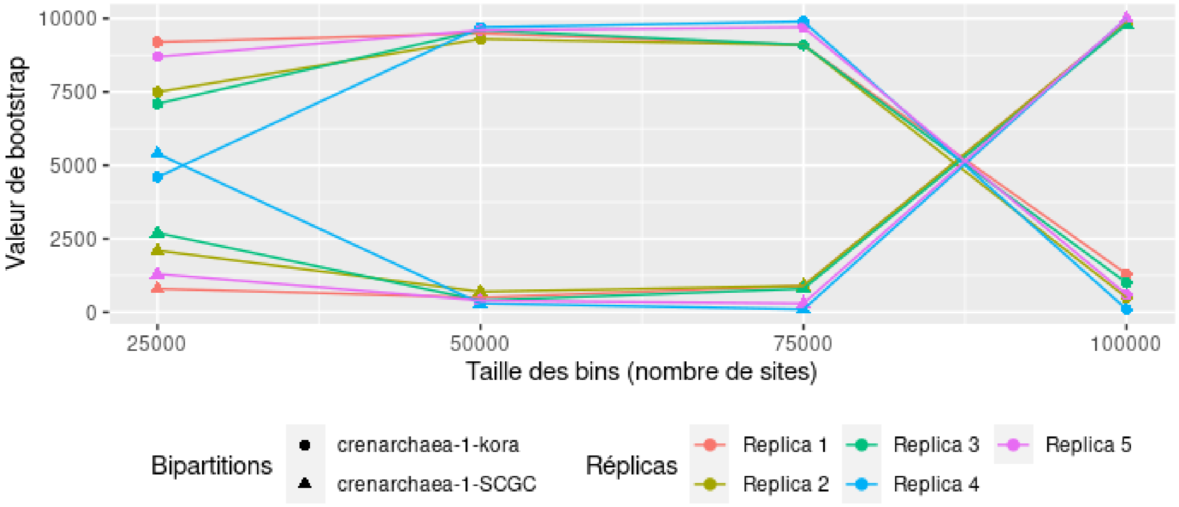


Figure 47. Proportion of trees indicating the 2 possible positions of Altiarchaeota (either with DPANNs or Methanomicrobiales) according to cumulative and non-cumulative bins size for each replica according to the LG4X model.

Data are provided in **Supp.Mat slow-fast**. Altiarchaeota tend to be closer to the DPANN group, except for replica 5.

We previously observed that the position of the Korarchaeota was uncertain and varied depending on the model used, sometimes at the base of the Crenarchaeota, sometimes at the base of the Crenarchaeota + SCGC (**Figure 48**). Our slow-fast analyses allow us to settle their position by favoring the Korarchaeota rather than the SCGC group as the sister group to the Crenarchaeota. This solution is strongly supported when using the cumulative bins (ultrafast-bootstrap > 71%). In contrast, the use of individual bins clearly shows that using fast sites favors the alternative hypothesis of SCGC at the base of all Crenarchaeota, indicating a phylogenetic reconstruction artifact. Therefore, the Korarchaeota and Crenarchaeota form a monophyletic group that roots within the SCGC group.



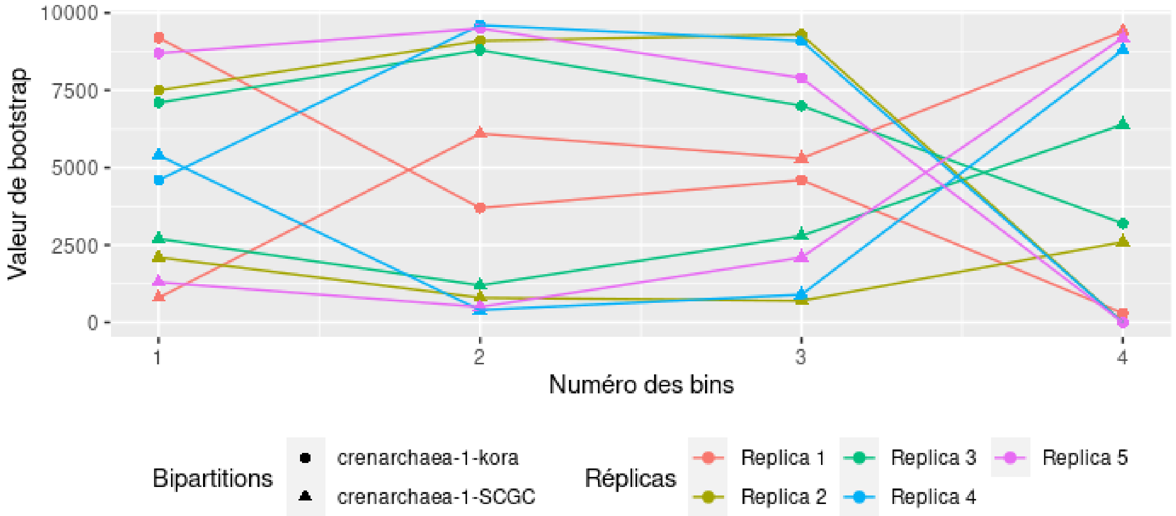


Figure 48. Proportion of trees indicating the 2 possible positions of Korarchaeota and SCGC in relation to Crenarchaeaota according to cumulative and non-cumulative bins size for each replica according to the LG4X model.

Data are provided in **Supp.Mat slow-fast**. Our results show that Korarchaeota and Crenarchaeota form a monophyletic group rooted in the SCGC group.

The Hadesarchaeota favor the Thermococci, Theionarchaea, and Methanomicrobia_Arc as the sister group when using slow sites. Only replica 4 struggles to support this hypothesis, with a maximum ultrafast-bootstrap of 7% when using separate bins, whereas the use of cumulative bins yields ultrafast-bootstrap values lower than 17%, except for the last bin, which reaches a high value of 89% (cf. **Supp.Mat slow-fast/SlowFast_bilan.xlsx**).

10 DISCUSSION

Several studies have shown that an increase in taxonomic sampling can improve phylogenetic accuracy (Jeffroy et al., 2006; Pollock et al., 2002; Rokas & Carroll, 2005; Zwickl & Hillis, 2002). This strategy is commonly used as a solution to resolve unstable nodes in the tree of life (Young & Gillung, 2020). In some cases, it has been suggested that conflicts between reported trees may result from differences in taxonomic representation (Da Cunha et al., 2017; Nasir et al., 2016) and it is unclear to what extent over-sampling some taxa compared to others can negatively affect the reconstruction of a tree (Martinez-Gutierrez & Aylward, 2021).

Bathyarchaeota phylogeny

Although the internal phylogeny of the Bathyarchaeota is not the precise focus of our study, we can mention that the first phylogeny of the Bathyarchaeota was established in 2012 based on 4,720 sequences from the SILVA database (Kubo et al., 2012). Sequences of at least 940 bp were used to construct the backbone of the tree, to which additional sequences were added without changing the general topology of the tree. Whether using distance methods or maximum likelihood, a total of 17 subgroups were identified, with 76% similarity shared by the most distant sequences. However, 12% of the sequences remained ungrouped and isolated. Subgroups 13 to 17 were unstable according to the reconstruction methods used and exhibited multifurcations. In 2016, another study added two new subgroups (18 and 19) with high ultrafast-bootstrap values (96% and 86%, respectively) (Fillol et al., 2016). Subgroup 5 was also divided into 5a and 5b, each with intra-group similarity greater than 90% according to maximum likelihood estimates. Group

5b itself was further subdivided into 5b and 5bb as new sequences were added. Bathyarchaeota are characterized by very high intra-group diversity. The limit of the 16S rRNA sequences of the Bathyarchaeota falls within the minimum sequence identity range considered for a phylum (74.95–79.9%), and each subgroup falls within the median sequence identity range of sequences from families and orders (91.65-92.9% and 88.25–90.1%, respectively) (Yarza et al., 2014). It has been proposed that this high diversity of the Bathyarchaeota reflects a high diversity of metabolisms within the different subgroups (Kubo et al., 2012). Currently, the classification of Bathyarchaeota based on 16S rRNA supports a total of 25 groups (Zhou et al., 2018).

Monophyly of the TACK group and relationship with the Asgard group

For more than a decade (1990-2002), Euryarchaeota and Crenarchaeota were the only two known archaeal groups. Between 2002 and 2011, thanks to the development of sequencing techniques (amplicon sequencing, shotgun metagenomics...), several new phyla were proposed based on phylogenetic and genomic analyses: Korarchaeota, Nanoarchaeota (cultivated in 2002), Thaumarchaeota (which oxidizes ammonia), and Aigarchaeota (Brochier-Armanet, Boussau, et al., 2008; Elkins et al., 2008; Huber et al., 2002; Nunoura et al., 2011). Together, they form the TACK superphylum (or Protoarchaeota) (Kozubal et al., 2013; Petitjean et al., 2014), to which the Geoarchaeota metagenomes (Guy et al., 2014; Kozubal et al., 2013), Bathyarchaeota (Evans et al., 2015; He et al., 2016; Lazar et al., 2016) and Verstraetearchaeota (Vanwonterghem et al., 2016) have recently been added. The monophyly of TACK is well-supported in our analyses. We consistently find the Asgard group at the base of TACK, regardless of the sampling strategy. This conclusion is supported by recent studies (Adam et al., 2017; Spang et al., 2015; T. A. Williams et al., 2017; Zaremba-Niedzwiedzka et al., 2017). However, some have suggested that the placement of Asgard archaea near TACK could be partly due to imbalanced taxon sampling (Da Cunha et al., 2017; Nasir et al., 2016). In our case, we have made sure that our sampling is well-balanced, which suggests that this result is likely correct.

Nomenclature and General Taxonomy of Archaeal Groups

Particular attention must also be given to the nomenclature of archaea. On two occasions, we noted cases of homonymy for species belonging to two unrelated genera. This is the case for species named (1) “thermoplasmatales” within two groups (the “true” Thermoplasmata and those forming the Theionarchaea group), and (2) Methanomicrobia called Arc, which are not related to the true Methanomicrobia found in our SANT group. It is difficult to say whether this confusion arises from an originally metabolism- and morphology-based description or if it is a phylogenetic error. The current taxonomy of recognized archaeal groups is inconsistent due to the lack of clear genome-based guidelines for microorganism classification (Gribaldo & Brochier-Armanet, 2012). These discrepancies can partly be explained by the limited resolution power of 16S rRNA, especially for the oldest nodes, by methodological artifacts, or by the poor quality or small size of sequences, all of which can explain these nomenclature issues as well as the increasing number of phyla described in recent years. The growing availability of genomic sequence data from previously unsampled lineages has led to a dramatic increase in proposals (at least 3 superphyla: TACK, DPANN, Asgard, and dozens of new phyla) that have been added to the historical dichotomy of Crenarchaeota and Euryarchaeota defined by Woese in 1990. Given the ongoing progress in sequencing technologies and the exploration of microbial diversity, there is a risk of an anarchic explosion of high-level taxa. It is therefore urgent to establish common rules that should be followed before formally proposing a new archaeal phylum. Only recently have efforts begun to be made to use well-defined criteria based on comparative phylogenomics and genomic frameworks (genomic distances in percentage) to standardize and help establish a robust taxonomy for archaea (and also for bacteria) (Parks et al., 2018; Rinke et al., 2021).

Importance of Species, Gene, and Model Selection

In our study, the results are strongly dependent on the species replica rather than the gene selection. For each method employed (super-tree vs super-matrix), the phylogenies obtained have congruent topologies overall, which is evidence of a coherent phylogenetic signal in these different datasets. A significant effort was made to ensure the use of reliable genes, and our results seem to support the robustness of the topologies when only considering the genes. However, incongruence is mainly observed at the replica and model levels. The variability between models is less pronounced and is mostly felt when using the LG4X model, which is more sensitive to gene sampling variations than categorical models due to its simpler nature. Additionally, super-trees are less robust than super-matrices. Finally, species replicas do not all behave the same way, highlighting the importance of having studied 5 different replicas. Therefore, the choice of species appears to be a critical criterion, along with the use of sophisticated models. Among our proposed phylogenetic trees for archaea, the one that seems most credible to us is shown in **Figure 49** below, corresponding to the trees of our replicas 2 and 3:

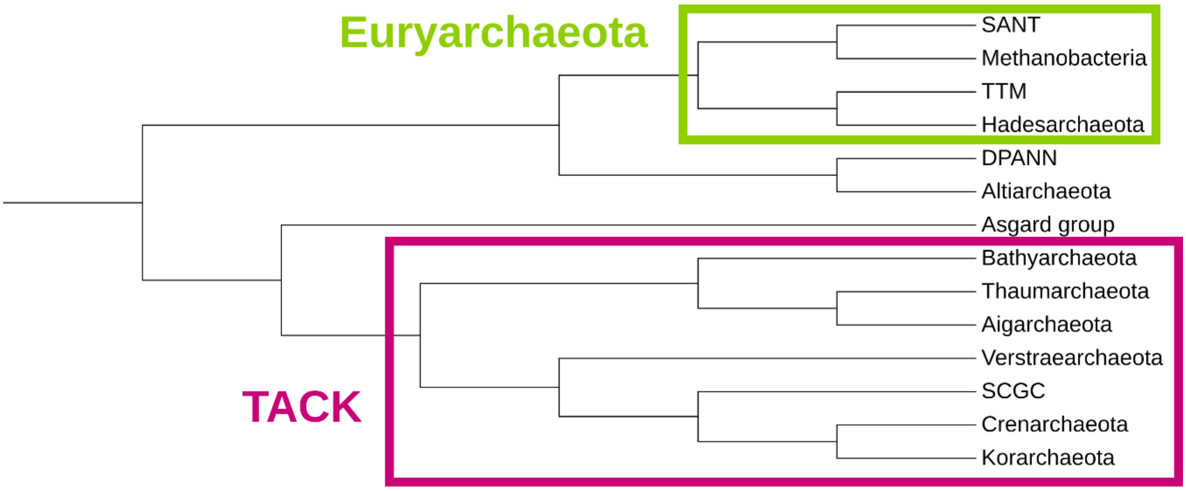


Figure 49. Unrooted phylogenetic trees of archaea appearing most credible according to our results.

The two main groups are the TACKs and the Euryarchaeota. DPANN and Altiarchaeota form a third group outside the Euryarchaeota.

CHAPTER 3

EUKARYOTES AND THEIR RELATIONSHIPS WITH ARCHAEA

1 INTRODUCTION

Recent studies suggest that eukaryotes share a common ancestor with the Asgard group. However, this result can be debated as it may stem from methodological biases, particularly a long-branch attraction artifact. To test this hypothesis, we will select genes that are present in both archaea and eukaryotes and rerun the phylogenetic analyses described previously to assess the impact of adding eukaryotes to our trees. In the previous chapter, we obtained a phylogeny of archaea for 5 species replicas. Our objective now is to insert a selection of eukaryotes into these trees.

2 SELECTION OF EUKARYOTES

We have expertise within our laboratory concerning eukaryotes (Van Vlierberghe, Philippe, et al., 2021). We focused on high-quality proteomes, particularly those with high completeness, which is essential for obtaining the most complete and balanced orthologous groups possible (Simão et al., 2015; Waterhouse et al., 2018), with low levels of contamination (Irisarri et al., 2017; Simion et al., 2017; Van Vlierberghe, Di Franco, et al., 2021). Within our laboratory, we have selected 73 high-quality eukaryotic proteomes (Van Vlierberghe, Philippe, et al., 2021). We then selected 11 species from these proteomes that best represent eukaryotic diversity. Ideally, this selection should allow us to hypothesize on the properties of LECA and avoid synapomorphies that would only concern certain subgroups of eukaryotes. The selection of eukaryotes we wish to add to our archaea is as follows:

Species	Sub-group	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

We added this selection of 11 eukaryotes with 42 to our 440 genes (i.e., before the congruence tests and the creation of chimeras) from our 5 archaeal replicas (121 species), for a total of 132 species per replica. Thus, for each replica, 204, 199, 206, 202, and 202 genes were enriched with at least one eukaryotic sequence.

3 GENES SELECTION

Now that we have added our selection of eukaryotes, we need to select the genes we will keep. Several criteria will be considered for the selection of our genes. We need genes where eukaryotes are sufficiently represented, both in number and in diversity. We also need to limit issues of paralogy that may arise when the same eukaryote is added multiple times within the same orthologous group. Indeed, some eukaryotic genes may have undergone numerous duplications, each with its own independent evolution over time. It is important to manage these paralogues and determine which correspond to archaeal homologues. We evaluated whether for a given MSA, some species were present multiple times, indicating possible paralogies. To do this, we counted the number of occurrences of each eukaryotic species within each MSA (**Figure 50**). Some organisms were added multiple times within the same OG, indicating possible paralogy phenomena. Specifically, the extreme cases of *Arabidopsis thaliana*, which is present 12 and 14 times in OG0000295 and OG0000825, respectively, and *Emiliania huxleyi*, which is present 13 times in OG0000295. Generally, this analysis shows that *Arabidopsis thaliana* and *Guillardia theta* are more likely to be added as multiple paralogues in our orthologous groups, resulting either from duplication events within their lineage, gene transfers, or secondary endosymbioses. Thus, *Guillardia theta* possesses a nucleomorph, a vestige of a 551 kbp nucleus from the red algae that was engulfed during its evolution, exacerbating this paralogy issue. The successive endosymbioses that gave rise to these microorganisms endow them with a particularly complex genomic organization, with four different genomes (Sibbald & Archibald, 2020) :

- two prokaryotic genomes, in the mitochondria and plastids of red and green algae;
- two eukaryotic genomes, in the host cell nucleus and in the nucleomorph.

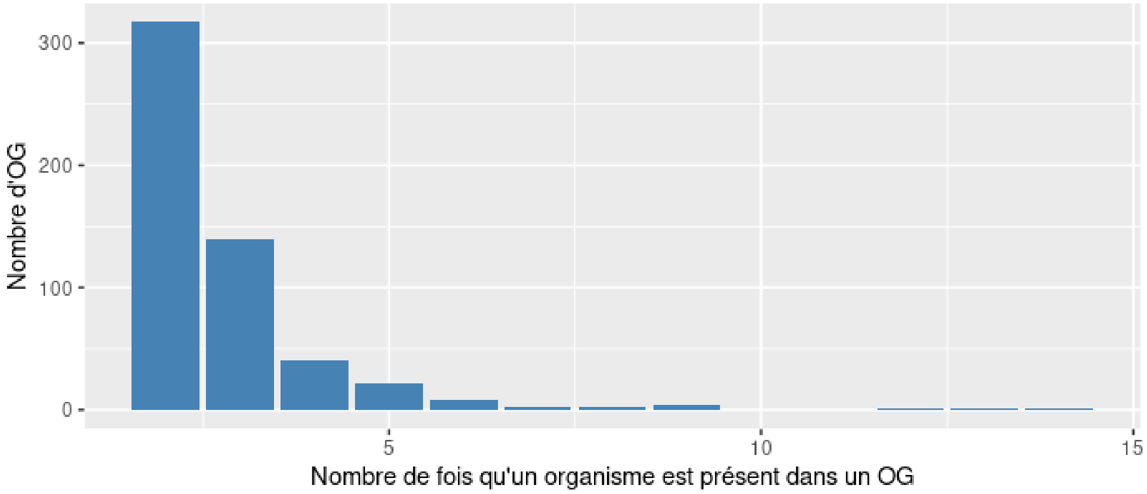


Figure 50. Distribution of duplicates within orthologous groups (Note that this is the distribution of duplicates, so by default the x-axis starts at 2).

Data are provided in **Supp.Mat selection-genes**. Some organisms have been added in several copies within the same OG, reflecting possible paralogy phenomena. Arabidopsis thaliana and Guillardia theta are most likely to be added as multiple paralogs in our orthologous groups.

Then, for each eukaryote, we counted the number of orthologous groups in which it is represented several times (**Figure 51**).

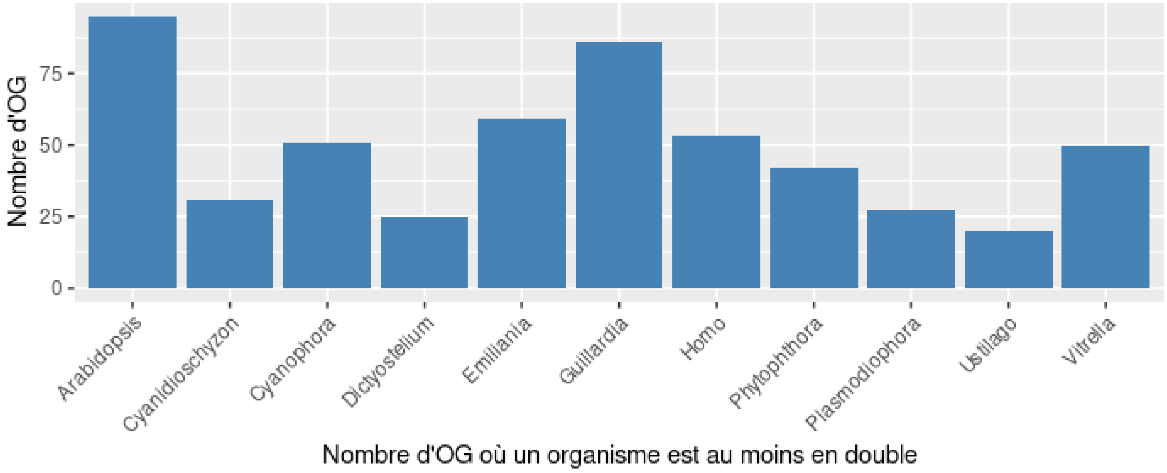


Figure 51. Number of orthologous groups with multiple occurrences of a given eukaryote.
Data are provided in **Supp.Mat selection-genes**.

We also evaluated for each orthologous group the number of eukaryotic species present at least twice (**Figure 52**). Here, we are looking for genes where eukaryotes (in general) are duplicated in order to properly manage duplication issues. Indeed, depending on the case, we will either select the correct copy of the eukaryotic gene to keep, or in cases where it is impossible to decide which one is the correct copy, eliminate the gene.

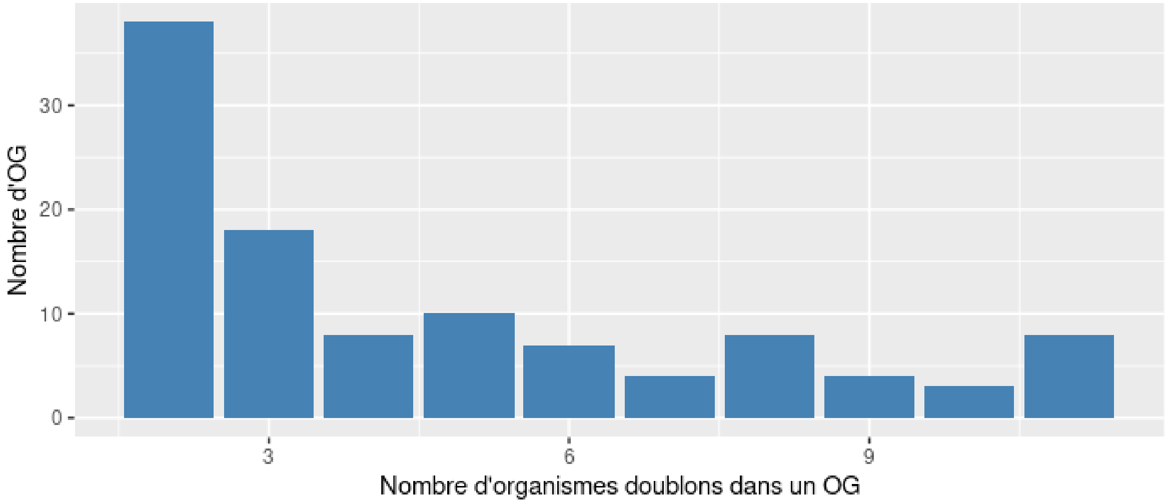


Figure 52. Distribution of the number of duplicate eukaryotic species per orthologous group (e.g. 10 OGs have 5 duplicate eukaryotes).
Data are provided in **Supp.Mat selection-genes**.

We then evaluated the distribution of the number of eukaryotes added within the orthologous groups (**Figure 53**). Long divided into Unikonts and Bikonts, phylogenetic analyses of protein super-matrices generally show a eukaryotic tree composed of five to eight "super-groups" that are further divided into three higher-order assemblages:

- Amorphea (Amoebozoa plus Obazoa, the latter including animals and fungi);
- Diaphoretickes (mainly Sar, Archaeplastida, Cryptista, and Haptophyta);
- Excavata (Discoba and Metanomonada) (Adl et al., 2012; M. W. Brown et al., 2017; Burki et al., 2020; Keeling & Burki, 2019).

The eukaryotic root would be located somewhere between Amorphea and the Diaphoretickes + Excavata group. This root is called "Opimoda-Diphoda" (Derelle et al., 2015). We have 24 alignments where only Diaphoretickes have been added, 3 where only Amorphea have been added, and 163 that contain both Amorphea and Diaphoretickes..

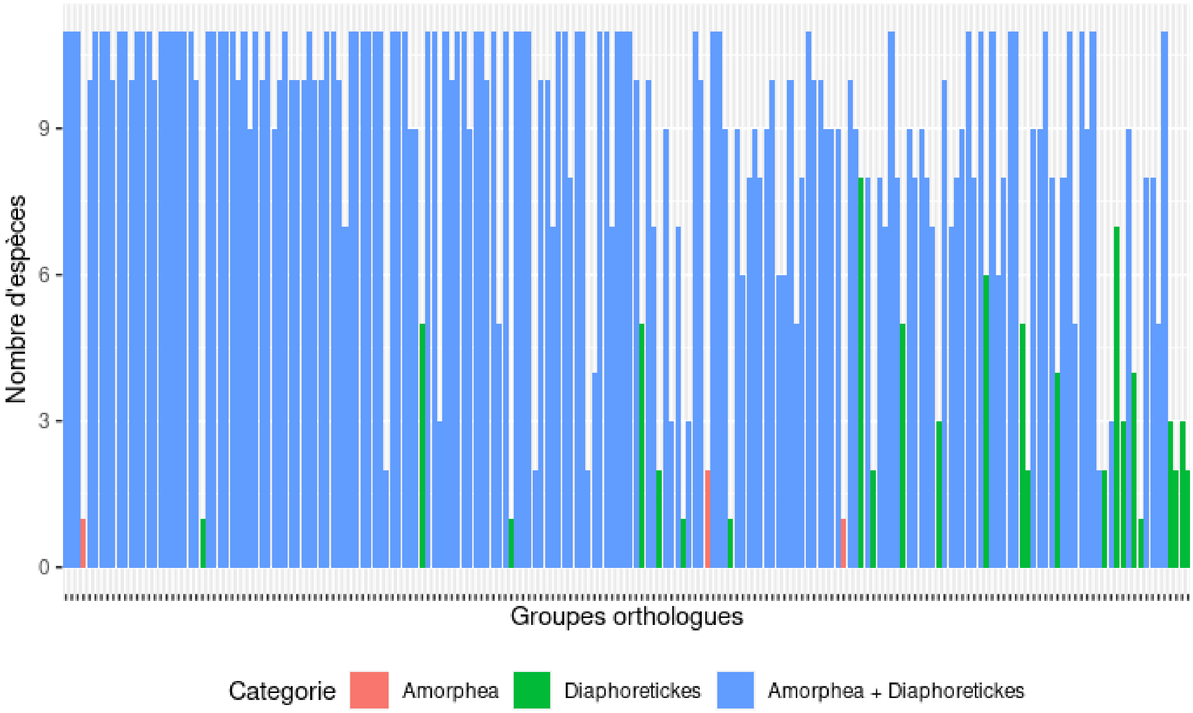


Figure 53. Distribution of the number of eukaryotes added within orthologous groups. Data are provided in **Supp.Mat selection-genes**.

We then sought to determine if there are orthologous groups more likely to be enriched in eukaryotes. To do this, we looked at how many orthologous groups added the same number of eukaryotes (**Figure 54**).

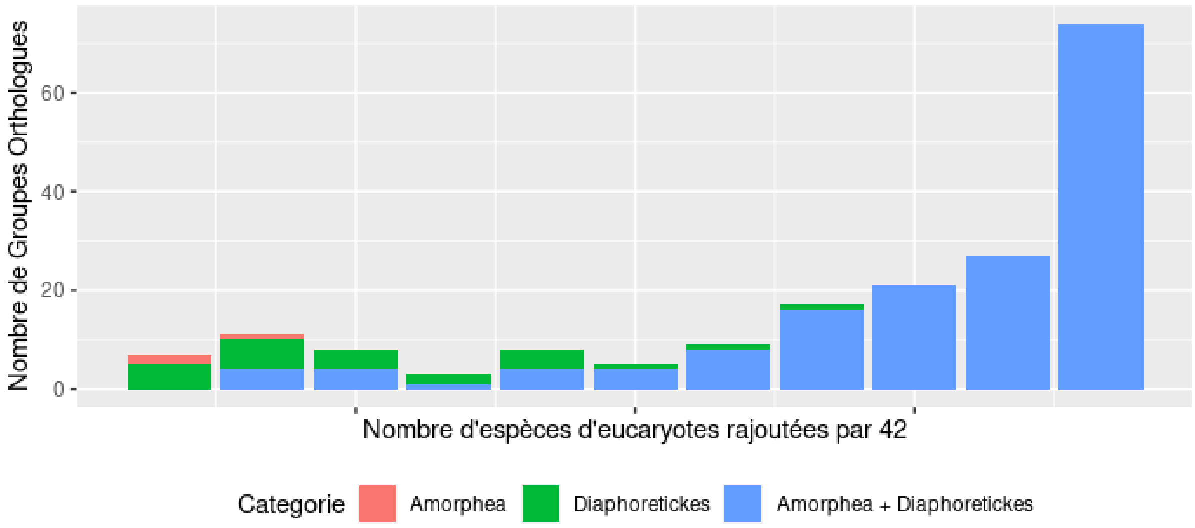


Figure 54. Number of orthologous groups having added the same number of eukaryotes. Data are provided in **Supp.Mat selection-genes**. The majority of our genes added at least 9 of our eukaryotes. Almost 74 genes added all our eukaryotes.

Finally, we assessed the proportion of Amorphea and Diaphoretickes for each orthologous group, in order to prioritize orthologous groups with a representative from each group (**Figure 55**). Indeed, we aim to get as close as possible to the state of LECA. Therefore, it is important to have orthologous groups for which we can legitimately infer that they were already present in LECA.

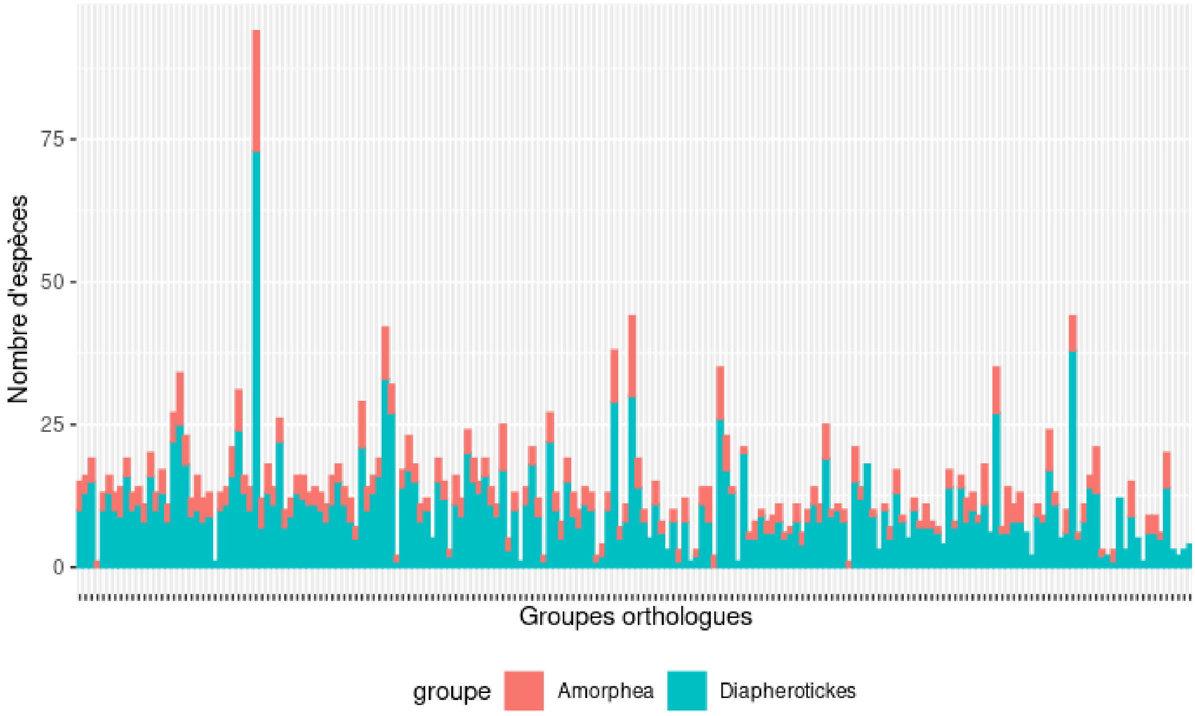


Figure 55. Proportion of Amorphea and Diaphoretickes added per 42 for each orthologous group. Data are provided in **Supp.Mat selection-genes**.

125 genes contain at least 10 eukaryotic organisms in one of the five replicates, and 102 genes have at least 10 eukaryotic organisms in all five replicates. We then removed individual sequences (based on their accession) that were not added in all five replicates, ensuring that these

replicates had the exact same set of eukaryotic sequences. We thus chose to keep orthologous groups with at least 8 eukaryotic organisms in all five replicates, i.e., 138 genes out of 440. We then calculated with IQTree and the LG4X model individual trees for our selection of 138 genes to verify the monophyly of the eukaryotes and avoid contamination from any sequence added by 42. Several cases arose:

1. the eukaryotic sequences are monophyletic in single copy (50 genes) (**Figure 56**);
2. the eukaryotic sequences are duplicated and polyphyletic in the archaeal tree (85 genes) (**Figure 57**); In this case, these are either isolated sequences that seem to be subject to the long-branch attraction artifact, or certain species (particularly *Arabidopsis thaliana*, *Cyanidioschyzon merolae*, *Cyanophora paradoxa*, *Emiliania huxleyi*, and *Guillardia theta*) that have numerous paralogs. Identifying and removing these sequences resolves this issue. Among these, 9 trees have 2 monophyletic eukaryotic subtrees, and we were unable to objectively determine which one is correct.
3. the eukaryotic sequences are in single copy but polyphyletic (3 genes) (**Figure 58**). These genes were systematically removed (trees were rooted in eukaryotes by *mid-point root*).

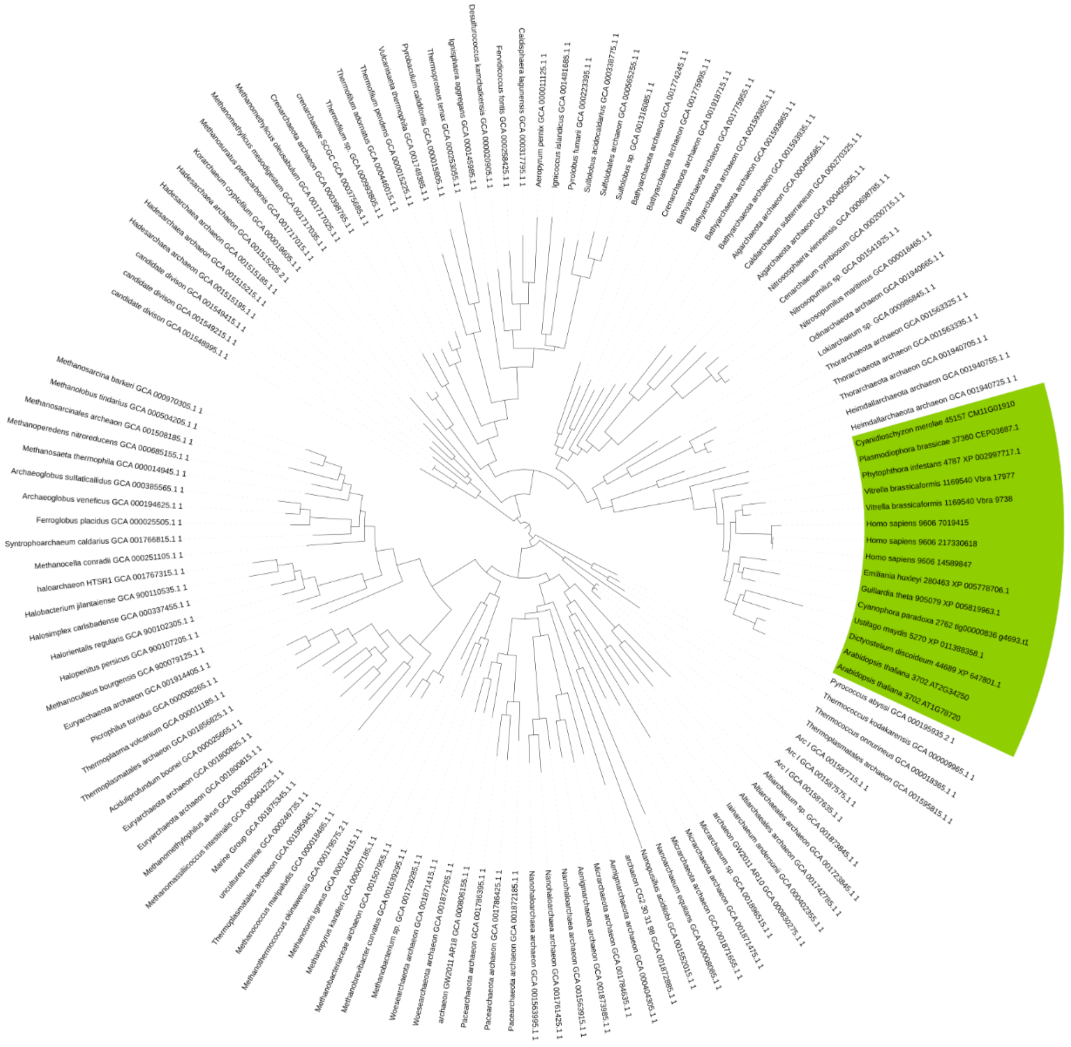

candidate divison GCA 001548995.1 1
candidate divison GCA 001549215.1 1
candidate divison GCA 001549415.1 1
Hadesarchaea archaeon GCA 001515195.1 1
Hadesarchaea archaeon GCA 001515215.1 1
Hadesarchaea archaeon GCA 001515185.1 1
Hadesarchaea archaeon GCA 001515205.2 1
Korarchaeum cryptofilum GCA 000019605.1 1
Methanosuratus petracarbonis GCA 001717015.1 1
Methanomethylicus mesodigestum GCA 001717035.1 1
Methanomethylicus oleusabulum GCA 001717025.1 1
Crenarchaeota archaeon GCA 000398765.1 1
crenarchaeote SCGC GCA 000375685.1 1
Thermofilum sp. GCA 000993805.1 1
Thermofilum adornatus GCA 000446015.1 1
Thermofilum pendens GCA 000015225.1 1
Vulcanisaeta thermophila GCA 001748385.1 1
Pyrobaculum calidifontis GCA 000015805.1 1
Thermoproteus tenax GCA 000253055.1 1
Ignisphaera aggregans GCA 000145985.1 1
Desulfurococcus kamchatkensis GCA 000020905.1 1
Fervidicoccus fontis GCA 000258425.1 1
Caldisphaera lagunensis GCA 000317795.1 1
Aeropyrum pernix GCA 000011125.1 1
Ignicoccus islandicus GCA 001481685.1 1
Pyrolobus fumarii GCA 000223395.1 1
Sulfolobus acidocaldarius GCA 000338775.1 1
Sulfolobales archaeon GCA 000565255.1 1
Sulfolobus sp. GCA 001316085.1 1
Bathyarchaeota archaeon GCA 001774245.1 1
Bathyarchaeota archaeon GCA 001775995.1 1
Crenarchaeota archaeon GCA 001918715.1 1
Bathyarchaeota archaeon GCA 001775955.1 1
Bathyarchaeota archaeon GCA 001593855.1 1
Bathyarchaeota archaeon GCA 001593865.1 1
Bathyarchaeota archaeon GCA 001593935.1 1
Aigarchaeota archaeon GCA 000405685.1 1
Caldiarchaeum subterraneum GCA 000270325.1 1
Aigarchaeota archaeon GCA 000405905.1 1
Nitrososphaera viennensis GCA 000698785.1 1
Cenarchaeum symbiosum GCA 000200715.1 1
Nitrosopumilus sp. GCA 001541925.1 1
Nitrosopumilus maritimus GCA 000018465.1 1
Odinarchaeota archaeon GCA 001940665.1 1
Lokiarchaeum sp. GCA 000986845.1 1
Thorarchaeota archaeon GCA 001563325.1 1
Thorarchaeota archaeon GCA 001563335.1 1
Thorarchaeota archaeon GCA 001940705.1 1
Heimdallarchaeota archaeon GCA 001940755.1 1
Heimdallarchaeota archaeon GCA 001940725.1 1
Cyanidioschyzon merolae 45157 CM11G01910
Plasmodiophora brassicae 37360 CEP03687.1
Phytophthora infestans 4787 XP 002997717.1
Vitrella brassicaformis 1169540 Vbra 17977
Vitrella brassicaformis 1169540 Vbra 9738
Homo sapiens 9606 7019415
Homo sapiens 9606 217330618
Homo sapiens 9606 14589847
Emiliania huxleyi 280463 XP 005778706.1
Guillardia theta 905079 XP 005819963.1
Cyanophora paradoxa 2762 tig00000836 g4693.t1
Ustilago maydis 5270 XP 011388358.1
Dictyostelium discoideum 44689 XP 647801.1
Arabidopsis thaliana 3702 AT2G34250
Arabidopsis thaliana 3702 AT1G78720
Pyrococcus abyssi GCA 000195935.2 1
Thermococcus kodakarensis GCA 000009965.1 1
Thermococcus onnurineus GCA 000018365.1 1
Thermoplasmatales archaeon GCA 001595815.1 1
Arc I GCA 001587715.1 1
Arc I GCA 001587575.1 1
Arc I GCA 001587635.1 1
Altiarchaeum sp. GCA 001873845.1 1
Altiarchaeales archaeon GCA 001723845.1 1
Altiarchaeales archaeon GCA 001742785.1 1
Iainarchaeum andersonii GCA 000402355.1 1
archaeon GW2011 AR10 GCA 000830275.1 1
Micrarchaeum sp. GCA 001896515.1 1
Micrarchaeota archaeon GCA 001871475.1 1
Micrarchaeota archaeon GCA 001871655.1 1
Nanoarchaeum equitans GCA 000008085.1 1
Nanopusillus acidilobi GCA 001552015.1 1
archaeon CG2 30 31 98 GCA 001872885.1 1
Aenigmarchaeota archaeon GCA 000404305.1 1
Micrarchaeota archaeon GCA 001784635.1 1
Aenigmarchaeota archaeon GCA 001873985.1 1
Nanohaloarchaea archaeon GCA 001563915.1 1
Nanohaloarchaea archaeon GCA 001761425.1 1
Nanohaloarchaea archaeon GCA 001563995.1 1
Pacearchaeota archaeon GCA 001872185.1 1
Pacearchaeota archaeon GCA 001786425.1 1
Pacearchaeota archaeon GCA 001786395.1 1
archaeon GW2011 AR18 GCA 000806155.1 1
Woesearchaeota archaeon GCA 001872765.1 1
Woesearchaeota archaeon GCA 001871415.1 1
Methanobacterium sp. GCA 001729285.1 1
Methanobrevibacter curvatus GCA 001639295.1 1
Methanobacteriaceae archaeon GCA 001507955.1 1
Methanopyrus kandleri GCA 000007185.1 1
Methanotorris igneus GCA 000214415.1 1
Methanothermococcus okinawensis GCA 000179575.2 1
Methanococcus maripaludis GCA 000018485.1 1
Thermoplasmatales archaeon GCA 001595945.1 1
uncultured marine GCA 000246735.1 1
Marine Group GCA 001875345.1 1
Methanomassiliicoccus intestinalis GCA 000404225.1 1
Methanomethylophilus alvus GCA 000300255.2 1
Euryarchaeota archaeon GCA 001800815.1 1
Euryarchaeota archaeon GCA 001800825.1 1
Aciduliprofundum boonei GCA 000025665.1 1
Thermoplasmatales archaeon GCA 001856825.1 1
Thermoplasma volcanium GCA 000011185.1 1
Picrophilus torridus GCA 000008265.1 1
Euryarchaeota archaeon GCA 001914405.1 1
Methanoculleus bourgensis GCA 900079125.1 1
Halopenitus persicus GCA 900107205.1 1
Halorientalis regularis GCA 900102305.1 1
Halosimplex carlsbadense GCA 000337455.1 1
Halobacterium jilantaiense GCA 900110535.1 1
haloarchaeon HTSR1 GCA 001767315.1 1
Methanocella conradii GCA 000251105.1 1
Syntrophoarchaeum caldarius GCA 001766815.1 1
Ferroglobus placidus GCA 000025505.1 1
Archaeoglobus veneficus GCA 000194625.1 1
Archaeoglobus sulfaticallidus GCA 000385565.1 1
Methanosaeta thermophila GCA 000014945.1 1
Methanoperedens nitroreducens GCA 000685155.1 1
Methanosarcinales archeaon GCA 001508185.1 1
Methanolobus tindarius GCA 000504205.1 1
Methanosarcina barkeri GCA 000970305.1 1

Figure 56. Example of a gene with monophyletic eukaryotes (OG0000228).

Eukaryotes are highlighted in green.

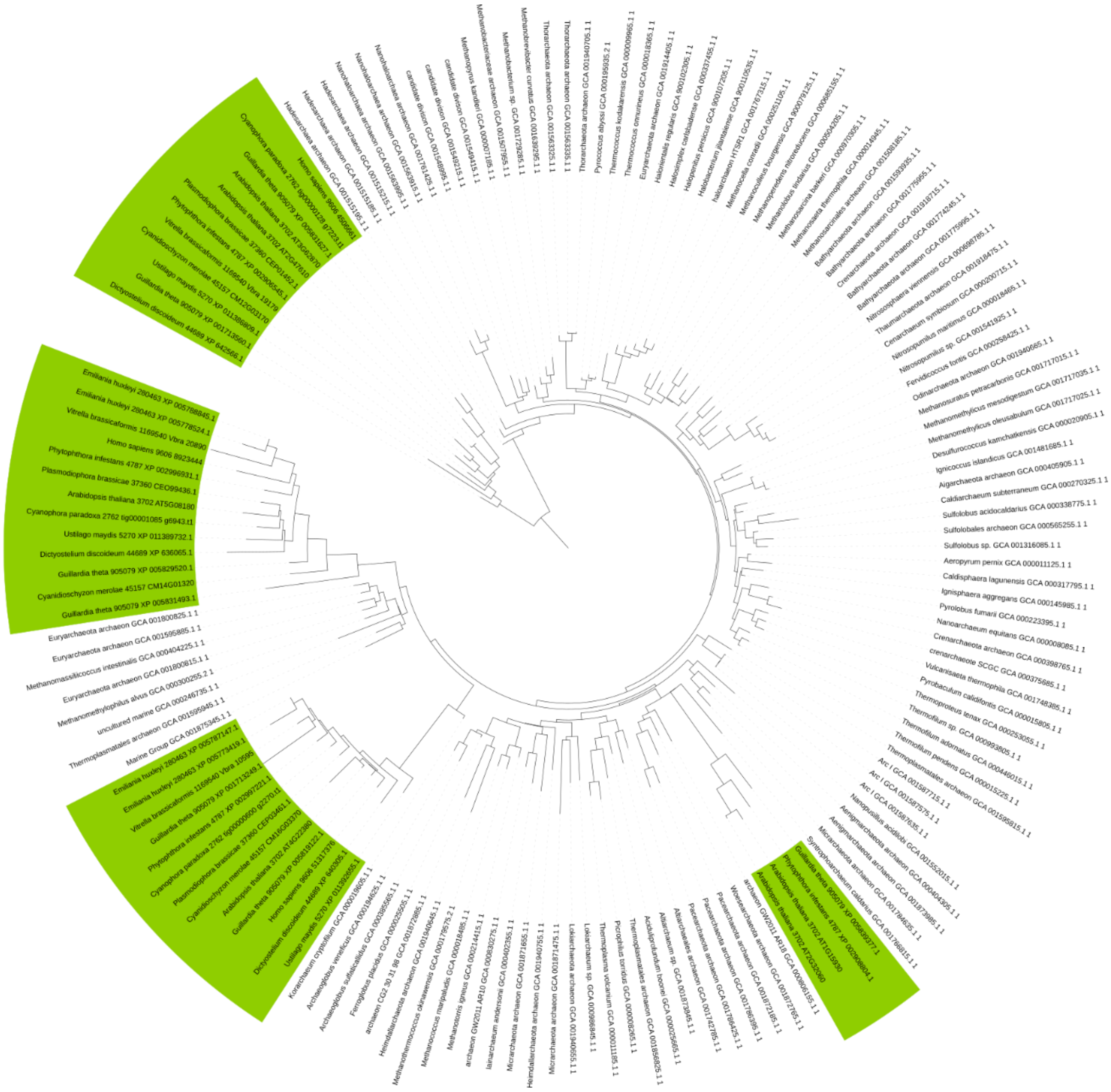

Dictyostelium discoideum 44689 XP 642566.1
Guillardia theta 905079 XP 001713560.1
Ustilago maydis 5270 XP 011386809.1
Cyanidioschyzon merolae 45157 CM12G03170
Vitrella brassicaformis 1169540 Vbra 19179
Phytophthora infestans 4787 XP 002906545.1
Plasmodiophora brassicae 37360 CEP01452.1
Arabidopsis thaliana 3702 AT2G47610
Arabidopsis thaliana 3702 AT3G62870
Guillardia theta 905079 XP 005831627.1
Cyanophora paradoxa 2762 tig00000128 g7223.t1
Homo sapiens 9606 4506661
Hadesarchaea archaeon GCA 001515195.1 1
Hadesarchaea archaeon GCA 001515185.1 1
Hadesarchaea archaeon GCA 001515215.1 1
Nanohaloarchaea archaeon GCA 001563995.1 1
Nanohaloarchaea archaeon GCA 001563915.1 1
candidate divison GCA 001761425.1 1
candidate divison GCA 001548995.1 1
candidate divison GCA 001549215.1 1
Methanopyrus kandleri GCA 001549415.1 1
Methanobacteriaceae archaeon GCA 000007185.1 1
Methanobacterium sp. GCA 001507955.1 1
Methanobrevibacter curvatus GCA 001729285.1 1
Thorarchaeota archaeon GCA 001639295.1 1
Thorarchaeota archaeon GCA 001563325.1 1
Thorarchaeota archaeon GCA 001563335.1 1
Thorarchaeota archaeon GCA 001940705.1 1
Pyrococcus abyssi GCA 000195935.2 1
Thermococcus kodakarensis GCA 000009965.1 1
Thermococcus omnurineus GCA 000018365.1 1
Euryarchaeota archaeon GCA 001914405.1 1
Halorientalis regularis GCA 900102305.1 1
Halosimplex carlsbadense GCA 000337455.1 1
Halopenitus persicus GCA 900107205.1 1
Halobacterium jilantaiense GCA 900110535.1 1
haloarchaeon HTSR1 GCA 001767315.1 1
Methanocella conradii GCA 000251105.1 1
Methanoculleus bourgensis GCA 900079125.1 1
Methanoperedens nitroreducens GCA 000685155.1 1
Methanolobus tindarius GCA 000504205.1 1
Methanosarcina barkeri GCA 000970305.1 1
Methanosaeta thermophila GCA 000014945.1 1
Methanosarcinales archaeon GCA 001508185.1 1
Bathyarchaeota archaeon GCA 001593935.1 1
Bathyarchaeota archaeon GCA 001775955.1 1
Crenarchaeota archaeon GCA 001918715.1 1
Bathyarchaeota archaeon GCA 001774245.1 1
Bathyarchaeota archaeon GCA 001775995.1 1
Nitrososphaera viennensis GCA 000698785.1 1
Thaumarchaeota archaeon GCA 001918475.1 1
Cenarchaeum symbiosum GCA 000200715.1 1
Nitrosopumilus maritimus GCA 000018465.1 1
Nitrosopumilus sp. GCA 001541925.1 1
Fervidicoccus fontis GCA 000258425.1 1
Odinarchaeota archaeon GCA 001940665.1 1
Methanosuratus petracarbonis GCA 001717015.1 1
Methanomethylicus mesodigestum GCA 001717035.1 1
Methanomethylicus oleusabulum GCA 001717025.1 1
Desulfurococcus kamchatkensis GCA 000020905.1 1
Ignicoccus islandicus GCA 001481685.1 1
Aigarchaeota archaeon GCA 000405905.1 1
Caldiarchaeum subterraneum GCA 000270325.1 1
Sulfolobus acidocaldarius GCA 000338775.1 1
Sulfolobales archaeon GCA 000565255.1 1
Sulfolobus sp. GCA 001316085.1 1
Aeropyrum pernix GCA 000011125.1 1
Caldisphaera lagunensis GCA 000317795.1 1
Ignisphaera aggregans GCA 000145985.1 1
Pyrolobus fumarii GCA 000223395.1 1
Nanoarchaeum equitans GCA 000008085.1 1
Crenarchaeota archaeon GCA 000398765.1 1
crenarchaeote SCGC GCA 000375685.1 1
Vulcanisaeta thermophila GCA 001748385.1 1
Pyrobaculum calidifontis GCA 000015805.1 1
Thermoproteus tenax GCA 000253055.1 1
Thermofilum sp. GCA 000993805.1 1
Thermofilum adornatus GCA 000446015.1 1
Thermofilum pendens GCA 000015225.1 1
Thermoplasmatales archaeon GCA 001595815.1 1
Arc I GCA 001587715.1 1
Arc I GCA 001587575.1 1
Arc I GCA 001587635.1 1
Nanopusillus acidilobi GCA 001552015.1 1
Aenigmarchaeota archaeon GCA 000404305.1 1
Aenigmarchaeota archaeon GCA 001873985.1 1
Micrarchaeota archaeon GCA 001784635.1 1
Syntrophoarchaeum caldarius GCA 001766815.1 1
Guillardia theta 905079 XP 005839377.1
Phytophthora infestans 4787 XP 002908804.1
Arabidopsis thaliana 3702 AT1G15930
Arabidopsis thaliana 3702 AT2G32060
archaeon GW2011 AR18 GCA 000806155.1 1
Woesearchaeota archaeon GCA 001872765.1 1
Pacearchaeota archaeon GCA 001872185.1 1
Pacearchaeota archaeon GCA 001786395.1 1
Pacearchaeota archaeon GCA 001786425.1 1
Altiarchaeales archaeon GCA 001742785.1 1
Altiarchaeum sp. GCA 001873845.1 1
Aciduliprofundum boonei GCA 000025665.1 1
Thermoplasmatales archaeon GCA 001856825.1 1
Picrophilus torridus GCA 000008265.1 1
Thermoplasma volcanium GCA 000011185.1 1
Lokiarchaeum sp. GCA 000986845.1 1
Lokiarchaeota archaeon GCA 001940655.1 1
Micrarchaeota archaeon GCA 001871475.1 1
Heimdallarchaeota archaeon GCA 001940755.1 1
Micrarchaeota archaeon GCA 001871655.1 1
Iainarchaeum andersonii GCA 000402355.1 1
archaeon GW2011 AR10 GCA 000830275.1 1
Methanotorris igneus GCA 000214415.1 1
Methanococcus maripaludis GCA 000018485.1 1
Methanothermococcus okinawensis GCA 000179575.2 1
Heimdallarchaeota archaeon GCA 001940645.1 1
archaeon CG2 30 31 98 GCA 001872885.1 1
Ferroglobus placidus GCA 000025505.1 1
Archaeoglobus sulfaticallidus GCA 000385565.1 1
Archaeoglobus veneficus GCA 000194625.1 1
Korarchaeum cryptofilum GCA 000019605.1 1
Ustilago maydis 5270 XP 011392655.1
Dictyostelium discoideum 44689 XP 640305.1
Homo sapiens 9606 51317376
Guillardia theta 905079 XP 005819122.1
Arabidopsis thaliana 3702 AT4G22380
Cyanidioschyzon merolae 45157 CM16G03370
Plasmodiophora brassicae 37360 CEP03461.1
Cyanophora paradoxa 2762 tig00000600 g2270.t1
Phytophthora infestans 4787 XP 002997221.1
Guillardia theta 905079 XP 001713249.1
Vitrella brassicaformis 1169540 Vbra 10595
Emiliania huxleyi 280463 XP 005773419.1
Emiliania huxleyi 280463 XP 005787147.1
Marine Group GCA 001875345.1 1
Thermoplasmatales archaeon GCA 001595945.1 1
uncultured marine GCA 000246735.1 1
Methanomethylophilus alvus GCA 000300255.2 1
Euryarchaeota archaeon GCA 001800815.1 1
Methanomassiliicoccus intestinalis GCA 000404225.1 1
Euryarchaeota archaeon GCA 001595885.1 1
Euryarchaeota archaeon GCA 001800825.1 1
Guillardia theta 905079 XP 005831493.1
Cyanidioschyzon merolae 45157 CM14G01320
Guillardia theta 905079 XP 005829520.1
Dictyostelium discoideum 44689 XP 636065.1
Ustilago maydis 5270 XP 011389732.1
Cyanophora paradoxa 2762 tig00001085 g6943.t1
Arabidopsis thaliana 3702 AT5G08180
Plasmodiophora brassicae 37360 CEO99436.1
Phytophthora infestans 4787 XP 002996931.1
Homo sapiens 9606 8923444
Vitrella brassicaformis 1169540 Vbra 20890
Emiliania huxleyi 280463 XP 005778524.1
Emiliania huxleyi 280463 XP 005788845.1

Figure 57. Example of a gene with duplicated and polyphyletic eukaryotes (OG0000343). Eukaryotes are highlighted in green.

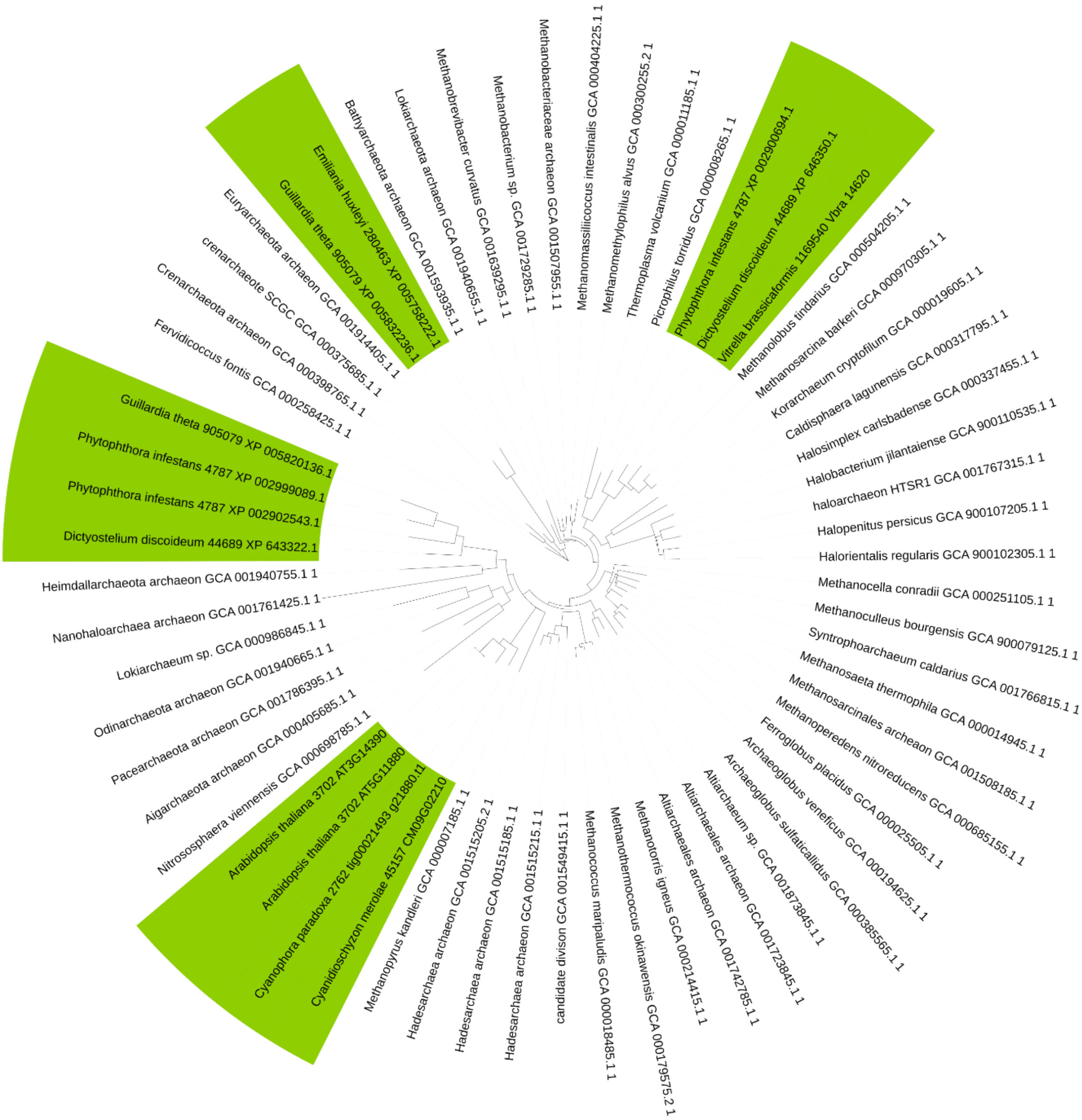

Methanobacteriaceae archaeon GCA 001507955.1 1
Methanomassiliicoccus intestinalis GCA 000404225.1 1
Methanomethylophilus alvus GCA 000300255.2 1
Thermoplasma volcanium GCA 000011185.1 1
Picrophilus torridus GCA 000008265.1 1
Phytophthora infestans 4787 XP 002900694.1
Dictyostelium discoideum 44689 XP 646350.1
Vitrella brassicaformis 1169540 Vbra 14620
Methanolobus tindarius GCA 000504205.1 1
Methanosarcina barkeri GCA 000970305.1 1
Korarchaeum cryptofilum GCA 000019605.1 1
Caldisphaera lagunensis GCA 000317795.1 1
Halosimplex carlsbadense GCA 000337455.1 1
Halobacterium jilantaiense GCA 900110535.1 1
haloarchaeon HTSR1 GCA 001767315.1 1
Halopenitus persicus GCA 900107205.1 1
Halorientalis regularis GCA 900102305.1 1
Methanocella conradii GCA 000251105.1 1
Methanoculleus bourgensis GCA 900079125.1 1
Syntrophoarchaeum caldarius GCA 001766815.1 1
Methanosaeta thermophila GCA 000014945.1 1
Methanosarcinales archeaon GCA 001508185.1 1
Methanoperedens nitroreducens GCA 000685155.1 1
Ferroglobus placidus GCA 000025505.1 1
Archaeoglobus veneficus GCA 000194625.1 1
Archaeoglobus sulfaticallidus GCA 000385565.1 1
Altiarchaeum sp. GCA 001873845.1 1
Altiarchaeales archaeon GCA 001723845.1 1
Altiarchaeales archaeon GCA 001742785.1 1
Methanotorris igneus GCA 000214415.1 1
Methanothermococcus okinawensis GCA 000179575.2 1
Methanococcus maripaludis GCA 000018485.1 1
candidate divison GCA 001549415.1 1
Hadesarchaea archaeon GCA 001515215.1 1
Hadesarchaea archaeon GCA 001515185.1 1
Hadesarchaea archaeon GCA 001515205.2 1
Methanopyrus kandleri GCA 000007185.1 1
Cyanidioschyzon merolae 45157 CMO9G02210
Cyanophora paradoxa 2762 tig00021493 g21880.t1
Arabidopsis thaliana 3702 AT5G11880
Arabidopsis thaliana 3702 AT3G14390
Nitrososphaera viennensis GCA 000698785.1 1
Aigarchaeota archaeon GCA 000405685.1 1
Pacearchaeota archaeon GCA 001786395.1 1
Odinarchaeota archaeon GCA 001940665.1 1
Lokiarchaeum sp. GCA 000986845.1 1
Nanohaloarchaea archaeon GCA 001761425.1 1
Heimdallarchaeota archaeon GCA 001940755.1 1
Dictyostelium discoideum 44689 XP 643322.1
Phytophthora infestans 4787 XP 002902543.1
Phytophthora infestans 4787 XP 002999089.1
Guillardia theta 905079 XP 005820136.1
Fervidicoccus fontis GCA 000258425.1 1
Crenarchaeota archaeon GCA 000398765.1 1
crenarchaeote SCGC GCA 000375685.1 1
Euryarchaeota archaeon GCA 001914405.1 1
Guillardia theta 905079 XP 005832236.1
Emiliania huxleyi 280463 XP 005758222.1
Bathyarchaeota archaeon GCA 001593935.1 1
Lokiarchaeota archaeon GCA 001940655.1 1
Methanobrevibacter curvatus GCA 001639295.1 1
Methanobacterium sp. GCA 001729285.1 1

Figure 58. Example of a gene where eukaryotes are polyphyletic (OG0000773). Eukaryotes are highlighted in green.

We also observed that the archaeon *Heimdallarchaeota archaeon* GCA001940645.1 could be a metagenome contaminated by a eukaryote. Indeed, this organism systematically inserts itself among the eukaryotes, making the *Heimdallarchaeota* polyphyletic. After individual analysis of each tree, the sequences were cleaned to correct these issues (cf. **Supp.Mat genes-polyphyletiques-causes.txt** Thus, to limit what could be an issue of incorrect sequences, we eliminated the genes where we could not obtain the monophyly of the eukaryotes (corresponding to 9 genes from case 2 and the 3 genes from case 3). We were left with 126 genes. We then added with 42 the genome of *Anaerobic archaeon_MK-D1*, a newly sequenced Asgard archaeon (strain MK-D1(Imachi et al., 2020)). Unlike the Asgard sequences we had previously, this one is not a metagenome but a true genome from a cultivated organism. This genome is of interest because it is the only Asgard to have been cultivated. Its sequencing can thus help assess the accuracy and completeness of sequences obtained from metagenomes.

4 REMOVAL OF SPECIES

Since the beginning of our study, we have observed a singularity in *Heimdallarchaeota archaeon* GCA001940645.1. Already in our archaeal phylogeny, this organism stood out due to the uncertainty regarding its phylogenetic position. In our trees including the eukaryotes, this archaeon seems to be closer to eukaryotes than to other *Heimdallarchaeota*. We suspect that this singularity may be the result of contamination or a new lineage. To evaluate the impact of this genome on our phylogeny, we created two new super-matrices: one where we removed *Heimdallarchaeota archaeon* GCA001940645.1 and one where we removed all *Heimdallarchaeota*. We then calculated trees using the PMSF method for our 5 replicates (cf. **Supp.Mat archaea+eukaryota_126_genes**).

We observe no differences in the position of the eukaryotes in our trees when removing the *Heimdallarchaeota*. Our results suggest that *Heimdallarchaeota archaeon* GCA001940645 is a metagenome contaminated by a eukaryote. This removal shows that independently of the *Heimdallarchaeota*, the position of the eukaryotes in this region of the tree does not change, even though the branch is elongated. If their position had changed, we would have questioned whether opposing signals were conflicting. But in our case, the addition of the *Heimdallarchaeota* tends to reinforce the position of the eukaryotes.

5 EFFECT OF OUR GENE SELECTION AND THE PRESENCE OF EUKARYOTES ON THE ARCHAEAL TOPOLOGIES OBTAINED IN CHAPTER 2

We wondered about the potential effects of adding eukaryotes to an archaeal super-matrix. If, on the same replicate with the same method, there is a topological change between the tree of archaeans alone and the tree with eukaryotes, is this due to the presence of eukaryotes or to gene sampling related to the inclusion of eukaryotes? To answer this, we decided to create for our 5 replicates a super-matrix of our 126 genes without the eukaryotes, then calculate the corresponding tree using PMSF to compare its topology with those previously obtained with 343 genes. We then calculated the differences between all our bipartitions of the datasets with 343 and 126 genes in order to assess the impact of this gene subsampling (**Figure 59** & **FigS59**). If these differences in absolute values tend toward 0, then the topologies coincide, and the differences observed after the addition of eukaryotes can be interpreted as a result of their inclusion. Otherwise, the observed differences can be interpreted as a result of a particular case of gene subsampling, possibly combined with the addition of eukaryotes. In this case, it would be difficult to draw a conclusion. As we concluded in the previous chapter, our gene subsampling has very little impact on the bipartitions found, reinforcing the stability of our topologies identified in Chapter 2. As seen in the previous chapter, the topology of the archaeans mainly depends on the replicate.

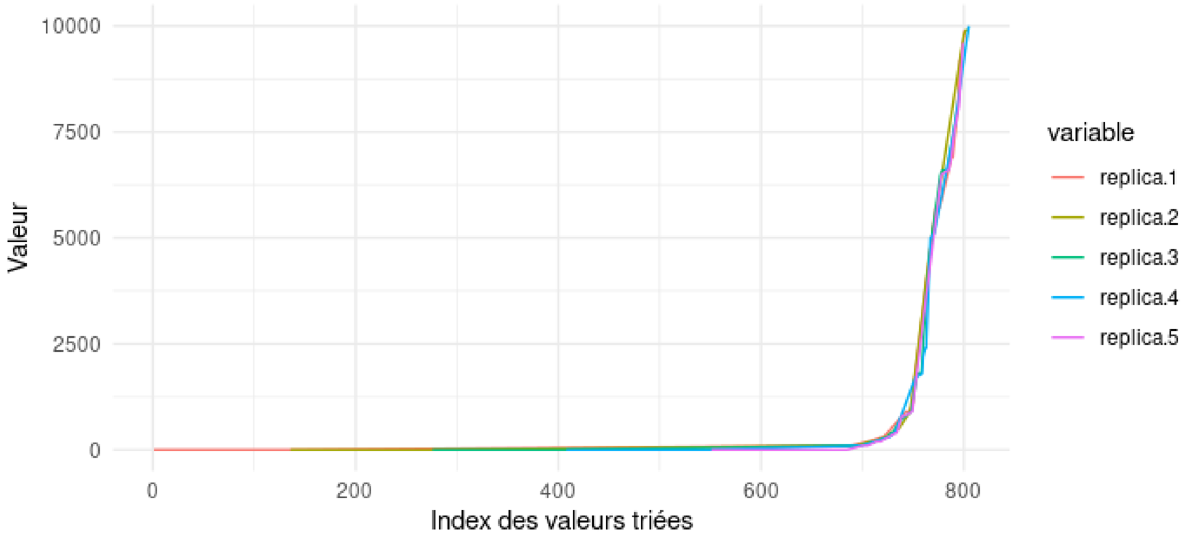


Figure 59. Differences between the bipartitions of our selection of 343 archaea genes with our selection of 126 genes.

Data are provided in **Supp.Mat FigS59**. Values tending towards 0 indicate that the topologies between the selection of 343 genes and the 126 genes coincide. The observed differences are interpreted either as the result of a particular case of gene subsampling.

We then calculated a collection of trees from our super-matrices, this time including the eukaryotes, comprising an LG4X tree, LG+C20+F+G, and LG+C60+F+G, and PMSF for each of the 5 replicates. Following the same protocol as before, evaluating the bipartitions of the archaeal groups found, we observed that in our case, the addition of eukaryotes does not change the topology of the archaeal tree found in the previous chapter. We obtain the same tree, simply with eukaryotes added as a sister group to the Asgard group.

However, this solution still seems too simplistic, and we suspect that this supposed relatedness between Asgards and eukaryotes is the result of underlying, hidden, and more complex processes. To verify, we will exploit our dataset using various strategies.

For the next steps of the protocol, the trees were systematically calculated using the PMSF method, with a tree computed using the LG+C60+F+G model as the guide tree.

6 HETEROGENEITY OF SUBSTITUTION OVER TIME

One of the issues that can arise in tree resolution is the portion of branches with different properties. These branches, with properties that differ so much from the rest of the tree, could impact the overall tree topology by attracting or repelling other branches. Thus, the phylogenetic signal of substitutions affecting certain columns in our alignments is overshadowed by deeper and subtler processes related to substitution heterogeneity over time. These could very well have occurred during the archaeal/eukaryotic transition, thus distorting the phylogenetic signal contained in each column, such as in cases where the proportions of invariant sites in unrelated species converge. Two properties can vary:

- the substitution rate over time (heterotachy or covarion);
- the profile, that is, the distribution of acceptable amino acids at a given position (heteropecilly).

6.1 HETEROTACHY (COVARION) : HETEROGENEITY OF SUBSTITUTION RATES AT A SITE OVER TIME

It is possible that if certain groups of archaea or eukaryotes have a very different substitution rate from other species in our tree, this could lead to phylogenetic reconstruction artifacts. To verify this hypothesis, we will infer site-specific evolutionary rates independently for archaea and eukaryotes. A calculation of rate deltas will then allow us to progressively filter columns to eliminate in order to create super-matrices with columns having more or less heterogeneous evolutionary rates. We will then calculate trees by progressively introducing sites as they evolve more differently.

We used the alignments from the dataset where we removed the genes where eukaryotes are polyphyletic. Columns having at least one representative from each major archaeal taxonomic group were retained, namely: one eukaryote, one DPANN, one Euryarchaeota, one TACK, and one Asgard. We thus have 126 genes, corresponding to approximately 33,000 positions for each replica. The goal is to maintain super-matrices of archaea and eukaryotes of identical size while playing with the subset of archaea used to calculate the rates. We then divided the super-matrices into two parts, separating archaea on one side and eukaryotes on the other. The rates for each position were calculated separately for the eukaryote and archaeal super-matrices using IQ-Tree, and the deltas between each site were then calculated between these two groups. We used 4 bins (one per gamma category) since the rates were an approximate version of these categories. For each site in our alignments, we calculated the substitution rate difference between our eukaryotes and archaea (taken either as a whole or after the removal of some of our previously mentioned subgroups) (cf. **rates**). We then also performed species removal manipulations on the archaeal super-matrices by alternately removing Euryarchaeota, DPANN, TACK, and Heimdallarchaeota archaeon GCA001940645 (which, as a reminder, tends to systematically insert itself within eukaryotes). These differences were then sorted from sites with the smallest (homotachy) to the largest (heterotachy) substitution rate differences. This allows us to distinguish sites with large rate differences between archaea and eukaryotes from sites with smaller differences. The results are given in **Figure 60** & **Supp.Mat FigS60**. No differences are observed in the distribution of substitution rate differences according to the archaeal set used. Therefore, there is no variation in substitution rates for certain subgroups. The overall distribution pattern of deltas does not seem affected by the subsampling of archaea, but this does not mean that the bin composition is identical in terms of sites from the original super-matrix. So, even though we do not expect variations a priori, we will verify this by calculating the different trees with all species, alternating substitution rate files corresponding to our different sets of archaea. We will then be able to use our entire archaeal set as a reference to create sub-super-matrices, without worrying about variations that may have been intrinsic to certain subgroups. Additionally, no differences seem to appear based on our replicas.

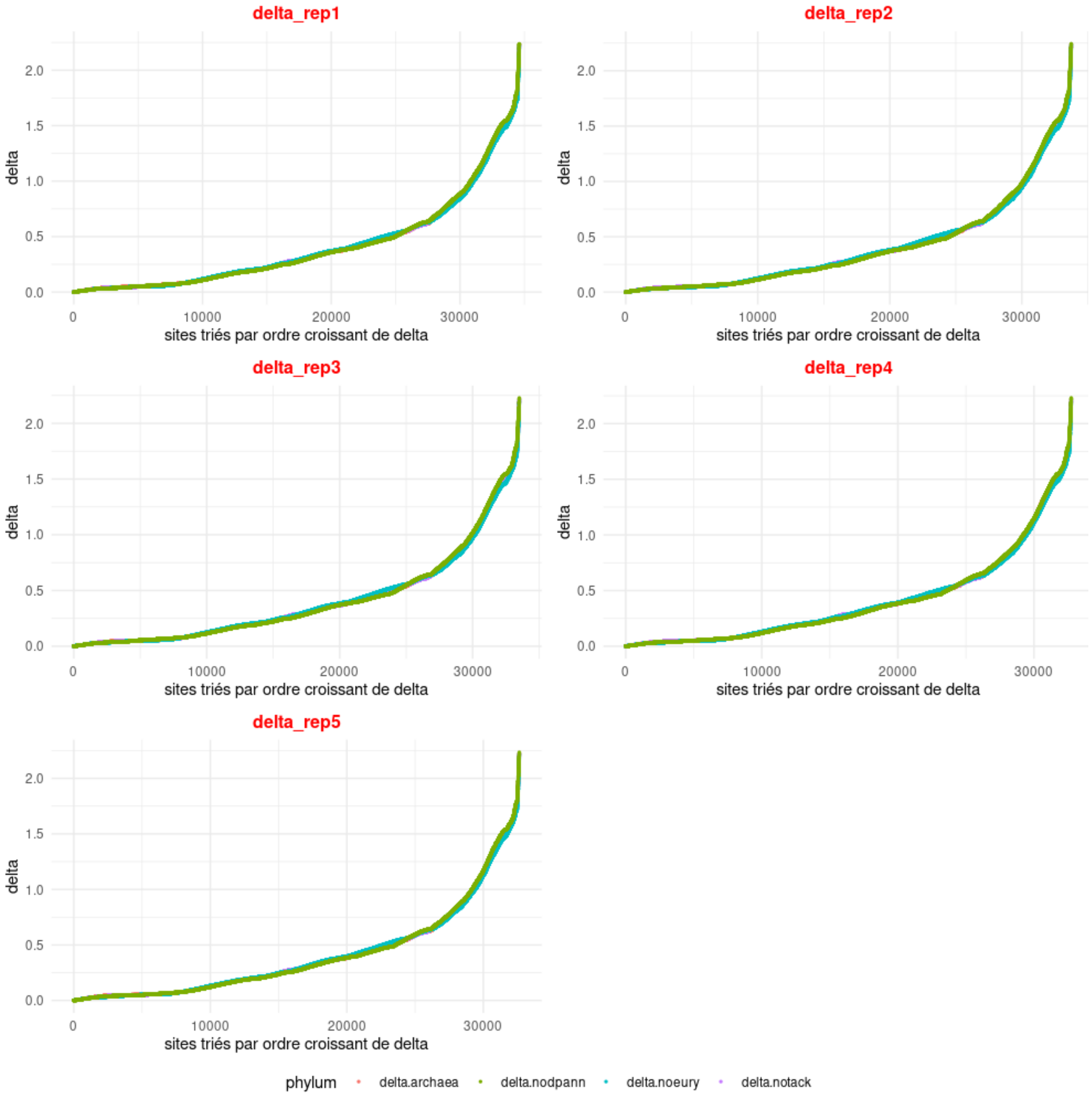


Figure 60. Variation in substitution rate per site between our different archaea sets and our eukaryotes.

Data are provided in **Supp.Mat FigS60**. The overall distribution pattern of the deltas does not seem to be affected by the archaea subsampling, but this does not mean that the composition of the bins is identical in terms of the sites of the original supermatrix. Moreover, there do not appear to be any differences between our replicas.

Our supermatrices range from 30,000 to 35,000 positions. We chose to create cumulative bins of approximately 6,000 positions based on these substitution rates, containing all species (both eukaryotes and archaea). Two types of bins were created. First, we used cumulative bins that progressively add columns as the deltas between substitution rates increase. Second, we did not cumulate the bins to observe trees built from alignments with different substitution rate deltas. These trees were calculated using LG4X. This model was chosen for two reasons: (1) first, we want to use a model that will be sensitive to artifacts so that they can be more easily identified, and (2) we are limited by computational time constraints. We observe that the same Heimdallarchaeota archaeon GCA001940645.1 is still causing problems (cf. **parse-consense**).

Analyses of our cumulative and non-cumulative bins (**Figure 61, Figure 62** & **Supp.Mat heterotachie**) show that for the most homogeneous sites in terms of rates between archaea and eukaryotes, the grouping of Asgards with eukaryotes is only very rarely represented (less than 2,500 trees out of 10,000), in favor of a grouping of Asgards with TACK (more than 7,500 trees out of 10,000). It is only when the super-matrices reach a size of 18,000 positions with the addition of sites with more heterogeneous substitution rates that the Asgards are systematically grouped with the eukaryotes. We can thus assume that the grouping of Asgards with eukaryotes could be the result of a phylogenetic reconstruction bias related to sites exhibiting poorly modeled heterotachy. In other words, this fortuitous grouping would be the consequence of a systematic error producing a false and statistically supported non-phylogenetic signal that competes with the true phylogenetic signal.

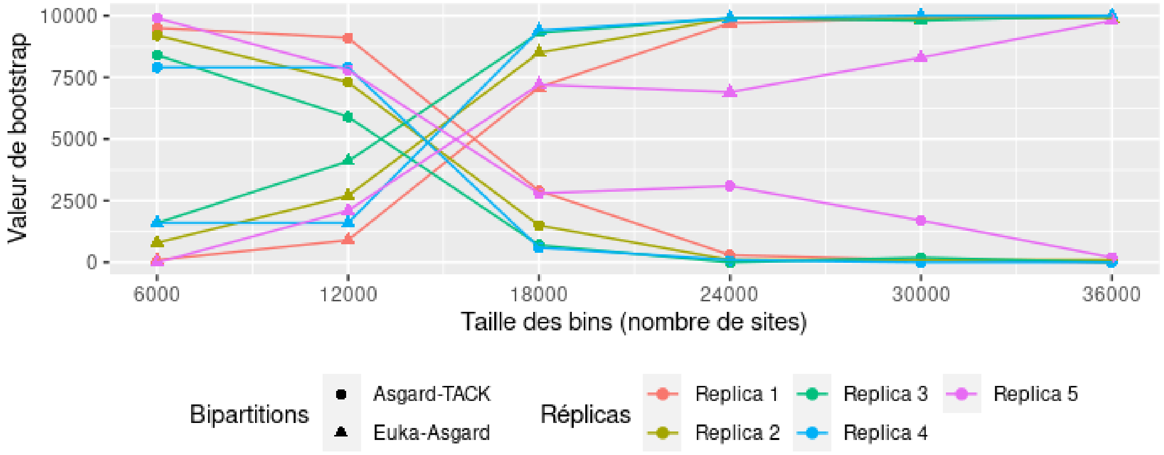


Figure 61. Proportion of trees indicating Asgard position according to cumulative bins size for each replica according to the LG4X model.

Data are provided in **Supp.Mat heterotachie**. A super-matrix size of 18,000 positions must be reached with the addition of sites with more heterogeneous substitution rates for Asgards to be systematically grouped with eukaryotes. We can therefore assume that the grouping of Asgard with eukaryotes may be the result of a phylogenetic reconstruction bias linked to poorly modelled heterotachy sites.

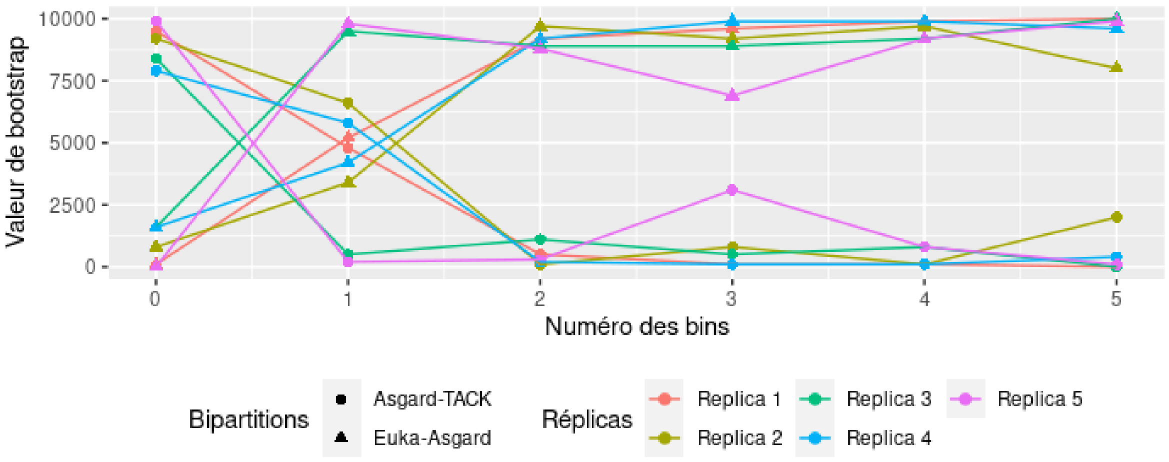


Figure 62. Proportion of trees indicating Asgard position according to non-cumulative bins size for each replica according to the LG4X model.

Data are provided in **Supp.Mat heterotachie**. More homogeneous bins tend to favor Asgards as a sister group to TACKs, while more heterogeneous sites support the hypothesis of eukaryotes as a sister group to Asgards.

Two examples of phylogenetic tree representations corresponding to the two possible topologies between eukaryotes, TACK and Asgard, are shown in **Figure 63** and **Figure 64**.

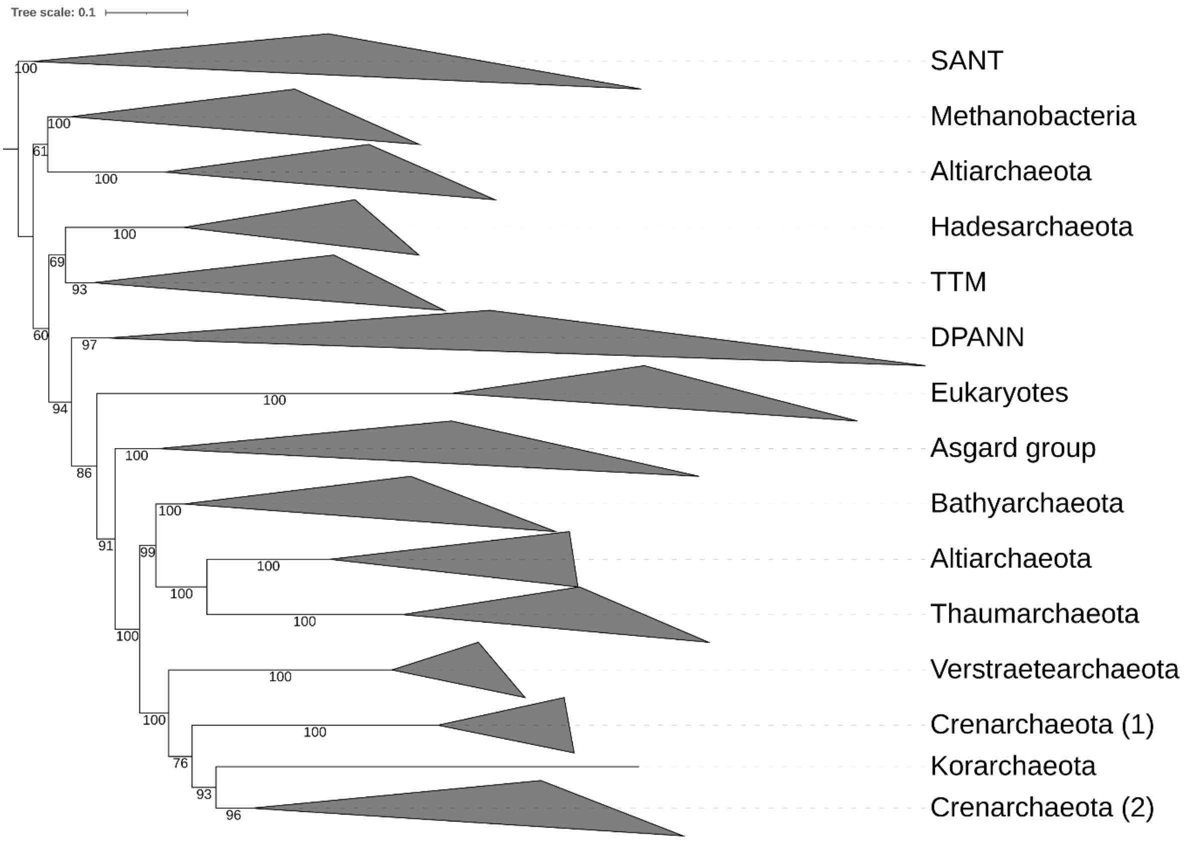


Figure 63. Tree corresponding to cumulative bins at 12,000 positions for replica 1 (132 species) according to the LG4X model.

Data are provided in **Supp.Mat heterotachie**. In this tree, eukaryotes are at the base of a group comprising Asgard and TACK, which are sister groups.

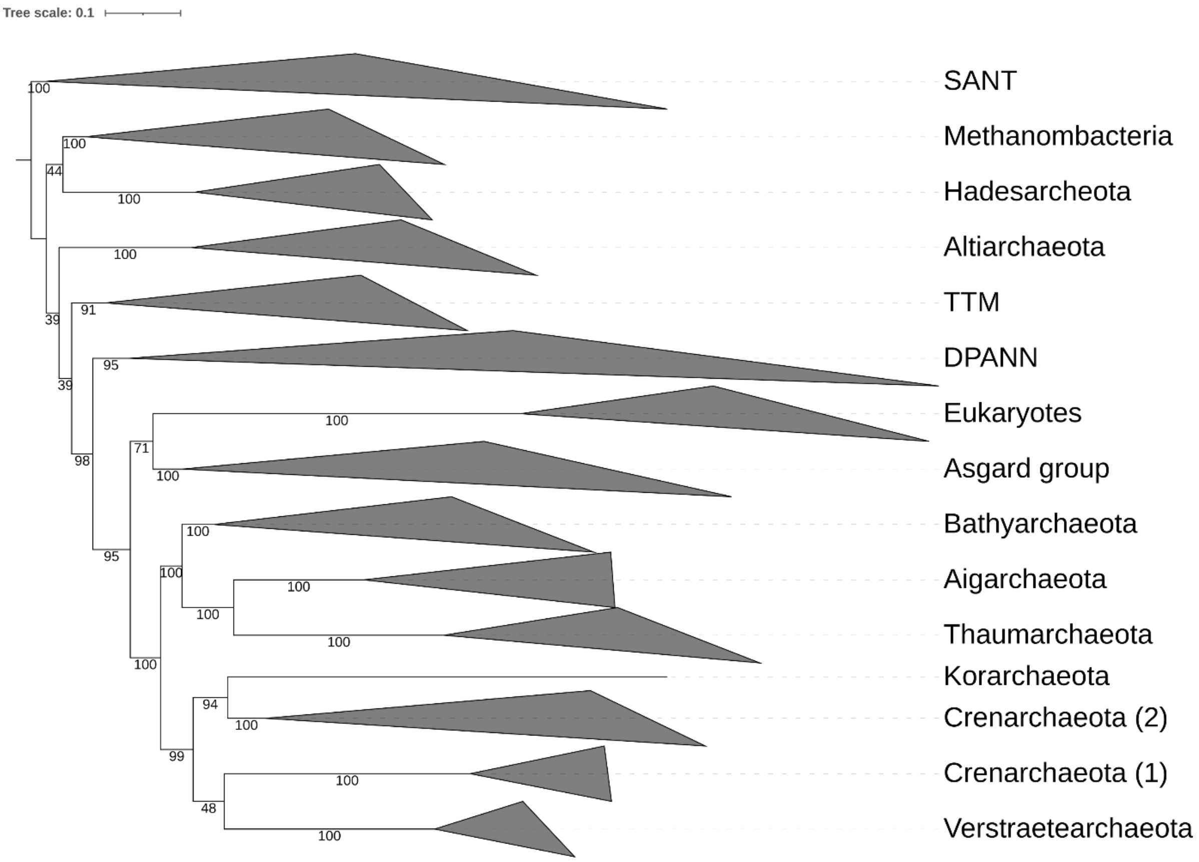


Figure 64. Tree corresponding to cumulative bins at 18,000 positions for replica 1 (132 species) according to the LG4X model.

Data are provided in **Supp.Mat heterotachie**. In this tree, eukaryotes are the sister group of the Asgard group.

We also constructed new trees using the PMSF model from cumulative bins whose rates were determined after the alternative removal of DPANN (**Figure 65**), then Euryarchaeota (**Figure 66**) and finally TACK (**Figure 67**). In all cases, we can clearly see that the rapprochement of eukaryotes with Asgards results from the use of heterotachous sites, supporting our interpretation that this hypothesis is erroneous and likely the result of an artifact.

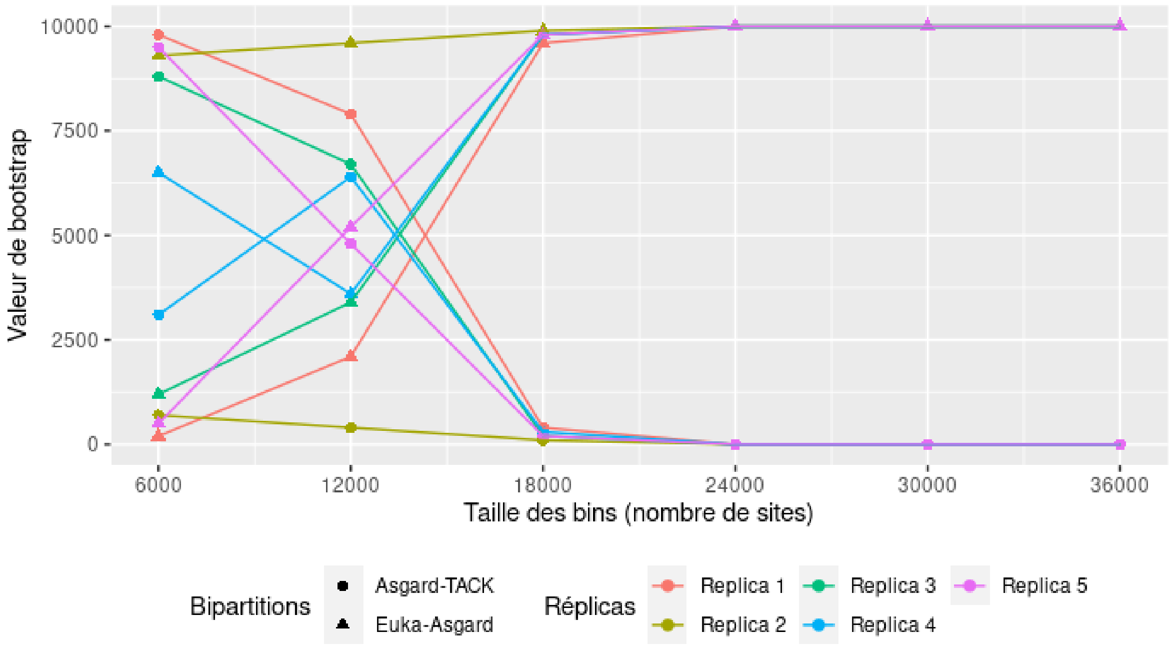


Figure 65. Proportion of trees according to PMSF model from cumulative bins whose velocities were determined after DPANN removal.

Data are provided in **Supp.Mat heterotachie**. With the PMSF model, we obtain the same result as the LG4X model, namely that we need to reach a super-matrix size of 18,000 positions with the addition of sites with more heterogeneous substitution rates for Asgards to be systematically grouped with eukaryotes.

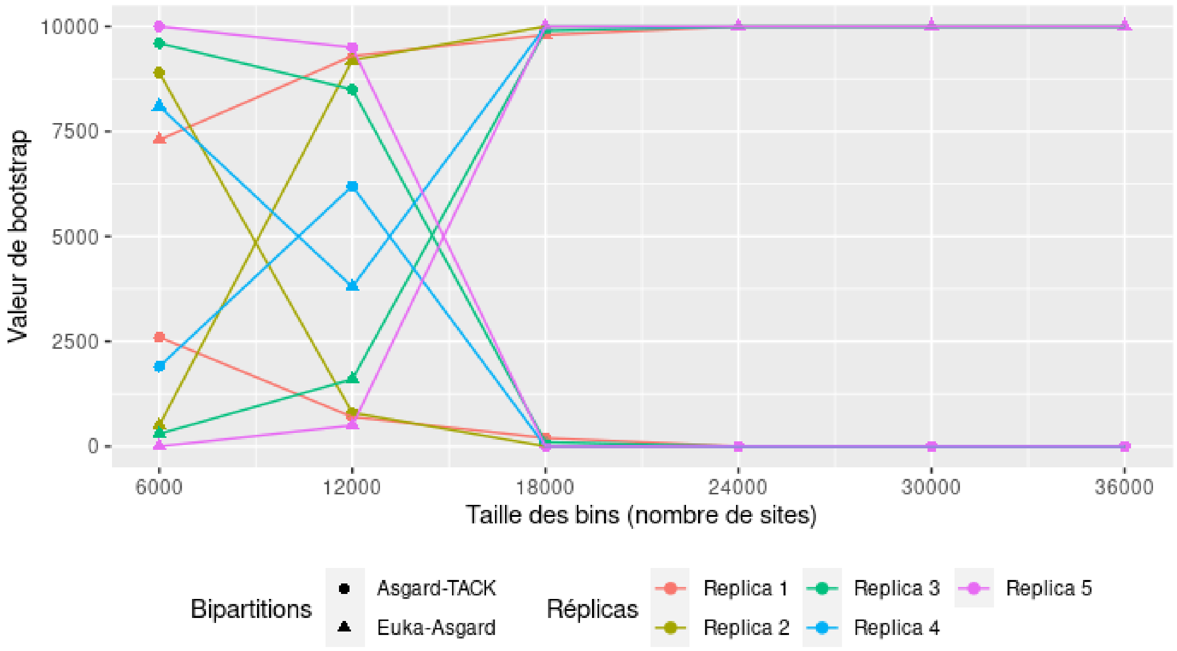


Figure 66. Proportion of trees according to PMSF model from cumulative bins whose velocities were determined after Euryarchaeota removal.

Data are provided in **Supp.Mat heterotachie**. Once again, the more homogeneous bins tend to favor Asgard as a sister group to TACK, while the more heterogeneous sites support eukaryotes as a sister group to Asgard. Only replica 1 supports eukaryotes as a sister group to Asgard. Replica 4, on the other hand, is difficult to interpret and seems to depend on the bin used.

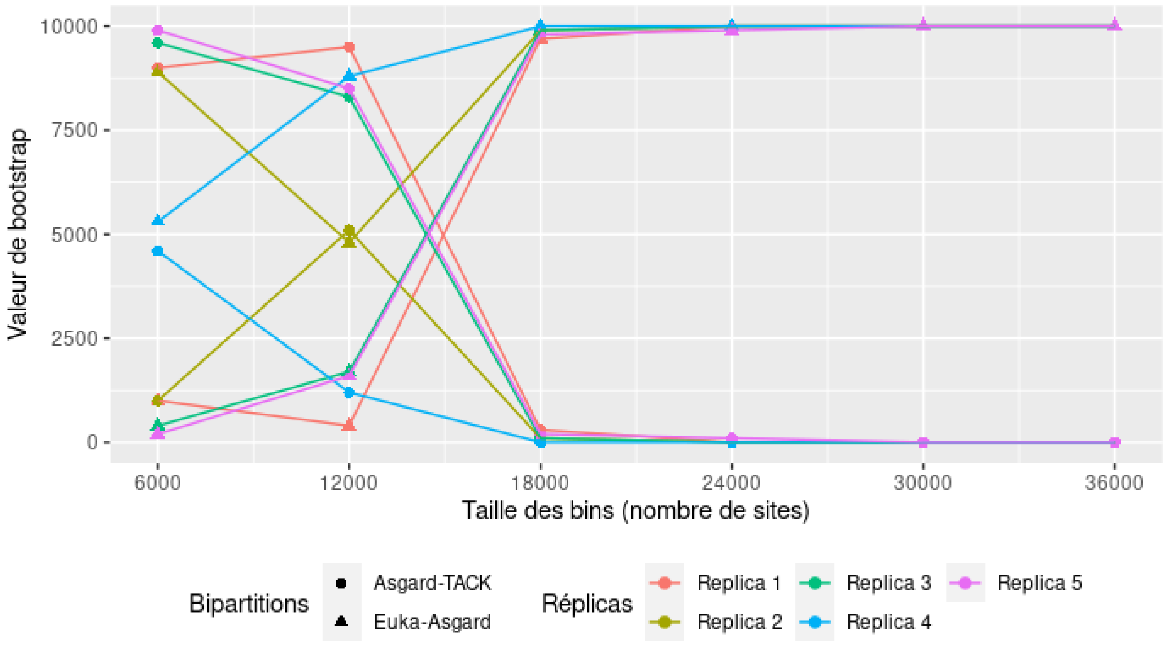


Figure 67. Proportion of trees according to PMSF model from cumulative bins whose velocities were determined after TACK removal.

Data are provided in **Supp.Mat heterotachie**. Once again, the more homogeneous bins tend to favor Asgard as a sister group to TACK, while the more heterogeneous sites support eukaryotes as a sister group to Asgard. Only replica 1 supports eukaryotes as a sister group to Asgard. Replica 4, on the other hand, is difficult to interpret and seems to depend on the bin used.

6.2 HETEROPECILLY: HETEROGENEITY OF SUBSTITUTION PROCESS OF A SITE OVER TIME

Finally, we wanted to check whether all amino acids are potentially possible for a given column. For our 5 species replicates, we separately estimated mixture model parameters for archaea and eukaryotes, and then deduced the frequency profile specific to each site. These profiles are based, for each site, on a Chi-square comparing the amino acid frequency distributions between the super-matrices. We then generated 6 cumulative bins per replicate and calculated the trees using the PMSF method (**Figure 68** & **Supp.Mat heteropecillie**). However, our results show no difference between our bins. The eukaryotes are consistently grouped as sister to the Asgards with maximal ultrafast-bootstrap values regardless of the replicate and bin.

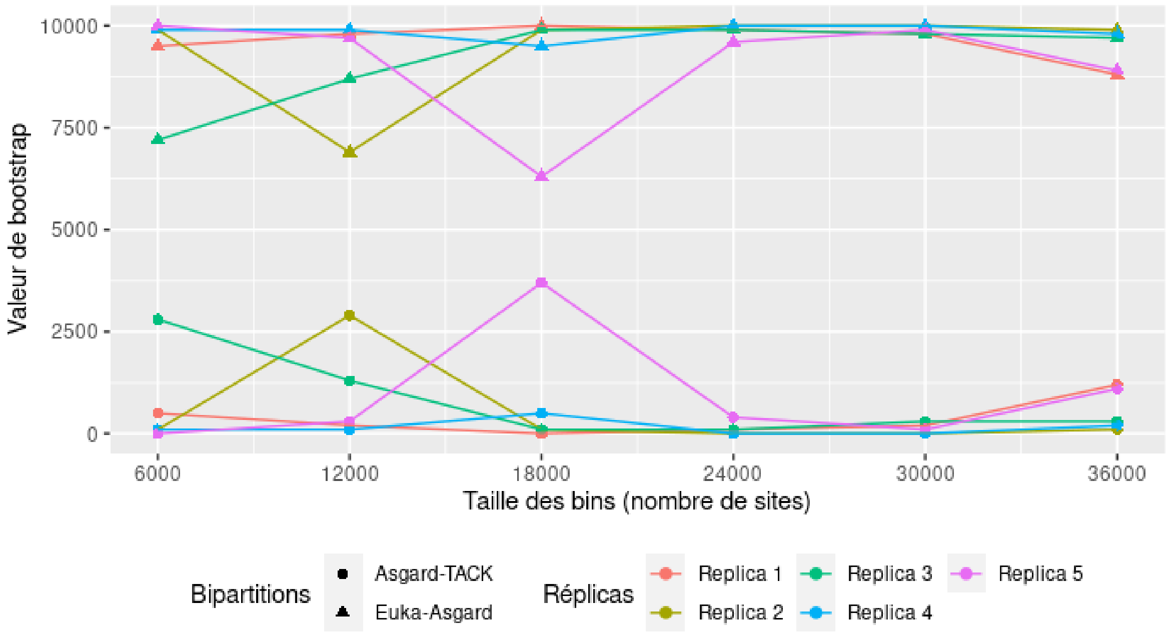


Figure 68. Proportion of trees indicating Asgard position according to cumulative bins size for each replica (132 species) according to the PMSF model.

Data are provided in **Supp.Mat heteropecillie**. Eukaryotes are systematically the sister group of Asgard.

We then wanted to check that all bins indeed gave the same result when taken separately. To do this, we decided to generate non-cumulative bins and calculate trees using the PMSF method, which is better suited to combat artifacts but theoretically more sensitive to heteropecilly (**Figure 69**). This time, our results show that the last bin does not recover the monophyly of eukaryotes + Asgards but rather favors the Asgards as sister group to the TACK (the ultrafast-bootstrap values for this grouping in the last bin are over 90%). It therefore appears that PMSF does not resist the phenomenon of heteropecilly. The replicate (and thus the species selection) and the model also seem to play a significant role. This shows that heteropecilly is an important phenomenon to consider when deciding to insert eukaryotes within archaea.

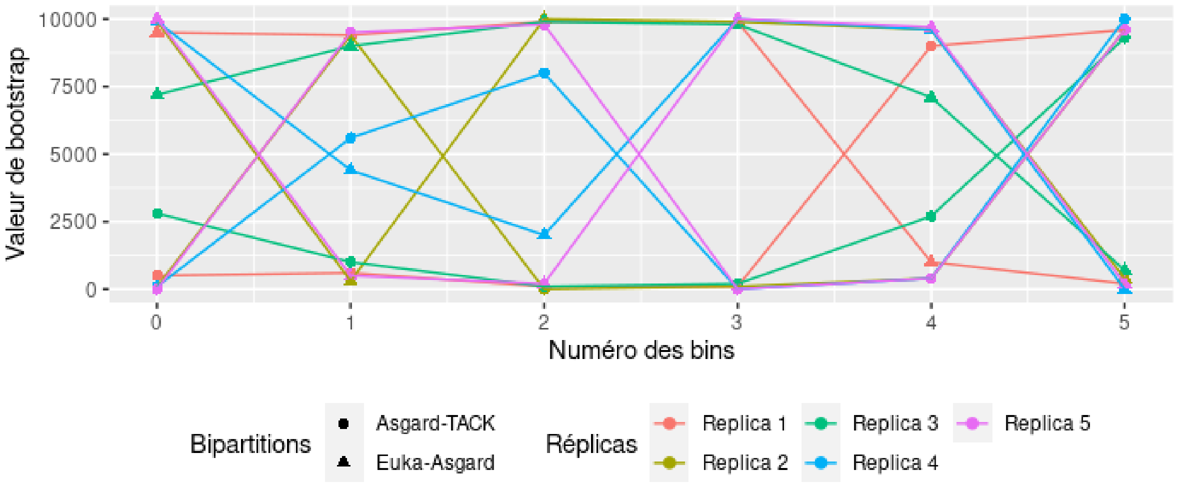


Figure 69. Proportion of trees indicating Asgard position according to non-cumulative bins for each replica (132 species) with the PMSF method.

Data are provided in **Supp.Mat heteropecillie**. Eukaryotes are grouped with Asgards for all bins except the last.

7 ROOTING OF ARCHAEA

Following my phylogenetic analyses of Archaea trees, it is crucial to root these trees in order to better understand the evolutionary relationships and determine the common origin of the different groups. Until now, my phylogenies were unrooted, which limits the interpretation of the data in terms of temporal and directional evolution. The objective now is to root these trees using robust methods to ensure the accuracy and reliability of the results.

We used two methods: the AU-test and rootstrap. The AU-test (*Approximately Unbiased test*) is a statistical method used to assess the confidence of rooting hypotheses by comparing different hypothetical rooted trees. This approach uses simulations to estimate the distribution of possible trees and calculates the probabilities of each rooting hypothesis. It allows for the selection of the most probable tree among several candidates, thus reinforcing the validity of the proposed rooting. Rootstrap is a complementary method that combines traditional bootstrap with specific rooting techniques to estimate the stability of the rooting through resampling of the data. This technique tests the robustness of the rooting by generating multiple bootstrap datasets, then observing the frequency with which a particular rooting is obtained. A high frequency of rooting in the bootstrap datasets indicates strong confidence in the position of the root. We applied these methods to two distinct datasets:

- A dataset composed solely of archaeal sequences. This allows for an initial estimation of the root without external influence;
- A dataset including eukaryotes to evaluate the impact of these sequences on the rooting results. Comparing the rooted trees obtained from both datasets will help determine if the inclusion of eukaryotes introduces artifacts or significantly alters the results.

Due to time constraints, we chose to work exclusively on the replica 2 supermatrix (33,959 positions), which, as mentioned earlier, seems to be the most credible tree topology. The use of the AU-test without eukaryotes, with the highest non-rejecting p-value (0.94), roots the tree within the Heimdallarchaeota (**Figure 70**). In this scenario, Asgards are paraphyletic. The evolution of Archaea would then proceed through secondary simplification. This scenario could be in agreement with a three-domain model. Archaea and eukaryotes would share a common ancestor that was already complex, and the evolution of Archaea would have proceeded through secondary simplification. This would explain the close phylogenetic proximity between the Asgard group and eukaryotes. This scenario could also be fully compatible with scenario 1D (D. P. Devos, 2021) which also proposes a common ancestor to Archaea and eukaryotes that was already complex. We then find two subgroups corresponding to the Euryarchaeota and TACK. The DPANN are sister group to the Altiarchaeota and thus insert within the Euryarchaeota.

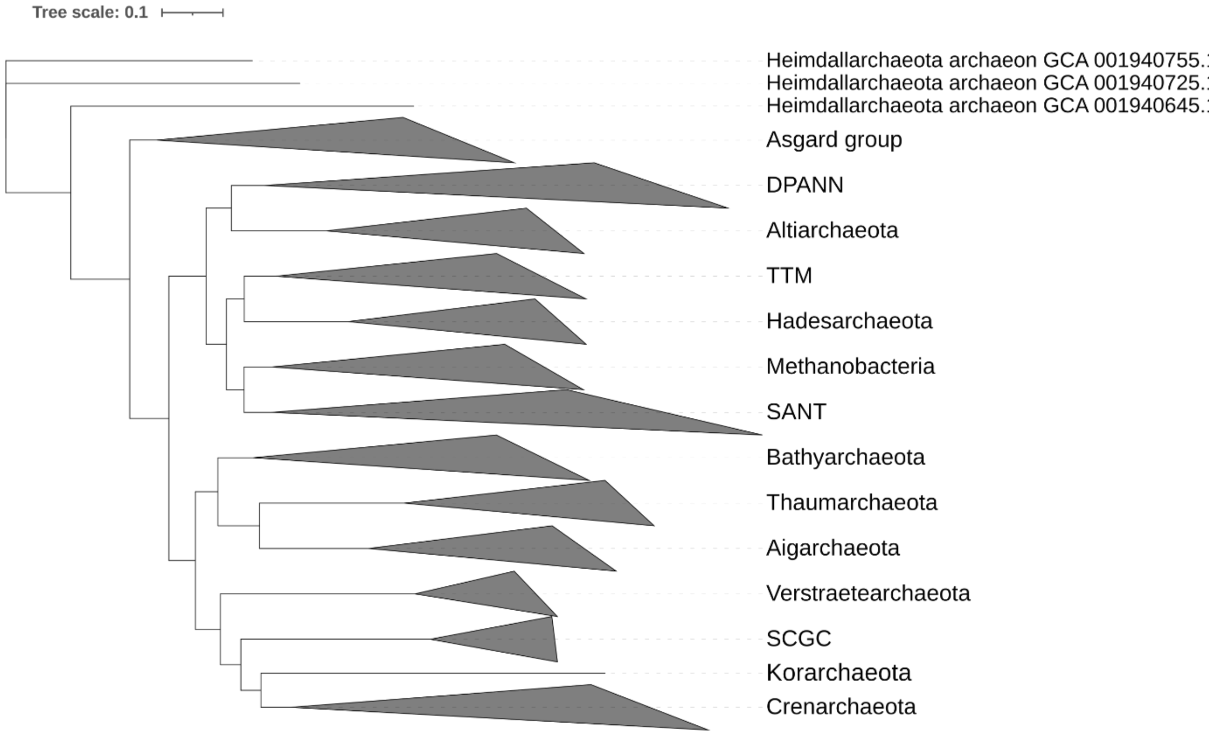


Figure 70. Tree obtained with the highest p-value (0.94) after an AU-test in the absence of eukaryotes on the super-matrix of replica 2 (33,959 positions, 121 species).
Data are provided in **Supp.Mat enracinement**. Asgards are paraphyletic and at the base of other archaea. In such a scenario, archaea would have evolved through secondary simplification, the ancestor of archaea being an already complex being.

Among the other possible roots of our confidence set, the AU-test without eukaryotes gives us 26 possible trees with 6 possible roots:

- Heimdallarchaeota (2 cases, with a p-value of 0.94)
- Altiarchaeota/DPANN (p-value = 0.47)
- SANT, which could be paraphyletic or include Methanobacteria as a sister group (p-values of 0.14, 0.70, 0.69, 0.09, 0.19)
- DPANN, which could or may not be paraphyletic (p-values of 0.22, 0.22, 0.17, 0.05, 0.06)
- TACK (3 cases, with p-values of 0.21, 0.17, 0.13, 0.12)
- Crenarchaeota + Korarchaeota, with TACK becoming paraphyletic (4 cases, with p-values of 0.16, 0.13, 0.08, 0.07)
- Bathyarchaeota (3 cases, with p-values of 0.06, 0.16, and 0.16)

- Finally, Archaea could be divided into two sister groups corresponding to the Euryarchaeota and the other Archaea (2 cases, with p-values of 0.77 and 0.49).

On the other hand, the inclusion of eukaryotes in the AU-test drastically alters the root position, resulting in two groups: the Euryarchaeota and a large group comprising DPANN + Asgard + Eukaryotes + TACK (p-value = 0.76) (**Figure 71**). The DPANN are no longer members of the Euryarchaeota but are placed at the base of the Asgard + Eukaryotes + TACK group. We also observe a paraphyly of the Heimdallarchaeota, with eukaryotes inserted into it. However, it is worth noting the length of the branch at the base of the eukaryotes, which could also suggest an artifact of long branch attraction (branch length: 0.38).

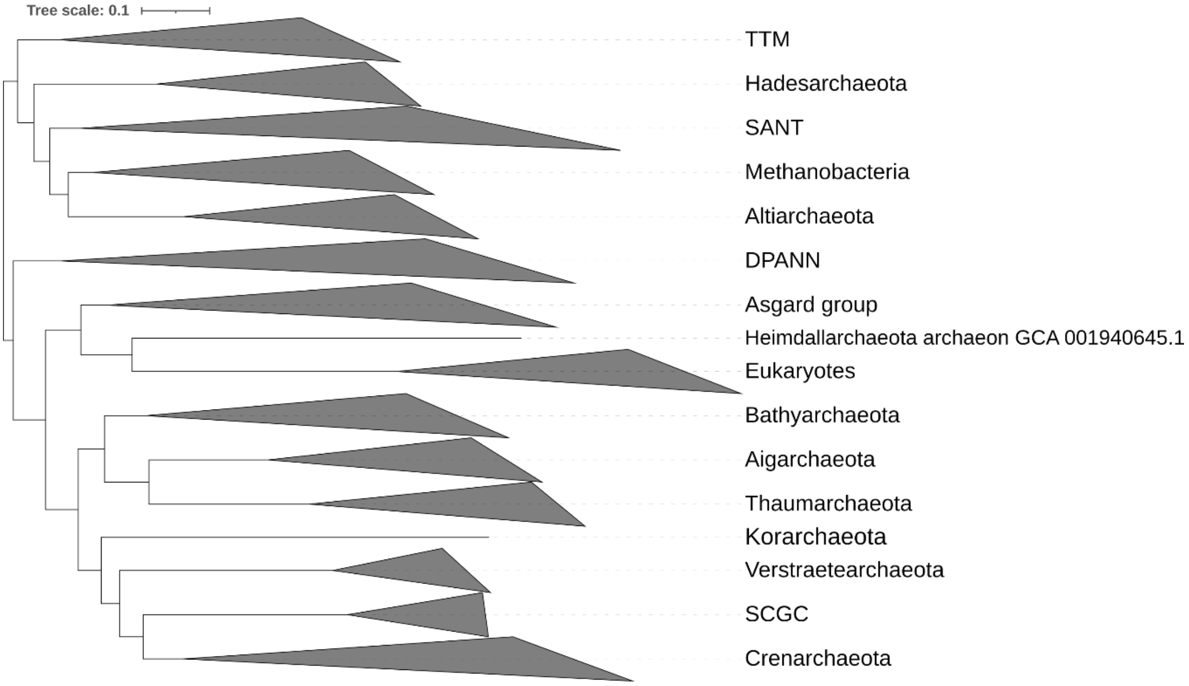


Figure 71. Tree obtained with the highest p-value (0.76) after an AU-test in the presence of eukaryotes on the super-matrix of replica 2 (33,959 positions, 132 species).
Data are provided in **Supp.Mat enracinement**. The addition of eukaryotes in this case favors two distinct groups of archaea: on the one hand, the Euryarchaeota, which are therefore monophyletic, and on the other, all other archaea, including eukaryotes.

Among the other possible roots of our confidence set, the AU-test with eukaryotes gives us 12 possible trees with 6 possible roots:

- the TTM (p-values = 0.05)
- the Thermoplasmatales (2 cases, with p-values of 0.37 and 0.19)
- the DPANN (p-values = 0.37)
- the Altiarchaeota (3 cases, with p-values of 0.41, 0.12, 0.08)
- Stenosarchaea + Archaeoglobi, making our SANT group paraphyletic (p-value = 0.14).
- Finally, Archaea could be divided into two sister groups corresponding to the Euryarchaeota and the other Archaea (4 cases, with p-values of 0.76, 0.68, 0.41, 0.19), with the DPANN either within the Euryarchaeota or at the base of all other Archaea. Additionally, in one case, the Hadesarchaeota and the TTM group are removed from the Euryarchaeota, making them paraphyletic, and are placed at the base of the DPANN and other Archaea.

When using the rootstrap method, whether with or without eukaryotes, the root is located within the SANT group (**Figure 72** & **Figure 73**). The Euryarchaeota, becoming paraphyletic, no longer exist as a valid group. The DPANN, as with the AU-test, are placed at the base of the (Asgard + Eukaryotes + TACK) group. The Verstraetearchaeota become the sister group of the SCGC. However, the position of the Altiarchaeota changes with the addition of eukaryotes, moving closer to part of the TTM group, rendering the latter polyphyletic. Similarly, the Asgards become paraphyletic, with the Heimdallarchaeota as the sister group of the eukaryotes. As with the AU-test, the long branch of the eukaryotes (branch length: 0.39) could result in a position influenced by long branch attraction.

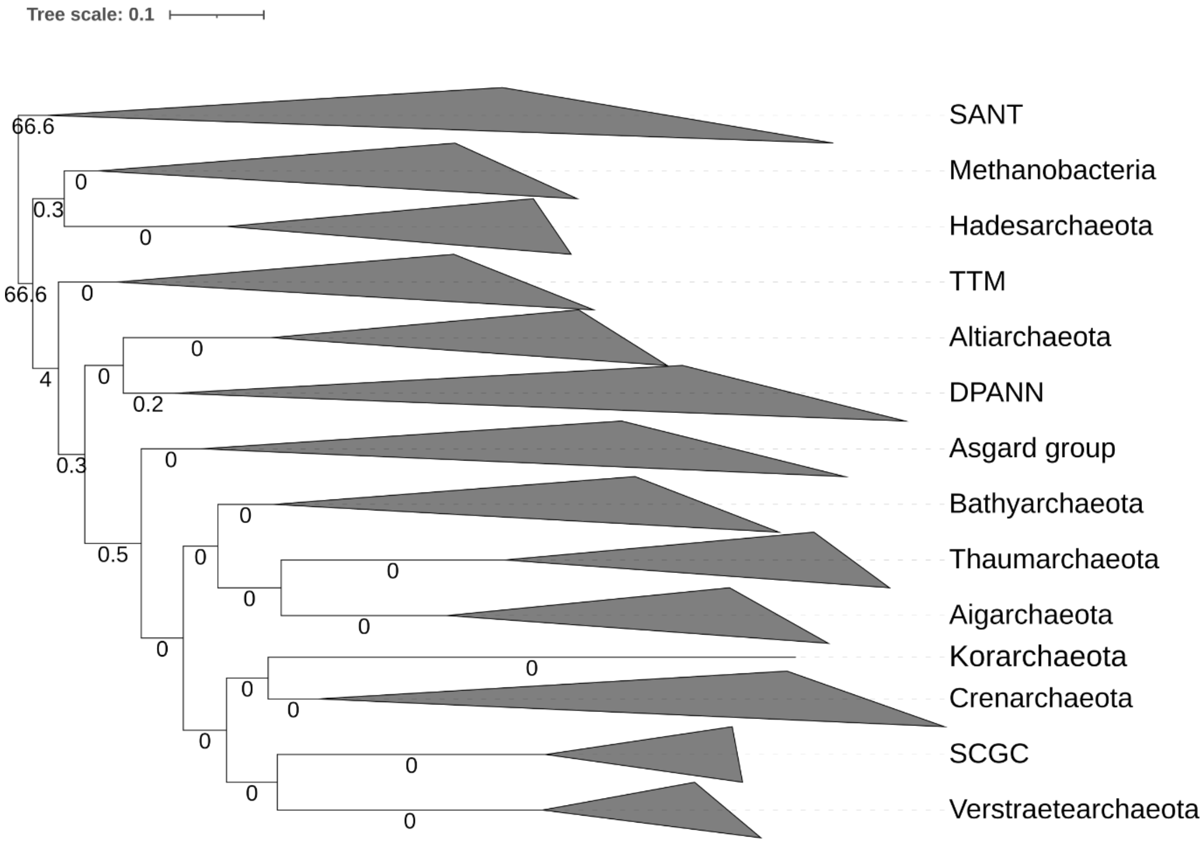


Figure 72. Tree obtained after rootstrap in the absence of eukaryotes on the super-matrix of replica 2 (33,959 positions, 121 species).

Data are provided in **Supp.Mat enracinement**. This tree roots archaea within the SANT group, making Euryarchaeota paraphyletic.

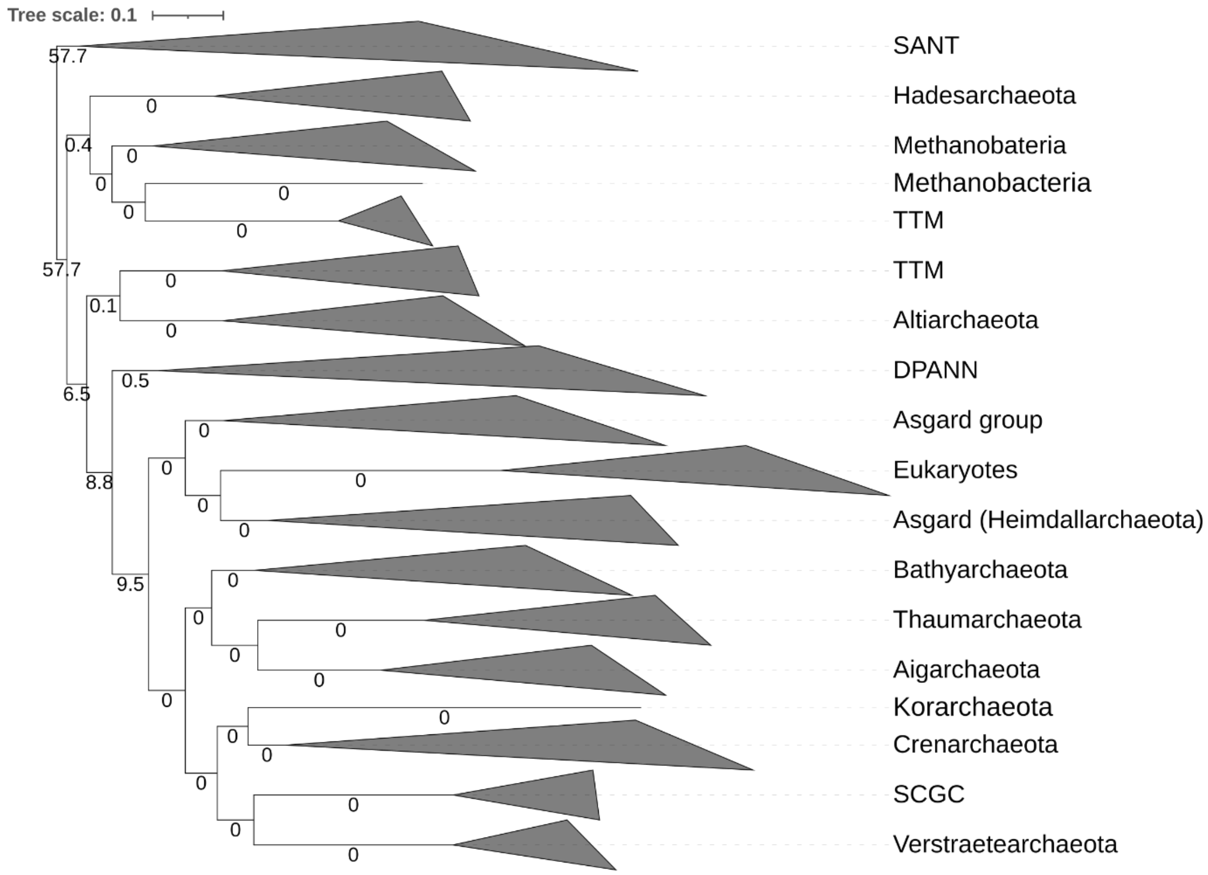


Figure 73. Tree obtained after rootstrap in the presence of eukaryotes on the super-matrix of replica 2 (33,959 positions, 132 species).

Data are provided in **Supp.Mat enracinement**. This tree roots the archaea within the SANT group, making the Euryarchaeota paraphyletic. The Altiarchaeota are close to certain members of the TTM group, making the latter polyphyletic.

8 PHYLOGENETIC INFÉRENCE

After performing phylogenetic analyses using IQ-TREE and the PMSF method on a dataset containing only Archaea and another dataset including Archaea and Eukaryotes, we chose to complement these analyses using the CAT model to perform Bayesian inferences with PhyloBayes. Indeed, the CAT model (Categories) is particularly well suited to capture the heterogeneity of equilibrium amino acid frequencies across the sites of a sequence. By redoing the analyses with the CAT model and PhyloBayes, we can compare the results with those previously obtained through IQ-TREE and the PMSF method. This comparison will help verify the consistency of the results and identify any potential artifacts introduced by the previous methods or models. These Bayesian inference analyses were conducted using PhyloBayes MPI 1.9a (Lartillot et al., 2013) according to the CAT-G and CAT-GTR-G models.

In accordance with our rooting results, we chose to root our trees within the SANT group. With both of our models (CAT-G and CAT-GTR-G), we obtain the same group of Ouranosarchaea (TACK + Asgard + Eukaryotes). Their phylogeny is the same (**Figure 74** & **Figure 75**), and we also obtain the same result with our PMSF analysis, namely the Eukaryotes as the sister group to the Asgard group. In both models, two out of four chains show a paraphyly of the Asgards, with either all Heimdallarchaeota or Heimdallarchaeota GCA001940645.1 becoming the sister group to the Eukaryotes.

Regarding the position of the Korarchaeota, which we discussed with the PMSF model, we predominantly find a position of the Korarchaeota as the sister group to the Crenarchaeota. However, with the CAT-G model, we find one chain (out of four) that places the Korarchaeota at the base of the TACK group. With the CAT-GTR-G model, out of four chains, one places the Korarchaeota at the base of TACK and another places them within TACK, at the base of the SCGC, Crenarchaeota, and Verstraetearchaeota.

Regarding the Hadesarchaeota, they are consistently at the base of the Ouranosarchaea with the CAT-G model and three out of four chains of the CAT-GTR-G model. This result differs from the analyses performed with IQ-TREE, which favor the DPANN at this position. Up to now, we had the Hadesarchaeota as an internal group within the Euryarchaeota. However, our rooting results do not seem to validate the Euryarchaeota as a valid monophyletic group. This result was only found during the AU-test with the eukaryotes. Additionally, in our PMSF analyses on our species replicas, only replica 4 places the Hadesarchaeota at the base of the Ouranosarchaea. The alternative hypothesis places the Hadesarchaeota within the Euryarchaeota. We then find the DPANN, which can have the Altiarchaeota as their sister group. If we root our trees within the SANT group, the Methanobacteria are consistently the group that comes just after in the tree, followed by the TTM group. The uncertainties remain regarding the positions of the DPANN, Altiarchaeota, and Hadesarchaeota groups, which insert between the Ouranosarchaea and the Methanobacteria.

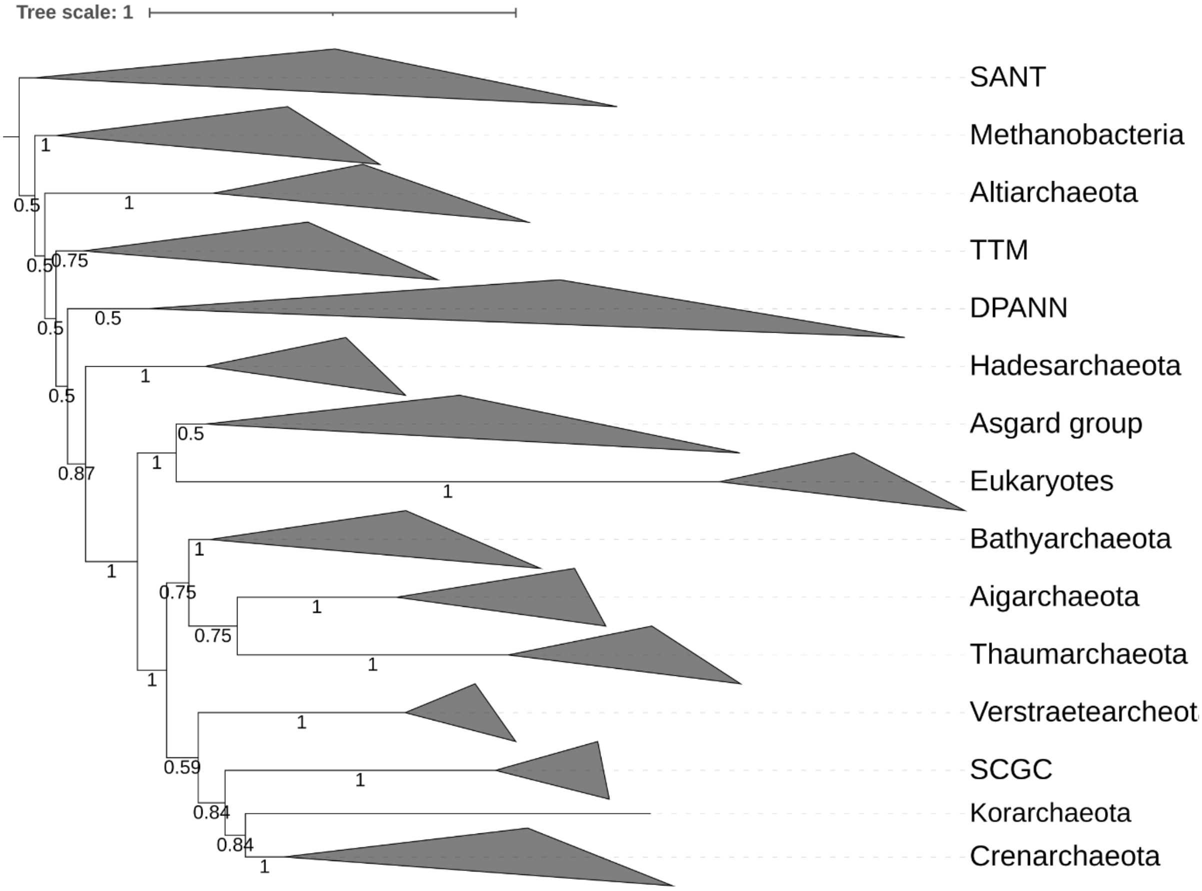


Figure 74. Consensus tree calculated with PhyloBayes according to the CAT-G model on replica 2 (132 species).

Data are provided in **Supp.Mat phyloBayes**. Eukaryotes are sister group to Asgard. Hadesarchaeota are systematically at the base of Ouranosarchaea, contrary to analyses carried out in PMSF with IQ-TREE, which favored DPANN in this position.

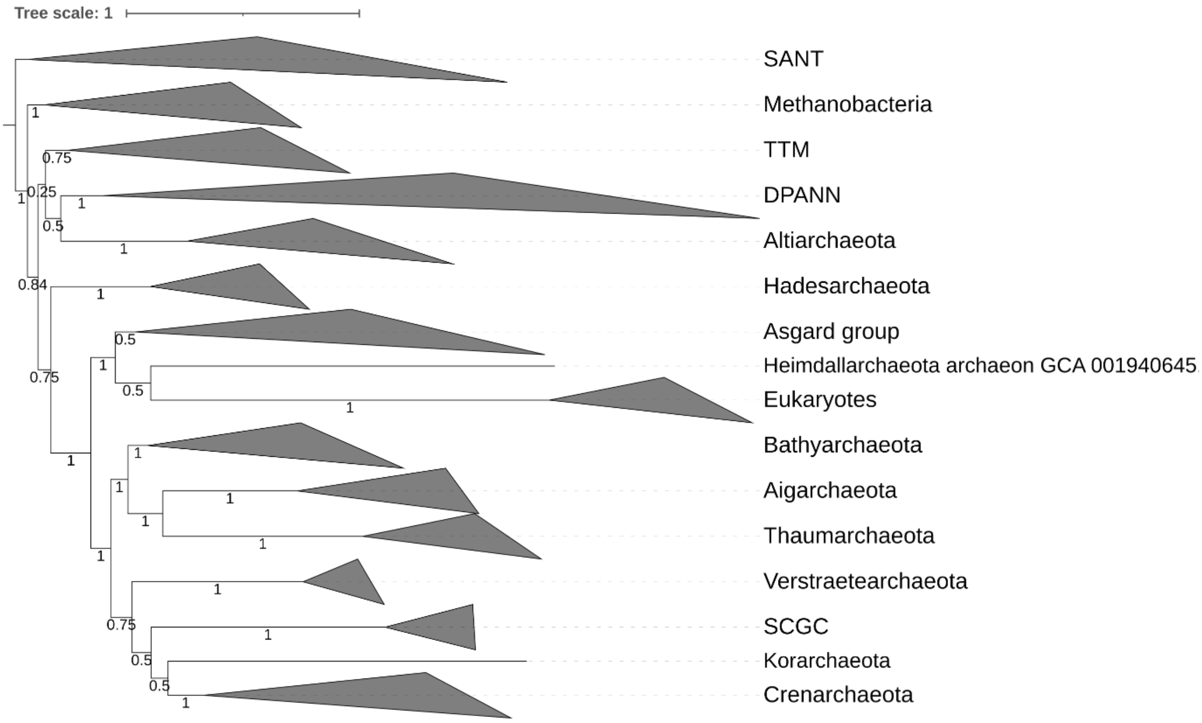


Figure 75. Consensus tree calculated with PhyloBayes according to the CAT-GTR-G model on replica 2 (132 species).

Data are provided in **Supp.Mat phyloBayes**. Eukaryotes are sister group to Asgard. Hadesarchaeota are at the base of Ouranosarchaea for 3 of our 4 chains, contrary to analyses performed in PMSF with IQ-TREE which favored DPANNs at this position.

Several factors, including variations in models (C60 vs CAT), software (PhyloBayes vs IQ-TREE), and implementation details (for example, the number of categories used to account for heterogeneity in rates (= gamma categories)) could explain the new variation in results observed here among the heterogeneous models at the site level. The number of categories inferred by CAT in PhyloBayes can be very high. This requires a very large number of additional parameters to be estimated. Finally, it is worth noting the significant branch length of the Eukaryotes (CAT-G: 1.48; CAT-GTR-G: 1.71), whose position could thus be subject to a long-branch attraction artifact.

9 DISCUSSION

The eukaryotic cell exhibits many seemingly unique characteristics, and there is a considerable gap between prokaryotic and eukaryotic cells in terms of complexity and development. Despite this apparent gap, few characteristics are truly eukaryotic-only, suggesting a mixed contribution from the other two domains. Indeed, the eukaryotic cell is an evolutionary mosaic composed of bacterial and archaeal traits, as well as eukaryotic innovations. The discovery of Asgard archaea (Eme et al., 2017; Spang et al., 2015, 2017; Zaremba-Niedzwiedzka et al., 2017) reignited an intense debate over the topology of the universal tree of life. It has been suggested that the branching of eukaryotes within the archaea in certain phylogenetic reconstructions is the result of phylogenetic reconstruction artifacts (notably long-branch attraction) associated with deep ancient divergence (D. P. Devos, 2021). Proponents of a two-domain tree (2D, Eocyte hypothesis) propose a scenario in which eukaryotes emerged within the archaea, more specifically as a subgroup of the Asgard super-phylum, while others support a tree where the three domains (3D) are monophyletic. In our study, we adopted a 2D model to test the supposed relationship between Asgards and eukaryotes. Nevertheless, a three-domain scenario could still be possible (**Figure 6**).

Furthermore, it is possible that horizontal gene transfers between ancestral Asgards and proto-eukaryotes could explain the very sparse distribution of certain Asgard ESPs and universal marker proteins (Da Cunha, Gaïa, et al., 2022). Indeed, it is possible that the variable position of Asgards in single-gene trees is due to a rapid evolutionary rate. Metabolic reconstructions have suggested that Asgards rely on symbiotic interactions for both anabolism and catabolism, which could explain why they are so difficult to cultivate (currently, 2 species are cultivated: *Prometheoarchaeum syntrophicum* (Imachi et al., 2020) and *Lokiarchaeum ossiferum* (Rodrigues-Oliveira et al., 2023). It is therefore possible that Asgards' adaptation to their partners increased the evolutionary rate of some of their proteins, just as with the formerly supposed Archezoa group. Additionally, these high rates of horizontal gene transfer could explain the disparate appearance of proteins previously thought to be ESPs within the archaea (Da Cunha, Gaïa, et al., 2022), something not mentioned in recent studies supporting an origin of eukaryotes within the Asgard group (Eme et al., 2023).

The abnormal behavior of Asgard proteins

Furthermore, in trees based on a single protein, the position of eukaryotes is variable, either nested within the Asgards, as a sister group to all Asgards, or as a sister group to other archaea. This phenomenon had already been observed with the first three published Asgards (Loki 1, 2, and 3) in 36 single-protein trees (Da Cunha et al., 2017; Da Cunha, Gaïa, et al., 2022). The Asgards were also often paraphyletic in these trees (with Heimdallarchaeota as the sister group to eukaryotes, while other Asgard groups formed three distinct, unrelated clades), whereas the monophyly of other major archaeal clades was generally recovered. This suggests that the scattered positions of Asgards were likely not due to a lack of resolution. Abnormal behavior of Asgard proteins has also been reported for ribosomal proteins (Garg et al., 2021). Indeed, the dispersion of Asgards in universal ribosomal protein trees could reflect errors in Metagenome-Assembled Genome (MAG) reconstruction, although in our case, the use of a cultured Asgard does not seem prone to this type of error and may therefore be quite reliable. However, this likely cannot explain all situations. Indeed, the same type of abnormal behavior has been observed in single-gene trees based on universal proteins encoded by rapidly evolving DPANN archaea. This could suggest that the variable position of Asgards in these archaeal trees results from a phenotype evolving rapidly (Da Cunha, Gaïa, et al., 2022).

The importance of protein selection

Another issue that may distort the topology of the tree of life is the set of marker data used. We tried to maximize the number of orthologous genes used, not limiting ourselves to ribosomal genes. Indeed, it has been shown that the transition from a 2D to a 3D tree may depend on just one protein. For example, a study based on an initial dataset of 30 proteins shows that the removal of just one protein, the bacterial ATPase YchF, turned a 3D tree into a 2D tree (Liu et al., 2021b). Similarly, the removal of the elongation factor EF2 from a dataset of 36 proteins turned a 2D tree into a 3D tree by breaking the monophyly of Eukaryotes-Lokiarchaeota (Da Cunha et al., 2017). The fact that a single protein can determine the topology of the universal tree of life from several dozen markers highlights the importance of carefully analyzing single-gene trees before performing data concatenation.

In marker selection, gene length is also crucial. Indeed, large proteins tend to support 3D trees, while short proteins tend to support 2D trees (Da Cunha et al., 2017). It is possible that short proteins do not harbor enough informative positions to detect the signal corresponding to the monophyly of the Archaea. Thus, the datasets used by Liu (Liu et al., 2021b), Xie (Xie et al., 2022) and their colleagues are both enriched in short proteins (80% and 40%, respectively), which potentially favors the 2D topology, and omit several large proteins that have individually given 3D trees in the past (Da Cunha et al., 2017) In particular, they are both lacking the large B subunit of RNA polymerase. However, large subunits of RNA polymerase have very high phylogenetic signal (i.e., they more accurately predict a true vertical lineage) (Martinez-Gutierrez & Aylward, 2021). Indeed, in a dataset of 41 conserved archaeal and bacterial markers, large RNA polymerase subunits performed better despite a shorter overall alignment length. In contrast, ribosomal proteins tend to individually exhibit low phylogenetic signal. The phylogenetic signal obtained with the concatenation of ribosomal proteins was much higher than that obtained with individual ribosomal proteins, but still lower than the signal obtained with the concatenation of both large subunits of RNA polymerase (Martinez-Gutierrez & Aylward, 2021). Similar results leading to 3D trees have also been obtained (Da Cunha et al., 2017; T. A. Williams et al., 2020). 2D trees were only obtained after recoding the amino acids from multiple sequence alignments. Given that amino acid recoding significantly weakens the signal in phylogenetic analysis, it is possible that the shift from a 3D tree to a 2D tree after recoding is due to a weaker signal supporting the monophyly of archaea (Da Cunha et al., 2017). For the purpose of obtaining maximum phylogenetic signal, we have attempted throughout our study to recover as many genes as possible in addition to ribosomal proteins. In our case, gene jackknifing performed on archaea and Robinson-Foulds distances allowed us to assess the impact of gene selection, which seemed low compared to species selection. Here, we assessed the impact of our gene selection by comparing our results with those obtained from our archaeal phylogenies. Our under-sampling of genes present in eukaryotes hardly affects the bipartitions found, neither helping us to decide between our archaeal topologies from the previous chapter nor adding new insights.

Horizontal Gene Transfer and ESPs and ESPs

The characterization of new Asgard lineages led to the identification of several new ESPs within them. However, the distribution of ESPs is highly uneven within the Asgards. Many ESPs are specific to a particular Asgard lineage, while others are absent from certain lineages (Liu et al., 2021b; Xie et al., 2022). For example, this is the case with tubulin, which is only present in the Odin lineage (Xie et al., 2022; Zaremba-Niedzwiedzka et al., 2017) In both 2D and 3D models, the extremely disparate phylogenetic distribution of these ESPs requires many losses and/or transfers between archaeal lineages. In the 2D scenarios, it is also assumed that eukaryotes emerged from an extinct Asgard lineage that contained all the ESPs currently distributed among the different existing Asgard lineages (Eme et al., 2017). In both models, it is particularly difficult to explain the existence of ESPs that are currently restricted to only one or a few Asgard lineages. However, if we consider that Eukaryotes are the base of Asgards + TACK, as suggested by our heterotachy work, then a complex ancestor to this group is conceivable. This ancestor would be from an extinct lineage that already possessed some of the genes and characteristics shared by Eukaryotes, Asgard, and TACK. Thus, up until this ancestor, evolution would have proceeded by increasing complexity. Then, this complexity would have continued in the Eukaryotes, while the Asgard + TACK group would have evolved by secondary simplification. The shared characteristics could very well be explained by vertical inheritance from this common ancestor, while other characteristics would have been lost or modified, explaining the disparate nature of certain features. Moreover, this situation is reminiscent of the pitfall of the Archézoa hypothesis, where it was once thought that these supposedly simple eukaryotes were at the base of Eukaryotes. The fact that some Asgards or TACK live in symbiosis or interact with other organisms could very well explain these phenomena of secondary simplification, as was the case for the Archezoa.

Another hypothesis that could more easily explain the uneven distribution of Asgard ESPs is that some ESPs were recruited by the Asgards from proto-eukaryotes (i.e., members of the eukaryotic lineages that preceded the last common ancestor of eukaryotes). Such horizontal gene transfers may have occurred either early in the evolution of the Asgards, with the corresponding ESPs being present in most or all Asgards, or later during the diversification of Asgard lineages, thus explaining the restricted distribution of the corresponding ESPs (Da Cunha et al., 2017). This could also explain why one of our three Heimdallarchaeota is almost systematically associated with the eukaryotes rather than the other Heimdallarchaeota. Recently, a study rooted eukaryotes within a new group called Hodarchaeales (Eme et al., 2023). We sought to determine what this Heimdallarchaeota corresponds to in their dataset. The analysis of our Heimdallarchaeota GCA_001940645.1 corresponds to LC-3 sp001940645, which is considered a Hodarchaeales, whereas it was previously considered a Heimdallarchaeota. Our thesis would therefore also highlight this new group.

It is important to note that this proto-eukaryote HGT hypothesis would explain why some Asgard ESPs are much more similar to their eukaryotic counterparts than to those found in archaea. This is the case with Odin tubulin, which is much closer to eukaryotic tubulins than to the tubulin found in Thaumarchaeota (Zaremba-Niedzwiedzka et al., 2017). Similarly, most Asgard actins are much more similar to eukaryotic actins and actin-related proteins (ARPs) than to archaeal crenactins (Da Cunha, Gaia, et al., 2022; Stairs & Ettema, 2020). In the case of actin, the proto-eukaryote HGT hypothesis is strongly supported by phylogenetic analyses showing that the different forms of actins found in Asgards branch between different clades of eukaryotic actin paralogs (cytoplasmic actin and ARP), which were already present in LECA (Da Cunha, Gaia, et al., 2022; Stairs & Ettema, 2020). Furthermore, corroborating this hypothesis, comparative genomic analyses have shown that the ESPs are of relatively recent origin and could correspond to late HGT events between archaea and eukaryotes (Nasir et al., 2021). Indeed, ESPs may be more widespread than previously thought, as suggested by the discovery of actin encoded by viruses (viractin) in several viral genomes or the presence of actin-like proteins in Bathyarchaeota (Dombrowski et al., 2018; Zhou et al., 2018). These discoveries suggest that the presence of ESPs in other archaeal lineages or viruses is likely underestimated (Nasir et al., 2021).

It is also important to keep in mind that phylogenetic trees are the final result of successive analytical steps. Among these, multiple sequence alignments are critical, and despite undeniable improvements in their methods, the risk of incorrect alignments increases with the size of the dataset. Protein alignment defects due to inversion/substitution of domains or mismatches have been detected in 42% of the universal markers used (36 arCOGs + ribosomal proteins) in the first Asgard paper, which were later used for subsequent studies (Nasir et al., 2021). It would therefore be important to verify new alignments to detect potential errors. Notably, the risk of incorrect alignments is increased when taxon oversampling occurs and when species with rapid evolution are included, both factors commonly observed in studies concerning the universal tree of life (Da Cunha, Gaïa, et al., 2022). Furthermore, our results support a monophyly of Asgards with the TACK group rather than with eukaryotes, particularly when keeping slow-evolving sites. Asgards could be highly sensitive to phenomena of heterotachy and heteropecillly. In fact, one of our Heimdallarchaeota is consistently grouped within eukaryotes instead of clustering with other Heimdallarchaeota; this could be explained if we hypothesize a strong heterotachy phenomenon that disrupts Heimdallarchaeota. Indeed, when considering the possibility of artifacts due to horizontal gene transfers and the use of fast-evolving sites (linked to their symbiotic or parasitic lifestyle with eukaryotes), it is entirely legitimate to question the close relationship between eukaryotes and Asgards.

Finally, one possibility is that these ESPs are in fact remnants of what was present in their common ancestor LAECA. This would suggest that this ancestor was already complex, and that evolution towards archaea occurred through simplification. This hypothesis is supported by the fact that the evolution of archaea probably involved massive losses from a complex ancestor (Koonin, 2015). In this scenario, archaea lost most of the eukaryotic features possessed by a complex LAECA. Asgards, having diverged less, then retain a greater number of ESPs as well as a closer phylogenetic relationship with eukaryotes. Therefore, 1D scenarios do not contradict the proximity of Asgards to eukaryotes, but rather interpret it differently (D. P. Devos, 2021). The scattered presence of certain proteins within archaea can then be interpreted as the result of secondary losses and simplifications, particularly in the TACK group, which Asgards cluster with in our heterotachy and heteropecilly analyses. The supposed large number of horizontal gene transfers from bacteria to archaea (Nelson-Sathi et al., 2015) may in fact mainly represent remnants of a lineage that diverged from a bacterial ancestor and was lost in some archaea while remaining divergent in others (D. P. Devos, 2021).

Moreover, some elements, although still limited, could support this hypothesis, notably the placement of PVC bacteria at the root of the lineage leading to LAECA. A phylogenetic signal has been detected in ribosomal proteins, supporting a 1D scenario with PVC bacteria at the base of the eukaryote-archaea domain (Cavalier-Smith & Chao, 2020). The signal is weak, but this is to be expected given that it pertains to one of the most ancient relationships in the tree of life. In fact, the signal between PVC bacteria and eukaryotes is expected to be much less clear than that between archaea and eukaryotes, as the former diverged much earlier than the latter (D. P. Devos, 2021).

Rooting of Archaea, Paraphyly of Euryarchaeota, Position of Altiarchaeota, and Relationship with the DPANN Group

Since the discovery of the first representative of DPANN (Rinke et al., 2013), the placement of this group within Archaea has been uncertain (T. A. Williams et al., 2015) due to their extremely reduced genome (490,885 base pairs for approximately 1,000 genes, the smallest among all archaea (Waters et al., 2003)) and their long branches (i.e., their rapid substitution rate) (Dombrowski et al., 2019). DPANN could instead belong to Euryarchaeota or be polyphyletic, occupying various positions within Euryarchaeota.

The basal placement of DPANN should be approached with caution. Several studies using bacteria as an outgroup and conserved protein markers have placed the root of Archaea between DPANN and all other archaea (Martinez-Gutierrez & Aylward, 2021; Petitjean et al., 2014; T. A. Williams et al., 2017). Advanced Bayesian models have also positioned DPANN at the root of Archaea (T. A. Williams et al., 2017). These models were based on genome evolution models that accounted for gene duplications, losses, and horizontal transfers. However, the dataset used in this study was relatively limited, comprising only 16S and 23S rRNA sequences, which are known to be compositionally biased (Galtier & Lobry, 1997). A multigene supertree was also rooted between DPANN and all other archaea using a recently developed genome evolution model (T. A. Williams et al., 2017). This study's metabolic reconstructions on the rooted tree suggest that the earliest archaea were anaerobes capable of reducing CO2 to acetate via the Wood-Ljungdahl pathway. Contrary to propositions that genome reduction was the dominant mode of archaeal evolution (Csűrös & Miklós, 2009), this study suggests a relatively small genome for the archaeal ancestor, which later increased in complexity through gene duplications and horizontal transfers. Genetic evidence for the ancient carbon fixation system (Wood-Ljungdahl pathway), methanogenic traits, and the ability to anaerobically oxidize methane and other short hydrocarbons has been found in diverse archaeal lineages inhabiting anaerobic environments (Spang et al., 2017). These findings support the hypothesis that all current archaea evolved from an autotrophic anaerobic ancestor utilizing the Wood-Ljungdahl pathway and potentially deriving energy from methanogenesis. Despite these hypotheses, the basal position and monophyly of DPANN remain highly uncertain, with low taxonomic sampling contributing to doubts about these results. Previous studies have shown that the phylogenetic position of DPANN is sensitive to the taxa included (Dombrowski et al., 2019; T. A. Williams et al., 2017). Our own analyses reached similar conclusions. The inclusion of eukaryotes tends to move DPANN away from Altiarchaeota, positioning them at the base of the Ouranosarchaea (TACK + Asgard (+ eukaryotes?)). This may also represent an artifact caused by long branch attraction (LBA) due to their high substitution rate and very small genomes (Aouad et al., 2018, 2022; Petitjean et al., 2014; Raymann et al., 2015; Roure & Philippe, 2011). Notably, DPANN are parasitic archaea—for example, *Nanoarchaeum equitans* is a symbiont of *Ignicoccus hospitalis*, and *Micrarchaeota spp.* are parasites of Sulfolobales. If DPANN represents the ancestral group of archaea, this would imply that the common ancestor of modern archaea was a symbiotic form that later gained complexity. This conclusion challenges the notion of an ancient origin for archaea. However, this reasoning may be circular, as a symbiotic ancestral lifestyle would imply a host of a different origin. In any case, our analyses never recover DPANN as basal to Archaea. We therefore have grounds to doubt the validity of this result. Additional sequencing of archaea, particularly of DPANN themselves, is likely to clarify their placement.

The Euryarchaeota have long been considered a monophyletic group (Petitjean et al., 2014; T. A. Williams et al., 2017). However, our PMSF analyses reveal that this result is unstable depending on taxonomic subsampling (species jackknife tests). Once again, the monophyly of Euryarchaeota is challenged when using the rootstrap approach, as they appear paraphyletic in our analyses. The Ouranosarchaea group (TACK + Asgard (+ eukaryotes?)) is consistently recovered, except in AU tests excluding eukaryotes, where Asgard archaea appear at the root. However, the monophyly of Gaiarchaea (Euryarchaeota) remains elusive. Based on our results, it is not possible to definitively support one hypothesis over another.

Questions surrounding the monophyly of Euryarchaeota, the placement of DPANN and its relationships to Altiarchaeales, and the basal position of DPANN within Archaea remain among the most significant unresolved issues in archaeal phylogeny. Addressing these questions will require assembling targeted datasets and implementing protocols specifically designed to investigate these topics (Aouad et al., 2022) as well as the discovery of new members of these groups. DPANN as the basal archaeal group may be analogous to the Archezoa in eukaryotes—symbiotic species with highly reduced genomes subject to long-branch attraction artifacts.

Our study tends to support the paraphyly of Euryarchaeota, a close relationship between DPANN and Altiarchaeota at the base of Ouranosarchaea + eukaryotes, and the rooting of Archaea within the SANT group. Considering our heterotachy analyses, eukaryotes could potentially be placed at the base of Ouranosarchaea. Ultimately, the cladogram we favor in the context of this thesis is shown in **Figure 76**.

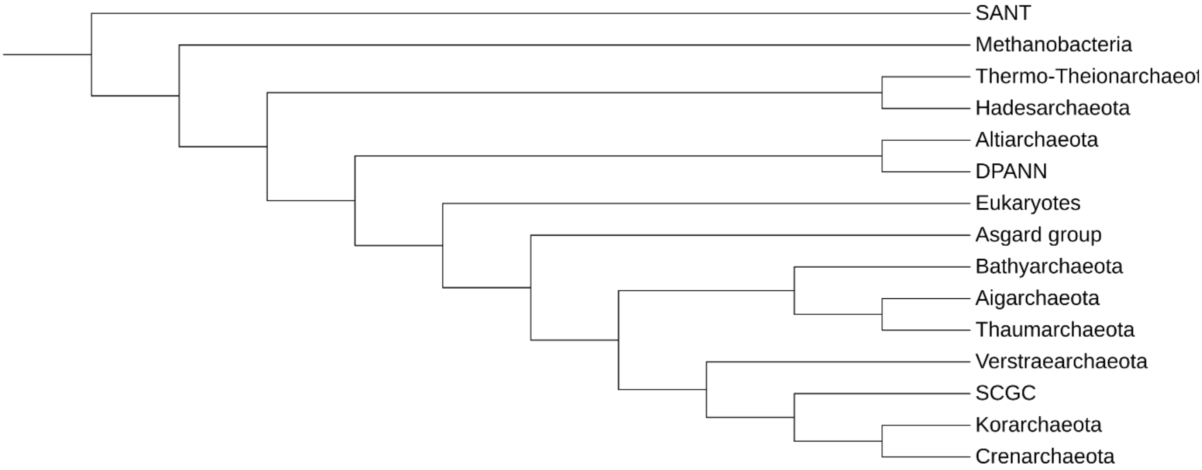


Figure 76. Root cladogram of archaea and eukaryotes based on all our results.
The root is located in the SANT group. In view of our heterotachy results, it would seem to us that Eukaryotes are at the base of Ouranosarchaea and that Euryarchaeota are polyphyletic.

GENERAL DISCUSSION & CONCLUSION

This doctoral research focused on archaea and their relationships with eukaryotes. Its objective was to evaluate the validity of the current trend to include eukaryotes within the domain of archaea. We leveraged the ever-growing availability of sequencing data to enhance the resolution of the archaeal evolutionary tree, subsequently incorporating eukaryotes. This massive influx of new sequences was utilized to construct the most reliable and robust dataset possible, which was then analyzed using various phylogenetic inference methods to identify potential methodological biases.

A significant portion of the work involved building a dataset that was as clean and comprehensive as possible to minimize any artifacts that might distort the results. This process included evaluating gene quality (contamination checks, orthology assessments, substitution rate evaluations) and making reasoned choices regarding species selection. Several datasets were created and tested, allowing for comparisons of results using jackknife procedures for both genes and species. The focus then shifted to assessing the quality of phylogenetic inference, encompassing everything from the alignment of orthologous sequences to the inference of species trees.

In general, resolving issues related to deep phylogeny requires analyzing large-scale genomic datasets, as the phylogenetic signal in distantly related sequences is weak. However, precautions must be taken when employing this approach, as a perfectly resolved tree does not necessarily mean that sequence evolution has been accurately reconstructed. In particular, bootstrap values are not useful for detecting systematic errors (false phylogenetic signals caused by signal saturation when substitutions are too abundant) in phylogenetic analyses. A relationship supported by a 100% bootstrap value does not guarantee its correctness.

My thesis confirms that phylogenetic inference artifacts become more significant when using a large number of positions, potentially obscuring the authentic phylogenetic signal (Da Cunha et al., 2017). This was evident in our site removal analyses, both in the Slow-Fast approach for archaea and in the heterotachy and heteropecilly analyses when eukaryotes were included. Bootstrap values were high in all cases, despite contradictory results depending on the alignment size and the choice of conserved sites.

To avoid artifacts, it would ideally be necessary to have evolutionary models capable of accounting for the complexity and heterogeneity of substitution processes. However, this is challenging, as it is simpler to model one process at a time than to simultaneously account for all possible processes. Pending the development of such a perfect model, this issue can be addressed using alternative approaches. To detect potential artifacts and systematic errors, the approach in my thesis involved testing the robustness of results to taxonomic sampling, the removal of fast-evolving sites, and the use of different sequence evolution models.

In the first chapter of my thesis, I reviewed the state of knowledge at the beginning of my research on the problem of rooting the tree of life. This chapter was published in the journal *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*. The article aimed to highlight the methodological challenges in rooting the tree of life and explore the underlying sociological reasons for the limited interest in this fundamental question.

I analyzed 126 articles published after 2000 to examine the difficulties impacting phylogenetic inference and ways to improve modeling of the highly heterogeneous substitution process, both across sites and over time. Surprisingly, less than half of the studies seemed to consider potential artifacts that could affect phylogenies. I also demonstrated that improved taxonomic sampling, better gene selection, and reasoned data removal strategies have led to numerous revisions of the tree of life, almost always shifting simpler organisms higher in the tree—provided that long-branch attraction artifacts were effectively neutralized.

Furthermore, despite the wealth of genomic data available between 2000 and 2015, there was surprisingly little interest in determining the tree of life's root. The few studies addressing this issue almost always relied on methods developed in the 1990s, whose limitations are now well understood. This led me to argue that the hypothesis of a bacterial root for the tree of life could be attributed to Aristotle's *Great Chain of Being* bias, which presupposes that simpler organisms are ancestors of more complex forms.

Finally, using the elongation factor, an anciently duplicated paralogous gene (Baldauf et al., 1996; Gouy et al., 2015), I demonstrated that even the best current models still cannot fully address the complexity of evolutionary processes, especially substitution processes. In our analyses, the two bacterial clades branched in different positions: as a sister group to archaea + eukaryotes for EF-Tu, and as a sister group to eukaryotes for EF-G. In both subtrees, archaea were paraphyletic, with Crenarchaeota closer to eukaryotes, but without statistical support. Clearly, stochastic and systematic errors deeply affect this phylogeny based on duplicated elongation factors (Gouy et al., 2015). I concluded that the bacterial root, commonly accepted at the start of my thesis in 2015, remains unproven and that the root of the tree of life should be revisited using phylogenomic supermatrices to correctly interpret the evolutionary processes that led to the emergence of the eukaryotic cell. While evaluating homologous characters (particularly through the removal of ambiguously aligned regions) and orthologous genes is relatively precise and not the most significant issue in deep phylogenetics, inference based on supermatrices appears robust to the inclusion of paralogous and xenologous (horizontally transferred) sequences. The main challenge lies in accurately modeling substitution processes, to which supermatrices are highly sensitive (Philippe et al., 2005, 2017; Young & Gillung, 2020).

Given that stochastic and systematic errors have a greater impact on the rooting of the tree of life than on resolving specific parts of it, rooting strategies should first be validated on less profound questions of similar difficulty. In our case, we used the example of the monophyly of bivalves. In our view, it is not advisable to directly apply new approaches, no matter how sophisticated, to locate the root of the tree of life without prior thorough validation on challenging questions whose answers are known. The necessary test data are relatively easy to assemble given the wealth of publications available today. Under these prerequisites, we argue that the supermatrix approach remains the method of choice for rooting the tree of life because it is the most widely used and validated strategy (Boussau et al., 2013; T. A. Williams et al., 2017).

Moreover, the discovery of the Asgard archaea group during my thesis encouraged the scientific community to consider a two-domain model of life, favoring a fusion scenario to explain the origin of the current eukaryotic cell. This led me, in the second chapter, to construct three datasets of 352 OTUs containing (1) 90 ribosomal proteins, (2) 343 orthologous genes, and (3) 117 "neo-orthologous" genes derived from the phylogenetic splitting of paralogous gene families. These datasets now represent a set of genes that can be used in future analyses to reconstruct the phylogeny of archaea in light of new genomes. One of the major challenges in assembling datasets for inferring species phylogeny was managing paralogy and xenology, both of which lead to multiple sequences per species for a given gene, with some not reflecting the true phylogeny of the organisms. I defined monophyletic groups based on our results and systematically varied taxonomic sampling by generating five replicas, followed by gene jackknives. I then inferred trees using two different strategies: supermatrices and supertrees. In each case, various sequence evolution models were employed to best assess the relevance of each and the biases (evolutionary process heterogeneity) and artifacts likely to affect our phylogenies. I observed that species selection plays a significant role in the resulting topology. However, in gene jackknife manipulations, RF distances were low as the trees were congruent. Gene selection had less influence on tree topology as long as sufficiently sophisticated models were used (particularly categorical models and PMSF). Finally, supermatrices tended to yield more robust results than supertrees. After analyzing the majority bipartitions, I identified four tree topologies that could be observed, revealing inconsistencies depending on species selection. In particular, we noted the following incongruences (**Figure 44**):

- The Korarchaeota, which position themselves either as a sister group to the Crenarchaeota or as a sister group to the Crenarchaeota + SCGC + Verstraearchaeota clade.
- The Hadesarchaeota, which appear either as a sister group to the Thermo Theionarchaeota group or as basal to all Euryarchaeota and DPANN.
- The Altiarchaeales, which position themselves either as a sister group to the DPANN or as a sister group to the Methanobacteria.

To reduce systematic error, I chose to test our dataset by eliminating sites with rapid evolutionary rates using a slow-fast approach (Brinkmann & Philippe, 1999; Roure & Philippe, 2011) to favor slower sites, which are less likely to contradict the assumptions of substitution process homogeneity. My analyses support the Korarchaeota rather than the SCGC (Single Cell Genomics Center) group as a sister group to Sulfolobales + Desulfurococcales + Thermofilaceae + Thermoproteaceae. However, it is difficult to determine which of our four topologies is the correct one. Indeed, species selection seems to be an important factor in the results obtained. My results highlight the difficulties encountered in tree inference. Improving the quality of this inference regarding the elimination of problematic data depends on a delicate balance between noise (stochastic error), non-phylogenetic signal (systematic error), and authentic phylogenetic signal (historical signal) (Baurain & Philippe, 2010) which is difficult for current automatic methods to account for, and thus requires further methodological developments. Our study would tend to favor the proximity of the DPANN and Altiarchaeota at the base of the Euryarchaeota, but this result appears to be dependent on taxonomic sampling. Additionally, the position of the Hadesarchaeota remains uncertain, although it seems to mostly cluster with what we referred to as the TTM group (Thermococci + Theionarchaea + Methanomicrobia Arc).

In the third chapter, I decided to include a selection of eukaryotes within our archaeal alignment to evaluate their supposed relationship with the Asgards. To address the numerous paralogous sequences within the eukaryotic subtrees, I performed preliminary phylogenetic analyses for each gene and manually selected orthologous sequences to the archaea. After numerous individual gene tree analyses, I constructed super-matrices of 126 genes. Since I had previously found four alternative tree topologies for the archaea based on our species selections, I checked that the addition of eukaryotes did not impact these topologies. Similarly, I verified and ensured that reducing from 343 genes to 126 genes did not change the topologies of my five species replicas. These results reassured me in the reliability of the gene selection I had made, as well as the rigorous work involved in constructing my dataset.

I then wanted to assess the impact of heterotachy (heterogeneity of substitution rates at a site over time) and heteropecilly (heterogeneity of substitution processes at a site over time) on the trees obtained. To do this, after inferring specific evolution rates for each site independently for archaea and eukaryotes, I gradually filtered out columns based on a delta speed calculation to create super-matrices with columns having more or less heterogeneous evolutionary rates. Regarding heterotachy, contrary to what recent literature suggests (Aouad et al., 2022; Eme et al., 2017), the clustering of Asgards with eukaryotes is very poorly represented when focusing on homotachy sites between archaea and eukaryotes, in favor of clustering Asgards with TACK. A large super-matrix size (18,000 positions) with the inclusion of sites with more heterogeneous substitution rates is required for Asgards to consistently cluster with eukaryotes, regardless of our taxonomic sampling. Thus, the grouping of Asgards with eukaryotes could be the result of a phylogenetic reconstruction bias linked to sites exhibiting poorly modeled heterotachy. On the other hand, heteropecilly analyses tend to show the opposite. When focusing on homotachy sites between archaea and eukaryotes, eukaryotes are positioned as a sister group to Asgards.

Thus, in my thesis, the removal of the most saturated sites proves to be an effective method for improving the extraction of phylogenetic signal, by avoiding the generation of non-phylogenetic signal and focusing solely on the historical signal, thereby supporting a solution that could be more relevant. However, when looking at the history of our understanding of the evolution of the eukaryotic cell, we may question whether some of the proposed solutions are not the result of problems that science has already encountered in the past. For example, the supposed archaeal root of the DPANN and the close relationship between the Asgards and eukaryotes could be a new version of what the Archezoa were to eukaryotes, namely symbiotic species with highly reduced genomes suffering from a long-branch attraction artifact. Thus, based on our results, we suspect that eukaryotes are actually at the base of the Ouranosarchaea, thereby explaining the sparse distribution of ESPs as a result of secondary simplifications.

The root of the tree of life remains an open question. Historically, two cellular structures (prokaryotes vs eukaryotes) were opposed, but the first molecular phylogenies of the tree of life highlighted the uniqueness of archaea and their similarities with eukaryotes (E. Chatton, 1925; É. P. L. Chatton, 1938; C. Woese et al., 1978; CarlR. Woese & Fox, 1977). This raised the question of whether the tree of life should include two or three domains. Many hypotheses have emerged regarding the origin of the eukaryotic cell, which seems to be a chimera between a bacterium and an archaeon, although it also possesses characteristics unique to it (T. M. Embley & Martin, 2006; T. M. Embley & Williams, 2015; Eme et al., 2017; Hartman & Fedorov, 2002; Lake, 2007). The discovery of the Asgard group was therefore significant (Eme et al., 2017; Spang et al., 2015; T. A. Williams et al., 2017; Zaremba-Niedzwiedzka et al., 2017). It confirmed the close relationship between eukaryotes and archaea and supported 2D scenarios (the Eocyte hypothesis). However, even though the phylogenetic gap between archaea and eukaryotes has been reduced, the proposed origin of eukaryotes within the Asgards still raises the same problems. Only two Asgards have been cultured so far. The first cell is *Candidatus Prometheoarchaeum syntrophicum* (Imachi et al., 2020), a very small Lokiarchaeota (0.5 mm) that shows no signs of cellular complexity or membranous organization. It is metabolically deficient, living in symbiosis with a sulfate-reducing bacterium of the *Desulfovibrio* genus and a methane-producing archaeon of the *Methanogenium* genus, which contradicts what we would expect from a complex proto-eukaryote. In 2022, a new species of Asgard collected from the mud of an estuary in Slovenia, *Lokiarchaeum ossiferum*, was cultivated (Rodrigues-Oliveira et al., 2023) (after seven years of efforts) and studied in detail by electron microscopy, revealing several dozen fine tentacles with thickenings and outgrowths as well as a cytoskeleton extending into the tentacles. The genome of this species and that of another species discovered in the same location were fully sequenced. They notably contain four genes encoding protein complexes that, in eukaryotes, are involved in folding, cutting, and assembling membranes to connect internal compartments. The hypothesis of horizontal gene transfers (HGT) in proto-eukaryotes is relevant in both 2D and 3D scenarios. This implies that, regardless of the correct scenario, the study of ESPs could provide important information on eukaryogenesis by identifying intermediate stages in the evolution of these proteins. In the modern biosphere, HGTs between eukaryotes and bacteria are particularly widespread among species living in close association. Symbiotic associations between archaea and eukaryotes have also been described, such as methanogens thriving in protists of various eukaryotic lineages (Husnik et al., 2021) or *Cenarchaeum symbiosum* living in marine sponges (Preston et al., 1996). It is therefore possible that some ancestors of the Asgard group were symbionts of proto-eukaryotes, exchanging genes with their hosts. Thus, if the Asgard ESPs are the result of horizontal gene transfers, they do not explain the origin of eukaryotes. Furthermore, this hypothesis implies that ancient Asgards were already diversified before the LECA and shared their biotopes with proto-eukaryotes (since they appear to be symbiotic). Therefore, studying the environmental distribution of modern Asgards could provide essential information on the nature of the biotopes in which proto-eukaryotes thrived. It could be interesting to study whether some Asgards live in symbiosis with modern eukaryotes by searching for Asgard signatures in various types of eukaryotic cells. Our work tends to favor the root of archaea within the SANT group. We do not systematically find the monophyly of the Euryarchaeota.

Finally, the issues related to the mosaic nature of genomes and the numerous biochemical, metabolic, and cellular differences between archaea and eukaryotes, including the biocompatibility of lipids, remain unresolved (Martin & Muller, 1998; Peretó et al., 2004; Valentine, 2007). The lipids of archaea consist of long chains of isoprenoid alcohols attached to glycerol by ether bonds, while eukaryotes and bacteria form their membrane lipids by assembling two fatty acid chains with a glycerol molecule through ester linkages. Eukaryotic chromosomes are linear and lack plasmids. Their genetic material is surrounded by a nucleus, and they possess numerous organelles. Archaea, on the other hand, exhibit a great metabolic diversity (chemolithotrophy, methanogenesis, nitrogen fixation, fermentation, anaerobic and aerobic respiration, etc.).

Nearly 50 years after their discovery, the genomic revolution has revealed that archaea are highly diverse organisms. Archaea have been discovered in almost all habitats (whether through metagenomes or cultured strains), and it is highly likely that many other archaeal lineages remain unknown. Archaea play key roles in important biogeochemical processes and are abundant in the microbiome of animals, including our own. Although several metagenomic analyses have contributed to revealing more about their diversity and prevalence, the fact that no archaeal pathogen has been discovered has not sparked much interest in their study (which, it should be noted, is quite curious and mysterious). It is unfortunate that most of the general public remains unaware of their existence. They are even present in our belly buttons, yet we still know little about them, despite their discovery in… 2012 (Hulcr et al., 2012). Their diversity within the microbiome remains understudied and poorly understood. It is only recently that a study revealed their importance across a broad spectrum of the animal kingdom, with their presence in 175 species out of 250 sampled (Thomas et al., 2022). Our lack of knowledge about the true diversity of archaea greatly hinders our ability to understand the relationship between archaea and eukaryotes. Now, one of the most important steps would be to further explore the diversity of life. With new technologies and sequencing techniques, the quantity of available genomes is impressive. The discovery and sequencing of unexplored branches of the tree of life could enhance the resolution of phylogenetic trees by mitigating some artifacts, particularly by breaking long branches. However, this massive influx of data that we now have access to must be harnessed appropriately with methods that can account for the complexity of biological processes if we are to hope to understand the relationships between bacteria, archaea, and eukaryotes.

BIBLIOGRAPHY

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

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