Reconstructing the evolutionary relationships between archaea and eukaryotes : a phylogenomic approach

Kasie Raymann

Université Pierre et Marie Curie

Complexité du Vivant

Biologie moléculaire du gène chez les extrêmophiles, Institut Pasteur

Reconstructing the evolutionary relationships between Archaea and Eukaryotes: a phylogenomic approach

Par Kasie RAYMANN

Thèse de doctorat de biologie

Dirigée par Simonetta GRIBALDO

Présentée et soutenue publiquement le 19 septembre 2014

Devant un jury composé de :

Emmanuel GOUY	**Directeur de recherche**	**Rapporteur**
Thorsten ALLERS	**Directeur de recherche**	**Rapporteur**
Frédéric DELSUC	Chargé de Recherche	Examinateur
Guennadi SEZONOV	Professeur à l'UPMC	Examinateur
Céline BROCHIER-ARMANET	Professeur à Université Lyon 1	Examinateur
Simonetta Gribaldo	Chargé de Recherche	Directeur de thèse

In loving memory of my grandmother, Margarie Kay Raymann Kocher
April 15, 1939- August 20, 2014

She was a very strong and inspirational woman who always encouraged and supported me throughout my life. She brought light and love to the lives of everyone she met. She will always be remembered and greatly missed.

Acknowledgements

Foremost, I would like to express my sincere gratitude to my advisor Simonetta Gribaldo for her continuous support, motivation, and patients throughout my Ph.D. As a successful young female scientist she has been an inspiration for me. She has helped me develop as a scientist and has taught me so many things. During difficult times when I struggled to be abroad she was always understanding and supportive. Her enthusiasm, knowledge, and great interest in many diverse topics kept me motivated even in rough times. She was a great mentor and advisor and I am honored that I had to opportunity to work with her during my studies.

I would also like to thank Patrick Forterre and Céline Brochier-Armanet whom I worked closely with over the last three years. They both helped me in many ways and provided me with scientific guidance and support. It was a great honor to work with them.

I would like to thank my thesis advisory committee, Isabelle Rosinski Chupin, Philippe Deschamps, Manolo Gouy, and Pierre Netter, for their time, guidance, and constructive evaluation during the last three years.

I gratefully acknowledge the members of my Ph.D jury, Manolo Gouy, Thorsten Allers, Frederic Delsuc, Guennadi Sezonov, and Céline Brochier-Armanet, for their time and valuable feedback on my thesis manuscript.

I am very thankful for all the amazing friends and colleagues that I have met in Paris. They have made my time here unforgettable and I hope we will always stay in touch.
** You are the people that I wanted to know and you're the places that I wanted to go. **
People as Places as People -- Modest Mouse

I want give a special thanks to Louis-Marie for all of his love and support. He not only provided me with emotional support but also scientific inspiration. Literally, I would not be where I am today without him. He makes me a better person and a better scientist. I especially want to thank him for all of our long discussions about life, culture, history, and science. He has taught me so much about France and French culture and made my time in France more than incredible. I look forward to our next journey together.

I want to acknowledge my best friend Kate whom I could never thank enough for all that she has done for me. I couldn't imagine a better person to have shared this experience with. I don't know how I survived 25 years without her. No matter how far apart we are I know we will also be close. I love my beautiful bird.

I am infinitely indebted to my parents who have always loved and supported me through good times and bad. As a small town girl from Indiana I know that I would never be where I am today without their continuous encouragement to always follow my dreams. I sincerely thank them for being such amazing parents. I also want to thank all the members of my family for their continuous love and support.

Finally, I want to thank the Pasteur-Paris University International doctoral program (PPU) for giving me the opportunity to do my doctoral research at Institut Pasteur. Thank you to all of the PPU committee members for all of your help and support. Also, I want to give a special thanks to Alice Zuccaire, daughter of Paul. W. Zuccaire and president of the Paul W. Zuccaire foundation (Pasteur Foundation), for providing the funding for my studies.

Introduction

Early taxonomic classification and the discovery of the Archaea

The classification of organisms is a very ancient practice, with informal classification of organisms dating back to Aristotle (384-322 BC) in his book *Historia Animalium* (Aristotle 1993). Although many other systems of classification were created following Aristotle, it wasn't until the 18th century that Carl Linnaeus published the first hierarchal system for naming species (Linné 1735). Linnaeus described three Kingdoms, animals, plants, and minerals, with all living organisms being classified as either plants or animal. These Kingdoms were further divided into classes, orders, families, and genera based on shared phenotypic characteristics (Linné 1735). Although bacteria had been observed over 100 years earlier by Antonie van Leeuwenhoek, they were classified in the plant kingdom until the 19th century (Sapp 2009). The publication of On the Origin of Species in 1859 by Charles Darwin paved the way for phylogenetic cladistics by introducing the theory that the diversity of life arose from common decent with modification through a branching pattern of evolution (Darwin 2006). In 1866, Ernst Haeckel, a strong advocate of Darwin's theories, defined a new kingdom that included all unicellular organisms called *Protista* (Haeckel 1866). This kingdom was subdivided into several groups, including what he called the monerans, structureless and homogeneous organisms. This group included bacteria and some small eukaryotes. Haeckel was the first to represent all living organisms in the framework of a phylogenetic tree, with Protista having equal rank to kingdoms Animalia and Plantae (Figure 1).

In 1937, Edouard Chatton, French biologist and mentor of André Lwoff, was the first to classify organisms with nucleated cells (eukaryotes) and organisms that do not have a distinct nucleus (prokaryotes) into two different groups (Sapp 2005), introducing the famous Prokaryote/Eukaryote dichotomy. In 1938, Herbert Copeland proposed a four-kingdom classification system that moved "prokaryotic organisms", or Heaeckel's Monera, into a separate kingdom (Copeland 1938).

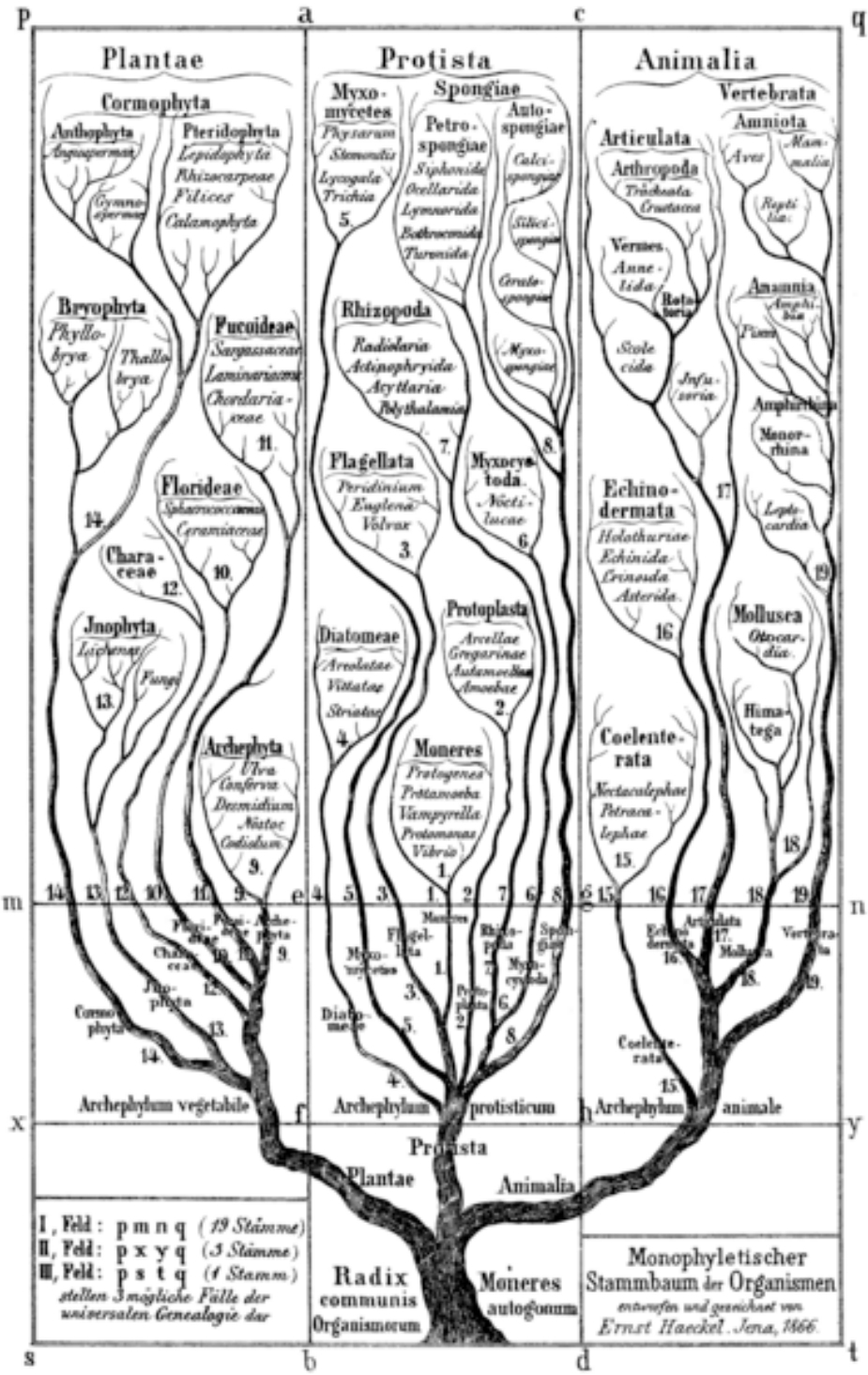


Figure 1. Phylogenetic tree of Ernst Haeckel from 'Generelle Morphologie der Organismen' (1866) with the three branches Plantae, Protista, Animalia. Protista were composed of the monerans (called Moneres in German). Haeckel's Monera included not only bacterial groups but also several small eukaryotic organisms.

Further developments in microscopy and microbiology after the Second World War led to the distinction of bacteria as separate from viruses and other cellular organisms. However, viruses and bacteria were classified together in a new kingdom called Protophyta (Breed et al. 1948), which was maintained until the late 1950's (Sapp 2005). Based on molecular structure

and physiology, in 1957 Andre Lwoff distinguished viruses from bacteria by demonstrating that viruses contained only one type of nucleic acid (RNA or DNA) and they do not reproduce by division like a cell (Lwoff 1957). "The Concept of a Bacterium" by Stanier and van Niel followed Lwoff's concept of viruses, however they could only define prokaryotes negatively in relation to eukaryotic organisms, with the main features being the absence of internal membranes and nuclear division by fission (Stanier and Niel 1962). The only positive definition of bacteria was the presence of a cell wall that contains a specific mucopeptide for its structure (Stanier and Niel 1962). The classification of prokaryotic and eukaryotic organisms into two separate superkingdoms or empires, which was at first only a cell organization distinction, quickly turned into a phylogenetic distinction (Sapp 2005). This was followed by formal proposals of the two kingdoms, *Prokaryae* and *Eukaryotae* (Stanier and Niel 1962; Allsopp 1969). However, the ability to classify bacteria in terms of natural or "phylogenetic" relationships was heavily debated because unlike plants and animals that could be classified based on physiology and comparative anatomy, bacteria lacked complex morphological traits and displayed a huge level of physiological diversity, which is difficult to interpret. In 1963, Stanier emphasized this by stating that bacteria as a whole were monophyletic, given their prokaryotic cell structure, but the evolutionary relationships between different bacteria could not be arranged phylogenetically (Stanier 1963). As Lwoff had proposed that there would be no transitional forms between viruses and cellular organisms, Stanier also proposed that there would be no transitional forms between bacteria and all other organisms (Stanier 1963).

Based on the clear distinction of prokaryotes from eukaryotes, in 1969 Robert Whittaker, who had originally not included Copeland's Monera as a kingdom but had divided autotrophs and heterotrophs into two different kingdoms (Fungi and Protists), published a new classification of organisms in which he proposed the first five-kingdom system, monera (bacteria and blue-green algae), fungi, protists, plants, and animals (Figure 2) (Whittaker 1969). As opposed to Haeckel's classification system, Whittaker's system placed much less importance on unicellular organisms by representing bacteria as the lowest ranking kingdom with progressive evolution toward multicellular organisms. This five-kingdom system was widely accepted, and the debate about the phylogenetic classification of bacteria started to change in light of advances in the field of molecular evolution.

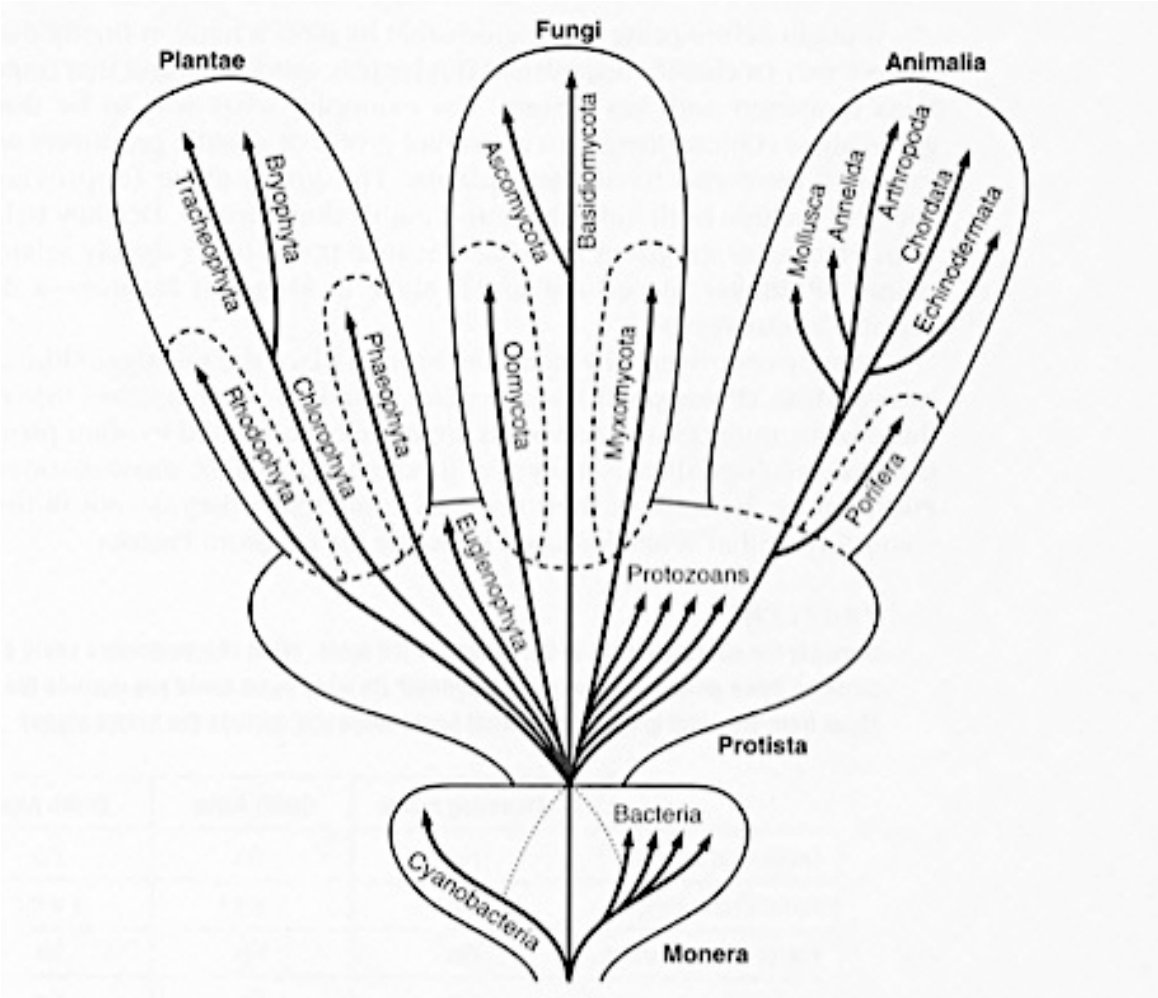


Figure 2. Whittakers five kingdom system from 1969 that separates prokaryotic organisms into the Kingdom Monera and places them at the base of the universal tree of life.

The Molecular Revolution

The molecular revolution began with the discovery of the double helix structure of DNA by Watson and Crick in 1953 (Watson and Crick 1953). A few years later, Fredrick Sanger and colleagues determined that the two-polypeptide chains of insulin had precise amino acid sequences, leading to the finding that every protein had a unique amino acid sequence (SANGER 1959). The protocols developed by Sanger made it tractable to determine protein sequences. The notion of using molecules as documents of evolutionary history was first created by Francis Crick in 1959 (Crick 1958), and more formally by Pauling and Zuckerkandl who pioneered the use of amino acid comparison to infer evolutionary relationships by using hemoglobin sequences to study primate phylogeny (Zuckerkandl and Pauling 1964). They noticed that the number of differences between lineages changes roughly linearly with time, indirectly introducing the notion of the “molecular clock” (Zuckerkandl and Pauling 1964).

Around the same time, Fitch and Margoliash inferred phylogenetic relationships between some animals and fungi by comparing amino acid similarities and differences in the cytochrome C molecule (Fitch and Margoliash 1967). They reinforced the molecular clock concept by suggesting that "differences between cytochrome c of any two species is mostly conditioned by the time elapsed since the lines of evolution leading to these two species originally diverged." (Fitch and Margoliash 1967). The notion of the molecular clock was a very important step in the ability to reconstruct the evolutionary history of organisms, however none of these studies addressed the phylogenetic relationships among all life. Moreover, early genetic approaches were not as successful in classifying microorganisms (reviewed in (Woese 1987)).

In 1977, a publication by Woese and Fox (Woese and Fox 1977) not only revolutionized the study of microbial evolution but the entire field of molecular phylogenetics. Carl Woese who was one of the main contributors to the study of the genetic code was interested in understanding how the translation of amino acid sequences from codons had evolved (Woese 1964). This interest led Woese to study the conservation and variation in the DNA of different organisms using molecular phylogeny. He focused on ribosomal RNA because rRNA is present in all cells, it's abundant, and its sequence changes slowly with time allowing the detection of relatedness among distantly related species (Woese and Fox 1977). Because DNA sequencing technologies were not yet available Woese and colleagues developed a 16S rRNA fingerprinting approach to analyze for the first time the phylogenetic relationships within the prokaryotic domain as well as the relationship between the two primary cell types using a comparative approach (Woese and Fox 1977). In doing so they revealed that "prokaryotes" and "eukaryotes" do not represent a dichotomy because the prokaryotes are divided into two distinct urkingdoms, the eubacteria and the archaebacteria (Woese and Fox 1977). At this time the proposed kingdom "archaebacteria" consisted of only four methane-generating organisms, *Methanobacterium thermoautotrophicum*, *Methanobrevibacter ruminantium*, *Methanobacterium* sp., and *Methanosarcina barkeri*, which had previously been described as possessing a very unique phenotype with cell walls lacking peptidoglycan (Cheeseman et al. 1972), special coenzymes (Fox et al 1977, Taylor et al 1974, Cheeseman et al 1972) (Cheeseman et al. 1972; Taylor et al. 1974; Fox et al. 1977), and distinct tRNA modification (Fox et al. 1977). At the time, these methanogens were an unknown class of anaerobes, but their unqiue methanogenic phenotype made them seem well suited to life on earth about 3-4 billion years ago, hence the suffix name –archae- (Woese and Fox 1977). Following this proposal it was found based on 16S rRNA cataloging that the archaebacteria also included

extreme halophiles, which live in hypersaline brines (Magrum et al. 1978), and thermophiles, which live in geothermal environments (Woese et al. 1978). This new method of phylogeny used by Woese was not immediately accepted, and his suggestion of a tripartite division of all life forms was highly controversial (Albers et al. 2013). Therefore, Zillig's discovery of the distinct subunit composition of archaeal RNAP, which differs dramatically from that of the less complex bacterial RNAP, was an important piece of evidence that strengthened the validity of Woese's phylogenetic classification (Zillig et al. 1978; Zillig et al. 1979). In addition, Zilligs work on RNAP subunit composition also demonstrated that the RNAPs of archaebacteria resembled that of eukaryotic RNAPs, providing the first hint that archaea might share specific features with eukaryotes (Prangishvilli et al. 1982; Zillig et al. 1991).

In 1980 Woese and Fox used 16S rRNA catalogs in an attempt to trace history back to the common ancestor of all extant life and to refine the phylogeny of prokaryotes (Fox et al. 1980) (Figure 3). Fox et al. created a general outline of the phylogeny seen from their comparisons of 16S rRNA catalogs (Figure 3). This phylogeny defined three lines of descent that diverge independently from a common ancestor and separated the archaebacteria into three groups, thermo-acidophiles, halophiles, and methanogens that form a coherent group to the exclusion of both bacteria and eukaryotes. In doing so they created, for the first time, the concept of the last archaebacterial common ancestor (Fox et al. 1980), which is now popularly known as LACA (Makarova et al. 2007). Fox and Woese reinforced the separation of archaebacteria from bacteria and eukaryotes by defining them by distinct characteristics such as "(i) a variety of cell walls, none of which conatin muramic acid (the hallmark of eubacterial walls), (ii) membranes whose major component is a branched chain (phytanyl), ether-linked lipid, (iii) transfer RNA's (tRNA) devoid of ribothymidine in the TΨC loop (T, thymine; Ψ, pseudouridine; C, cytidine), (iv) distinctive RNA polymerase subunit structures, and perhaps (v) an unusual but still not fully elaborated spectrum of coenzymes. (All archaebacteria demonstrate the first four of these properties; the fifth so far is confined largely to methanogens)"(Fox et al. 1980).

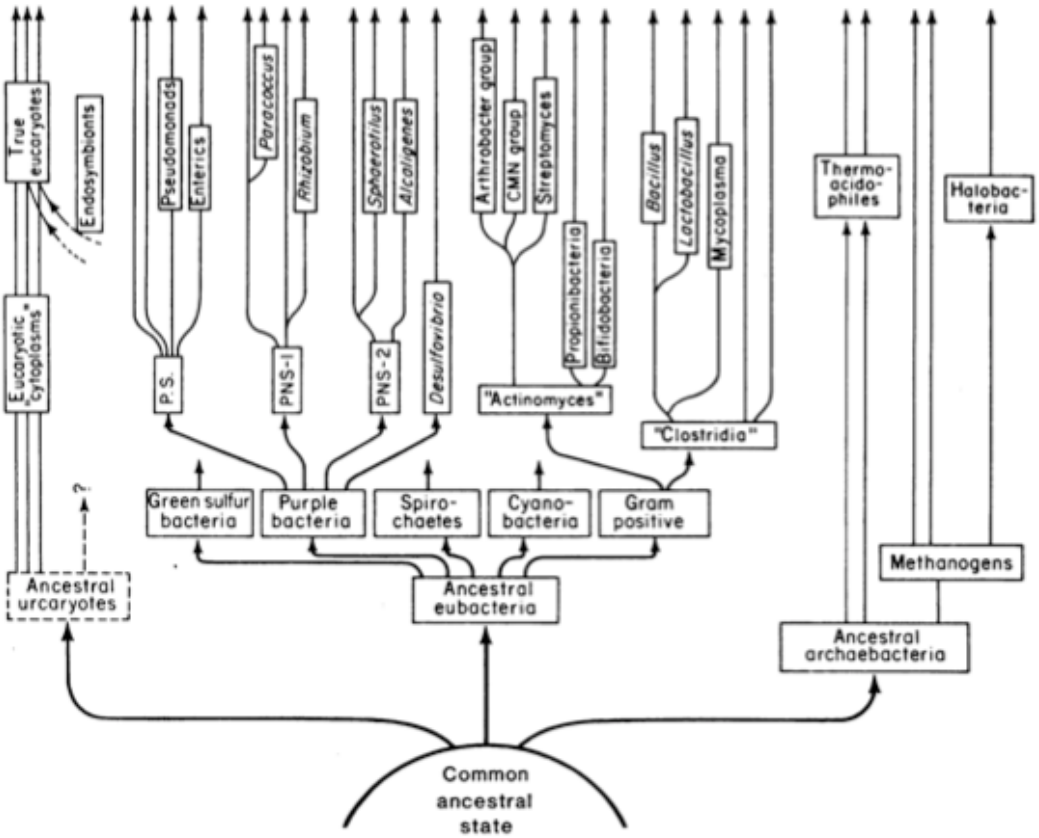


Figure 3. Schematic representation from Fox et al 1980 of the major lines of prokaryotic descent based on results from 16S rRNA catalog comparisons. Three major lines of descent emerge from a common ancestor, ancestral urcaryotes, ancestral eubacteria, and ancestral archaeobacteria.

The first sequence-based phylogeny of the archaebacteria was constructed in 1985 when the complete 16S rRNA sequences for more than 10 archaebacterial species became available (Woese and Olsen 1985). This helped to refine and confirm the earlier concepts of the archaebacterial phylogeny and supported their separation into two divisions, Division I- the thermophilc sulfur dependent species which included the genera Sulfolobus, Thermoproteus, Pyrodictium, and Desulfurococcus and Division II- the three major methanogen groups (Methanomicrobiales, Methanococcales, and Methanobacteriales), the extreme halophiles, and two thermoacidophiles (*Thermoplasma acidophilum* and *Thermococcus celer*) (Woese and Olsen 1985). From this phylogeny the close relationship between the halophiles and methanomicrobiales was revealed as well as the early branching of Thermococcus and the grouping of Thermoplasmatales with the methanogens and extreme halophiles (Figure 4).

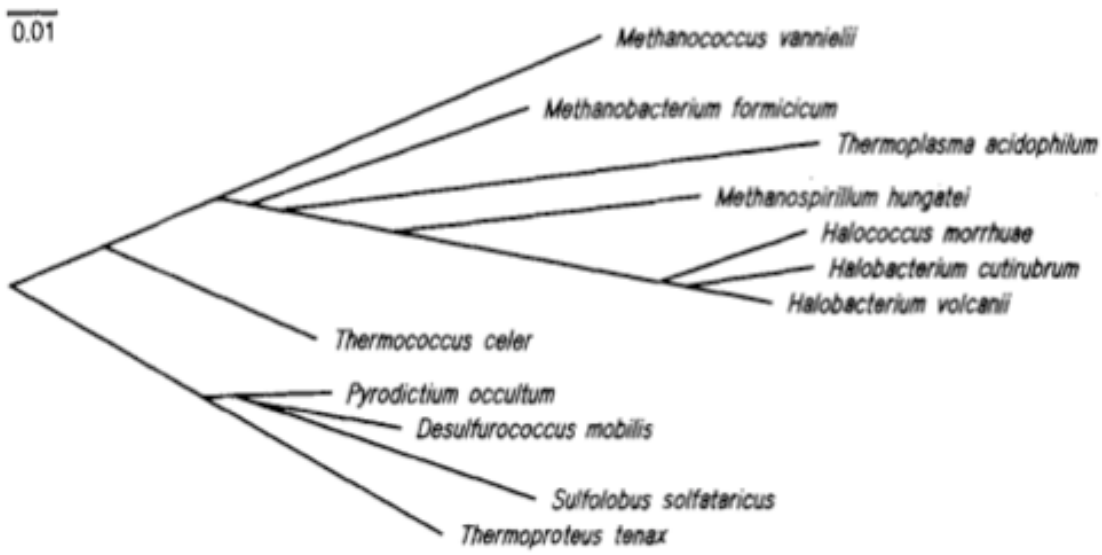


Figure 4. Phylogenetic tree derived from a distance-treeing algorithm based on an alignment of 16S sequences taken from Woese and Olsen 1986. The root was chosen based on various bacterial and eukaryotic sequences. The distance bar represents the number of mutational events per sequences position.

Following the rooting of the tree of life in 1989 based on universal paralogous proteins (Gogarten et al. 1989; Iwabe et al. 1989), Woese and colleagues officially classified the archaebacteria as a separate domain of life (Woese et al. 1990). By placing the root on the branch leading to the bacteria, the archaea and eukaryotes became sister lineages, which entitled Woese to remove the suffix bacteria (Figure 5) and three separate domains of life were defined, the "Domain Eucarya [Greek adjective eu- (good; true in modern common usages); and Greek noun karuōtos (nut or kernel; refers to the nucleus in modem biological usage)]: cells eukaryotic; cell membrane lipids predominantly glycerol fatty acyl diesters; ribosomes containing a eukaryotic type of rRNA. Domain Bacteria [Greek noun βακτηρία (small rod or staff)]: cells prokaryotic; membrane lipids predominantly diacyl glycerol diesters; ribosomes containing a (eu)bacterial type of rRNA. Domain Archaea [Greek adjective ἀρχαῖος (ancient, primitive)]: cells prokaryotic; membrane lipids predominantly isoprenoid glycerol diethers or diglycerol tetraethers; ribosomes containing an archaeal type of rRNA" (Woese et al. 1990). The Archaea were divided into two major lineages, the Crenarchaeota and the Euryarchaeota: the Kingdom Crenarchaeota (Greek noun for spring, fount) comprising most "thermoacidophiles" including the genera Thermoproteus (thermophilic sulphur-dependent organisms) and Pyrodictium (hyperthermophilic organisms) and the Kingdom Euryarchaeota (Greek adjective for broad, wide, spacious) comprising more diverse

organisms; the three methanogenic lineages (genera Methanococcales, Methanobacteriales, and Methanomicrobiales), extreme halophiles, sulfur-reducing species (genus Archaeoglobus), and thermophiles (genus Thermoplasma and the Thermococcus- Pyrococcus group) (Figure 5) (Woese et al. 1990). At the time, the split between the Crenarchaeaota and Euryarchaeota was based solely on differences in rRNA (Woese et al. 1990).

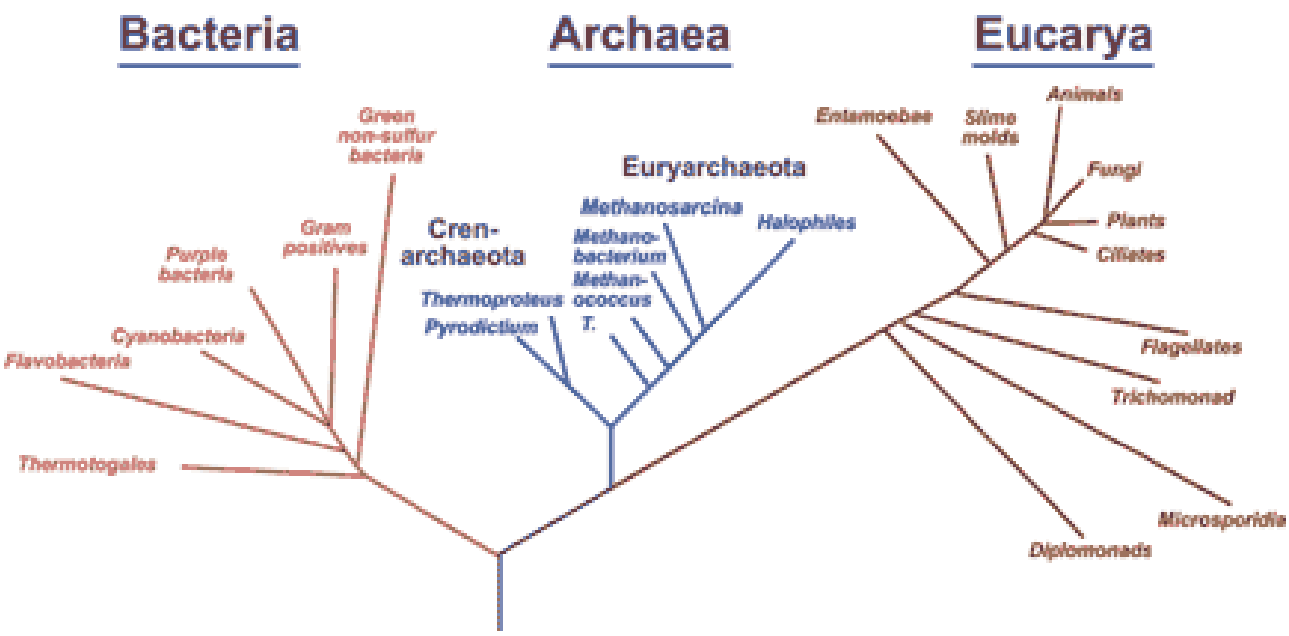


Figure 5. Universal phylogenetic tree of Woese et al 1990. A matrix of evolutionary distances was calculated from an alignment of 16S rRNA sequences and used to construct a distance tree. The position of the root was placed on the branch leading to the bacteria as suggested by comparisons of paralogous gene pairs (Iwabe et al 1989). This root separates the bacteria from the other two primary groups making the archaea and eukaryotes specific relatives.

Section 1. Diversity of the Archaea, from 16S to complete genomes

The first phylogenies of the Archaea

The fact that only a few cultivated strains were available for many archaeal lineages prior to the 21st century made it difficult to understand the true diversity of the archaea (Schleper et al. 2005). With advances in environmental genomic studies, it became possible to characterize novel archaeal lineages using 16S rRNA sequences from PCR-based molecular ecological surveys. From the first ecological surveys it became apparent that the archaea are a very diverse and widespread group on Earth, and that the organisms that have been cultivated in the laboratory and studied in detail, such as methanogens, thermophiles and halophiles, likely represent a minority of archaeal phenotypes (reviewed in (Schleper et al. 2005)). As of 2005, most of the newly discovered lineages fell into one of the two major phyla, Euryarchaeota and

Crenarchaeota (Figure 6). However, the discovery of a few distant and deeply branching lineages, such as the Korarchaeota (Barns et al. 1996), indicated that there was likely a much greater archaeal diversity and that more lineages might be recovered through improved molecular ecological searches, more sophisticated cultivation techniques, and by metagenomic approaches (Schleper et al. 2005).

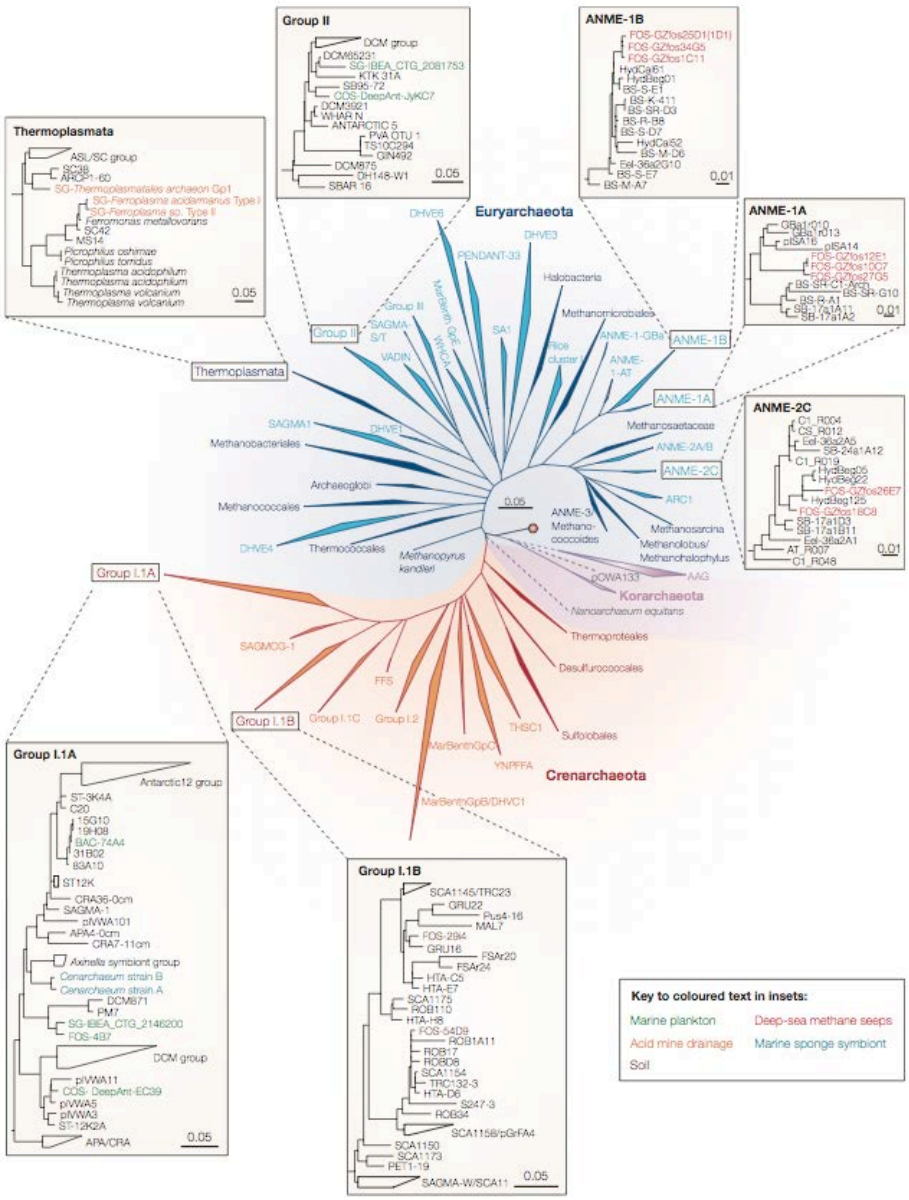


Figure 6. 16S rRNA maximum likelihood tree of 1,344 archaeal 16S rRNA sequences taken from Schleper et al 2005. Groups of uncultivated species that have been targeted in genomic studies are emphasized in boxes. The scale bar for the whole tree represents 0.05 changes per nucleotide. Triangles in light colours represent branches with exclusively uncultivated species, and dark triangles show branches with cultivated species.

Moreover, although 16S rRNA phylogenies are helpful in revealing the diversity of the archaea, they are unable to resolve their deep evolutionary relationships (Figure 6). In fact,

the inability to resolve the deepest nodes of the archaeal tree using 16S rRNA was highlighted by the group of Norman Pace (Figure 7) when they showed that an archaeal phylogeny including cultured and uncultured environmental samples is completely unresolved with regard to the relationships between lineages (Robertson et al. 2005). This irresolution has been attributed to unequal rates of evolution and differences in base composition (Philippe and Laurent 1997). It is also well known that phylogenies based on single genes do not contain enough signal for ancient events (Gribaldo and Philippe 2002).

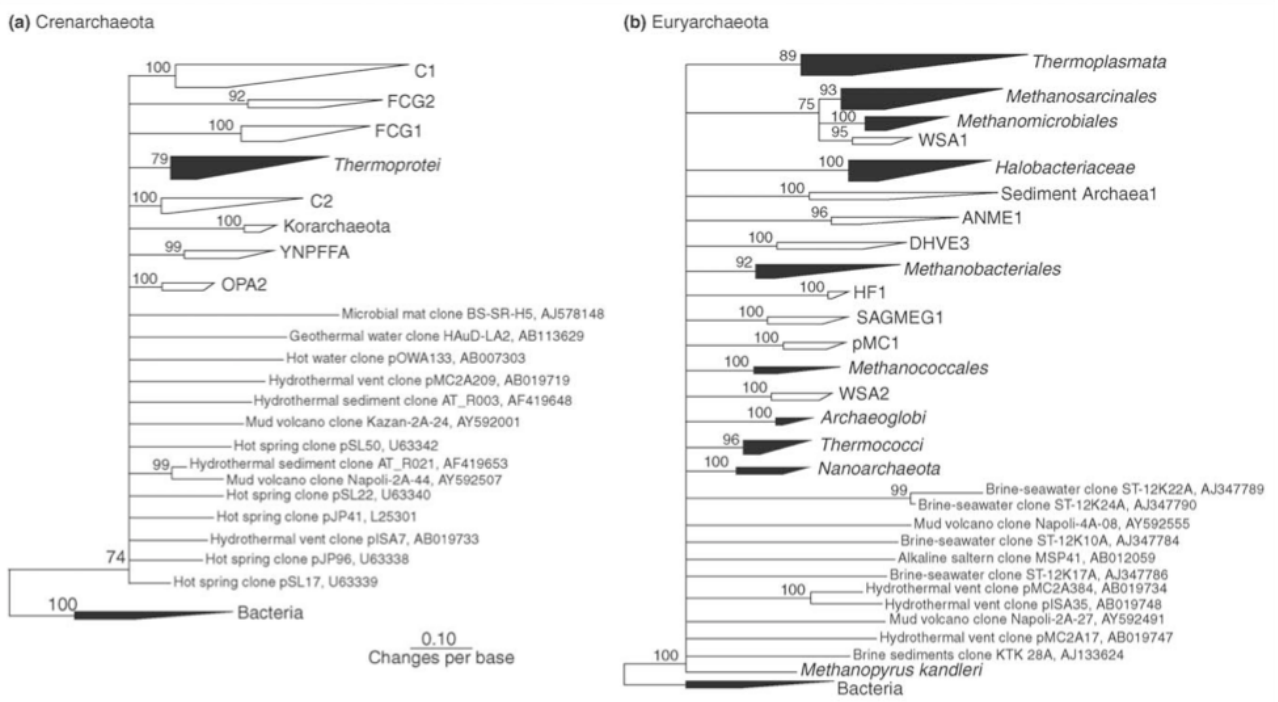


Figure 7. Maximum likelihood phylogenetic trees from Robertson et al 2005 based on 712 archaeal rRNA sequences, generated from RAxML with 100 bootstrap replicates. Three bacterial sequences were used as outgroup. Any nodes in the tree that had less than 70% bootstrap support were collapsed. Solid colored groups have at least one cultured representative; others are known only from environmental samples.

In 1996 the first complete archaeal genome was sequenced, that of *Methanocaldococcus jannaschii* (formerly *Methanococcus jannaschii*) a thermophilc methanogen (Bult et al. 1996). Shortly after the first archaeal genome was sequenced, an exponential rate of genome sequencing occurred, with a doubling time of approximately 34 months for archaea (Koonin and Wolf 2008). The availability of additional completely sequenced genomes made it possible to use alternative markers to 16S rRNA sequences. Phylogenomic approaches, or the use of phylogenetic principles to make sense of genomic data, were popularized as the number of complete genomes began to increase (Delsuc et al. 2005). One approach in phylogenomics is the simultaneous use of many genes to reconstruct early phylogenetic

events (reviewed in (Delsuc et al. 2005)). Large concatenated datasets, or supermatrix approaches, were first used in studies of the early evolution of eukaryotes (Philippe and Germot 2000; Baldauf 2003; Simpson and Roger 2004). However, phylogenomic studies in prokaryotes were limited because as complete genome sequences became available for a larger number of bacterial and archaeal organisms, it became apparent that considerable horizontal gene transfer (HGT) has occurred between prokaryotes (Doolittle 1998; Doolittle 1999; Ochman et al. 2000). This caused some to suggest that there could not be a sequence-based universal phylogenetic tree (Doolittle 1998; Doolittle 1999; Bapteste et al. 2003; Bapteste and Brochier 2004; Bapteste, Susko, et al. 2005; Koonin and Wolf 2008), and specifically that a phylogeny of prokaryotes may not exist (Bapteste and Boucher 2007; Papke 2008). Despite the impact of HGT on prokaryotic evolution, some early phylogenomic methods based on whole-genome analysis resulted in phylogenetic trees similar to the rRNA tree (Snel et al. 1999; Korbel et al. 2002; Wolf et al. 2003; Dutilh et al. 2004). In addition, in 2002 several analyses based on supertrees (Daubin et al. 2002) and supermatrices (Brown et al. 2001; Brochier et al. 2002; Matte-Tailliez et al. 2002) identified a core set of genes that are rarely affected by HGT that can be used to reconstruct the phylogeny of prokaryotes. One of the first phylogenomic studies aimed specifically at reconstructing the evolutionary history of the archaea using a supermatrix approach involved the concatenation of ribosomal proteins (Matte-Tailliez et al. 2002). Ribosomal proteins are highly conserved at the functional and sequence level and are universal, two qualities that had made the SSU rRNA the marker of choice for the phylogeny of organisms. Additionally, they were shown to be very rarely horizontally transferred between organisms (Matte-Tailliez et al. 2002). Subsequent comparative phylogenomic analyses of ribosomal proteins and the components of the transcription machinery including increased taxonomic samplings yielded robust, largely resolved, and congruent trees for the Archaea (Brochier et al. 2004; Brochier, Forterre, et al. 2005; Brochier, Gribaldo, et al. 2005; Gribaldo and Brochier-Armanet 2006; Brochier-Armanet, Boussau, et al. 2008; Brochier-Armanet et al. 2011). The same was also found for components of other macromolecular systems conserved in archaeal genomes such as the exosome and the signal recognition particle (Gribaldo and Brochier-Armanet 2006). These analyses demonstrated that a core of orthologous proteins exists which can be used to infer a robust phylogeny of the archaea. The use concatenation of ribosomal proteins has now become the standard for inferring the phylogenetic positions of newly sequenced bacterial and archaeal genomes (Brochier et al. 2004; Brochier, Forterre, et al. 2005; Brochier, Gribaldo, et al. 2005; Brochier-Armanet et al. 2011; Ramulu et al. 2014).

Over one hundred genomes later

In 2010, just a little more than 30 years after the discovery of the Archaea, 100 complete archaeal genomes had been sequenced and made publically available (Brochier-Armanet et al. 2011). The availability of a large number of archaeal genomes revealed many new insights on archaeal evolution and diversity. The phylum Crenarchaeota consisted of three major orders, Thermoproteales, Sulfolobales, and Desulfurococcales, all represented by hyperthermopilic organisms. The Euryarchaota, being more diverse, comprised ten major orders, Thermococcales, Methanopyrales, Methanobacteriales, Methanococcales, Thermoplasmatales, Archaeoglobales, Methanosarcinales, Methanocellales, Methanomicrobiales, and Halobacteriales.

A phylogeny based on concatenated ribosomal proteins published in 2011 (Brochier-Armanet et al. 2011) provided a very robust tree that recovered all of the major archaeal lineages and orders (Figure 8). In particular several relationships were confirmed such as the early emergences of the Thermococcales, and the split between Methanogens Class I and Class II, supporting the emergence of methanogenesis just after the Thermococcales with subsequent loss of this metabolism in various lineages. The inclusion of two uncultured species distantly related to Thermoplasmatales based on 16S rRNA, '*Candidatus* Aciduliprofundum boonei' and uncultured marine Group II in the protein phylogeny (Figure 8) confirmed their close relationship with Thermoplasmatales and strengthened the late emergence of the Thermoplasmatales which were sometimes indicated as the deepest branching euryarchaeal lineage (Kelly et al. 2011) likely due to HGT from Crenarchaeota and Bacteria (Fütterer et al. 2004). Within the Crenarchaeota the basal branching of the Thermoproteales was recovered, with Sulfolobales and Desulfurobales sharing a close relationship (Figure 8).

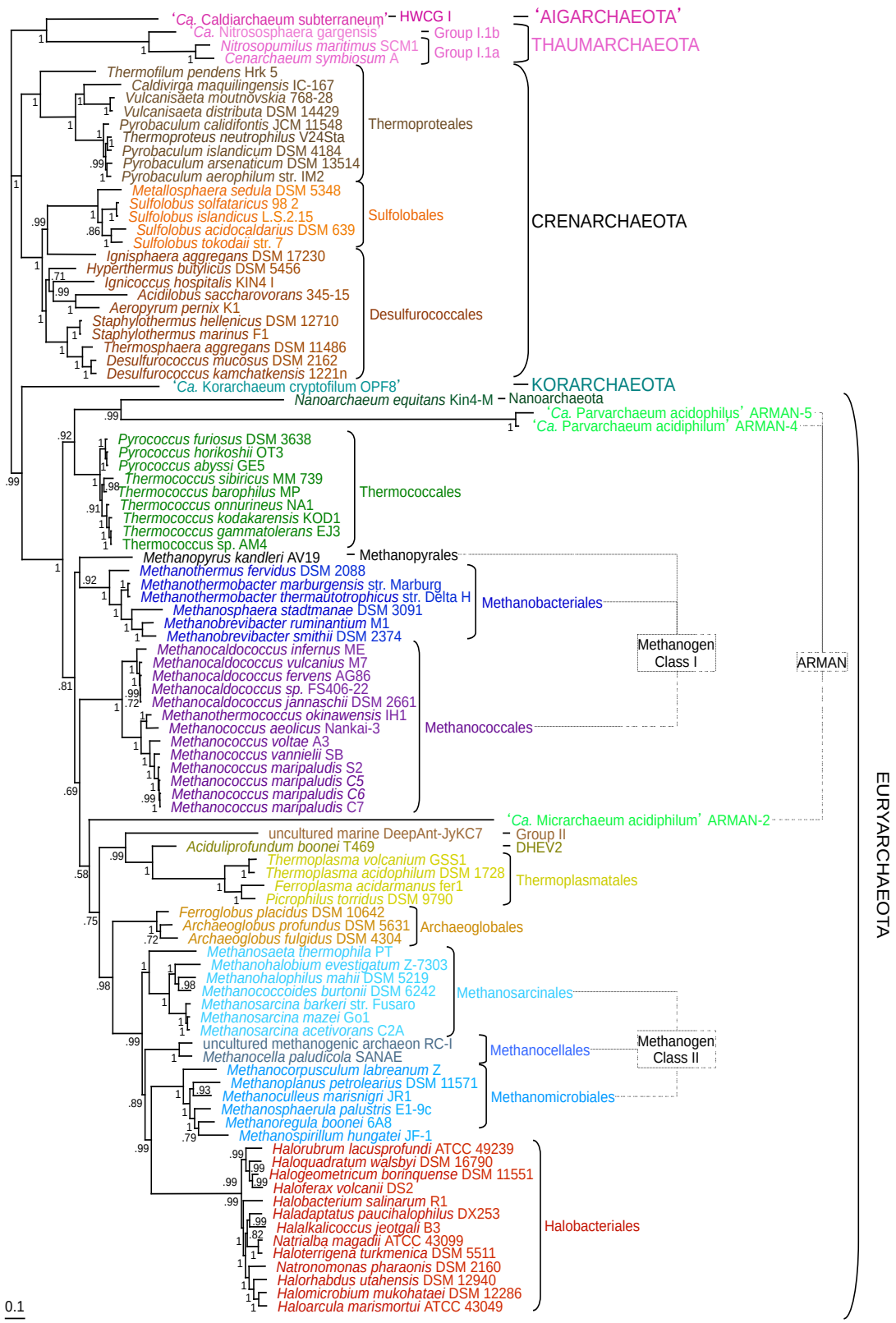


Figure 8. Unrooted Bayesian tree of 99 archaeal species based on the concatenation of 56 ribosomal proteins taken from Brochier-Armanet et al., 2011. The tree was inffered with the CAT+Γ model to take into account evolutionary rate site variations among sites. The scale bar indicates the average number of substitutions per site. Numbers at branches represent posterior probabilities.

The Korarchaeota

Six years after the classification of the Archaea as a separate Domain of life (Woese et al. 1990), a new archaeal phylum was proposed, the Korarchaeota (greek noun for young man or young woman) (Barns et al. 1996). This proposal was based on environmental rRNA sequence analyses of two organisms from a hydrothermal pool (Figure 9) in Yellowstone National Park (Barns et al. 1996). The Korarchaeota exhibit an ultrathin filamentous morphology and inhabit geographically isolated terrestrial and marine thermal environments (Elkins et al. 2008) (Figure 9). The genome of '*Candidatus* Korarchaeum cryptofilum' (Elkins et al. 2008), the first representative of Korarchaeota, wasn't sequenced until over a decade later. This led to the observation that it lacks many biosynthetic pathways, and it encodes ribosomal proteins and RNA polymerases that are closely related to those of Crenarchaeota, but its tRNA maturation, DNA replication/repair, and cell division systems more closely resemble those of Euryarchaeota (Elkins et al. 2008). These unique features were consistent with the proposal that Korarchaeota represents a distinct phylum of the archaea (Barns et al. 1996). Different analyses have disagreed on the placement of Korarchaeota in the archaeal phylogeny, either indicating a basal lineage (Barns et al. 1996) or a deep relationship with the Crenarchaeota (Elkins et al. 2008). However with only one sequenced representative the relationship to the other main archaeal divisions is unclear (Brochier-Armanet et al. 2011).

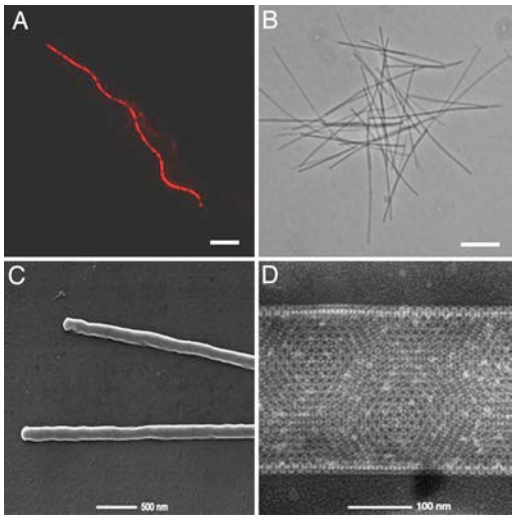


Figure 9. 1) Obsidian Pool, Yellowstone National Park, Wyoming, the hydrothermal pool where Korarchaeaota was first discovered. 2) Microscopy of Ca. K. cryptofilum taken from Elkins et al 2008. (A) FISH analysis with Korarchaeota- specific Cy3-labeled oligonucleotide probes. The cell shape results from drying of the specimen on gelatin coated slides before hybridization. (Scale bar, 5 nm.) (B) Phase-contrast image of korarchaeal filaments after physical enrichment. (Scale bar, 5 nm.) (C) Scanning electron micrograph of purified cells. (D) Transmission electron micrograph after negative staining displaying the paracrystalline S layer.

Thaumarchaeota and 'Aigarchaeota'

Archaea were for a long time considered as only 'extremophiles', or organisms living in extreme environmental conditions such as low and high pH, very high temperatures, and high salinity. This idea was challenged when the first environmental archaeal sequences were detected in marine environments (DeLong 1992; Fuhrman et al. 1992). These mesophilic archaea could be separated into two groups (Group I and Group II). Group I was found to be a sister group of the hyperthermophillic Crenarchaeota (Fuhrman et al. 1992) and Group II emerged within the Euryarchaeota (DeLong 1992; Schleper et al. 1997; Schleper et al. 2005) (see Figure 8). The discovery of mesophilic crenarchaeota revealed that archaea not only inhabit extreme environments but are also a major fraction of the oceans biomass (Pace 1997). A few years later a marine Group I archaeon was isolated living in association with a temperate sponge, *Cenarchaeum symbiosum* (Preston et al. 1996) (Figure 10).

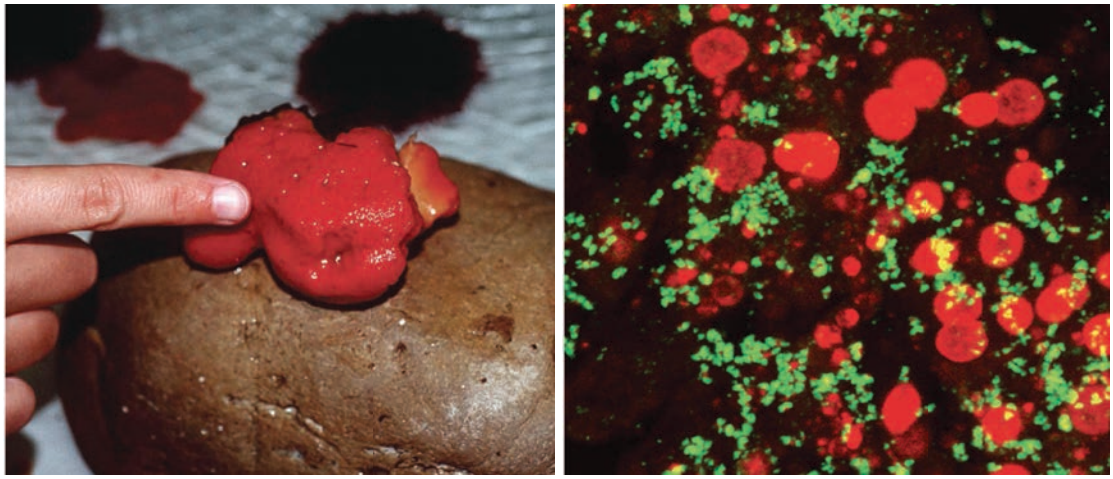


Figure 10. 1) The marine sponge Axinella mexicana and its archaeal symbiont, C. symbiosum taken from (DeLong 2003). 2) Fluorescence Microscopy of C. symbiosum (green) and host A. mexicana (nuclei in red). Image courtesy of C. Preston, Monterey Bay Aquarium Research Institute, Moss Landing, USA.

Small subunit ribosomal RNA analyses placed this organism within the Crenarchaeota (Schleper et al. 1997), strengthening the classification of Group I as mesophilic Crenarchaeota. One of the most important discoveries was finding of mesophilic Crenarchaeota capable of aerobic ammonia oxidation, representing the first observation of nitrification in the Archaea (Könneke et al. 2005). Isolation of *Nitrosopumilus maritimus* (Figure 11), a representative of the marine Group I archaea, revealed that it grows chemolithoautotrophically by aerobically oxidizing ammonia to nitrite (Könneke et al. 2005), a key step in the nitrogen cycle that was previously thought to be restricted to beta- Proteobacteria and gamma-Proteobacteria (Fuhrman et al. 1992). Ammonia oxidizing Group I archaea were subsequently discovered in

high abundance in marine freshwaters, soils, surface sediments, and geothermal habitats suggesting they play a prominent role in the global nitrogen cycle (Figure 11). Furthermore, comparative genomics and physiological studies have revealed a novel primarily copper-based pathway for ammonia oxidation and respiration distinct from that of the known ammonia-oxidizing bacteria (Stahl and la Torre 2012).

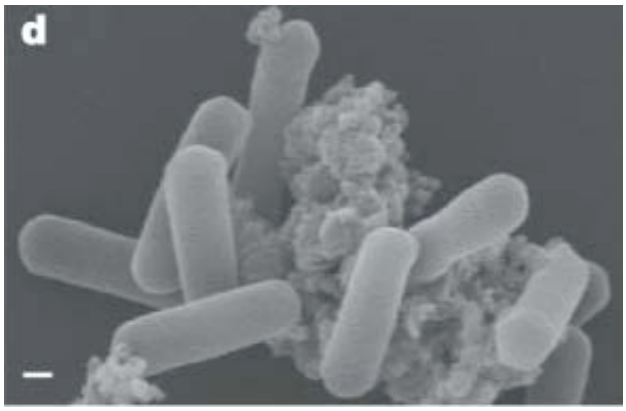


Figure 11. A) Schematic representation of the nitrogen cycle, showing the archaea as contributers to the first step, ammonia oxidation. B) Scanning electron micropgrah of *N. maritimus* taken from Konneke et al 2005 Nature.

Comparative genomic and phylogenetic analysis of the first Group I genome from *Cenarchaeum symbiosum* led to the proposal that mesophilic crenarchaeota represent a third archaeal phylum, the Thaumarchaeota (Greek noun for a wonder) (Brochier-Armanet, Boussau, et al. 2008). This was based on their diversity, genomic distinctiveness, and the fact that they did not branch within any other previously known archaeal phyla in phylogenetic analysis (Brochier-Armanet, Boussau, et al. 2008). The complete genome sequences of other Group I members appropriated the definition of this lineage as a new phylum (Brochier-Armanet et al. 2012). A specific character was also found which seems to be restricted to Thaumarchaeota, a eukaryotic-like topoisomerase IB (TopoIB), a topoisomerase capable of relaxing both negative and positive superturns in supercoiled DNA (Brochier-Armanet, Gribaldo, et al. 2008). Homologs of Topo IB have not been found in any other archaeal lineage, and phylogenetic analysis suggested that this gene was not acquired by HGT from eukaryotes (Brochier-Armanet, Gribaldo, et al. 2008).

The Thaumarchaeota have now been found to constitute one of the most abundant and diverse phyla of the archaea with not only mesophilic lineages from fresh water (Sauder et al. 2011), oceans (Francis et al. 2005; Beman et al. 2008), soil (Ochsenreiter et al. 2003; Leininger et al. 2006), and also human skin (Probst et al. 2012), but also with some lineages from thermophilic environments (la Torre et al. 2008). It is worth noting that the deepest thaumarchaeal branches are composed of uncultivated groups from various hot environments

(la Torre et al. 2008), suggesting that the ancestor of Thaumarchaeota was a thermophile, and that mesophily in some thaumarchaeal lineages is a derived character (Brochier-Armanet et al. 2012).

In addition to the Thaumarchaeota, another new phylum was recently proposed, the 'Aigarchaeota', based on genome sequencing of the first representative of the uncultivated lineage Hot Water Crenarchaeotic Group I (HWCG I), '*Candidatus* Caldiarchaeum subterraneum' (Nunoura et al. 2011). HWCG I is a group of uncultivated archaea that comprises thermophiles detected in high-temperature environments, such as terrestrial surface and subsurface hot springs and deep-sea hydrothermal vents (Barns et al. 1996; Marteinsson et al. 2001; Nunoura et al. 2005; la Torre et al. 2008; Nunoura et al. 2011). The composite genome of '*Ca.* C. subterraneum' was obtained from a metagenomic library of a geothermal water stream (Nunoura et al 2011). Based on phylogenetic analysis, comparative genomics, and the presence of components of the eukaryotic ubiquitin-like protein modifier system previously not detected in archaea or bacteria, '*Ca*. C. subterraneum' and other members of this group were proposed to constitute a new phylum of archaea, the Aigarchaeota (greek noun for dawn and aurora) (Nunoura et al. 2011). However, the proposal of the Aigarchaeota as a new phylum is debated because '*Ca.* C. subterraneum' also encodes features typical of the Thaumarchaeota, such as TopoIB, and emerges at the base of the Thaumarchaeota in phylogenies based on concatenated ribosomal proteins (Figure 8) (Brochier-Armanet et al. 2011; Nunoura et al. 2011). Furthermore, the distances inferred between '*Ca*. C. subterraneum' and the thaumarchaeal lineages are very similar to those seen between different euryarchaeal orders, suggesting that it could represent a deeply branching thaumarchaeal order instead of a new archaeal phylum (Brochier-Armanet et al. 2011).

Nano-sized archaea

In 2002, the first nanosized archaeon was discovered, *Nanoarchaeum equitans*, a hyperthermophilic archaeon from a submarine hot vent (Figure 12) off the coast of Iceland on the Kolbeinsey Ridge (Huber et al. 2002). It is only 400 nm in diameter and possesses the smallest sequenced archaeal genome (0.5 Mb). *N. equitans* grows attached to the surface of *Ignicoccus hospitalis* (Figure 12) an obligate chemolithoautotrophic sulfur reducer within the crenarchaeal family of the Desulfurococcaceae (Huber et al. 2002). *N. equitans* cannot be cultivated in the absence of its host *I. hospitalis*, whereas the latter thrives well without its putative symbiont (Huber et al. 2002). Surprisingly, *N. equitans* could not be detected with the

general primers used for detection of other archaea because it harbors many base-pair exchanges even in the so-called 'highly conserved regions' that are usually employed as primer targets for SSU rDNA PCR (Huber et al. 2002). Based on its basal position in archaeal trees of SSU rRNA (Huber et al. 2002) and concatenated ribosomal proteins (Waters et al. 2003), N. equitans was suggested to represent a new phylum of archaea, the Nanoarchaeota. This proposal was first accepted because of the atypical features found in N. equitans such as split tRNA genes that were interpreted as ancestral characters (Di Giulio 2008). However more detailed phylogenetic analyses suggest that N. equitans is most likely a very fast evolving and derived euryarchaeal lineage closely related to Thermococcales (Brochier-Armanet et al. 2011) see Figure 8.

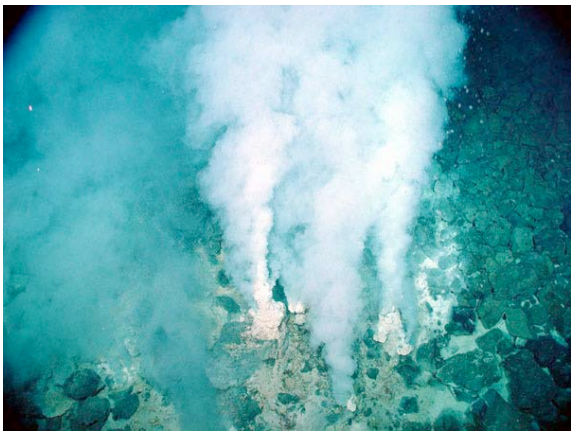


Figure 12. 1) Photo of a submarine geothermal vent, the habitat of N. equitans 2) Microscopy of N.equitains and its host Ignicoccus taken from Huber et al 2002. a) Freeze-etched cell of *Ignicoccus* and four attached cells of N. equitans. b) Ultrathin section of two cells of N.equitans attached to the outer membrane of Ignicoccus. c) Cell of Ignicoccus, with several cells of N. equitans attached d) Confocal laser scanning micrograph Nanoarchaeum (red) and Ignicoccus (green). a–d, Scale bar, 1.0 microm.

In 2010, three new nanosized archaeal lineages were discovered in chemoautotrophic biofilms of the acidic metal rich Richmond Mine of Iron Mountain California ('*Candidatus* Microarchaeum acidiphilum ARMAN-2', '*Candidatus* Parvarchaeum acidiphilum ARMAN-4' and *Candidatus* Parvarchaeum acidophilus ARMAN-5') (Baker et al. 2010). These lineages called ARMANS (archaeal Richmond Mine acidophilic nanoorganisms) live in association with Thermoplasmatales (Figure 13) and contain split genes and high AT contents, typical of streamlined and fast evolving symbionts (Baker et al. 2010). Phylogenetic analyses of several individual genes suggested that the ARMANS branch near the root of the Euryarchaeota (Baker et al. 2010), but phylogenetic analysis of r-proteins rather suggested that ARMAN-4 and ARMAN-5 might be closely related to *N. equitans*, whereas and ARMAN-2 emerges later within the Euryarchaeota (Brochier-Armanet et al. 2011) (see Figure 8). However, the ARMANS are very

fast evolving lineages and their placement in the archaeal phylogeny is still unclear (Brochier-Armanet et al. 2011).

Figure 13. 1) Richmod Mine of Iron Moutain in Redding, California were the ARMANS were discovered and isolated 2) Microscopy of ARMAN and Thermoplasma taken from Baker et al 2010. The ARMAN cell (center, orange) is penetrated by a needle-like protrusion from Thermoplasma (lower left), an Archaea that lives in the same acidic pools as ARMAN. The much smaller, yellow lemon-shaped cells are viruses that also infect ARMAN cells.

Recently discovered archaeal lineages

Since 2010 the number of archaeal genomes available has continued to increase. Recent progress in genomic sequencing technologies and cultivation-independent methods have allowed for uncultivated lineages to be more easily sequenced. Just in the last few years a number of exciting new archaeal lineages have been sequenced which have revealed an even greater diversity of the archaea.

Nanohaloarchaea

High saline habitats are found all over the world in the form of hypersaline lakes, salt ponds, and solar salterns. Although salt-adapted organisms derive from all three domains of life, most extreme hypersaline habitats are dominated by halophilic archaea belonging to the class Halobacteria (Oren 2007). However, recently two novel halophlic archaea ('*Candidatus* Nanosalina sp. J07AB43' and '*Candidatus* Nanosalinarum sp. J01AB56') were discovered in a hypersaline lake in Australia (Figure 14) and near-complete genomes were constructed from a de novo metagenomic assembly of multiple, deeply sequenced libraries (Narasingarao et al. 2012). These halophilc archaea represent another group of nano-sized archaea with very small genome and cell sizes (0.8-1.0 nm with 1.2 Mb genomes) (Narasingarao et al. 2012).

They float on the water surface and display unusual characteristics such as unique amino acid composition, absence of gas vesicle coding proteins, and atypical archaeal metabolic pathways such as genes supporting the entire pentose phosphate pathway including both oxidative and non-oxidative branches (Narasingarao et al. 2012). Phylogenetic analysis of these two organisms showed that they are a distinct lineage distantly related to Halobacteriales (Narasingarao et al. 2012). Similar to haloarchaea, they are suggested to be aerobic heterotrophs incapable of anaerobic respiration (Narasingarao et al. 2012). Given their high abundance and unique characteristics they have been classified as a new class within the Euryarchaeota, called the Nanohaloarchaea (Narasingarao et al. 2012).

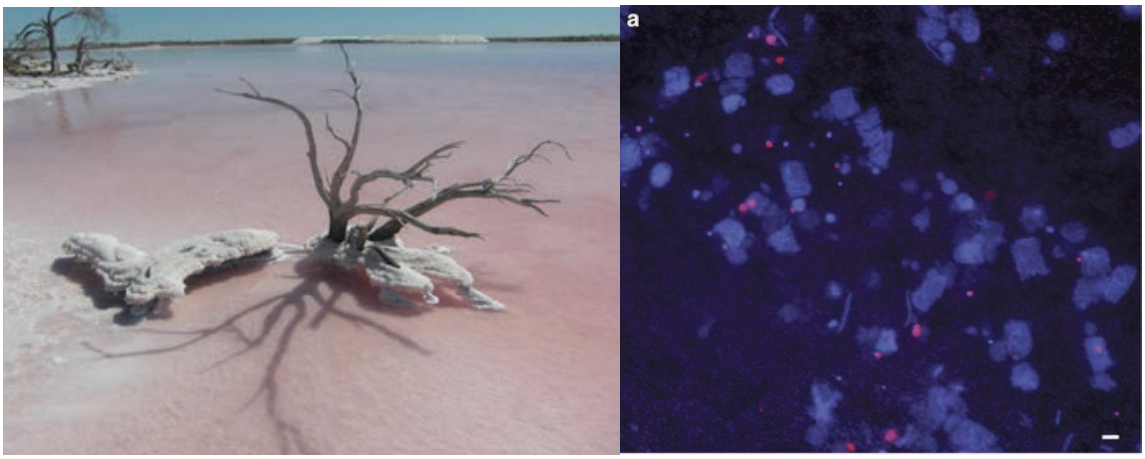


Figure 14. 1) Lake Tyrrell, Victoria Australia where Nanohaloarchaea were discovered. 2) FISH micrographs of water sample from Lake Tyrrell (0.1 to 3 µm filter fraction) taken from Narasingarao et al 2012. Nanohaloarchea cells are stained in red and all other cells in blue.

The 7th order of Methanogens

The first archaeal species associated with the human gut were cultivated and described over 30 years ago, *Methanobrevibacter smithii* and *Methanosphaera stadtmanae*, from the order Methanobacteriales (Miller and Wolin 1982; Miller and Wolin 1985). During the last decade as metagenomic sequencing has become more tractable, a much larger diversity of human associated archaea have been identified. Members related to the Crenarchaeota (Rieu-Lesme et al. 2005), Halobacteriales (Oxley et al. 2010), and more recently to a Thermoplasmatales-related lineage (Mihajlovski et al. 2007; Scanlan and Marchesi 2008; Halachev et al. 2011) have been discovered in the human gut based on 16S rRNA sequences. One member of the Thermoplasmatales-related lineage, *Methanomassiliicoccus luminyensis,* was recently isolated and described as a methanogen that grows on methanol and hydrogen only (Dridi et al. 2012). It was sequenced in 2012 (Dridi et al. 2012). Subsequently, a second member of this lineage, '*Candidatus* Methanomethylophilus alvus', was enriched from the human gut and was found to

be a dominant species in the microbial community (Borrel et al. 2012). Furthermore, uncultured archaea with more than 97% identity in 16S rRNA to '*Ca*. Methanomethylophilus alvus' have also been identified in the digestive tract of various animals (Janssen and Kirs 2008). Two other closely related strains have been enriched from termite gut, MpT1 and MpM2 (Paul et al. 2012), as well as from waste treatment sludge, '*Candidatus* Methanogranum caenicola' (Iino et al. 2013). These lineages were initially given the name "Methanoplasmatales" due to their phylogenetic proximity to the wall-less thermoacidophilic Thermoplasmatales, but a new name has now been proposed for the order, "Methanomassiliicoccales", by Iino et al. 2013 based on the name given to the first isolated species (Iino et al. 2013).

The availability of the first two genomes of Methanomassiliicoccales, *M. luminyensis* and "*Ca*. M. alvus", made it possible to clarify their evolutionary relationships with the other euryarchaeal lineages using other markers than 16S rRNA (Borrel et al. 2013). An analysis of concatenated ribosomal proteins from 84 euryarchaeal genomes firmly demonstrates that M. luminyensis and "*Ca*. M. alvus" represent a monophyletic lineage that is not associated with the previously known orders of methanogens (Figure 15). As previously noted based on 16S rRNA phylogenies, these two lineages share a close relationship with Thermoplasmatales, but also appear to be specifically related to two uncultured lineages: the planktonic Marine Group II (MG-II) and the Marine Benthic Group D (MBG-D) (Figure 15). These results support the proposal of a new order of methanogens, resulting in a total of seven different orders of methanogens within the archaea (Borrel et al. 2013). Several previous phylogenetic studies have supported a single ancient origin of methanogenesis in the Euryarchaeota likely after the Thermococcales with subsequent loss of this metabolism in non-methanogenic euryarchaeal orders (Bapteste, Brochier, et al. 2005; Brochier, Forterre, et al. 2005; Gribaldo and Brochier-Armanet 2006). The existence of this additional order of methanogens now further strengthens this notion and indicates that most of the non-methanogenic euryarchaeal lineages actually derive from methanogenic ancestors through loss of this metabolism (red crosses in Figure 15).

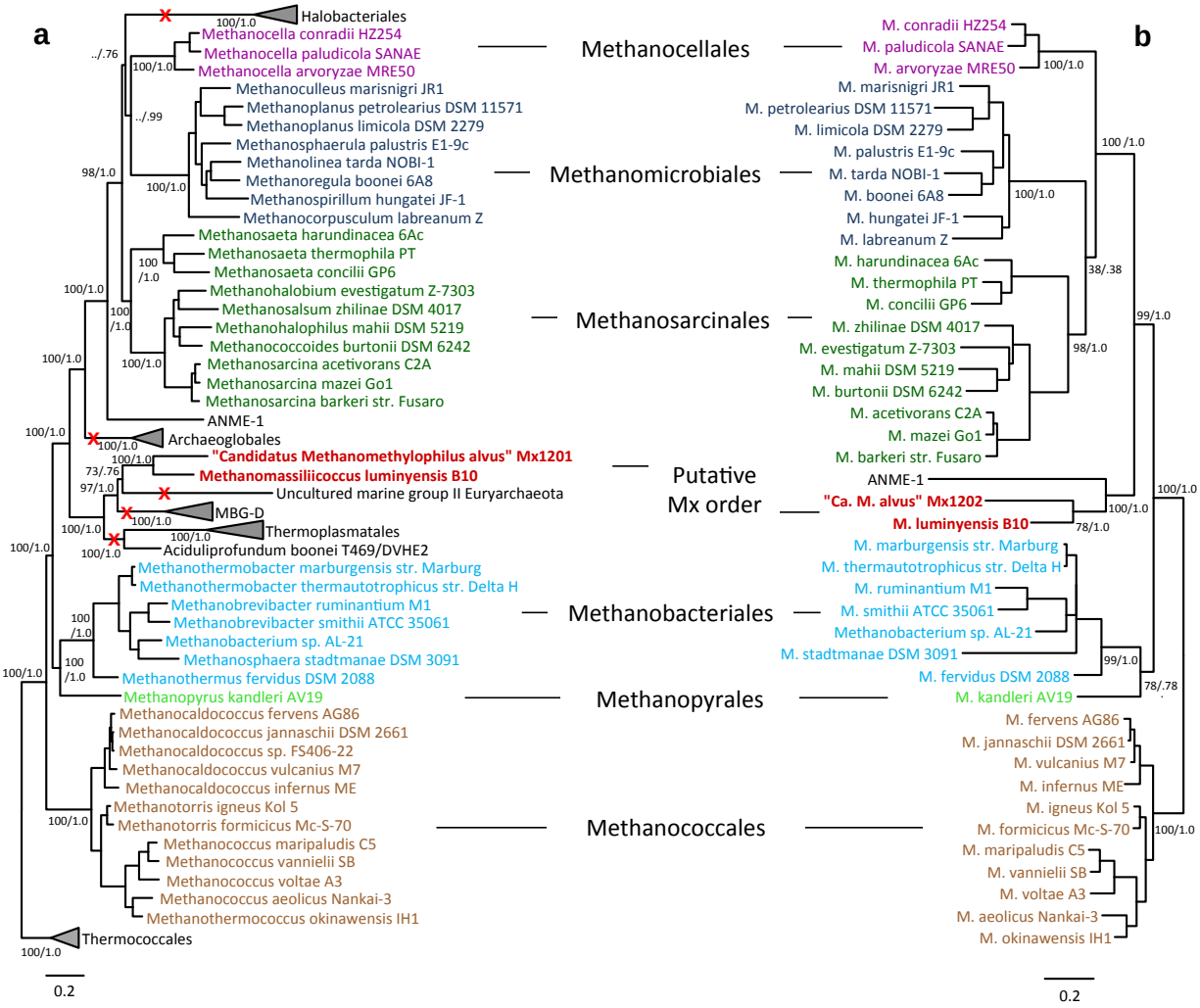


Figure 15. Phylogenetic position of the Mx order and inferred losses of methanogenesis adapted from Borrel 2013. (a) Bayesian phylogeny of Euryarchaeota based on a concatenation of 57 ribosomal proteins. Values at nodes represent Bayesian posterior probabilities and bootstrap values based on maximum likelihood analysis and 100 resamplings of the original data set. The scale bar represents the average number of substitutions per site. Red crosses indicate complete loss of methanogenesis. (b) Maximum likelihood phylogeny of methanogens based on a concatenation of McrA-B-C-D-G protein sequences. The tree based on a concatenation of the five markers of methanogenesis (McrA-B-C- D-G) that are shared by all methanogens is largely consistent with the ribosomal protein-based phylogeny, demonstrating that these genes have not been horizontally transferred.

SM1 archaeaon

The SM1 lineage was first discovered in 2001 living in association with sulfide-oxidizing bacteria in cold sulfurous marsh water of the Sippenauer Moor near Regensburg, Bavaria, Germany (Rudolph et al. 2001). This association involves the sulfide-oxidizing bacteria forming a string-of-pearls community around the SM1 colonies (Figure 16) (Rudolph et al. 2001). In these communities, the bacterial partner and the SM1 euryarchaeon are present in almost equal abundance. SM1 archaea have now been found in sulfide-containing fresh and marine waters all over Europe and can thrive in both anaerobic and aerobic enviroments

(Rudolph et al. 2003). SM1 archaea also possess very unique cell surface appendages called hami beacuse they closely resemble manmade fishhooks or anchors (Moissl et al. 2005) (Figure 16). These appendages have only been observed in SM1 and are believed to be involved in cell anchoring or adhesion (Moissl et al. 2005). Based on 16S rRNA phylogenetic analysis, SM1 branches deeply within the euryarchaeal tree with no close relationship to any of known archaeal lineages (Rudolph et al. 2001). More recently, SM1 has been discovered in biofilms found 1 meter below the water table of a sulfidic spring where the upwelling water is not yet mixed with atmospheric oxygen (Henneberger et al. 2005). This represents the only known example of a naturally occurring archaea-dominated biofilm, with up to 95% being composed only of SM1 (Henneberger et al. 2005). The bacteria found in the biofilm are either randomly distributed or form dense microcolonies (Henneberger et al. 2005). A few other archaea have been detected in the biofilm that appear to be closely related to SM1, but they are present only in very small abundance (Probst et al. 2013).

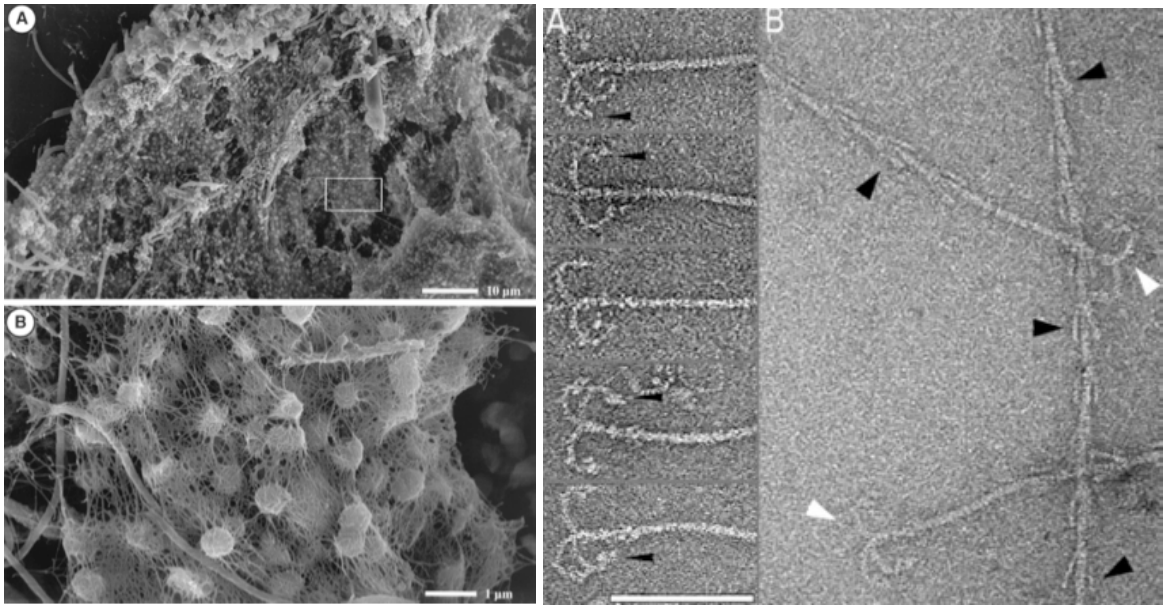


Figure 16. 1) Scanning electron micrographs of a single pearl from the Sippenauer Moor taken from Rudolph et al 2001. A) The inner part of a pearl, showing large numbers of small cocci. B) Detail of the inset in panel A. Single round cocci are embedded in a fibrous matrix. 2) Electron micrographs of SM1 hami appendages taken from Moissel et al 2005. A) Grappling hooks located at the distal ends of the hami taken from XX. Arrowheads indicate locations of barbs. B) High level structured SM1 hami. The hami show prickles (black arrowheads) and grappling hooks (white arrowheads).

Open questions concerning the phylogeny of Archaea

Although the phylogeny of the Archaea is overall well resolved, it is by no means definitive because there are still a few nodes that remain to be clarified. The relationship between some groups needs to be further investigated, for example, the monophyly of Desulfurococcales, the relationships among Class I methanogens, and the relationships among Class II methanogens and their link with Halobacteriales (Brochier-Armanet et al. 2011). The phylogenetic position of some of the very recently discovered lineages, such as SM1, will also need to be investigated once genome sequences are available. Additionally, the position of the nano-sized archaea in the archaeal phylogeny, i.e the Nanoarchaeota, ARMANS, and Nanohaloarchaea, has still not been clarified (Brochier-Armanet et al. 2011). The discovery of nano-sized archaea from a very diverse range of habitats (hydrothermal, acidic, hypersaline) and lifestyles (parasitic/symbiotic, freeliving) may represent convergent evolution, suggesting that small genome and cell size has been selected for multiple times throughout the archaeal diversity. However, some phylogenetic analyses have recovered the monophyly of all of the nano-sized archaea (Rinke et al. 2013; Spang et al. 2013; Williams and Embley 2014). In fact, Rinke et al. (2013) have proposed the existence of a monophyletic superphylum called DPANN whose members would be characterized by small cell and genome sizes and would include the ARMANS, Nanohaloarchaea, and Nanoarchaeota (Rinke et al. 2013). However, given the fast evolutionary rates of these nano-sized lineages further analyses need to be carried out to further investigate their position (Brochier-Armanet et al. 2011).

Aside from resolving the relationships among some lineages, the relationships between the major phyla also need to be clarified. Some of the proposed phyla are only represented by one uncultivated member such as Korarchaeota (Elkins et al. 2008) and Aigarchaeota (Nunoura et al. 2011). The lack of genomic data from other members makes it difficult to determine their relationship to the other phyla, and moreover, to determine if they actually warrant their status as a phylum (Brochier-Armanet et al. 2011). Furthermore, the root of the archaeal tree is still unknown, making it difficult to truly understand the exact relationships between each of the phyla (Brochier-Armanet et al. 2011). The archaea were traditionally rooted in between the Crenarchaeota and Euryarchaeota in trees based rRNA using bacteria as an out-group (Woese and Olsen 1985; Woese 1987). Subsequent studies using rRNA have also rooted the archaeal tree using bacteria which resulted in different rootings such as on the branch leading

to the Thaumarchaeota, the Korarchaeota, or the Euryarchaeota (Brochier-Armanet, Gribaldo, et al. 2008; Elkins et al. 2008; Groussin and Gouy 2011). A few archaeal phylogenies were rooted using the eukaryotes as an out-group, which lead to the suggestion that the Thaumarchaeota might be the deepest branching archaeal lineage (Brochier-armanet et al 2008, Spang Schleper 2010) (Brochier-Armanet, Gribaldo, et al. 2008; Spang et al. 2010). However, most published archaeal phylogenies are left unrooted (Brochier-Armanet et al. 2011). Finding the root is very important to our understanding of the deep evolution of the archaea and the nature of the ancestor.

The Last Archaeal Common Ancestor

The availability of sequenced genomes from a larger diversity of archaea as well as a robust phylogeny of the domain has provided a good framework to reconstruct the evolutionary history and nature of the last archaeal common ancestor (LACA). For example, phylogenomic analyses have been used to reconstruct the evolutionary history of key cellular processes. One study addressed the origin of aerobic respiration by performing a phylogenomic analysis of the four types of dioxygen reductases (O_2Red), the key enzymes of aerobic respiratory chains (Brochier-Armanet et al. 2009). These four systems have very different evolutionary histories, with some of them being transferred from bacteria to archaea or vice versa (Brochier-Armanet et al. 2009). It was shown that family A-O_2Red and is an ancient dioxygen reductase that likely originated prior to the divergence between archaea and bacteria and therefore was present in the archaeal ancestor suggesting it may have been aerotolerant or even capable of oxygen respiration (Brochier-Armanet et al. 2009). Furthermore, a study of the origin of archaeal flagella, which is completely unrelated to bacterial flagella, predicted the presence of two different types of flagella in the last archaeal common ancestor (Brochier-Armanet and Gribaldo 2007). Addtionally, phylogenomic analyses of the enzymes involved in methanogensis have suggested that this metabolism arose late in the archaea, likely after the divergence of thermococcales, with subsequent loss in non-methanogenic members of euryarchaeota (Bapteste, Brochier, et al. 2005; Borrel et al. 2013).

A hyperthermophilic (optimal growth temperatures of >80°C) ancestor of archaea was suggested very early based on the abundance of hyperthermophiles in the archaeal domain and their early branching in rooted archaeal SSU rRNA trees (Woese 1987). The hypothesis of a hyperthermophilic LACA has also been supported by the evolutionary history and distribution of reverse gyrase, a hyperthermophilic specific DNA topoisomerase that produces

positive supercoiling into circular DNA (Déclais et al. 2000; Forterre 2002; Brochier-Armanet and Forterre 2007). A study which specifically aimed to reconstruct the evolutionary history of environmental temperatures in the archaeal domain by analyzing the G+C and amino acid content of bacterial and archaeal ribosomal DNA genes and proteins also predicted that the ancestor of archaea was hyperthermophilic (Figure 17), with an estimated optimal growth temperatures around 82 degrees Celsius (Groussin and Gouy 2011). These results are consistent with previous studies that used phylogenetics to infer ancestral conditions of life (Galtier et al. 1999; Boussau et al. 2008; Gaucher et al. 2008).

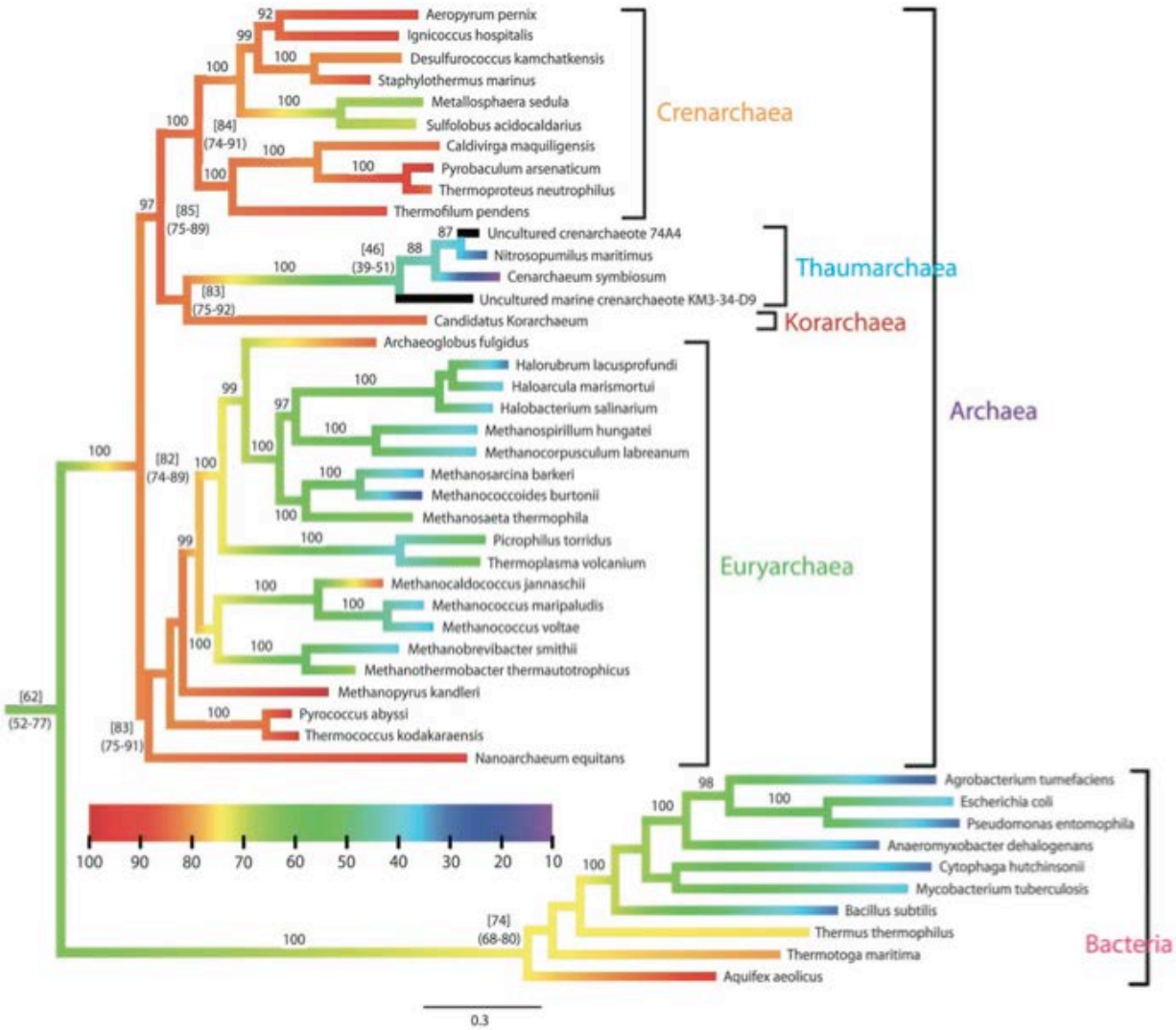


Figure 17. Evolution of OGT in Archaea taken from Groussin and Gouy 2011. The phylogenetic reconstruction is based on 72 genes and on a 9,799 amino acid long alignment. Branch lengths correspond to temperature estimates at nodes, with a linear gradient of color between nodes. No evolution of OGT is represented in the vertical tree lines. The Thaumarchaea branches are colored black because OGTs for uncultured organisms are not available. The branch length scale is in substitution per site. The color scale is in degree Celsius. Mean estimates of temperature at key nodes are given between square brackets with the confidence intervals (95%) for estimates of ancestral OGTs between round brackets. Bootstrap values higher than 85% are represented.

Large-scale comparative genomic studies have also attempted to reconstruct the gene content of the archaeal ancestor (Makarova et al. 2007; Csuros and Miklos 2009; Wolf et al. 2012). According to these reconstructions, each of the major archaeal lineages underwent some degree of genome reduction, suggesting that the ancestor of the archaea was more gene rich and complex than current day representatives (Makarova et al. 2007; Csuros and Miklos 2009; Wolf et al. 2012). The latest reconstruction based on a comparative analysis of 120 archaeal genomes inferred between 1,400 and 1,800 gene families in the last common ancestor of the extant archaea (Figure 18). The patterns of gene gain and loss were mapped on a phylogeny based on ribosomal proteins with the root on the branch leading to the Euryarchaeota with Nanoarchaeota left unresolved (Wolf et al. 2012). After taking into account paralogous proteins and lineage-specific genes in modern archaeal genomes, these authors predicted that approximately 2,500 genes were present in the ancestral genome, which is a larger genome than those of extant archaea (Wolf et al. 2012).

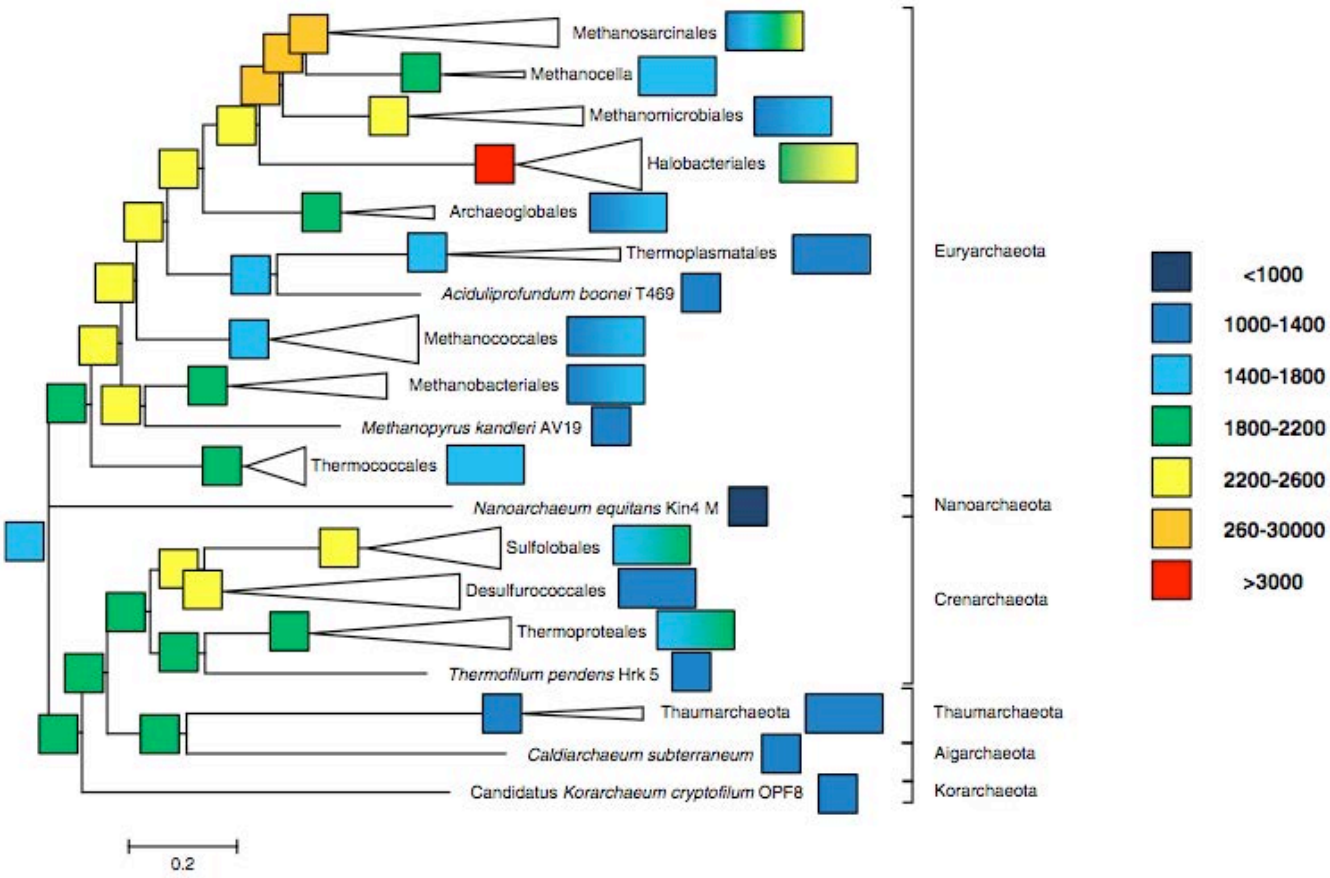


Figure 18. Reconstruction of the evolution of genome size in the archaea taken from Wolf et al 2012. Square boxes at the bases of clades indicate the number of families in the inferred ancestral genomes; the rectangles at the tips of clades indicate the number of families in the extant genomes within the clade. Square boxes at the tips indicate single genomes; rectangular boxes indicate multiple genomes.

This trend of genome streamlining has also been supported by reconstructions of the ancestral state of specific functional systems. For example, comparative analysis of the cell divisions machineries indicates that the common ancestor of extant archaea might have possessed all three varieties of the cell division systems found in modern forms, the FtsZ-based bacterial-type system, the ESCRT-III-based eukaryote-like system and a putative novel system that uses an archaeal actin-related protein (Makarova et al. 2010). The archaeal ancestor was also predicted to have possessed more ribosomal proteins than most extant lineages, suggesting there has even been streamlining of key cellular processes (Desmond et al. 2011).

Section 2. The relationship between Archaea and Eukaryotes

Archaea and the origin of the eukaryotic cell

Eukaryotic cells differ greatly from prokaryotic cells by a number of features such as: membrane-bound nucleus, numerous membrane-enclosed organelles (e.g., mitochondria, lysosomes, Golgi apparatus), cytoskeletal elements, and division by mitosis (Katz 2012). It has been proposed that the last eukaryotic common ancestor (LECA) already possessed a modern nucleus (Ben J Mans et al. 2004) and associated features, such as the nucleolus (Staub et al. 2004), capped and polyadenylated mRNA, introns (Collins and Penny 2005), nuclear pore complexes (Bapteste, Charlebois, et al. 2005; Neumann et al. 2010), chromatin (Iyer et al. 2007), and linear chromosomes and centromeres (Cavalier-Smith 2009). Additional features such the mitochondrion (Embley, van der Giezen, Horner, Dyal, and Foster 2003; Gabaldón and Huynen 2007), a complete membrance-trafficking system (Dacks et al. 2009; De Craene et al. 2012), a cytoskeleton based on actin and microtubules (Hammesfahr and Kollmar 2011), and a modern cell cycle (Eme et al. 2010) are also predicted to have already existed in LECA (Figure 19). How, when, and in what order these features originated during the path leading to the LECA remains one of the most important unanswered questions in Evolutionary Biology.

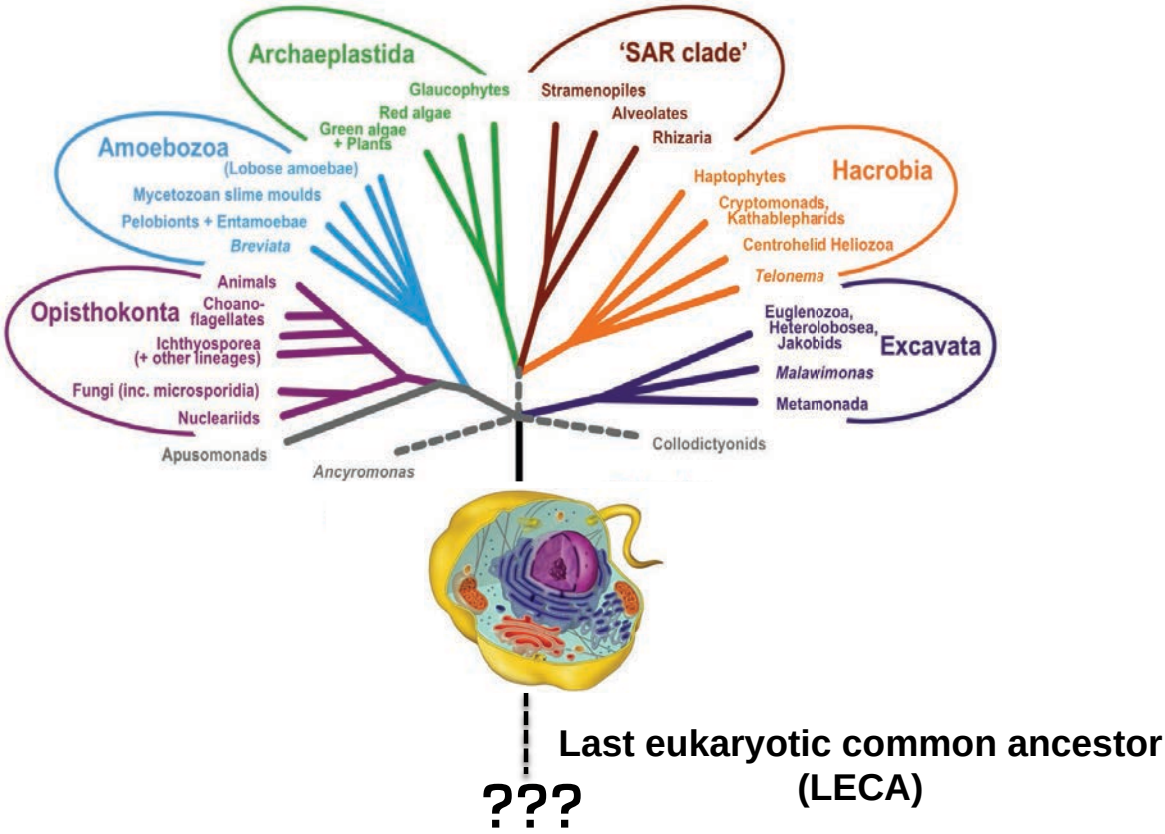


Figure 19. Schematic representation of the phylogeny of eukaryotes adapted from ABG Simpson 2013 unpublished. The eukaryotic cell at the base of the tree represents an already complex eukaryotic ancestor, with the dotted line leading to this cell representing the unknown history of the events that gave rise to the LECA.

Before the 1970's the origin of eukaryotes was thought of in two ways, either it was linked to the origin of life itself, or eukaryotes derived from prokaryotes (Sapp 2005). Several endosymbiotic theories were put forth in the early 1900's to explain the origin of the eukaryotic organelles (Wallin 1923; Wallin 1927). However none of these theories were seriously taken into consideration (Jan Sapp 1994). In 1966, Lynn Margulis reincarnated the endosymbiont theory in her article *The Origin of Mitosing Eukaryotic Cells* in which she postulated that a bacterial endosymbiont established itself inside a proto-eukaryotic host and became the mitochondrion (Sagan 1967). Later, endoysmbiotic theories for the origin of mitochondria and chloroplasts were widely accepted and supported not only by comparative cytology (Sagan 1967; Margulis 1970; Cavalier-Smith 1987) but also by molecular data (Gray and Doolittle 1982; Yang et al. 1985; Cavalier-Smith and Chao 1996). The absence of mitochondria in some eukaryotic lineages led to the hypothesis that they were early branching amitochondriate eukaryotes that originated prior to mitochondrial endosymbiosis, the "Archezoa" (Cavalier-Smith 1987). The Archezoa included the obligate intracellular

parasites Microsporidia and a number of parasitic protists including Entamoeba, Giardia, and Trichomonas (Cavalier-Smith 1987). However, later evidence demonstrated that all of these organisms possess homologs of known mitochondrial proteins targeted to internal organelles, such as the hydrogenosome or mitosome, that share common ancestry with mitochondria (van der Giezen et al. 2002; Embley, van der Giezen, Horner, Dyal, and Foster 2003; Embley, van der Giezen, Horner, Dyal, Bell, et al. 2003). This led to the notion that LECA already contained a mitrochondrion and that its absence in archezoan is due to loss (Martin and Koonin 2006; Lane and Martin 2010).

The eukaryotic cell is a chimera of bacterial, archaeal, and eukaryotic specific features. Eukaryotic information systems (i.e. transcription, translation, and DNA replication) resemble those found in the Archaea (Sarmiento et al. 2014), whereas eukaryotic biosynthetic and metabolism functions are generally assumed to be bacterial-like (Canback and Andersson 2002). This mixture of bacterial and archaeal features has led to several diverse hypotheses on the origin of eukaryotes. These mostly differ in regard to which bacterial and archaeal lineages contributed to the eukaryotic cell, how many partners were involved, and the nature of their contribution (i.e. symbiosis, engulfment, ancestry, fusion), but they can be grouped into to major categories, the "amitochondriate", and the "mitochondriate" scenarios (Embley and Martin 2006) (Figure 20). The amitochondriate models suggest that the LECA contained many eukaryotic features, such as the nucleus, the endoplasmic reticulum (ER), and the cytoskeleton, which evolved as the result of a primary fusion or symbiosis of specific archaeal and bacterial partners (Zillig 1991; Lake and Rivera 1994; Gupta and Golding 1996; López-García and Moreira 1999; Margulis et al. 2006). This cell then served as the host for the endosymbiontic event that gave rise to the mitochondria. In contrast, the mitochondriate models postulate that all of the eukaryotic cell characteristics, including the nucleus, cytoskeleton, ER, and the mitochondria, are the result of a single symbiotic event involving a bacterial and an archaeal cell (Searcy and Hixon 1991; Martin and Müller 1998). These hypotheses invoke a 2 Domains topology for the Tree of Life composed of two primary Domains, the Archaea and the Bacteria, and a secondary Domain, the Eukaryotes. However, these hypotheses differ in the mechanisms involved in the emergence of eukaryotes, the nature of the archaeal and bacterial lineages inferred to be at the origin, and the host of the symbiosis, bacterium versus archaeon (Embley and Martin 2006).

One of the first fusion scenarios was proposed by Zillig in 1991, which stated that eukaryotes originated from a physical fusion of two cells leading to a single cellular compartment and a single integrated genome (Zillig 1991). This process is different from the origin of the

mitochondrion, meaning that modern eukaryotes emerge from three cells, a fusion followed by endosymbiosis (Figure 20). Another hypothesis proposed by Lake and Rivera in 1994, known as the eocyte hypothesis, suggests that eukaryotes originated from the engulfment of an archeaon, specifically a crenarchaeota, by a bacterium (Lake and Rivera 1994) (Figure 20). Engulfment of a eoycte archaeon by a bacterial host (a gram-negative bacterium) was also proposed by Gupta and Golding in 1996 (Gupta and Golding 1996) (Figure 20). Alternative hypotheses propose that a member of the Euryarchaeota, rather than a crenarchaeote, was involved in a symbiosis with a bacterium. For example, some models suggest that a member of the genus Thermoplasma, a wall-less sulfur-respiring thermoacidophile, was the archaeal parent in the symbiosis that gave rise to the eukaryotic cell (Searcy and Hixon 1991; Margulis et al. 2006). The serial symbiosis theory of Lynn Margulis involves a Thermoplasma as a host to a Spirochete (which would give rise to eukaryotic flagella), with the mitochondrion is the resulting from a secondary symbiosis (Margulis et al. 2006) (Figure 20). However, in the model by Searcy, which also involves a Thermoplasma, there is only one symbiosis and that is with the bacterium that gave rise to the mitochondria (Searcy and Hixon 1991) (Figure 20). Two other models, the hydrogen hypothesis (Martin and Müller 1998) and the syntrophy hypothesis (López-García and Moreira 1999), predict that the archaeal parent was a methanogenic archaeon (Figure 20). The syntrophy hypothesis is a mitochondria late or amitochondriate model, as it suggests that two proteobacteria endosymbionts were involved. First a methanogenic archaeon and a deltaproteobacterium (an ancestral sulphate-reducing myxobacteria) came together by syntrophy (two organisms that are mutually depenedent on each other and as a result either exchange substantial genetic material or they fuse together). The resulting fusion cell then evolved into an amitochondrial eukaryote which phagocytosed the second endosymbiont, an alphaproteobacterial methanotroph, giving rise to the mitochondria (López-García and Moreira 1999). In contrast the hydrogen hypothesis, supposes there was only one endosymbiosis between an alphaproteobacterium and a strictly anaerobic autotrophic methanogenic archaeon that required hydrogen and carbon dioxide from the environment (Martin and Müller 1998). This alphaproteobacterial symbiont gave rise to the nuclear envelope, mitochondria, and other features early eukaryotes all at once (Figure 20).

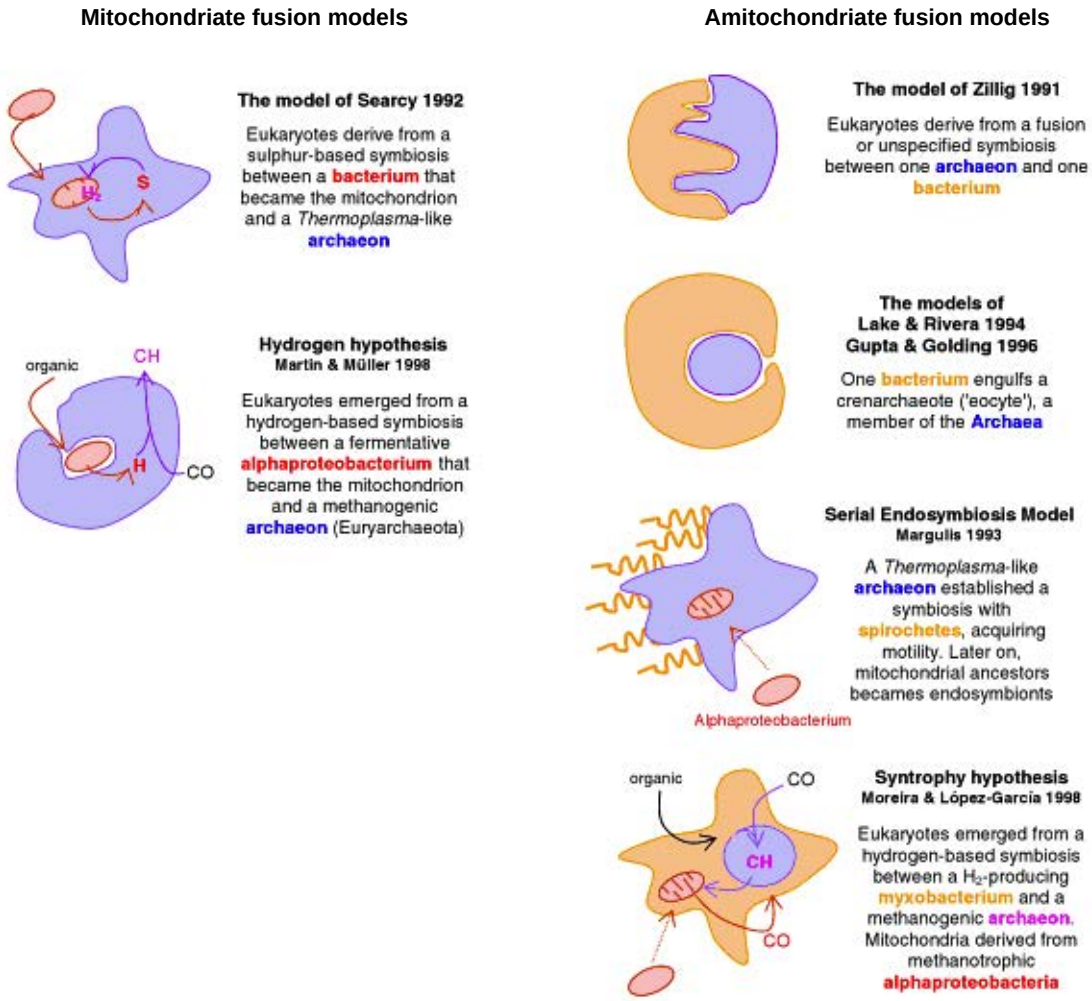


Figure 20. Schematic figure showing some examples of fusion hypotheses for the origin of eukaryotes adapted from the text book From Suns to Life: A Chronological Approach to the History of Life on Earth (Gargaud et al. 2007). In the left panel are the mitochondriate fusion models and on the right are the amitochondriate models.

Each of these hypotheses is attractive, but none are free of problems. First of all, there are no known examples of symbioses between bacteria and archaea in nature. Also, there is the problem of the lipids. The plasma membranes of Bacteria and Eukaryotes contain phospholipids in which fatty acids are covalently bound to *sn*-glycerol-3-phosphate via an ester linkage, whereas all the Archaea studied so far contain phospholipids with isoprenoid chains linked to *sn*-glycerol-1-phosphate via an ether bond (reviewed in (Lombard et al. 2012)). Therefore, all models invoking an archaeal host have to explain which process and what selective pressure led to the archaeal-to-bacterial membrane transition. This problem is rarely addressed by the different hypotheses that have been proposed to explain the emergence of eukaryotes from an archaeal host cell. One explanation is that there was progressive replacement of archaeal membranes by the phospholipids of the bacterial partner of the symbiosis (Martin and Müller 1998). However, it is not clear what kind of selective

pressure would invoke this replacement of the host archaeal membrane specifically because components of bacterial phospholipids are present in some archaea and vice versa, but the transformation from one membrane type to another has never been observed in nature (Lombard et al. 2012). Models that invoke a bacterial host cell (Lake and Rivera 1994; López-García and Moreira 1999) avoid the problem of the lipids because this would imply that the outer membranes of the host cell would have been bacterial, while the archaeal membrances were subsequently lost. Additionally, most of these hypotheses are not backed up by phylogenetic analyses because most universal rooted trees (Gogarten et al. 1989; Iwabe et al. 1989; Brown and Doolittle 1995; Lawson et al. 1996; Brown et al. 1997; Gribaldo and Cammarano 1998) have recovered Archaea and Eukaryotes as sister lineages.

Universal gene trees and origin of Eukaryotes

At the end of the 80's, two studies independently obtained the first rooted universal trees of life by using ancient paralogous gene couples. The first one used the catalytic and regulatory subunits of the V- and F-type ATPases (Gogarten et al. 1989), and the second one the translation elongation factors EF-Tu/1a and EF-G/2 (Iwabe et al. 1989). Both studies indicated that the first divergence from the last universal common ancestor separate a lineage leading to Bacteria from a lineage leading to the common ancestor of Archaea and Eukaryotes. These results were endorsed by Woese to propose his 3 Domain topology for the Tree of Life (Woese et al. 1990). Ancient couples of paralogous genes that are largely distributed and not transferred among domains are very rare. Nevertheless, a few of them were found and analyzed in the following years. Brown and Doolittle (1995) used tRNA synthetases (lle-tRS) versus paralogous Val- and Leu-tRS, Lawson et al. (1996) used an internal gene duplication in carbamoyl phosphate synthetase (CPS), Brown et al. (1997) used aminoacyl-tRNA synthetases Tyr-tRS versus paralogous Trp-tRS, and Gribaldo and Cammarano (1998) used the Signal Recognition Particle (SRP) 54kD protein and its paralogous SRP receptor SR-a. All of these analyses confirmed a root in the branch leading to the Bacteria and recovered the monophyly of the archaeal domain as sister to the Eukaryotes (Brown and Doolittle 1995; Lawson et al. 1996; Brown et al. 1997; Gribaldo and Cammarano 1998).

The debate over the monophyly of the Archaea started nearly 10 years before the archaea were officially classified as a separate Domain of life by Woese and colleagues (Woese et al. 1990). One of the most famous challenges was provided by Lake et al. 1984 based on ribosome morphology (Figure 21). Lake and colleagues broke up the Archaea (then

archaebacteria) by placing the Crenarchaeota, which they called eocytes, in a clade with eukaryotes, and all other archaebacteria in a clade with the eubacteria (Lake et al. 1984). Similarities found in RNA polymerase structures also supported a close relationship between the "eocytes" and the eukaryotes. By analyzing RNA polymerase structural differences, Wolfram Zillig and coworkers supported a division of the archaebacteria into two different groups -the methanogens/halophiles and the thermoacidophiles (eocytes), and showed that the RNA polymerases of the 'eocytes' more closely resemble the RNA polymerases of eukaryotes in terms of their complex component patterns and structure at the level of single components (Huet et al. 1983; Zillig et al. 1989; Zillig et al. 1991). Lake used this similarity in RNA polymerases to further support his eocyte hypothesis (Lake et al. 1984; Lake and Rivera 1994).

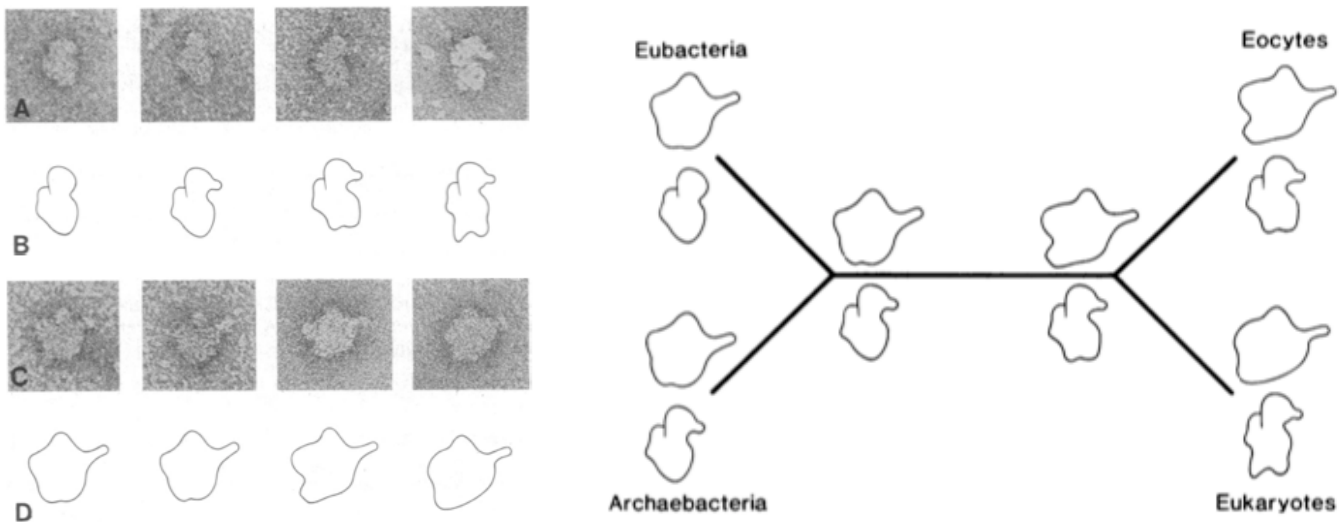


Figure 21. Ribosome structures as support for the eocyte hypothesis. Figures were taken from Lake et al 1984. 1) Elecotron micrographs of the ribosomes of representatives of eubacteria, archaebacteria, eocytes, and eukaryotes. The organisms included in the analysis (from left to right), Synechocystis (eubacteria); Halobacterium cutirubrum (archaebacteria); Thermoproteus tenax (eocyte); and Saccharomyces cerevisiae (eukaryote). The small subunits are shown in panel A and large subunits shown in panel C, with schematic diagrams of their profiles shown below the micrographs in panels B and D. 2) An unrooted dendrogram relating the steps in the evolution of taxa based on ribosome structure from the four proposed lineages. The relationship between the groups was based on parsimony.

Woese and colleagues heavily refuted the claims that the archaea were not monophyletic by pointing out that the use of ribosomal morphology as well as RNA polymerase complexity and structure was not solid evidence for a specific relationship between sulfur-dependent archaea and eukaryotes. They argued that the ribosomal features used by Lake and colleagues, such as the "lobes of various sizes", were of unknown molecular basis because they were in a region that had been characterized by others as lacking intrinsic ribosomal proteins (Woese and

Olsen 1985). Additionally, the lobe of "archaebacteria" was only characterized from one representative, *Halobacterium cutirubrum*, making the claimed phylogenetic distribution of shapes very biased. Moreover, as more ribosome structures were determined, it was discovered that the so-called unique structural characteristics of "eoyctes" are also present in archaebacteria, such as *Methanococcus vannielii* (Stöffler and Stöffler-Meilicke 1985; Stöffler-Meilicke et al. 1986) and *Halobacterium marismortui* (HARAUZ et al. 1987).

In 1987, Lake developed a new phylogenetic approach called evolutionary parsimony, a technique related to parsimony analysis which applies to four lines of descent (each able to contain multiple species) and for each of the possible unrooted trees a χ^2 value is calculated and the tree is retained if the associated χ^2 value is significant (Lake 1987). By performing a phylogenetic analysis of rRNA sequence using his evolutionary parsimony method, Lake recovered the eocyte tree, with the eocytes and eukaryotes as a sister group that he called Karyotes, and the Methanogens, halophiles, and members of the Bacteria forming a monophyletic group which he called the Parkaryotes (Lake 1988). Note that the root of this universal tree lies within the Archaea, with Bacteria emerging from them (Figure 22).

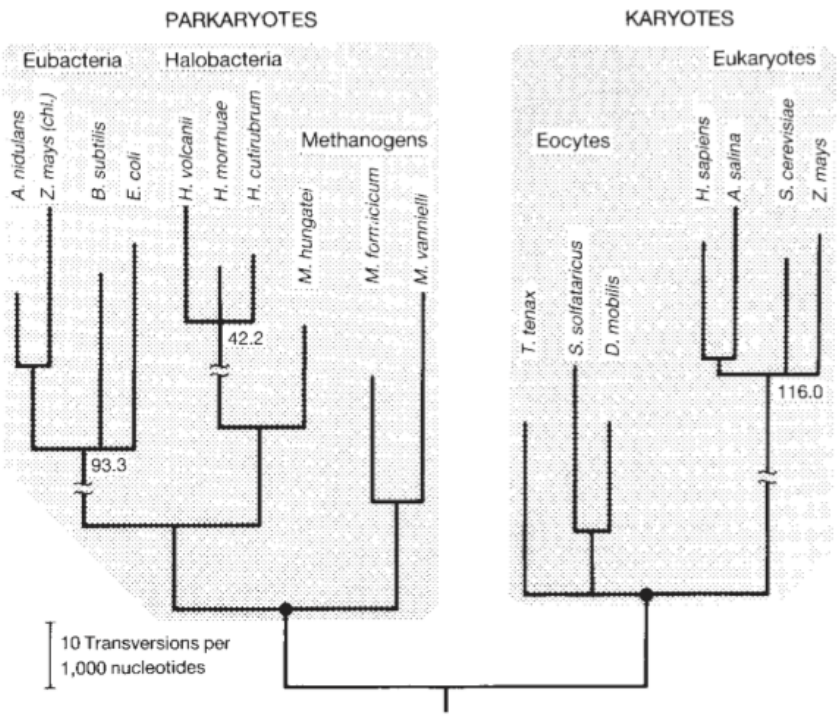


Figure 22. Rooted evolutionary tree based on rRNA sequences taken from Lake 1988. The tree was constructed using the evolutionary parsimony method and the root was selected by parsimony rooting. The branch lengths represent the number of transversion differences rather than substitutions. The parkaryotes contain the eubacteria, halobacteria, and methanogens (including Thermoplasma and Archaeoglobus), and the karyotes consist of the eocytes and the eukayotes.

However, the "eocyte" tree obtained by Lake based on rRNA sequences was challenged by phylogenetic analyses of SSU rRNA using Maximum parsimony and Neighbor Joining methods, which showed a classical 3Domain topology (Gouy and Li 1989). A few years later, Lake and Rivera used the identification of an 11-amino acid insertion within the elongation factor EF-Tu/1α that occurs in both the eocytes and the eukaryotes and is not present in Euryarchaeota (Figure 23) as another argument to support the eocyte hypothesis (Rivera and Lake 1992).

Taxon	Organism	Left primer	11–amino acid segment	4–amino acid segment	Right primer
Eukaryotes	Human	KNMITG TSQADCAVLIVAAGV	GEFEAGISKNG		QTREH
Eukaryotes	Tomato	KNMITG TSQADCAVLIIDSTT	GGFEAGISKDG		QTREH
Eukaryotes	Yeast	KNMITG TSQADCAILIIAGGV	GEFEAGISKDG		QTREH
Eocytes	*P. occu.*	KNMITG ASQADAAILVVSARK	GEFEAGMSAEG		QTREH
Eocytes	*D.muco.*	KNMITG ASQADAAILVVSARK	GEFEAGMSAEG		QTREH
Eocytes	*A.infe.*	KNMITG ASQADAAIIAVSAKK	GEFEAGMSEEG		QTREH
Eocytes	*Su.acid.*	KNMITG ASQADAAILVVSAKK	GEYEAGMSAEG		QTREH
Methanogens and relatives	*T.celer*	KNMITG ASQADAAVLVVAVTD		---GVMP	QTKEH
	Mc.van.	KNMITG ASQADAAVLVVNVDD		AKSGIQP	QTREH
Halobacteria	*H.maris.*	KNMITG ASQADNAVLVVAADD		---GVQP	QTQEH
Eubacteria	*Th.mar.*	KNMITG AAQMDGAILVVAATD		---GPMP	QTREH
Eubacteria	*S.plat.*	KNMITG AAQMDGAILVVSAAD		---GPMP	QTREH
Eubacteria	Mitoch.	KNMITG AAQMDGAIIVVAATD		---GQMP	QTREH

Figure 23. Figure showing the 11 amino acid region shared between eocytes and eukaryotes taken from Lake et al 1992. The figure shows a comparison of the elongation factors EF-Tu/1α near the KNMTG region (a conserved region conserved responsible for terminating the helix).

Some subsequent phylogenetic analyses supported Lake's eocyte hypothesis, but these were based almost solely on phylogenetic analyses of elongation factor genes. For example, reanalysis of the EF-2/G and EF-Tu/1a data sets by Baldauf et al. (1996), albeit obtaining a root between the bacteria and the archaea/eukaryotes, favored the placement of the eukaryotes within the archaea, sister to the Crenarchaeota (Baldauf et al. 1996). However, it was shown that depending on the phylogenetic methods, the amino acid positions, and taxonomic sampling used, unrooted phylogenies based upon the elongation factor proteins yielded conflicting results (Cammarano et al. 1999). Cammarano et al. constructed an EF-2/G gene tree using new sequences from the archaea and various methods (distance-matrix, maximum likelihood, and parsimony) and obtained strong support for the monophyly of the three domains, with the bifurcation of the kingdoms Crenarchaeota and Euryarchaeota (Cammarano et al. 1992). Creti et al. also showed that the archaea are strongly monophyletic for EF-2/G genes but only weakly for EF-Tu/1a genes (Creti et al. 1994). At the same time, additional support for the eocyte tree came from a phylogenetic analysis of LSU and SSU rRNAs and of concatenated nucelotide sequences of the two largest subunits of RNA polymerase (Tourasse and Gouy 1999). By using phylogenetic methods that take into account

the variability of the substitution rate among sites (Yang 1995; Tourasse and Gouy 1997), they recovered support for the eocyte-eukaryote relationship for both datasets (Figure 24).

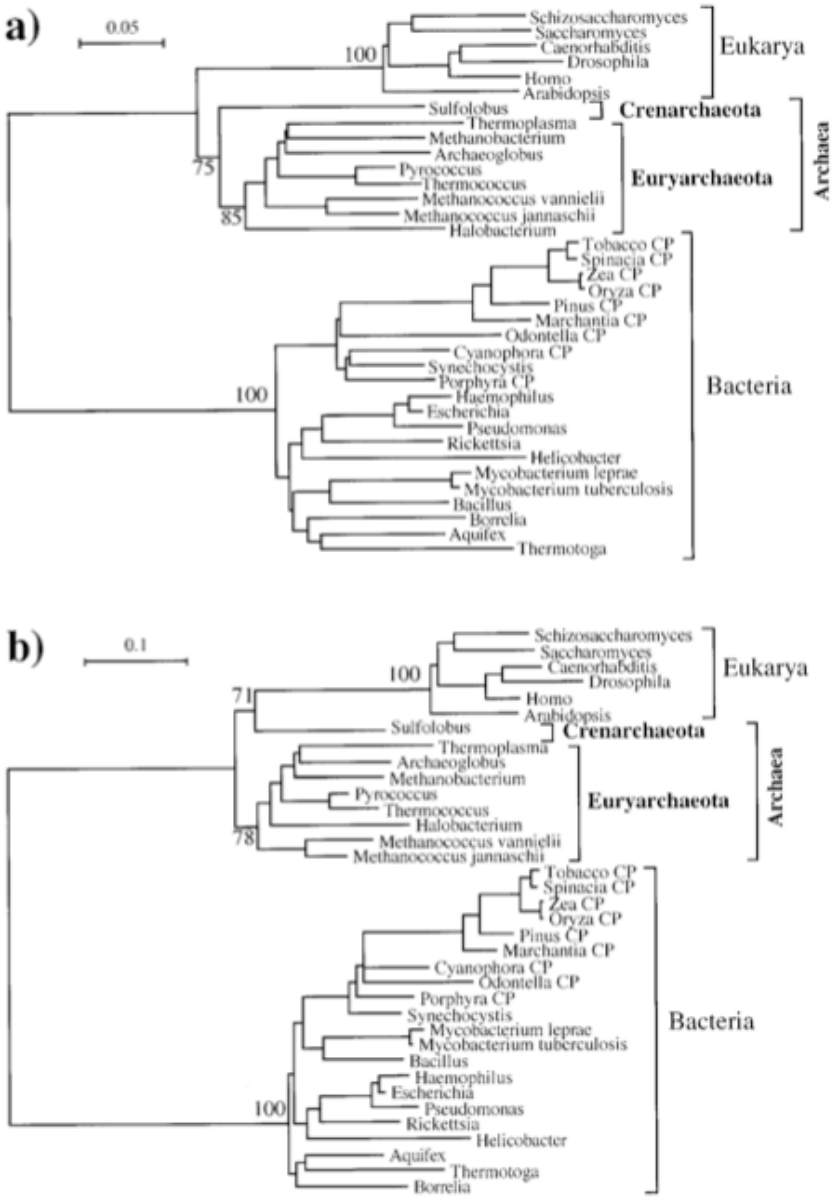


Figure 24. Unrooted universal phylogeny deduced from concatenation of the two largest subunits of RNA polymerase taken from Tourasse and Gouy 1999. NJ trees, based on 717 codons were evaluated by the maximum likelihood method under Kimura's substitution model (Kimura 1980), A) without and B) with a gamma distribution for rate variation among sites. By taking into account rate variation across sites (B), the authors recovered the eocyte tree.

One of the earliest attempts to analyze the relationship between the primary domains of life using a large number of genes came from Brown and Doolittle (Brown and Doolittle 1997). They performed phylogenetic analyses on 56 proteins involved in various cellular systems (e.g. informational, metabolism, amino acid biosynthesis, cofactors, respiration), considering

only gene trees with two or more species from each domain (Brown and Doolittle 1997). Of the 56 genes analyzed, only 21 could be used to test the relationship between archaea and eukaryotes (i.e. members from both archaeal domains were present and statistical support for the archaea/eukaryotic grouping exisited). More than half of these 21 trees depicted the archaea as monophyletic, only three showed a branching order consistent with the eocyte hypothesis (Ef-Tu.1/1a, RP L11, and RP S11), and the remaining eight trees had either methanogens or halophiles as the closest relative to eukaryotes (Brown and Doolittle 1997). The authors noted that phylogenetic evidence from both duplicated and unrooted gene trees is generally consistent with the 3 Domains topology tree in which archaea and eukaryotes are sister lineages (Brown and Doolittle 1997). However, they also stated that most of their single gene trees were likely to be unreliable due to horizontal gene transfer events between the primary domains of life, unrecognized paralogy, or unequal evolutionary rates (Roger and Brown 1995; Brown and Doolittle 1997; Feng et al. 1997). In addition, they emphasized that in order to fully understand the relationship between the three domains, an evolutionary paradigm had to be created where genomes, biochemistry, and organisms are all considered in together (Brown and Doolittle 1997).

Phylogenomic approaches

Starting from the beginning of the 21st century several large-scale phylogenetic analyses were performed that investigated the relationships between Archaea and Eukaryotes either directly or indirectly. These studies used different approaches for assembly of the datasets and different methods of analysis and supported either 3D or 2D scenarios (Figure 25). However, it was remarked that there is an overlap in the datasets used (Gribaldo et al. 2010). This overlap is due to the use of universal markers because of the need for a bacterial outgroup, which significantly restricts the number of markers that can be used to investigate the relationship between archaea and eukaryotes. The conflicting results obtained, despite the use of overlapping markers, may be explained by the various accuracy level of different methodologies, different approaches used to assemble the datasets, taxonomic sampling, and also in the interpretation of the results (Gribaldo et al. 2010). Phylogenomic approaches taken to resolve deep evolutionary relationships have to be performed with great care as they are prone many biases and artifacts (Delsuc et al. 2005; Gribaldo et al. 2010). Some of these issue are inherent to biological sequence data such as compositional bias and varying evolutionary rates, but the outcome of phylogenomic studies are also strongly influenced by the quality of the data analyzed, which heavily depends on selection of markers, taxonomic sampling,

multiple alignment and site selection, phylogenetic methods, and evolutionary models (Delsuc et al. 2005). The lack of congruence among different studies has prompted the idea of a 'phylogenomic impasse', were we might be trapped in an endless cycle of reanalyzing very similar, limited sets of universal genes without ever reaching a consensus (Gribaldo et al. 2010). Despite this warning, a large number of similar analyses have been published since 2010, mostly using supermatrices of universal markers, although with an effort to use additional sophisticated evolutionary models.

Publication	Number of markers used to infer relationships among the three domains	Taxonomic sampling	Number of amino acid positions used	Method	Model supported
Harris et al. **2003** *Genome Res.*	50	25 bacteria 1 crenarchaeote 7 euryarchaeotes 3 eukaryotes	variable, depending on the gene	Single-gene analysis Maximum likelihood Maximum parsimony Distance	**3D**
Ciccarelli et al. **2006** *Science*	31	150 bacteria 4 crenarchaeotes 14 euryarchaeotes 23 eukaryotes	8,09	Concatenation Maximum likelihood	**3D**
Yutin et al. **2008** *Mol.Biol.Evol*	136	variable, depending on the gene	variable, depending on the gene	Single-gene analysis Maximum likelihood	**3D**
Rivera & Lake **2004** *Nature*	complete genome of *A. fulgidus*	2 bacteria 1 crenarchaeote 2 euryarchaeotes 2 eukaryotes	Not applicable	Genome content (conditioned reconstruction)	**2D** (the Eukarya sister of the crenarchaeota)
Pisani et al. **2007** *Mol.Biol.Evol*	Data not available	97 bacteria 4 crenarchaeotes 17 euryarchaeotes 17 eukaryotes	variable depending on gene	Supertree	**2D** (the Eukarya sister of the Thermoplasmatales)
Cox et al. **2008** *PNAS*	45	10 bacteria 3 crenarchaeotes 11 euryarchaeotes 16 eukaryotes	5,521	Concatenation: Bayesian Maximum likelihood Maximum parsimony	**2D** (the Eukarya sister of the crenarchaeota)
Foster et al. **2009** *Phil.Trans.R.Soc.B*	41	8 bacteria 8 crenarchaeotes 2 thaumarchaeotes 6 euryarchaeotes 11 eukaryotes	5,222	Concatenation: Bayesian Maximum likelihood Maximum parsimony	**2D** (the Eukarya sister of a group comprising the crenarchaeota and the Thaumarchaeota)

Figure 25. Large-scale phylogenomic analyses performed from 2003 to 2010. The table was adapted from Gribaldo et al. 2010. Analyses supporting the 3D scenario are highlighted in blue and those supporting the 2D in purple. The two analyses highlighted in the black box were both performed by the group of Martin Embley using new evolutionary models and are further discussed in the text below.

New and improved evolutionary models

The first analysis that implemented more realistic models aimed at capturing sequence evolution more accurately was performed by Cox et al. 2007 (Figure 25 and 26). In a Bayesian framework, they used the node-discrete composition heterogeneity (NDCH) model (Foster 2004), which allows composition to change in different lineages over time and the CAT mixture model (Lartillot and Philippe 2004; Lartillot et al. 2009) that accommodates among-site compositional heterogeneity using multiple substitution classes, each with its own composition profile. However, due to computational burden, the NDCH model could only be applied to the universal protein dataset after recoding according to Dayhoff groups (Hrdy et al. 2004), which defines six groups of amino acids corresponding to the PAM matrix: 1,

cysteine; 2, alanine, serine, threonine, pro- line, glycine; 3, asparagine, aspartic acid, glutamic acid, glutamine; 4, histidine, arginine, lysine; 5, meth- ionine, isoleucine, leucine, valine; 6, phenylalanine, tyrosine, tryptophan. They also analyzed the data using standard one-matrix models in both a Maximum Likelihood and Bayesian framework. Both Bayesian and Maximum Likelihood analyses of the universal protein dataset recovered a 2D topology in which the Eukaryotes emerge from within the Archaea as sister to the Crenarchaeota, consistent with the eocyte hypothesis (Figure 26) (Cox, Foster, et al. 2008).

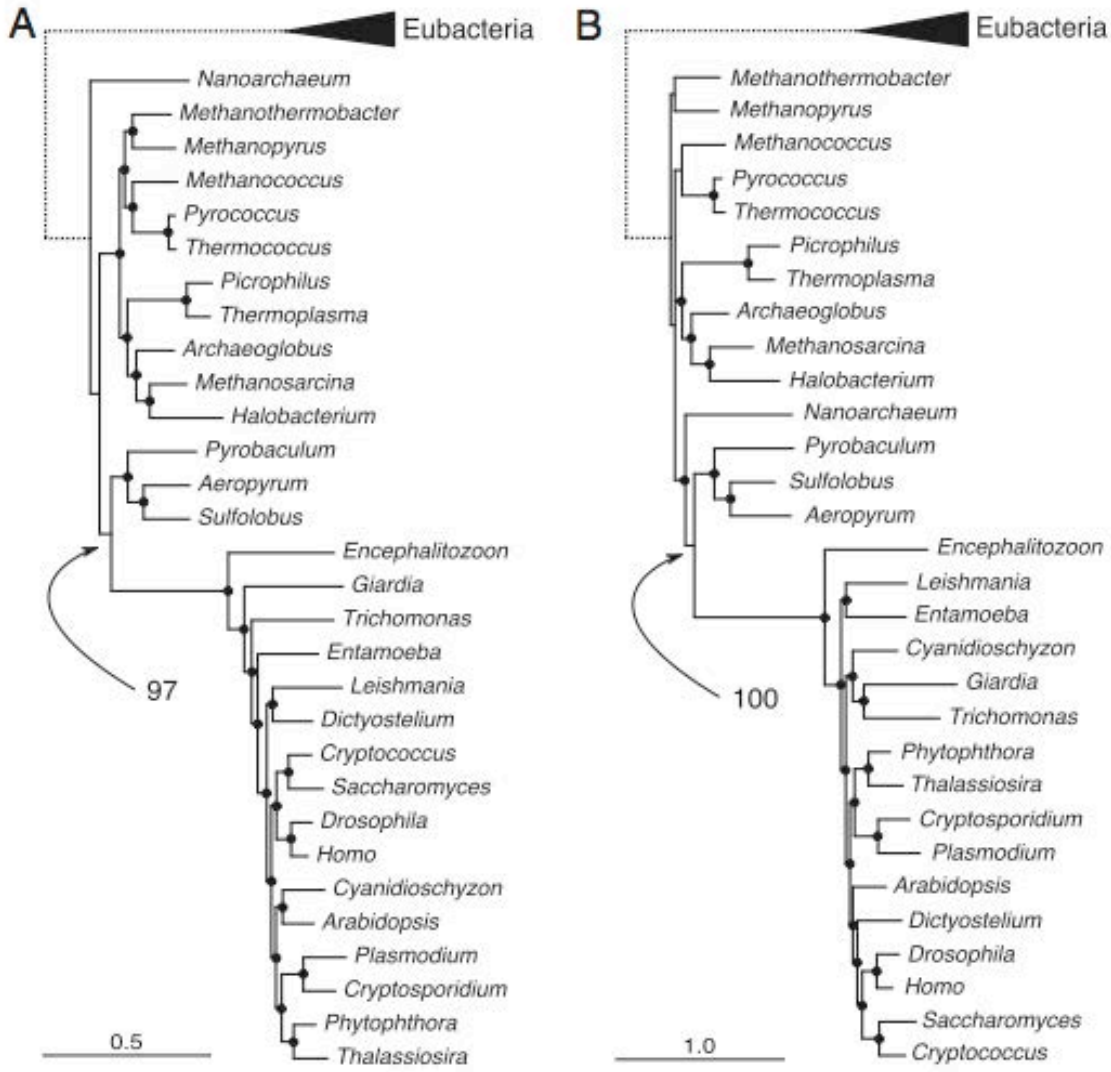


Figure 26. Phylogenetic analysis of 45 concatenated proteins universal proteins taken from Cox et al 2008. Scale bars indicate substitutions per site. The dotted branches leading to eubacteria are arbitrary lengths. Nodes highlighted with dots were supported by >95% PP. The 2 values indicate support (PP) for the eocyte hypothesis. (*A*) NDCH+Γ4 model with Dayhoff-recoded data; (*B*) CAT+Γ4 model with standard amino acid coded data

Using the same methods, a subsequent analysis was done by the same authors but with the addition of representatives from newly sequenced archaeal taxa. Fox et al. 2009 reanalyzed

the Cox et al. 2007 dataset using two available genomes from the phylum Thaumarchaeota as well as five additional crenarchaeotes. For this analysis 41 of the universal proteins used by Cox et al. 2008 were analyzed, after excluding previously unidentified paralogs (Foster, Cox, and Embley 2009a). This updated analysis recovered a sister relationship between the Eukaryotes and a cluster comprising Crenarchaeota and Thaumarchaeota (Figure 27) (Foster, Cox, and Embley 2009a).

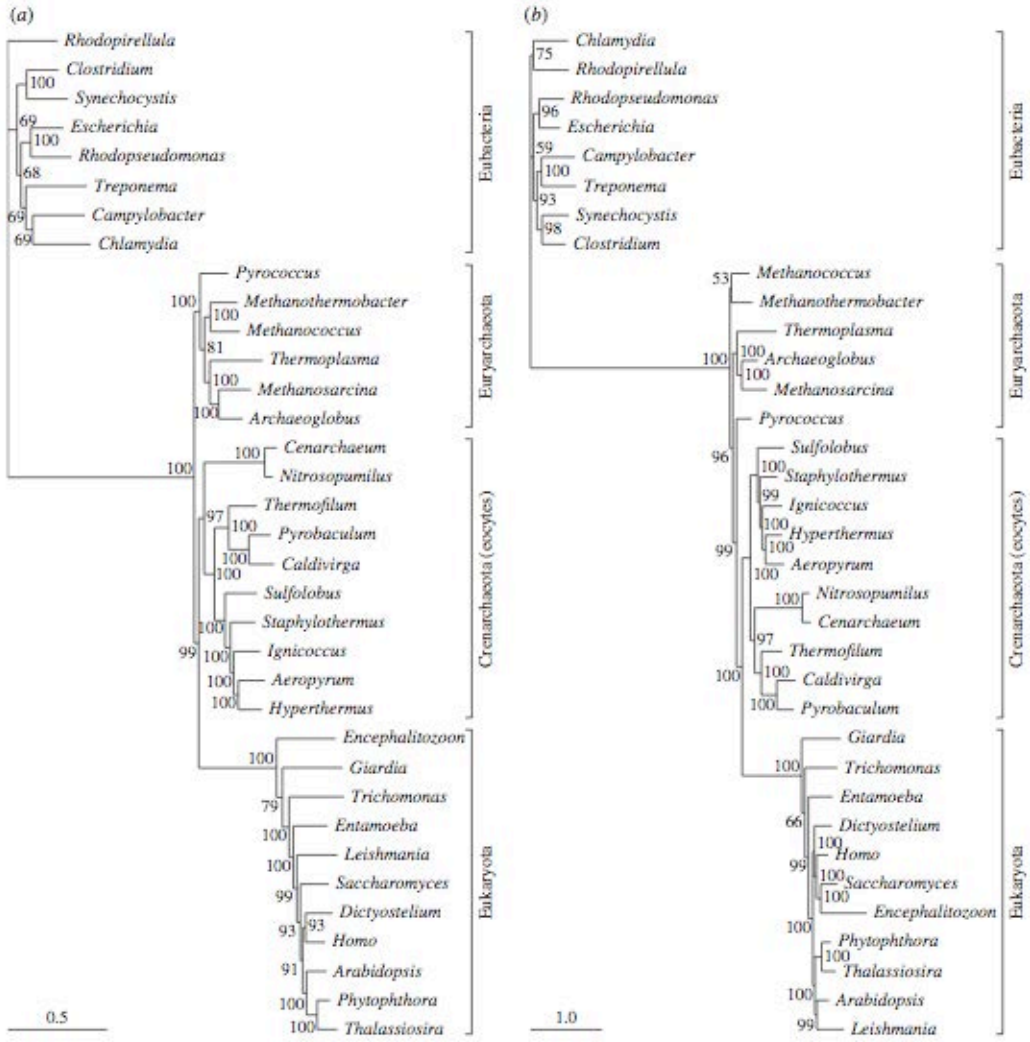


Figure 27. Bayesian phylogenetic analyses of concatenated amino acid data. The analysis in panel (a) used Dayhoff-recoded data with a GTR +Γ+ NDCH tree-heterogeneous substitution model in P4. This is the analysis summarized in table 3, row I. The analysis shown in panel (b) used standard amino acid-coded data with a CAT-Poisson substitution model in Phylobayes.

Although these studies used evolutionary models that were shown to fit better the data, it is noteworthy that the authors also obtained a 2D topology with less sophisticated homogeneous models, suggesting that their results may be linked more to the data set used (i.e. the genes and the taxonomic sampling) than to the new evolutionary models (Gribaldo et al. 2010).

Eukaryotes and the TACK superphylum

Over the past few years a large number of analyses have appeared in the literature recovering the 2 Domains tree. Most of these analyses have been performed on nearly the same set of universal proteins with a limited number of taxa and a majority of them have been done by the same group (Figure 28).

Publication	Number of markers used to infer relationships among the three domains	Taxonomic sampling	Number of amino acid positions used	Method	Programs and Models	Model supported
Cox et al. **2008** *PNAS*	45 universal	10 bacteria 3 crenarchaeotes 11 euryarchaeotes 16 eukaryotes	5,521	Concatenation: Bayesian Maximum likelihood Maximum parsimony Neighbor Joining	RAxML:WAG+G MrBayes: WAG+G P4: GTR+G+NDCH recoded Dayhoff6 * Phylobayes: CAT+G + recoded Dayhoff6	**2D** (Eukarya sister of the crenarchaeota)
Foster et al. **2009** *Phil.Trans.R.Soc.B*	41 universal	8 bacteria 8 crenarchaeotes 2 thaumarchaeotes 6 euryarchaeotes 11 eukaryotes	5,222	Concatenation: Bayesian Maximum likelihood Maximum parsimony Neighbor Joining	RAxML: GTR+G MrBayes: WAG+G P4: GTR+G & GTR+G+NDCH recoded Dayhoff6 * Phylobayes: CAT+G + recoded Dayhoff6	**2D** (Eukarya sister of a group comprising the crenarchaeota and the Thaumarchaeota)
Guy and Ettema **2011** *Trends in Micro.*	26 universal	10 bacteria 14 crenarchaeotes 2 thaumarchaotes 1 aigarchaeote 1 korarchaeote 7 euryarchaeotes 7 eukaryotes	Not specified	Concatenation: Bayesian & Maximum likelihood	RaxML: LG+G+I Phylobayes: LG+G & CAT+G nh-phylobayes: CATBP	**2D** (Eukarya originating or sister group of the TACK superphyla)
Williams et al. **2012** *Proc.R.Soc.B*	29 universal 63 AE proteins	8 bacteria 8 crenarchaeotes 2 thaumarchaeotes 1 aigarchaeote 1 korarchaeote 6 euryarchaeotes 10 eukaryotes	U: 3983 AE: 8438	Concatenation: Bayesian	Phylobayes: LG+G & CAT+G	**2D** (Eukarya originating or sister group of the TACK superphyla)
Lasek-Nesselquist & Gogarten **2013** *Mol Phylogenet Evol.*	85 ribosomal proteins	different depending on the dataset	abe1: 13,432 abe2: 13941 abe3: 17876	Concatenation: Bayesian & Maximum likelihood	RAxML: LGF+G PhyML-4X: LG4M & LG4X PhyML-Structure: UL3 PhyloBayes CAT+G	**2D** (Eukarya sister to a Kor/Thaum or Cren/Thaum/Kor clade)
Rochette et al. **2013** *MBE*	28 universal 121 AE proteins	variable, depending on the gene	variable, depending on the gene	Single-gene analysis: Bayesian & Maximum likelihood	RAxML: CAT LG+G Phylobayes LG+G PhyML- structure: UL3+G	Deep **2D or** close very close **3D**
Williams et al. **2014** *GBE*	20 universal 29 universal	8 bacteria 8 crenarchaeotes 2 thaumarchaeotes 1 aigarchaeote 1 korarchaeote 13 euryarchaeotes 10 eukaryotes	not specified	Concatenation: Bayesian	Phylobayes: LG+G & CAT+GTR	**2D** (Eukarya originating or sister group of the TACK superphyla)

Figure 28. Methods and datasets used in recent large-scale phylogenomic analyses supporting the 2D topology. Rows highlighted in purple were performed by the group of Martin Embley.

In 2011, Guy and Ettema re-evaluated the phylogenetic position of eukaryotes using 26 universal proteins defined as the intersection between the datasets used in the phylogenomic studies of Cox et al. 2007 and Ciccarelli et al. 2006. The phylogenies were constructed using the LG model (Le and Gascuel 2008) in the Maximum Likelihood and Bayesian framework, and Bayesian analysis were also performed using the models CAT (Lartillot and Philippe 2004) and CATBP (a model that accounts for variations of the evolutionary process both along the sequence and across lineages) (Blanquart and Lartillot 2008). All methods supported a strong affiliation of the Eukaryotes with the Thaumarchaeota, 'Aigarchaeota', Crenarchaeota, and Korarchaeota (Guy and Ettema 2011). Based on this affiliation the authors proposed the existence of a new superphylum of archaea composed of the Thaumarchaeota, 'Aigarchaeota', Crenarchaeota, and Thaumarchaeota, collectively called the TACK (Guy and Ettema 2011).

This relationship was further highlighted by a specific enrichment of eukaryotic-like features within the TACK superphylum (Figure 29).

Figure 29. Figure showing the evolutionary links between TACK and Eukaryotes taken from Guy and Ettema 2011. The schematic phylogenies shown in a-d represent the results obtained from the phylogenetic analyses of 26 universal proteins using different phylogenetic methods (see Figure 25). The bacterial out-group is not shown. Schematic phylogenies (a-b) were rooted based on the root obtained with the bacterial out-group and schematic phylogenies (c-d) were rooted on the branch leading to euryarchaoeta. Panel d shows the phylogenetic distribution of orthologs of eukaryotic core genes based on those that have been reported in the literature. The rooted phylogenetic tree of the archaea was based on the 26 universal proteins with the eukaryotes omitted. The tree was inferred with Phylobayes under the LG model with a continuous gamma distribution.

However, these eukaryotic-like characters are not uniquely present in the TACK, and are also heterogeneously distributed among the TACK lineages. For example, the small RPB8 subunit of the eukaryotic RNA polymerase (RpoG), is present in Crenarchaeota and Korarchaeota but is absent in Thaumarchaeota and 'Aigarchaeota' (Koonin et al. 2007). Components of a eukaryotic-type ubiquitin (Ubl) modifier system are present in 'Aigarchaeota' (Nunoura et al.

2011) but are absent in all other members of the TACK sequenced so far, and the archaeal actin ortholog (Makarova et al 2010, Yutin et al 2009) is only present in the crenarchaeal order Thermoproteales and in '*Ca*. Korarchaeum cryptofilum' (Korarchaeota). The authors reconcile this patchy distribution by suggesting that a complex ancestral TACK lineage from which eukaryotes emerged existed only transiently in time, that a more complex archaeal lineage remains to be discovered, or that the extant TACK lineages have undergone differential reductive evolution (Guy and Ettema 2011).

In 2012, a third analysis was performed by the group of Martin Embley, with a further increased taxonomic sampling including Korarchaeota and 'Aigarchaeota' (Williams et al. 2012). The dataset for Williams et al. included 29 universal proteins, i.e. the 41 used in Foster et al. with the exclusion of 12 proteins that were removed either because of paralogy, potential HGT, or absence in some of the newly added genomes (Williams et al. 2012). It should be noted that each subsequent analyses of these authors resulted in the reduction of the number of universal markers and amino acid positions used for phylogenetic reconstruction (Figure 28). Williams et al. also analyzed 64 genes conserved in their sample of archaea and eukaryotes to investigate the in-group relationship between the eukaryotes and a specific archaeal lineage (Williams et al. 2012). To analyze the Archaea/Eukaryote phylogeny they placed the root in the branch leading to the Euryarchaeota as indicated by the universal tree, albeit not exactly in the same place (Figure 30). The universal and archaea/eukaryote datasets were analyzed under the Bayesian framework using the CAT model (Lartillot and Philippe 2004) and in both the Maximum Likelihood and Bayesian framework using the homogeneous substitution model LG (Le and Gascuel 2008). Altogether, the results of these analyses support either a sister relationship or origin from within the TACK superphylum (Figure 30) (Williams et al. 2012).

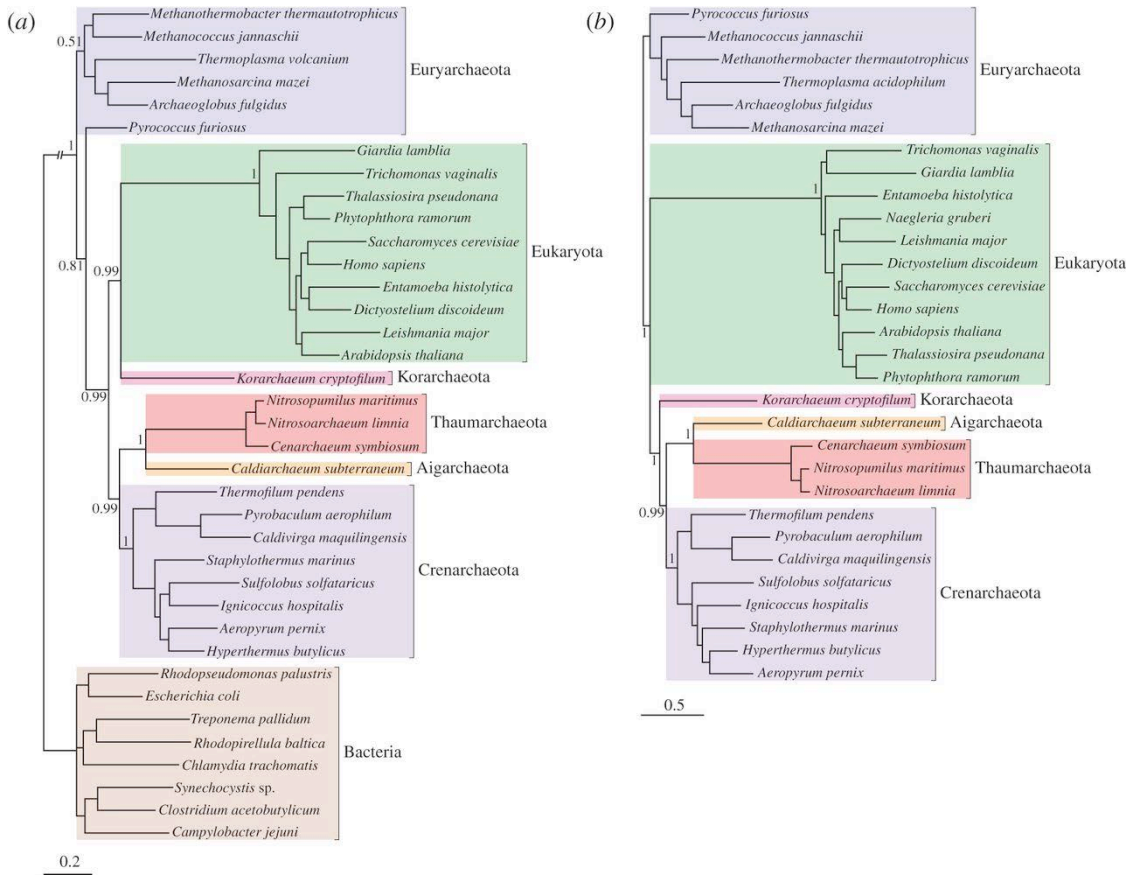


Figure 30. Bayesian phylogenies taken from Williams et al 2012 (a) Bayesian phylogeny inferred from 29 concatenated proteins using conserved between Bacteria, Archaea and eukaryotes (CAT+G4). The eukaryotes emerge as the sister group of Korarchaeum, within the TACK superphylum. (b) Bayesian phylogeny inferred from 63 concatenated proteins shared between Archaea and eukaryotes with the root placed on the branch leading to the Euryarchaeota (CAT+G4). The tree shows the eukaryotes emerging as the sister group to the TACK superphylum, including Korarchaeum. Branch lengths are proportional to substitutions per site, except the truncated bacterial branch in (a).

In 2013, Lasek-Nesselquist and Gogarten performed an analysis to test the effects of model choice on the ribosomal tree of life (Lasek-Nesselquist and Gogarten 2013). To do this they tested several models of varying sophistication on three different datasets and employed different strategies to remove compositional heterogeneity in order to examine their effects on the topological outcome. They concatenated 86 ribosomal proteins and analyzed three datasets with different taxonomic samplings and with different amounts of missing data. The datasets were analyzed in both a Maximum likelihood and Bayesian framework using different models of evolution, the CAT model (Lartillot and Philippe 2004) (Figure 31), two mixture models LG4M and LG4X implemented in PhyML-4X that allow changes in the rate matrix across sites but not across different lineages (Le et al. 2012), UL3 from PhyML-structure (Le et al. 2008), a mixture model that employs three different substitution matrices based on solvent accessibility of residues and secondary structure, as well as the LG model (Le and Gascuel 2008) with an estimation of amino acid frequencies from the data (Figure 31).

From this study they found that different models displayed varying proficiencies at resolving different areas of the tree, but all of the models recovered topologies where Eukarya emerge from within Archaea as sister to Korarchaeota/Thaumarchaeota (KT) or Crenarchaeota/KT clade under all or at least one of the strategies employed (Lasek-Nesselquist and Gogarten 2013). Although 'Aigarchaeota' was not included in this study, their results are compatible with a TACK-Eukaryote relationship (Figure 31).

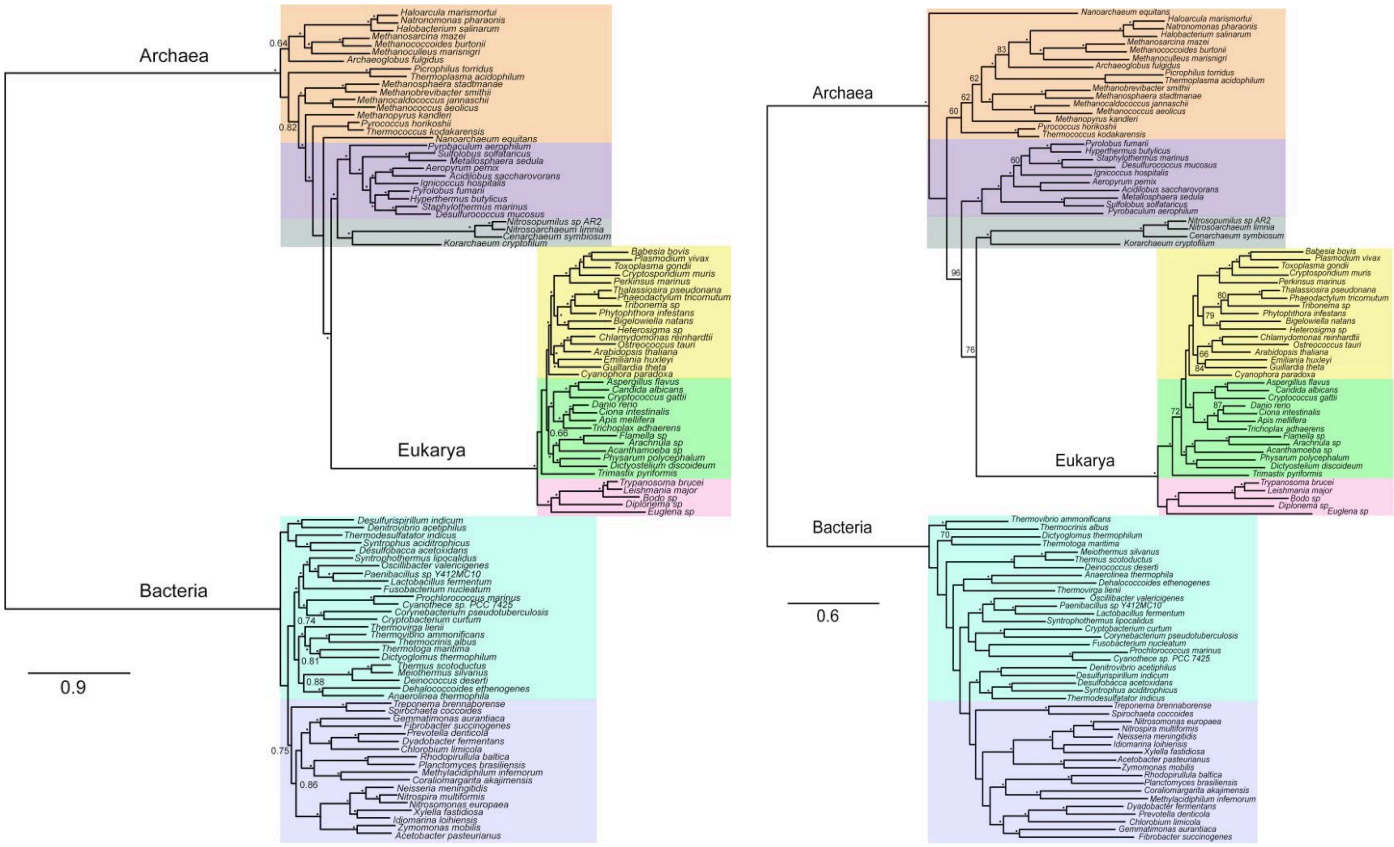


Figure 31. Phylogenetic trees based on the aeb 1 datset (86 concatenated ribosomal proteins- 13432 aa positions) taken from Lasek- Nesselquist and Gogarten 2013 (a) Bayesian phylogeny inferred by Phylobayes under the CAT model. Dots indicate posterior probabilities of 0.9–1.0. Posterior probabilities less than 0.60 are not shown. The tree recovers the Eukaryotes as sister to Thaumarchaeota/Crenarchaeota/Korarchaeota (b) Maximum Likelihood phylogeny inferred by RAxML under the LGF model. Dots indicate bootstrap support values of 90–100%. Bootstrap values less than 60% are not shown. This tree recovers a sister relationship between Thaumarchaeota/Korarchaeota and Eukaryotes.

Another recent analysis by Rochette et al. 2014 focused on dissecting the origins of eukaryotic genes in much more detail than previous studies. In this study the authors distinguished genes whose phylogeny supports a relationship between eukaryotes and a particular prokaryotic taxonomic group, genes that have been transferred among prokaryotes, and genes that do not contain sufficient phylogenetic signal (Rochette et al. 2014). To do so, they used clusters of homologs provided by the HOGENOM protein family database (Penel et al. 2008) to determine proteins families that date back to the LECA (Last Eukaryotic Common

Ancestor) which have homologs in prokaryotic genomes. These protein families or clades were determined by analysis of individual Maximum Likelihood trees using extended topological criteria, which they term configurations (Rochette et al. 2014). They identified 554 LECA-traceable protein clades with prokaryotic homologs, and of these one-third of them appeared archaea-related and two-thirds appeared bacteria-related. First, they used 28 universal makers to test the monophyly of the three domains. From this they found that monophyly of the bacteria was strongly supported in all but one tree, whereas the monophyly of the archaea was only recovered in four trees. However, they found that for many LECA clades the 3-Domain topology and the best paraphyletic-Archaea topology were equivalent. The authors also analyzed the relationship between archaea and eukaryotes using all of the archaea-related LECA clades (121 genes). They did not recover a specific branching order for the archaeal phyla or a particular position of eukaryotes relative to them. Therefore, they conclude that their results are compatible with the view that Eukaryotes branch deeply within the Archaea or close to the root (Rochette et al. 2014).

The latest large-scale phylogenetic analysis was performed in response to a universal phylogeny published in Nature by Rinke et al 2013. These authors recovered a 3 Domains universal tree of life by including many uncultured lineages that they obtained from a large single-cell sequencing project (Rinke et al. 2013). In addition, they obtained a monophyletic clade comprising of all nano-sized archaea, which they propose to represent a distinct phylum called the DPANN (Rinke et al. 2013). Subsequent reanalysis of the dataset used in Rinke et al. 2013 showed that they did not use the best-fit model for the data and they included mitochondrial and plastid sequences in the eukaryotic dataset (Williams and Embley 2014). Reanalysis of the dataset after removal of problematic genes and the use of the better-fitting CAT+GTR+Γ4 model resulted in the recovery the eocyte tree (TACK/eukaryote relationship) rather than a 3D topology (Williams and Embley 2014). The same results were obtained when the tree was reconstructed using the 29 universal proteins from Williams et al. 2012. The monophyly of the DPANN was also recovered with high support, but the position of the clade as a whole within the archaea was not resolved (Williams and Embley 2014).

Last year a review appeared in Nature by the group of Marin Embley (Williams et al. 2013) which largely popoularized the notion of 2D topology of the tree of life and a chimeric origin for Eukaryotes. At about the same time, Joran Martijn and This Ettema proposed a new chimeric hypothesis for the origin of eukaryotes (Figure 32). The phagocytosing archaeon

theory (PhAT), which poses that eukaryotes evolved from an ancestral archaeal lineage belonging to the recently proposed 'TACK superphylum' (Martijn and Ettema 2013). In the PhAT model, a cellular fusion between a TACK archaeon and an alphaproteobacterium resulted in a mitochondrion containing but nucleus-lacking cell, similar to the scenario recently proposed by the group of Eugene Koonin (Yutin et al. 2009).

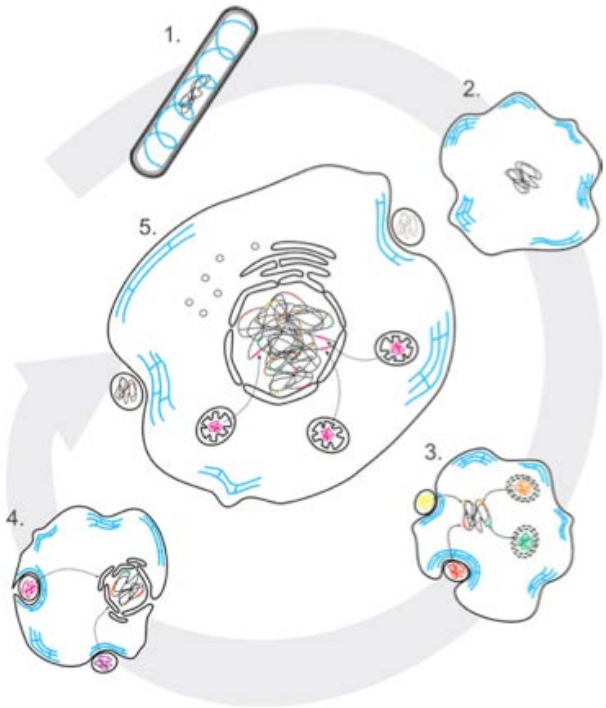


Figure 32. Schematic step-wise overview of the crucial steps of the proposed PhAT taken from Martijn and Ettema 2013. (1) An archaeon probably belonging to the recently proposed 'TACK superphylum' contains an actin-based cytoskeleton (blue). (2) After losing its cell wall, the archaeon evolves a more flexible actin-based cytoskeleton, which supported the formation of cellular protrusions. (3) The archaeon's cytoskeleton matures into a primitive phagocytotic machinery and the digestion of other prokaryotic cells exposes the archaeon to increasing amounts of 'foreign' DNA destabilizing the archaeal host genome. (4) A protective membrane boundary is formed to protect the genetic integrity of the host cell genome via invagination events, giving rise to a primitive karyotic cell type. An ancient alphaproteobacterium is taken up, which establishes an endosymbiotic interaction with the 'parakaryotic' host cell. (5) The transition of the alphaproteobacterial endosymbiont into an energy-producing mitochondrion forms the basis of the emergence of cellular complexity typical for eukaryotes.

This hypothesis starts with a transient TACK archaeon that contained a full set of the eukaryotic specific proteins (ESPs). Then this ancient TACK lineage lost its cell wall and the cytoskeleton matured into a primitive phagocytosis machinery resulting in rampant HGT into the archaeal genome. In order to maintain genetic integrity, a protective membrane boundary was formed via invagination and an ancient alphaproteobacterium was phagocytosed but not digested. This alphaproteobacterium then established an endosymbiotic relationship with the host cell and eventually evolved into an ATP generating organelle, the mitochondrion. The

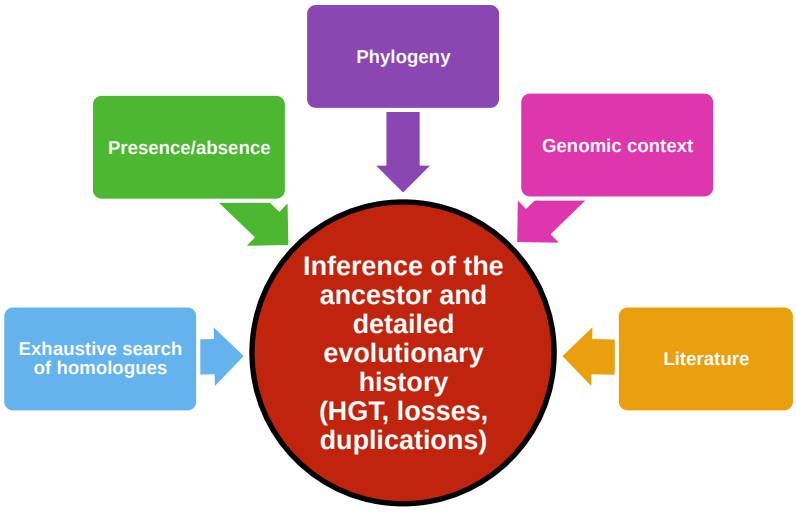
energetic advantage brought by the presence of the mitochondrion then allowed the host cell to evolve cellular complexity (Martijn and Ettema 2013). It has to be noted that this scenario is not very different from classical eukaryogenesis models proposed in the past, except that the starting point is a modern archaeon. However, one important point that is not addressed in the PhAT model is the replacement of the archaeal-lipids, as discussed earlier.

An origin of Eukaryotes from a modern archaeon represents a dramatic shift of paradigm on the tree of life that surely deserves the most exhaustive investigation. Although it is true that many recent analysis using more sophisticated models of evolution have recovered support for the 2 Domains tree, the bulk of this evidence comes from universal proteins, which are prone to suffer from a number of drawbacks. First of all, the number of proteins that can be used is very small because of the need for a bacterial out-group (Gribaldo et al. 2010). This also drastically limits the number of positions that can be unambiguously aligned between the three domains. Moreover, only a limited taxonomic sampling of bacteria Archaea and eukaryotes can be used due to the need for widespread presence in representatives of the three domains as well as the computational burden of large-scale analysis using sophisticated models (Delsuc et al. 2005). Finally, it is also well known that universal proteins are heavily mutationally saturated leading to a loss of phylogenetic signal (Gribaldo and Philippe 2002).

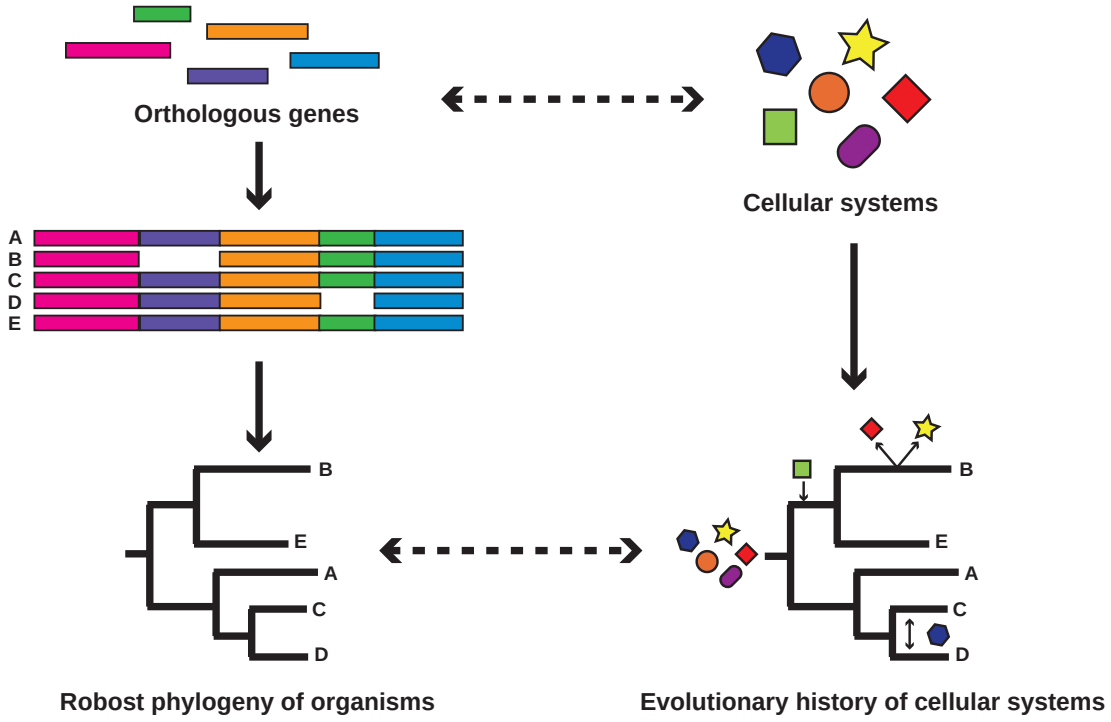
Objectives and Approach

Phylogenomics involves the analysis of genomic data to reconstruct the evolutionary history of organisms and genes. In particular, a branch of phylogenomics uses the combined analysis of orthologous genes, those that were inherited from a common ancestor, to reconstruct the evolutionary history of organisms. One approach consists in concatenating orthologous genes into one large dataset (supermatrix) for phylogenetic analysis. Concatenation of multiple genes increases the phylogenetic signal for ancient events and allows a certain amount of data to be missing because species that are lacking some sequence data are still represented by a large number of informative characters. These approaches are particularly suited to resolving deep evolutionary relationships, which are difficult to solve. A whole range of new approaches and computing facilities are now available and new evolutionary models have been created that better account for sequence evolution in both the Maximum Likelihood and Bayesian frameworks. In addition, hypothesis testing can be performed on different topologies using sophisticated models of evolution.

A second branch of phylogenomics aims at the fine dissection of the evolutionary history of specific cellular processes. Starting from a characterized process, it involves the exhaustive identification of homologs in complete genomes, phylogenetic analysis, genome context analysis, presence and absence patterns, and functional and structural information from the literature. Altogether this process allows for the inference of the ancestor as well as a detailed evolutionary history of the cellular system.



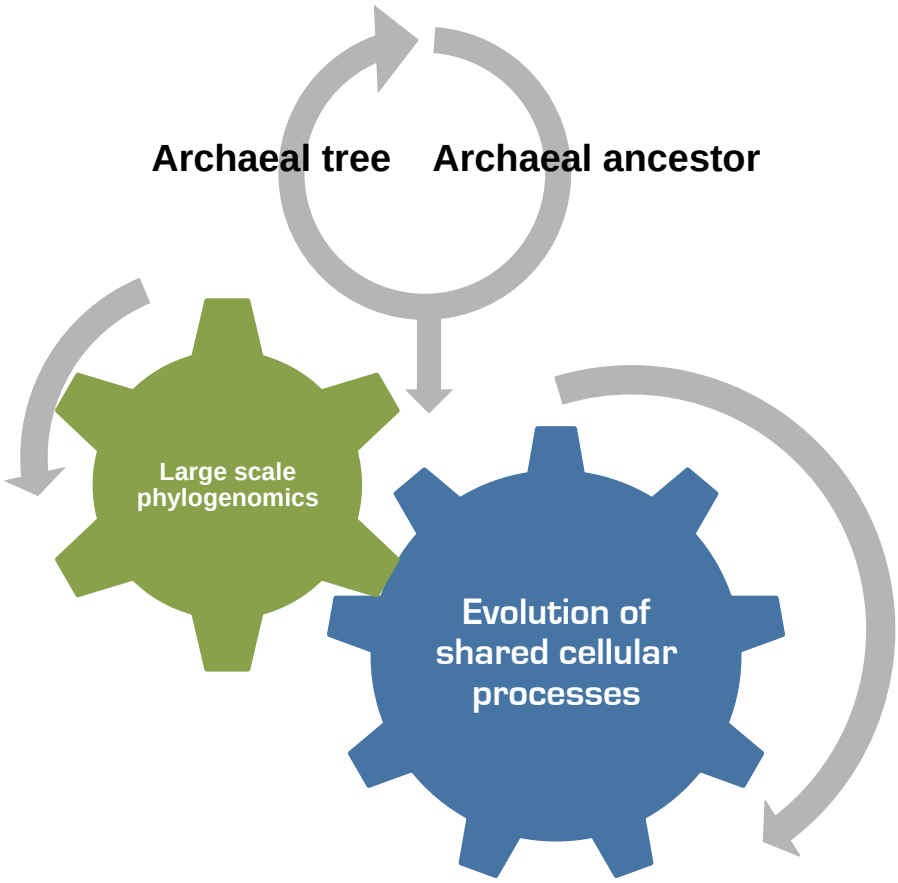
These two phylogenomic branches are highly complementary because the phylogeny obtained from the first approach can be used as a frame of reference to retrace the history of individual genes or cellular systems. In turn, by analyzing the history of individual genes, additional phylogenetic markers can be identified which can be used to establish or clarify evolutionary relationships.



For my thesis I have decided to investigate the relationship between archaea and eukaryotes using two complementary phylogenomic approaches: (i) the analysis of a specific archaeal cellular system with an evolutionary link to eukaryotes, and (ii) a large-scale phylogenomic analysis at the level of the three domains of life.

The key to understanding the relationship between archaea and eukaryotes is to first have a solid picture of the evolutionary history and diversity of the archaeal domain. Not only is it important to have a robust phylogeny, but also it is important to understand the nature of the archaeal ancestor and the subsequent evolution of this domain. For this reason, I first focused on the evolution of DNA replication in the archaea, a key cellular system shared with eukaryotes. The objective of this study was to retrace the evolutionary history of DNA replication in the archaea, reconstruct the archaeal phylogeny, and gain insight on the relationship between archaea and eukaryotes.

Next, I performed a large-scale analysis that involved studying protein families shared between archaea and eukaryotes. By studying protein families shared between these two domains I aimed to identify new phylogenetic markers that could be used to test the archaeal/eukaryote relationship as well as proteins shared between eukaryotes and a specific archaeal lineage. Additionally, I identified phylogenetic markers shared between archaea and bacteria in order to root the archaeal tree, the key to understanding the relationship between archaea and eukaryotes as well as the nature of the archaeal ancestor and the evolution of the entire archaeal domain.



Results

1. Phylogeneomic analysis of DNA replication in the Archaea

My first analysis focused on the fine dissection of an archaeal cellular process evolutionarily linked to eukaryotes. Archaeal informational systems (translation, transcription, replication) harbor a specific evolutionary link with Eukaryotes. Compared to previous analyses of translation and transcription components, phylogenomic analysis of DNA replication appears more complex due to the existence of multiple paralogs which are sometimes highly divergent and probably issued from integrative elements. I performed an exhaustive phylogenomic analysis of the 22 known components of DNA replication in over 140 complete archaeal genomes. This allowed me to accurately assign them in terms of orthology, paralogy, horizontal gene transfers, and copies originating from mobile elements. My results provide a full picture of the diversity of DNA replication among different lineages, and allowed me to infer the presence of a modern-type DNA replication machinery in the last archaeal common ancestor. I was able to clarify the precise evolutionary history that shaped this key cellular machinery during archaeal diversification. This appears incredibly dynamic and involved multiple independent gene losses, duplications, horizontal transfers, non-orthologous replacements, and a significant impact of mobile elements. My study allowed me to highlight a new set of markers that provide information on yet unclear evolutionary relationships within archaea, such as those among the recently discovered lineages with nano-sized representatives. In addition, the precise identification of orthologs allowed me to analyze, for the first time, the phylogenetic signal carried by DNA replication components. This is highly consistent with that harbored by two other key informational machineries, translation and transcription, strengthening the existence of a robust organismal tree for the Archaea. Finally, most of the components inferred to have been present in the archaeal ancestor are shared with eukaryotes, where some were reassigned to other cellular functions, such as repair, allowing discussion on the evolutionary relationships between Archaea and Eukaryotes. Additionally, my results provide important and useful information for future functional studies on DNA replication components in the archaea.

The results from this analysis have been published in GBE in January 2014 and highlighted by the Faculty of 1000 Prime.

Article 1

Global Phylogenomic Analysis Disentangles the Complex Evolutionary History of DNA Replication in Archaea

Kasie Raymann[1,2], Patrick Forterre[1], Céline Brochier-Armanet[3], and Simonetta Gribaldo[1,*]

[1]Département de Microbiologie, Institut Pasteur, Unité Biologie Moléculaire du Gene chez les Extrêmophiles, Paris, France

[2]Université Pierre et Marie Curie, Cellule Pasteur UPMC, Paris, France

[3]Université de Lyon, Université Lyon 1, CNRS, UMR5558, Laboratoire de Biométrie et Biologie Evolutive, Villeurbanne, France

*Corresponding author: E-mail: simonetta.gribaldo@pasteur.fr.

Accepted: December 29, 2013

Abstract

The archaeal machinery responsible for DNA replication is largely homologous to that of eukaryotes and is clearly distinct from its bacterial counterpart. Moreover, it shows high diversity in the various archaeal lineages, including different sets of components, heterogeneous taxonomic distribution, and a large number of additional copies that are sometimes highly divergent. This has made the evolutionary history of this cellular system particularly challenging to dissect. Here, we have carried out an exhaustive identification of homologs of all major replication components in over 140 complete archaeal genomes. Phylogenomic analysis allowed assigning them to either a conserved and probably essential core of replication components that were mainly vertically inherited, or to a variable and highly divergent shell of extra copies that have likely arisen from integrative elements. This suggests that replication proteins are frequently exchanged between extrachromosomal elements and cellular genomes. Our study allowed clarifying the history that shaped this key cellular process (ancestral components, horizontal gene transfers, and gene losses), providing important evolutionary and functional information. Finally, our precise identification of core components permitted to show that the phylogenetic signal carried by DNA replication is highly consistent with that harbored by two other key informational machineries (translation and transcription), strengthening the existence of a robust organismal tree for the Archaea.

Key words: Cdc6/Orc1, RPA/SSB, DNA gyrase, primase, phylogeny, nanosized archaea.

Introduction

Replication of the genetic material is a crucial step of the cell cycle. All three domains of life replicate their DNA semiconservatively (Meselson and Stahl 1958) and follow basically the same sequence of events (for a recent review see DePamphilis and Bell [2010]): The replication fork is assembled by a specific protein or initiation complex that recognizes the origin of replication on the chromosome and opens up the double-stranded DNA. A helicase is then recruited, producing a replication bubble that is protected by single-stranded DNA-binding proteins. The core replication machinery then assembles at the fork with the help of the sliding clamp, a ring-shaped factor that tethers it to the DNA template. The main replicative polymerase extends DNA replication bidirectionally from short RNA primers made by a primase, with one strand being synthesized continuously (leading strand), and the other discontinuously (lagging strand). The Okazaki fragments produced during synthesis of the lagging strand are joined together by a DNA ligase after excision of the RNA primers. During the whole process, a number of topoisomerases act to resolve topological problems arising from DNA supercoiling in front of the replication fork and chromosome entangling at the end of replication. Despite the overall conservation of these major steps, the machinery used for DNA replication in Archaea and Eukaryotes exhibits striking differences to the bacterial replication machinery, which uses nonhomologous proteins belonging to completely different families (fig. 1) (Grabowski and Kelman 2003; Barry and Bell 2006).

The archaeal replication machinery is generally considered to be a simplified version of the eukaryotic apparatus, which usually harbors more components (fig. 1). However, it too has its own peculiar characteristics. Along with a PolB polymerase, most archaea also possess a PolD polymerase whose catalytic subunit has no homologs in Bacteria or Eukaryotes (Cann et al. 1998). Furthermore, to relax positive superturns arising during replication and decatenate the chromosome at the end of

A

		Bacteria	Archaea	Eukaryotes
Initiation				
	Origin recognition	*Dna A*	Cdc6-Orc1	ORC(1), *ORC(2-6)*, Cdc6
	Helicase loading	*Dna C*	Cdc6-Orc1	Cdc6
				Cdt1
	Replicative helicase	*Dna B (E.coli)*	MCM	MCM (2-7)
			GINS-51, GINS-23	GINS (Sld5, Psf1), GINS (Psf2, Psf3)
			RecJ(Cdc45)?	Cdc45
Elongation				
	Single stranded binding protein	*SSB*	SSB	*
			RPA1	RPA70, *RPA34, RPA14*
	Polymerase/exonuclease	*Pol III (Family C)*	Pol B (Family B)	Pol δ, Pol e (Family B)
			Pol D (DPL and DPS)	
	Clamp loader	*γ complex*	RFC-L, RFC-S	RFC-L, RFC-S(1-4)
	Sliding clamp	*β clamp*	PCNA	PCNA
	Primase	Dna G	Primase (PriS, PriL), DnaG?	Primase (PriS, PriL)
				B subunit
				DNA pol α
	Primer excision (lagging)	*RNase HI*	RNase HII	RNase H2A, *RNase H2B, RNase H2C*
		DNA Pol I	FEN-1	FEN-1
			Dna2?	Dna2
	Maturation (lagging)		ATP-dependent DNA ligase	ATP-dependent DNA ligase
		NAD+-dependent DNA ligase	NAD+-dependent DNA ligase	
DNA relaxation				
	Topoisomerases		*Topo VI (Type IIB)*	*
			Topo IB (Type IB)	Topo IB (Type IB)
		DNA gyrase, Topo IV (Type IIA)	DNA gyrase (Type IIA)	Topo IIA (Type IIA)

B

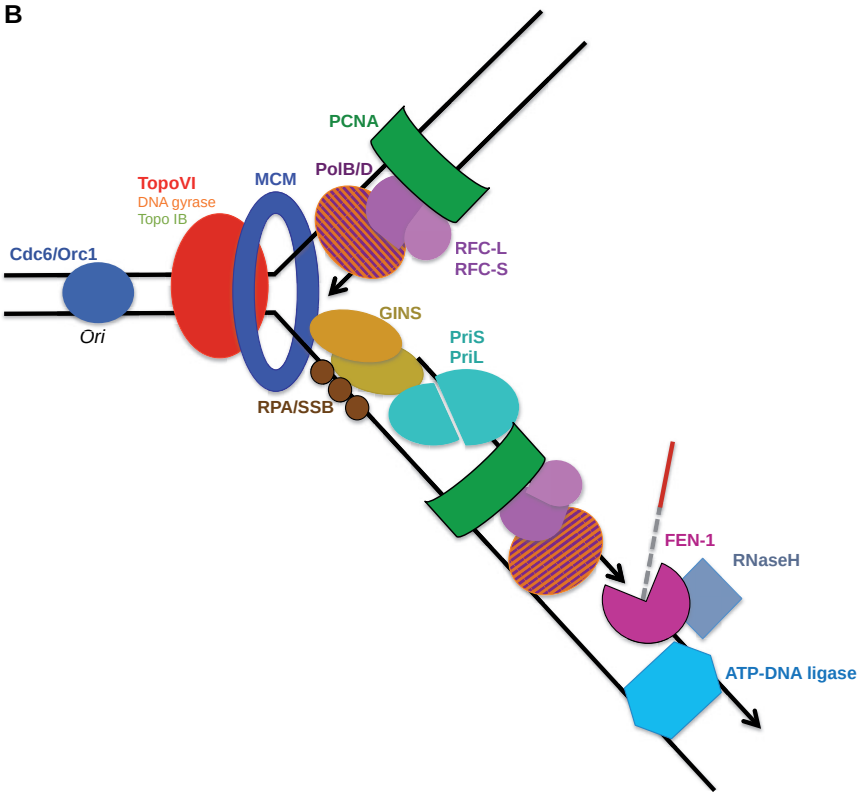


Fig. 1.—(*A*) General overview of the components of DNA replication in the Archaea compared to the other two domains of life. Same color in a given row indicates homology; gray shading indicates that the bacterial version has only structural similarity with the archaeal/eukaryal component; question marks represent components with unclear implication in archaeal replication, i.e., DnaG, Dna2, and RecJ homologs; asterisks indicate that a eukaryotic homolog exist but is not involved in replication, i.e., SSB and TopoVI. See main text for details. (*B*) Sketch of the DNA replication machinery in the Archaea. Colors corresponds to those in (*A*).

replication, two tasks which are performed by Type IIA enzymes in Eukaryotes and Bacteria, most archaea use a topoisomerase of the type IIB family (TopoVI; Forterre et al. 2007). Some archaeal components have homologs in eukaryotes that are not involved in DNA replication (fig. 1). For example, eukaryotic homologs of the catalytic subunit of archaeal TopoVI (Spo11) are involved in the initiation of meiotic recombination (Bergerat et al. 1997). Additionally, homologs of the archaeal single-stranded binding (SSB) proteins were identified in eukaryotes several years ago (Robbins et al. 2005) and are the subject of growing appeal due to their probable yet poorly understood role in genome integrity (Richard et al. 2008; Shi et al. 2012). The role of some homologs of eukaryotic replication components in archaea is not clear and remains to be confirmed by functional studies. For example, Dna2 may be involved in Okazaki fragment maturation, performing the same function as in eukaryotes (Higashibata et al. 2003). Similarly, the role of the archaeal RecJ, a 5′-3′ exonuclease (also found in bacteria and a distant homolog of eukaryotic Cdc45) remains to be verified experimentally, but may fulfill the same function in an archaeal CMG (Cdc45, MCM, GINS) complex (Makarova et al. 2012). Archaea also harbor a few homologs of bacterial replication components such as NAD+-dependent DNA ligase, DNA gyrase, and DnaG (fig. 1). Although ATP-dependent ligases are ubiquitous in Archaea and Eukaryotes (Wilkinson et al. 2001; Martin and MacNeill 2002), bacterial-like NAD+-dependent ligases have been identified in some members of Halobacteriales (Zhao et al. 2006). DNA gyrase, a topoisomerase belonging to the Topo IIA family, is present in a number of euryarchaeal lineages (Forterre et al. 2007). In the case of archaeal homologs of bacterial primase DnaG (Aravind and Koonin 1998), the proposal that they are involved in replication (Bauer et al. 2013) is weakened by strong evidence that suggests a role in RNA metabolism (Hou et al. 2013).

Remarkably, the machinery for DNA replication appears to vary greatly among archaeal lineages, which can harbor various combinations of key components. This variation includes different main replicative polymerases (PolB and PolD), single or multiple replication origins and initiator proteins (Cdc6/Orc1), different SSB proteins (SSB, RPA), and alternative multimeric complexes (PCNA, RFC, and GINS); (Grabowski and Kelman 2003; Barry and Bell 2006; McGeoch and Bell 2008; Bell 2011; Beattie and Bell 2011). There have also been reports of possible replacements of components by nonhomologous proteins, such as the putative initiator protein MJ0774 in *Methanococcus jannaschii* (Zhang RR and Zhang C-TC 2004) and the putative single-stranding binding protein ThermoDPB in Thermoproteales (Paytubi et al. 2012). Moreover, archaeal genomes can display additional copies of replication components that are often embedded in integrative elements of plasmid and/or viral origin. For example, the integrated element TKV3 of *Thermococcus kodakarensis* KOD1 encodes a homolog of PCNA (Fukui et al. 2005); *Haloferax volcanii* harbors three replication origins and nine Cdc6/Orc1 coding genes, with one pair embedded in a 50 kb prophage region (Hartman et al. 2009); Sulfolobales contain three replication origins and three Cdc6/Orc1 copies, one of which is associated with the second origin of replication that was contributed by an integrative element (Samson et al. 2013). Finally, a number of additional divergent MCM homologs originating from integrative elements or plasmids are present in various archaeal taxa (Krupovic, Gribaldo, et al. 2010).

Such extreme diversity has made it particularly challenging to dissect the evolutionary history of archaeal DNA replication. Although some components have been previously analyzed (Chia et al. 2010; Krupovic, Gribaldo, et al. 2010), no attempt has been made to perform a global survey of the complete machinery. Here, we have carried out an in depth phylogenomic analysis of all components of DNA replication in over 140 complete archaeal genomes. We specifically assess the taxonomic distribution of homologs in each of these genomes. In addition, we precisely identify copies arising from integrative elements/decaying paralogs/horizontal gene transfers as opposed to those that constitute a conserved and vertically inherited core replication machinery, providing important information for further evolutionary and functional analysis of these components. Phylogenetic analysis of the core components allowed us to infer the nature of DNA replication in the last archaeal common ancestor (LACA) and the subsequent evolutionary history that shaped this machinery. Finally, our analysis enabled us to investigate, for the first time, the phylogenetic signal carried by DNA replication. It shows remarkable consistency with that harbored by the two other main informational processes (transcription and translation), confirming the existence of a robust phylogenomic core that can be used to reconstruct the tree of the Archaea.

Materials and Methods

Identification of Homologs of DNA Replication Components

Homologs of each archaeal DNA replication component were retrieved from the reference sequence database at the National Center for Biotechnology Information (NCBI) using the BlastP (Altschul et al. 1997) program with different seeds from each archaeal order. The top 100 best hits for each order were then used to create hidden Markov model (HMM) profiles (Johnson et al. 2010; http://www.hmmer.org, last accessed January 16, 2014) that allowed an iterative search of a local database of 142 archaeal genomes including 98 plasmid sequences and a local database of 56 complete archaeal virus genomes downloaded from the Viral Genomes database of NCBI (as of June 20, 2013) (supplementary table S3, Supplementary Material online). The absence of a given homolog in a specific genome was verified by performing additional tBlastN (Altschul et al. 1997) searches. Genomic

context was investigated using MaGe (Vallenet et al. 2005), MGV2 (Kerkhoven et al. 2004), and STRING (Szklarczyk et al. 2011).

Phylogenetic Analysis

Multiple alignments were performed with MUSCLE v3.8.31 (Edgar 2004) and manually inspected using the ED program from the MUST package (Philippe 1993) to verify that all sequences retrieved at the first step were homologous. Final single protein data sets were trimmed using the software BMGE (Criscuolo and Gribaldo 2010) with default parameters and subjected to phylogenetic analyses by Maximum Likelihood and Bayesian methods. Maximum likelihood analyses were performed with Treefinder (Jobb et al. 2004; version of 2008). For each protein data set, the best-fit parameters and model of amino acid substitution were chosen using the Akaike information criterion with a correction (AICc) for finite sample sizes (Hurvich and Tsai 1989) as implemented in Treefinder (Jobb et al. 2004). Bootstrap supports were calculated based on 100 resamplings of the original alignment. Bayesian analyses were run with MrBayes 3.2 (Ronquist et al. 2012), using the mixed amino acid substitution model and four categories of evolutionary rates. Two independent runs were performed for each data set, and runs were stopped when they reached a standard deviation of split frequency below 0.01 or the log likelihood values reached stationary. The majority rule consensus trees were obtained after discarding first 25% samples as burn-in.

For the analysis of DNA gyrase, alternative tree topologies were statistically evaluated using the following paired-sites tests: expected-likelihood weights, bootstrap probability (BP; Felsenstein 1985), Kishino and Hasegawa (Kishino and Hasegawa 1989), Shimodaira and Hasegawa (SH; Shimodaira and Hasegawa 1999; Goldman et al. 2000), Weighted SH test (Shimodaira and Hasegawa 1999; Buckley et al. 2001), and approximately unbiased (AU) test (Shimodaira 2002) as implemented in Treefinder (Jobb et al. 2004). A total of 500000 RELL (Kishino et al. 1990) replicates were run. Three alternative topologies were tested and it was determined that the data did not reject the topology if the *P* value was greater than 0.05 for all tests.

Supermatrix Analyses

Fourteen DNA core replication proteins that were present in at least 60% of the archaeal genomes (PriS, MCM, PCNA, Cdc6/Orc1, DPL, DPS, PolB, TopoVI-A, TopoVI-B, RFC-s, RFC-l, RNaseH, DNA ligase, and FEN-1) were retained for supermatrix analysis. To handle species-specific paralogs, we chose one paralog, and the slowest evolving if applicable, to limit possible artifacts due to fast evolutionary rates. In the case of ancient paralogs, we also chose those representing the cluster with larger taxonomic representation and/or showing the slowest evolutionary rates. For example, we chose the Cdc6/Orc1-1 paralog (see Results). Each multiple alignment was independently realigned, trimmed, and concatenated into a character supermatrix comprised of 4,295 amino acid positions and 129 archaeal taxa (after keeping only one representative strain of the same species). PhyloBayes 3.3b (Lartillot et al. 2009) was used to perform Bayesian analysis using the CAT + GTR model and a gamma distribution with four categories of evolutionary rates was used to model the heterogeneity of site evolutionary rates. The concatenated datasets were also recoded using Dayhoff 6 and Dayhoff4 recoding schemes as implemented in PhyloBayes 3.3b (Lartillot et al. 2009) and analyzed with the same model parameters. For each data set, two independent chains were run until convergence (max diff < 0.01). The first 25% of trees were discarded as burn in and the posterior consensus was computed by selecting one tree out of every two to compute the 50% majority consensus tree. Maximum likelihood analysis was carried out by PhyML (Guindon et al. 2010), the LG model and a gamma correction with four categories of evolutionary rates. Bootstrap support was calculated based on 100 resamplings of the original alignment.

Results

Archaeal DNA Replication: The Core Component and the Variable Shell

We performed an exhaustive search for homologs of the 16 major components of the DNA replication machinery (22 proteins considering subunits) in 142 complete archaeal genomes (fig. 2; supplementary table S1, Supplementary Material online). The taxonomic distribution of these proteins shows a highly dynamic pattern along the different archaeal lineages. Some components are present essentially in a single copy and in the majority of genomes (e.g., GINS 51, TopoVI A and B, RFC-L, DNA ligase 1, Fen1, RNase HII, PriS, and PriL), whereas others are missing altogether from a number of archaeal lineages (e.g., Cdc6/Orc1 in Methanococcales and Methanopyrales, TopoVI in Thermoplasmatales, PolD and RPA in Crenarchaeota, SSB in most Euryarchaeota and Thermoproteales). Incomplete assembly of some genomes, such as the Nanohaloarchaea, uncultured marine group II, *Candidatus* Caldiarchaeum subterraneum (Aigarchaeota), and the ARMANS (Archaeal Richmond Mine Acidophilic Nanoorganisms) suggests that some absences in these taxa must be taken with caution. Finally, a few components display a large number of extra copies in some taxa (e.g., Cdc6/Orc1 in Halobacteriales, MCM in Methanococcales, RPA in many Euryarchaeota, PolB in many Euryarchaeota and Crenarchaeota, PCNA in Crenarchaeota).

Inspection of multiple alignments, phylogenies, and genome synteny allowed us to highlight two categories of homologs: 1) slow-evolving homologs lying within chromosomal regions that are syntenic among closely related taxa and whose phylogeny is overall consistent with the archaeal

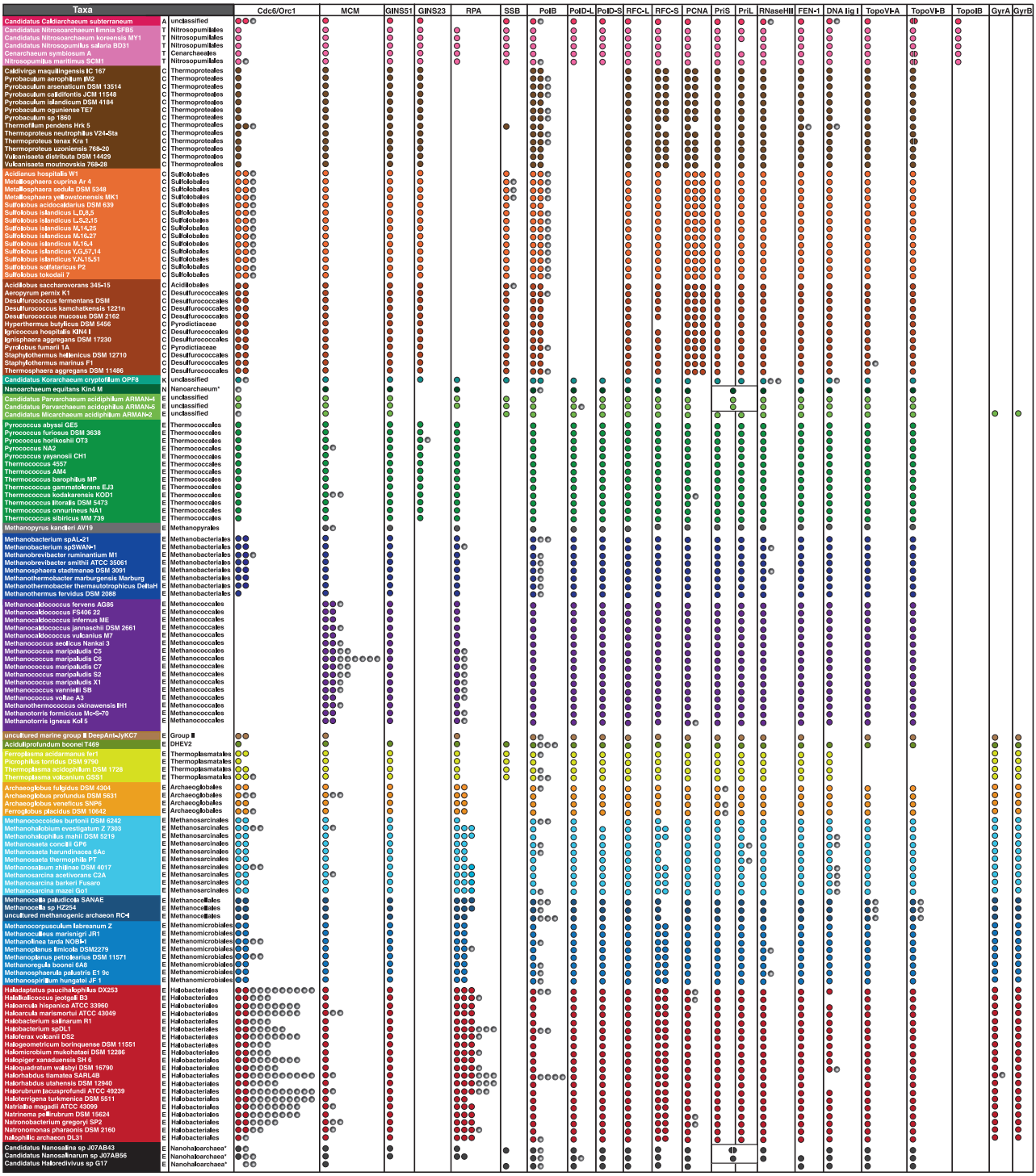


Fig. 2.—Distribution of homologs of 22 main replication components in 142 archaeal genomes. Filled circles represent homologs that we assigned to the core replication machinery, whereas gray circles represent homologs assigned to the shell component (see text for details). Split genes are indicated by half circles, and the fused primases by a box (see text for details). Letters in first column indicate the phylum (A, Aigarchaeota; T, Thaumarchaeota; C, Crenarchaeota; K, Korarchaeota; N, Nanoarchaeota; E, Euryarchaeota). Asterisks indicate classes instead of orders. Full accession numbers are given in supplementary table S1, Supplementary Material online.

phylogeny, as opposed to 2) highly divergent copies that lie within nonconserved genetic contexts and/or display more restricted taxonomic sampling and inconsistent phylogenetic affiliations. We reasoned that the first category represents components that were primarily vertically inherited during archaeal diversification and form what we called the conserved core replication components (fig. 2, filled circles; for full accession numbers see supplementary table S1, Supplementary Material online), whereas the second category represents horizontally transferred genes, decaying paralogs, or homologs arising from integration of extrachromosomal elements that form a variable pool of proteins that we called the shell replication components (fig. 2, open circles; for full accession numbers see supplementary table S1, Supplementary Material online).

An example of our approach is provided by the analysis of Cdc6/Orc1. Except for the previously mentioned absence in Methanococcales and Methanopyrales, all archaeal genomes contain at least one homolog of the initiation protein Cdc6/Orc1. Most lineages harbor at least two copies, and a very large number of homologs are present in Halobacteriales (fig. 2). We found that in each genome only one or two Cdc6/Orc1 homologs are slow evolving and show conserved synteny among closely related taxa. Additional copies, when present, are very divergent and display nonconserved genomic contexts. When a phylogenetic tree was built from all homologs (not shown) the first category formed two clearly distinct clusters representing a large taxonomic coverage, which, albeit not completely resolved, is globally consistent with archaeal phylogeny. In contrast, the second category fell into an unresolved group showing very long branches, restricted taxonomic coverage and highly inconsistent phylogenetic relationships. The first category was therefore assigned to the core replication machinery (fig. 2, filled circles; supplementary table S1, Supplementary Material online), and the second to the shell (fig. 2, open circles; supplementary table S1, Supplementary Material online). For validation, among the three Cdc6/Orc1 copies present in Sulfolobales, we correctly assigned the copy corresponding to the origin of replication embedded in an integrative element as a shell component (Robinson and Bell 2007). Similarly, among the large number of Cdc6/Orc1 copies present in Halobacteriales, only two were identified as part of the core replication, whereas all others fell into the shell component (fig. 2; supplementary table S1, Supplementary Material online).

The identification of the fast-evolving shell components allowed for a finer analysis of the precise evolutionary history of core Cdc6/Orc1 proteins (fig. 3). Although the tree was not completely resolved due to the limited number of positions analyzed, the monophyly of the two clusters was strongly supported, each displaying robust monophyletic groups corresponding to the major archaeal phyla and orders (fig. 3*A*). In particular, when two copies are present in a given taxon, they generally correspond to either one cluster or the other. For instance, this is the case of the two core paralogs of Sulfolobales; one corresponds to the first cluster (Cdc6/Orc1-1) and the other to the second cluster (Cdc6-Orc1-2). The same is true for Halobacteriales, where only two core paralogs belonging to each of the two clusters could be identified. This suggests that Cdc6/Orc1-1 and Cdc6/Orc1-2 are ancient paralogs that arose from gene duplication and were both likely present in the LACA. Therefore, the absence of one of the two copies in present day genomes must be interpreted as the consequence of gene loss (fig. 3*B*). This trend of gene loss is observed across the whole archaeal tree, with different lineages having lost either one paralog or the other. For example, we can infer loss of Cdc6/Orc1-2 in the ancestor of Thaumarchaeota and in the ancestor of Thermococcales, and loss of Cdc6/Orc1-1 in the ancestor of Thermoproteales and Korarchaeota (fig. 3*B*). Methanococcales and Methanopyrales have pushed this trend to the extreme by losing both copies, likely in parallel to replacement by a nonorthologous protein (Zhang RR and Zhang C-CT 2004; Berthon et al. 2008). The Cdc6/Orc1-2 cluster appears to evolve faster than the Cdc6/Orc1-1 cluster and exhibits a few inconsistencies with the archaeal phylogeny, such as the branching of Korarchaeota and Aigarchaeota within Thermoproteales, and of Thermoplasmatales/uncultured marine group II at the base of Crenarchaeota (fig. 3*A*). More data from these lineages will be necessary to clarify whether these taxa acquired their Cdc6/Orc1-2 via horizontal gene transfer from Crenarchaeota, or if these placements are the result of a tree artifact. Indeed, a number of horizontal gene transfers from Crenarchaeota are known to have occurred during adaptation of Thermoplasmatales to thermoacidic environments (Fütterer et al. 2004). Finally, Halobacteriales have kept both Cdc6-Orc1 and Cdc6/Orc1-2 paralogs, but most genomes have acquired multiple extra copies arising from integration of mobile elements (fig. 2). It has to be noted that Cdc6/Orc1-1 coincides with one of the three origins of replication identified in *H. volcanii* (Hawkins et al. 2013), but Cdc6-Orc1-2 does not. The same is true for *Sulfolobus solfataricus,* where only Cdc6/Orc1-1 coincides with one of the three origins of replication (Samson et al. 2013).

The Cdc6/Orc1 case is not unique. By using the same approach, we identified shell copies for most replication components, with an apparent preference for Cdc6/Orc1, MCM, PCNA, and PolB (fig. 2). Remarkably, the components that appear enriched in shell copies are also specifically present in plasmid and viral sequences, particularly from Halobacteriales (fig. 4; supplementary table S2, Supplementary Material online). This suggests that the shell replication homologs may come predominantly from extrachromosomal elements. In addition, it appears that extrachromosomal entities are enriched with different replication proteins, for example, Cdc6/Orc1 is more abundant in plasmids and PolB is particularly present in viruses (fig. 4). Although the current taxonomic covering of viral and plasmid sequences from archaea is

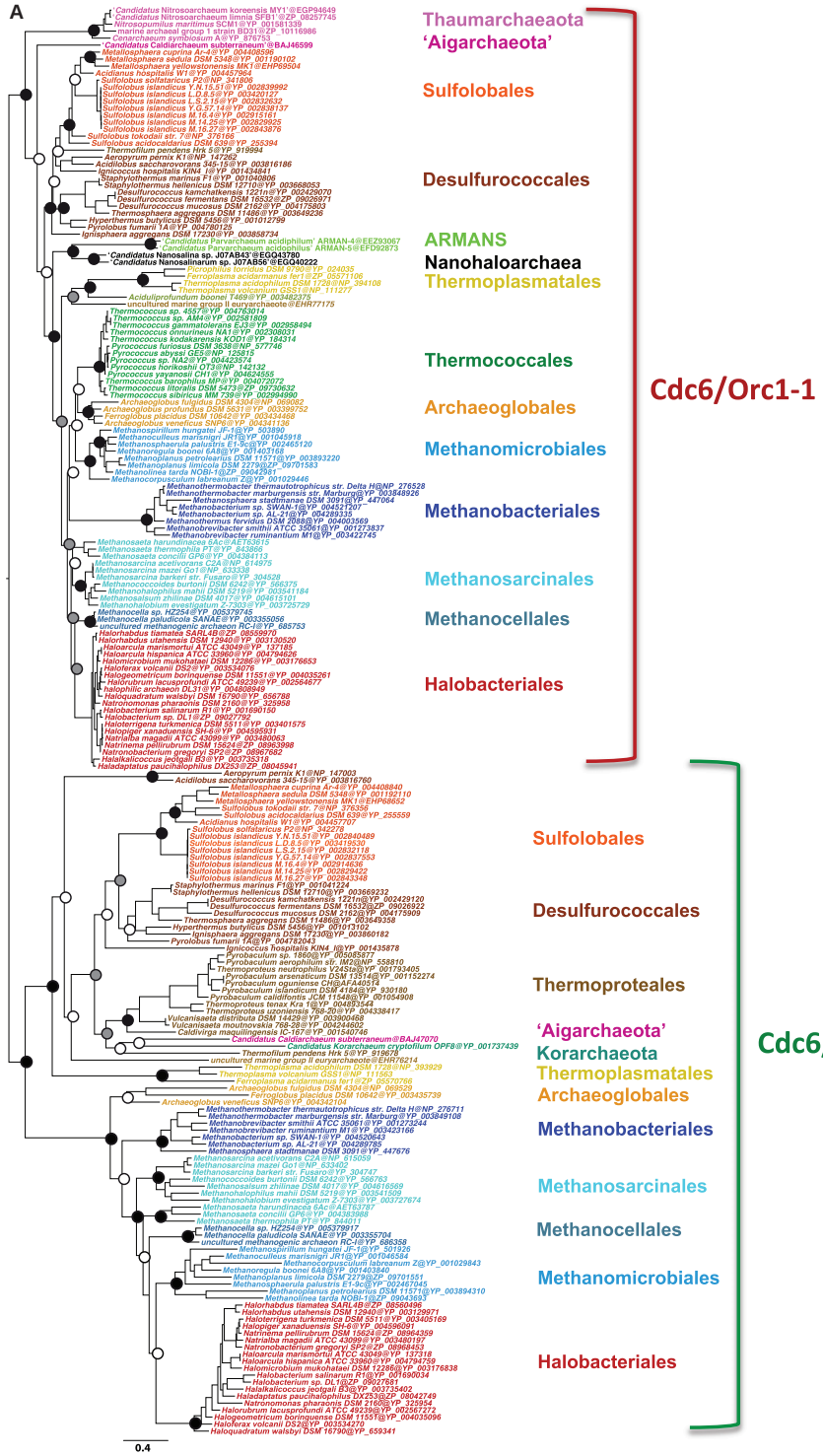


Fig. 3.—(*A*) Maximum likelihood phylogeny of Cdc6/Orc1 core components. The tree was calculated by Treefinder (MIX model + gamma4) based on 261 unambiguously aligned amino acid positions. The scale bar represents the average number of substitutions per site. Dots represent bootstrap values (BV) based on 100 replicates of the original alignment. For clarity, supports are shown for major lineages only: black dots indicate BV > 90%, gray dots BV 80–90%, and white dots BV <80%. (*B*) Evolutionary scenario for Cdc6/Orc1. The two Cdc6/Orc1 paralogs 1 (red) and 2 (green) arose from ancestral gene duplication in the Last Common Archaeal Ancestor. Independent gene losses occurred subsequently in a number of lineages, involving either one paralog (red crosses) or the other (green crosses), and in some cases both. See text for details.

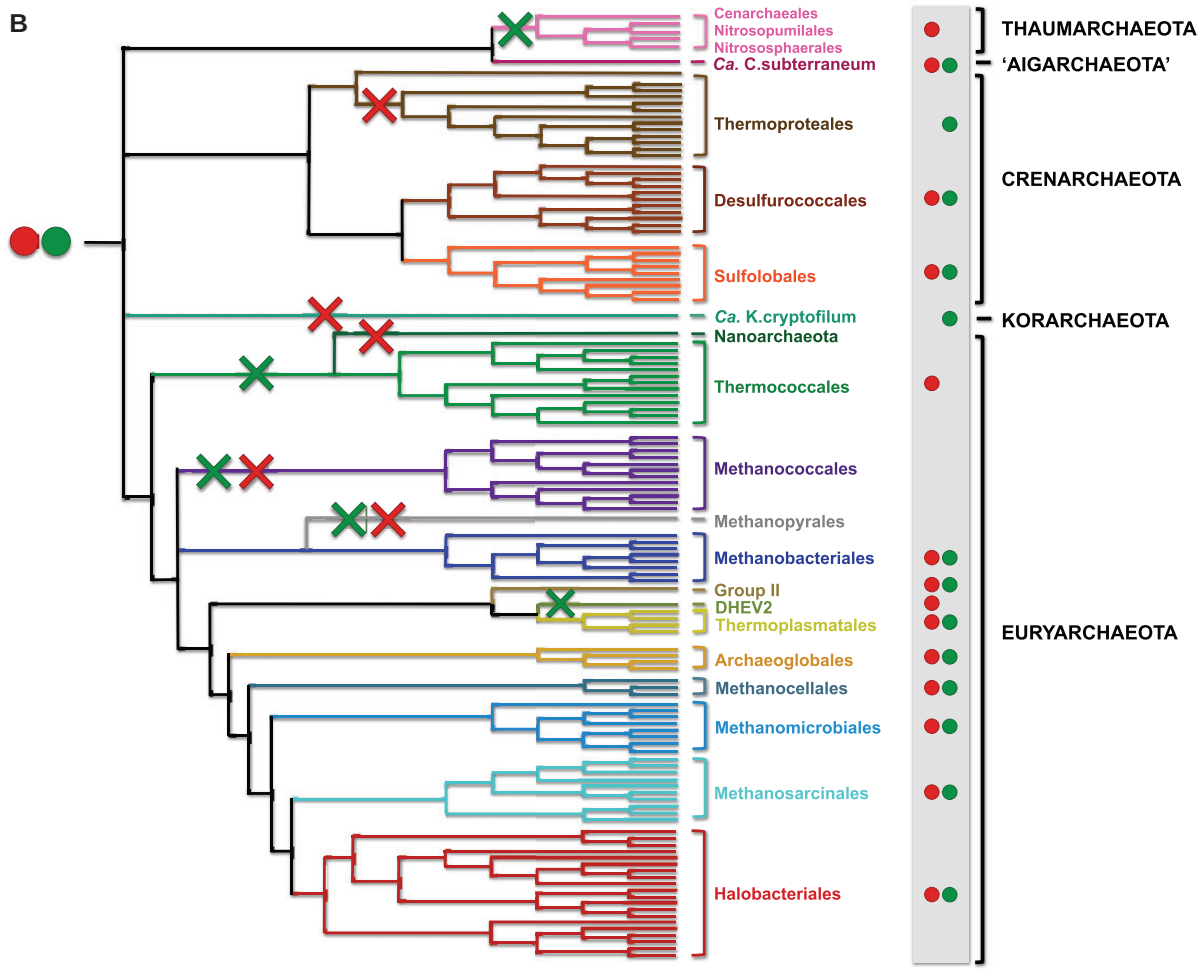


Fig. 3.—Continued.

narrow (supplementary table S3, Supplementary Material online), these data suggest that replication proteins are frequently exchanged between extrachromosomal elements and cellular genomes.

The precise identification of core and shell replication components can be important for functional studies on archaeal replication, as proteins belonging to the core may have essential roles while shell components may keep functions linked to their extrachromosomal entity. For instance, of the three MCM present in *T. kodakarensis*, we assigned the gene encoding MCM3 (TK1620) to the core (supplementary table S2, Supplementary Material online); in fact, experimental data have shown that this is the only essential copy and is likely the only MCM involved in genome replication (Pan et al. 2011). Additionally, of the two PCNA homologs in *T. kodakarensis*, we designated PCNA1 (TK0535) as the core component and PCNA2 (TK0582) as the shell, consistent with the finding that only PCNA1 is required for cell viability (Pan et al. 2013).

The analysis of each replication protein allowed us to precisely reconstruct the global evolutionary history of DNA replication in the Archaea and the dynamics that shaped this key cellular machinery from the LACA throughout the subsequent diversification of this Domain of Life. Some of our results also provide interesting evolutionary and functional information, and are detailed hereafter.

Complex Evolutionary History of SSB and RPA Proteins

It is commonly assumed that SSB proteins with a single OB fold and a flexible C-terminal tail (SSB) are typical of Crenarchaeota (Wadsworth and White 2001) and that SSB proteins with multiple OB folds (RPA) are typical of Euryarchaeota (Grabowski and Kelman 2003; Kerr et al. 2003). The high degree of sequence divergence among archaeal SSB proteins makes the assignment of homologs particularly challenging. According to sequence similarity and the presence of single or

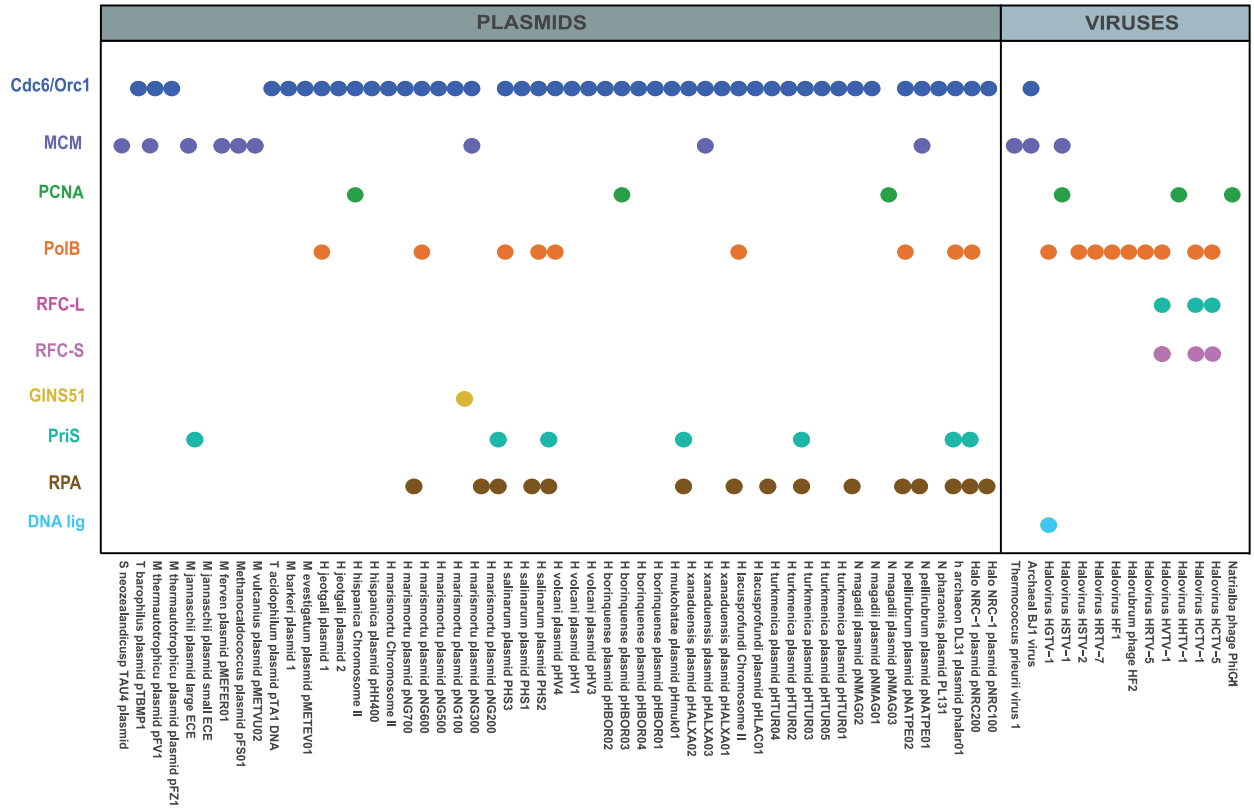


Fig. 4.—Homologs of DNA replication proteins found in archaeal plasmids and viruses. Colors correspond to those used in figure 1. Accession numbers are given in supplementary table S2, Supplementary Material online.

multiple OB folds, we now clarified the distribution of SSB and RPA homologs in all archaeal genomes (fig. 2; supplementary table S1, Supplementary Material online).

Euryarchaeal RPAs can display different domain architectures and form various structural conformations. For example, *Methanococcus jannaschii* encodes a unique SSB protein, homologous to eukaryotic RPA70 that functions as a monomer in solution (Kelly et al. 1998). *Methanosarcina acetivorans* encodes a homolog of eukaryotic RPA70 called MacRPA1, along with two divergent homologs, MacRPA2 and MacRPA3, each able to self-assemble into a homomultimeric complex (Robbins et al. 2004; Skowyra and MacNeill 2012). In addition, many archaeal genomes encode proteins that are not homologous to RPA but are found close by and therefore were called RPA-associated proteins (Berthon et al. 2008) (hereafter referred to as RAP). In *H. volcanii* these RPA-associated proteins are thought to be cotranscribed with the adjacent RPA2 and RPA3 genes (Skowyra and MacNeill 2012) and have been shown to interact with them (Stroud et al. 2012). We found that homologs related to Methanosarcina RPA1 are largely distributed in archaeal genomes (in yellow in fig. 5, see also supplementary table S2 [Supplementary Material online] for full accession numbers) and their phylogeny, although not completely resolved, is consistent with the archaeal tree (not shown). Therefore, these likely represent the core RPA component and are likely essential. In fact, among the three RPA copies present in *H. volcanii*, the copy that we assigned to the core is the only one that is essential (Skowyra and MacNeill 2012).

A number of late emerging euryarchaeal lineages also display one or two additional and divergent RPA homologs that we classified as RPA2 and RPA3 according to their sequence similarity to *Methanosarcina acetivorans* MacRPA2 and MacRPA3 (in green in fig. 5, see also supplementary table S2 [Supplementary Material online] for full accession numbers). Their specific distribution in late emerging euryarchaeal lineages and phylogenetic analysis (not shown) indicates that RPA2 and RPA3 are paralogs that arose via gene duplication in Euryarchaeota, after the divergence of Thermococcales, Methanococcales, and Methanobacteriales. We found that RPA2 and RPA3 always lie close to RAP2 and RAP3 proteins (in red in fig. 5). RAP2 and RAP3 proteins are homologous and phylogenetic analysis showed a consistent topology to that of RPA2/RPA3 (not shown) suggesting that they also arose by gene duplication in the same ancestor. Such similar evolutionary history and genomic association strongly points to an ancient and important functional linkage of RPA and their associated proteins in these euryarchaeota.

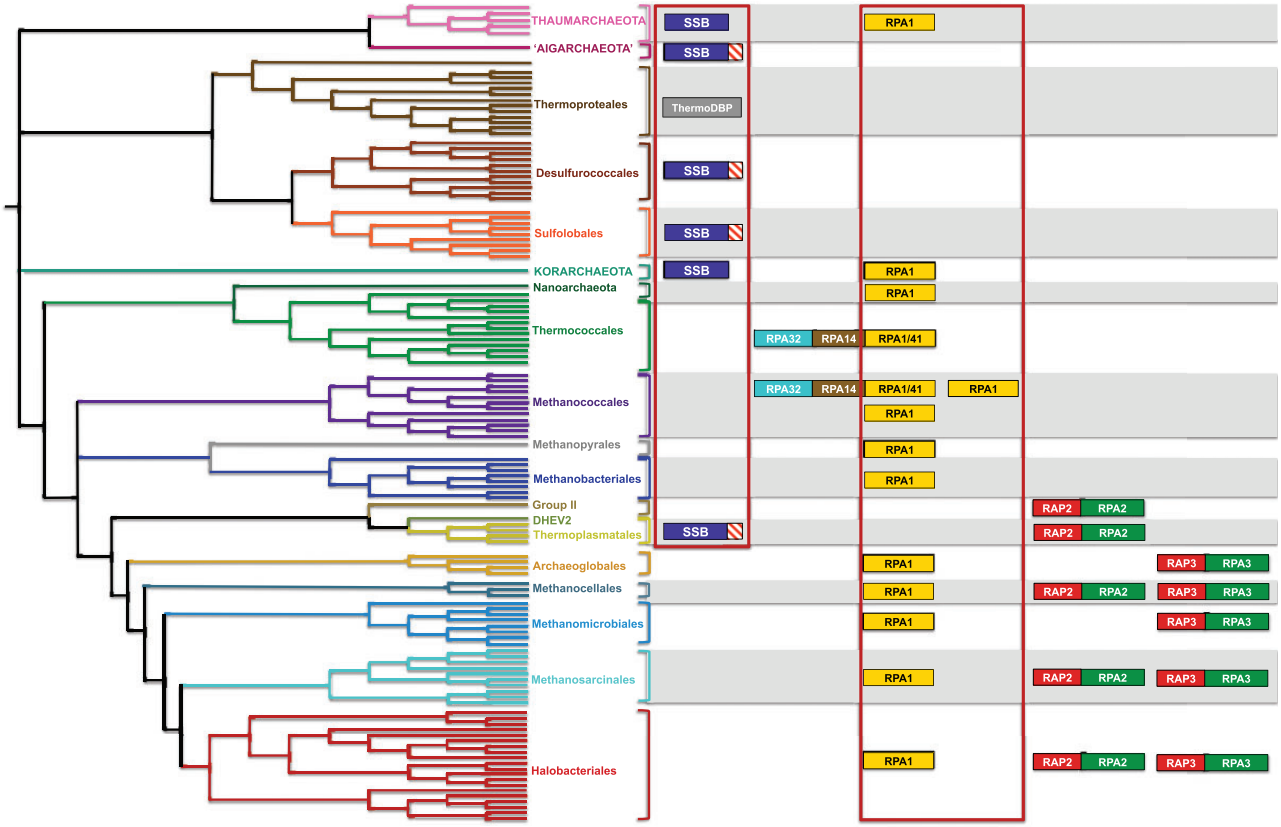


Fig. 5.—Taxonomic distribution and diversity of archaeal SSB and RPA homologs plus the associated proteins (RAP2 and RAP3). ThermoDP, the proposed replacement for the native SSB of Thermoproteales, is shown in gray. See text for details.

Thermococcales display very peculiar characteristics concerning their SSB proteins. *Pyrococcus furiosus* harbors three nonhomologous SSB proteins: RPA41, RPA14 (which, despite its name, is not homologous to eukaryotic RPA14), and RPA32. Together these form a stable heterotrimeric complex, and their encoding genes are adjacent in the genome (Komori and Ishino 2001). RPA41 is only distantly related to other archaeal RPA1 homologs, and closely related homologs of RPA32, RPA14, and RPA41 are also found in Methanococcales where they maintain the same genomic arrangement (fig. 5). Because these two orders do not share an exclusive common ancestor according to ribosomal protein trees (Matte-Tailliez et al. 2002; Brochier et al. 2004, 2005; Brochier-Armanet et al. 2011), the presence of such a unique three-protein RPA system may be explained with a horizontal gene transfer, either directly or through a common mobile element, which possibly displaced the original RPA1. In fact, some Methanococcales genomes still harbor an RPA1 homolog that may represent the original protein (fig. 5; supplementary table S2, Supplementary Material online).

In contrast to RPA, SSB homologs have a much more restricted taxonomic distribution and are mostly present in a single copy (fig. 2; supplementary table S2, Supplementary Material online). The presence of an SSB in *Thermophilum pendens*, an early emerging lineage in the Thermoplasmatales, testifies to the ancestral presence of this protein in this lineage prior to its replacement by the nonhomologous ThermoDPB (Paytubi et al. 2012). The distribution of SSB appears complementary to that of RPA, with the notable exception of Thaumarchaeota, Korarchaeota, Thermoplasmatales/DHEV2, two Nanohaloarchaea, and ARMAN, which harbor both an RPA1 and an SSB homolog (fig. 2). The function of SSB homologs outside the Crenarchaeota is unknown, as is their possible interaction or division of labor in the taxa that harbor an RPA homolog. We noticed that the SSB homologs of Aigarchaeota and Thermoplasmatales/DHEV2 harbor the flexible C-terminal tail typical of crenarchaeal SSB. In Crenarchaeota, this tail appears to be involved in repair and recombination (Cubeddu and White 2005) (schematically represented by a striped box in fig. 5, for a full alignment see supplementary fig. S1, Supplementary Material online). This tail is absent from the SSB of Thaumarchaeota and Korarchaeota, which harbor an RPA1 homolog (fig. 5; supplementary fig. S1, Supplementary Material online). This may hint at a change in function of SSB in these taxa or even a potential interaction with RPA1.

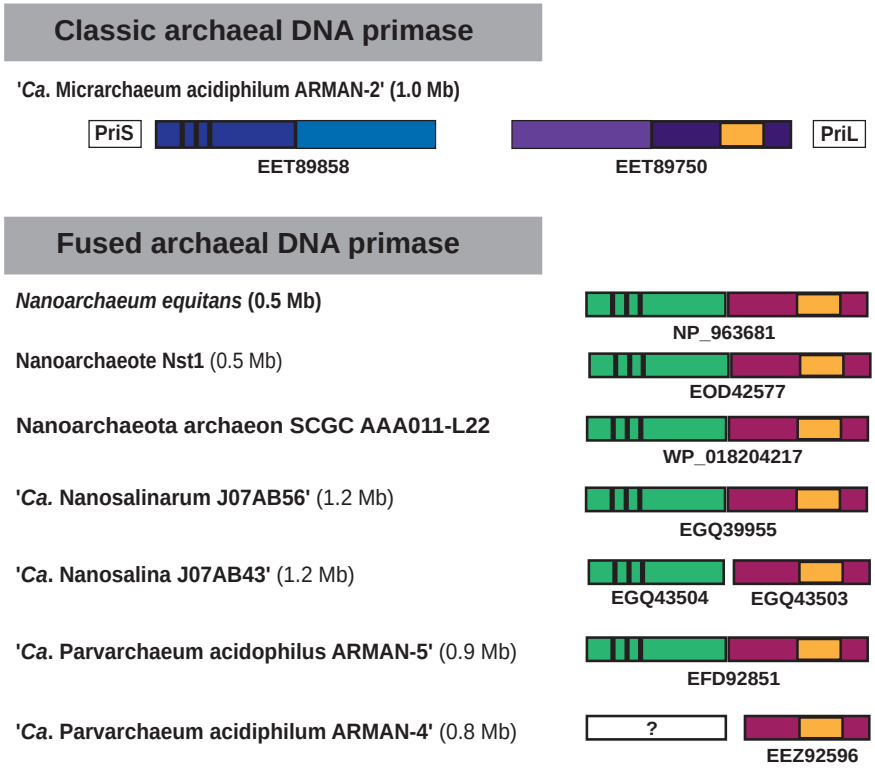


Fig. 6.—Schematic representation of the classic archaeal DNA primase genes encoding for the two subunits PriS and PriL, as opposed to the single genes encoding for fused archaeal primases that we found in some nanosized lineages. The presence of a PriS in *Ca.* Parvarchaeum acidophilum ARMAN-4 is unknown (question mark). The genome sizes are given in parentheses. See text for details.

Indeed, in the genomes of *Candidatus* Parvarchaeum acidophilum ARMAN-4 and *Candidatus* Parvarchaeum acidophilus ARMAN-5′ the gene coding for RPA1 lies next to the gene coding for SSB (supplementary table S1, Supplementary Material online). Phylogenetic analysis of SSB homologs (supplementary fig. S1, Supplementary Material online) suggests that Thermoplasmatales and Aigarchaeota may have acquired their SSB via horizontal gene transfer from Crenarchaeota, an event possibly linked with the loss of the native RPA1 in both lineages. Intriguingly, this putative transfer displays a similar pattern to the one that is likely at the origin of the Cdc6/Orc1-2 of these lineages, as discussed earlier. It is therefore not excluded that both Cdc6/Orc1-2 and RPA1 where transferred together, indicating a possible direct functional linkage of these two components.

Fused Archaeal DNA Primases: A Shared Derived Character for Nanosized Archaea?

Archaeal DNA primases (PriS and PriL) show low sequence similarity with their eukaryotic counterparts and even within Archaea. Most archaea contain a classic primase, made of a catalytic subunit PriS and an accessory subunit PriL (fig. 6). The PriL subunit contains a conserved Fe-S cluster-binding domain that plays an important role in primase activity (Klinge et al. 2007) (fig. 6, yellow box). The activity of PriS lies in an N-terminal catalytic domain with a conserved motif (fig. 6, black bars). It has been previously observed that *Nanoarchaeum equitans* contains a short atypical primase encoded by a single gene, which is composed of a fusion of the catalytic domain of PriS and the Fe–S cluster-binding domain of PriL (Iyer et al. 2005). We identified this same type of primase in the recently sequenced Nanoarchaeote Nst1 (Podar et al. 2012) and in an uncultured nanoarchaeon from a recent single cell genomics survey (Rinke et al. 2013).

Besides Nanoarchaeota, two novel uncultured archaeal lineages characterized by reduced genomes and very small cell sizes have been highlighted recently: a candidate class called Nanohaloarchaea represented by three metagenomic assemblies isolated from a highly saline lake in Australia (Narasingarao et al. 2012), and the Archaeal Richmond Mine Acidophilic Nanoorganisms or ARMAN lineage represented by three metagenomic assemblies isolated from an acidic iron-rich mine in the United States (Baker et al. 2010). Interestingly, we found that *Candidatus* Parvarchaeum acidophilus ARMAN 5 and the nanohaloarchaeon *Candidatus* Nanosalinarum sp. J07AB56 contain a single gene encoding a fused PriS/PriL whose sequences are closely related to that of

N. equitans but are very divergent in comparison to other archaeal primases. The second available nanohaloarchaeum *Candidatus* Nanosalina sp. J07AB43 harbors two adjacent genes encoding for a short primase that clearly align with the other fused primases (fig. 6). *Candidatus* Parvarchaeum acidiphilum ARMAN 4 has a PriL homolog that aligns well with the C-terminal metal binding domain of the short PriL, but appears to lack the N-terminal catalytic PriS domain (fig. 6). However, it is located at the end of a contig in this nonassembled genome, and therefore the presence of the PriS domain cannot be excluded at present. In contrast, *Candidatus* Micrarchaeum acidiphilum ARMAN 2 possesses a classic primase (fig. 6).

It could be argued that these peculiar fused primases arose from evolutionary convergence following genome streamlining in these nanosized lineages. However, the hypothesis of convergence can be excluded because they are related at the sequence level. This leaves two possibilities: either the lineages harboring a fused primase share a common ancestor or the fused primases have replaced the original primases via horizontal gene transfer. Based on phylogenetic analysis of 38 universal protein markers, Rinke et al. (2013) have proposed the existence of a monophyletic superphylum called DPANN whose members would be characterized by small cell and genome sizes and would include the ARMANS, Nanohaloarchaea, and Nanoarchaeota. The sharing of fused primases may appear consistent with the existence of a DPANN clade. However, it is not consistent with *Ca*. Micrarchaeum acidiphilum ARMAN 2 harboring a classical primase. Moreover, the grouping of nanosized archaeal lineages in phylogenetic trees should be interpreted with caution given that robust clustering of fast evolving lineages is a well-known artifact of phylogenetic reconstruction (Gribaldo and Philippe 2002). Indeed, recent ribosomal protein trees support the clustering of *Ca.* Parvarchaeum acidiphilum ARMAN 4, *Ca.* Parvarchaeum acidophilus ARMAN 5 and Nanoarchaeota to the exclusion of *Ca.* Micrarchaeum acidiphilum ARMAN 2 (Brochier-Armanet et al. 2011), and the grouping of Nanohalobacteria with Halobacteriales (Narasingarao et al. 2012).

Alternatively, it may be hypothesized that these fused primases have replaced the original primase via horizontal gene transfer among these lineages, possibly through related integrative elements. Fused DNA primases might be frequent in integrative elements, as suggested by the DNA polymerase/primase recently highlighted in the plasmid pTN2 from *Thermococcus nautilus* (Soler et al. 2010) that harbors a similar PriS/PriL fusion. However, we observe that this fused primase displays no sequence similarity with the primases of nanosized archaea, indicating an independent origin. Moreover, organisms belonging to nanosized lineages thrive in very different environments (hyperthermophilic [Huber et al. 2002], extreme halophilic [Narasingarao et al. 2012], or extreme acidic [Baker et al. 2006]), making the hypothesis of a horizontal gene transfer puzzling. Undoubtedly, more data are needed to clarify the issue and further understand the diversity and evolutionary history of these fascinating lineages.

Acquisition of Bacterial DNA Gyrase: When and How Many Times?

To resolve topological conflicts arising during replication, archaea use a TopoVI that relaxes both positive and negative supercoils. Previous phylogenetic analysis has indicated that bacterial-like DNA gyrases were acquired in a number of euryarchaeota through horizontal gene transfer (Forterre et al. 2007). Because bacterial DNA gyrases actively introduce negative DNA supercoiling, this transfer event likely had a significant impact, changing the overall genome topology and all associated cellular processes, such as the pattern of gene expression (Forterre et al. 2007; Forterre and Gadelle 2009). In most of these euryarchaea, DNA gyrase now coexists with the endogenous TopoVI. In contrast, Thermoplasmatales have lost their original TopoVI and now must solely rely on DNA gyrase for replication and chromosome decatenation (Forterre et al. 2007; Forterre and Gadelle 2009). With the availability of an expanded taxonomic sampling covering more euryarchaeal diversity, we sought to address the timing and number of events that introduced DNA gyrase into this phylum. Consistent with previous reports, we found both DNA gyrase subunits in all genomes from the orders Archaeoglobales, Methanosarcinales, and Halobacteriales (Bergerat et al. 1997; Forterre et al. 2007; Berthon et al. 2008; Forterre and Gadelle 2009). We also identified both subunits in all analyzed genomes of the orders Methanomicrobiales and Methanocellales (which together with Methanosarcinales form the methanogen class II), as well as in DHEV2, uncultured marine group II, and *Ca.* Micrarchaeum acidiphilum ARMAN 2 (fig. 2; supplementary table S1, Supplementary Material online).

Given that these lineages form a late emerging monophyletic cluster in the archaeal phylogeny, and that DNA gyrase is most likely rarely acquired because of its biological consequences, we speculated that this horizontal gene transfer occurred only once at the base of this group. Albeit not completely resolved, a phylogenetic tree of concatenated large and small DNA gyrase subunits shows that archaeal sequences form a monophyletic cluster (fig. 7) supporting a single acquisition of DNA gyrase in these archaea via horizontal gene transfer from an unidentified bacterium. The uncultured marine group II is an exception and likely represents an independent horizontal transfer. However, the weak phylogenetic signal makes this monophyletic group very unstable, as it can be broken up in two clusters depending on the bacterial taxonomic sampling used (not shown). In this case, one cluster corresponds to Halobacteriales and Methanogens class II, and the other to Thermoplasma/DHEV2/Archaeoglobales/ARMAN-2. This would indicate that two independent

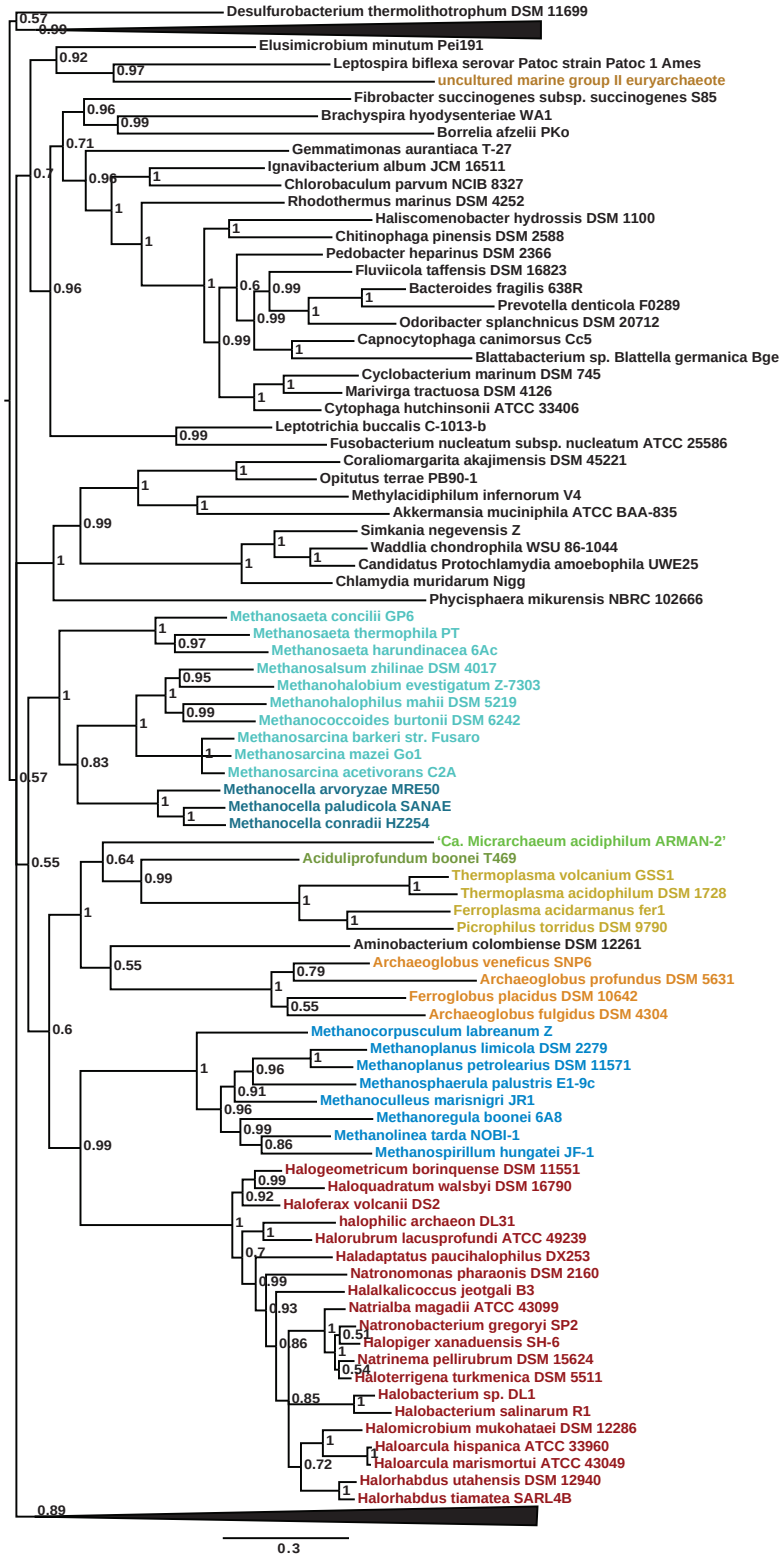


Fig. 7.—Bayesian phylogeny of a concatenation of archaeal DNA gyrase small and large subunits and a selection of bacterial homologs (1,083 amino acid positions). The tree was calculated by MrBayes (MIX model + gamma4). The scale bar represents the average number of substitutions per site. Supports at nodes indicate posterior probabilities. Colors correspond to archaeal orders according to those used in figure 2. The tree is collapsed for clarity. See supplementary table S1 (Supplementary Material online) for accession numbers and taxonomic information.

horizontal gene transfers from bacteria are at the origin of DNA gyrases in the two groups of archaea. However, we speculate that the second transfer would have been possible only because the newly introduced DNA gyrase replaced an already present bacterial-type enzyme. The two alternative scenarios (a single transfer or two successive transfers) remain possible, as statistical tests showed that the data do not reject either of the two topologies ($P > 0.48$ for all tests, see Materials and Methods for details).

DNA gyrase is likely essential in all species that harbor it, suggesting that it may be difficult to lose this enzyme once acquired. We could not find any homologs of DNA gyrase in the genomes of Nanohaloarchaea nor of ARMAN-4 and ARMAN-5 (fig. 2). This may be consistent with an emergence of these lineages prior to the alleged first horizontal gene transfer introducing DNA gyrase in the Thermoplasma/DHEV2/Archaeoglobales/ARMAN-2.

DNA Replication Proteins Harbor a Robust Signal for Archaeal Phylogeny

Fourteen core DNA replication orthologs present in more than 60% of the taxa (PriS, MCM, PCNA, Cdc6/Orc1, DPL, DPS, PolB, TopoVI-A, TopoVI-B, RFC-s, RFC-l, RNaseH, DNA ligase, and FEN-1) were concatenated into a large supermatrix of 4,295 amino acid positions from 129 complete or nearly complete archaeal genomes (keeping only one genome per species, see Materials and Methods). The amount of missing data from the concatenation was analyzed, and except for phyla or orders displaying specific losses or absences (e.g., both small and large subunits of PolD absent in all Crenarchaeota) there are no specific species that are underrepresented (supplementary fig. S2, Supplementary Material online). The phylogeny obtained from this supermatrix (fig. 8) is highly consistent with the previous archaeal phylogenies inferred from transcription and translation components (Matte-Tailliez et al. 2002; Brochier et al. 2004, 2005; Brochier-Armanet et al. 2011). The monophylies of Crenarchaeota, Euryarchaeota, Korarchaeota, and Thaumarcheaota are all recovered with strong support as well as those of all major orders. The phylogeny solidifies the clustering of uncultured marine group II and the DHEV2 representative with the Thermoplasmatales (Brochier-Armanet et al. 2011) and the monophyly of Methanogens class I (i.e., *Methanopyrus kandleri* + Methanobacteriales + Methanococcales) (Bapteste et al. 2005). The robust monophly of Thaumarchaeota and Aigarchaeota observed in the replication tree is in agreement with the proposal that Aigarchaeota represent an early emerging thaumarchaeotal lineage (Brochier-Armanet et al. 2011). Other important points that should be underlined are 1) the emergence of *Acidilobus* within Desulfurococcales, which refutes the recent proposal of the new order Acidilobales (Prokofeva et al. 2009); 2) the clustering of Halobacteriales with Methanogens class II, with a specific grouping of Methanomicrobiales and Halobacteriales; 3) the grouping of Methanogens class II + Halobacteriales with Archaeaoglobales and Thermoplasmatales (fig. 8).

A few differences were observed between the replication phylogeny and the previous trees based on ribosomal proteins (Brochier-Armanet et al. 2011). For example, the robust monophyly of Methanogens class I and Thermococcales, the grouping of Korarchaeota with Thaumarchaeota, and the early emergence of Methanocellales within Methanogens class II (fig. 8). Finally, all of the nanosized archaea (Nanoarchaeota, ARMAN-5, ARMAN-4, and the three Nanohaloarchaea), except for ARMAN-2, form a monophyletic clade that emerges after the divergence of Thermococcales and Methanogens class I (fig. 8). Considering the very fast evolutionary rate of these lineages, it cannot be excluded that this grouping is due to a tree reconstruction artifact. To test this possibility, we created several versions of the concatenated dataset containing different combinations of taxa (i.e., we removed all nanosized lineages from the concatenation and reintroduced them one by one) and we recoded the amino acid supermatrix using Dayhoff 6 and Dayhoff4 recoding schemes, a procedure known to alleviate certain artifacts due to fast evolutionary rates (Delsuc et al. 2005). However, no major differences were observed.

Discussion

Dynamic History of a Key Cellular System

Through our precise identification and phylogenetic analysis of core replication components, we reconstructed the global evolutionary history of the DNA replication machinery in Archaea. In particular, we inferred the presence of a complete and modern type machinery in the LACA (table 1). The LACA would have harbored two Cdc6/Orc1 paralogs, two GINS paralogs (GIN23 and GIN51), and one homolog each of the MCM helicase, the sliding clamp PCNA and its loader RFC with both subunits, the polymerase PolB, the archaeal primase with both subunits, the Okazaki fragment processing flap endonuclease Fen1 and RNaseH II, the ATP-dependent DNA ligase, and the topoisomerase Topo VI with both subunits. Although the involvement of DnaG in replication is dubious, this protein must have an important and conserved role because it is universally present in archaea. Moreover, the phylogeny is robustly supported and is strikingly consistent with the archaeal species tree (not shown). This indicates that the presence of DnaG in archaea is not due to horizontal gene transfer from bacteria but instead was harbored by the LACA and was subsequently strictly vertically inherited up to present. For the few remaining components (PolD, SSB, and RPA1), their presence in the LACA strongly depends on the root of the archaeal tree, which is presently unclear (Brochier-Armanet et al. 2011; table 1). TopoIB represents a special case because its presence in LACA relies on whether

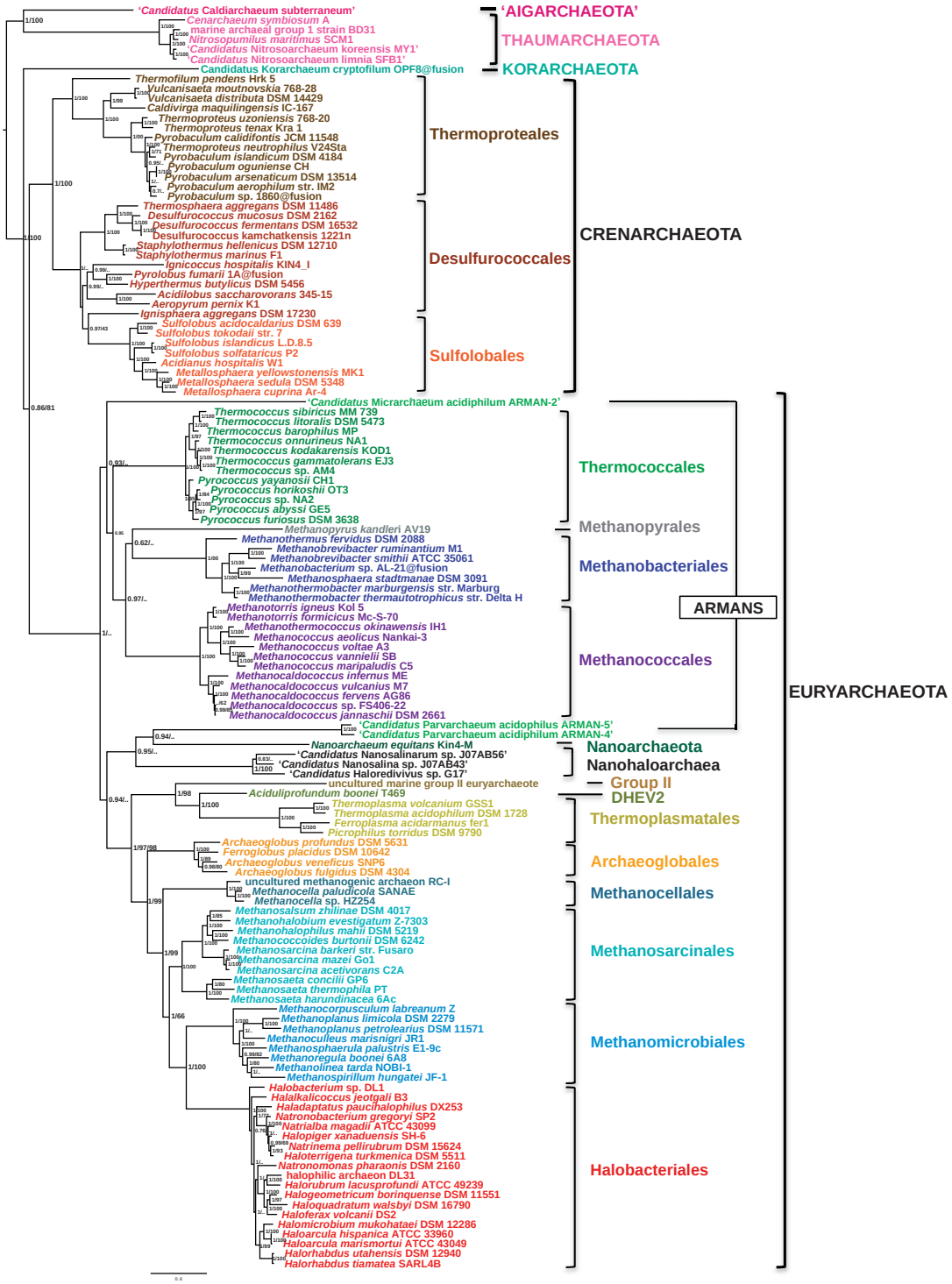


Fig. 8.—Bayesian phylogeny of a concatenated data set of 14 replication components (4,295 amino acid positions). The tree was calculated by Phylobayes (CAT + GTR + gamma4). The scale bar represents the average number of substitutions per site. Values at nodes represent posterior probabilities and BV based on 100 resamplings of the original data set calculated by PhyML (LG model + gamma4), when the same node was recovered.

Table 1
Inferred Components of DNA Replication in the LACA and in the Ancestor of Each Major Phylum

LACA	Thaumarchaeota/Aigarchaeota	Korarchaeaota	Crenarchaeota	Euryarchaeota
Cdc6/Orc1-1	**Cdc6/Orc1-1**		**Cdc6/Orc1-1**	**Cdc6/Orc1-1**
Cdc6/Orc1-2	**Cdc6/Orc1-2**	**Cdc6/Orc1-2**	**Cdc6/Orc1-2**	**Cdc6/Orc1-2**
MCM	**MCM**	**MCM**	**MCM**	**MCM**
GINS51	**GINS51**	**GINS51**	**GINS51**	**GINS51**
GINS23	**GINS23**	**GINS23**	**GINS23**	**GINS23**
RPA1	**RPA1**	**RPA1**		**RPA1**
SSB	**SSB**	**SSB**	**SSB**	
PolB	**PolB**	**PolB (X2)**	**PolB (X2)**	**PolB**
PolD-L/S	PolD-L/S	PolD-L/S		DP-L/S
RFC-S/L	**RFC-S/L**	**RFC-S/L**	**RFC-S/L**	**RFC-S/L**
PCNA	**PCNA**	**PCNA**	**PCNA (X2)**	**PCNA**
Pri-S/L	**Pri-S/L**	**Pri-S/L**	**Pri-S/L**	**Pri-S/L**
RNaseH II	**RNaseH II**	**RNaseH II**	**RNaseH II**	**RNaseH II**
FEN-1	**FEN-1**	**FEN-1**	**FEN-1**	**FEN-1**
ATP DNA ligase	**ATP DNA ligase**	**ATP DNA ligase**	**ATP DNA ligase**	**ATP DNA ligase**
TopoIV-A/B	**TopoVI-A/B**	**TopoVI-A/B**	**TopoVI-A/B**	**TopoVI-A/B**
TopoIB	**TopoIB**			

Root-dependent components
- **Thaumarchaeota/"Aigarchaeota"** → PolD-L/S, **RPA, SSB, TopoIB**
- **Korarchaeota** → PolD-L/S, **RPA, SSB**
- **Crenarchaeota** → **SSB**
- **Euryarchaeota** → PolD-L/S, **RPA**

Note.—Additional components that would have been present in the LACA according to a rooting in each of the four major phyla are indicated. Components shown in bold have homologs in eukaryotes and those shown in gray are root dependent.

Archaea and Eukaryotes are sister lineages, a currently unsettled matter (see below).

The core components inferred in the ancestor of each phylum are overall very similar (table 1). Major differences appear most evident in the ancestor of Crenarchaeota, with a number of specific characters such as the presence of at least two PCNA and PolB paralogs, the absence of PolD, and the presence of SSB but not RPA. The subsequent evolutionary history of the DNA replication machinery appears very dynamic. In particular, the absence in any present day lineage of a component inferred to have been present in the LACA has to be interpreted as a consequence of gene loss. We observed many independent gene losses frequently involving one of two ancestral paralogs, for example, Cdc6/Orc1 and GINS. A similar phenomenon of gene loss has been observed in archaeal ribosomes, which appear to have experienced independent losses of components in different lineages (Desmond et al. 2010; Yutin et al. 2012), as well as on a global genomic scale (Csuros and Miklos 2009). Our results are therefore consistent with a growing consensus on a complex LACA (Makarova et al. 2007; Csuros and Miklos 2009; Wolf et al. 2011).

However, there is not a unique trend toward gene loss in regard to the replication machinery. We highlighted the occurrence of a number of component accretions throughout archaeal diversification. Examples are the multiplication of RPA copies in Euryarchaeota and the expansion of the MCM family in Methanococcales. These are both due to gene duplication of core components and acquisition of additional shell components from extrachromosomal elements. Some of these events also led to increased complexity of multiprotein machineries involved in replication. For example, whereas most archaeal RFC are composed of four identical RFC small subunits (RFC-S) and one RFC large subunit (RFC-L) (Barry and Bell 2006), some species contain two RFC-S homologs (RFC-S1 and RFC-S2). In these cases, three RFC-S1 subunits and one RFC-S2 subunit assemble with RFC-L to form the pentameric RFC complex (Chen et al. 2005). Similarly, Crenarchaeota contain two or three copies of PCNA that have arisen from gene duplication and form a heterotrimeric structure in which each subunit has specific binding functions to different replication proteins (Grabowski and Kelman 2003; Barry and Bell 2006). It is noteworthy that, according to current knowledge, these accretions of components in multisubunit complexes appear to be due to gene duplication rather than integration of shell components or horizontal gene transfer. However, it will be very interesting to study if extra copies arising from integrative elements may, in some instances, replace the native component or integrate complexes made of core components.

As opposed to the high dynamics of shell components, horizontal gene transfers involving core components appear to be relatively rare. A few cases can been seen which are

consistent with known exchanges amongst archaea thriving in the same environments such as from Crenarchaeota to Thermoplasmatales. Moreover, we show that horizontal gene transfer events involving bacterial replication components, albeit rare, have occurred during archaeal diversification. For example, other than the previously discussed case of DNA gyrase, we observed a single horizontal gene transfer introducing a bacterial-type NAD+-dependent DNA ligase in the ancestor of Halobacteriales (not shown), which may have in some cases replaced the native archaeal/eukaryal ATP- dependent DNA ligase (fig. 2; supplementary table S1, Supplementary Material online).

Why So Many DNA Replication Components in Extracellular Elements?

An evident phenomenon affecting archaeal DNA replication is the presence of many divergent extra copies particularly those involved in the first steps of replication, such as Cdc6/Orc1, MCM, RPA1, and PolB (fig. 2). Moreover, different archaeal viruses, proviruses, and plasmids are known to encode homologs of Cdc6/Orc1 and MCM (Pagaling et al. 2006; Yamashiro et al. 2006; Krupovic, Forterre, et al. 2010). Similarly, an archaeal homolog of eukaryotic Ctd1 called WhiP was recently identified in the integrative element that contributed the third origin of replication in Sulfolobales (Robinson and Bell 2007). Precise identification of all extra copies of replication components that reside in integrative elements in archaeal genomes requires extensive work and is beyond the scope of this article. Nevertheless, our study strongly suggests that extrachromosomal elements have had an impact on the evolution of the archaeal DNA replication machinery and actively modeled its composition, both by picking up and transferring components to and from cellular genomes. Considering the small number and taxonomic coverage of viral sequences presently available in public databases (supplementary table S3, Supplementary Material online) our analysis suggests that the world of archaeal extrachromosomal entities may be particularly enriched in genes encoding for replication proteins. Moreover, the presence of highly divergent and related components in Thermococcales and Methanococcales, such as their DNA primase and the RPA three-gene cluster, may indicate potential avenues of gene sharing through a common pool of plasmids and viruses (Soler et al. 2010).

Archaeal plasmids and viruses rarely encode components of the transcription machinery and, to our knowledge, no translation components. The targeting of DNA replication by virus/plasmid entities to hijack the host machinery provides a strong advantage and is a well-known phenomenon. However, it is much less known that, upon viral/plasmid integration, many DNA replication proteins of extrachromosomal origin became residents (either transient or permanent) of cellular genomes. This can confuse the phylogeny of these proteins if the difference between real and false cellular genes is not correctly assessed. Finally, it will be interesting to carry out a similar global analysis in Bacteria and Eukaryotes to understand whether this phenomenon is particularly evident in the Archaea or is a more general trend.

An Archaeon at the Origin of Eukaryotes?

A recent analysis inferred the core DNA replication components in the last eukaryotic common ancestor (Aves et al. 2012). Aves et al. predicted that LECA (the Last Eukaryotic Common Ancestor) would have possessed all of the components that we have inferred in the archaeal ancestor, with the exclusion of PolD (table 1). This is coherent with the classical scenario indicated by ancient paralogous protein pairs where Archaea are a sister lineage to Eukaryotes (Gogarten et al. 1989; Iwabe et al. 1989; Gribaldo and Cammarano 1998). In contrast, recent analyses support the emergence of Eukaryotes from within the archaeal radiation (Cox et al. 2008; Foster et al. 2009; Guy and Ettema 2011; Williams et al. 2012; Alvarez-Ponce et al. 2013; Lasek-Nesselquist and Gogarten 2013). In particular, a deep branching within a cluster composed of Thaumarchaeota, Aigarcharchaeota, Korarchaeota, and Crenarchaeota seems to be predominant, and would be consistent with an apparent enrichment of eukaryotic-like characters in these phyla with respect to Euryarchaeota (Guy and Ettema 2011).

Unfortunately, archaeal DNA replication components are very divergent from their eukaryotic homologs, preventing the reconstruction of reliable phylogenies to test the evolutionary relationship between these two domains of life. Nonetheless, our reconstruction of the evolution of the DNA replication machinery along archaeal diversification sheds new light on this issue. The absence of eukaryotic core components from the replication machinery of the ancestor of a given archaeal lineage would exclude the emergence of eukaryotes from one of its members (unless invoking an extremely unparsimonious scenario where the component was independently lost in all members of the lineage but only kept in the one that would have given rise to eukaryotes). By this rationale, we can exclude the emergence of eukaryotes from within the radiation of any of the major archaeal phyla. For example, the lack of GINS 23 and SSB in the ancestor of Euryarchaeota (table 1) would exclude an emergence of Eukaryotes from within this phylum. Similarly, the absence of RPA in the ancestor of Crenarchaeota would also exclude an emergence of Eukaryotes from within the radiation of this phylum. Furthermore, an origin of Eukaryotes from within Crenarchaeota also seems unlikely given the presence of a peculiar heterotrimeric PCNA derived from an ancestral homotrimeric structure. In this situation, the complex would have reverted back into the homo-trimeric form observed in present day eukaryotes, an improbable scenario. Among the four major archaeal phyla, none seem to be particularly enriched in characters shared with Eukaryotes, perhaps with the

exception of Thaumarchaeota (table 1). However, this kind of argument should not be used to infer a specific evolutionary link between Eukaryotes and Thaumarchaeota. In fact, gene loss appears to be a common process that has substantially affected DNA replication, along with many other cellular processes during the diversification of Archaea.

Irrespective of the different evolutionary scenarios for the origin of eukaryotes, our study indicates that the ancestral replication machinery of these two domains of life was very similar (table 1). Therefore, our analysis provides a key starting point for understanding the subsequent evolutionary history of the eukaryotic DNA replication machinery. For example, specific gene duplications would have occurred in the eukaryotic ancestor giving rise to paralogous components such as MCM(2-7) and GINS (Sld5, Psf1, Psf2, and Psf3), or the addition of multiple nonhomologous subunits like ORC(1-6), RPA(70, 34, 14), and RNaseH2 (A, B, C). A few components with homology to archaea are not involved in replication in eukaryotes, and it can be speculated that they were reassigned to other cellular functions. For example, most eukaryotes encode a homolog of the A subunit of archaeal TopoVI called Spo11 (Bergerat et al. 1997), which is not involved in replication but instead induces the double stand breaks that initiate meiotic recombination (Bergerat et al. 1997; Martini and Keeney 2002). In contrast, members of the Archaeplastida (land plants and green, red, and glaucocystophyte algae) possess homologs of both subunits (A and B) of archaeal TopoVI, where they combine into a functional enzyme that appears to play a role in DNA endoreduplication, a process required for polyploidization (Hartung and Puchta 2001). The presence of both subunits in some protist lineages such as Kinetoplastids opens up the possibility that a functional TopoVI was present in the ancestor of Eukaryotes (Malik et al. 2007), and was subsequently lost in most lineages. The same logic applies to the archaeal-like SSB that we identified in representatives of most eukaryotic phyla (supplementary table S4, Supplementary Material online), where it may have an important and possibly ancestral role in (Robbins et al. 2005; Richard et al. 2008; Shi et al. 2012).

On the other hand, a few of the core components of eukaryotic DNA replication are not present in Archaea and therefore would have arisen specifically in the lineage leading to Eukaryotes. This is the case of DNA pol-α and the B-subunit of the primase complex, topoisomerase IIA, and the FACT complex (Aves et al. 2012). The emergence of DNA pol-α is particularly fascinating. In Bacteria and Archaea the RNA primer is directly extended by the main replicative DNA polymerase, but in Eukaryotes Pol-α adds 10-30 nt DNA stretches to the RNA primer, and only then does the complex hand-off to the main replicative DNA polymerase (DePamphilis and Bell 2010). These 10–30 nucleotides therefore need to be removed during Okazaki fragment maturation (Stillman 2008), raising the question of the origin of this polymerase (Forterre 2013). The future availability of both genomic and experimental data from a larger fraction of eukaryotic diversity will surely allow a better understanding of the diversity and evolutionary history of DNA replication in this Domain of Life.

Finally, further exploration of diversity and function of archaeal replication may uncover unsuspected links with their eukaryotic cousins. It is not excluded that some of these components/functions were ancestrally present in the Archaea and subsequently lost.

Increasing the Conserved Phylogenomic Core for Archaea

In the past, we have shown that the components of the transcription and translation machineries contain a consistent and robust phylogenetic signal that reflects the history of archaeal diversification (Brochier et al. 2005; Gribaldo and Brochier-Armanet 2006; Gribaldo and Brochier 2009). The third major informational system that remained to be analyzed was the DNA replication machinery. However, the complex evolutionary history of DNA replication components and the occurrence of multiple highly divergent copies of unclear origin rendered the application of phylogenomic approaches to this cellular machinery particularly challenging. Our precise identification of orthologs has now made it possible to perform such analysis, and indeed, archaeal DNA replication carries a robust phylogenetic signal that is largely consistent with that of the two other informational systems. Moreover, reconstructing the evolution of DNA replication brings novel information to the archaeal phylogeny. It consolidates important relationships such as Aigarchaeota as a sister lineage of Thaumarchaeota, and the monophyly of Methanogens class I. The clustering of Thermococcales and Methanococcales merits further study, because it is not apparent in trees based on ribosomal proteins or transcription components (Matte-Tailliez et al. 2002; Brochier et al. 2004, 2005; Brochier-Armanet et al. 2011) but is in agreement with some common peculiarities in their replication machinery (see above). Therefore, this relationship in the tree based on replication components may reflect a bias introduced by undetected independent transfers from related mobile elements, viruses, and/or plasmids. The phylogenetic placement of nanosized archaea remains unclear. Their grouping in our trees may indicate common ancestry, but only partially supports the recently proposed DPANN cluster (Rinke et al. 2013). In fact, one member of the ARMANS (*Ca.* Micrarchaeum acidiphilum ARMAN-2) does not cluster with the other nanosized lineages, consistent with the analysis of ribosomal proteins (Brochier-Armanet et al. 2011). This is congruous with a number of additional observations: the absence of a fused primases (figs. 2 and 6), the presence of bacterial DNA gyrase (figs. 2 and 5), the presence of an SSB with an N-terminal tail (figs. 2 and 5; supplementary fig. S1, Supplementary Material online), and the absence of RPA. Targeted phylogenomic analyses combined with novel

genomic data from these peculiar lineages will bring important insights into this issue.

It is important to highlight that a detailed analysis such as ours allows for the identification of novel phylogenetic markers that would most likely be discarded by more automated analyses. A commonly used approach to build concatenated data sets for phylogenetic analysis is to choose genes present in a single copy in all (or nearly all) genomes to avoid problems arising from a mixture of orthologs and paralogs. Such a strategy drastically reduces the number of usable markers, especially when dealing with deep evolutionary relationships. In addition, this type of strategy biases our understanding of prokaryotic evolution, by underrepresenting vertical inheritance (tree-like process) with respect to horizontal gene transfers (net or forest-like process) (Dagan and William Martin 2006). Had we applied such strategy, we would have essentially discarded all replication components. Instead, we have shown that reliable phylogenetic information can be extracted even from proteins that are not universally distributed or exist in multiple paralogs—allowing the tree to appear from the forest. Even if a strict core of vertically inherited genes might be limited, our results clearly demonstrate the existence of a soft core of cellular components involved in different processes whose genes have similar histories and can therefore be used to trace back the evolutionary relationships among the organisms that carry them (Gribaldo and Brochier-Armanet 2006; Gribaldo and Brochier 2009). It is likely that this soft phylogenomic core is much richer than usually assumed.

Conclusions

The emergence of novel techniques grants rapid access to an ever-wider fraction of microbial diversity, both from a genomic and functional point of view. In this context, the integration of evolutionary studies will be of primary importance, not only to provide key information for experimental work but also to uncover general trends in the global evolutionary history of the largest fraction of the biosphere.

Supplementary Material

Supplementary figures S1 and S2 and tables S1–S4 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

The authors acknowledge the article by Makarova and Koonin (2013) that was published while this manuscript was under review and which presents a comparison of DNA replication in Archaea and Eukaryotes. The authors thank the PRABI (Pôle Rhône-Alpes de Bioinformatique) for providing computing facilities. K.R. is a scholar from the Pasteur–Paris University (PPU) International PhD program and received a stipend from the Paul W. Zuccaire Foundation. C.B.A. is member of the Institut Universitaire de France. This work was supported by the Investissement d'Avenir grant "Ancestrome" (ANR-10- BINF-01-01).

Literature Cited

Altschul SF, et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25: 3389–3402.

Alvarez-Ponce D, Lopez P, Bapteste E, McInerney JO. 2013. Gene similarity networks provide tools for understanding eukaryote origins and evolution. Proc Natl Acad Sci U S A. 110:E1594–E1603.

Aravind L, Koonin EV. 1998. Phosphoesterase domains associated with DNA polymerases of diverse origins. Nucleic Acids Res. 26:3746–3752.

Aves SJ, Liu Y, Richards TA. 2012. Evolutionary diversification of eukaryotic DNA replication machinery. Subcell Biochem. 62:19–35.

Baker BJ, et al. 2006. Lineages of Acidophilic Archaea Revealed by Community Genomic Analysis. Science 314:1933–1935.

Baker BJ, et al. 2010. Enigmatic, ultrasmall, uncultivated Archaea. Proc Natl Acad Sci U S A. 107:8806–8811.

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

Barry ER, Bell SD. 2006. DNA replication in the archaea. Microbiol Mol Biol Rev. 70:876–887.

Bauer RJ, Graham BW, Trakselis MA. 2013. Novel interaction of the bacterial-Like DnaG primase with the MCM helicase in archaea. J Mol Biol. 425:1259–1273.

Beattie TR, Bell SD. 2011. Molecular machines in archaeal DNA replication. Curr Opin Chem Biol. 15:614–619.

Bell SD. 2011. DNA replication: archaeal oriGINS. BMC Biol. 9:36.

Bergerat A, et al. 1997. An atypical topoisomerase II from Archaea with implications for meiotic recombination. Nature 386:414–417.

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

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

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

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

Buckley TRT, Simon CC, Shimodaira HH, Chambers GKG. 2001. Evaluating hypotheses on the origin and evolution of the New Zealand alpine cicadas (Maoricicada) using multiple-comparison tests of tree topology. Mol Biol Evol. 18:223–234.

Cann I, Komori K, Toh H, Kanai S, Ishino Y. 1998. A heterodimeric DNA polymerase: evidence that members of Euryarchaeota possess a distinct DNA polymerase. Proc Natl Acad Sci U S A. 95: 14250–14255.

Chen YH, et al. 2005. Biochemical and mutational analyses of a unique clamp loader complex in the archaeon *Methanosarcina acetivorans*. J Biol Chem. 280:41852–41863.

Chia N, Cann I, Olsen GJ. 2010. Evolution of DNA replication protein complexes in eukaryotes and Archaea. PLoS One 5:e10866.

Cox CJC, Foster PGP, Hirt RPR, Harris SRS, Embley TMT. 2008. The archaebacterial origin of eukaryotes. Proc Natl Acad Sci U S A. 105: 20356–20361.

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 Evol Biol. 10: 210.
Csuros M, Miklos I. 2009. Streamlining and large ancestral genomes in Archaea inferred with a phylogenetic birth-and-death model. Mol Biol Evol. 26:2087–2095.
Cubeddu L, White MF. 2005. DNA damage detection by an archaeal single-stranded DNA-binding protein. J Mol Biol. 353:10–10.
Dagan T, Martin W. 2006. The tree of one percent. Genome Biol. 7:118.
Delsuc FF, Brinkmann HH, Philippe HH. 2005. Phylogenomics and the reconstruction of the tree of life. Nat Rev Genet. 6:361–375.
DePamphilis ML, Bell SD. 2010. Genome duplication. London and New York: Garland Publications.
Desmond E, Brochier-Armanet C, Forterre P, Gribaldo S. 2010. On the last common ancestor and early evolution of eukaryotes: reconstructing the history of mitochondrial ribosomes. Res Microbiol. 162:18–18.
Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.
Felsenstein J. 1985. Confidence limits on phylogenies: an approach using the Bootstrap. Evolution 39:783–791.
Forterre P. 2013. Why are there so many diverse replication machineries? J Mol Biol. 425:4714–4726.
Forterre P, Gribaldo S, Gadelle D, Serre M-C. 2007. Origin and evolution of DNA topoisomerases. Biochimie. 89:427–446.
Forterre PP, Gadelle DD. 2009. Phylogenomics of DNA topoisomerases: their origin and putative roles in the emergence of modern organisms. Nucleic Acids Res. 37:679–692.
Foster PG, Cox CJ, Embley TM. 2009. The primary divisions of life: a phylogenomic approach employing composition-heterogeneous methods. Philos Trans R Soc Lond B Biol Sci. 364:2197–2207.
Fukui T, et al. 2005. Complete genome sequence of the hyperthermophilic archaeon *Thermococcus kodakaraensis* KOD1 and comparison with *Pyrococcus* genomes. Genome Res. 15:352–363.
Fütterer OO, et al. 2004. Genome sequence of *Picrophilus torridus* and its implications for life around pH 0. Proc Natl Acad Sci U S A. 101: 9091–9096.
Gogarten JP, et al. 1989. Evolution of the vacuolar H+-ATPase: implications for the origin of eukaryotes. Proc Natl Acad Sci U S A. 86: 6661–6665.
Goldman N, Anderson JP, Rodrigo AG. 2000. Likelihood-based tests of topologies in phylogenetics. Syst Biol. 49:652–670.
Grabowski B, Kelman Z. 2003. Archeal DNA replication: eukaryal proteins in a bacterial context. Annu Rev Microbiol. 57:487–516.
Gribaldo S, Brochier C. 2009. Phylogeny of prokaryotes: does it exist and why should we care? Res Microbiol. 160:513–521.
Gribaldo S, Brochier-Armanet C. 2006. The origin and evolution of Archaea: a state of the art. Philos Trans R Soc Lond B Biol Sci. 361: 1007–1022.
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. J Mol Evol. 47:508–516.
Gribaldo SS, Philippe HH. 2002. Ancient phylogenetic relationships. Theor Popul Biol. 61:391–408.
Guindon S, et al. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 59:307–321.
Guy L, Ettema TJ. 2011. The archaeal 'TACK' superphylum and the origin of eukaryotes. Trends Microbiol. 19:8–8.
Hartman ALA, et al. 2009. The complete genome sequence of *Haloferax volcanii* DS2, a model archaeon. PLoS One 5:e9605–e9605.
Hartung FF, Puchta HH. 2001. Molecular characterization of homologues of both subunits A (SPO11) and B of the archaebacterial topoisomerase 6 in plants. Gene 271:81–86.
Hawkins M, Malla S, Blythe MJ, Nieduszynski CA, Allers T. 2013. Accelerated growth in the absence of DNA replication origins. Nature 1–16.
Higashibata H, Kikuchi H, Kawarabayasi Y, Matsui I. 2003. Helicase and nuclease activities of hyperthermophile *Pyrococcus horikoshii* Dna2 inhibited by substrates with RNA segments at 5′-end. J Biol Chem. 278:15983–15990.
Hou L, Klug G, Evguenieva-Hackenberg E. 2013. The archaeal DnaG protein needs Csl4 for binding to the exosome and enhances its interaction with adenine-rich RNAs. RNA Biol. 10: 415–424.
Huber H, Hohn MJ, Rachel R, Fuchs T, Wimmer VC. 2002. A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont. Nature 417:63–67.
Hurvich CM, Tsai C-L. 1989. Regression and time series model selection in small samples. Biometrika 72:297–307.
Iwabe NN, Kuma KK, Hasegawa MM, Osawa SS, Miyata TT. 1989. Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. Proc Natl Acad Sci U S A. 86:9355–9359.
Iyer LM, Koonin EV, Leipe DD, Aravind L. 2005. Origin and evolution of the archaeo-eukaryotic primase superfamily and related palm-domain proteins: structural insights and new members. Nucleic Acids Res. 33: 3875–3896.
Jobb G, Haeseler von A, Strimmer K. 2004. TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics. BMC Evol Biol. 4:18.
Johnson LS, Eddy SR, Portugaly E. 2010. Hidden Markov model speed heuristic and iterative HMM search procedure. BMC Bioinformatics 11:431.
Kelly TJT, Simancek PP, Brush GSG. 1998. Identification and characterization of a single-stranded DNA-binding protein from the archaeon *Methanococcus jannaschii*. Proc Natl Acad Sci U S A 95: 14634–14639.
Kerkhoven R, van Enckevort FHJ, Boekhorst J, Molenaar D, Siezen RJ. 2004. Visualization for genomics: the Microbial Genome Viewer. Bioinformatics 20:1812–1814.
Kerr ID, et al. 2003. Insights into ssDNA recognition by the OB fold from a structural and thermodynamic study of Sulfolobus SSB protein. EMBO J. 22:2561–2570.
Kishino H, Takashi M, Hasegawat M. 1990. Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. J Mol Evol. 31: 151–160.
Kishino HH, Hasegawa MM. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. J Mol Evol. 29: 170–179.
Klinge S, Hirst J, Maman JD, Krude T, Pellegrini L. 2007. An iron-sulfur domain of the eukaryotic primase is essential for RNA primer synthesis. Nat Struct Mol Biol. 14:875–877.
Komori KK, Ishino YY. 2001. Replication protein A in *Pyrococcus furiosus* is involved in homologous DNA recombination. J Biol Chem. 276: 25654–25660.
Krupovic M, Forterre P, Bamford DH. 2010. Comparative analysis of the mosaic genomes of tailed archaeal viruses and proviruses suggests common themes for virion architecture and assembly with tailed viruses of bacteria. J Mol Biol. 397:17–17.
Krupovic M, Gribaldo S, Bamford DH, Forterre P. 2010. The evolutionary history of archaeal MCM helicases: a case study of vertical evolution combined with hitchhiking of mobile genetic elements. Mol Biol Evol. 27:2716–2732.
Lartillot N, Lepage T, Blanquart S. 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics 25:2286–2288.

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.
Makarova KS, Koonin EV. 2013. Archaeology of eukaryotic DNA replication. Cold Spring Harb Perspect Biol. 5:a012963.
Makarova KS, Koonin EV, Kelman Z. 2012. The CMG (CDC45/RecJ, MCM, GINS) complex is a conserved component of the DNA replication system in all archaea and eukaryotes. Biol Direct. 7:7.
Makarova KS, Sorokin AV, Novichkov PS, Wolf YI, Koonin EV. 2007. Clusters of orthologous genes for 41 archaeal genomes and implications for evolutionary genomics of archaea. Biol Direct. 2:33.
Malik S-BS, Ramesh MAM, Hulstrand AMA, Logsdon JMJ. 2007. Protist homologs of the meiotic Spo11 gene and topoisomerase VI reveal an evolutionary history of gene duplication and lineage-specific loss. Mol Biol Evol. 24:2827–2841.
Martin IV, MacNeill SA. 2002. ATP-dependent DNA ligases. Genome Biol. 3: REVIEWS3005.
Martini EE, Keeney SS. 2002. Sex and the single (double-strand) break. Mol Cell. 9:700–702.
Matte-Tailliez O, Brochier C, Forterre P, Philippe H. 2002. Archaeal phylogeny based on ribosomal proteins. Mol Biol Evol. 19:631–639.
McGeoch AT, Bell SD. 2008. Extra-chromosomal elements and the evolution of cellular DNA replication machineries. Nat Rev Mol Cell Biol. 9: 569–574.
Meselson M, Stahl FW. 1958. The replication of DNA in *Escherichia coli*. Proc Natl Acad Sci U S A. 44:671–682.
Narasingarao P, et al. 2012. De novo metagenomic assembly reveals abundant novel major lineage of Archaea in hypersaline microbial communities. ISME J. 6:81–93.
Pagaling E, et al. 2006. Sequence analysis of an Archaeal virus isolated from a hypersaline lake in Inner Mongolia, China. BMC Genomics 8: 410.
Pan M, et al. 2013. *Thermococcus kodakarensis* has two functional PCNA homologs but only one is required for viability. Extremophiles 17: 453–461.
Pan M, Santangelo TJ, Li Z, Reeve JN, Kelman Z. 2011. *Thermococcus kodakarensis* encodes three MCM homologs but only one is essential. Nucleic Acids Res. 39:9671–9680.
Paytubi S, et al. 2012. Displacement of the canonical single-stranded DNA-binding protein in the Thermoproteales. Proc Natl Acad Sci U S A. 109: E398–E405.
Philippe H. 1993. MUST, a computer package of management utilities for sequences and trees. Nucleic Acids Res. 21:5264–5272.
Podar M, et al. 2012. Insights into archaeal evolution and symbiosis from the genomes of a nanoarchaeon and its inferred crenarchaeal host from Obsidian Pool, Yellowstone National Park. Biol Direct. 8:9.
Prokofeva MI, et al. 2009. Isolation of the anaerobic thermoacidophilic crenarchaeote *Acidilobus saccharovorans* sp. nov. and proposal of Acidilobales ord. nov., including Acidilobaceae fam. nov. and Caldisphaeraceae fam. nov. Int J Syst Evol Microbiol. 59:3116–3122.
Richard DJD, et al. 2008. Single-stranded DNA-binding protein hSSB1 is critical for genomic stability. Nature 453:677–681.
Rinke C, et al. 2013. Insights into the phylogeny and coding potential of microbial dark matter. Nature 499:431–437.
Robbins JB, et al. 2005. The euryarchaeota, nature's medium for engineering of single-stranded DNA-binding proteins. J Biol Chem. 280: 15325–15339.
Robbins JBJ, et al. 2004. Functional analysis of multiple single-stranded DNA-binding proteins from *Methanosarcina acetivorans* and their effects on DNA synthesis by DNA polymerase BI. J Biol Chem. 279: 6315–6326.
Robinson NP, Bell SD. 2007. Extrachromosomal element capture and the evolution of multiple replication origins in archaeal chromosomes. Proc Natl Acad Sci U S A. 104:5806–5811.
Ronquist F, et al. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst Biol. 61: 539–542.
Samson RY, et al. 2013. Specificity and functionof archaeal DNA replication initiator proteins. Cell Rep. 3:485–496.
Shi W, et al. 2012. Essential developmental, genomic stability, and tumour suppressor functions of the mouse orthologue of hSSB1/NABP2. PLoS Genet. 9:e1003298–e1003298.
Shimodaira H. 2002. An approximately unbiased test of phylogenetic tree selection. Syst Biol. 51:492–508.
Shimodaira H, Hasegawa M. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol Biol Evol. 16: 1114–1116.
Skowyra A, MacNeill SA. 2012. Identification of essential and non-essential single-stranded DNA-binding proteins in a model archaeal organism. Nucleic Acids Res. 40:1077–1090.
Soler N, et al. 2010. Two novel families of plasmids from hyperthermophilic archaea encoding new families of replication proteins. Nucleic Acids Res. 38:5088–5104.
Stillman B. 2008. DNA polymerases at the replication fork in eukaryotes. Mol Cell. 30:259–260.
Stroud A, Liddell S, Allers T. 2012. Genetic and biochemical identification of a novel single-stranded DNA-binding complex in *Haloferax volcanii*. Front Microbiol. 3:224–224.
Szklarczyk D, et al. 2011. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Res. 39:D561–D568.
Vallenet D, et al. 2005. MaGe: a microbial genome annotation system supported by synteny results. Nucleic Acids Res. 34:53–65.
Wadsworth RIR, White MFM. 2001. Identification and properties of the crenarchaeal single-stranded DNA binding protein from *Sulfolobus solfataricus*. Nucleic Acids Res. 29:914–920.
Wilkinson A, Day J, Bowater R. 2001. Bacterial DNA ligases. Mol Microbiol. 40:1241–1248.
Williams TAT, Foster PGP, Nye TMWT, Cox CJC, Embley TMT. 2012. A congruent phylogenomic signal places eukaryotes within the Archaea. Proc Biol Sci. 279:4870–4879.
Wolf YIY, Makarova KSK, Yutin NN, Koonin EVE. 2011. Updated clusters of orthologous genes for Archaea: a complex ancestor of the Archaea and the byways of horizontal gene transfer. Biol Direct. 7:46–46.
Yamashiro KK, Yokobori S-IS, Oshima TT, Yamagishi AA. 2006. Structural analysis of the plasmid pTA1 isolated from the thermoacidophilic archaeon *Thermoplasma acidophilum*. Extremophiles 10:327–335.
Yutin N, Puigbò P, Koonin EV, Wolf YI. 2012. Phylogenomics of prokaryotic ribosomal proteins. PLoS One 7:e36972.
Zhang RR, Zhang C-TC. 2004. Identification of replication origins in the genome of the methanogenic archaeon, *Methanocaldococcus jannaschii*. Extremophiles 8:253–258.
Zhao A, Gray FC, MacNeill SA. 2006. ATP- and NAD+-dependent DNA ligases share an essential function in the halophilic archaeon *Haloferax volcanii*. Mol Microbiol. 59:743–752.

Associate editor: Martin Embley

2. Large-scale phylogenomic analysis

My second major project involved the investigation of the relationships between Archaea and Eukaryotes through a large-scale analysis. Over the last few years, there have been a growing number of studies indicating an archaeal origin of eukaryotes. However, all these studies used the same approach, i.e. the reconstruction of universal trees through the use of a very similar and small dataset of markers. Moreover, universal phylogenies may be criticized for a number of reasons. Universal proteins can be heavily mutationally saturated and they present several limitations such as: the number of proteins can be used because of the need for a bacterial out-group, the number of positions that can be unambiguously aligned between the three domains, and the taxonomic sampling because of the need for widespread presence in representatives of the three domains. I applied a novel strategy, which involved extracting and separately analyzing two sets of markers conserved between archaea/eukaryotes on one side, and archaea/bacteria on the other. This way, I could obtain new and more markers, better alignments, and use a larger taxonomic sampling. I carried out an extensive analysis on 132 archaeal genomes involving the extraction of core protein families, detection of orthologs, and exhaustive searches on bacteria and eukaryal genomes. I could identify 72 reliable phylogenetic makers shared between archaea and eukaryotes and 46 shared between bacteria and archaea. These markers were assembled into the largest supermatrices ever used for testing the archaea/eukaryote relationship. By analyzing the archaeal/eukaryote dataset I was able to exclude the origin of eukaryotes from within any of the archaeal orders or phyla. When I analyzed the bacterial/archaeal dataset I could highlight a new and original rooting for the archaeal tree that lies within the Euryarchaeota. These results were consistently obtained using a wide range of phylogenetic approaches, evolutionary models, statistical tests, and mutational desaturation analyses. My results would support an emergence of Eukaryotes from within the Archaea as sister to the TACK superphylum but also change dramatically our view on the phylogeny and evolution of the archaeal domain. Finally, I performed an exhaustive search for homologs of eukaryotic proteins inferred to date back to the LECA (the Last Eukaryotic Common Ancestor) that are not universally present in the Archaea. I identified 10 markers that were not reported in previous analyses that belong to a wide range of cellular processes. This opens up new avenues of research on the origin of these markers.

We are waiting for a few last results but the manuscript is ready to be submitted within the next month.

Article 2

The quest for Eukaryotic origins turns the tree of Archaea upside down

Kasie Raymann[1], Céline Brochier-Armanet[2], Simonetta Gribaldo[1*]

1. Institut Pasteur, Departement de Microbiologie, Unite Biologie Moleculaire du Gene chez les Extremophiles, Paris, France

2. Universite Lyon 1, CNRS, UMR5558, Laboratoire de Biometrie et Biologie Evolutive, Villeurbanne, France

*to whom correspondence should be addressed: simonetta.gribaldo@pasteur.fr

How the lineage leading to modern Eukaryotes originated is one of the most fundamental questions in Evolutionary Biology. A number of recent analyses based on universal proteins have indicated an emergence of Eukaryotes from a deep branch within the Archaea, specifically sister to or within the TACK (Thaumarchaeota/Aigarchaeota/Crenarchaeota/Korarchaeota) superphylum. This has important consequences because it indicates that the earliest steps of eukaryogenesis involved a *bona fide* archaeal ancestor with precise characteristics, such as G1P membrane lipids and a hyperthermophilic lifestyle. Therefore, it becomes crucial to test the robustness of this scenario using new proteins and approaches. Here, we have used a strategy alternative to universal proteins that allows having more markers and better quality phylogenetic signal. First, a dataset of proteins shared between Archaea and Eukaryotes robustly excludes the emergence of Eukaryotes from within the TACK superphylum or the Euryarchaeota. Second, we root the tree of Archaea using a bacterial outgroup and highlight the possibility for an alternative and original rooting within Euryarchaeota, leading on one side to a group comprising Methanogens Class I/Thermococcales/TACK and on the other side to all other euryarchaeal lineages. These results are robust against a range of statistical tests including the use of a nonhomogeneous model of protein evolution and mutational desaturation analysis. This novel rooting would support an emergence of Eukaryotes from within Archaea as sister to the TACK superphylum and is also obtained in trees obtained from a universal protein dataset. Therefore, if we are to embrace an archaeal origin for Eukaryotes, we also have to accept a dramatic rethinking of archaeal evolution and systematics over the last 40 years.

Introduction

How the lineage leading to the fantastic diversity of eukaryotic forms, including humans, came into being is one of the most fascinating open questions in Evolutionary Biology. Since their discovery in 1977, the Archaea have not stopped intriguing evolutionary biologist due to the presence of a number of eukaryotic traits, a feature that clearly distinguishes them from Bacteria. This has given rise to a variety of hypotheses on their involvement in the origin of Eukaryotes (Guy et al. 2014). The presence of these features has been traditionally interpreted as having been inherited from a common ancestor shared with Eukaryotes, whose nature remained indetermined. Initially divided in two major phyla, Euryarchaeota and Crenarchaeota, based on 16S rRNA trees (Woese et al. 1990; Guy et al. 2014), the availability of genomic sequences from a larger fraction of archaeal diversity has progressively blurred this line and introduced at least three new phyla, the Thaumarchaeota, the Aigarchaeota, and the Korarchaeota (Brochier-Armanet et al. 2011). Albeit suggested by a few old phylogenetic analysis (Lake et al. 1984; Lake 1988; Baldauf et al. 1996; Tourasse and Gouy 1999), over the past five years a number of universal trees of life rooted in the branch leading to Bacteria have supported an emergence of Eukaryotes from within the radiation of modern Archaea (Cox, Foster, et al. 2008; Foster, Cox, and Embley 2009a; Guy and Ettema 2011; Williams et al. 2012; Lasek-Nesselquist and Gogarten 2013; Williams and Embley 2014), with a specific link to a group comprising Thaumarchaeota/Aigarchaeota/Crenarchaeota and Korarchaeota (also called the TACK superphylum) (Guy and Ettema 2011). This has very important consequences because it clearly defines that an organism endowed with characteristics of a modern archaeon was the starting point for the process of eukaryogenesis. It is therefore critical to test the robustness of this hypothesis against the widest possible range of approaches, in particular to avoid all potential criticisms on the reliability of the phylogenetic signal contained in datasets used for comparing such distant domains which are prone to tree reconstruction artifacts (Gribaldo and Philippe 2002). In fact, universal datasets are heavily saturated by multiple substitutions that occurred at the same position over such long evolutionary time. Moreover, multiple alignments among the three domains reduce the set of markers and the number of unambiguously aligned positions that can be used. Finally, bacterial markers are frequently very distant from their archaeal/eukaryal counterparts. This introduces a very long branch leading to Bacteria, which contrast to the small one leading to Archaea/Eukaryotes, making it very difficult to resolve and introducing potential biases.

Here, we have applied the alternative strategy to universal proteins which we proposed a few years ago (Gribaldo et al. 2010). We assembled two separate datasets, one containing

only Archaea and Eukaryotes and one containing only Archaea and Bacteria. This strategy allowed us to gather more markers than previously used, obtain better quality alignments, and use a larger taxonomic sampling. We first tested if Eukaryotes are related to a specific archaeal lineage. Then, we analyzed the second dataset to root the tree of Archaea using a sure outgroup. Comparison of these two datasets provides new and original information on the topology of the universal tree of life, but also has important consequences on the origin and early evolution of the Archaea.

Results

Assembly of three high quality datasets of markers

In order to gather high quality datasets (i.e. well aligned verified orthologs with large taxonomic coverage), we applied a very strict semi-automatic protocol (see M&M for details). We first identified 235 protein families present in at least 95% of 132 complete archaeal genomes. We specifically excluded fast evolving lineages such as nanosized archaea, which may introduce artifacts. Because a large and representative taxonomic sampling is essential to verify the quality of the phylogenetic signal and detect eventual tree reconstruction artifacts, we searched for their homologs in a local databank of 211 and 31 genomes representative of major bacterial and eukaryotic phyla, respectively. We extracted the protein families that were present in at least 90% of Bacteria and/or 90% of Eukaryotes. We removed partial sequences and very distant homologs and built preliminary phylogenetic trees. We identified and removed the protein families showing clear signs of horizontal gene transfer (i.e. non-monophyly of archaea and/or bacteria), as well as eukaryotic sequences of bacterial/mitochondrial origin. This led to 92 candidate protein datasets (Archaea/Eukarya, Archaea/Bacteria, and Universal), for which we chose a smaller representative taxonomic sampling of 49 archaea, 67 bacteria, and 18 eukaryotes. In order to further identify potential horizontal gene transfer within Archaea or Bacteria, we extracted and realigned the archaeal and bacterial sequences from the three datasets, and Prunier (Abby et al. 2009) was used to identify and remove sequences causing incongruence. Following this step, the datasets maintaining at least 90% of taxonomic coverage for bacteria and/or archaea were retained. Sequences from the three domains were reunited and realigned, and datasets containing at least 70 amino acid positions after trimming were kept, leading to a final list of 79 protein families. We then assembled three concatenated datasets allowing for a maximum 10% of missing taxa for each protein: Archaea/Bacteria (AB, 46 proteins, 10986 amino acid positions), Archaea/Eukarya (AE, 72 proteins, 17892 aa

positions), and universal (ABE, 37 proteins and 9090 aa positions) (Figure 1). Finally, we analyzed the mutational saturation level of each dataset. As expected for proteins with very ancient relationships, the three supermatrices show substantial degree of mutational saturation. This saturation is a result of multiple substitutions occurring at the same position over time, which progressively erases the original phylogenetic signal (Figure 1). The high level of saturation in the datasets indicates that extra care needs to be taken in order to extract their phylogenetic signal.

A

Dataset	Number of proteins	Number of positions	Number of taxa	Average number of missing positions per taxa
ABE	37	9090	134	2,51%
AE	72	17892	67	2,88%
AB	46	10986	116	2,46%

B

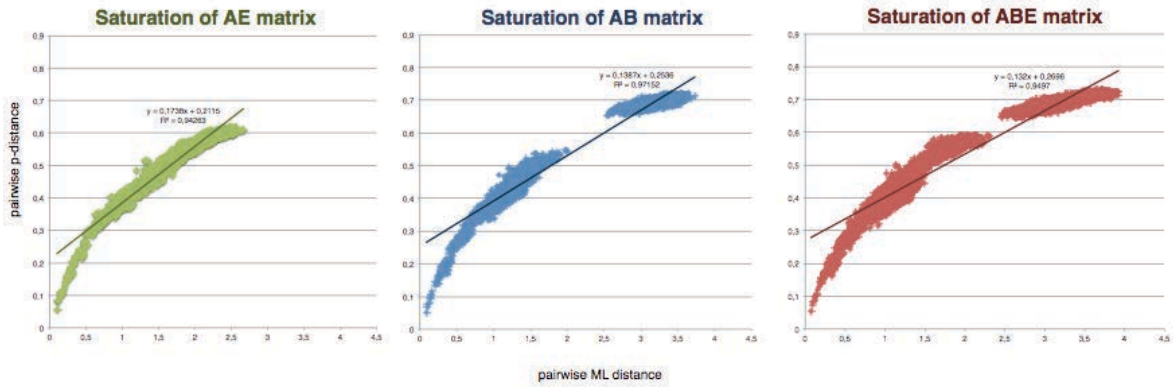


Figure 1. Summary of datasets A) A list of the number of proteins, amino acid positions, taxa, and average number of missing positions per taxa for the concatenated datasets. B) Graphs showing the mutational saturation level of each dataset. The mutational saturation level was estimated by comparing the evolutionary distance deduced from ML trees inferred with PhyML (x-axis) to the p-distance (i.e. observed divergence) deduced from the multiple alignment between each pair of sequences (x-axis) (Philippe et al. 1994; Philippe and Forterre 1999; Chiari et al. 2011) (Chiari et al., 2012, Philippe and Forterre, 1999 and Philippe et al., 1994). The slope of the linear regression indicates the amount of mutational saturation: the lower the slope, the greater the number of inferred multiple substitutions. Conversely, the higher the slope, the lower the number of inferred multiple substitutions

Are Eukarya affiliated to a specific archaeal lineage?

Universal trees obtained in previous analyses have indicated a specific affiliation of Eukaryotes as either sister to the TACK or within the TACK but sister to Korarchaeon, although it has been suggested that this may be an artifact due to the long and lonely branch leading to this phylum (Williams et al. 2012). Therefore, we sought to investigate this placement by using our AE dataset. It contains 72 proteins (37 universal and 35 uniquely shared between Archaea and Eukaryotes) and 17892 aa positions. With respect to the most recent analysis by Williams et al. 2012, we have 10 new AE markers and over twice as many

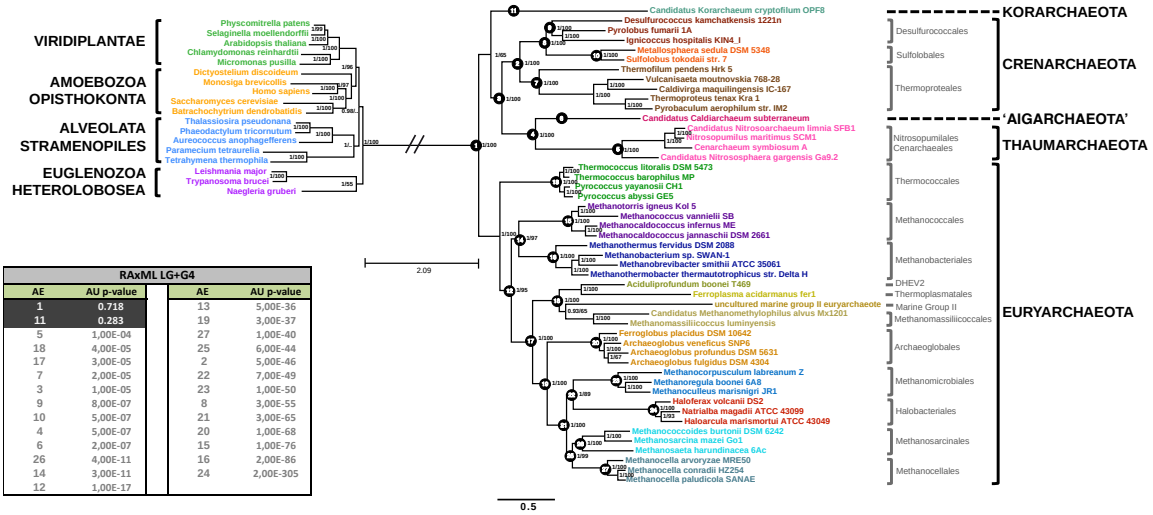
positions, 8438 versus 17892, possibly due to the fact that we used a better-performing software for alignment trimming and selection of informative positions (Criscuolo and Gribaldo 2010). Moreover, 10 markers used in the AE dataset of Williams et al. 2012 did not pass our strict protocol of selection because of evidence of HGT amongst Archaea (in gray in Table 1).

	Archaea/Eukaryote	
COG1718	Serine/threonine protein kinase involved in cell cycle control	Signal transduction mechanisms, Cell cycle control; cell division; chromosome partitioning
COG1632	Ribosomal protein L15E	Translation; ribosomal structure and biogenesis
COG1889	Fibrillarin-like rRNA methylase	Translation; ribosomal structure and biogenesis
COG1269	Archaeal/vacuolar-type H+-ATPase subunit I	Energy production and conversion
COG0051	Ribosomal protein S10	Translation; ribosomal structure and biogenesis
COG1890	Ribosomal protein S3AE	Translation; ribosomal structure and biogenesis
COG1500	Predicted exosome subunit	Translation; ribosomal structure and biogenesis
COG1471	Ribosomal protein S4E	Translation; ribosomal structure and biogenesis
COG2147	Ribosomal protein L19E	Translation; ribosomal structure and biogenesis
COG0197	Ribosomal protein L16/L10E	Translation; ribosomal structure and biogenesis
COG1093	Translation initiation factor 2, alpha subunit (eIF-2alpha)	Translation; ribosomal structure and biogenesis
COG1976	Translation initiation factor 6 (eIF-6)	Translation; ribosomal structure and biogenesis
COG1867	N2,N2-dimethylguanosine tRNA methyltransferase	Translation; ribosomal structure and biogenesis
COG2101	TATA-box binding protein (TBP), component of TFIID and TFIIIB	Transcription
COG0638	20S proteasome, beta subunit	Posttranslational modification; protein turnover; chaperones
COG1155	Archaeal/vacuolar-type H+-ATPase subunit A	Energy production and conversion
COG0470/1	ATPase involved in DNA replication	Replication; recombination and repair
COG0423	Glycyl-tRNA synthetase (class II)	Translation; ribosomal structure and biogenesis
COG0186	Ribosomal protein S17	Translation; ribosomal structure and biogenesis
COG0097	Ribosomal protein L6P/L9E	Translation; ribosomal structure and biogenesis
COG0091	Ribosomal protein L22	Translation; ribosomal structure and biogenesis
COG1358	Ribosomal protein HS6-type (S12/L30/L7a)	Translation; ribosomal structure and biogenesis
COG0456	Acetyltransferases	General function prediction only
COG0522/3	Ribosomal protein S4 and related proteins	Translation; ribosomal structure and biogenesis
COG1095	DNA-directed RNA polymerase, subunit E'	Transcription
COG1717	Ribosomal protein L32E	Translation; ribosomal structure and biogenesis
COG1394	Archaeal/vacuolar-type H+-ATPase subunit D	Energy production and conversion
COG1631	Ribosomal protein L44E	Translation; ribosomal structure and biogenesis
COG1498	Protein implicated in ribosomal biogenesis, Nop56p homolog	Translation; ribosomal structure and biogenesis
COG0585	Uncharacterized conserved protein	Nucleotide transport and metabolismunction unknown
COG0592	DNA polymerase sliding clamp subunit (PCNA homolog)	Replication; recombination and repair
COG2139	Ribosomal protein L21E	Translation; ribosomal structure and biogenesis
COG1537	Predicted RNA-binding proteins	General function prediction only
COG1241	Predicted ATPase involved in replication control, Cdc46/Mcm family	Replication; recombination and repair
COG0258	5'-3' exonuclease (including N-terminal domain of PolI)	General function prediction only eplication; recombination and repair
COG0016	Phenylalanyl-tRNA synthetase alpha subunit	Translation; ribosomal structure and biogenesis
COG0148	Enolase	Energy production and conversionarbohydrate transport and metabolism
COG0180	Tryptophanyl-tRNA synthetase	Translation; ribosomal structure and biogenesis
COG0495	Leucyl-tRNA synthetase	Translation; ribosomal structure and biogenesis
COG1793	ATP-dependent DNA ligase	General function prediction only eplication; recombination and repair
COG1899	Deoxyhypusine synthase	Posttranslational modification; protein turnover; chaperones
COG0184	Ribosomal protein S15P/S13E	Translation; ribosomal structure and biogenesis
COG0089	Ribosomal protein L23	Translation; ribosomal structure and biogenesis
COG0185	Ribosomal protein S19	Translation; ribosomal structure and biogenesis
COG0256/1	Ribosomal protein L18	Translation; ribosomal structure and biogenesis
	Universal	
COG0017	Aspartyl/asparaginyl-tRNA synthetases	Translation; ribosomal structure and biogenesis
COG5256	Translation elongation factor EF-1alpha (GTPase)	Translation; ribosomal structure and biogenesis
COG0541	Signal recognition particle GTPase	Intracellular trafficking; secretion; and vesicular transport
COG0459	Chaperonin GroEL (HSP60 family)	Posttranslational modification; protein turnover; chaperones
COG0052	Ribosomal protein S2	Translation; ribosomal structure and biogenesis
COG0103	Ribosomal protein S9	Translation; ribosomal structure and biogenesis
COG0102	Ribosomal protein L13	Translation; ribosomal structure and biogenesis
COG0049	Ribosomal protein S7	Translation; ribosomal structure and biogenesis
COG0086	DNA-directed RNA polymerase, beta' subunit/160 kD subunit	Transcription
COG0085	DNA-directed RNA polymerase, beta subunit/140 kD subunit	Transcription
COG0090	Ribosomal protein L2	Translation; ribosomal structure and biogenesis
COG0092	Ribosomal protein S3	Translation; ribosomal structure and biogenesis
COG0093	Ribosomal protein L14	Translation; ribosomal structure and biogenesis
COG0094	Ribosomal protein L5	Translation; ribosomal structure and biogenesis
COG0098	Ribosomal protein S5	Translation; ribosomal structure and biogenesis
COG0201	Preprotein translocase subunit SecY	Intracellular trafficking; secretion; and vesicular transport
COG0080	Ribosomal protein L11	Translation; ribosomal structure and biogenesis
COG0081	Ribosomal protein L1	Translation; ribosomal structure and biogenesis
COG0480	Translation elongation factors (GTPases)	Translation; ribosomal structure and biogenesis
COG0533	Metal-dependent proteases with possible chaperone activity	Posttranslational modification; protein turnover; chaperones
COG0532	Translation initiation factor 2 (IF-2; GTPase)	Translation; ribosomal structure and biogenesis
COG1156	Archaeal/vacuolar-type H+-ATPase subunit B	Energy production and conversion
COG0470	ATPase involved in DNA replication	General function prediction only eplication; recombination and repair
COG0048	Ribosomal protein S12	Translation; ribosomal structure and biogenesis
COG0100	Ribosomal protein S11	Translation; ribosomal structure and biogenesis
COG0552	Signal recognition particle GTPase	Intracellular trafficking; secretion; and vesicular transport
COG0099	Ribosomal protein S13	Translation; ribosomal structure and biogenesis
COG0087	Ribosomal protein L3	Translation; ribosomal structure and biogenesis
COG0088	Ribosomal protein L4	Translation; ribosomal structure and biogenesis
COG0096	Ribosomal protein S8	Translation; ribosomal structure and biogenesis
COG5257	Translation initiation factor 2, gamma subunit (eIF-2gamma; GTPase)	Translation; ribosomal structure and biogenesis
COG0030	Dimethyladenosine transferase (rRNA methylation)	Translation; ribosomal structure and biogenesis
COG0200	Ribosomal protein L15	Translation; ribosomal structure and biogenesis
COG0130/1	Pseudouridine synthase	Translation; ribosomal structure and biogenesis
COG1222	ATP-dependent 26S proteasome regulatory subunit	Posttranslational modification; protein turnover; chaperones
COG0020	Undecaprenyl pyrophosphate synthase	General function prediction only eplication; recombination and repair ipid transport and metabolism
COG0062/1	Uncharacterized conserved protein	Nucleotide transport and metabolismunction unknown
COG0008	Glutamyl- and glutaminyl-tRNA synthetases	Translation; ribosomal structure and biogenesis
COG0024	Methionine aminopeptidase	Translation; ribosomal structure and biogenesis
COG1245	Predicted ATPase, RNase L inhibitor (RLI) homolog	General function prediction only
	Archaea/Bacteria	
COG0528	Uridylate kinase	Nucleotide transport and metabolism
COG0358/3	DNA primase (bacterial type)	General function prediction only eplication; recombination and repair
COG0343/1	Queuine/archaeosine tRNA-ribosyltransferase	Translation; ribosomal structure and biogenesis
COG0462	Phosphoribosylpyrophosphate synthetase	Nucleotide transport and metabolism Amino acid transport and metabolism
COG0195	Transcription elongation factor	Transcription
COG0012	Predicted GTPase, probable translation factor	Translation; ribosomal structure and biogenesis
COG0250	Transcription antiterminator	Transcription
COG0126	3-phosphoglycerate kinase	Energy production and conversionarbohydrate transport and metabolism
COG0149	Triosephosphate isomerase	Energy production and conversionarbohydrate transport and metabolism

Table 1. List of proteins used in the concatenated datasets. Proteins highlighted in green were used in this analysis. Proteins in black font were also used in the analysis of Williams et al 2012. Proteins in blue font are new proteins that were identified and used in this study. The proteins highlighted in white are those that were used in the dataset of Williams et al. 2012 but were not used in this analysis.

We ran two independent analyses, one in the Maximum Likelihood framework with an empirical 20x20 homogeneous model of amino acid substitution LG+Γ4 (Le and Gascuel 2008), and one in the Bayesian framework with a more realistic model of protein evolution that takes into account the heterogeneity of the substitution process across sites CAT+GTR+Γ4 (Lartillot and Philippe 2004).

Both Bayesian and Maximum Likelihood analyses gave a consistent and robust unrooted phylogeny (Figure 2). The quality of our dataset is underlined by the nicely resolved internal phylogenies for both Archaea and Eukaryotes, recovering the monophyly of the major archaeal orders and eukaryotic phyla, which are frequently difficult to recover, especially in such deep phylogenies. However, Bayesian analysis appears to better resolve the internal eukaryotic topology, notably concerning the monophyly of Euglenozoa/Heterolobosea, which has a very low bootstrap support in the ML tree, and that of Stramenopiles/Alveolata, which is not recovered in the ML tree. Also, the ML support for the monophyly of TACK is very low (bootstrap value 65%) as compared to Bayesian analysis (posterior probability 1). These differences may be due to a better capture of the substitution process for the A/E dataset by the CAT+GTR model, as a Bayesian analysis with CAT alone was largely unresolved (not shown). Nevertheless, although the tree is unrooted, Eukaryotes appear robustly external to both TACK and Euryarchaeota, meaning that any branching within their radiation can be excluded. This placement does not appear to be affected by a bias in amino acid composition because it is also recovered when the dataset is recoded according to 6 Dayhoff categories (SM Figure 1).



RAxML LG+G4			
AE	AU p-value	AE	AU p-value
1	0.718	13	5,00E-36
11	0.283	19	3,00E-37
5	1,00E-04	27	1,00E-40
18	4,00E-05	25	6,00E-44
17	3,00E-05	2	5,00E-46
7	2,00E-05	22	7,00E-49
3	1,00E-05	23	1,00E-50
9	8,00E-07	8	3,00E-55
10	5,00E-07	21	3,00E-65
4	5,00E-07	20	1,00E-68
6	2,00E-07	15	1,00E-76
26	4,00E-11	16	2,00E-86
14	3,00E-11	24	2,00E-305
12	1,00E-17		

Figure 2. Unrooted Bayesian phylogeny of a concatenated data set of 72 proteins shared between archaea and eukaryotes (17892 amino acid positions). The tree was calculated by Phylobayes (CAT+GTR+Γ4). The scale bar represents the average number of substitutions per site. Values at nodes represent posterior probabilities and BV based on 100 resamplings of the original data set calculated by PhyML (LG+Γ4), when the same node was recovered. The table shows results from AU tests on 27 alternative placements for Eukaryotes with respect to Archaea (see text and M&M for details).

In order to further test this external placement of Eukaryotes with respect to TACK and Euryarchaeota, we fixed the archaeal and eukaryal internal phylogenies and manually created 27 alternative topologies by moving the eukarya within each of the internal archaeal branches (i.e. all branches excluding within orders, see numbered nodes on Figure 2). We then we ran an AU test with RAxML v 7.2.8 (Stamatakis 2006) with LG+G4 to see if they are statistically rejected as compared to the one indicated by the Bayesian and ML trees (i.e. #1). All alternative topologies were rejected with statistical confidence except #1, and a single less supported alternative hypothesis indicating a branching of Eukaryotes with Korarchaeota (#11).

As the A/E dataset includes 37 markers that are universal and 35 that are specifically shared between Archaea and Eukaryotes, we sought to verify if one of the two groups of proteins were responsible for the signal obtained from the whole dataset. However, this does not seem the case as separate Bayesian and Maximum Likelihood analyses of these two sets of proteins (SM Figure 2 and 3) give trees consistent with those obtained by the whole supermatrix.

In order to assess whether the phylogenetic signal provided by the AE dataset is robust against its heavy mutational saturation level (Figure 1), we applied the slow-fast method (Brinkmann and Philippe 1999). Slow-fast is a desaturation strategy (see M&M for details) that identifies and progressively removes the most saturated positions in a sequence alignment. Using this method one can follow the variation in the phylogenetic signal as more saturated positions are added to the matrix. For each position of the supermatrix, we the evolutionary rates within established monophyletic groups in both archaea and eukaryotes were estimated. Then, the overall rate of a position was calculated as the sum of the internal rates over all groups. This strategy avoids circularity because it does not make any assumptions about the evolutionary relationships among groups. We built 41 supermatrices including positions showing progressively higher rates of substitutions among groups, from 0 to 41 (Figure 3) and calculated both Bayesian (CAT+G4) and ML (LG+G4) trees for each of them. We followed the evolution of the topologies and support at nodes in order to identify the window were the matrices provide the most reliable results as a tradeoff between too little signal (few saturation but few positions) and too much noise (more positions but more saturation).

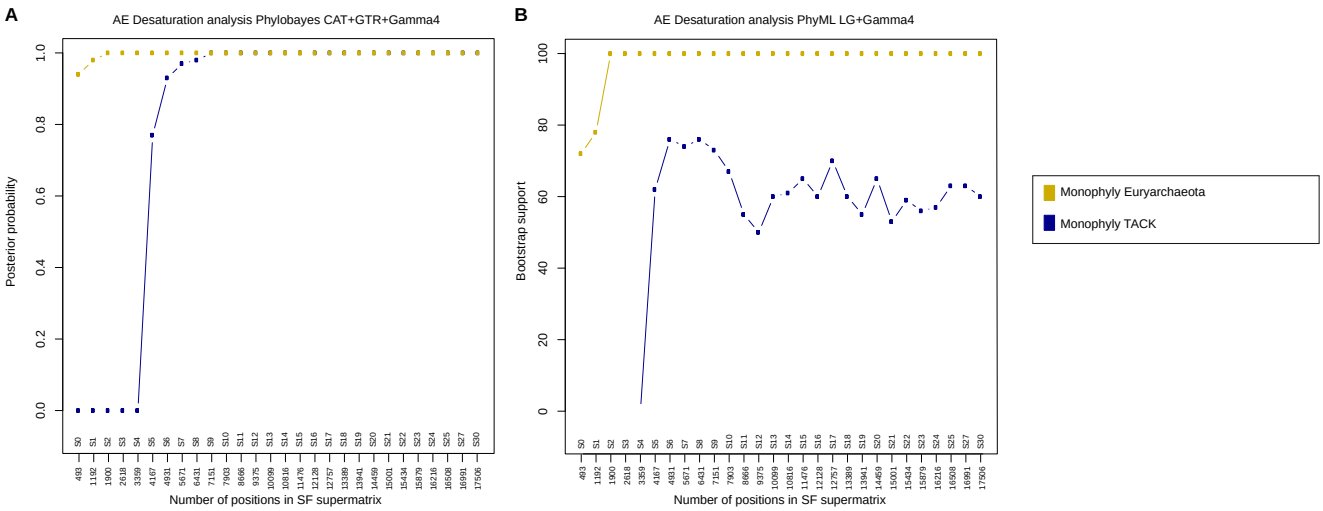


Figure 3. Desaturation analysis. On the x-axis the names and number of positions of matrices containing a progressively larger number of saturated positions are shown (see text and M&M for details). The graphs represent the support of each matrix for an external placement of Eukaryotes with respect to the TACK or Euryarchaeota as indicated by the support at the nodes corresponding to the monophylies of these two archaeal groups. The phylogenies were inferred by Phylobayes with the CAT+GTR+Γ4 model (A) and with PhyML with the LG+Γ4 model (B).

We then analyzed the evolution of the signal for the relationship between Archaea and Eukaryotes as assessed by the monophyly of Euryarchaeota and that of the TACK. Consistent with the analysis of the complete dataset, we observe robust support for the monophyly of Euryarchaeota starting from the very first matrices and a progressive increase for the monophyly of TACK with the CAT+GTR model (Figure 3A). With the LG model we observe strong support for monophyly of the Euryarchaeota and a much less robust support for the monophyly of the TACK (Figure 3B). Indeed, alternative topologies are present in the bootstrap sampling of trees where the TACK are paraphyletic due to an instable branching of Korarchaeon (not shown). AU tests on two matrices with different level of saturation (S9 and S15) confirm that topology #1 receives the highest support, with two alternative topologies not rejected but much less supported where Eukaryotes are specifically affiliated with either Korarchaeota (#11) or the TAC (#3) (Table 2).

A

RAxML LG + Γ4					
AE S09	AU p-value	Std. error	AE S15	AU p-value	Std. error
11	0.568	(0.010)	1	0.647	(0.009)
1	0.545	(0.010)	11	0.359	(0.009)
3	0.088	(0.007)	9	0.001	(0.010)
4	0.002	(0.001)	10	5,00E-04	(0.001)
17	2,00E-04	(0.001)	18	1,00E-04	(0.001)
20	1,00E-04	(0.000)	3	7,00E-05	(0.000)
25	1,00E-04	(0.001)	5	5,00E-05	(0.000)
22	1,00E-04	(0.001)	23	5,00E-05	(0.000)
5	7,00E-05	(0.001)	6	4,00E-05	(0.000)
27	5,00E-05	(0.000)	8	4,00E-05	(0.000)
21	3,00E-05	(0.000)	14	2,00E-05	(0.000)
12	6,00E-06	(0.000)	15	2,00E-05	(0.000)
6	4,00E-06	(0.000)	16	1,00E-05	(0.000)
13	7,00E-07	(0.000)	21	6,00E-06	(0.000)
26	6,00E-07	(0.000)	19	5,00E-06	(0.000)
2	3,00E-09	(0.000)	20	3,00E-06	(0.000)
15	1,00E-42	(0.000)	22	7,00E-07	(0.000)
16	2,00E-44	(0.000)	13	3,00E-08	(0.000)
7	1,00E-45	(0.000)	2	7,00E-09	(0.000)
23	3,00E-50	(0.000)	26	6,00E-10	(0.000)
24	2,00E-54	(0.000)	12	6,00E-40	(0.000)
8	6,00E-57	(0.000)	24	6,00E-41	(0.000)
19	5,00E-60	(0.000)	27	3,00E-46	(0.000)
14	9,00E-64	(0.000)	4	9,00E-57	(0.000)
9	4,00E-71	(0.000)	7	4,00E-63	(0.000)
18	3,00E-82	(0.000)	25	6,00E-71	(0.000)
10	2,00E-99	(0.000)	17	3,00E-163	(0.000)

B

NH–NS LG+COaLA[2]+Γ4					
AE S09	AU p-value	Std. error	AE S15	AU p-value	Std. error
11	0.595	(0.009)	1	0.514	(0.010)
1	0.484	(0.010)	11	0.493	(0.010)
3	0.037	(0.005)	3	0.001	(0.001)
5	0.002	(0.007)	2	0,0002	(0.001)
10	0.001	(0.005)	21	0,00004	(0.000)
4	0.001	(0.000)	10	0,00004	(0.000)
18	4,00E-04	(0.005)	4	3,00E-05	(0.000)
6	3,00E-05	(0.000)	12	6,00E-06	(0.000)
9	5,00E-07	(0.000)	9	8,00E-07	(0.000)
7	1,00E-08	(0.000)	23	7,00E-07	(0.000)
8	8,00E-09	(0.000)	27	7,00E-07	(0.000)
12	1,00E-09	(0.000)	25	1,00E-07	(0.000)
19	3,00E-10	(0.000)	22	7,00E-08	(0.000)
24	1,00E-11	(0.000)	17	7,00E-09	(0.000)
13	4,00E-14	(0.000)	26	4,00E-09	(0.000)
2	2,00E-33	(0.000)	14	3,00E-11	(0.000)
27	7,00E-35	(0.000)	24	1,00E-21	(0.000)
16	4,00E-37	(0.000)	20	8,00E-30	(0.000)
17	1,00E-44	(0.000)	13	2,00E-34	(0.000)
21	3,00E-48	(0.000)	5	8,00E-43	(0.000)
25	3,00E-58	(0.000)	15	8,00E-45	(0.000)
23	2,00E-67	(0.000)	6	5,00E-45	(0.000)
26	1,00E-94	(0.000)	16	2,00E-48	(0.000)
15	5,00E-99	(0.000)	7	2,00E-48	(0.000)
20	3,00E-99	(0.000)	19	5,00E-52	(0.000)
14	2,00E-99	(0.000)	18	3,00E-58	(0.000)
22	8,00E-114	(0.000)	8	1,00E-108	(0.000)

Table 2. Hypothesis testing. The table shows results from AU tests on 27 alternative placements for Eukaryotes with respect to Archaea obtained from the per-site log likelihoods calculated by A) RAxML (LG+Γ4) and B) COaLA (nH model + Γ4) for the AE SF matrices SF09 and SF15. See text and M&M for details.

To sum up, our results show that the CAT+GTR+Γ4 model appears more robust than the LG+Γ4 model for extracting phylogenetic signal from the AE dataset. Furthermore, our results do not support an emergence of Eukaryotes from either within the TACK or within the Euryarchaeota, consistent with the analysis of the global supermatrix (Figure 2). These results are robust against desaturation analysis and different statistical tests.

Where is the root of the Archaea?

If a branching of Eukaryotes is excluded from within the TACK or within the Euryarchaeota, this leaves three possibilities: (i) they are sister to the TACK; (ii) they are sister to the Euryarchaeota; or (iii) they are sister to all Archaea (TACK+Euryarchaeota). The choice between these three alternatives therefore depends on where the root of the Archaea lies. In fact, a root within Euryarchaeota would favor scenario (i), a root within TACK would favor scenario (ii), and a root between Euryarchaeota and TACK would leave open all three possibilities.

Therefore, we proceeded to root the archaeal tree by using the AB dataset, because Bacteria are a sure outgroup. The AB dataset contains 46 proteins (37 universal and 9 uniquely shared between Archaea and Bacteria) and 10,986 aa positions. Both Bayesian (CAT+GTR+Γ4) and Maximum likelihood (LG+Γ4) analysis provide a largely resolved tree with robust internal branching for both Archaea and Bacteria, albeit ML has lower support at nodes (Figure 4 and 5, respectively). In particular, we observe the monophyly of all major bacterial phyla,

including for example that of Proteobacteria, frequently difficult to recover (Ramulu et al. 2014), and Planctomycetales/Verrucomicrobiales/Chlamydia (PVC) (Wagner and Horn 2005). Although the aim of this study was not to resolve the deep phylogeny of the bacterial domain, these data testify for the good quality of our dataset. Concerning Archaea, we recover all major orders, which is consistent with the internal phylogeny obtained with the A/E dataset. These results do not appear to be affected by biases of amino acid compositions because they are also supported with a recoded dataset (SM Figure 4). However, we observe one major discrepancy between the Bayesian and the ML analysis concerning the root of the Archaea. In fact, Bayesian analysis with the CAT+GTR model (Figure 4) recovers a root within Euryarchaeota, making this phyum non-monophyletic. With this root the Methanococcales/Methanobacteriales/Thermococcales are clustered with the TACK (which we will call hereafter Group I, pp 1) on one side, and on the other side all other euryarchaeal lineages are together (hereafter called Group II pp 1). In contrast, ML analysis with the LG model (Figure 5) shows strong support for a root of Archaea in between Euryarchaeota (BV 97%) and TACK (BV 98%). Because the root within Euryarchaeota would favor scenario (i) where Eukaryotes are sister to the TACK would leave open all three possible scenarios for eukaryotic origins, this discrepancy needs to be carefully tested.

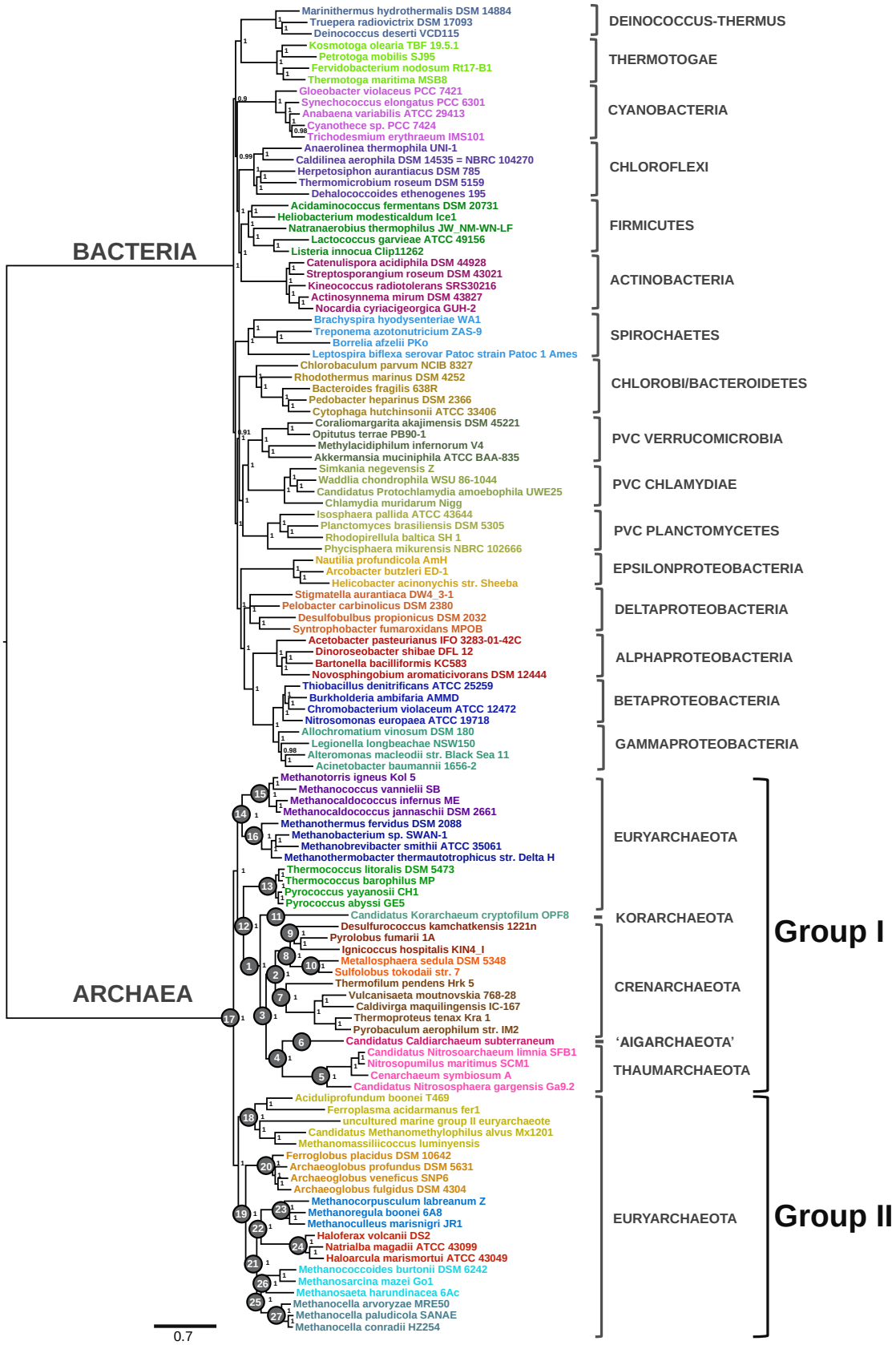


Figure 4. Unrooted Bayesian phylogeny of the AB supermatrix (10986 amino acid positions). The tree was calculated by Phylobayes (CAT+GTR+Γ4). The scale bar represents the average number of substitutions per site. Values at nodes represent posterior probabilities. The numbered nodes on the archaeal phylogeny represent the 27 alternative rootings that were tested (see text and M&M for details).

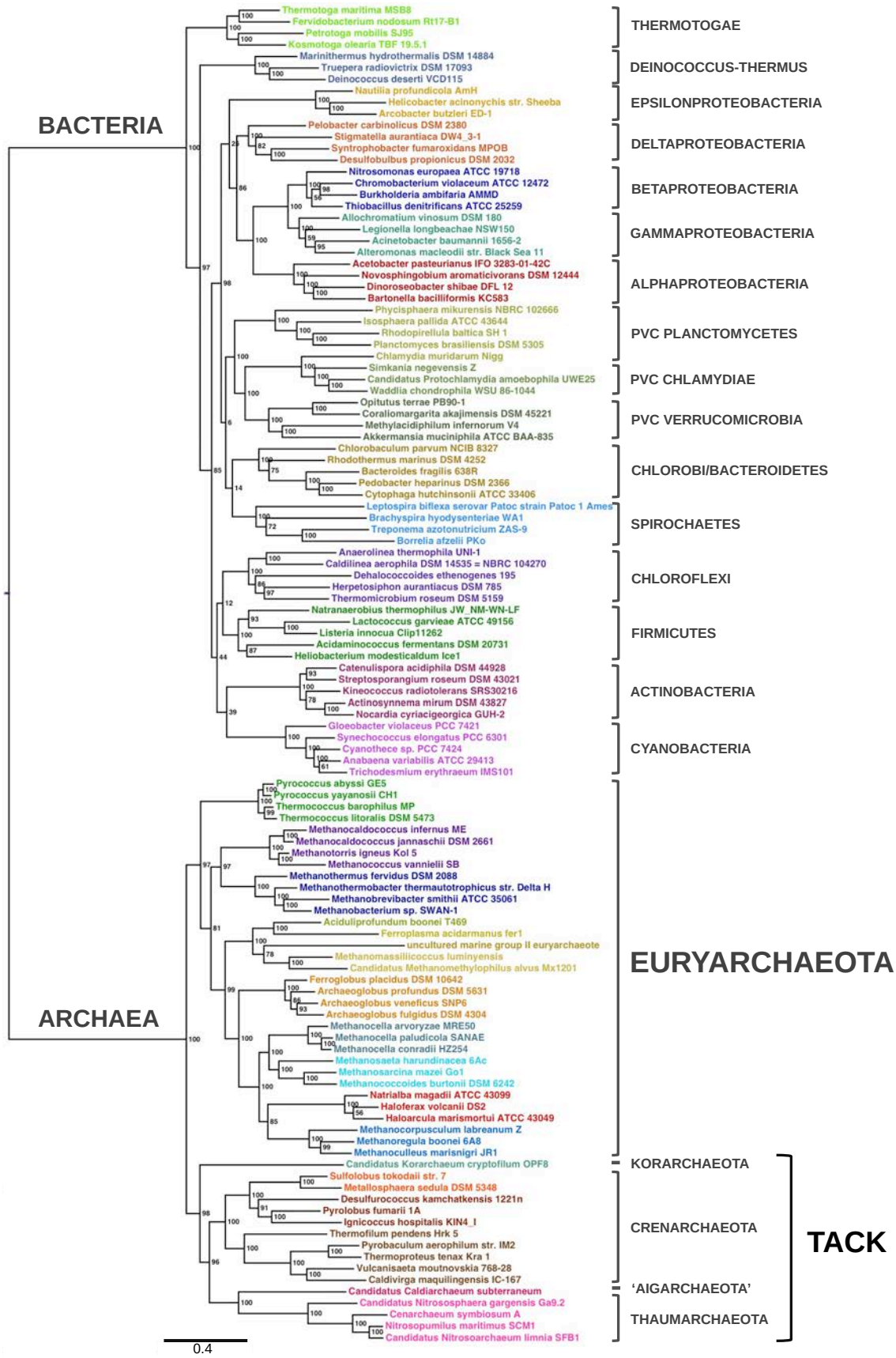


Figure 5. Unrooted ML phylogeny of the AB supermatrix (10986 amino acid positions). The tree was calculated by PhyML (LG+Γ4). The scale bar represents the average number of substitutions per site. Values at nodes represent supports obtained by non parametric bootstrapping based on 100 resamplings of the original alignment. See text and M&M for details.

Similar to the A/E dataset analysis, we fixed the archaeal and bacterial topologies and carried out an AU test (LG+G4) on all 27 possible locations of the root (i.e. all branches except within orders) (Figure 4). Our results show that only four alternative rootings are statistically not rejected, with highest support for root # 1 (between TACK and Euryarchaeota, as observed in the ML tree in Figure 4B), but immediately followed by root #17 (within Euryarchaeota, as observed in the PBayes tree in Figure 4).

We therefore sought to investigate the impact of mutational saturation on these alternative rootings by applying the SF approach to the A/B dataset. We built different supermatrices including positions with progressively higher rates of substitutions among groups, from zero to 53 (S0 to S53). Bayesian (CAT+GTR+ Γ4) and ML (LG+ Γ4) analyses were performed on each matrix and we followed the evolution of the phylogenetic signal (Figure 6). When we examine the evolution of the phylogenetic signal for the archaeal topology for the Bayesian (CAT+GTR+ Γ4) we never observe a rooting between TACK and Euryarchaeota (root #1) but systematically a rooting deep within Euryarchaota (root #17), as assessed by the non-monophyly of Euryarchaeota (Figure 6).

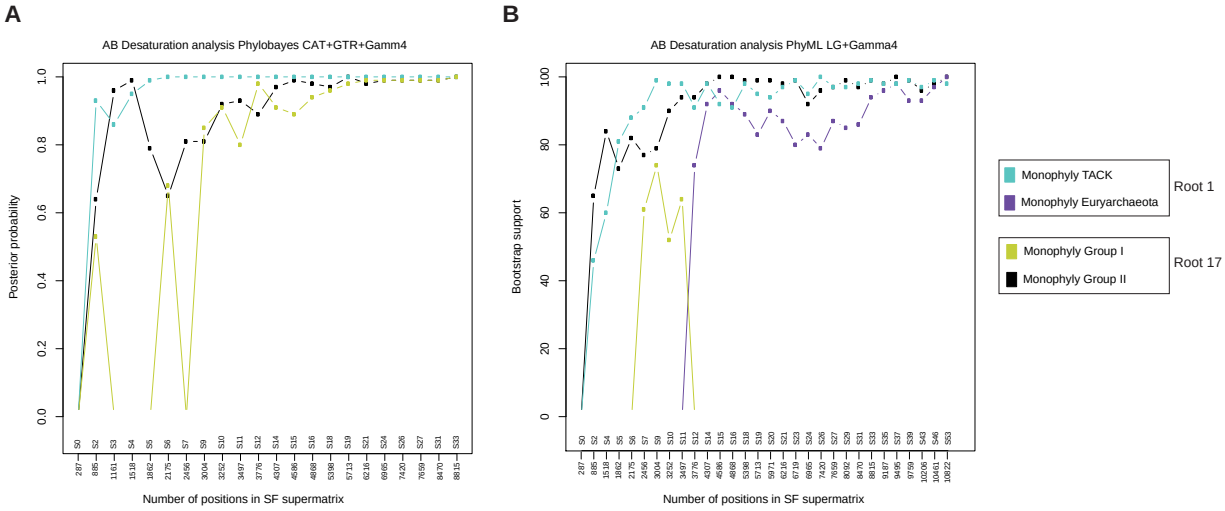


Figure 6. Desaturation analysis. On the x-axis are shown the names and number of positions of matrices containing a progressively larger number of saturated positions (see text and M&M for details). The graphs represent support of each matrix for the monophyly of the Euryarchaeota and the TACK (supporting root #1) and the monophyly of Group I and Group II (supporting root #17) for each destaturated supermatrix as calculated by Phylobayes with the CAT+GTR+Γ4 model (A) and PhyML with the LG+Γ4 model (LG).

As observed for the A/E dataset, Maximum likelihood +LG analysis is overall less robust than Bayesian (CAT+GTR+ Γ4) analysis, in particular concerning the monophyly of Proteobacteria, which is recovered only with matrices S12 to S21 (between 3776 and 5971 aa positions) and rather low support (not shown). When examining the evolution of the archaeal topology, we observe a robust support for root #1 as assessed by the joint monophylies of Euryarchaeota

and TACK, consistent with the complete dataset (Figure 6B). However, we also note a low support for the alternative root #17 in the less saturated matrices S6 to S11 as assessed by the joint monophylies of Group I and Group II, and a switch to root #1 at matrix S12 (Figure 6B and SM Figure 5 and 6).

We ran AU tests for all possible 27 roots on three SF matrices, S11 (just before the switch between root #17 and #1), S15 (within the les saturated window supporting root #1) and the more saturated S27 matrix. Results show that the three matrices significantly reject all rootings except root #1 (between TACK and Euryarchaeota), root #17 (within Euryarchaeota) as well as a few alternative rootings deep within Euryarchaeota (Table 3A). Finally, for all the seven roots not rejected by the AU tests on both the complete dataset and the SF matrices (i.e. #1, 17, 18, 14, 15, 16, 19) we ran a more realistic model of protein evolution that allows both amino acid composition as well as evolutionary rates to change along tree branches (branch-nonhomogeneous and nonstationary) (Groussin et al. 2013). The results confirm those of the AU test, with only root #14 rejected (Table 3B).

A

RAxML LG + Γ4								
AB_SF11	AU p-value	Std. error	AB_SF15	AU p-value	Std. error	AB_SF27	AU p-value	Std. error
17	0.930	(0.003)	1	0.855	(0.006)	1	0.705	(0.008)
14	0.403	(0.012)	17	0.447	(0.011)	17	0.571	(0.010)
1	0.282	(0.010)	14	0.141	(0.010)	18	0.193	(0.009)
18	0.147	(0.008)	18	0.059	(0.006)	11	0.039	(0.004)
15	0.127	(0.009)	11	0.038	(0.004)	19	0.008	(0.004)
16	0.071	(0.009)	12	0.028	(0.007)	14	0.007	(0.003)
19	0.071	(0.007)	19	0.021	(0.005)	12	0.004	(0.002)
11	0.038	(0.005)	15	0.016	(0.005)	4	0.001	(0.001)
12	0.007	(0.003)	16	0.009	(0.004)	3	3,00E-05	(0.000)
21	0.004	(0.002)	27	0.001	(0.002)	6	2,00E-05	(0.000)
20	0.003	(0.003)	3	5,00E-04	(0.001)	13	2,00E-05	(0.000)
13	0.001	(0.001)	26	5,00E-04	(0.001)	20	5,00E-06	(0.000)
4	1,00E-04	(0.001)	13	3,00E-04	(0.001)	15	2,00E-06	(0.000)
24	3,00E-05	(0.000)	20	2,00E-05	(0.000)	21	1,00E-06	(0.000)
27	3,00E-05	(0.000)	8	1,00E-05	(0.000)	16	5,00E-08	(0.000)
26	3,00E-06	(0.000)	4	7,00E-06	(0.000)	27	1,00E-14	(0.000)
9	5,00E-07	(0.000)	7	2,00E-06	(0.000)	26	2,00E-35	(0.000)
10	7,00E-09	(0.000)	21	2,00E-07	(0.000)	9	1,00E-40	(0.000)
3	4,00E-09	(0.000)	23	5,00E-40	(0.000)	5	1,00E-44	(0.000)
8	1,00E-45	(0.000)	24	8,00E-49	(0.000)	3	1,00E-44	(0.000)
25	7,00E-46	(0.000)	5	4,00E-50	(0.000)	10	3,00E-45	(0.000)
5	5,00E-55	(0.000)	9	2,00E-53	(0.000)	24	4,00E-52	(0.000)
23	4,00E-55	(0.000)	3	7,00E-64	(0.000)	22	7,00E-55	(0.000)
22	8,00E-56	(0.000)	6	5,00E-70	(0.000)	7	6,00E-55	(0.000)
3	5,00E-64	(0.000)	25	1,00E-70	(0.000)	8	2,00E-65	(0.000)
7	9,00E-81	(0.000)	22	2,00E-79	(0.000)	25	4,00E-67	(0.000)
6	2,00E-81	(0.000)	10	3,00E-106	(0.000)	23	1,00E-84	(0.000)

B

NH–NS LG+COaLA[2]+Γ4								
AB S11	AU p-value	Std. error	AB S15	AU p-value	Std. error	AB S27	AU p-value	Std. error
17	0.911	(0.004)	1	0.891	(0.005)	1	0.695	(0.009)
14	0.505	(0.012)	17	0.422	(0.011)	17	0.605	(0.010)
1	0.258	(0.010)	14	0.170	(0.011)	18	0.156	(0.008)
15	0.169	(0.011)	11	0.053	(0.004)	11	0.061	(0.005)
16	0.120	(0.010)	18	0.052	(0.007)	12	0.027	(0.007)
18	0.100	(0.007)	12	0.036	(0.008)	14	0.009	(0.004)
19	0.069	(0.007)	15	0.034	(0.006)	19	0.007	(0.003)
11	0.030	(0.004)	19	0.015	(0.004)	15	0.001	(0.002)
12	0.010	(0.004)	16	0.014	(0.004)	6	0.001	(0.002)
20	0.001	(0.001)	4	5,00E-04	(0.002)	23	1,00E-04	(0.003)
6	0.001	(0.004)	6	5,00E-04	(0.004)	20	6,00E-05	(0.000)
13	0.001	(0.001)	26	5,00E-04	(0.001)	16	1,00E-05	(0.000)
4	0.001	(0.002)	27	3,00E-04	(0.001)	24	7,00E-06	(0.000)
3	0.001	(0.003)	20	9,00E-05	(0.000)	9	6,00E-06	(0.000)
2	4,00E-04	(0.001)	21	7,00E-05	(0.000)	10	5,00E-06	(0.000)
5	2,00E-04	(0.001)	2	3,00E-05	(0.000)	7	2,00E-06	(0.000)
21	1,00E-04	(0.000)	13	1,00E-05	(0.000)	2	2,00E-06	(0.000)
7	2,00E-06	(0.000)	3	2,00E-08	(0.000)	8	9,00E-08	(0.000)
9	3,00E-07	(0.000)	9	4,00E-42	(0.000)	21	7,00E-08	(0.000)
10	2,00E-08	(0.000)	22	1,00E-46	(0.000)	25	5,00E-10	(0.000)
22	1,00E-39	(0.000)	25	4,00E-48	(0.000)	22	5,00E-10	(0.000)
24	5,00E-42	(0.000)	23	2,00E-49	(0.000)	5	9,00E-11	(0.000)
25	3,00E-48	(0.000)	5	6,00E-53	(0.000)	26	1,00E-37	(0.000)
26	5,00E-49	(0.000)	24	3,00E-57	(0.000)	13	5,00E-44	(0.000)
8	3,00E-57	(0.000)	8	4,00E-80	(0.000)	27	1,00E-44	(0.000)
23	2,00E-65	(0.000)	10	6,00E-100	(0.000)	3	8,00E-55	(0.000)
27	2,00E-94	(0.000)	7	2,00E-115	(0.000)	4	2,00E-78	(0.000)

Table 3. Hypothesis testing. AU test p-values obtained from the per-site log likelihoods calculated by A) RAxML (LG+Γ4) for each of the 27 root positions from the full AB matrix as well as the SF matrices SF11, SF25, and SF 27 and B) COaLA (nH model + Γ4). See text and M&M for details.

In conclusion, our analyses of the AB dataset highlight the existence of an original new rooting of the archaeal tree deep within Euryarchaeota, which is robustly supported by Bayesian (CAT+GTR+Γ4) analysis and is not rejected by Maximum likelihood (LG+Γ4) analysis, statistical tests, and a desaturation approach.

Is the phylogenetic signal of universal markers consistent with the AE and AB datasets?

As mentioned in the Introduction section, the trees obtained from universal datasets may be criticized. However, previous analyses included a very limited taxonomic sampling of bacteria and eukarya and were never tested against alternative topologies. Our ABE dataset contains a

very large taxonomic sampling and 9090 aa positions, around twice the size of previously published analyses, and a much larger taxonomic sampling. Moreover, although the ABE dataset partially overlaps with both the AE and AB datasets, we can specifically analyze the reliability of its phylogenetic signal combined over the three domains by comparing it to the results obtained with the AE and AB datasets.

Bayesian (CAT+GTR+ Γ4) and Maximum likelihood (LG+ Γ4) analysis provide largely resolved trees (Figure 7 and 8, respectively) with monophylies of major archaeal orders and bacterial and eukaryal phyla that are consistent with those obtained by the separate AE and AB analyses. Both trees indicate an emergence of Eukaryotes from within the TACK, specifically sister to Korarchaeota, albeit with very low support (pp 0.61 Figure 7 and 38% BV, Figure 8). Our extensive analysis of the AE dataset now indicates that the sisterhood between Eukaryotes and Korarchaeota observed in our and previously published universal trees is likely artifactual. Interestingly, when Korarchaeon is removed from the dataset, the ML tree switches to a classical 3Domains topology, with the Archaea becoming monophyletic (albeit with low confidence, BV 82%, not shown), revealing that the ML support for Eukaryotes within Archaea in universal trees containing Korarchaeota is due to an artifact. In contrast, the Bayesian tree remains stable when Korarchaeon is removed from the dataset, with Eukarya staying sister to TAC (not shown). So, the emergence of Eukaryotes from within Archaea in Bayesian analysis of universal markers with CAT+GTR+Γ4 appears robust against changes in the taxonomic sampling.

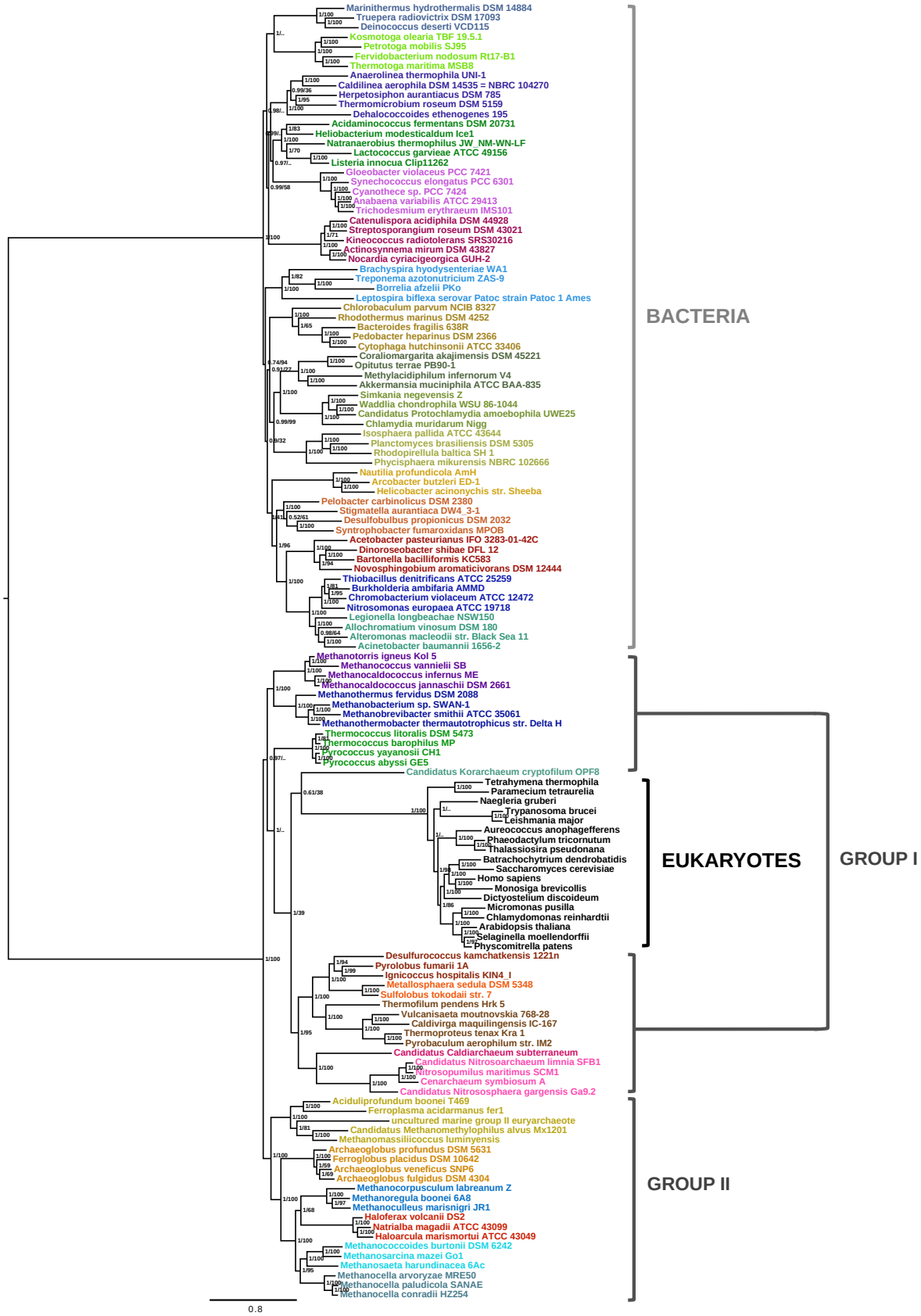


Figure 7. Bayesian phylogeny of a concatenated data set of the ABE supermatrix (9090 amino acid positions). The tree was calculated by Phylobayes (CAT+GTR+Γ4). The scale bar represents the average number of substitutions per site. Values at nodes represent posterior probabilities. See text and M&M for details.

Figure 8. Unrooted ML phylogeny of the ABE supermatrix (9090 amino acid positions). The tree was calculated by PhyML (LG+Γ4). The scale bar represents the average number of substitutions per site. Values at nodes represent supports obtained by non parametric bootstrapping based on 100 resamplings of the original alignment. See text and M&M for details.

Concerning the root of the Archaea, both ML and Bayesian analyses on the ABE dataset give results that are consistent with those obtained with the AB dataset. In fact, Maximum likelihood with the LG+Γ4 model still supports monophily of the Euryarchaeota (BV 99%), whereas Bayesian analysis with the CAT+GTR+Γ4 model supports paraphyly of Euryarchaeota with a split corresponding to root #17 observed in the AB tree (pp 0.97 and pp 1). However, when analyzing the desaturated ABE datasets with ML (LG+Γ4), we observe support for root #17 in the lowest saturation matrices and a clear switch to monophyly of Euryarchaeota in the more saturated matrices (not shown).

The consistency of the phylogenetic signal of the ABE dataset with that contained in the larger AB and AE datasets confirm its high quality, which may become a reference for future analyses aimed at reconstructing universal trees. More importantly, our results strongly indicate that an emergence of Eukaryotes from within Archaea by using universal markers is strongly supported by Bayesian+CAT+GTR+Γ4 and is systematically accompanied by a root within Euryarchaeota, which is also brought to light when running ML on the less saturated datasets.

Discussion

Although debated in early reports (Lake et al. 1984; Lake 1988; Baldauf et al. 1996; Hashimoto and Hasegawa 1996; Tourasse and Gouy 1999), it is only recently that an emergence of Eukaryotes from within Archaea has been brought back to the frontline by a number of analyses on universal proteins (Williams et al. 2013). Largely considered to be the result of better performing models of protein evolution, this “2 Domains” topology of the Tree of Life is more likely the fruit of the availability of new genomes from a larger representative of archaeal diversity, as it is also obtained with a wide range of classical approaches. The biological plausibility of 2D scenarios and their consequences have been extensively discussed in the literature (Embley and Martin 2006; Lombard et al. 2012; Guy et al. 2014). In particular, if Eukaryotes emerged from within the diversification of modern archaea, this unequivocally implies that the process of eukaryogenesis was initiated in an ancestor that harbored specific archaeal characteristics. It was recently proposed that such ancestor would have been an archaeon with that already possessed some eukaryotic characteristics such as the ability of phagocytosis (Martijn and Ettema 2013). In this respect, this model is similar to classical ones for eukaryogenesis apart from the starting point. However, it remains to be discussed what biological process and selection pressure would have accompanied the transition from specific G1P archaeal lipids to the classical G3P eukaryotic lipids. Moreover, it

has to be noted that basal archaeal lineages are all hyperthermophilic, including the TACK superphylum. Therefore, if Eukaryotes branched off deep within the Archaea, we also need to take into account how the earliest steps of eukaryogenesis may have taken place in such hot environments. If anything, a hyperthermophilic origin for Eukaryotes excludes the hypothesis that eukaryogenesis was triggered by mitochondrial symbiosis, as to our knowledge alphaproteobacteria are not usual inhabitants of such hot environments. Finally, we have to understand perhaps one of the most intriguing consequences of an archaeal origin for eukaryotes, i.e. the scattered distribution of eukaryotic characters in different archaeal lineages. We exhaustively searched for homologs of eukaryotic proteins inferred to date back to the LECA (the Last Eukaryotic Common Ancestor) that are not universally present in the Archaea (see M&M for details) (Figure 9). With respect to a previous analysis (Guy and Ettema 2011), we identified 10 novel markers (GTPases Arf1, NOG1, RPAP4/GPN1, and SAR1, PP2 metallophosphoesterase, a H/ACA RNA-protein complex component Nop10p, ribosomal L11 methyltransferase, 50S ribosomal protein L39e, RNase PH Rrp41, DNA helicase TIP49, and a RIO-like serine/threonine protein), opening future avenues of phylogenomic investigation on these cellular processes. In contrast, the dicer helicase (ERCC4), ribosomal protein L30e, and the multiprotein bridging factor 1 (MBF1) were not included because they are present in all phyla. RpoG was also not included because it did not meet our threshold requirement.

Importantly, no archaeal lineage appears to be specifically enriched in eukaryotic characters, except perhaps for a few ribosomal proteins in Crenarchaeota (Figure 9). Therefore, this argument cannot be used to indicate a particular archaeal group at the origin of Eukaryotes. Further analysis will be needed to assess if this scattered distribution is due to ancient horizontal gene transfer between eukaryotes and archaea, followed by a dynamic evolutionary process involving, for example, additional transfer among archaea possibly coupled with gene losses.

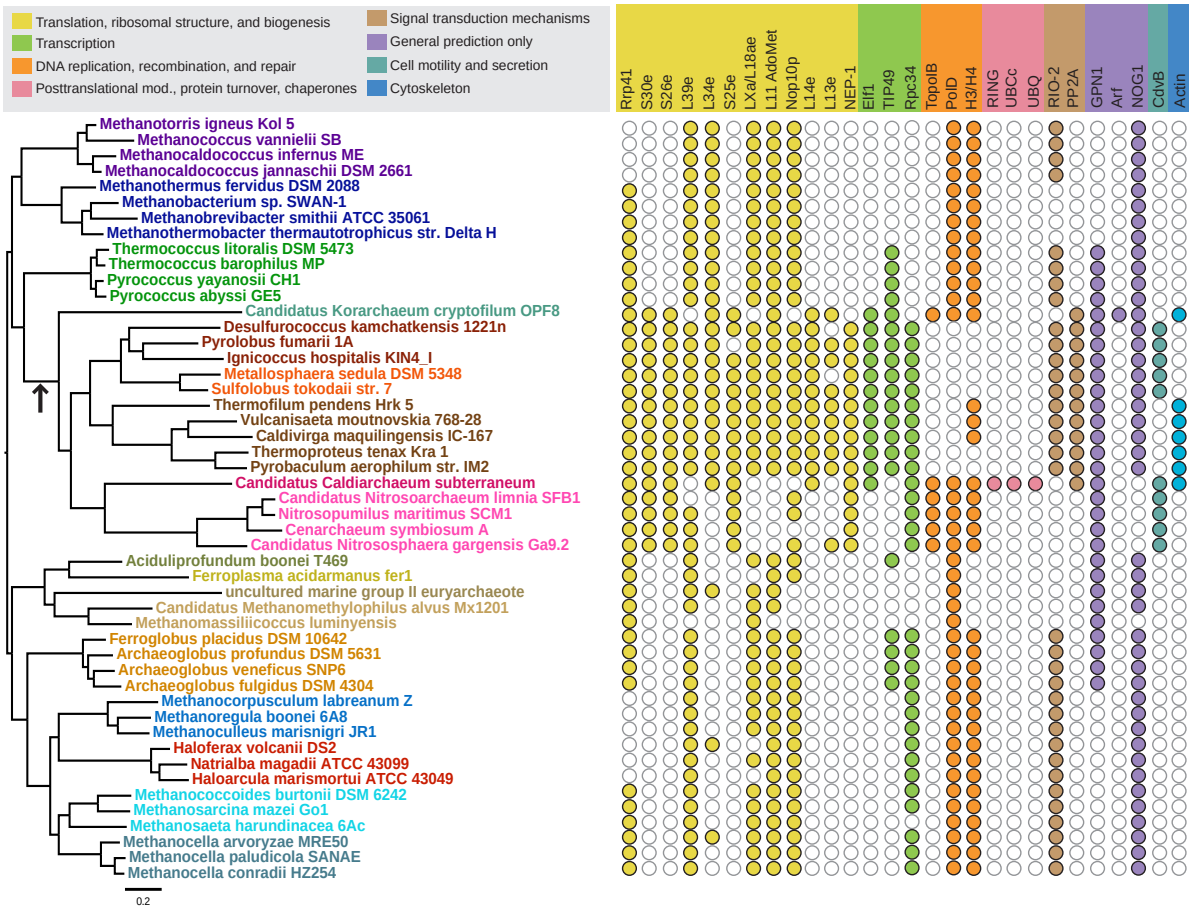


Figure 9. Scattered distribution of eukaryotic characters in the Archaea. The arrow indicates the potential branching of Eukaryotes.

The recent phylogenomic support for an archaeal origin of Eukaryotes represents therefore an important shift in paradigm that requires confirmation by the most thorough and accurate approaches. Here, we aimed at tackling the issue with a novel approach alternative and complementary to the use of universal proteins. Our results with the A/E dataset robustly exclude an emergence of Eukaryotes from within Euryarchaeota or TACK. In contrast, the phylogenetic signal harbored by the A/B dataset appears more delicate to extract and is prone to conflict depending on the model and approach used. Indeed, Bayesian analysis with CAT+GTR seems robust in indicating an original and novel rooting of the Archaeal tree within Euryarchaeota, which would definitely indicate an emergence of Eukaryotes as sister to the TACK. ML with the LG model shows evidence of a more unstable and conflicting signal but nevertheless does not reject this root, and even supports it in the less saturated datasets. Moreover, analysis of the ABE dataset strongly indicates that support for an emergence of Eukaryotes as sister to the TACK is systematically accompanied with this original rooting

within Euryarchaeota. Therefore, if we want to embrace an origin of Eukarya from within Archaea, we also need to accept this dramatically novel root for the third Domain of Life.

A rooting within Euryarchaeota has sometimes been observed in previous universal trees, albeit in in different branches depending on the dataset and approach, and never discussed(Cox, Cox, et al. 2008; Foster, Cox, and Embley 2009b; Williams et al. 2012; Lasek-Nesselquist et al. 2013). These analyses were based on universal proteins with a very limited taxonomic sampling of bacteria and eukarya, and frequently included very fast evolving archaeal taxa, which may lead to concerns of tree reconstruction artifacts. Here, this new rooting is revealed with an original approach, high quality datasets of markers, and a large array of strategies to extract reliable phylogenetic signal.

If confirmed, these results will provoke a real revolution in our view of the evolution of the third domain of life, not to mention the systematics, as the archaeal tree would virtually turn upside down with respect to what usually assumed (Figure 9). Classically considered as the earliest emergence in the Euryarchaeota, the Thermococcales would move away from their basal position to become the closest relatives to the branch from which the TACK (and Eukaryotes) originated. The nature of the last archaeal common ancestor (LACA) will have to be reconsidered. This may have harbored a number of features now present in modern Euryarchaeota, whereas the characters specific of TACK or TACK+Eukaryotes would have originated later and will not need to be inferred back to the LACA. The capacity to perform methanogenesis would become an ancestral characteristic in the Archaea, which would have been subsequently lost a large number of times independently along their diversification, even more than previously suspected (Borrel et al. 2013). Similarly, all phylogenomic analyses aimed at the reconstruction of ancestral features and their evolution along archaeal diversification (e.g. optimal growth temperatures, dynamics of cellular processes and structures, emergence of various metabolisms, inference of gene losses, duplications, horizontal gene transfers, etc.) may sensibly change.

A nonmonophyly of Euryarchaeota does not appear shocking from a genomic point of view. In fact, all the features initially considered as defining the line between Euryarchaeota and Crenarchaeota such as for instance the cell division protein FtsZ, the main replicative DNA polymerase PolD, eukaryotic-like histones, have now been found in other phyla, notably the Thaumarchaeota or the Korarchaeota. Rather, we now may want to search for markers that are specific for either Group I or Group II archaea. One intriguing character that may support this split is the presence of a DNA gyrase of bacterial type in Group II archaea, inferred to have arisen from a single horizontal gene transfer event from bacteria and supposed to have had a

dramatic effect on the global DNA topology of the recipient archaeon (Raymann et al. 2014). However, under the new archaeal root, no horizontal gene transfer would have to be inferred because this gyrase may have been present in the ancestor of all archaea and inherited only in Group II. Finally, it has to be stressed that a root within Euryarchaeota does not provide a solution to the puzzling distribution of eukaryotic characters in the archaea because it still requires a number of horizontal/gene transfers and/or losses.

Conclusions

The quest for eukaryotic origins is a difficult one. The availability of novel data from the archaea combined with the development of original approaches and improved evolutionary models has brought much progress to the issue. Unexpectedly, it also seems to blow a wind of novelty on our image of the very evolution of the Archaea, which, 40 years after their discovery, have not yet revealed all of their secrets. Our study confirms that the study of ancient evolution cannot be disconnected from a global vision on the whole Tree of Life, and will bring many exciting discoveries in the years to come. In particular, it will be critical to further explore not only the diversity of the TACK as probably the closest relatives of eukaryotes, but also that of the basal euryarchaeota lineages, in particular methanogens, as this will bring new information of the ancient history of one the most enigmatic of the three Domains of Life.

Materials and Methods

Dataset construction

Local databases were constructed using 132, 211, and 31 complete or nearly complete archaeal, bacterial, and eukaryotic genomes respectively which were downloaded from the National Center for Biotechnology Information (NCBI). An all-against-all BLASTp (Altschul et al. 1997) search was performed on the local archaeal database. Homologous sequences were clustered into protein families using SiLiX (Miele et al. 2011) a minimum thresholds of 70% length overlap and 30% identity. Large paralogous protein families were split using the MCL clustering algorithm as implemented in BioLayout (Theocharidis et al. 2009). Taxonomic distribution of each protein family in archaea was determined and the eventual absence of proteins was verified by additional Blastp and tBlastn searches. Protein families present in more than 95% of the analyzed archaeal genomes were kept for further analysis. This

resulted in a total of 230 protein families which were then used as seeds to perform HMMER searches (Johnson et al. 2010) on the local database of eukaryotes and bacteria. If present in more than 90% of the eukaryotic and/or bacterial genomes, multiple alignments were constructed, trimmed and subjected to phylogenetic analysis (archaea alone, bacteria alone, archaea/bacteria, archaea/eukaryotes, and archaea/bacteria/eukaryotes). All multiple alignments were performed with MUSCLE v3.8.31 (Edgar 2004) and manually inspected using SEAVIEW (Gouy et al. 2010). Single protein datasets were trimmed using the software BMGE (Criscuolo and Gribaldo 2010) with a BLOSUM30 matrix. Proteins that contained more than 70 amino acid positions after trimming were not retained for further analysis.

All individual trees were manually inspected. A protein family was retained if it met three criteria: the monophyly of archaea in AB trees, the monophyly of bacteria in AB trees, and the monophyly of all the major phyla of archaea and bacteria. Additionally, the trees were inspected for eukaryotic proteins of bacterial/mitochondrial origin. When multiple copies of a protein were present, the less divergent homolog was retained. This resulted in a total of 49 Archaea/Bacteria markers, 74 Archaea/Eukarya markers, and 39 Universal markers. In order to decrease the computational time of the analysis the taxonomic sampling was then reduced to 49 archaea, 67 bacteria, and 18 eukaryotes, keeping at least three representatives for each order when possible. To ensure that all proteins contained congruent signal, Prunier (Abby et al. 2009) was used (bootstrap threshold 70%, depth 2). The phylogenetic trees tested by Prunier were built using RAxML v 7.2.8 with 1000 rapid bootstraps (Stamatakis 2006). All sequences detected as transfers affecting nodes external to orders were removed from the alignments and the 49 Archaea/Bacteria markers, 74 Archaea/Eukarya markers, and 39 Universal markers were concatenated into three large supermatrices, allowing up 10 percent of missing taxa for each protein. The final concatenated datasets consisted of, AB (46 proteins and 10986 positions), AE (72 proteins and 17892 positions), and ABE (37 proteins and 9090 positions).

Phylogenetic analysis

Phylogenies of individual proteins were inferred using PhyML v. 3.1 (Guindon et al. 2010) and RAxML v 7.2.8 (Stamatakis 2006). Single matrix substitution models were chosen using the ProteinModelSelection script available from the RAxML website (http://sco.h-its.org/exelixis/software.html). In order to take into account the heterogeneity of evolutionary rates across sites, we used a gamma distribution with four discrete classes of sites (Γ4) and an estimated alpha parameter. The branch robustness of the ML trees was

estimated with the non-parametric bootstrap procedure implemented in PhyML (100 replicates of the original dataset), and the rapid bootstrap method implemented in RAxML (500 replicates of the original dataset).

Phylogenetic trees of the supermatrices were subjected to Maximum Likelihood and Bayesian analysis. Single matrix substitution models were chosen using the ProteinModelSelection script available from the RAxML website (http://sco.h-its.org/exelixis/software.html). PhyloBayes 3.3b (Lartillot et al. 2009) was used to perform Bayesian analysis using the CAT+Γ4 and CAT+GTR+Γ4 models and a gamma distribution with four categories of evolutionary rates was used to model the heterogeneity of site evolutionary rates. The concatenated datasets were also recoded using the Dayhoff6 recoding scheme as implemented in PhyloBayes 3.3b (Lartillot et al. 2009) and analyzed with the same model parameters. For each data set, two independent chains were run until convergence. Convergence was assessed by evaluating the discrepancy of bipartition frequencies and between independent runs. The first 25% of trees were discarded as burn in and the posterior consensus was computed by selecting one tree out of every two to compute the 50% majority consensus tree. Maximum likelihood analyses were performed using PhyML v 3.1 (Guindon et al. 2010) and RAxML v 7.2.8 (Stamatakis 2006), the LG+Γ4 model. The branch robustness of the ML trees was estimated with the non-parametric bootstrap procedure implemented in PhyML (100 replicates of the original dataset), and the rapid boostrap method implemented in RAxML (500 replicates of the original dataset). The ABE dataset was also analyzed with and without Korarchaeota using all methods described.

Hypothesis testing

Candidate topologies were created using TreeGraph2 (Stoever and Mueller 2010). The per-site log likelihood values were calculated for each topology under the LG+Γ4 model using RAxML, and under a branch heterogeneous substitution model with +Γ4 using COaLA (Groussin et al. 2013). The alternative topologies were then statistically evaluated using the approximately unbiased (AU) test (Shimodaira 2002) as implemented in CONSEL (Shimodaira and Hasegawa 2001).

Desaturation analysis

A site-by-site desaturation of all three supermatrices was carried out using the Slow-Fast method (Brinkmann and Philippe 1999). To do this, the sequences from each dataset were subdivided into known monophyletic groups (16 bacterial phyla, 12 archaeal orders/phyla,

and 4 eukaryotic groups) (see supplementary Table 1 for details). All of the considered groups contained 3 or more taxa except Korarchaoeota, which was considered alone. The evolutionary rate at sites was estimated with the program SlowFaster, which uses Maximum parsimony to count the number of changes of every alignment position within the selected groups (Kostka et al. 2007). For the AE dataset 42 (SF_0 to SF_{41}) matrices of increasing size were built by progressively incorporating more fast-evolving positions into the alignment. The same was done for the AB dataset (SF_0 –SF_{65}) and ABE dataset (S_0-S_{75}). We then extracted the matrices with a fixed increase in the number of positions (AB +350 aa, AE +500 aa, ABE +250). This resulted in 30 matrices for AB, 28 matrices for AE, 34 matrices for ABE that were subjected to Maximum Likelihood (LG+Γ4) and Bayesian analysis (CAT and CAT+GTR with Γ4). The support values (bootstrap value or posterior probability) at defined nodes of interest were recorded and plotted using R (R Development Core Team 2014).

Acknowledgments

The authors thank the PRABI (Pole Rhone-Alpes de Bioinformatique) for providing computing facilities. K.R. is a scholar from the Pasteur–Paris University (PPU) International PhD program and received a stipend from the Paul W. Zuccaire Foundation. C.B.A. is member of the Institut Universitaire de France. This work was supported by the Investissement d'Avenir grant "Ancestrome" (ANR-10- BINF- 01-01).

References

Abby SS, Tannier E, Gouy M, Daubin V. 2009. Detecting lateral gene transfers by statistical reconciliation of phylogenetic forests. BMC Bioinformatics 11:324–324.

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

Baldauf SL, Palmer JD, Doolittle WF. 1996. The root of the universal tree and the origin of eukaryotes based on elongation factor phylogeny. Proc. Natl. Acad. Sci. U.S.A. 93:7749–7754.

Borrel G, O'Toole PW, Harris HMB, Peyret P, Brugere J-F, Gribaldo S. 2013. Phylogenomic Data Support a Seventh Order of Methylotrophic Methanogens and Provide Insights into the Evolution of Methanogenesis. Genome biology and evolution 5:1769–1780.

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

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

Chiari Y, Cahais V, Galtier N, Delsuc F. 2011. Phylogenomic analyses support the position of turtles as the sister group of birds and crocodiles (Archosauria). BMC Biol. 10:65–65.

Cox CJC, Cox CJC, Foster PGP, et al. 2008. The archaebacterial origin of eukaryotes. Audio, Transactions of the IRE Professional Group on 105:20356–20361.

Cox CJC, Foster PGP, Hirt RPR, Harris SRS, Embley TMT. 2008. The archaebacterial origin of eukaryotes. Audio, Transactions of the IRE Professional Group on 105:20356–20361.

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 Evol. Biol. 10:210.

Edgar RC. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5:113.

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

Foster PG, Cox CJ, Embley TM. 2009a. The primary divisions of life: a phylogenomic approach employing composition-heterogeneous methods. Philos Trans R Soc Lond B Biol Sci 364:2197–2207.

Foster PGP, Cox CJC, Embley TMT. 2009b. The primary divisions of life: a phylogenomic approach employing composition-heterogeneous methods. Philos Trans R Soc Lond B Biol Sci 364:2197–2207.

Gouy M, Guindon S, Gascuel O. 2010. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol. Biol. Evol. 27:221–224.

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

Gribaldo SS, Philippe HH. 2002. Ancient phylogenetic relationships. Theor Popul Biol 61:391–408.

Groussin M, Boussau B, Gouy M. 2013. A Branch-Heterogeneous Model of Protein Evolution for Efficient Inference of Ancestral Sequences. Syst. Biol. 62:523–538.

Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. 59:307–321.

Guy L, Ettema TJ. 2011. The archaeal “TACK” superphylum and the origin of eukaryotes. Trends Microbiol. 19:8–8.

Guy L, Saw JH, Ettema TJG. 2014. The Archaeal Legacy of Eukaryotes: A Phylogenomic Perspective. Cold Spring Harb Perspect Biol.

Hashimoto T, Hasegawa M. 1996. Origin and early evolution of eukaryotes inferred from the amino acid sequences of translation elongation factors 1α/Tu and 2/G. Advances in biophysics.

Johnson LS, Eddy SR, Portugaly E. 2010. Hidden Markov model speed heuristic and iterative HMM search procedure. BMC Bioinformatics 11:431.

Kostka M, Uzlikova M, Cepicka I, Flegr J. 2007. SlowFaster, a user-friendly program for slow-fast analysis and its application on phylogeny of Blastocystis. BMC Bioinformatics 9:341–341.

Lake JA. 1988. Origin of the eukaryotic nucleus determined by rate-invariant analysis of rRNA sequences. CORD Conference Proceedings 331:184–186.

Lake JAJ, Henderson EE, Oakes MM, Clark MWM. 1984. Eocytes: a new ribosome structure indicates a kingdom with a close relationship to eukaryotes. Proc. Natl. Acad. Sci. U.S.A. 81:3786–3790.

Lartillot N, Lepage T, Blanquart S. 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics 25:2286–2288.

Lartillot N, Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol. Biol. Evol. 21:1095–1109.

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.

Lasek-Nesselquist E, Pisani DD, Gogarten JP, Cotton JAJ, McInerney JOJ. 2013. The effects of model choice and mitigating bias on the ribosomal tree of life. Mol Phylogenet Evol 69:17–38.

Le SQ, Gascuel O. 2008. An improved general amino acid replacement matrix. Mol. Biol. Evol. 25:1307–1320.

Lombard J, López-García P, Moreira D. 2012. The early evolution of lipid membranes and the three domains of life. Nat. Rev. Microbiol. 10:507–515.

Martijn J, Ettema TJG. 2013. From archaeon to eukaryote: the evolutionary dark ages of the eukaryotic cell. Biochem. Soc. Trans. 41:451–457.

Miele V, Penel S, Duret L. 2011. Ultra-fast sequence clustering from similarity networks with SiLiX. BMC Bioinformatics 12:–116.

Philippe H, Forterre P. 1999. The rooting of the universal tree of life is not reliable. J. Mol. Evol. 49:509–523.

Philippe H, Sörhannus U, Baroin A. 1994. Comparison of molecular and paleontological data in diatoms suggests a major gap in the fossil record. Journal of ….

Ramulu HG, Groussin M, Talla E, Planel R, Daubin V, Brochier-Armanet C. 2014. Ribosomal proteins: Toward a next generation standard for prokaryotic systematics? Mol Phylogenet Evol 75:103–117.

Raymann K, Forterre P, Brochier-Armanet C, Gribaldo S. 2014. Global phylogenomic analysis disentangles the complex evolutionary history of DNA replication in archaea. Genome biology and evolution 6:192–212.

Shimodaira H, Hasegawa M. 2001. CONSEL: for assessing the confidence of phylogenetic tree selection. Bioinformatics 17:1246–1247.

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

Stoever BC, Mueller KF. 2010. TreeGraph 2: Combining and visualizing evidence from different phylogenetic analyses. BMC Bioinformatics 11:–7.

Theocharidis A, van Dongen S, Enright AJ, Freeman TC. 2009. Network visualization and analysis of gene expression data using BioLayout Express(3D). Nat Protoc 4:1535–1550.

Tourasse NJN, Gouy MM. 1999. Accounting for Evolutionary Rate Variation among Sequence Sites Consistently Changes Universal Phylogenies Deduced from rRNA and Protein-Coding Genes. Mol Phylogenet Evol 13:10–10.

Wagner M, Horn M. 2005. ThePlanctomycetes,Verrucomicrobia,Chlamydiaeand sister phyla comprise a superphylum with biotechnological and medical relevance. Curr Opin Biotechnol 17:241–249.

Williams TA, Embley TM. 2014. Archaeal “Dark Matter” and the Origin of Eukaryotes. Genome biology and evolution 6:474–481.

Williams TA, Foster PG, Cox CJ, Embley TM. 2013. An archaeal origin of eukaryotes supports only two primary domains of life. Nature 504:231–236.

Williams TAT, Foster PGP, Nye TMWT, Cox CJC, Embley TMT. 2012. A congruent phylogenomic signal places eukaryotes within the Archaea. Proc Biol Sci 279:4870–4879.

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. U.S.A. 87:4576–4579.

Discussion and Perspectives

Nearly 40 years after their discovery, the genomic revolution has revealed that the archaea are very diverse organisms. The archaea have now been found in nearly every habitat imaginable, and it is very likely that many more archaeal lineages are yet unknown. Unfortunately, in some ways, the fact that no archaea pathogens have been discovered has affected the interest to study this domain, and it is sad that most of the general public is still ignorant of the existence of the archaea. However, the discovery that the archaea play key roles in important biogeochemical process and are abundant in the microbiome of animals, including ours, has sparked interest in studying them, and in the last few years several metagenomic analyses have helped in revealing more about the diversity and prevalence of the archaea.

During my thesis I have contributed to our understanding of archaeal diversity and evolution in several ways. First I analyzed a key cellular system, DNA replication, which has a very complex evolutionary history making it difficult to dissect. By accurately identifying orthologous replication proteins I was able to obtain a robust archaeal phylogeny that is consistent with two other key cellular systems, translation and transcription, which have been successfully used to reconstruct the archaeal phylogeny. The robust phylogeny obtained from the core DNA replication proteins further testifies that a phylogeny of the archaea does exist even if it can be difficult to extract. This dataset now constitutes an additional gene set that can be used in further analyses to reconstruct the phylogeny in archaea in light of new genomes. In fact, my dataset has already been used in two collaborations for to determine the placement of newly discovered lineages.

By analyzing the components of DNA replication I was able to gain more information on some unresolved nodes in the archaeal tree, such as the evolutionary relationships and placement of fast evolving nanosized lineages However, these relationships clearly need to be analyzed further by specific analyses such as whole genome comparisons and additional markers. It is also possible that in the future the availability of additional complete genomes from slow evolving related lineages may help clarify the issue.

During this analysis, I also made a few interesting observations that should be investigated experimentally. For example, I highlighted the existence of a conserved second origin

recognition protein Cdc6/Orc-1 copy that is ancestral and present in nearly all archaeal genomes and may therefore have a key and possibly regulatory function. Also, the presence of both single stranded binding proteins RPA and SSB in some archaeal lineages such as the Thaumarchaeota is intriguing and should be experimentally studied. In fact many metazoans, including men, also possess in addition to RPA proteins archaeal-like SSB homologs with a potential and unclear involvement in disease, but their possible interaction has never been evoked.

Integrative elements appear to specifically target the archaeal DNA replication machinery and are likely major players in the dynamic evolutionary history of this cellular process. It would be interesting to see if the same is true for Eukaryotes and Bacteria, or if this is a unique phenomenon in the Archaea. Finally, given the complex history of this system, the data obtained during this study may represent a good dataset for testing the accuracy of DTL (duplication, loss, and transfer) models (Szöllosi et al. 2012; Szöllosi and Daubin 2012; Szöllosi et al. 2013; Boussau et al. 2013).

Our lack of knowledge about the true diversity of the archaea greatly hinders the ability of understand the relationship between archaea and eukaryotes. Since the discovery of the archaea researchers have been trying to understand their exact relationship to eukaryotes. Knowing more about the diversity of archaea can aid in our understanding of this relationship. However, as eukaryotic organisms ourselves we focus on how we have come into existence and we sometimes forget that we must look much deeper to get the full picture. When I started my thesis I myself was focusing specifically on understanding how the eukaryotic cell originated, but during the course of my Ph.D I found myself realizing that in order to discern the origin of eukaryotes we first need to have a full picture of the organism that are more closely related to them and from which they even might originate.

Part of my thesis involved applying a novel approach to investigate the relationships between archaea and eukaryotes. For the first time, I separately analyzed the Archaea and Eukaryotes on one side and the Archaea and Bacteria on the other. This resulted in three very high quality datasets that included more amino acid positions, proteins, and taxa than previously used in large-scale universal studies. The progresses in technology allowed me to analyze these large datasets using sophisticated approaches and evolutionary models that would have been unthinkable to process in a reasonable computing time until recently. Nevertheless, these analyses took a very large computational power and several months of computing time and

some of the more sophisticated nonhomogeneous evolutionary models could not be applied because of the size of the dataset. With the exponential increase in computer technologies it is likely that in a few years we will be able to analyze these datasets with even more sophisticated models of evolution and include many more taxa.

Although larger than previously used, my datasets were still limited in taxonomy coverage because I only aimed at resolving the overall relationships among archaea and eukaryotes. However, they resulted in very well resolved phylogenies for all three domains and can be a reference for many future studies. As of now, the root of the eukaryotic tree and the evolutionary relationships among the major phyla are still unclear. The same is true for the bacterial phylogeny. My Archaea/Bacteria and Archaea/Eukaryote datasets could be used not only to root but also refine the phylogenies of both the Bacteria and the Eukaryotes.

During this study I had an unexpected result: I highlighted an original root of the archaeal tree that supports an emergence of eukaryotes from within the archaeal domain, but also dramatically changes our perspectives on the origin and evolution of the archaea and opens a new avenue of research on eukaryotic evolution and the early steps of eukaryogenesis. This new root also has a large impact on all future studies aimed at understanding the nature of the Last Archaeal Common Ancestor (LACA). For example, by placing the Methanogens Class I as early branchings in the archaeal phylogeny, it opens up the possibility that methanogenesis is more ancient than previously thought. This new root could also have an impact on the inferred optimal growth temperature (OGT) of the LACA. In fact, previously published predictions of the OGT were mapped onto a tree rooted between the Euryarchaeota and the TACK, resulting in the inference of a hyperthermophilic ancestor (approx. 86 degrees) as well as a hyperthermophilic ancestor of euryarchaeota with a progressive decrease in OGT in the euryarchaeal lineage (Groussin and Gouy 2011). However, if the root is placed within the euryarchaeota, in particular around lineages where the optimal growth temperature is much lower (between 40-60) then this may lead to lower inferred OGT in the LACA. This root would have an impact on all other phylogenomic studies, including studies predicting ancestral genome content ancestral gene content. Given the groundbreaking nature of this result, it will of course be necessary to support it with additional studies. In particular, the need for the bacterial out-group limits the number of proteins that can be used for phylogenetic analysis. The limitation in the number of proteins that can be used is mostly due to the extensive amount of horizontal gene transfer that has occurred among and between these two prokaryotic domains. In fact, of the approximately 250 genes that I identified as present in

over 95% of the archaeal genomes, more than 30% of them were discarded either because of horizontal transfers detected within the archaea or between archaea and bacteria. Although these transferred genes could not be used in my analysis, they could be indeed very useful with different approaches. In fact, new methods have recently been developed which attempt to jointly infer gene trees and species trees, which account for duplication, loss, and HGT (Szöllosi et al. 2012; Szöllosi and Daubin 2012; Szöllosi et al. 2013; Boussau et al. 2013). One way to test our new root for the archaea would be to apply these methods and determine whether this new topology reduces the amount of HGT that is needed to reconcile the gene trees with the species tree. In addition, it may be envisaged to use the whole core of the archaea and methods that do no need an outgroup, for example branch-heterogeneous models that allow lineages to diverge toward different amino acid compositions making the position of the root affect the likelihood value.

My analysis also involved defining a "relaxed" core genome for each of the major archaeal phyla. This analysis needs to be refined and expanded to include the cores of each lineage, especially if the Euryarchaeota are confirmed not to be a monophyletic group. In particular, the thresholds used for defining these cores were very stringent because it was important to use proteins present in nearly all archaeal genomes, and also because I wanted to insure the accurate identification of orthologs for phylogenetic analysis. By comparing core genes present in different lineages many important things can be revealed. This would for example allow the identification of lineage specific genes, which might be important for understanding specific features such as adaptation to particular environments. It would also help to better understand the relationships between lineages, specifically those that are still unresolved, such as the relationship among the nano-sized archaea. Reconstructing the core genome of each lineage can also reveal the impact of horizontal gene transfer on different lineages either from other archaeal lineages or from bacteria. Furthermore, the identification of proteins shared among different lineages could aid in clarifying the root of the archaeal tree.

Another important result of my thesis was the identification of a number of new markers inferred to have been present in the last common ancestor of eukaryotes (LACA) but instead found in some archaeal lineages only. This opens up whole new avenues of research involving the full phylogenomic dissection of the cellular processes they are involved in, from both the archaeal and eukaryotic sides. Moreover, it is unclear whether these proteins were present in the archaeal ancestor and were subsequently independently lost in different lineages, or if

they are the result of ancient horizontal gene transfers between archaea and eukaryotes. This needs to be investigated by further phylogenomic analyses, of these proteins in both eukaryotes and archaea. Careful analysis of these scattered proteins could help test the hypothesis of a more complex ancestor as well as help reveal the frequency of transfers between eukaryotes and prokaryotes. Moreover, if the eukaryotes indeed originated from within the archaea we really need new hypotheses to explain this from a biological point of view. Although some models have been put forth there are still unresolved issues such as the transition from archaeal-type to eukaryotic-type lipids and a hyperhtermophilic origin for eukaryotes.

Now, one of the most important steps is to further explore the diversity of life. With new technology the amount of genomes available is overwhelming, but there needs to be more focus on discovering and sequencing unexplored branches of the tree of life. Only by understanding the real diversity of life will we ever begin to unravel the most fundamental questions in evolutionary biology.

References

Abby SS, Tannier E, Gouy M, Daubin V. 2009. Detecting lateral gene transfers by statistical reconciliation of phylogenetic forests. BMC Bioinformatics 11:324–324.

Albers S-V, Forterre P, Prangishvili D, Schleper C. 2013. The legacy of Carl Woese and Wolfram Zillig: from phylogeny to landmark discoveries. Nat. Rev. Microbiol. 11:713–719.

Allsopp A. 1969. SCBBZY. New Phytologist.

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

Aristotle. 1993. Historia animalium. Loeb Classical Library

Baker BJ, Baker BJ, Comolli LR, et al. 2010. Enigmatic, ultrasmall, uncultivated Archaea.

Baldauf SL, Palmer JD, Doolittle WF. 1996. The root of the universal tree and the origin of eukaryotes based on elongation factor phylogeny. Proc. Natl. Acad. Sci. U.S.A. 93:7749–7754.

Baldauf SL. 2003. The deep roots of eukaryotes. Science 300:1703–1706.

Bapteste E, Boucher Y, Leigh J, Doolittle WF. 2003. Phylogenetic reconstruction and lateral gene transfer. Trends Microbiol. 12:406–411.

Bapteste E, Boucher Y. 2007. Lateral gene transfer challenges principles of microbial systematics. Trends Microbiol. 16:200–207.

Bapteste E, Brochier C. 2004. On the conceptual difficulties in rooting the tree of life. Trends Microbiol. 12:9–13.

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

Bapteste E, Charlebois RL, MacLeod D, Brochier C. 2005. The two tempos of nuclear pore complex evolution: highly adapting proteins in an ancient frozen structure. Genome Biol. 6:R85–R85.

Bapteste E, Susko E, Leigh J, MacLeod D, Charlebois RL, Doolittle WF. 2005. Do orthologous gene phylogenies really support tree-thinking? BMC Evol. Biol. 5:33.

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

Beman JM, Popp BN, Francis CA. 2008. Molecular and biogeochemical evidence for ammonia oxidation by marine Crenarchaeota in the Gulf of California. ISME J 2:429–441.

Ben J Mans, Anantharaman V, Aravind L, Koonin EV. 2004. Comparative genomics, evolution and origins of the nuclear envelope and nuclear pore complex. Cell Cycle 3:1612–1637.

Blanquart S, Lartillot N. 2008. A site- and time-heterogeneous model of amino acid replacement. Mol. Biol. Evol. 25:842–858.

Borrel G, Harris HMB, Tottey W, et al. 2012. Genome sequence of "Candidatus Methanomethylophilus alvus" Mx1201, a methanogenic archaeon from the human gut belonging to a seventh order of methanogens. J. Bacteriol. 194:6944–6945.

Borrel G, O'Toole PW, Harris HMB, Peyret P, Brugere J-F, Gribaldo S. 2013. Phylogenomic Data Support a Seventh Order of Methylotrophic Methanogens and Provide Insights into the Evolution of Methanogenesis. Genome biology and evolution 5:1769–1780.

Boussau B, Blanquart S, Necsulea A, Lartillot N, Gouy M. 2008. Parallel adaptations to high temperatures in the Archaean eon. Nature 456:942–945.

Boussau, Bastien, Gergely J Szöllosi, Laurent Duret, Manolo Gouy, Eric Tannier, and Vincent Daubin. 2013. "Genome-Scale Coestimation of Species and Gene Trees.." *Genome Research* 23 (2): 323–30. doi:10.1101/gr.141978.112.

Breed RS, Hitchens AP, Murray EGD. 1948. Bergey's Manual of Determinative Bacteriology.

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

Brochier C, Bapteste E, Moreira D, Philippe H. 2002. Eubacterial phylogeny based on translational apparatus proteins. Trends Genet. 18:1–5.

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

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

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

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

Brochier-Armanet C, Forterre P, Gribaldo S. 2011. Phylogeny and evolution of the Archaea:

one hundred genomes later. Curr. Opin. Microbiol. 14:274–281.

Brochier-Armanet C, Forterre P. 2007. Widespread distribution of archaeal reverse gyrase in thermophilic bacteria suggests a complex history of vertical inheritance and lateral gene transfers. Archaea 2:83–93.

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. Biol. Direct 3:54.

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

Brochier-Armanet C, Gribaldo S. 2007. Phylogenomics of the archaeal flagellum: rare horizontal gene transfer in a unique motility structure. BMC evolutionary \ldots.

Brochier-Armanet C, Talla E, Gribaldo S. 2009. The multiple evolutionary histories of dioxygen reductases: Implications for the origin and evolution of aerobic respiration. Mol. Biol. Evol. 26:285–297.

Brown JR, Doolittle WF. 1995. Root of the universal tree of life based on ancient aminoacyl-tRNA synthetase gene duplications. Proc. Natl. Acad. Sci. U.S.A. 92:2441–2445.

Brown JR, Douady CJ, Italia MJ, Marshall WE, Stanhope MJ. 2001. Universal trees based on large combined protein sequence data sets. Nature Genetics 28:281–285.

Brown JR, Robb FT, Weiss R, Doolittle WF. 1997. Evidence for the early divergence of tryptophanyl- and tyrosyl-tRNA synthetases. J. Mol. Evol. 45:9–16.

Brown JRJ, Doolittle WFW. 1997. Archaea and the prokaryote-to-eukaryote transition. Microbiol. Mol. Biol. Rev. 61:456–502.

Bult CJC, White OO, Olsen GJG, et al. 1996. Complete genome sequence of the methanogenic archaeon, Methanococcus jannaschii. Science 273:1058–1073.

Cammarano P, Creti R, Sanangelantoni AM, Palm P. 1999. The archaea monophyly issue: A phylogeny of translational elongation factor G(2) sequences inferred from an optimized selection of alignment positions. J. Mol. Evol. 49:524–537.

Cammarano P, Palm P, Creti R, Ceccarelli E, Sanangelantoni AM, Tiboni O. 1992. Early evolutionary relationships among known life forms inferred from elongation factor EF-2/EF-G sequences: phylogenetic coherence and structure of the archaeal domain. J. Mol. Evol. 34:396–405.

Canback B, Andersson S. 2002. The global phylogeny of glycolytic enzymes.

Cavalier-Smith T, Chao EE. 1996. Molecular phylogeny of the free-living archezoan

Trepomonas agilis and the nature of the first eukaryote. J. Mol. Evol. 43:551–562.

Cavalier-Smith T. 1987. Eukaryotes with no mitochondria. Nature 326:332–333.

Cavalier-Smith T. 2009. Origin of the cell nucleus, mitosis and sex: roles of intracellular coevolution. Biol. Direct 5:7–7.

Cheeseman P, Toms-Wood A, Wolfe RS. 1972. Isolation and properties of a fluorescent compound, factor 420 , from Methanobacterium strain M.o.H. J. Bacteriol. 112:527–531.

Chiari Y, Cahais V, Galtier N, Delsuc F. 2011. Phylogenomic analyses support the position of turtles as the sister group of birds and crocodiles (Archosauria). BMC Biol. 10:65–65.

Ciccarelli FD, Doerks T, Mering von C, Creevey CJ, Snel B, Bork P. 2006. Toward automatic reconstruction of a highly resolved tree of life. Science 311:1283–1287.

Collins L, Penny D. 2005. Complex spliceosomal organization ancestral to extant eukaryotes. Mol. Biol. Evol. 22:1053–1066.

Copeland HF. 1938. The kingdoms of organisms. The Quarterly Review of Biology.

Cox CJC, Cox CJC, Foster PGP, et al. 2008. The archaebacterial origin of eukaryotes. Audio, Transactions of the IRE Professional Group on 105:20356–20361.

Cox CJC, Foster PGP, Hirt RPR, Harris SRS, Embley TMT. 2008. The archaebacterial origin of eukaryotes. Audio, Transactions of the IRE Professional Group on 105:20356–20361.

Creti R, Ceccarelli E, Bocchetta M, Sanangelantoni AM, Tiboni O, Palm P, Cammarano P. 1994. Evolution of translational elongation factor (EF) sequences: reliability of global phylogenies inferred from EF-1 alpha(Tu) and EF-2(G) proteins. Proc. Natl. Acad. Sci. U.S.A. 91:3255–3259.

Crick F. 1958. On Protein Synthesis. The Symposia of the Society for Experimental Biology 12:138–163.

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 Evol. Biol. 10:210.

Csuros M, Miklos I. 2009. Streamlining and Large Ancestral Genomes in Archaea Inferred with a Phylogenetic Birth-and-Death Model. Mol. Biol. Evol. 26:2087–2095.

Dacks JB, Peden AA, Field MC. 2009. Evolution of specificity in the eukaryotic endomembrane system. The International Journal of Biochemistry & Cell Biology 41:330–340.

Darwin C. 2006. On the Origin of Species. Mundus Publishing

Daubin V, Gouy M, Perriere G. 2002. A phylogenomic approach to bacterial phylogeny: Evidence of a core of genes sharing a common history. Genome Res 12:1080–1090.

De Craene J-O, Ripp R, Lecompte O, Thompson JD, Poch O, Friant S. 2012. Evolutionary analysis of the ENTH/ANTH/VHS protein superfamily reveals a coevolution between membrane trafficking and metabolism. BMC Genomics 13:297.

DeLong EF. 1992. Archaea in coastal marine environments. Proc. Natl. Acad. Sci. U.S.A. 89:5685–5689.

DeLong EF. 2003. Oceans of Archaea. Asm News 69:503–511.

Delsuc FF, Brinkmann HH, Philippe HH. 2005. Phylogenomics and the reconstruction of the tree of life. Nat Rev Genet 6:361–375.

Desmond E, Brochier-Armanet C, Forterre P, Gribaldo S. 2011. On the last common ancestor and early evolution of eukaryotes: reconstructing the history of mitochondrial ribosomes. Res. Microbiol. 162:53–70.

Déclais AC, Marsault J, Confalonieri F, La Tour de CB, Duguet M. 2000. Reverse gyrase, the two domains intimately cooperate to promote positive supercoiling. J. Biol. Chem. 275:19498–19504.

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

Doolittle W. 1998. Lateral genomics. Trends Genet. 15:M5–M8.

Doolittle WF. 1999. Phylogenetic classification and the universal tree. Science 284:2124–2129.

Dridi B, Fardeau M-L, Ollivier B, Raoult D, Drancourt M. 2012. Methanomassiliicoccus luminyensis gen. nov., sp. nov., a methanogenic archaeon isolated from human faeces. Int J Syst Evol Microbiol 62:1902–1907.

Dutilh BE, Huynen MA, Bruno WJ, Snel B. 2004. The consistent phylogenetic signal in genome trees revealed by reducing the impact of noise. J. Mol. Evol. 58:527–539.

Edgar RC. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5:113.

Elkins JGJ, Podar MM, Graham DED, et al. 2008. A korarchaeal genome reveals insights into the evolution of the Archaea. Proc. Natl. Acad. Sci. U.S.A. 105:8102–8107.

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

Embley TM, van der Giezen M, Horner DS, Dyal PL, Bell S, Foster PG. 2003. Hydrogenosomes, mitochondria and early eukaryotic evolution. Iubmb Life 55:387–395.

Embley TM, van der Giezen M, Horner DS, Dyal PL, Foster P. 2003. Mitochondria and

hydrogenosomes are two forms of the same fundamental organelle. Philos Trans R Soc Lond B Biol Sci 358:191–192.

Eme L, Trilles A, Moreira D, Brochier-Armanet C. 2010. The phylogenomic analysis of the anaphase promoting complex and its targets points to complex and modern-like control of the cell cycle in the last common ancestor of eukaryotes. BMC Evol. Biol. 11:265–265.

Feng DFD, Cho GG, Doolittle RFR. 1997. Determining divergence times with a protein clock: update and reevaluation. Proc. Natl. Acad. Sci. U.S.A. 94:13028–13033.

Fitch WM, Margoliash E. 1967. Construction of phylogenetic trees. Science 155:279–284.

Forterre P. 2002. A hot story from comparative genomics: reverse gyrase is the only hyperthermophile-specific protein. Trends Genet. 18:236–237.

Foster PG, Cox CJ, Embley TM. 2009a. The primary divisions of life: a phylogenomic approach employing composition-heterogeneous methods. Philos Trans R Soc Lond B Biol Sci 364:2197–2207.

Foster PG. 2004. Modeling compositional heterogeneity. Syst. Biol. 53:485–495.

Foster PGP, Cox CJC, Embley TMT. 2009b. The primary divisions of life: a phylogenomic approach employing composition-heterogeneous methods. Philos Trans R Soc Lond B Biol Sci 364:2197–2207.

Fox GE, Magrum LJ, Balch WE, Wolfe RS, Woese CR. 1977. Classification of methanogenic bacteria by 16S ribosomal RNA characterization. Proc. Natl. Acad. Sci. U.S.A. 74:4537–4541.

Fox GE, Stackebrandt E, Hespell RB, et al. 1980. The phylogeny of prokaryotes. Science 209:457–463.

Francis CA, Roberts KJ, Beman JM, Santoro AE, Oakley BB. 2005. Ubiquity and diversity of ammonia-oxidizing archaea in water columns and sediments of the ocean. Proc. Natl. Acad. Sci. U.S.A. 102:14683–14688.

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

Fütterer OO, Angelov AA, Liesegang HH, Gottschalk GG, Schleper CC, Schepers BB, Dock CC, Antranikian GG, Liebl WW. 2004. Genome sequence of Picrophilus torridus and its implications for life around pH 0. Proc. Natl. Acad. Sci. U.S.A. 101:9091–9096.

Gabaldón T, Huynen MA. 2007. From endosymbiont to host-controlled organelle: the hijacking of mitochondrial protein synthesis and metabolism. PLoS computational biology 3:e219.

Galtier N, Tourasse N, Gouy M. 1999. A nonhyperthermophilic common ancestor to extant life forms. Science 283:220–221.

Gargaud M, Claeys P, López-García P, Martin H, Montmerle T, Pascal R, Reisse J. 2007. From

Suns to Life: A Chronological Approach to the History of Life on Earth. Springer Science & Business Media

Gaucher EA, Govindarajan S, Ganesh OK. 2008. Palaeotemperature trend for Precambrian life inferred from resurrected proteins. Nature 451:704–707.

Gogarten JP, Kibak H, Dittrich P, et al. 1989. Evolution of the vacuolar H+-ATPase: implications for the origin of eukaryotes. Proc. Natl. Acad. Sci. U.S.A. 86:6661–6665.

Goodman M, Czelusniak J, Moore GW. 1979. Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. Systematic \ldots.

Gouy M, Guindon S, Gascuel O. 2010. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol. Biol. Evol. 27:221–224.

Gouy M, Li WH. 1989. Phylogenetic analysis based on rRNA sequences supports the archaebacterial rather than the eocyte tree. Nature 339:145–147.

Gray MW, Doolittle WF. 1982. Has the endosymbiont hypothesis been proven? Microbiological reviews. Baltimore 46:1–42.

Gribaldo S, Brochier-Armanet C. 2006. The origin and evolution of Archaea: a state of the art. Philos Trans R Soc Lond B Biol Sci 361:1007–1022.

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. J. Mol. Evol. 47:508–516.

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

Gribaldo SS, Philippe HH. 2002. Ancient phylogenetic relationships. Theor Popul Biol 61:391–408.

Groussin M, Boussau B, Gouy M. 2013. A Branch-Heterogeneous Model of Protein Evolution for Efficient Inference of Ancestral Sequences. Syst. Biol. 62:523–538.

Groussin MM, Gouy MM. 2011. Adaptation to environmental temperature is a major determinant of molecular evolutionary rates in archaea. Mol. Biol. Evol. 28:2661–2674.

Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. 59:307–321.

Gupta RSR, Golding GBG. 1996. The origin of the eukaryotic cell. Trends Biochem. Sci. 21:166–

171.

Guy L, Ettema TJ. 2011. The archaeal “TACK” superphylum and the origin of eukaryotes. Trends Microbiol. 19:8–8.

Guy L, Saw JH, Ettema TJG. 2014. The Archaeal Legacy of Eukaryotes: A Phylogenomic Perspective. Cold Spring Harb Perspect Biol.

Haeckel E. 1866. Generelle Morphologie der Organismen.

Halachev MR, Loman NJ, Pallen MJ. 2011. Calculating Orthologs in Bacteria and Archaea: A Divide and Conquer Approach.Badger JH, editor. PLoS ONE 6:e28388.

Hammesfahr B, Kollmar M. 2011. Evolution of the eukaryotic dynactin complex, the activator of cytoplasmic dynein. BMC Evol. Biol. 12:95–95.

HARAUZ G, STOEFFLERMEILICKE M, VANHEEL M. 1987. Characteristic Views of Prokaryotic-50s Ribosomal-Subunits. J. Mol. Evol. 26:347–357.

Hashimoto T, Hasegawa M. 1996. Origin and early evolution of eukaryotes inferred from the amino acid sequences of translation elongation factors 1α/Tu and 2/G. Advances in biophysics.

Henneberger R, Moissl C, Amann T, Rudolph C, Huber R. 2005. New insights into the lifestyle of the cold-loving SM1 euryarchaeon: natural growth as a monospecies biofilm in the subsurface. Applied and Environmental Microbiology 72:192–199.

Hrdy I, Hirt RP, Dolezal P, Bardonová L, Foster PG, Tachezy J, Embley TM. 2004. Trichomonas hydrogenosomes contain the NADH dehydrogenase module of mitochondrial complex I. Nature 432:618–622.

Huber H, Hohn MJ, Rachel R, Fuchs T, Wimmer VC. 2002. A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont. Nature.

Huet J, Schnabel R, Sentenac A, Zillig W. 1983. Archaebacteria and eukaryotes possess DNA-dependent RNA polymerases of a common type. EMBO J. 2:1291–1294.

Iino T, Tamaki H, Tamazawa S, Ueno Y, Ohkuma M, Suzuki K-I, Igarashi Y, Haruta S. 2013. Candidatus Methanogranum caenicola: a novel methanogen from the anaerobic digested sludge, and proposal of Methanomassiliicoccaceae fam. nov. and Methanomassiliicoccales ord. nov., for a methanogenic lineage of the class Thermoplasmata. Microbes and environments / JSME 28:244–250.

Iwabe NN, Kuma KK, Hasegawa MM, Osawa SS, Miyata TT. 1989. Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. Proc. Natl. Acad. Sci. U.S.A. 86:9355–9359.

Iyer LM, Anantharaman V, Wolf MY, Aravind L. 2007. Comparative genomics of transcription

factors and chromatin proteins in parasitic protists and other eukaryotes. International Journal for Parasitology 38:1–31.

Jan Sapp Department of Science Studies York University. 1994. Evolution by Association : A History of Symbiosis. Oxford University Press

Janssen PH, Kirs M. 2008. Structure of the archaeal community of the rumen. Applied and Environmental Microbiology 74:3619–3625.

Johnson LS, Eddy SR, Portugaly E. 2010. Hidden Markov model speed heuristic and iterative HMM search procedure. BMC Bioinformatics 11:431.

Katz LA. 2012. Origin and diversification of eukaryotes. Annu. Rev. Microbiol. 66:411–427.

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. … of the Royal ….

Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. 16:111–120.

Koonin EV, Makarova KS, Elkins JG. 2007. Orthologs of the small RPB8 subunit of the eukaryotic RNA polymerases are conserved in hyperthermophilic Crenarchaeota and. Biol. Direct.

Koonin EV, Wolf YI. 2008. Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. Nucleic Acids Res. 36:6688–6719.

Korbel JO, Snel B, Huynen MA, Bork P. 2002. SHOT: a web server for the construction of genome phylogenies. Trends Genet. 18:158–162.

Kostka M, Uzlikova M, Cepicka I, Flegr J. 2007. SlowFaster, a user-friendly program for slow-fast analysis and its application on phylogeny of Blastocystis. BMC Bioinformatics 9:341–341.

Könneke M, Bernhard AE, la Torre de JR, Walker CB, Waterbury JB, Stahl DA. 2005. Isolation of an autotrophic ammonia-oxidizing marine archaeon. Nature 437:543–546.

la Torre de JR, Walker CB, Ingalls AE, Könneke M, Stahl DA. 2008. Cultivation of a thermophilic ammonia oxidizing archaeon synthesizing crenarchaeol. Environ. Microbiol. 10:810–818.

Lake JA, Rivera MC. 1994. Was the nucleus the first endosymbiont? Proc. Natl. Acad. Sci. U.S.A. 91:2880–2881.

Lake JA. 1987. A rate-independent technique for analysis of nucleic acid sequences: evolutionary parsimony. Mol. Biol. Evol. 4:167–191.

Lake JA. 1988. Origin of the eukaryotic nucleus determined by rate-invariant analysis of rRNA sequences. CORD Conference Proceedings 331:184–186.

Lake JAJ, Henderson EE, Oakes MM, Clark MWM. 1984. Eocytes: a new ribosome structure indicates a kingdom with a close relationship to eukaryotes. Proc. Natl. Acad. Sci. U.S.A. 81:3786–3790.

Lane N, Martin W. 2010. The energetics of genome complexity. Nature 467:929–934.

Lartillot N, Lepage T, Blanquart S. 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics 25:2286–2288.

Lartillot N, Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol. Biol. Evol. 21:1095–1109.

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.

Lasek-Nesselquist E, Pisani DD, Gogarten JP, Cotton JAJ, McInerney JOJ. 2013. The effects of model choice and mitigating bias on the ribosomal tree of life. Mol Phylogenet Evol 69:17–38.

Lawson FS, Charlebois RL, Dillon JA. 1996. Phylogenetic analysis of carbamoylphosphate synthetase genes: complex evolutionary history includes an internal duplication within a gene which can root the tree of life. Mol. Biol. Evol. 13:970–977.

Le SQ, Dang CC, Gascuel O. 2012. Modeling protein evolution with several amino acid replacement matrices depending on site rates. Mol. Biol. Evol. 29:2921–2936.

Le SQ, Gascuel O. 2008. An improved general amino acid replacement matrix. Mol. Biol. Evol. 25:1307–1320.

Le SQ, Lartillot N, Gascuel O. 2008. Phylogenetic mixture models for proteins. Philos Trans R Soc Lond B Biol Sci 363:3965–3976.

Leininger S, Urich T, Schloter M, Schwark L, Qi J, Nicol GW, Prosser JI, Schuster SC, Schleper C. 2006. Archaea predominate among ammonia-oxidizing prokaryotes in soils. Nature 442:806–809.

Linné von C. 1735. Carolus Linnaeus Systema Naturae, 1735.

Lombard J, López-García P, Moreira D. 2012. The early evolution of lipid membranes and the three domains of life. Nat. Rev. Microbiol. 10:507–515.

López-García P, Moreira D. 1999. Metabolic symbiosis at the origin of eukaryotes. Trends Biochem. Sci.

LWOFF A. 1957. The concept of virus. Journal of general microbiology 17:239–253.

Magrum LJ, Luehrsen KR, Woese CR. 1978. Are extreme halophiles actually ``bacteria''? J. Mol. Evol. 11:1–8.

Makarova KS, Sorokin AV, Novichkov PS, Wolf YI, Koonin EV. 2007. Clusters of orthologous

genes for 41 archaeal genomes and implications for evolutionary genomics of archaea. Biol. Direct 2:33.

Makarova KSK, Yutin NN, Bell SDS, Koonin EVE. 2010. Evolution of diverse cell division and vesicle formation systems in Archaea. Nat. Rev. Microbiol. 8:731–741.

Margulis L, Chapman M, Guerrero R, Hall J. 2006. The last eukaryotic common ancestor (LECA): acquisition of cytoskeletal motility from aerotolerant spirochetes in the Proterozoic Eon. Proc. Natl. Acad. Sci. U.S.A. 103:13080–13085.

Margulis L. 1970. Origin of Eukaryotic Cells.

Marteinsson VT, Hauksdóttir S, Hobel CF, Kristmannsdóttir H, Hreggvidsson GO, Kristjánsson JK. 2001. Phylogenetic diversity analysis of subterranean hot springs in Iceland. Applied and Environmental Microbiology 67:4242–4248.

Martijn J, Ettema TJG. 2013. From archaeon to eukaryote: the evolutionary dark ages of the eukaryotic cell. Biochem. Soc. Trans. 41:451–457.

Martin W, Koonin EV. 2006. A positive definition of prokaryotes. Nature 442:868–868.

Martin WW, Müller MM. 1998. The hydrogen hypothesis for the first eukaryote. Nature 392:37–41.

Matte-Tailliez O, Brochier C, Forterre P, Philippe H. 2002. Archaeal phylogeny based on ribosomal proteins. Mol. Biol. Evol. 19:631–639.

Miele V, Penel S, Duret L. 2011. Ultra-fast sequence clustering from similarity networks with SiLiX. BMC Bioinformatics 12:–116.

Mihajlovski A, Alric M, Brugere J-F. 2007. A putative new order of methanogenic Archaea inhabiting the human gut, as revealed by molecular analyses of themcrAgene. Res. Microbiol. 159:516–521.

Miller TL, Wolin MJ. 1982. Enumeration of Methanobrevibacter smithii in human feces. Arch. Microbiol. 131:14–18.

Miller TL, Wolin MJ. 1985. Methanosphaera stadtmaniae gen. nov., sp. nov.: a species that forms methane by reducing methanol with hydrogen. Arch. Microbiol. 141:116–122.

Moissl C, Rachel R, Briegel A, Engelhardt H, Huber R. 2005. The unique structure of archaeal "hami," highly complex cell appendages with nano-grappling hooks. Mol. Microbiol. 56:361–370.

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

Nei M. 1987. Molecular Evolutionary Genetics. Columbia University Press

Neumann N, Lundin D, Poole AM. 2010. Comparative genomic evidence for a complete nuclear pore complex in the last eukaryotic common ancestor. PLoS ONE 5:e13241.

Nunoura T, Hirayama H, Takami H, et al. 2005. Genetic and functional properties of uncultivated thermophilic crenarchaeotes from a subsurface gold mine as revealed by analysis of genome fragments. Environ. Microbiol. 7:1967–1984.

Nunoura T, Takaki Y, Kakuta J, et al. 2011. Insights into the evolution of Archaea and eukaryotic protein modifier systems revealed by the genome of a novel archaeal group. Nucleic Acids Res. 39:3204–3223.

Ochman H, Lawrence JG, Groisman EA. 2000. Lateral gene transfer and the nature of bacterial innovation. Nature 405:299–304.

Ochsenreiter T, Selezi D, Quaiser A, Bonch-Osmolovskaya L, Schleper C. 2003. Diversity and abundance of Crenarchaeota in terrestrial habitats studied by 16S RNA surveys and real time PCR. Environ. Microbiol. 5:787–797.

Oren A. 2007. Microbial life at high salt concentrations: phylogenetic and metabolic diversity. Saline Systems 4:2–2.

Oxley APA, Lanfranconi MP, Würdemann D, Ott S, Schreiber S, McGenity TJ, Timmis KN, Nogales B. 2010. Halophilic archaea in the human intestinal mucosa. Environ. Microbiol. 12:2398–2410.

Pace NR. 1997. A molecular view of microbial diversity and the biosphere. Science 276:734–740.

Papke RT. 2008. A critique of prokaryotic species concepts. Methods Mol. Biol. 532:379–395.

Paul K, Nonoh JO, Mikulski L, Brune A. 2012. “Methanoplasmatales,” Thermoplasmatales-related archaea in termite guts and other environments, are the seventh order of methanogens. Applied and Environmental Microbiology 78:8245–8253.

Penel S, Arigon A-M, Dufayard J-F, Sertier A-S, Daubin V, Duret L, Gouy M, Perrière G. 2008. Databases of homologous gene families for comparative genomics. BMC Bioinformatics 10 Suppl 6:S3–S3.

Philippe H, Forterre P. 1999. The rooting of the universal tree of life is not reliable. J. Mol. Evol. 49:509–523.

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.

Philippe H, Laurent J. 1997. How good are deep phylogenetic trees? Curr. Opin. Genet. Dev. 8:616–623.

Philippe H, Sörhannus U, Baroin A. 1994. Comparison of molecular and paleontological data in diatoms suggests a major gap in the fossil record. Journal of ….

Prangishvilli D, Zillig W, Gierl A, Biesert L, Holz I. 1982. DNA-dependent RNA polymerase of thermoacidophilic archaebacteria. European Journal of Biochemistry 122:471–477.

Preston CM, Wu KY, Molinski TF, DeLong EF. 1996. A psychrophilic crenarchaeon inhabits a marine sponge: Cenarchaeum symbiosum gen. nov., sp. nov. Proc. Natl. Acad. Sci. U.S.A. 93:6241–6246.

Probst AJ, Auerbach AK, Moissl-Eichinger C. 2012. Archaea on human skin. PLoS ONE 8:e65388–e65388.

Probst AJ, Holman H-YN, DeSantis TZ, et al. 2013. Tackling the minority: sulfate-reducing bacteria in an archaea-dominated subsurface biofilm. ISME J 7:635–651.

Ramulu HG, Groussin M, Talla E, Planel R, Daubin V, Brochier-Armanet C. 2014. Ribosomal proteins: Toward a next generation standard for prokaryotic systematics? Mol Phylogenet Evol 75:103–117.

Raymann K, Forterre P, Brochier-Armanet C, Gribaldo S. 2014. Global phylogenomic analysis disentangles the complex evolutionary history of DNA replication in archaea. Genome biology and evolution 6:192–212.

Rieu-Lesme F, Delbès C, Sollelis L. 2005. Recovery of partial 16S rDNA sequences suggests the presence of Crenarchaeota in the human digestive ecosystem. Current Microbiology 51:317–321.

Rinke C, Schwientek P, Sczyrba A, et al. 2013. Insights into the phylogeny and coding potential of microbial dark matter. Nature 499:431–437.

Rivera MCM, Lake JAJ. 1992. Evidence that eukaryotes and eocyte prokaryotes are immediate relatives. Science 257:74–76.

Robertson CE, Harris JK, Spear JR, Pace NR. 2005. Phylogenetic diversity and ecology of environmental Archaea. Curr. Opin. Microbiol. 8:638–642.

Rochette NC, Brochier-Armanet C, Gouy M. 2014. Phylogenomic Test of the Hypotheses for the Evolutionary Origin of Eukaryotes. Mol. Biol. Evol. 31:832–845.

Roger AJ, Brown JR. 1995. A chimeric origin for eukaryotes re-examined. Trends Biochem. Sci. 21:370–371.

Rudolph C, Moissl C, Henneberger R, Huber R. 2003. Ecology and microbial structures of archaeal/bacterial strings-of-pearls communities and archaeal relatives thriving in cold sulfidic springs. FEMS Microbiology Ecology 50:1–11.

Rudolph C, Wanner G, Huber R. 2001. Natural communities of novel archaea and bacteria growing in cold sulfurous springs with a string-of-pearls-like morphology. Applied and Environmental Microbiology 67:2336–2344.

Sagan L. 1967. On the origin of mitosing cells. Journal of Theoretical Biology 14:255–274.

SANGER F. 1959. Chemistry of insulin; determination of the structure of insulin opens the way to greater understanding of life processes. Science 129:1340–1344.

Sapp J. 2005. The prokaryote-eukaryote dichotomy: meanings and mythology. Microbiol. Mol. Biol. Rev. 69:292–305.

Sapp J. 2009. The New Foundations of Evolution: On the Tree of Life. Oxford University Press

Sarmiento F, Long F, Cann I, Whitman WB. 2014. Diversity of the DNA Replication System in the Archaea Domain. Archaea 2014:–675946.

Sauder LA, Engel K, Stearns JC, Masella AP. 2011. Aquarium nitrification revisited: Thaumarchaeota are the dominant ammonia oxidizers in freshwater aquarium biofilters. PLoS ONE.

Scanlan PD, Marchesi JR. 2008. Micro-eukaryotic diversity of the human distal gut microbiota: qualitative assessment using culture-dependent and -independent analysis of faeces. ISME J 2:1183–1193.

Schleper C, Holben W, Klenk HP. 1997. Recovery of crenarchaeotal ribosomal DNA sequences from freshwater-lake sediments. Applied and Environmental Microbiology 63:321–323.

Schleper C, Jurgens G, Jonuscheit M. 2005. Genomic studies of uncultivated archaea. Nat. Rev. Microbiol. 3:479–488.

Searcy DG, Hixon WG. 1991. Cytoskeletal origins in sulfur-metabolizing archaebacteria. Biosystems 25:1–11.

Shimodaira H, Hasegawa M. 2001. CONSEL: for assessing the confidence of phylogenetic tree selection. Bioinformatics 17:1246–1247.

Simpson A, Roger AJ. 2004. The real “kingdoms” of eukaryotes. Curr Biol.

Snel B, Bork P, Huynen MA. 1999. Genome phylogeny based on gene content. Nature Genetics 21:108–110.

Spang A, Hatzenpichler R, Brochier-Armanet C, et al. 2010. Distinct gene set in two different lineages of ammonia-oxidizing archaea supports the phylum Thaumarchaeota. Trends Microbiol. 18:331–340.

Spang A, Martijn J, Saw JH, Lind AE, Guy L, Ettema TJG. 2013. Close encounters of the third domain: the emerging genomic view of archaeal diversity and evolution. Archaea 2013:202358–202358.

Stahl DA, la Torre de JR. 2012. Physiology and Diversity of Ammonia-Oxidizing Archaea. Annu. Rev. Microbiol. 66:83–101.

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

Stanier RY, Niel CB. 1962. The concept of a bacterium. Arch. Microbiol.

Stanier RY. 1963. The microbial world.

Staub E, Fiziev P, Rosenthal A, Hinzmann B. 2004. Insights into the evolution of the nucleolus by an analysis of its protein domain repertoire. Bioessays 26:567–581.

Stoever BC, Mueller KF. 2010. TreeGraph 2: Combining and visualizing evidence from different phylogenetic analyses. BMC Bioinformatics 11:–7.

Stöffler G, Stöffler-Meilicke M. 1985. Electron microscopy of archaebacterial ribosomes. Syst Appl Microbiol 7:123–130.

Stöffler-Meilicke M, Böhme C, Strobel O, Böck A, Stöffler G. 1986. Structure of Ribosomal Subunits of M. vannielii: Ribosomal Morphology as a Phylogenetic Marker. Science 231:1306–1308.

Szöllosi, Gergely J, and Vincent Daubin. 2012. “Modeling Gene Family Evolution and Reconciling Phylogenetic Discord..” *Methods in Molecular Biology (Clifton, N.J.)* 856: 29–51. doi:10.1007/978-1-61779-585-5_2.

Szöllosi, Gergely J, Bastien Boussau, Sophie S Abby, Eric Tannier, and Vincent Daubin. 2012. “Phylogenetic Modeling of Lateral Gene Transfer Reconstructs the Pattern and Relative Timing of Speciations..” *Proceedings of the National Academy of Sciences of the United States of America* 109 (43): 17513–18. doi:10.1073/pnas.1202997109.

Szöllosi, Gergely J, Wojciech Rosikiewicz, Bastien Boussau, Eric Tannier, and Vincent Daubin. 2013. “Efficient Exploration of the Space of Reconciled Gene Trees..” *Systematic Biology* 62 (6): 901–12. doi:10.1093/sysbio/syt054.

Taylor CD, McBride BC, Wolfe RS, Bryant MP. 1974. Coenzyme M, essential for growth of a rumen strain of Methanobacterium ruminantium. J. Bacteriol. 120:974–975.

Theocharidis A, van Dongen S, Enright AJ, Freeman TC. 2009. Network visualization and analysis of gene expression data using BioLayout Express(3D). Nat Protoc 4:1535–1550.

Tourasse NJ, Gouy M. 1997. Evolutionary distances between nucleotide sequences based on the distribution of substitution rates among sites as estimated by parsimony. Mol. Biol. Evol.

14:287–298.

Tourasse NJN, Gouy MM. 1999. Accounting for Evolutionary Rate Variation among Sequence Sites Consistently Changes Universal Phylogenies Deduced from rRNA and Protein-Coding Genes. Mol Phylogenet Evol 13:10–10.

van der Giezen M, Slotboom DJ, Horner DS, Dyal PL, Harding M, Xue G-P, Embley TM, Kunji ERS. 2002. Conserved properties of hydrogenosomal and mitochondrial ADP/ATP carriers: a common origin for both organelles. EMBO J. 21:572–579.

Wagner M, Horn M. 2005. ThePlanctomycetes,Verrucomicrobia,Chlamydiaeand sister phyla comprise a superphylum with biotechnological and medical relevance. Curr Opin Biotechnol 17:241–249.

Wallin IE. 1923. Symbionticism and prototaxis, two fundamental biological principles. The Anatomical Record.

Wallin IE. 1927. Symbionticism and the origin of species.

Waters E, Hohn MJ, Ahel I. 2003. The genome of Nanoarchaeum equitans: insights into early archaeal evolution and derived parasitism.

Watson JD, Crick F. 1953. A Structure for Deoxyribose Nucleic Acid. Nature:737–738.

Whittaker RH. 1969. New concepts of kingdoms or organisms. Evolutionary relations are better represented by new classifications than by the traditional two kingdoms. Science 163:150–160.

Williams TA, Embley TM. 2014. Archaeal “Dark Matter” and the Origin of Eukaryotes. Genome biology and evolution 6:474–481.

Williams TA, Foster PG, Cox CJ, Embley TM. 2013. An archaeal origin of eukaryotes supports only two primary domains of life. Nature 504:231–236.

Williams TAT, Foster PGP, Nye TMWT, Cox CJC, Embley TMT. 2012. A congruent phylogenomic signal places eukaryotes within the Archaea. Proc Biol Sci 279:4870–4879.

Woese CR, Fox GE. 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc. Natl. Acad. Sci. U.S.A. 74:5088–5090.

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. U.S.A. 87:4576–4579.

Woese CR, Magrum LJ, Fox GE. 1978. Archaebacteria. J. Mol. Evol. 11:245–252.

Woese CR, Olsen GJ. 1985. Archaebacterial phylogeny: Perspectives on the Urkingdoms. Syst Appl Microbiol 7:161–177.

Woese CR. 1964. UNIVERSALITY IN THE GENETIC CODE. Science 144:1030–1031.

Woese CR. 1987. Bacterial evolution. Microbiological reviews. Baltimore 51:221–271.

Wolf YI, Rogozin IB, Koonin EV. 2003. Coelomata and not Ecdysozoa: evidence from genome-wide phylogenetic analysis. Genome Res 14:29–36.

Wolf YIY, Makarova KSK, Yutin NN, Koonin EVE. 2012. Updated clusters of orthologous genes for Archaea: a complex ancestor of the Archaea and the byways of horizontal gene transfer. Biol. Direct 7:46–46.

Yang D, Oyaizu Y, Oyaizu H, Olsen GJ, Woese CR. 1985. Mitochondrial origins. Proc. Natl. Acad. Sci. U.S.A. 82:4443–4447.

Yang Z. 1995. A space-time process model for the evolution of DNA sequences. Genetics 139:993–1005.

Yutin N, Wolf MY, Wolf YI, Koonin EV. 2009. The origins of phagocytosis and eukaryogenesis. Biol. Direct 4:9–9.

Zillig W, Klenk HP, Palm P, Pühler G, Gropp F, Garrett RA, Leffers H. 1989. The Phylogenetic Relations of Dna-Dependent Rna-Polymerases of Archaebacteria, Eukaryotes, and Eubacteria. Can. J. Microbiol. 35:73–80.

Zillig W, Palm P, Klenk HP, Pühler G, Gropp F. 1991. Phylogeny of DNA-dependent RNA polymerases: testimony for the origin of eukaryotes. General and applied \ldots.

Zillig W, STETTER KO, JANEKOVIC D. 1979. Dna-Dependent Rna-Polymerase From the Archaebacterium Sulfolobus-Acidocaldarius. European Journal of Biochemistry 96:597–604.

Zillig W, STETTER KO, TOBIEN M. 1978. Dna-Dependent Rna-Polymerase From Halobacterium-Halobium. European Journal of Biochemistry 91:193–199.

Zillig W. 1991. Comparative biochemistry ofArchaeaandBacteria. Curr. Opin. Genet. Dev. 1:544–551.

Zuckerkandl E, Pauling L. 1964. Molecules as documents of evolutionary history. Journal of Theoretical Biology 8:357–366.

Appendix 1. Article 3

Genomic survey of *reb* genes

My undergraduate research project was on the isolation, characterization and sequencing of killer bacterial endosymbionts of various Paramecium strains. These bacteria produce characteristic intracellular structures called R-bodies, which confer them with a killing phenotype towards sensitive strains, and whose role and mechanism of action is still unclear. A few months before starting my Ph.D with Simonetta Gribaldo, while I was still in the US, we started collaboration to perform a phylogenomic analysis of the genes responsible for the formation of R-bodies, the *reb* genes. This project allowed me to get a first introduction to bioinformatics and phylogenomics and was finalized during the first months of my thesis in Paris. By performing an exhaustive genomic survey I identified highly conserved Reb homologs in the genomes of a very large number of proteobacterial taxa displaying a wide variety of lifestyles, from free-living to mutualistic or pathogenic association eukaryotes. My phylogenomic analysis indicates a complex evolutionary history involving spread of entire reb loci across distantly related proteobacterial families along with species-specific reb gene duplications, and allowed us to propose a number of potential candidates for additional components of the R-body system.

This work was published in G3 (Genes, Genomes, Genetics), a new journal of the Genetics Society of America, in March 2013.

Article 3

A Genomic Survey of Reb Homologs Suggests Widespread Occurrence of R-Bodies in Proteobacteria

Kasie Raymann,*,† Louis-Marie Bobay,†,‡,§ Thomas G. Doak, Michael Lynch,** and Simonetta Gribaldo*,1**

*Institut Pasteur, Unité Biologie Moleculare du Gene chez les Extremophiles, Departement de Microbiologie, Paris, 75724 Cedex 15, France, †Université Pierre et Marie Curie, Cellule Pasteur UPMC, Paris, 75724 Cedex 15, France, ‡Institut Pasteur, Microbial Evolutionary Genomics, Departement Genomes et Genetique, Paris, 75724 Cedex 15, France, §Centre National de la Recherche Scientifique, Unité Mixte de Recherche 3525, Paris, F-75015 France, and **Indiana University, Department of Biology, Bloomington, Indiana 47405

ABSTRACT Bacteria and eukaryotes are involved in many types of interaction in nature, with important ecological consequences. However, the diversity, occurrence, and mechanisms of these interactions often are not fully known. The obligate bacterial endosymbionts of Paramecium provide their hosts with the ability to kill sensitive Paramecium strains through the production of R-bodies, highly insoluble coiled protein ribbons. R-bodies have been observed in a number of free-living bacteria, where their function is unknown. We have performed an exhaustive survey of genes coding for homologs of Reb proteins (R-body components) in complete bacterial genomes. We found that *reb* genes are much more widespread than previously thought, being present in representatives of major Proteobacterial subdivisions, including many free-living taxa, as well as taxa known to be involved in various kinds of interactions with eukaryotes, from mutualistic associations to pathogenicity. Reb proteins display very good conservation at the sequence level, suggesting that they may produce functional R-bodies. Phylogenomic analysis indicates that *reb* genes underwent a complex evolutionary history and allowed the identification of candidates potentially involved in R-body assembly, functioning, regulation, or toxicity. Our results strongly suggest that the ability to produce R-bodies is likely widespread in Proteobacteria. The potential involvement of R-bodies in as yet unexplored interactions with eukaryotes and the consequent ecological implications are discussed.

KEYWORDS
kappa particles
Caedibacter
phylogenomics

During more than two billion years of coexistence, prokaryotes have established various forms of interaction with eukaryotes. Examples include the mutualistic symbioses that benefit eukaryotic host by providing nutrients, defense, competition, and adaptation to new environments (Gast *et al.* 2009). At the same time, bacteria have developed various ways to defend themselves against grazing by eukaryotes (Matz and Kjelleberg 2005), with potential implications for the emergence of pathogens (Brown and Barker 1999). However, the extent and diversity of bacterial/eukaryotic interactions in nature remains largely underexplored. As a growing amount of genomic data covering a large fraction of bacterial diversity becomes available, hints of such relationships may be gathered from *in silico* analyses. These can be linked to experimental observations, providing useful directions for further work.

A fascinating example of a bacterial/eukaryote relationship is provided by the killer endosymbionts of the ciliate Paramecium. In the 1930s Tracey Sonneborn discovered that some strains of the *Paramecium aurelia* complex have a killer phenotype toward sensitive strains (Beale and Preer 2004; Preer 2006; Sonneborn 1938). Sonneborn could show that this phenomenon is not controlled by nuclear genes, providing one of the first examples of cytoplasmic inheritance (Sonneborn 1943). It was later discovered that the killer phenotype is conveyed by an obligate endosymbiotic bacterium (also referred to as

doi: 10.1534/g3.112.005231
Manuscript received December 3, 2012; accepted for publication January 9, 2013

Supporting information is available online at http://www.g3journal.org/lookup/suppl/doi:10.1534/g3.112.005231/-/DC1.
[1]Corresponding author: Unité Biologie Moleculare du Gene chez les Extremophiles, Departement de Microbiologie, Institut Pasteur, 25-28 rue du Dr Roux 75724, Paris Cedex 15, France. E-mail: simonetta.gribaldo@pasteur.fr

a kappa particle), and each killer paramecium strain harbors its own specific endosymbiont that usually resides in the cytoplasm but can also be in the nucleus (Preer 1975). Whereas all these obligate endosymbionts initially were placed in the genus Caedibacter, they were later shown to belong to different proteobacterial lineages (Beier *et al.* 2002).

In Caedibacter, the killer trait is directly linked to the production of R-bodies, unusual cytoplasmic refractile inclusion bodies (Dilts and Quackenbush 1986) [for review, see (Pond *et al.* 1989)]. What is known about R-bodies comes primarily from *Caedibacter taeniospiralis*, which belongs to the gammaproteobacteria family of Thiotrichales (Figure 1) (Beier *et al.* 2002). R-bodies are highly insoluble protein ribbons that are typically coiled into cylindrical structures. They are produced by only a fraction of the endosymbiont population, which then stop dividing. When R-body−containing bacteria are released into the environment and captured by sensitive strains, killing occurs very rapidly. The exact mechanism for killing is not known, but it is thought that internalization of the R-body-containing bacteria into the food vacuole triggers unrolling of the R-body, which penetrates the phagosomal membrane and delivers a killer toxin to the cytoplasm (Figure 1) (Jurand *et al.* 1971; Preer *et al.* 1974). Isolated R-body−containing bacteria are capable of killing sensitive Paramecium strains in various ways, with prelethal symptoms such as paralysis, vacuolization, and opposite swimming rotation. On the contrary, exposure to non-R-body−containing Caedibacter is not lethal [for review, see (Pond *et al.* 1989)], and a mutant Caedibacter strain unable to make R-bodies loses its killing ability (Dilts and Quackenbush 1986). Different strains of Caedibacter produce different types of R-bodies, which vary in diameter (0.25−0.8 mm), length (<10−30 mm), ribbon morphology (tapered or blunt termini), mode of unrolling (from the outside or from the inside in a telescopic fashion), and the nature of the stimulus for unrolling (changes in pH, temperature, ionic strength) [for review, see (Pond *et al.* 1989; Sanchez-Amat 2006)]. Interestingly, the unrolling of type 51 R-bodies has been shown to be reversible (they unroll when the pH is dropped <6.5 and reroll when the pH is again raised >7.0) (Pond *et al.* 1989).

Studies of the genetic determinants of R-bodies began in the early 1980s, and in *C. taeniospiralis* the R body−coding region was found to lie on a plasmid (Quackenbush and Burbach 1983). When a region from the pKAP47 plasmid (from *Paramecium teraurelia* California strain 47) was cloned into *Escherichia coli*, R-bodies were produced, but the clones did not exhibit toxicity toward sensitive Paramecium strains (Quackenbush and Burbach 1983). Therefore, whereas production of R-bodies is necessary for killing by *C. taeniospiralis*, as described previously, it is not sufficient for killing by recombinant *Escherichia coli*. This excludes a direct cytotoxic effect of R-bodies and indicates a requirement for an essential unknown toxin encoded either by the plasmid or the *C. taeniospiralis* genome (Preer and Stark 1953; Quackenbush and Burbach 1983). These data have been recently confirmed, and it has been shown that recombinant *E. coli* expressing the four *reb* genes of the *C. taeniospiralis* pKAP298 plasmid (from *Paramecium teraurelia* Panama strain 298) were capable of producing R-bodies but were not toxic toward sensitive Paramecium strains (Schrallhammer *et al.* 2012).

Very little information is available on the assembly process of R-bodies. At least three polypeptides of 10, 13, and 18 kDa were found to be involved in the structure and assembly of type 51 R-bodies (Kanabrocki *et al.* 1986) and were later shown to be encoded by three genes: *reb*C, *reb*B, and *reb*A, respectively, the last two being homologous (Heruth *et al.* 1994). These early data proposed that the major structural protein is RebB and that RebA may act as a scaffold to facilitate the polymerization process whereas RebC may act as a transcriptional regulator (Heruth *et al.* 1994). Finally, it was suggested that RebB might be modified posttranslationally, with the possible involvement of RebC (Heruth *et al.* 1994). The role of a fourth gene in the *reb* locus, *reb*D, coding for a homolog of RebA and RebB, is unclear but it was shown not to be necessary for R-body production in *E. coli* (Heruth *et al.* 1994).

The complete sequence of the Reb-harboring pKAP298 plasmid of *C. taeniospiralis* strain 298 was obtained in 2005 (Jeblick and Kusch 2005). It was found that this plasmid contains 63 open reading frames, 23 only having similarity with proteins with known function, and

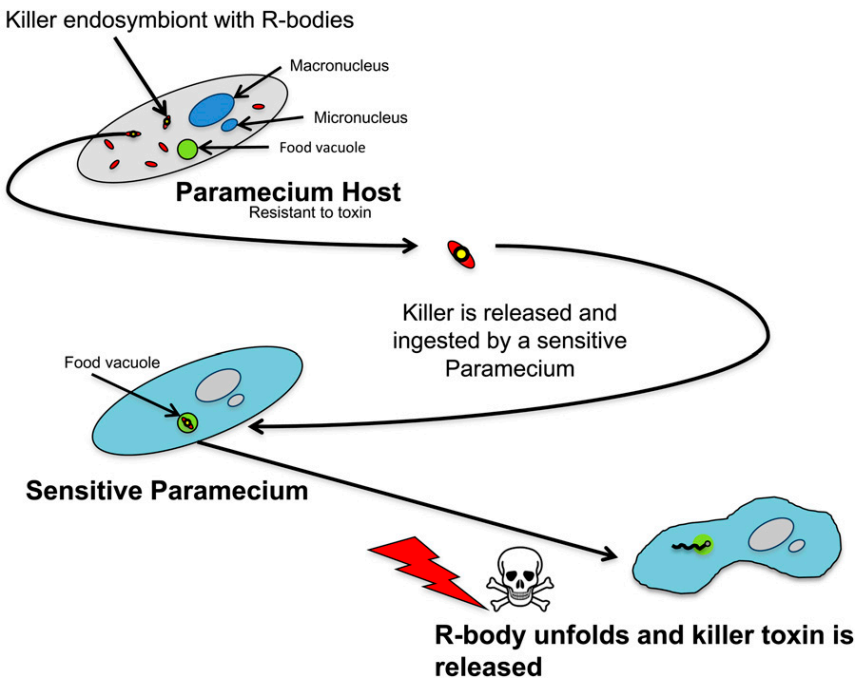


Figure 1 Illustration of the *C. taeniospiralis* R-body toxin delivery system (see main text for details and references).

a few being similar to proteins encoded by phages or prophages, which led to the suggestion that the plasmid originated from a bacteriophage (Jeblick and Kusch 2005), which is consistent with early observations of the association of phage-like particles with R-bodies (Preer *et al.* 1974). A protein with homology to the Soj-ParA family of membrane-associated ATPases was suggested as a possible candidate for the toxin, which would kill the host by somehow affecting its membrane, although a precise mechanism was not proposed (Jeblick and Kusch 2005).

The harboring of an endosymbiont that produces R-bodies gives a competitive advantage to its killer Paramecium host with respect to sensitive strains (Kusch *et al.* 2002). In turn, R-body production seems to play a role in defense against predation and creates a benefit for the Caedibacter strains at the population level (Sanchez-Amat 2006). However, many important questions remain to be clarified. For example, it is not known how obligate symbiosis is established in the first place or how sensitive strains can pick up Caedibacter and become killers, nor how killer Paramecium strains are protected from their own specific R-body producing endosymbionts (Gibson 1973; Preer *et al.* 1974).

Interestingly, casual observations of coiled R-body structures of various types have been reported in several free-living bacteria: the hydrogen-oxidizing β-proteobacterium *Pseudomonas taeniospiralis* (Lalucat and Mayer 1978), now known as *Hydrogenophaga taeniospiralis*; the soil β-proteobacterium *Pseudomonas avenae* (Wells and Horne 1983), now known as *Acidovorax avenae* subsp. *avenae*; the soil strain *Pseudomonas* sp. EPS-5028 (Fusté *et al.* 1986); the anoxigenic photosynthetic N2-fixing α-proteobacterium *Rhodospirillum centenum* (Favinger *et al.* 1989); the soil strain *Pseudomonas aeruginosa* 44T1 (Espuny *et al.* 1991); and the melanin-producing marine γ-proteobacterium *Marinomonas mediterranea* (Hernandez-Romero *et al.* 2003). However, no further study on these R-body structures has been reported for any of these species [for review, see (Sanchez-Amat 2006)], nor have they been linked to the presence of Reb homologs in their genomes. Therefore, the role of these R-bodies in these diverse bacterial remains puzzling. A recent study has shown the presence of Reb homologs in the genome of the rhizobiale *Azorhizobium caulinodans*, a microsymbiont of the tropical legume *Sesbania rostrata* (Akiba *et al.* 2010). Interestingly, deletion of the putative transcription factor *praR* caused aberrant nodule formation and was linked to greater expression of the *reb* locus. On the contrary, a double *reb* and *praR* mutant had a restored wild-type nodule formation. The authors hypothesized that *praR* is essential to suppress the killer trait conferred by the *reb* locus and establish symbiosis between *A. caulinodans* and *S. rostrata* (Akiba *et al.* 2010). However, it is not known whether *A. caulinodans* is able to make R-bodies. The authors also reported the presence of Reb homologs in a number of Proteobacteria and in the Bacteroidetes member *Kordia algicida* OT-1 (Akiba *et al.* 2010).

Here, we have performed an exhaustive phylogenomic analysis of Reb homologs in currently available bacterial genomes. Reb homologs are widely distributed in members of *Proteobacteria*, comprising many free-living taxa as well as symbionts or pathogens of various eukaryotes, including humans. The evolutionary history of *reb* genes appears very dynamic, involving vertical inheritance, horizontal gene transfers, and gene duplications. By combining phylogenetic, genome synteny, and genomic content analyses, we highlight a few potential candidate partners of Reb proteins. Finally, we found no clear signs of *reb* loci originating from defective prophages, or from recent transfer via mobile elements. Ecological implications are discussed.

MATERIALS AND METHODS

Homology searches

Reb proteins (A-D) encoded in the plasmid pKAP298 from *C. taeniospiralis* (AAR87077.1, AAR87076.1, AAR87131.1, AAR87075.1) were used as seeds to search for Reb homologs in the nonredundant protein database at the National Center for Biotechnology Information (NCBI). Homology searches were performed by BlastP (Altschul *et al.* 1997) and all hits within an *e*-value cutoff of 1×10^{-3} were retained. PSI-BLAST and tBLASTn programs (Altschul *et al.* 1997) also were used to search for highly divergent or misannotated Reb homologs. Searches were reiterated by using a number of seeds from various taxa. We also performed targeted searches against the metagenome and the viral sequence databases at the NCBI, and against all eukaryotic genome sequence available at the Joint Genome Institute (http://genome.jgi.gov). Sequences were aligned using Muscle 3.8.31 (Edgar 2004). Poorly aligned or divergent sequences were manually removed. Finally, HMM searches were performed with HMMER 3.0 (hmmer.org) against a local databank of 841 complete bacterial genomes (only one representative per species), including 435 from Proteobacteria downloaded from the NCBI ftp Genomes server using a model built on the multiple alignment of all previously recovered Reb proteins, but no additional homologs were found.

Sequence analysis

Sequence secondary structures were predicted using PSIPRED (http://bioinf.cs.ucl.ac.uk/psipred/) (Buchan *et al.* 2010). Alpha helical wheel diagrams were created using the tool created by Don Armstrong and Raphael Zidovetzki (http://rzlab.ucr.edu/scripts/wheel/wheel.cgi). PredictProtein was used to search for additional structural features (http://www.predictprotein.org/). We searched for Reb homologs with available 3D structures using sequence based-PSI-BLAST (Altschul *et al.* 1997) searches at the PFAM and Uniprot databases by using HHpred (http://hhpred.tuebingen.mpg.de/hhpred) and FFAS03 (http://ffas.ljcrf.edu/ffas-cgi/cgi/ffas.pl). WebLogo analysis was performed at http://weblogo.berkeley.edu/.

Phylogenetic analysis

The final dataset of 203 Reb homologs was aligned using Muscle 3.8.31 (Edgar 2004) and trimmed using BMGE (Criscuolo and Gribaldo 2010) with the less-stringent parameters (Blosum30), giving a dataset of 73 unambiguously aligned amino acid positions for phylogenetic analysis. A Bayesian tree was obtained using Phylobayes 3.3 (Lartillot *et al.* 2009). (See figure legends for details on analyses.)

16s rRNA sequences from Proteobacterial taxa representative of the diversity of this phylum were downloaded from the NCBI, as well as from the specialized Silva (http://www.arb-silva.de/) and the Ribosomal Database Project (http://rdp.cme.msu.edu/) databases. Sequences were aligned using Muscle 3.8.31 (Edgar 2004) and manually trimmed using the ED program of the MUST suite (Philippe 1993). Maximum likelihood trees were obtained using Treefinder (Jobb *et al.* 2004). (See figure legends for details on analyses.)

Genome synteny and genome content analysis

For genome synteny analysis, we retrieved the five open reading frames upstream and downstream of the *reb* loci (or extracted from contigs when the whole genome sequence was not available). Protein sequences were defined as homologous when sharing at least 40% similarity and less than 20% difference in length. Pairs of homologous proteins were then expanded to homologous protein families by including all proteins homologous to at least one member of the family.

Taxon Name	Class	Accession number of Reb homologues
Azorhizobium caulinodans ORS 571	Alphaproteobacteria	YP_001526697 YP_001526698 YP_001526699 YP_001526702
Azospirillum brasilense Sp245 (megaplasmid) ★	Alphaproteobacteria	CCD03670.1 CCD03671.1 CCD03672.1 CCD03673.1
Azospirillum sp. B510	Alphaproteobacteria	
Brevundimonas sp. BAL3	Alphaproteobacteria	ZP_05031814 ZP_05032498 ZP_05033395 ZP_05033502 ZP_05033545
Brevundimonas subvibrioides ATCC 15264	Alphaproteobacteria	
Labrenzia alexandrii DFL-11	Alphaproteobacteria	ZP_05112322 ZP_05116504
Maricaulis maris MCS10	Alphaproteobacteria	YP_756025
Oceanicaulis alexandrii HTCC2633	Alphaproteobacteria	ZP_00953290 ZP_00953295
Polymorphum gilvum SL003B-26A1	Alphaproteobacteria	ADZ69851 ADZ69856 ADZ69857
Pseudovibrio sp. FO-BEG1	Alphaproteobacteria	
Pseudovibrio sp. JE062	Alphaproteobacteria	ZP_05083395 ZP_05083514 ZP_05083638 ZP_05083799 ZP_05083855
Rhodobacteraceae bacterium HTCC2150	Alphaproteobacteria	ZP_01741096
Rhodospirillum centenum SW	Alphaproteobacteria	YP_002298200 YP_002298201 YP_002298202 YP_002298203 YP_002298204 YP_002298205
Roseibium sp. TrichSKD4	Alphaproteobacteria	ZP_07661110 ZP_07661112 ZP_07661117
Ruegeria pomeroyi DSS-3 (megaplasmid) ★	Alphaproteobacteria	YP_165220 YP_165225 YP_165226
Ruegeria sp. TM1040	Alphaproteobacteria	
Acidovorax avenae subsp. avenae ATCC 19860	Betaproteobacteria	YP_004236845 YP_004236846 YP_004236847 YP_004236848
Burkholderia ambifaria AMMD	Betaproteobacteria	YP_773103 YP_773104 YP_773105 YP_773108
Burkholderia ambifaria IOP40-10	Betaproteobacteria	ZP_02889068 ZP_02889069 ZP_02889070 ZP_02889073
Burkholderia ambifaria MC40-6	Betaproteobacteria	YP_001807945 YP_001807946 YP_001807947 YP_001807950
Burkholderia ambifaria MEX-5	Betaproteobacteria	ZP_02905799 ZP_02905800 ZP_02905801 ZP_02905804
Burkholderia gladioli BSR3	Betaproteobacteria	AEA62614 AEA62615 AEA62616 AEA62617 AEA63937
Burkholderia glumae BGR1	Betaproteobacteria	YP_002909241 YP_002909242 YP_002909243 YP_002909244 YP_002909245 YP_002909246
Burkholderia mallei ATCC 23344	Betaproteobacteria	AAU50114
Burkholderia mallei NCTC 10229	Betaproteobacteria	ABN03246
Burkholderia mallei NCTC 10247	Betaproteobacteria	ABN03246
Burkholderia mallei SAVP1	Betaproteobacteria	ABM49889
Burkholderia oklahomensis C6786	Betaproteobacteria	ZP_02361292 ZP_02361293 ZP_02361294 ZP_02361295
Burkholderia oklahomensis EO147	Betaproteobacteria	ZP_02354115 ZP_02354116 ZP_02354117 ZP_02354118
Burkholderia pseudomallei 1106a	Betaproteobacteria	YP_001064557
Burkholderia pseudomallei 1710b	Betaproteobacteria	
Burkholderia pseudomallei 305	Betaproteobacteria	ZP_01770366
Burkholderia pseudomallei 576	Betaproteobacteria	ZP_03456481
Burkholderia pseudomallei 668	Betaproteobacteria	YP_001057315
Burkholderia pseudomallei K96243	Betaproteobacteria	YP_106893
Burkholderia sp. 383	Betaproteobacteria	YP_366668 YP_366669 YP_366670 YP_366673
Burkholderia sp. CCGE1001	Betaproteobacteria	
Burkholderia sp. CCGE1002	Betaproteobacteria	
Burkholderia sp. CCGE1003	Betaproteobacteria	YP_003910733 YP_003910734 YP_003910737
Burkholderia sp. YI23	Betaproteobacteria	
Burkholderia thailandensis E264	Betaproteobacteria	
Burkholderia thailandensis MSMB43	Betaproteobacteria	ZP_02461981 ZP_02461983 ZP_02461984 ZP_02461985 ZP_02461986
Burkholderia ubonensis Bu	Betaproteobacteria	ZP_02381581 ZP_02381584 ZP_02381585 ZP_02381586
Chromobacterium violaceum ATCC 12472	Betaproteobacteria	NP_900391 NP_900392 NP_900393 NP_900394 NP_900403 NP_901057
Ralstonia syzygii R24	Betaproteobacteria	CCA86915.1 CCA86916.1 CCA86917.1 CCA86918.1
Desulfatibacillum alkenivorans AK-01	Deltaproteobacteria	YP_002433824 YP_002433825 YP_002433826 YP_002433827 YP_002433828 YP_002433829 YP_002433830
Plesiocystis pacifica SIR-1	Deltaproteobacteria	ZP_01912570 ZP_01912575 ZP_01912576 ZP_01912577 ZP_01912578 ZP_01912579 ZP_01912580 ZP_01912581
Caedibacter taeniospiralis (plasmid) ★	Gammaproteobacteria	AAR87075.1 AAR87076.1 AAR87077.1
Marinomonas mediterranea MMB-1	Gammaproteobacteria	ADZ91901 ADZ91905 ADZ91906 ADZ91907 ADZ91908 ADZ91953 ADZ91954 ADZ91955 ADZ91956
Marinomonas sp. MED121	Gammaproteobacteria	ZP_01077806 ZP_01077807 ZP_01077808 ZP_01077809 ZP_01077810 ZP_01077811
Marinomonas sp. MWYL1	Gammaproteobacteria	
Pseudomonas aeruginosa LESB58	Gammaproteobacteria	YP_002439849 YP_002439850 YP_002439855
Pseudomonas aeruginosa M18	Gammaproteobacteria	
Pseudomonas aeruginosa NCGM2.S1	Gammaproteobacteria	
Pseudomonas aeruginosa PACS2	Gammaproteobacteria	ZP_01366280 ZP_01366286 ZP_01366287
Pseudomonas aeruginosa PA7	Gammaproteobacteria	
Pseudomonas aeruginosa PAO1	Gammaproteobacteria	
Pseudomonas aeruginosa UCBPP-PA14	Gammaproteobacteria	YP_790355 YP_790354
Pseudomonas aeruginosa 39016	Gammaproteobacteria	ZP_07794130
Pseudomonas fluorescens F113	Gammaproteobacteria	
Pseudomonas fluorescens Pf-5	Gammaproteobacteria	YP_257326 YP_257329 YP_257330
Pseudomonas fluorescens Pf0-1	Gammaproteobacteria	
Pseudomonas fluorescens SBW25	Gammaproteobacteria	
Saccharophagus degradans 2-40	Gammaproteobacteria	YP_528473
Shewanella denitrificans OS217	Gammaproteobacteria	YP_564147 YP_564148 YP_564149 YP_564150
Stenotrophomonas maltophilia K279a	Gammaproteobacteria	
Stenotrophomonas maltophilia R551-3	Gammaproteobacteria	YP_002028483 YP_002028484 YP_002028485 YP_002028486
Stenotrophomonas sp. SKA14	Gammaproteobacteria	ZP_05132942 ZP_05135487 ZP_05135740
Vibrio coralliilyticus ATCC BAA-450	Gammaproteobacteria	ZP_05885798 ZP_05885799 ZP_05885800 ZP_05885804
Vibrio fischeri ES114	Gammaproteobacteria	YP_207085 YP_207090 YP_207091 YP_207092 YP_207093 YP_207094 YP_207095
Vibrio fischeri MJ11	Gammaproteobacteria	
Vibrio nigripulchritudo ATCC 27043	Gammaproteobacteria	ZP_08730919 ZP_08730923 ZP_08730924 ZP_08730925
Xanthomonas axonopodis pv. citri str. 306	Gammaproteobacteria	NP_643323 NP_643324 NP_643325 NP_643326 NP_643396
Xanthomonas axonopodis pv. citrumelo F1	Gammaproteobacteria	
Xanthomonas campestris pv. campestris str. 8004	Gammaproteobacteria	YP_242290
Xanthomonas campestris pv. campestris str. ATCC 33913	Gammaproteobacteria	NP_638256
Xanthomonas campestris pv. campestris str. B100	Gammaproteobacteria	
Xanthomonas campestris pv. raphani 756C	Gammaproteobacteria	
Xanthomonas campestris pv. vesicatoria str. 85-10	Gammaproteobacteria	YP_364948
Xanthomonas campestris pv. vasculorum NCPPB702	Gammaproteobacteria	ZP_06486467
Xanthomonas fuscans subsp. aurantifolii str. ICPB 11122	Gammaproteobacteria	ZP_06705918
Xanthomonas oryzae pv. oryzae KACC 10331	Gammaproteobacteria	
Xanthomonas oryzae pv. oryzae MAFF 311018	Gammaproteobacteria	
Xanthomonas oryzae pv. oryzae PXO99A	Gammaproteobacteria	YP_001914301
Xanthomonas oryzae pv. oryzicola BLS256	Gammaproteobacteria	AEQ97370
Xanthomonas perforans 91-118	Gammaproteobacteria	ZP_08187121
Kordia algicida OT-1	Bacteroidetes	ZP_02161959 ZP_02161960 ZP_02161961 ZP_02161966

For genome content analysis, families of homologous proteins were built from 861 fully sequenced bacterial genomes downloaded from the NCBI ftp Genomes server. Protein sequences were defined as homologous if sharing at least 50% similarity and less than 20% difference in length. Pairs of homologous proteins were then expanded to homologous protein families by including all proteins homologous to at least one member of the family.

RESULTS

Taxonomic distribution

Although the production of R-bodies has been observed in a few free-living bacteria (see Introduction), the effective distribution of Reb homologs in prokaryotes has not been clear. We carried out an exhaustive search for Reb homologs in current sequence databases (see Materials and Methods) (Figure 2). We found no additional homologs of RebC other than the *C. taeniospiralis* pKAP298 plasmid, indicating that this protein is specific to the Caedibacter Reb system. On the contrary, RebB, RebA and RebD are homologous and widely distributed. We identified 203 Reb homologs from 64 taxa belonging exclusively to *Proteobacteria*, with the one exception of *Kordia algicida* OT-1, which belongs to the phylum *Bacteroidetes*, as recently noticed (Akiba *et al.* 2010). Reb homologs are widely distributed among representatives of four of the six subdivisions of *Proteobacteria*, Alpha, Beta, Gamma, and Delta (Figure 3). We found between one and nine Reb homologs in each genome (Figure 2). Although *reb* genes were first identified on the *C. taeniospiralis* pKAP298 plasmid, we found Reb homologs on only two additional plasmids: the megaplasmid from the α-proteobacterium *Ruegeria pomeroyi* DSS-3, and the AZOBR_p4 plasmid from the α-proteobacterium *Azospirillum brazilense* Sp245 (Figure 2). The availability of a complete genome for these two taxa indicates that no additional homologs are present on the chromosome. We could not find any other Reb homologs in viruses, Archaea, or Eukarya, apart from four homologs from the Global Ocean Sampling marine metagenome sequence database that are closely related to Proteobacteria (not shown).

Of importance, for three taxa with available sequence data, we could link for the first time the previously reported observation of R-bodies (see *Introduction* and references therein) with the presence of Reb homologs. In particular, *R. centenum* has six copies, *A. avenae* has four copies, and *M. mediterranea* has nine copies (Figure 2). Reb-containing taxa display a wide variety of lifestyles. Albeit many taxa harboring Reb homologs harbor free-living lifestyles in a wide variety of environments, from marine to terrestrial, a few taxa other than *C. taeniospiralis* appear to have an interaction with eukaryotes. For example, *Pseudovibrio sp. JE062* is a symbiont of sponges; *Vibrio fischeri ES114* is the specific bioluminescent symbiont in the light-emitting organs of certain squids and fishes; *Labrenzia alexandrii* and *Oceanicaulis alexandrii* have been isolated from dinoflagellates; *Stenotrophomonas maltophilia* R551-3, *Azospirillum brasilense* Sp245, and *Pseudomonas fluorescens* Pf-5 are plant growth−promoting endophytes; various strains of *Burkholderia ambifaria* are important in the biocontrol of pea plant phytopathogens; and *Azorhizobium caulinodans* is a nitrogen-fixing proteobacterium involved in mutualistic rhizobiale symbioses with plant roots.

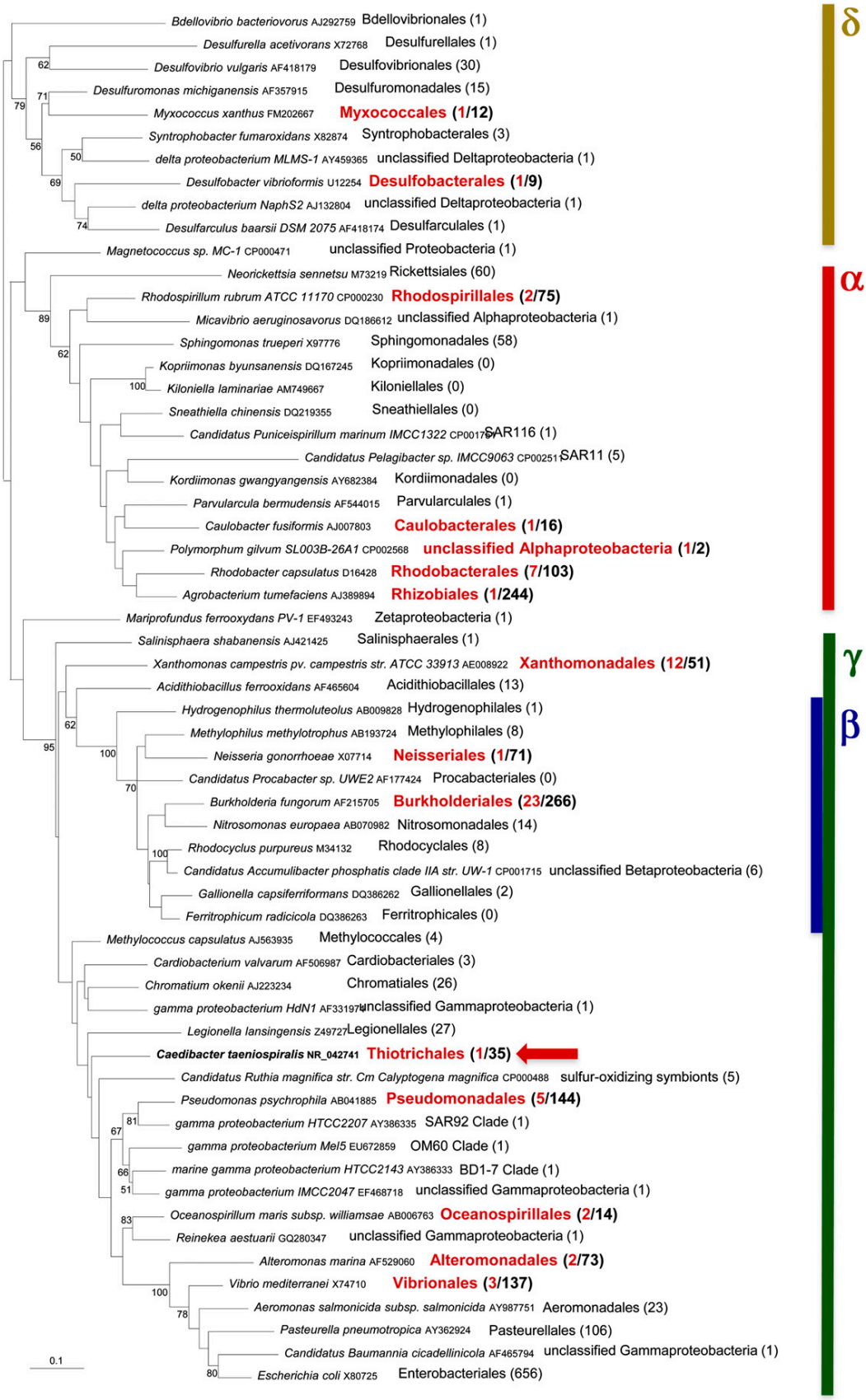
Other than in the algae pathogen *Kordia algicida*, we also found Reb homologs in a number of important pathogens of plants such as *Acidovorax avenae subsp. avenae* ATCC 19860; various strains of *Xanthomonas*; *Burkholderia gladioli and B. glumae*; *Ralstonia syzygii* R24; but also in important pathogens of aquatic animals such as shrimp and corals (*Vibrio nigripulchritudo* ATCC27043; *Vibrio coralliilyticus* ATCC-BAA450). Reb homologs were also found in the genomes of opportunistic pathogens of humans, such as various strains of *Pseudomonas aeruginosa*, including the hypervirulent LESB58 strain; in various strains of *Burkholderia pseudomallei*, the causative agent of meilodiosis, in *B. mallei*, that causes glanders and pneumonia; and *Stenotrophomonas maltophila*, a rare but serious threat to patients who require catheterization. We also observed some interesting patterns by looking at the distribution of Reb homologs in closely related strains (Figure 2). For example, although the sponge symbiont *Pseudovibrio sp*. JE062 harbors *reb* genes, its closely related free-living relative *Pseudovibrio sp*. FO-BEG1 does not. Similarly, *Burkholderia thailandensis* MSMB43, an opportunistic pathogen that causes meilodiosis, has *reb* genes, whereas the closely related *B. thailandensis* E264, a common soil and avirulent strain, does not (Figure 2).

Sequence analysis

Despite the widespread presence of Reb homologs in many bacterial taxa, it remains to be proven experimentally that these are responsible for producing R-bodies. However, some hints can already be gained from sequence analysis.

Reb homologs are 95 amino acids long on average. They show good conservation at the sequence level and all display a basic alpha helical secondary structure with no significant structural difference among sequences (Figure 4A). Very little is known about the regulation and mechanism of R-body assembly [for review see, (Sanchez-Amat 2006)]. We could not observe any particular pattern in the sequences that allows distinguishing the equivalents of RebA, B, and D of *C. taeniospiralis* in other taxa, and this was also confirmed by phylogenetic analysis. Overall sequence conservation is high, suggesting that these Reb homologs are likely functional and conserved at the structural level. A WebLogo analysis highlighted highly conserved positions (Figure 4B) that may be important for R-body assembly and/or unrolling-rolling, and should be the target of choice of future mutation studies. Unfortunately, we could not identify any homologous proteins with a solved crystal structure in extant databases (see *Materials and Methods*). Small proteins assembling into structures frequently display amphipathic helices, which consist of hydrophobic amino acids concentrated on one side and hydrophilic or polar amino acids on the other, and these can be highlighted using helical wheel diagrams (see *Materials and Methods*). However, we found no evidence that the helices of Reb homologs display amphipathic character (data not shown). Obtaining the 3-D structure of an R-body will therefore be essential to understand how Reb homologs assemble and function.

Figure 2 Distribution of Reb homologs. Presence/absence of Reb homologs in proteobacteria and *Kordia algicida*. Colors indicate the different proteobacterial subdivisions. For each genome that harbors Reb homologs, we included complete genomes of closely related taxa without any *reb* genes, when available. Taxa with no available complete genome sequence but harboring Reb homologs are highlighted in gray. For these taxa, the presence of extra Reb copies cannot be excluded. When present, Reb homologs are indicated by their corresponding accession number. Reb homologs located on plasmids are indicated by an asterisk. See main text for discussion.


Bdellovibrio bacteriovorus AJ292759 Bdellovibrionales (1)
Desulfurella acetivorans X72768 Desulfurellales (1)
Desulfovibrio vulgaris AF418179 Desulfovibrionales (30)
Desulfuromonas michiganensis AF357915 Desulfuromonadales (15)
Myxococcus xanthus FM202667 Myxococcales (1/12)
Syntrophobacter fumaroxidans X82874 Syntrophobacterales (3)
delta proteobacterium MLMS-1 AY459365 unclassified Deltaproteobacteria (1)
Desulfobacter vibrioformis U12254 Desulfobacterales (1/9)
delta proteobacterium NaphS2 AJ132804 unclassified Deltaproteobacteria (1)
Desulfarculus baarsii DSM 2075 AF418174 Desulfarculales (1)
Magnetococcus sp. MC-1 CP000471 unclassified Proteobacteria (1)
Neorickettsia sennetsu M73219 Rickettsiales (60)
Rhodospirillum rubrum ATCC 11170 CP000230 Rhodospirillales (2/75)
Micavibrio aeruginosavorus DQ186612 unclassified Alphaproteobacteria (1)
Sphingomonas trueperi X97776 Sphingomonadales (58)
Kopriimonas byunsanensis DQ167245 Kopriimonadales (0)
Kiloniella laminariae AM749667 Kiloniellales (0)
Sneathiella chinensis DQ219355 Sneathiellales (0)
Candidatus Puniceispirillum marinum IMCC1322 CP001073 SAR116 (1)
Candidatus Pelagibacter sp. IMCC9063 CP002511 SAR11 (5)
Kordiimonas gwangyangensis AY682384 Kordiimonadales (0)
Parvularcula bermudensis AF544015 Parvularculales (1)
Caulobacter fusiformis AJ007803 Caulobacterales (1/16)
Polymorphum gilvum SL003B-26A1 CP002568 unclassified Alphaproteobacteria (1/2)
Rhodobacter capsulatus D16428 Rhodobacterales (7/103)
Agrobacterium tumefaciens AJ389894 Rhizobiales (1/244)
Mariprofundus ferrooxydans PV-1 EF493243 Zetaproteobacteria (1)
Salinisphaera shabanensis AJ421425 Salinisphaerales (1)
Xanthomonas campestris pv. campestris str. ATCC 33913 AE008922 Xanthomonadales (12/51)
Acidithiobacillus ferrooxidans AF465604 Acidithiobacillales (13)
Hydrogenophilus thermoluteolus AB009828 Hydrogenophilales (1)
Methylophilus methylotrophus AB193724 Methylophilales (8)
Neisseria gonorrhoeae X07714 Neisseriales (1/71)
Candidatus Procabacter sp. UWE2 AF177424 Procabacteriales (0)
Burkholderia fungorum AF215705 Burkholderiales (23/266)
Nitrosomonas europaea AB070982 Nitrosomonadales (14)
Rhodocyclus purpureus M34132 Rhodocyclales (8)
Candidatus Accumulibacter phosphatis clade IIA str. UW-1 CP001715 unclassified Betaproteobacteria (6)
Gallionella capsiferriformans DQ386262 Gallionellales (2)
Ferritrophicum radicicola DQ386263 Ferritrophicales (0)
Methylococcus capsulatus AJ563935 Methylococcales (4)
Cardiobacterium valvarum AF506987 Cardiobacteriales (3)
Chromatium okenii AJ223234 Chromatiales (26)
gamma proteobacterium HdN1 AF331974 unclassified Gammaproteobacteria (1)
Legionella lansingensis Z49727 Legionellales (27)
Caedibacter taeniospiralis NR_042741 Thiotrichales (1/35)
Candidatus Ruthia magnifica str. Cm Calyptogena magnifica CP000488 sulfur-oxidizing symbionts (5)
Pseudomonas psychrophila AB041885 Pseudomonadales (5/144)
gamma proteobacterium HTCC2207 AY386335 SAR92 Clade (1)
gamma proteobacterium Mel5 EU672859 OM60 Clade (1)
marine gamma proteobacterium HTCC2143 AY386333 BD1-7 Clade (1)
gamma proteobacterium IMCC2047 EF468718 unclassified Gammaproteobacteria (1)
Oceanospirillum maris subsp. williamsae AB006763 Oceanospirillales (2/14)
Reinekea aestuarii GQ280347 unclassified Gammaproteobacteria (1)
Alteromonas marina AF529060 Alteromonadales (2/73)
Vibrio mediterranei X74710 Vibrionales (3/137)
Aeromonas salmonicida subsp. salmonicida AY987751 Aeromonadales (23)
Pasteurella pneumotropica AY362924 Pasteurellales (106)
Candidatus Baumannia cicadellinicola AF465794 unclassified Gammaproteobacteria (1)
Escherichia coli X80725 Enterobacteriales (656)
0.1
δ
α
γ
β

When present in multiple copies, Reb homologs are clustered on the genome, mostly lying side-by-side or separated by a few intervening genes (Figure 5). We generally found only one *reb* locus per genome, with the exception of *M. mediterranea* whose nine homologs are organized into two different genomic regions, and *Xanthomonas axonopodis* and *Chromobacterium violaceum*, which both have an extra Reb homolog located far from the main cluster (Figure 5). Given their short length, the phylogeny of Reb homologs is globally poorly resolved, but a few monophyletic groups are apparent which are consistent with genomic synteny (Supporting Information, Figure S1 and Figure 5). This allowed us to infer the evolutionary history of these Reb homologs (Figure 5). In some cases, Reb homologs from the same taxon are more closely related to each other than to Rebs of other taxa, suggesting that these have arisen from species-specific duplication events.

This is for, instance, the case of five *reb* genes from *Marinomonas sp*. MED121, which are all more closely related to each other than to any other Reb (Figure 5). The same can be said for the six of the seven *reb* genes from *Vibrio fischeri* ES114 (Figure 5). In other cases, there is clear evidence for vertical inheritance of Reb proteins from the ancestor of a specific *Proteobacterial* family (*e.g.*, *Xanthomonas*) (Figure 5). In yet other instances, Reb proteins are most closely related among distantly related lineages, suggesting horizontal transfer of the whole locus, for example in the case of Marinomonas mediterranea MMB-1 and Shewanella denitrificans OS217 (Figure 5). Horizontal gene transfer was also suggested for extra *reb* copies in a few taxa (Figure 5). Finally, phylogeny could not help in assigning the equivalents in other taxa of the RebA-B-D of *C. taeniospiralis*, as these are more closely related among themselves (Figure S1). It is therefore difficult to make analogies between the previously reported data on the role of the RebA-B-D of *C. taeniospiralis* in the assembly process of its R-bodies and what occurs in the other taxa. Moreover, because we found no homologs of RebC outside *C. taeniospiralis*, it is possible that other proteins have analogous function in Reb-harboring taxa. This finding would be consistent with RebC being a transcription regulator and therefore potentially species-specific.

In search for potential partners of Reb proteins

It has been shown that RebA, B and C from *C. taeniospiralis* are sufficient for production of type 51 R-body in *E. coli* but not for the killing phenotype (see *Introduction*), which indicates that yet-unidentified partners coded on either the plasmid or the chromosome of *C. taeniospiralis* are involved in the killing. Interestingly, we found no homologs of the 63 proteins encoded in the *C. taeniospiralis* pKAP298 plasmid in any of the Reb-harboring taxa. This may indicate that none of these proteins is a likely candidate for the killing toxin, which would be then encoded in the *C. taeniospiralis* genome (yet unavailable). Alternatively, the *C. taeniospiralis* toxin may well be on the plasmid but is not conserved in other bacteria, which may either display no killing activity or use nonhomologous toxins.

To search for candidate partners of Reb proteins, we carried out a genome synteny analysis of the *reb* locus in 41 taxa for which a complete genome or sufficient genomic structure information is available (Figure 6, see *Materials and Methods*). Two mutually exclusive synteny patterns could be observed (hereafter referred to as Group 1 and 2, respectively; Figure 6). Strikingly, the genes included in these conserved synteny patterns are exclusively present in Reb-harboring taxa, strongly suggesting a functional link with the Reb system. The Group 1 synteny pattern is defined by four proteins annotated as hypothetical: HP1.1 (red), HP1.2 (blue), HP1.3 (yellow), and HP1.4 (orange), which are only found in the surroundings of the *reb* locus and are only present in Reb-harboring taxa. The HP1.4 (~60−80 aa) and HP1.3 (!170 aa) proteins appear to be distant Reb homologs. However, they lack some of the conserved amino acid positions characteristic of other Reb proteins, and the HP1.3 protein is approximately twice as long as a typical Reb (data not shown). The HP1.1 (~360 aa) and the HP1.2 (~110−120 aa) proteins display no putative conserved domains. These four proteins often exhibit the same genomic organization (HP1.1,HP1.2,HP1.3, HP1.4; Figure 6). In three cases, another hypothetical protein (HP1.5, purple) is associated with this context, and is distantly related to the HP1.3 protein (Figure 6).

The Group 2 synteny pattern is defined by the presence of two proteins annotated as hypothetical: HP2.1 (light blue) and HP2.2 (fuchsia) (Figure 6). These proteins are approximately the same size (~205−220 aa) and display no putative conserved domains. They frequently co-occur, but their genomic organization varies in different taxa. An interesting characteristic of Group 2 synteny pattern is the frequent association with a putative RNA polymerase sigma-factor protein (HP2.3, light pink) and a transcriptional regulator/cyclic nucleotide binding protein (HP2.4, dark pink), which might be involved in transcription regulation of *reb* genes in these taxa. It should be noted that the conserved Group 1 and Group 2 synteny patterns are generally consistent with the Reb clusters highlighted by phylogenetic analysis (Figure 5) and have phylogenies similar to the Reb one (not shown), indicating a common evolutionary history and providing further suggestion of a functional link between these proteins and Reb proteins. As an additional strong indication, the plasmid sequence of *A. brasilense* contains the four proteins characteristic of the Group 1 synteny pattern.

A few taxa did not present any particularly conserved genomic context nor did they harbor the conserved genes found in Group 1 or Group2 synteny patterns. It is therefore possible that other genetic elements important for *reb* function are located in different positions of the genome in these taxa. To this end, we sought to identify additional proteins specific to Reb-harboring taxa by carrying out a whole-genome content analysis (see *Material and Methods*). Using the complete genomes of 841 bacterial taxa—including 25 Reb-harboring complete genomes—we constructed protein families having at least 50% identity and 80% size conservation (see *Material and Methods*). This analysis confirmed that the only protein family exclusively present in Reb-harboring taxa is the Reb family itself, along with the protein families specific to Group 1 and 2 synteny patterns (Figure 7). Although genomes of taxa not containing Reb homologs harbored distant homologs of the HP2.3 and HP2.4 (light pink and dark pink) proteins of genome synteny patterns, these fell outside of subfamilies which are exclusively present in Reb-harboring taxa.

Figure 3 Distribution of Reb-harboring taxa across Proteobacteria. Unrooted Maximum likelihood phylogenetic tree of 16s rRNA sequences from 60 taxa representative of proteobacterial diversity. Proteobacterial orders that include members containing Reb homologs are highlighted in red. The number of Reb-harboring taxa over the total number of available complete genomes is indicated in parenthesis. *Caedibacter taeniospiralis* belongs to the gammaproteobacterial family of Thiotrichales (indicated by a red arrow). The tree was obtained using Treefinder with the J1 model of nucleotide substitution and a discrete gamma distribution with four categories to take into account among-site rate variation. Numbers at nodes indicate bootstrap values (BV) for 100 replicates of the original dataset. For clarity, only BVs greater than 50% are shown. The scale bar represents the average number of substitutions per site.

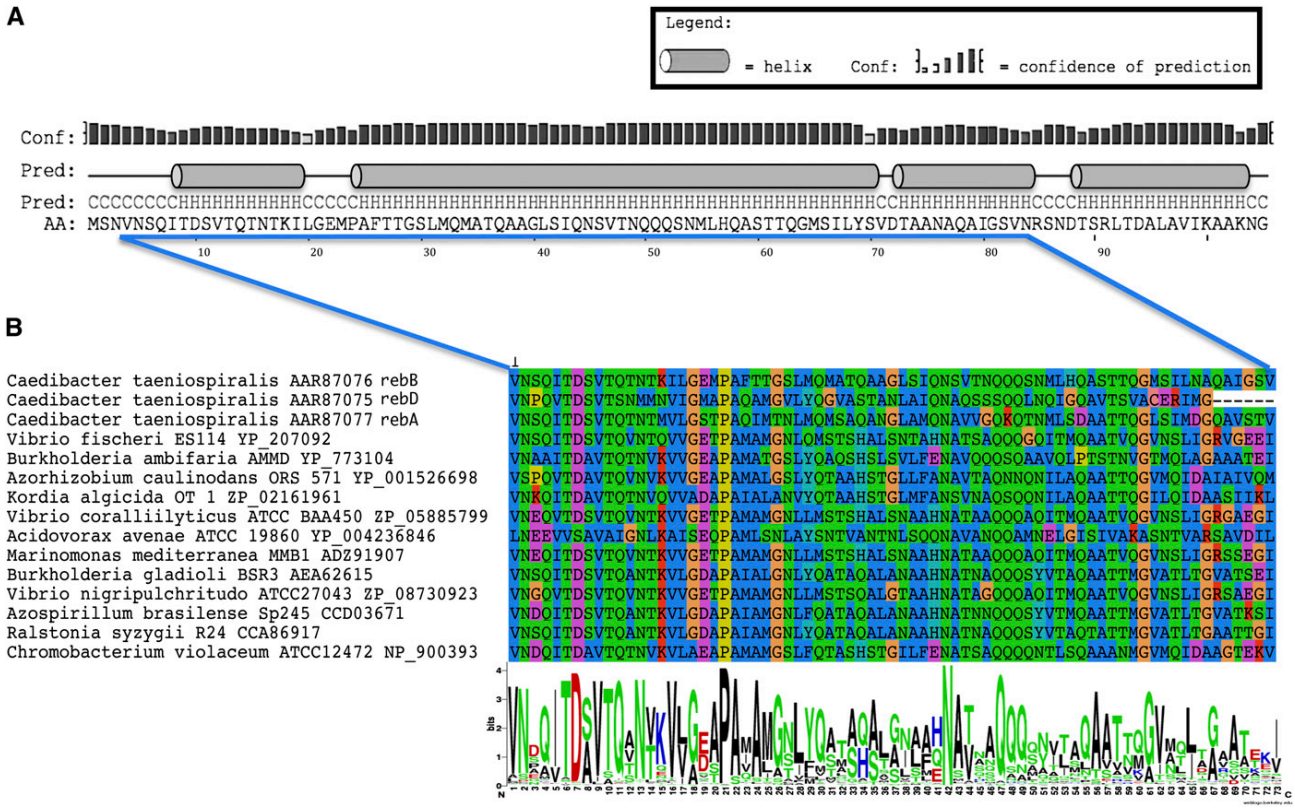


Figure 4 Sequence analysis. (A) Secondary structure of the RebB of *C. taeniospiralis* predicted by PSIPRED [http://bioinf.ucl.ac.uk/psipred (Buchan *et al.* 2010)]. The same structure was substantially conserved in all other Reb homologs. (B) Conserved amino acid positions identified using Weblogo on an unambiguously aligned excerpt of the entire alignment of the 203 identified Reb homologs. For clarity, only 15 representative Reb sequences are shown. Position numbers refer to the RebB of *C. taeniospiralis*.

These few conserved proteins might be either involved in regulation of R-body assembly and function, or represent the toxin, and should be priority targets for future studies. It will also be interesting to test whether the Reb-harboring taxa that harbor none of these candidate partners are able to make R-bodies or display killing activity. Finally, it should be noted that none of these proteins belong to the Soj-ParA family or any annotated membrane-associated ATPase, weakening the previous hypothesis that these types of protein may represent the toxin responsible for killing (Jeblick and Kusch 2005).

A phage origin?

Jeblick and Kush emphasized the presence of phage-related genes on the *reb* carrying plasmid (Jeblick and Kusch 2005) and Preer (Preer *et al.* 1974) observed an association of phage-like particles with R-bodies, suggesting that R-bodies may be encoded by defective phage genes. Moreover, the evolutionary analysis of Reb families (Figure 5) and their genomic context (Figure 6) suggest horizontal gene transfer events, for which bacteriophages are known to be major contributors. However, we found no Reb homologs in genomic sequences from phages. We therefore sought to see if Reb homologs are part of integrated elements or prophages. We examined the 40 kbp on each side of the *reb* locus in the 25 Reb-harboring taxa for which complete genome sequences are available for the presence of prophages (integrated phages) or other phage-related elements. First, we searched using the PHAST database [http://phast.wishartlab.com (Zhou *et al.* 2011)], which contains phage proteins that have been associated with a clear phage function. However, none of these regions were positive for prophage sequences. As a complementary analysis, we specifically searched a comprehensive local databank of 1130 bacteriophage sequences downloaded from GenBank (December 2011), which included 248 phages isolated from 33 proteobacterial genera. These regions did not display any specific similarity to phage elements. We also looked at whether *reb* loci are embedded in genomic islands by running searches on the IslandViewer server [http://www.pathogenomics.sfu.ca/islandviewer (Langille and Brinkman 2009)], which combines several prediction methods: (1) atypical dinucleotide content; or (2) codon usage; (3) identification of unique regions not present in closely related genomes; and (4) presence of genes that are functionally related to mobile elements. However, none of the analyzed genomes displayed identified potential genomic islands adjacent to or surrounding the *reb* locus. We also verified from the literature whether Reb homologs were present in any previously reported genomic region of potential exogenous origin. For example, *reb* genes did not fall into any of the two atypical regions identified in the genome of *Xanthomonas oryzae pv. oryzae* PXO99A (Salzberg *et al.* 2008), nor in the four atypical regions highlighted in the genome of *Xanthomonas campestris pv. campestris* str. ATCC 33913 (Vorholter *et al.* 2003), and were not included in any of the prophage islands identified in the genome of the *Pseudomonas aeruginosa* hypervirulent LESB58I strain (Winstanley *et al.* 2009). Finally, among 1062 plasmid sequences available from *Proteobacteria*, we detected *reb* genes only on two plasmids other than the *C. taeniospiralis* plasmid pKAP298: the *Azospirillum brasilense* Sp245 plasmid AZOBR_p4 and the *Ruegeria pomeroyi* DSS-3 megaplasmid.

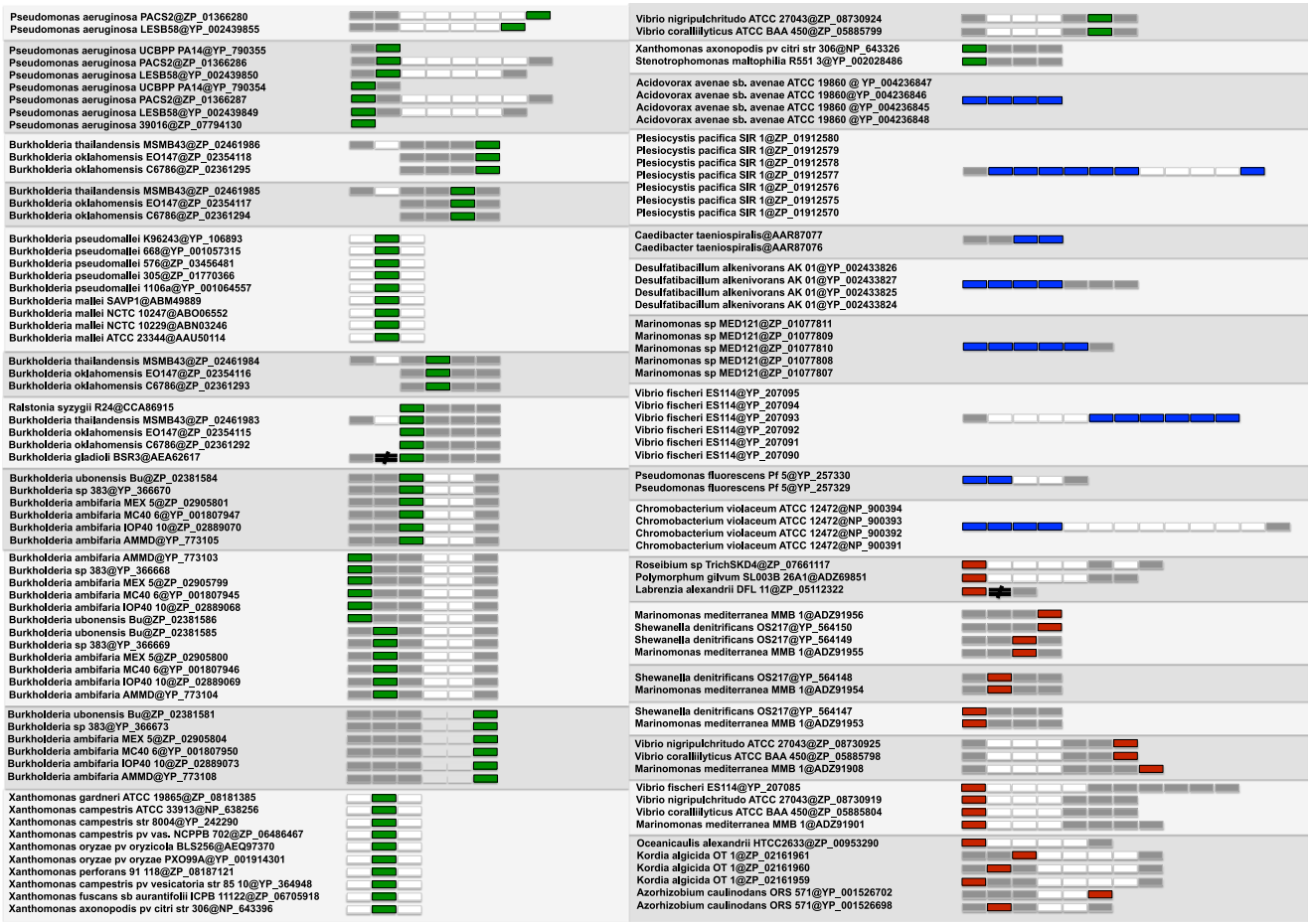


Figure 5 Evolutionary inference of Reb homologs based on phylogenetic analysis of the 203 Reb homologs (Figure S1). Here we have highlighted a few of the monophyletic groups. For each taxon, the genome locations of the corresponding Reb proteins are shown. Reb homologs highlighted in green are orthologs that were inferred to have been inherited through speciation events; those highlighted in blue represent paralogs issued from species-specific gene duplications; and those in red are the Reb homologs that have likely originated via horizontal transfer. Adjacent *reb* genes are indicated in gray. Open reading frames between *reb* genes are shown in white, and black slash-like symbols represent large intervening regions between *reb* genes.

DISCUSSION

Despite having been continuously and intensively studied from the 1930s through the 1980s, recent data on R-bodies have been scarce. With a whole array of novel technology, studies on the diversity and role of these puzzling bacterial structures can now be fully tackled. Our exhaustive analysis shows that Reb homologs are widely present in Proteobacterial genomes spanning the diversity of this major bacterial phylum, indicating that they are much more widespread than previously known. In the perspective of obtaining experimental data, our analysis remains for the time being largely descriptive, but nevertheless provides a number of interesting hints for discussion and future work.

Sequence analysis suggests structural and functional conservation, indicating that Reb homologs are likely responsible for the production of functional R-bodies in all the taxa where we found them, although this needs to be verified experimentally. Moreover, the presence of Reb homologs in bacteria where R-bodies have been previously observed is already a good hint. The occurrence of R-bodies in a wide range of bacteria harboring Reb homologs should be tested, with priority given to those that have medical, agricultural, and ecological implications. Our data will also help direct mutational studies to characterize the system further through structural and functional analysis of *reb* genes from Caedibacter but also other taxa, including those that harbor multiple *reb* genes and those that have only one copy. Also, it will be important to verify the involvement in R-body production, assembly, regulation, and killing of the likely partners that we have identified by genome context and whole genome content analysis. Because none of the Reb-harboring genomes identified in this study possess homologs of the genes carried by the *C. taeniospiralis* plasmid, it is possible that the killing toxin is encoded in the *C. taeniospiralis* genome and is perhaps one of the candidate genes that we have identified. The completion of this genome will therefore be very important. Another possibility is that the R-bodies are delivery systems for species-specific toxins.

Our analysis of taxonomic distribution shows that Reb homologs are present in taxa displaying very different lifestyles, suggesting that the role of R-bodies in nature could be quite diverse. Moreover, we show that intact *reb* loci have been spread among Proteobacterial taxa via horizontal transfer, indicating that an advantage exists in acquiring and keeping R-bodies. However, we found no clear signs of a phage origin for the *reb* loci. It remains possible that this is due to an undersampling of phages from the Reb-harboring taxa or that these horizontal transfers correspond to events old enough to have allowed sequences to adapt to the new genome, or that all traces of the transfer vectors have been erased from the genome after transfer.

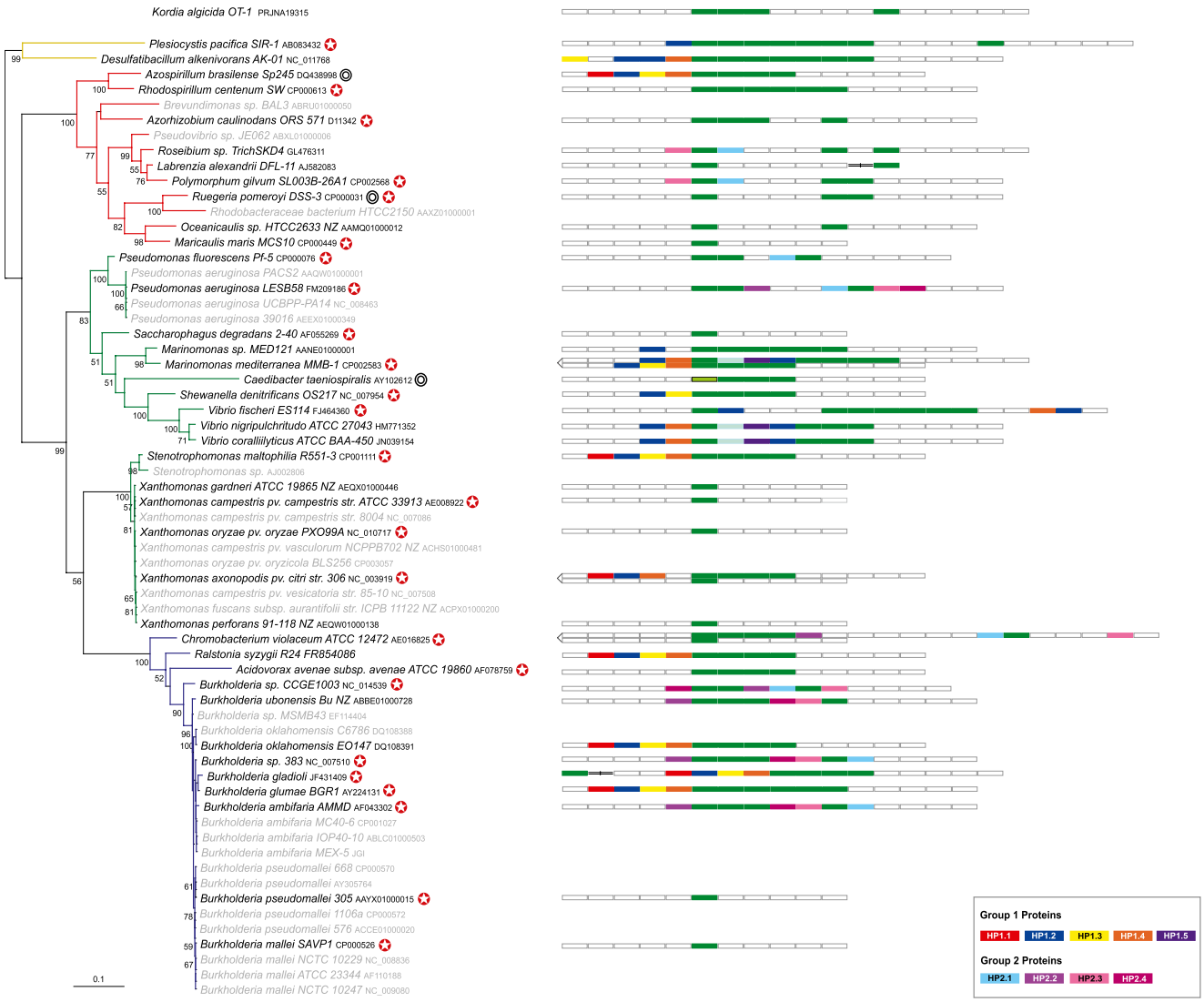


Figure 6 Genome synteny analysis of the *reb* locus mapped onto an Unrooted Maximum likelihood phylogenetic tree of 16s rRNA sequences from the 64 *reb*-containing Proteobacteria. The tree was created using Treefinder with the GTR model of nucleotide substitution and a discrete gamma distribution with four categories to take into account among-site rate variation. Numbers at nodes indicate BVs for 100 replicates of the original dataset. For clarity, only BVs greater than 50% are shown. The scale bar represents the average number of substitutions per site. Species where Reb homologs are located on a plasmid are marked by a black circle. A white star in a red circle marks the fully sequenced genomes used in the analysis. Reb homologs are shown in green. Homologous genes are represented by the same color. For clarity, only genes discussed in the text are indicated. Black slash-like symbols represent large regions in between genes. The RebC of *C.taeniospiralis* (AAR87131) is shown in light green and outlined in black to indicate its lack of homology with the other *rebs*. The genome context for the *Flavobacterium K. algicida* is shown separately.

Some of the Reb-harboring taxa have very important ecological, agricultural, and medical relevance. In addition, by observing the pattern of presence/absence of *reb* genes in closely related strains, we found intriguing links between virulence and presence of *reb* genes that will surely be worthy of further investigation. R-bodies may be involved in mediating interactions of these bacteria with eukaryotic cells, perhaps through the triggering of unrolling when ingested in the vacuole, similarly to what observed in the Caedibacter/Paramecium interaction.

We found Reb homologs in many free-living bacteria. R-bodies in these bacteria may be used as a defense mechanism against grazing by eukaryotes. Bacteria have in fact developed various strategies against protozoan predation in nature (Matz and Kjelleberg 2005). Many examples have been reported of cytotoxicity responses against grazing by eukaryotes in different bacteria such as *Pseudomonas* (Matz *et al.* 2004; Weitere *et al.* 2005) and *Vibrio* (Erken *et al.* 2011). Also, it is tempting to speculate that some of these free-living, Reb-harboring taxa can establish transient "killing" symbioses with ciliates or other protists, as seems to be indicated by the fact that some Reb-harboring taxa have indeed been isolated from aquatic microbial eukaryotes. It will be interesting to perform feeding experiments to test their killing potential in Paramecium and also other protistan taxa, such as algae. Indeed, we confirmed the presence of *reb* genes in *Kordia algicida*, a planktonic bacterium recently highlighted as a killer of diatoms by a yet unclear mechanism involving an unidentified protease (Paul and Pohnert 2011). Our study suggests that delivery of this protease could

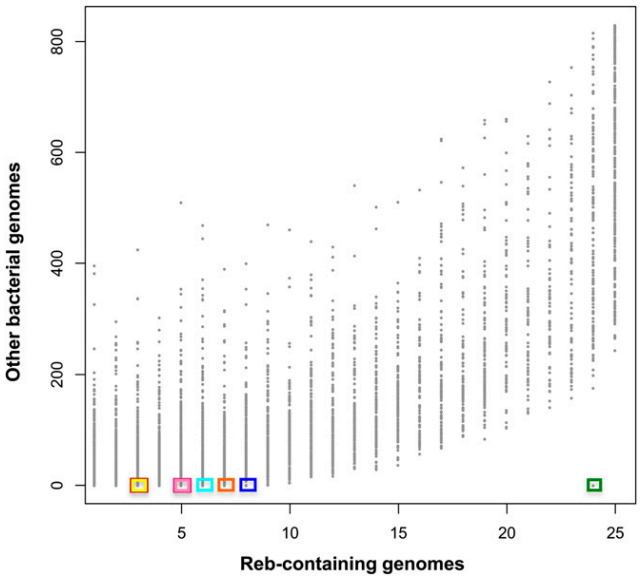


Figure 7 Whole genome content analysis. Graphical representation of protein families created using the R software (R Development Core Team 2011). The 25 fully sequenced Reb-harboring taxa are represented on the x-axis and the other 816 fully sequenced bacterial taxa analyzed are represented on the y-axis. Each point on the graph represents a protein family (see *Materials and Methods* for details on how protein families were defined). For example, the Reb family (indicated by a green box) is present in 24 fully sequenced Reb-harboring taxa but in none of the remaining genomes. The 4 Rebs of *Acidovorax avenae* subsp. Avenae ATCC 19860 did not fall into the Reb protein family because they are very divergent (see Figure 5 and Figure S1). The other five unique protein families specific to Reb-harboring taxa (see main text) are shown with boxes corresponding to colors as defined in the legend to Figure 6.

be performed via R-bodies, which would therefore be important players in the regulation of algal blooms. Similarly, it would be interesting to verify whether the presence of Reb homologs in the powerful coral pathogen *Vibrio coralliilyticus* is linked to R-body production and if these are somehow involved in delivery of the killing toxin. The killing factor produced by *K. algicida* is triggered independently of the presence of the diatom target but rather likely depends on a quorum sensing mechanism when the population size reaches a certain density (Paul and Pohnert 2011). Similarly, the triggering of R-body production in a fraction of the Caedibacter endosymbiont population in Paramecium, a phenomenon that is not yet understood, may be regulated by a quorum sensing mechanism.

In addition, our study suggests that R-bodies may be involved in the interaction of Proteobacteria with several multicellular organisms, such as plants and animals. The recently reported involvement of *reb* genes in the regulation of a rhizobial symbiosis (Akiba *et al.* 2010) and the presence of Reb homologs in a number of Proteobacteria known to interact with plant roots is intriguing, and it is not excluded that R-bodies may help the bacterium to move through plant tissues, via delivery of specific lytic compounds. Indeed, we found Reb homologs in a number of bacterial strains known to be able to penetrate the xylem of plants. A similar mechanism may be used to move through tissues by some Proteobacteria that interact with animals, such as *Vibrio fischeri* with its squid host. Finally, the presence of *reb* genes in important pathogens of eukaryotes, including humans, some of which are responsible for emerging and poorly characterized infections, should prompt the study of their potential involvement in the infection process, perhaps by helping tissue invasion.

If our predictions are verified, bacteria may represent a largely overlooked role in the regulation of microbial eukaryotic abundance and distribution, in addition to the much more studied impact of viruses. This regulation may be performed at different levels, by direct killing of eukaryotic grazers, but also by providing mechanisms used for defense among eukaryotes, as is the example of the Paramecium/Caedibacter symbiosis. Interestingly, it was recently reported that the thricocysts of eukaryotic algae belonging to the Cryptomonads, ejectile organelles that are probably used with a defensive role against predation, are composed of four proteins that share similarity with Reb proteins (Yamagishi *et al.* 2012). The authors proposed that these proteins were acquired horizontally from Proteobacteria. R-body−harboring bacteria could therefore play a larger role in the origin and spread of defense mechanisms in eukaryotic microorganisms. Finally, elucidation of the mechanism of rolling/unrolling/toxin delivery of R-bodies will surely open the way to interesting biotechnological applications.

ACKNOWLEDGMENTS

We would like to thank Céline Brochier-Armanet for kindly sharing 16S rRNA trees of Proteobacteria and for useful comments on the manuscript, along with Patrick Forterre and Eduardo Rocha for interesting discussion. We also wish to thank Mart Krupovic for advice on secondary structure analysis. K.R. is the recipient of a Pasteur Ph.D. International grant and a Paul Zuccaire fellow and would like to thank John R. and Louise B. Preer for early training in Paramecium and killer endosymbiont biology. After completion of this work, we became aware that an analysis partially similar to ours was independently carried out and discussed in the Ph.D. thesis of Martina Schrallhammer (2010, unpublished) under the supervision of Giulio Petroni, whom we thank for bringing it to our attention. We also thank two anonymous referees for very useful comments that helped improve the manuscript.

LITERATURE CITED

Akiba, N., T. Aono, H. Toyazaki, S. Sato, and H. Oyaizu, 2010 phrR-like gene praR of *Azorhizobium caulinodans* ORS571 is essential for symbiosis with Sesbania rostrata and is involved in expression of reb genes. Appl. Environ. Microbiol. 76: 3475–3485.

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang *et al.*, 1997 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25: 3389–3402.

Beale, G., and J. R. Preer Jr., 2004 *Paramecium: Genetics and Epigenetics*, CRC Press, Boca Raton, FL.

Beier, C. L., M. Horn, R. Michel, M. Schweikert, H. D. Gortz *et al.*, 2002 The genus Caedibacter comprises endosymbionts of Paramecium spp. related to the Rickettsiales (Alphaproteobacteria) and to *Francisella tularensis* (Gammaproteobacteria). Appl. Environ. Microbiol. 68: 6043–6050.

Brown, M. R., and J. Barker, 1999 Unexplored reservoirs of pathogenic bacteria: protozoa and biofilms. Trends Microbiol. 7: 46–50.

Buchan, D. W., S. M. Ward, A. E. Lobley, T. C. Nugent, K. Bryson *et al.*, 2010 Protein annotation and modelling servers at University College London. Nucleic Acids Res. 38: W563–568.

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

Dilts, J. A., and R. L. Quackenbush, 1986 A mutation in the R body-coding sequence destroys expression of the killer trait in P. tetraurelia. Science 232: 641–643.

Edgar, R. C., 2004 MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5: 113.

Erken, M., M. Weitere, S. Kjelleberg, and D. McDougald, 2011 In situ grazing resistance of *Vibrio cholerae* in the marine environment. FEMS Microbiol. Ecol. 76: 504–512.

Espuny, M. J., C. Andres, M. E. Mercade, M. Robert, M. A. Manresa *et al.*, 1991 R-bodies in *Pseudomonas aeruginosa* strain 44T1. Antonie van Leeuwenhoek 60: 83–86.

Favinger, J., R. Stadtwald, and H. Gest, 1989 Rhodospirillum centenum, sp. nov., a thermotolerant cyst-forming anoxygenic photosynthetic bacterium. Antonie van Leeuwenhoek 55: 291–296.

Fusté, M., M. Simon-Pujol, A. Marques, J. Guinea, and F. Congregrado, 1986 Isolation of a new free-living bacterium containing R-bodies. J. Gen. Microbiol. 132: 2801–2805.

Gast, R. J., R. W. Sanders, and D. A. Caron, 2009 Ecological strategies of protists and their symbiotic relationships with prokaryotic microbes. Trends Microbiol. 17: 563–569.

Gibson, I., 1973 Transplantation of killer endosymbionts in paramecium. Nature 241: 127–129.

Hernandez-Romero, D., P. Lucas-Elio, D. Lopez-Serrano, F. Solano, and A. Sanchez-Amat, 2003 *Marinomonas mediterranea* is a lysogenic bacterium that synthesizes R-bodies. Microbiology 149: 2679–2686.

Heruth, D. P., F. R. Pond, J. A. Dilts, and R. L. Quackenbush, 1994 Characterization of genetic determinants for R body synthesis and assembly in *Caedibacter taeniospiralis* 47 and 116. J. Bacteriol. 176: 3559–3567.

Jeblick, J., and J. Kusch, 2005 Sequence, transcription activity, and evolutionary origin of the R-body coding plasmid pKAP298 from the intracellular parasitic bacterium *Caedibacter taeniospiralis*. J. Mol. Evol. 60: 164–173.

Jobb, G., A. von Haeseler, and K. Strimmer, 2004 TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics. BMC Evol. Biol. 4: 18.

Jurand, A., B. M. Rudman, and J. R. Preer Jr., 1971 Prelethal effects of killing action by stock 7 of *Paramecium aurelia*. J. Exp. Zool. 177: 365–387.

Kanabrocki, J. A., R. L. Quackenbush, and F. R. Pond, 1986 Organization and expression of genetic determinants for synthesis and assembly of type 51 R bodies. J. Bacteriol. 168: 40–48.

Kusch, J., L. Czubatinski, S. Wegmann, M. Hubner, M. Alter *et al.*, 2002 Competitive advantages of Caedibacter-infected Paramecia. Protist 153: 47–58.

Lalucat, J., and F. Mayer, 1978 "Spiral bodies"–intracytoplasmic membraneous structures in a hydrogen oxidizing bacterium. Z. Allg. Mikrobiol. 18: 517–521.

Langille, M. G., and F. S. Brinkman, 2009 IslandViewer: an integrated interface for computational identification and visualization of genomic islands. Bioinformatics 25: 664–665.

Lartillot, N., T. Lepage, and S. Blanquart, 2009 PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics 25: 2286–2288.

Matz, C., and S. Kjelleberg, 2005 Off the hook–how bacteria survive protozoan grazing. Trends Microbiol. 13: 302–307.

Matz, C., T. Bergfeld, S. A. Rice, and S. Kjelleberg, 2004 Microcolonies, quorum sensing and cytotoxicity determine the survival of Pseudomonas aeruginosa biofilms exposed to protozoan grazing. Environ. Microbiol. 6: 218–226.

Paul, C., and G. Pohnert, 2011 Interactions of the algicidal bacterium *Kordia algicida* with diatoms: regulated protease excretion for specific algal lysis. PLoS ONE 6: e21032.

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

Pond, F. R., I. Gibson, J. Lalucat, and R. L. Quackenbush, 1989 R-body-producing bacteria. Microbiol. Rev. 53: 25–67.

Preer, J. R. Jr., 1975 The hereditary symbionts of *Paramecium aurelia*. Symp. Soc. Exp. Biol. 29: 125–144.

Preer, J. R. Jr., 2006 Sonneborn and the Cytoplasm. Genet. Soc. Am. 172: 1373–1377.

Preer, J. R. Jr., and P. Stark, 1953 Cytological observations on the cytoplasmic factor "kappa" in *Paramecium aurelia*. Exp. Cell Res. 5: 478–491.

Preer, J. R. Jr., L. B. Preer, and A. Jurand, 1974 Kappa and other endosymbionts in *Paramecium aurelia*. Bacteriol. Rev. 38: 113–163.

Quackenbush, R. L., and J. A. Burbach, 1983 Cloning and expression of DNA sequences associated with the killer trait of *Paramecium tetraurelia* stock 47. Proc. Natl. Acad. Sci. USA 80: 250–254.

R Development Core Team, 2011 R: *A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna. Available at: http://www.R-project.org.

Salzberg, S. L., D. D. Sommer, M. C. Schatz, A. M. Phillippy, P. D. Rabinowicz *et al.*, 2008 Genome sequence and rapid evolution of the rice pathogen Xanthomonas oryzae pv. oryzae PXO99A. BMC Genomics 9: 204.

Sanchez-Amat, A., 2006 R-bodies. Microbiol. Monogr. 1: 331–341.

Schrallhammer, M., S. Galati, J. Altenbuchner, M. Schweikert, H. D. Gortz *et al.*, 2012 Tracing the role of R-bodies in the killer trait: Absence of toxicity of R-body producing recombinant *E. coli* on paramecia. Eur. J. Protistol. 48: 290–296.

Sonneborn, T. M., 1938 Mating types in *P. aurelia*: diverse conditions for mating in different stocks; occurrence, number and interrelations of the types. Proc. Am. Philos. Soc. 79: 411–434.

Sonneborn, T. M., 1943 Gene and cytoplasm: II. The bearing of the determination and inheritance of characters in *Paramecium aurelia* on the problems of cytoplasmic inheritance, pneumococcus transformations, mutations and development. Proc. Natl. Acad. Sci. USA 29: 338–343.

Vorholter, F. J., T. Thias, F. Meyer, T. Bekel, O. Kaiser *et al.*, 2003 Comparison of two Xanthomonas campestris pathovar campestris genomes revealed differences in their gene composition. J. Biotechnol. 106: 193–202.

Weitere, M., T. Bergfeld, S. A. Rice, C. Matz, and S. Kjelleberg, 2005 Grazing resistance of *Pseudomonas aeruginosa* biofilms depends on type of protective mechanism, developmental stage and protozoan feeding mode. Environ. Microbiol. 7: 1593–1601.

Wells, B., and R. W. Horne, 1983 The ultrastructure of Pseudomonas avenae II. Intracellular refractile (R-body) structure. Micron 14: 329–344.

Winstanley, C., M. G. Langille, J. L. Fothergill, I. Kukavica-Ibrulj, C. Paradis-Bleau *et al.*, 2009 Newly introduced genomic prophage islands are critical determinants of in vivo competitiveness in the Liverpool epidemic strain of *Pseudomonas aeruginosa*. Genome Res. 19: 12–23.

Yamagishi, T., A. Kai, and H. Kawai, 2012 Trichocyst ribbons of a cryptomonads are constituted of homologs of R-body proteins produced by the intracellular parasitic bacterium of Paramecium. J. Mol. Evol. 74:147–157.

Zhou, Y., Y. Liang, K. H. Lynch, J. J. Dennis, and D. S. Wishart, 2011 PHAST: a fast phage search tool. Nucleic Acids Res. 39: W347–352.

Communicating editor: A. Rokas

Appendix 2: Collaborations

During my Ph.D I had the opportunity to be involved in three collaborative projects, through my phylogenetic skills and extensive knowledge of the archaeal phylogeny, as well as the use of my newly identified set of DNA replication components. Two of these collaborative projects have now been published and one is currently under revision in Nature Communications.

Collaboration 1

Comparative genomics of seventh order of Methanogens

Although methanogens are known to inhabit the animal digestive tract, not much is known about their diversity and evolutionary history. As described in the Introduction section, a new lineage of methanogens has been only recently identified from the gut of animals and humans that is not related to any previously known archaeal methanogens but rather affiliated to the Thermoplasmatales. The group of Jean-François Brugere in Clermont-Ferrand, France, has contacted us for collaboration in the frame of a comparative genomic study on the first three complete genomes from this order. I identified and performed phylogenetic analysis on the components of DNA replication to determine their evolutionary history and gain more insight on the relationship of these three lineages to the other archaeal lineages. The replication gene set is very similar to that of the most closely related lineages, i.e. Thermoplasmatales, Marine Group II and DHVE2. However, some unique features were found in the DNA replication components of Methanomassiliicoccales such as some species-specific horizontally acquired replication proteins. Phylogenetic analysis of Cdc6 proteins reveals the presence of the two predicted ancestral copies orc1/cdc6-1 and orc1/cdc6-2 in all three genomes. This study reveals many insights into genome organization and metabolic traits of the seventh order of methanogens, and suggests contrasted evolutionary history among the three analyzed Methanomassiliicoccales representatives.

These results were published in BMC Genomics on the 13 August 2014.

Collaboration Article 1

Borrel *et al. BMC Genomics* 2014, **15**:679
http://www.biomedcentral.com/1471-2164/15/679

BMC Genomics

RESEARCH ARTICLE Open Access

Comparative genomics highlights the unique biology of Methanomassiliicoccales, a Thermoplasmatales-related seventh order of methanogenic archaea that encodes pyrrolysine

Guillaume Borrel[1,2†], Nicolas Parisot[1,3†], Hugh MB Harris[2], Eric Peyretaillade[1], Nadia Gaci[1], William Tottey[1], Olivier Bardot[4], Kasie Raymann[5,6], Simonetta Gribaldo[5,6], Pierre Peyret[1], Paul W O'Toole[2] and Jean-François Brugère[1*]

Abstract

Background: A seventh order of methanogens, the Methanomassiliicoccales, has been identified in diverse anaerobic environments including the gastrointestinal tracts (GIT) of humans and other animals and may contribute significantly to methane emission and global warming. Methanomassiliicoccales are phylogenetically distant from all other orders of methanogens and belong to a large evolutionary branch composed by lineages of non-methanogenic archaea such as Thermoplasmatales, the Deep Hydrothermal Vent Euryarchaeota-2 (DHVE-2, *Aciduliprofundum boonei*) and the Marine Group-II (MG-II). To better understand this new order and its relationship to other archaea, we manually curated and extensively compared the genome sequences of three Methanomassiliicoccales representatives derived from human GIT microbiota, "*Candidatus* Methanomethylophilus alvus", "*Candidatus* Methanomassiliicoccus intestinalis" and *Methanomassiliicoccus luminyensis*.

Results: Comparative analyses revealed atypical features, such as the scattering of the ribosomal RNA genes in the genome and the absence of eukaryotic-like histone gene otherwise present in most of Euryarchaeota genomes. Previously identified in Thermoplasmatales genomes, these features are presently extended to several completely sequenced genomes of this large evolutionary branch, including MG-II and DHVE2. The three Methanomassiliicoccales genomes share a unique composition of genes involved in energy conservation suggesting an original combination of two main energy conservation processes previously described in other methanogens. They also display substantial differences with each other, such as their codon usage, the nature and origin of their CRISPRs systems and the genes possibly involved in particular environmental adaptations. The genome of *M. luminyensis* encodes several features to thrive in soil and sediment conditions suggesting its larger environmental distribution than GIT. Conversely, "*Ca.* M. alvus" and "*Ca.* M. intestinalis" do not present these features and could be more restricted and specialized on GIT. Prediction of the *amber* codon usage, either as a termination signal of translation or coding for pyrrolysine revealed contrasted patterns among the three genomes and suggests a different handling of the Pyl-encoding capacity.
(Continued on next page)

* Correspondence: jf.brugere@udamail.fr
†Equal contributors
[1]EA-4678 CIDAM, Clermont Université, Université d'Auvergne, 28 Place Henri Dunant, BP 10448, 63000 Clermont-Ferrand, France
Full list of author information is available at the end of the article

(Continued from previous page)

Conclusions: This study represents the first insights into the genomic organization and metabolic traits of the seventh order of methanogens. It suggests contrasted evolutionary history among the three analyzed Methanomassiliicoccales representatives and provides information on conserved characteristics among the overall methanogens and among Thermoplasmata.

Keywords: Archaea, Methanomassiliicoccales, *Methanomethylophilus*, *Methanomassiliicoccus*, Origin of replication (ORI) binding (ORB) motif, Genome streamlining, CRISPR, Pyrrolysine Pyl, H_2-dependent methylotrophic methanogenesis, Energy conservation

Background

Methanogenic archaea are distributed worldwide in anaerobic environments and account for a large proportion of methane emissions into the atmosphere, partly due to anthropogenic activity (*e.g.* rice fields and livestock). Over the last ten years, sequences of novel archaeal lineages distantly related to all orders of methanogens have recurrently been found in diverse anaerobic environments. One of these lineages, phylogenetically related to the Thermoplasmatales, was first reported in the rumen [1,2] and was thereafter referred as Rumen Cluster-C in this environment [3]. The methanogenic nature of these archaea was subsequently strongly supported by the co-occurrence in human stool samples of 16S rRNA affiliated to this lineage and *mcrA* genes (a functional marker of methanogens) distantly related to any other methanogens [4,5]. The final evidence that they represent a new order of methanogens was recently given with the isolation of *Methanomassiliicoccus luminyenis B10* from human feces [6] and the culture in consortia of several strains of this order: "*Candidatus* Methanomethylophilus alvus" [7] and "*Candidatus* Methanomassiliicoccus intestinalis" [8] from human feces samples, MpT1 and MpM2 [9] from termite gut and "*Candidatus* Methanogranum caenicola" [10] from waste treatment sludge. All the culture-based studies agreed on a common methanogenic pathway relying on the obligate dependence of the strains on an external H_2 source to reduce methyl-compounds into methane. The restriction to this metabolism was previously only observed in two methanogens from digestive tract (*Methanosphaera stadtmanae* and *Methanomicrococcus blatticola*) and considered an exception [11]. The apparently large distribution of this obligate metabolism among this novel order of methanogens turns this exception into one of the important pathways among the overall methanogens. It also highlights the need for a more cautious utilisation of the term of "hydrogenotrophic methanogens" which is generally used to refer to methanogens growing on $H_2 + CO_2$, but also fits for an increasing number of described methanogens growing on H_2 + methyl-compounds. Two names were proposed for this order, Methanoplasmatales [9] and Methanomassiliicoccales [10], the latter being now validated by the International Committee on Systematics of Prokaryotes [12]. For this reason, the name of Methanomassiliicoccales will be used in the current publication to refer to this novel order of methanogens.

The global contribution of Methanomassiliicoccales representatives to methane emission could be large, considering that it constitutes one of the three dominant archaeal lineages in the rumen [3] and in some ruminants it represents half or more of the methanogens [13-15]. Using *mcrA* and 16S rRNA sequences, several studies have also highlighted the broad environmental distribution of this order, not limited to digestive tracts of animals but also retrieved in rice paddy fields, natural wetlands, subseaflor and freshwater sediments for example [9,10,16,17]. Methanomassiliicoccales were split into three large clusters, the "*Ca.* M. alvus" cluster, grouping sequences mostly retrieved from digestive tract of animals, the *M. luminyensis* cluster, mainly composed of sequences from soils and sediments and to a lesser extent from digestive tracts, and the Lake Pavin cluster formed by sequences retrieved from diverse environments but not digestive tracts [16].

The genome sequences of three different Methanomassiliicoccales members cultured from human stool samples, *M. luminyensis* B10 [18], "*Ca.* M. intestinalis Mx1-Issoire" [8] and "*Ca.* M. alvus Mx1201" [7], have recently been made available [19]. *M. luminyensis* shows 98% identity with "*Ca.* M. intestinalis" over the whole 16S rRNA gene and only 87% with "*Ca.* M. alvus". According to the environmental origin of the sequences constituting the large cluster to which they belong, *M. luminyensis* and "*Ca.* M. intestinalis" might be more recently adapted to gut condition than "*Ca.* M. alvus". Moreover the important difference in genome size and [G + C] % content between the two *Methanomassiliicoccus* spp. genomes suggests a rapid evolution of one of them in response to its adaptation from soil or sediment to digestive tract conditions [8]. Despite the important phylogenetic distance between "*Ca.* M. alvus" and the *Methanomassiliicoccus* spp., these genomes uncover common unique genomic characteristics. In particular, the analysis of "*Ca.* M. alvus" and *M. luminyensis* methanogenic pathways revealed they lack the 6 step C_1-pathway forming methyl-CoM by the reduction of CO_2 with H_2, otherwise present

in all previously sequenced methanogens, fitting with their restriction to H_2-dependent methylotrophic methanogenesis [16]. Moreover, these analyses helped define putative alternative substrates to methanol by identification of genes involved in the use of methylated-amines and dimethyl-sulfide. Methylated-amines utilization by Methanomassiliicoccales representatives has also been proposed in a metatranscriptomic study on rumen methanogens [17]. The use of tri-, di- and monomethylamine, with the obligate dependence on H_2, has subsequently been validated *in vivo* with *M. luminyensis* [20]. This property could be significant for human health since gut-produced TMA could be implied in two different diseases [19-22]. The presence of pyrrolysine (Pyl, O), the 22nd proteinogenic amino acid, is associated to this metabolism as it is incorporated in methyltransferases involved in utilization of methylated-amines through an *amber* codon suppression by a Pyl-tRNA [23,24]. All the necessary genetic machinery is found in the three genomes of the Methanomassiliicoccales, including the genes for pyrrolysine synthesis (pylBCD), the *amber* suppressor tRNA^{Pyl} (pylT) and the dedicated amino-acyl tRNA synthetase (pylS). Their structure and unusual features, together with the evolutionary implications of this system have been recently described elsewhere [25].

These original metabolic and genetic characteristics, as well as the closer phylogenetic proximity of this order with Thermoplasmatales than other orders of methanogens prompted us to perform a more comprehensive analysis of these three genomes. We provide here their general characteristics, including comparisons to phylogenetic neighbor genomes, and derived potential metabolism and adaptation to environmental conditions from their gene composition. In the particular context of the missing genes of the CO_2 reduction-pathway otherwise shared by all other methanogens, we reevaluate the global core of enzymes that are unique and specific to all methanogens and highlight the atypical composition of genes likely involved in energy conservation. The potential usage of the *amber* codon as a translational stop signal or encoding a Pyl in proteins was analyzed and suggests a differential handling of the Pyl-encoding capacity among the three Methanomassiliicoccales representatives.

Results and discussion

General genomic features

Genome size, [G + C] %, CDS and tRNA numbers were separately reported in the announcement of these genomes [7,8,18]. Data are gathered in the Table 1 with other newly defined general features.

The tRNA gene complement present in the genomes is in part redundant and covers the usual 20 amino acids, with the exception of Lys in *M. luminyensis*, for which no tRNA was detected: this amino acid is likely encoded in the remnant ~17 kbp from this genome which are currently not available (Table 1). An archaeal complete set of amino-acyl tRNA synthetases is found in all three genomes, Asn- and Gln- tRNAs being obtained by an Asp-/Glu- tRNA (Asn/Gln) amidotransferase [26]. As previously described [25], an important feature is the presence of a tRNA^{Pyl} in all the three genomes. Several small non-coding RNAs (ncRNAs, complete list in Additional file 1: Table S1) were detected. Among them are found a Group II catalytic intron (only in "*Ca.* M. alvus"), the RNA component of the archaeal signal recognition particle (aSRP RNA) and the archaeal RNAse P.

Strikingly, 16S and 23S rRNA genes are not clustered and do not form a transcriptional unit as found in most bacterial and archaeal genomes. Among archaea, this unusual characteristic was first documented in Thermoplasmatales [27], but is also found in related lineages such as the uncultured Marine Group II (MG-II) and *Aciduliprofundum boonei* (Figure 1). This particular organization of the rRNA genes is consistent with the phylogenetic position of the seventh order of methanogens determined using a concatenation of ribosomal proteins [16] and constitutes a distinctive characteristic of Thermoplasmatales and related lineages. On a practical point of view, this also indicates that the Ribosomal Intergenic Transcribed Spacer Analysis, recently proposed as a tool to study the diversity of the methanogenic archaea in digesters [28], will likely fail to detect the Methanomassiliicoccales representatives.

As previously reported [8], the three genomes show significant size heterogeneity, with a variation of 58% (from around 1.7 Mbp to 2.6 Mbp, Table 1). Such heterogeneity is found even within the same genus with 36% size variation between the genomes of "*Ca.* M. intestinalis" and *M. luminyensis* (1.9 to 2.6 Mbp). The number of genes is highly variable and ranges from 1,705 ("*Ca.* M. alvus") to 2,713 (*M. luminyensis*). The average CDS size and gene density is very close among the three genomes (around 900 bp and a protein coding gene every 984 to 1,054 bp). The main translation initiation codon is methionine (AUG) for which two copies of the corresponding tRNA are detected in "*Ca.* M. alvus" and three copies in "*Ca.* M. intestinalis" and *M. luminyensis*. In a lower extent, GUG and UUG are also found as translation start codons (Additional file 1: Table S2). Nucleotide composition [G + C] % ranges from 41.3% to 60.5% (Table 1) [7,8,18]. Codon usage patterns among CDS primarily reflect this [G + C] % variation, "*Ca.* M. intestinalis" primarily using AT-rich codons for a given amino acid (Additional file 1: Table S2). Two of the three stop codons follow this usage pattern, the *ochre* codon UAA accounts for 45% of the stop codons in the genome of "*Ca.* M. intestinalis" and respectively only 17% and 14% in the genomes of "Ca. M. alvus" and *M. luminyensis* and a

Table 1 Genome statistics

Feature	"*Ca.* M. alvus"	"*Ca.* M. intestinalis"	*M. luminyensis*
Genome size[a]	1,666,795	1,931,651	2,637,810[d]
			(2,620,233)
DNA G + C content	55.6%	41.3%	60.5%
% DNA coding region	89.5%	88.4%	87.6%
Intergenic regions mean size (SD)[a]	102 (175)	119 (264)	121 (238)
Genes mean G + C content	56.3%	42.4%	61.0%
Putative replicons	1(+1)[b]	1(+1)[b]	1 (+1)[b]
Extrachromosomal elements	NA[c]	NA[c]	NA[c]
Total genes	1,705	1,882	2,713
RNA genes	52	50	52
rRNA genes (5S-16S-23S)	4 (2 - 1 - 1)	4 (2 - 1 - 1)	4 (2 - 1 - 1)
tRNA genes	48	46	48
Protein coding genes	1,653	1,832	2,661
Mean size of protein coding genes (SD)[a]	901 (667)	930 (890)	859 (676)
Median size of protein coding genes[a]	771	780	732
Gene products with function prediction	1,335	1,476	2,002
Gene products assigned to arCOGs	1,271	1,438	2,065
Gene products assigned Pfam domains	123	125	204
Gene products with signal peptides	247	336	512
Gene products with transmembrane helices	281	389	585
CRISPR repeats	1[e]	1	1

[a]Sizes are given in bp.
[b]Presence of two different *cdc6* genes per genome. See the text for more information.
[c]Not available.
[d]Data from [8]: in bracket stands the total bp (26 contigs) available from database [GenBank: CAJE01000001 to CAJE01000026], analyzed in this study.
[e]Presence of CRISPR repeats split into two neighboring loci (see Additional file 1: Table S3) surrounding a DNA sequence containing one gene encoding a putative transposase.

same trend is observed for *opal* codon UGA (Additional file 1: Table S2). However, a different pattern is observed for the *amber* codon UAG and could be the result of a different selection process (see dedicated section on *amber* codon usage and putative Pyl-containing proteins). All the ribosomal RNA genes of the three genomes have a [G + C] % above 50%. In "*Ca.* M. intestinalis", they thus have a largely higher [G + C] % than the genome average. When compared to *M. luminyensis*, general characteristics of the "*Ca.* M. intestinalis" genome suggest streamlining accompanied by a sharp [G + C] % reduction as previously observed in free-living *Prochlorococcus* [29]. This potential genomic evolution could be related to the recent colonization of digestive tract by "*Ca.* M. intestinalis" from soil or sediment environments.

CRISPR elements

The CRISPR system confers to prokaryotes a highly adaptive and heritable resistance to foreign genetic elements such as plasmids and phages [30-33]. CRISPR loci are composed of genome-specific conserved Direct Repeats (DRs) separated by small sequences (spacers) which constitute a record of past infections. CRISPR-associated (Cas) proteins are responsible for integration of new spacers borrowed from invasive DNA and use the small antisense RNA transcript of these spacers to protect the cell from new invasions. CRISPR loci were previously notified in the three Methanomassiliicoccales genomes [7,8,18] and are characterized in the present study. The CRISPR DRs are concentrated in one genomic unit in "*Ca.* M. intestinalis" and *M. luminyensis* but are interrupted by a gene encoding a putative IS4-type transposase (AGI85628.1) in "*Ca.* M. alvus". The DRs of the three genomes differ from each other in length (31 and 36 bp, Additional file 1: Table S3), sequence and associated 2D-structure (Additional file 2: Figure S1), and belong to three different superclasses. A CRISPR map analysis [34] attributed the *M. luminyensis* DRs to the superclass D, family 3 and the "*Ca.* M. intestinalis" DRs to the superclass A (no family) with a partial motif #27 which is exclusively shared with Methanococcales sequences (from *Methanothermus okinawensis*, *Methanocaldococcus jannaschii* and *Methanocaldococcus fervens*, Additional file 2: Figure S1). The "*Ca.* M. alvus" DRs (ATCTACACTAGTAGAAATTCTGAATGAGTTTT

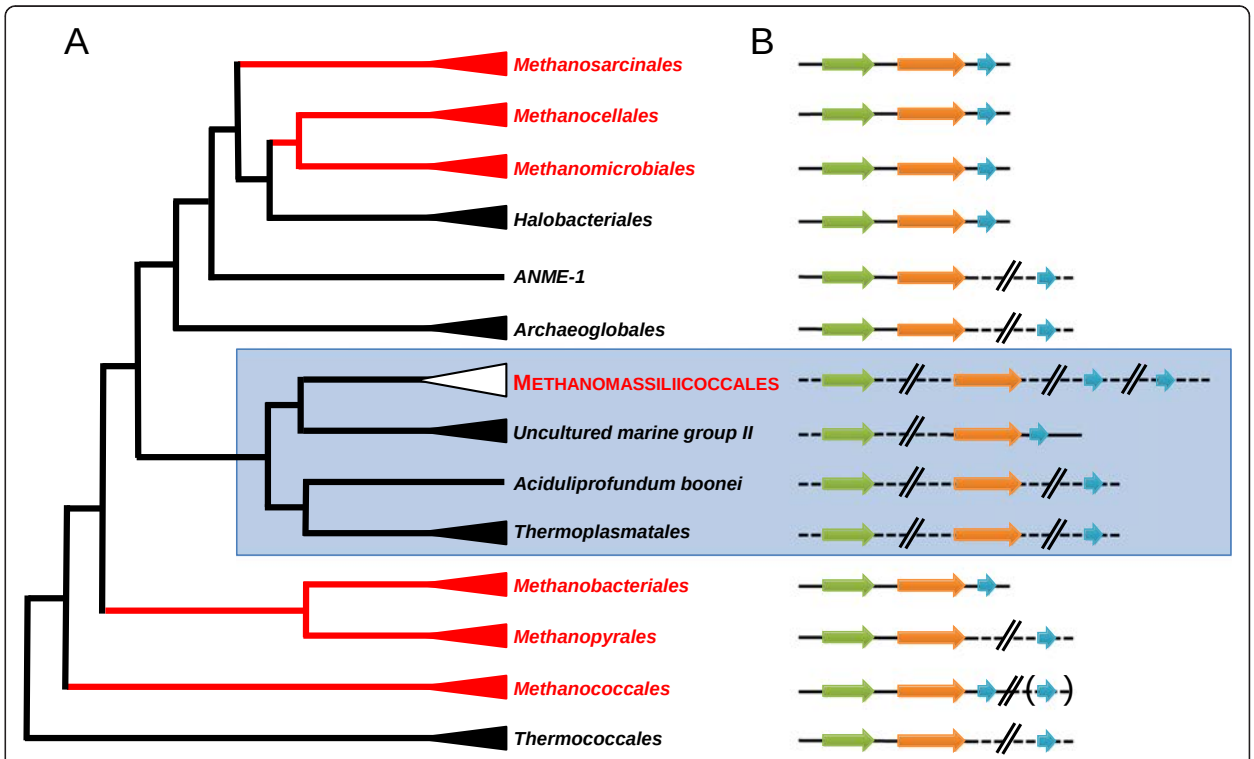


Figure 1 Genomic features of ribosomal genes in Euryarchaeota. (A) Phylogeny of Euryarchaeota highlighting the position of the Methanomassiliicoccales (according to [16]). The seven orders of methanogens are in red. **(B)** Genomic organization of ribosomal genes in Euryarchaeota: 5S, 16S and 23S rRNA genes are symbolized by blue, green and orange arrows, respectively. They are indicated irrespectively of the (+) or (-) DNA strand carrying them. A plain line defines an operon organization where tRNAs (when present) are not shown, nor the number of genes encoding rRNA with the exception of the Methanomassiliicoccales. The 5S rRNA gene in bracket refers to a second 5S rRNA copy isolated from the 16S-23S-5S rRNA gene operon in *Methanococcus maripaludis* C5.

AGAC, superclass E) could not be classified in any sequence/structure family and likely represents a new family of CRISPR DR elements. The number of spacers within DRs ranges from 12 to 113 per locus (from 59 to 113 per genome). Each spacer has a particular size range, from 25 to 28 bp in "*Ca.* M. alvus" to 35 to 40 bp in *M. luminyensis* (Additional file 1: Table S3). A few other CRISPR-like elements are also found in as many as three copies and their functional role remains unknown (Additional file 1: Table S3).

According to the CRISPR system classification proposed by Makarova *et al.* [35] on the basis of organization and composition of the Cas protein-coding genes found in the neighborhood of the CRISPRs, *M. luminyensis* presents a CRISPR-Cas system subtype I-C (WP_019177384.1 to WP_019177390.1). The CRISPR-Cas system of "*Ca.* M. intestinalis" is a hybrid of the subtypes I-A and I-B since its organization corresponds to subtype I-B, but contains the signature gene of the subtype I-A (Cas8a) (AGN26276 to AGY50180.1). The recently defined PreFran subtype (for Prevotella and Francisella) is present in "*Ca.* M. alvus" (AGI85629 to AGI85632). Notably, the Cas1 protein of "*Ca.* M. alvus" is predicted to contain a pyrrolysine (see section on *amber* codon usage and putative Pyl-containing proteins).

As suggested by the different superclass assignments of the repeats and the different types of CRISPR-Cas system, these CRISPRs likely result from non-vertical inheritance among the three species. The PreFran type, only found in 20 bacterial genomes so far is rather uncommon in comparison to the type I of the *Methanomassiliicoccus* spp. Bacteria that hold the PreFan type are generally found in tight association with animals and the genus *Prevotella* is one of the dominant in rumen [36] and human gut [37] suggesting that "*Ca.* M. alvus" may have acquired this system through other gut bacteria. Moreover, the spacers are specific to each of the three genomes suggesting they undergone different histories of infection. In "*Ca.* M. alvus", one of the spacers is 93% similar (25 of 27 nt) to a ssDNA virus isolated from pig feces (JX305998.1).

With the exception of viruses from the families of *Myoviridae* and *Siphoviridae* (head-tail viruses) which also infect bacteria, archaeal viruses sequenced to date have almost no significant residue identity with each other and sequences in public databases [38,39]. Accordingly,

the lack of detection of prophage sequences by dedicated software does not imply the absence of prophages in these three genomes: some clusters of 10-30 adjacent genes with few significant matches in public databases might represent still unknown prophages. Furthermore, genes distantly related to phage ones are found in the three genomes and could belong to unknown prophages or represent residual traces of past infection. This is for example the case of two contiguous genes, present in the vicinity of the "*Ca.* M. intestinalis" CRISPR locus, which encode putative proteins (YP_008071639.1 & YP_008071640.1) with similarity to phage capsid synthesis proteins.

Genome replication

Origins of replication were identified with a consensus Origin Recognition Box (ORB) motif recently identified from active replication origins of Thaumarchaeota (*Nitrosopumilus maritimus*), Crenarchaeota and Euryarchaeota [40]. Several ORB motifs were found in the three genomes, most of them gathered by pairs (Table 2). A consensus sequence for a Methanomassiliicoccales ORB motif was deduced and shows little difference with the archaeal consensus recently proposed [40] (Table 2).

Each of the three genomes possesses two copies of the *orc1/cdc6* (Origin Recognition complex/Cell division cycle 6) gene (Table 3). At least two ORB motifs are found in the vicinity of only one of the two *orc1/cdc6* genes. In the draft genome of *M. luminyensis*, these two genes are associated in the same contig (CAJE01000021), allowing comparison with the other two genomes. In every case, the *orc1/cdc6* genes are each located on a different strand (Additional file 2: Figure S2). They are close together within the *M. luminyensis* and "*Ca.* M. intestinalis" genomes (respectively around 70 and 90 kbp), and more distant in "*Ca.* M. alvus" (around 695 kbp). They are inversely oriented in the three genomes. Consistent with a recent study [41], phylogenetic analysis reveals that these genes correspond to two paralogs, *orc1/cdc6.1* and *orc1/cdc6.2* (Additional file 2: Figure S3). *orc1/cdc6.1* lies close to the predicted origin of replication, displays a conserved genomic context (Figure 2) is slow-evolving and groups phylogenetically with Thermoplasmatales/DHVE2/uncultured Marine Group II (Additional file 2: Figure S3), consistent with vertical inheritance. On the other hand, *orc1/cdc6.2* copies display much faster evolutionary rates, lies in a non-conserved genomic context (Figure 2), and show inconsistent phylogenetic placement close to Crenarchaeota (Additional file 2: Figure S3). This may be due to a tree reconstruction artifact or may represent a possible horizontal gene transfer from an unspecified crenarchaeon. Given its higher conservation, its conserved genomic context and its vicinity to ORB motifs, Orc1/Cdc6.1 is likely the main initiator protein and Orc1/Cdc6.2 may represent an inactive or accessory copy, possibly active in different environmental conditions.

The replication gene set is similar to that of the most closely related lineages (Table 3). However, some interesting features are present in the three genomes. For example, they do not harbor any homologs of the single-stranded binding protein SSB similarly to MG-II, whereas Thermoplasmatales and DHVE2 have both RPA and SSB. The absence of SSB may strengthen the sister relationship of the Methanomassiliicoccales and MG-II lineages as observed in a phylogenetic reconstruction based on ribosomal proteins [16]. The Methanomassiliicoccales,

Table 2 ORBs motifs found in the Methanomassiliicoccales genomes

ORB	Sequence	Position	Spacing	Orientation	Comment
"*Ca.* M. alvus" ORB1	**GTTCCAGTGGAAATGG**-T**GGGGT**	78 - 99	39	inverted	downstream *orc1/cdc6.1*
"*Ca.* M. alvus" ORB2	**GTTCCA**C**TGGAAA**CA**G-AGGGGT**	138 - 159		inverted	downstream *orc1/cdc6.1*
"*Ca.* M. alvus" ORB3	T**TTCCA**C**TGGAAA**CA**G-AGGGGT**	1977 - 1998	47		upstream *orc1/cdc6.1*
"*Ca.* M. alvus" ORB4	**GTTCCA**C**TGGAAATGG**-T**GGGGT**	2045 - 2066			upstream *orc1/cdc6.1*
"*Ca.* M. intestinalis" ORB1	A**TT**A**CAGTGGAAATGA-AGGGGT**	15 - 36	256	inverted	downstream *orc1/cdc6.1*
"*Ca.* M. intestinalis" ORB2	T**TT**G**CAGTGGAAATGA-AGGGGT**	292 - 313			downstream *orc1/cdc6.1*
"*Ca.* M. intestinalis" ORB3[a]	**GTT**C**CAGTGGAAATGA-AGGGGT**	795626 - 795647			downstream *fstZ*
"*Ca.* M. intestinalis" ORB4[a]	TC**T**G**CA**C**TGGAAATGA-AGGGGT**	1576211 -1576232		inverted	downstream fused *nifH/nifE*
M. luminyensis ORB1	**GTT**C**CA**T**TGGAAAT**C**G**-GCA**GG**A	73488 - 73475[b]	113		downstream *orc1/cdc6.1*
M. luminyensis ORB2	**GTT**C**CAGTGGAAAT**A**A-AGGGGT**	73341 - 73362[b]		inverted	downstream *orc1/cdc6.1*
Methanomassiliicoccales consensus ORB	**GTTCCAGTGGAAATGG-AGGGGT A**				
Archaea consensus ORB	C**TTCCAGTGGAAA**C**G**AA**AGGGGT**				Pelve *et al.*, [40]

Bases in bold indicate consensual bases of the ORB sequence in the Methanomassiliicoccales. The "*Ca.* M. alvus" ORBs, and the ORB2 of *M. luminyensis* and "*Ca.* M. intestinalis" might be extended by a "GGGGGT" sequence otherwise not conserved in the 4 other Methanomassiliioccales ORBs and the Archaea consensus ORB.
[a]Not found in close association to another ORB.
[b]Contig [GenBank: CAJE01000021.1].

Table 3 DNA replication proteins compared to the corresponding components in Thermoplasmatales, MG-II and DHEV2

	"*Ca.* M. alvus"	"*Ca.* M. intestinalis"	*M. luminyensis*	MG-II	DHEV2	Thermoplasmatales
ATP-dependent DNA ligase	AGI85913	AGN25909	WP_019176428	X	■	■
Orc1/Cdc6	AGI84758 (1)	AGN25419 (1)	WP_019178385 (1)	■■	■	■■
	AGI85775 (2)	AGN27158 (2)	WP_019178317 (2)			
DNA Pol D large subunit (DPL)	AGI85099	AGN26720	WP_019177373	■	■	■
DNA Pol D small subunit (DPS)	AGI84772	AGN27082	WP_019178373	■	■	■
FEN-1	AGI85207	AGN26626	WP_019176843	■	■	■
GINS 51	AGI84890	AGN27100	X	X	■	■
GINS 23	X	X	X	X	X	X
DNA Gyrase subunit B	[AGI86382]	[AGY50228]	[WP_019178436]	[■]	[■]	[■]
DNA Gyrase subunit A	[AGI86381]	[AGN27159]	[WP_019178437]	[■]	[■]	[■]
MCM	AGI86392	AGN26346	WP_019178416	■	■	■
		AGN27203				
PCNA	AGI84935	AGN27068	WP_019176118	■	■	■
DNA Pol B	AGI86264	AGN26701	WP_019177962	■	■	■
			WP_019177491			
Primase large subunit (PriL)	AGI84820	AGN27177	WP_019178297	■	■	■
Primase small subunit (PriS)	AGI86400	AGY50234	WP_019178400	■	■	■
RFC large subunit	AGI85559	AGN26596	WP_019176873	■	■	■
RFC small subunit	AGI85778	AGN26166	WP_019177244	■	■	■
RNaseH II	AGI86158	AGN25790	WP_019177553	■	■	■
TopoVI subunit A	AGI85998	AGN26743	WP_019177592	■	■	X
TopoVI subunit B	AGI85997	AGN26742	WP_019177591	■	■	X
Topo IB	X	X	X	X	X	X
SSB	X	X	X	X	■	■
RPA2	AGI84916	AGN25568	WP_019178149	■	■	■
		AGY50184	*WP_019177069*			
rpa2A (rp associated protein)	AGI84915	AGN25567	WP_019178150	■	■	■
NAD-dependent DNA ligase	[AGI85455]	X	X	[■]	X	X

Proteins in brackets indicate horizontal transfers from bacteria; Proteins in italics indicate fast evolving additional copies likely representing decaying paralogs, genes horizontally transferred among archaea, or homologs arising from integration of foreign elements. Absent proteins (or unavailable due to genome incompleteness) are indicated by an X. (1) and (2) in front of the Orc1/Cdc6 protein accession numbers indicate the Orc1/Cdc6.1 and Orc1/Cdc6.2, respectively.

Marine Group II, and DHVE2 harbor both subunits of the archaeal topoisomerase TopoVI, strengthening a specific loss of this gene in Thermoplasmatales, which replaced it by a bacterial-type DNA gyrase [42]. Moreover, all three Methanomassiliicoccales representatives also harbor a bacterial-like DNA gyrase, known to have been acquired from bacteria in late emerging Euryarchaeota [41]. Some components are present as extra copy in the three genomes (in bold in Table 3), for example the Minichromosome Maintenance Protein (MCM) in the genome of "*Ca.* M. intestinalis", which is highly divergent with respect to the other MCM coding genes and lies in a genome region with no synteny with the other closely related genomes. This is also the case for an extra PolB coding gene identified in the genome of *M. luminyensis*. Finally, genes coding for two additional OB-fold containing proteins (RPA-like) were identified in the genomes of "*Ca.* M. intestinalis" and *M. luminyensis*. All these extra copies are very divergent and likely represent decaying paralogs or homologs arising from integration of foreign elements. In addition, we found a bacterial type NAD-dependent DNA ligase homolog in the genome of "*Ca.* M. alvus" that appears to originate via a specific and recent horizontal gene transfer from a bacterium of the *Prevotella* genus, which is abundant in the human gut microbiota (Additional file 2: Figure S4A).

An important feature shared by the three Methanomassiliicoccales representatives, the Thermoplasmatales and other related lineages is the lack of Eukaryotic-like histone found in other Euryarchaeota [43], suggesting that the loss of this gene occurred early in the evolution of the whole lineage. Surprisingly, no gene coding for homologues of the bacterial-type HTa histones known to

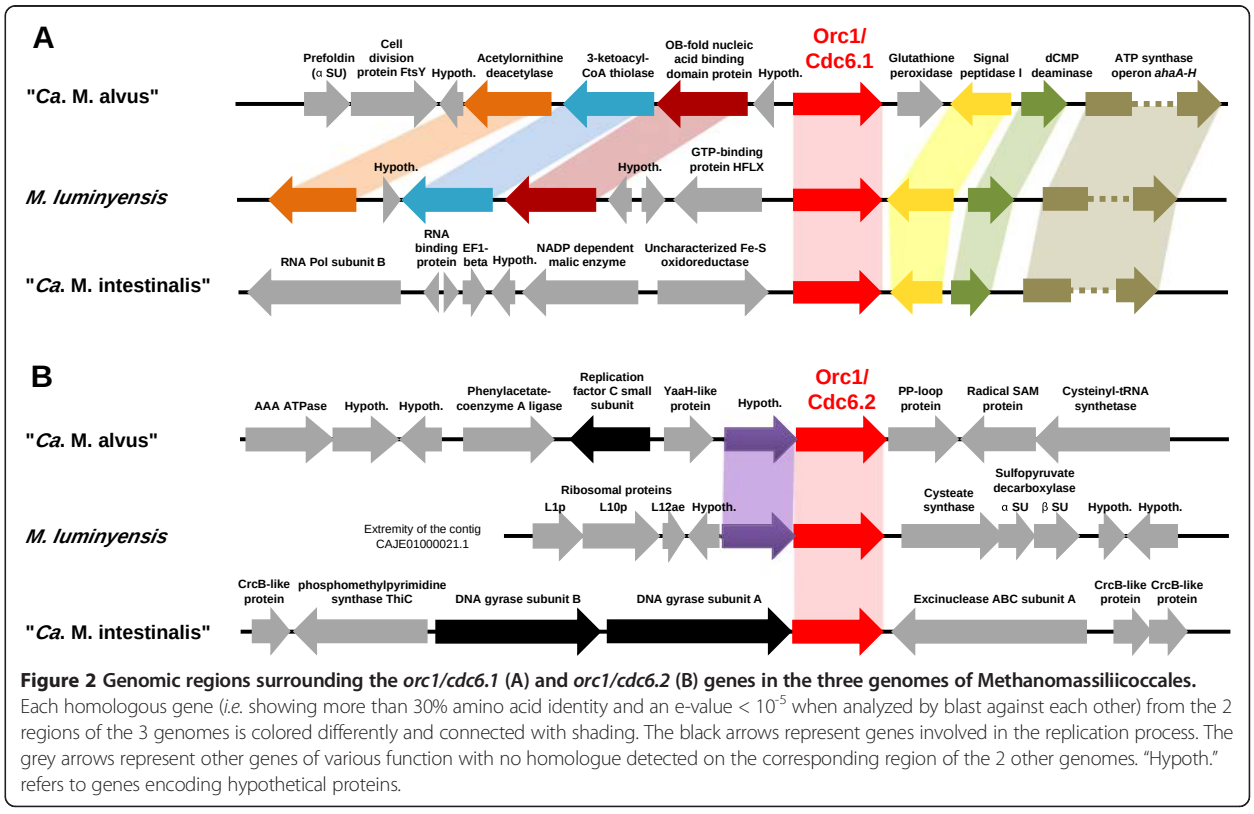


Figure 2 Genomic regions surrounding the *orc1/cdc6.1* (A) and *orc1/cdc6.2* (B) genes in the three genomes of Methanomassiliicoccales. Each homologous gene (*i.e.* showing more than 30% amino acid identity and an e-value < 10^{-5} when analyzed by blast against each other) from the 2 regions of the 3 genomes is colored differently and connected with shading. The black arrows represent genes involved in the replication process. The grey arrows represent other genes of various function with no homologue detected on the corresponding region of the 2 other genomes. "Hypoth." refers to genes encoding hypothetical proteins.

have replaced the native histone in Thermoplasmatales and DHVE2 are present in the Methanomassiliicoccales genomes and the MG-II genome. The DNA packaging function could be fulfilled in *M. luminyensis* by an Alba protein (WP_019176109.1) also presents in *Thermoplasmatales* [44] and MG-II, but absent in "*Ca.* M. alvus" and "*Ca.* M. intestinalis". Few candidate proteins with a very weak similarity to bacterial histones and a Lys- and Arg-rich tail were identified in *M. luminyensis* (WP_019177894.1) and "*Ca.* M. intestinalis" (AGN26805.1) but not in "*Ca.* M. alvus". While the proteins responsible for this crucial function remain elusive, a homologue of the histone acetyltransferase of the ELP3 family was identified in the three genomes (WP_019178580.1, AGN27049 and AGI86364). Only *M. luminyensis* possesses a histone deacetylase HdaI, related to Crenarchaeota and not found in other Thermoplasmatales (WP_019177579.1).

Core genome

The best BLAST hits of the CDS from the three genomes were most frequently found in other archaeal members (70% to 82%), around 18% to 30% to Bacteria, and less than 0.3% to Eukaryota (Additional file 1: Table S4). It is likely that some of these reflect lateral gene transfer events, consistent with the presence of genomic islands with different [G + C] % composition from the genome average, as observed in "*Ca.* M. alvus" and, more pronounced, in "*Ca.* M. intestinalis" (Additional file 2: Figure S2).

The core genome of the three species is composed of 658 CDS. While the number of CDS shared between genome pairs reflects partly their phylogenetic relatedness, an impressive proportion of CDS are specific to each one, in particular for *M. luminyensis* (Figure 3, Additional file 1: Table S5 for a complete list). Of the core genome, 173 genes are not found in the closest lineages (*Ferroplasma acidarmanus*, *Thermoplasma acidophilum*, *Thermoplasma volcanium*, uncultured Marine Group II and *Aciduliprofundum boonei* (Table 4, complete data in Additional file 1: Table S5). A part of these genes could correspond to specific traits of the Methanomassiliicoccales, at least for 20 of them which have no close homologue sequence in the databases (Additional file 1: Table S5). Another part of these genes reflects the metabolic pathway of the Methanomassiliicoccales representatives, methanogenesis, not shared with the Thermoplasmatales and any of the other related lineages for which genomic or physiological data are available. As discussed below, some of these genes are unique to methanogens. Among the predicted core proteins, 227 have no homologues in the two other methanogens commonly found in the same environment, the human gut (*Methanobrevibacter smithii* and *Methanosphaera stadtmanae*). Some of these differences rely on the particular methanogenic pathway of the Methanomas-

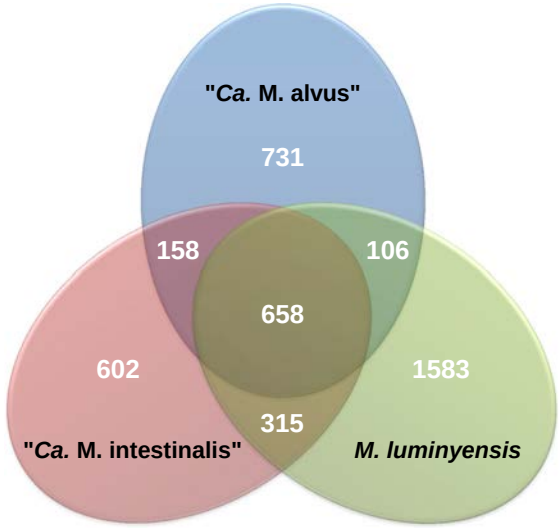


Figure 3 Shared and unique CDS among the three genomes. Venn diagram indicating the core genome at its center, deduced from a BLAST analysis of the CDS from the 3 genomes of the Methanomassiliicoccales. Unique and shared CDS among genome pairs are also given.

siliicoccales which can use methylated amines as substrate [20], which is not the case of *M. smithii* and *M. stadtmanae*. One hundred and two core proteins have no homologues in either the closely related lineages or the two gut methanogens (Table 4, complete data in Additional file 1: Table S6). Some show hits to other methanogens (Methanocellales, Methanomicrobiales and Methanosarcinales), and are specific for methanogenesis/energy conservation. Others likely reflect ancient lateral gene transfer events (LGTs) in the ancestor of the Methanomassiliicoccales. They include proteins involved in carbohydrate metabolism (glycosyl transferases, sugar transporters), nitrogen metabolism, and several proteins specific to the Methanosarcinaceae and involved in methanogenesis (see below).

General metabolism and adaptations to environment

Analysis of archaeal clusters of orthologous groups (ArCOG [45]) resulted in 1,271; 1,438 and 2,065 assigned functions for "*Ca.* M. alvus", "*Ca.* M. intestinalis" and *M. luminyensis* respectively (representing between 77-79% of all CDS) (Additional file 1: Table S7). Components of cell wall/membrane and envelope biogenesis (class M) were less abundant when compared to the other gut methanogens *M. smithii* and *M. stadtmanae*. Indeed, comparatively to these Methanobacteriales, electron micrographies of *M. luminyensis* did not show a prominent cell-wall-like structure [6]. However, it seems that the synthesis of activated mannose is likely possible from fructose-6-P, therefore allowing the biosynthesis of N-glycans potentially associated to a cell-wall. A specific enrichment was observed for inorganic ion transport and metabolism (class P) and, as noted for other methanogens, for coenzyme transport and metabolism (class H): when analyzed in more details, many of the predicted transporters are ABC transporter permease proteins with homology to those identified in other methanogens (Additional file 1: Table S8). Noteworthy is the presence of quaternary ammonium compound efflux pumps as well as specialized systems involved in substrate acquisition for specialized methanogenesis-related functions (H_2-dependent methylotrophic methanogenesis, see below): this includes putative transporters for dimethylamine (AGI85872.1/AGI85374.1/AGI85246.1 for "*Ca.* M. alvus", AGN26255.1 for "*Ca.* M. intestinalis", WP_019178528.1 for *M. luminyensis*) and trimethylamine (AGI85867.1,

Table 4 Proteome of the three Methanomassiliicoccales representatives compared to their phylogenetic neighbors, human gut methanogens and NCBI nr proteins

Core genome of Methanomassiliicoccales: 658 protein sequences	Specific[a]	Shared[b]
Phylogenetic neighbors	173	485
		125 absent from human gut methanogens
Human gut methanogens	227	431
		63 absent from phylogenetic neighbors
Phylogenetic neighbors and human gut methanogens	102	556 shared with at least one
		Encompassing:
		125 absent from human gut methanogens
		71 absent from phylogenetic neighbors
		360 shared with the two groups
NCBI non-redundant protein sequences database	20 (21)[c]	637

[a]Number of deduced proteins of the core genome of Methanomassiliicoccales that are not found in the corresponding organisms.
[b]Number of deduced proteins of the core genome of Methanomassiliicoccales that are also found in the corresponding organisms.
[c]The value of 21 encompasses CDS that are specific of the proteome of the Methanomassiliicoccales together with either the ones of the phylogenetic neighbors or of the human gut methanogens, without any other blast hits with the NCBI nr protein sequences database.

AGN26256.1, WP_019178522.1). The following part of the section focuses on several genomic features of the three Methanomassiliicoccales representatives that suggest metabolic adaptations to their environment. An overview of the inferred general metabolism is given in Additional file 2: Figure S5. As usually observed in methanogens, the three species harbors an incomplete reductive TCA cycle [46]. Further details on lipid, amino-acid and purine synthesis pathways, as well as molecular nitrogen fixation are also presented in Additional file 3.

Similarly to other methanogens and differently from the Thermoplasmatales representatives, the three Methanomassiliicoccales lack PurK for purine synthesis pathway. Two purE-like enzymes were identified (AGI84793.1, AGI85002.1, AGN25661.1, AGN26431.1, WP_019178351.1, WP_019177087.1) without clear assignment to class I or class II PurE (Additional file 3). Depending on the assignment of these PurE, the ATP-dependent activity of PurK might be substituted by a class I PurE in presence of high concentration of CO_2 or a class II PurE, both avoiding the hydrolysis of ATP [47]. The former possibility could represent an adaptation to the high CO_2 concentrations in anaerobic environments as proposed for other methanogens [47].

Two possible sources of ammonia are predicted to be common in the three Methanomassiliicoccales, a direct uptake from the environment by dedicated transporters (Additional file 1: Table S8) and an intracellular production, as a by-product of methanogenesis from monomethylamine. The presence of some of these transporters in close association to the genes involved in methanogenesis from monomethylamine suggests that they could alternatively be used to export ammonium when monomethylamine is used for methanogenesis. Ammonia could also be derived from urea in "*Ca*. M. intestinalis" which possesses a *ureA-G* operon encoding a urease (AGN27148.1 to AGN27154.1) and a urea transporter (AGN27055.1). Ammonia is likely assimilated by a glutamine synthetase GlnN, one in "*Ca*. M. alvus" and "*Ca*. M. intestinalis" (AGI86325.1; AGN25771.1) and two in *M. luminyensis* (WP_019177566.1; WP_019177539.1, this second one likely acquired through LGT from bacteria). *M. luminyensis* is predicted to be diazotroph with a putative flexibility upon the dependency on Molybdenum, while "*Ca.* M. alvus" and "*Ca.* M. intestinalis" probably lack the capacity to fix N_2 (Additional file 3). N_2 fixation capacity has been found among soil and sediment methanogens but not in common gut methanogens (Additional file 3) [48-50]. Accordingly, the potential capacity of *M. luminyensis* to fix N_2 could reflect an adaptation to soil or sediment conditions and a facultative association to digestive tracts.

Each Methanomassiliicoccales genome encodes at least one catalase (*katE*), peroxiredoxin (*prx*), rubredoxin (*rub*) and rubrerythrin (*rbr*) to resist to oxygen exposure (Additional file 1: Table S9). *M. luminyensis* presents the highest antioxidant capacity, in particular with 8 copies of a peroxiredoxins (*prx*) gene, against 4 and 2 copies in "*Ca.* M. intestinalis" and "*Ca.* M. alvus" respectively. *M. luminyensis* is also the only one to harbor homologues of superoxide dismutase (*sodA*) and desulfoferrodoxin (*dfx*). A large diversity and redundancy of the antioxidant systems was previously reported for dominant rice field soil methanogens, Methanocellales, and described as a specific adaptation of these methanogens to oxic episodes regularly occurring in these environments [48,49]. In line with its probable diazotrophic capacity, the larger number and diversity of genes encoding antioxidant enzymes in *M. luminyensis* argue for a greater adaptation to soil environments than "*Ca.* M. alvus" and "*Ca.* M. intestinalis". A glycine-betaine ABC transporter (WP_019176328.1, WP_019176329.1, WP_019176330.1) was also found in *M. luminyensis*. This kind of transporter helps to cope with external variations in salt concentration by accumulating glycine-betaine as an osmoprotectant and was previously identified in Methanosarcinales [51,52]. No similar transporter of glycine-betaine was identified in "*Ca.* M. alvus" or "*Ca.* M. intestinalis".

Interestingly, among the three Methanomassiliicoccales representatives, "*Ca.* M. alvus" is the only one to encode a choloylglycine hydrolase (YP_007713843.1), which confers resistance to bile salts encountered in the gastro-intestinal tracts (GIT). This gene is also present in the genome of the two other dominant human gut methanogens, *M. smithii* and M. *stadtmanae* [53,54], and could have been transferred from other gut bacteria (Additional file 2: Figure S4B). Another adaptation to GIT could be inferred through the presence of a conserved amino acid domain corresponding to COG0790 (TPR repeat, SEL1 subfamily) in at least one protein of each Methanomassiliicoccales representative. This conserved domain has been previously identified in proteins involved in interactions between bacteria and eukaryotes and was never reported in archaea [55] suggesting an adaptation to digestive tracts unique to Methanomassiliicoccales among archaea. In that case, the occurrence of the genes encoding proteins with this domain in the Methanomassiliicoccales genomes, 28 in "*Ca.* M. alvus", 6 in "*Ca.* M. intestinalis" and one in *M. luminyensis* would support a higher adaptation of "*Ca.* M. alvus" to digestive tracts.

Methanogenesis and core enzymes specific to methanogens

It was previously reported that *M. luminyensis* and "*Ca.* M. alvus" lack the genes that encode the 6 step C_1-pathway leading to methyl-CoM by the reduction of CO_2 with H_2 [16]. Our current analysis revealed a similar lack of these genes in "*Ca.* M. intestinalis" (Figure 4). It also reveals that "*Ca.* M. intestinalis" does not harbor the

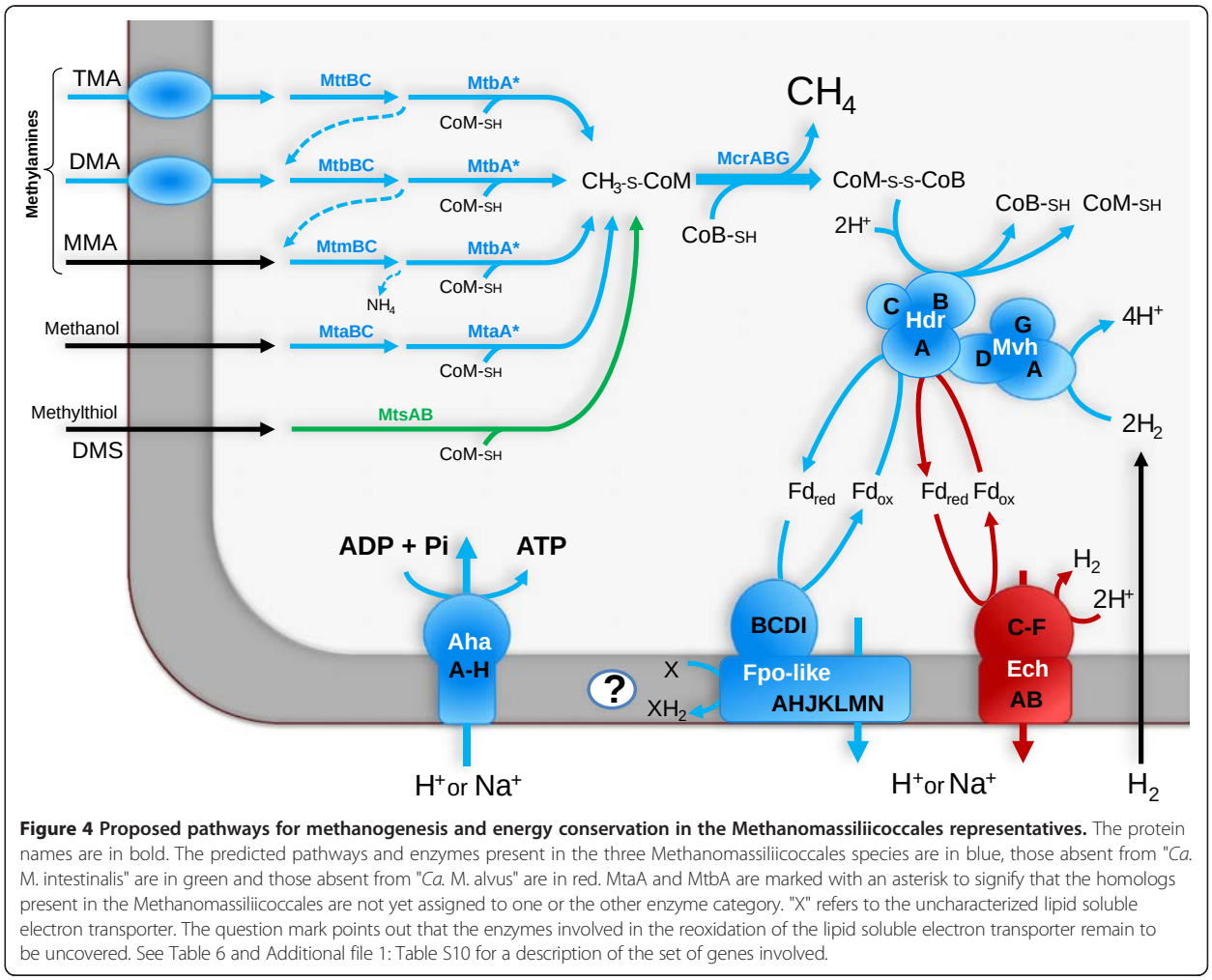


Figure 4 Proposed pathways for methanogenesis and energy conservation in the Methanomassiliicoccales representatives. The protein names are in bold. The predicted pathways and enzymes present in the three Methanomassiliicoccales species are in blue, those absent from "*Ca.* M. intestinalis" are in green and those absent from "*Ca.* M. alvus" are in red. MtaA and MtbA are marked with an asterisk to signify that the homologs present in the Methanomassiliicoccales are not yet assigned to one or the other enzyme category. "X" refers to the uncharacterized lipid soluble electron transporter. The question mark points out that the enzymes involved in the reoxidation of the lipid soluble electron transporter remain to be uncovered. See Table 6 and Additional file 1: Table S10 for a description of the set of genes involved.

genes *mtsAB* (Figure 4) which code for enzymes likely involved in methanogenesis from dimethylsulfide [56]. The composition of the methyltransferases involved in the H_2-dependent methylotrophic methanogenesis from the three genomes was partially determined before [7,8,16,17] and is compiled in the Additional file 1: Table S10, with their relative genomic position displayed in the Additional file 2: Figure S6.

A pool of genes conserved among all methanogens and not found in any other archaea was recently determined by Kaster *et al.* [57]. These genes encode the subunits of two enzymatic complexes unique and shared by all methanogens, the methyl-H4MPT: coenzyme M methyltransferase (Mtr) and the methyl coenzyme reductase (two complexes of isoenzymes Mcr and Mrt), as well as proteins of unknown function. Being unique to methanogens, these uncharacterized proteins likely have an important role for methanogenesis and could be directly associated to the functioning of Mcr and Mtr [57]. The lack of Mtr and the other genes of the CO_2-reductive pathway in the three Methanomassiliicoccales described here, prompted us to reevaluate the overall methanogenesis markers. In addition to the five genes coding for subunits of the Mtr enzymatic complex, two former methanogenesis markers (annotated as methanogenesis markers 10 and 14 in the databases and belonging to arCOG00950 and arCOG04866, respectively) are absent from the three Methanomassiliicoccales genomes (Table 5). One of these genes (belonging to arCOG04866) is present in the vicinity of the operon coding for Mtr in *Methanosaeta thermophila*, Methanobacteriales, Methanopyrales and Methanocellales genomes. Its genomic position in methanogens encoding Mtr and its absence in Methanomassiliicoccales suggests its involvement in the functioning of Mtr. Fifteen genes present in the three Methanomassiliicoccales genomes have homologues (and/or paralogs in the case of *atwA* and the *mcr/mrt* operons) conserved in all other methanogens and not in other archaea and could still be considered as methanogenesis markers (Table 5). Interestingly, 13 of these genes, including the *mcr* operon, are clustered on a small genomic portion (~16 Kb)

Table 5 Core proteins of methanogenesis

Annotation	"*Ca.* M. alvus"	"*Ca.* M. intestinalis"	*M. luminyensis*	Distribution	arCOG
Nitrogenase molybdenum-iron like protein (NifD-like/NflD)	*AGI86050*	*AGN27015*	*WP_019176684.1*	1	arCOG04888
UDP-N-acetylmuramyl pentapeptide synthase like protein (MurF-like)	*AGI86051*	*AGN27016*	*WP_019176685.1*	1	arCOG02822
Methyl-coenzyme M reductase operon associated like protein (McrC-like)	**AGI85157**	***AGN26013***	**WP_019176790.1**	1	arCOG03226
Conserved hypothetical protein	**AGI85156**	***AGN26012***	**WP_019176789.1**	1	arCOG04904
CoA-substrate-specific enzyme activase	**AGI85155**	***AGN26011***	**WP_019176788.1**	1	arCOG02679
Conserved hypothetical protein	**AGI85154**	***AGN26010***	**WP_019176787.1**	1	arCOG04903
Conserved hypothetical protein	**AGI85153**	***AGN26009***	**WP_019176786.1**	1	arCOG04901
Peptidyl-prolyl cis-trans isomerase related protein	**AGI85152**	***AGN26008***	**WP_019176785.1**	1	arCOG04900
Methyl coenzyme M reductase operon associated protein (McrC)	**AGI85151**	***AGN26006***	**WP_019176783.1**	1	arCOG03225
Methyl-coenzyme M reductase, component A2 (AtwA)	**AGI85150**	***AGN26005***	**WP_019176782.1**	1*†	arCOG00185
Methyl coenzyme M reductase, beta subunit (McrB/MrtB)	**AGI85141**	**AGN26874**	**WP_019176771.1**	1*	arCOG04860
Methyl coenzyme M reductase, protein D (McrD/MrtD)	**AGI85142**	**AGN26873**	**WP_019176772.1**	1*	arCOG04859
Methyl coenzyme M reductase, gamma subunit (McrG/MrtG)	**AGI85143**	**AGN26872**	**WP_019176773.1**	1*	arCOG04858
Methyl coenzyme M reductase, alpha subunit (McrA/MrtA)	**AGI85144**	**AGN26871**	**WP_019176774.1**	1*	arCOG04857
SH3 fold protein	**AGI85145**	**AGN26870**	**WP_019176775.1**	1	arCOG04846
Conserved hypothetical protein	**AGI85146**	**AGN26876**	**WP_019176769.1**	2*†	arCOG02882
AIR synthase-like protein	AGI85549	AGN26462	WP_019176932.1	2	arCOG00640
Predicted DNA-binding protein containing a Zn-ribbon domain	AGI84948	AGN25597	WP_019176187.1	2*	arCOG01116
Methyltransferase related protein (MtxX)	AGI85117	AGN26654	WP_019177314.1	3	arCOG00854
Conserved hypothetical protein	AGI84870	AGN25885	WP_019178690.1	3*	arCOG04893
Fe-S oxidoreductase, related to NifB/MoaA family	-	-	-	4*	arCOG00950
Conserved hypothetical protein	-	-	-	4	arCOG04866
N5-methyltetrahydromethanopterin: coenzyme M methyltransferase, subunit A (MtrA)	-	-	-	4*	arCOG03221
N5-methyltetrahydromethanopterin: coenzyme M methyltransferase, subunit B (MtrB)	-	-	-	4	arCOG04867
N5-methyltetrahydromethanopterin: coenzyme M methyltransferase, subunit C (MtrC)	-	-	-	4	arCOG04868
N5-methyltetrahydromethanopterin: coenzyme M methyltransferase, subunit D (MtrD)	-	-	-	4	arCOG04869
N5-methyltetrahydromethanopterin: coenzyme M methyltransferase, subunit E (MtrE)	-	-	-	4	arCOG04870
Soluble P-type ATPase	-	-	-	5	arCOG01579
Uncharacterized conserved protein	-	-	-	5*	arCOG04844
Conserved hypothetical protein (putative kinase)	-	-	-	6	arCOG04885

Protein accession numbers with the same font (bold, italics or bold-italics) are encoded by genes situated close to each other in their respective genomes.
*Paralogues.
†Related to a bacterial cluster with same conserved domain.
1, Methanogenesis marker, present in and unique to all sequenced methanogens and not in other archaea.
2, Present in all sequenced methanogens and less than 5% of other sequenced archaea.
3, Present in more than 90% of sequenced methanogens including Methanomassiliicoccales and less than 5% of other sequenced archaea.
4, Absent from the Methanomassiliicoccales but present and unique to all other methanogens.
5, Absent from the Methanomassiliicoccales but present in more than 90% other methanogens and not in other archaea.
6, Absent from the Methanomassiliicoccales but present in more than 90% of sequenced methanogens and less than 5% of other sequenced archaea.

of *M. luminyensis* and "*Ca.* M. alvus". At the exception of *mcrABG* and *atwA* [58], they encode for proteins of unknown function. One of these proteins (WP_019176775.1, AGN26870, AGI85145), not previously reported as a methanogenesis marker, might be associated to the functioning of Mcr as it is encoded by a gene located directly upstream *mcrA* in the three Methanomassiliicoccales genomes. The nifD-like (NflD) gene previously proposed

to be involved in the biosynthesis of the coenzyme F_{430}, the prosthetic group of Mcr/Mrt, is also present in the three genomes [59]. It forms a cluster with a UDP-N-acetylmuramyl pentapeptide synthase like gene (Table 5) and a *nifH-like* gene also suggested to be involved in coenzyme F_{430} biosynthesis. Several uncharacterized proteins are shared by almost all methanogens, while present in very few other archaea, suggesting a tight relationship with methanogenesis (Table 5). This is for example the case of a putative methyltransferase MtxX [60] only missing in *Methanosaeta concilii* GP6 (but still present as a pseudogene, MCON_2260) among methanogens and only present in *Ferroglobus placidus* DSM-10642 among non-methanogens.

Other genes present in the three genomes are more widely distributed than in methanogens but play a crucial role in methanogenesis. This is the case of genes required for the biosynthesis of the coenzyme M and coenzyme B involved in the last step of methanogenesis. Inferred CoM biosynthesis uses sulfopyruvate, which originates from 3-phosphoserine converted to cysteate by a cysteate synthase and then to sulfopyruvate (ComDE), as observed in Methanosarcinales, Methanomicrobiales [61] and Methanocellales (Additional file 1: Table S11). An alternative pathway takes place in other methanogens, where CoM originates from phosphoenolpyruvate and sulfite to produce sulfolactate, which is then oxidized [62-64]. These steps require the activity of enzymes encoded by the *comABC* genes which are absent in the three genomes, similar to what is observed in Methanosarcinales and Methanomicrobiales (Additional file 1: Table S11).

Energy conservation

Methanogenesis is coupled to energy conservation through the establishment of a proton and/or sodium ion electrochemical gradient across the cytoplasmic membrane that drives an archaeal-type A_1A_0 ATP synthase complex to form ATP [65]. The genes coding for this complex are found in close association with the putative origin of replication in the three genomes (Figure 2, Table 6). The exergonic reduction of the heterodisulfide CoM-S-S-CoB formed by the Mcr complex is a crucial step for energy conservation conserved in all methanogens. The three genomes harbor at least one copy of *hdrA, hdrB* and *hdrC* homologues encoding a soluble heterodisulfide reductase (Table 6), HdrB representing the catalytic activity for CoM-S-S-CoB reduction. The current HdrA differs from its homologues present in other methanogens by its longer size and the presence of two predicted FAD-binding sites instead of one, and three 4Fe-4S centers instead of four. The three genomes also contain homologues of *hdrD,* encoding the catalytic site of a second class of heterodisulfide reductase (HdrDE), but no homologues of *hdrE* encoding the membrane bound cytochrome subunit of this complex. Similarly to the Methanococcales, Methanobacteriales and Methanopyrales, the *hdrB* and *hdrC* genes are adjacent whereas the *hdrA* gene is located apart and in close association with *mvhDGA* encoding the cytoplasmic F_{420}-non-reducing hydrogenase, absent from members of the Methanosarcinales and some Methanomicrobiales [66]. MvhA contains the Ni-Fe domain for activation of H_2. MvhADG and HdrABC were shown to form a complex that couples the reduction of CoM-S-S-CoB and a ferredoxin with H_2 through a flavin-based electron bifurcation in *Methanothermobacter marburgensis* [67]. Presence of MvhADG and HdrABC in the three Methanomassiliicoccales representatives suggests a similar process (Figure 4). Energy conservation may likely result from the subsequent reoxidation of ferredoxin coupled to translocation of H^+ (or possibly Na^+) across the membrane by a membrane associated enzymatic complex (Figure 4), as proposed by Thauer *et al.* [68] for *M. stadtmanae*. However the Ehb complex likely responsible for the translocation Na^+ in *M. stadtmanae* is not present in the three Methanomassiliicoccales representatives.

The only identified complex shared by the three genomes which could fulfil this role corresponds to the 11-subunits respiratory complex I found in a large number of archaea and bacteria [69]. This complex is homologous to the Fpo complex ($F_{420}H_2$ dehydrogenase) of Methanosarcinales [70]. Characterized respiratory complex I and Fpo catalyze the exergonic transfer of electrons from a cytoplasmic electron transporter to a membrane soluble electron transporter coupled to the translocation of ions across the membrane [69,70]. A similar process in Methanomassiliicoccales would thus imply a membrane associated electron transport chain which was so far only observed in *Methanosarcinales* among methanogens. The currently predicted enzymatic complex is truncated as compared to the Fpo of *Methanosarcina* spp. with the lack of homologues of the FpoO and FpoF subunits, forming an FpoABCDHIJKLMN like complex (Figure 4, Table 6). The lack of the FpoF subunit is similar to the Fpo complex of *Methanosaeta* representatives which were proposed to use ferredoxin instead of $F_{420}H_2$ as electron donor [71] (Table 6). The three genomes also harbor genes required for biosynthesis of a liposoluble electron transporter (Additional file 3, Table 6), whose role may be to accept electrons from the Fpo complex [72]. This membrane-soluble electron carrier, whose biochemical nature has to be determined experimentally, would drive electron transfer in the membrane, linking the Fpo complex to another membrane bound protein/complex, possibly a second coupling site reducing the heterodisulfide. The energy-converting hydrogenase EchA-F is another membrane enzymatic complex which could also translocate ions by the re-oxidation of the ferredoxin [73] but it only

Table 6 Genes involved in energy conservation in "*Ca*. M. alvus", "*Ca*. M. intestinalis" and *M. luminyensis* and accession numbers of the proteins they encode

	"*Ca*. M. alvus"	"*Ca*. M. intestinalis"	*M. luminyensis*	Transmembrane helices
ATP synthase				
ahaH	AGI84762.1	AGN25422.1	WP_019178382.1	no
ahaI	AGI84763.1	AGN25423.1	WP_019178381.1	yes
ahaK	AGI84764.1	AGN25424.1	WP_019178380.1	yes
ahaE	AGI84765.1	AGN25425.1	WP_019178379.1	no
ahaC	AGI84766.1	AGN25426.1	WP_019178378.1	no
ahaF	AGI84767.1	AGN25427.1	WP_019178377.1	no
ahaA	AGI84768.1	AGN25428.1	WP_019178376.1	no
ahaB	AGI84769.1	AGN25429.1	WP_019178375.1	no
ahaD	AGI84770.1	AGN25430.1	WP_019178374.1	no
Membrane-bound proton-translocating pyrophosphatase				
hppA	/	AGN26077.1	WP_019176822.1	yes
Heterodisulfide reductase				
hdrA	AGI85054.1	AGN25863.1	WP_019177460.1	no
hdrB1	AGI86093.1	AGN25718.1	WP_019177711.1	no
hdrB2	AGI85474.1	AGN25916.1	WP_019176125.1	no
hdrC1	AGI86094.1	AGN25719.1	WP_019177712.1	no
hdrC2	/	/	WP_019176126.1	no
hdrD1	AGI86375.1	AGN25510.1	WP_019178460.1	no
hdrD2	AGI86212.1	AGN25649.1	WP_019177852.1	no
hdrD3	/	/	WP_019177557.1	no
hdrE	/	/	/	/
Methyl-viologen-reducing hydrogenase				
mvhD1	AGI85055.1	AGN25864.1	WP_019177459.1	no
mvhD2	/	AGN25453.1	WP_019176201.1	no
mvhD3	/	/	WP_019176130.1	no
mvhG	AGI85056.1	AGN25865.1	WP_019177458.1	no
mvhA	AGI85057.1	AGN25866.1	WP_019177457.1	no
$F_{420}H_2$ dehydrogenase-like/11-subunit respiratory complex 1				
fpoA	AHA34030.1	AGN25601.1	WP_019176183.1	yes
fpoB	AGI84952.1	AGN25602.1	WP_019176182.1	no
fpoC	AGI84953.1	AGN25603.1	WP_019176181.1	no
fpoD	AGI84954.1	AGN25604.1	WP_019176180.1	no
fpoF	/	/	/	
fpoH	AGI84955.1	AGN25605.1	WP_019176179.1	yes
fpoI	AGI84956.1	AGN25606.1	WP_019176178.1	no
$fpoJ_N$	AGI84957.1	AGN25607.1	WP_019176177.1	yes
$fpoJ_C$	AGI84958.1	AGN25608.1	WP_019176176.1	yes
fpoK	AGI84959.1	AGN25609.1	WP_019176175.1	yes
fpoL	AGI84960.1	AGN25610.1	WP_019176174.1	yes
fpoM	AGI84961.1	AGN25611.1	WP_019176173.1	yes
fpoN	AGI84962.1	AGN25612.1	WP_019176172.1	yes

Table 6 Genes involved in energy conservation in "*Ca*. M. alvus", "*Ca*. M. intestinalis" and *M. luminyensis* and accession numbers of the proteins they encode *(Continued)*

fpoO	/	/	/	
Energy-converting hydrogenase				
echA1	/	AGN25511.1	WP_019178471.1	yes
echA2	/	AGN26997.1	WP_019176386.1	yes
echB1	/	AGN25512.1	WP_019178472.1	yes
echB2	/	AGN26998.1	WP_019176385.1	yes
echC1	/	AGN25513.1	WP_019178473.1	no
echC2	/	AGN26999.1	WP_019176384.1	no
echD1	/	AGN25514.1	WP_019178474.1	no
echD2	/	AGN27000.1	WP_019176383.1	no
echE1	/	AGN25515.1	WP_019178475.1	no
echE2	/	AGN27001.1	WP_019176382.1	no
echF1	/	AGN25516.1	WP_019178476.1	no
echF2	/	AGN27002.1	WP_019176381.1	no
Liposoluble electron transporter synthesis				
ispA[a]	AGI84964.1	AGN25614.1	WP_019176170.1	/
ubiA[b]	AGI85875.1	AGN26416.1	WP_019178349.1	/
		AGN26109.1		
ubiE[c]	AGI85874.1	AGN26417.1	WP_019178072.1	/
		AGN25541.1	WP_019178198.1	
			WP_019176998.1	

[a]encoding a geranylgeranyl pyrophosphate synthase (GGPPS).
[b]encoding a1,4-dihydroxy-2-naphthoate octaprenyltransferase (DHNOPT).
[c]encoding a 2-heptaprenyl-1,4-naphthoquinone methyltransferase (HPNQMT).

occurs in *M. luminyensis* and "*Ca.* M. intestinalis" (Figure 4). Nevertheless EchA-F could also operate in reverse and exploit the chemosmotic gradient for anabolic reactions [74]. Finally, a gene encoding a membrane-bound pyrophosphatase is found in the genomes of *M. luminyensis* and "*Ca.* M. intestinalis" (Table 6) but not in "*Ca.* M. alvus". This protein is predicted to allow the translocation of protons across the cytoplasmic membrane by hydrolysis of PPi to phosphate [75,76].

The three genomes share an original combination of genes likely involved in energy conservation, suggesting a different process than what is observed in other methanogens. The predicted flavin-based electron bifurcation in MvhADG/HdrABC complex is a feature shared by most methanogens with the exception of Methanosarcinales and some Methanomicrobiales representatives, while the putative membrane associated electron transport chain related to the activity of the Fpo-like complex was so far a unique feature of Methanosarcinales among methanogens. However, no membrane-bound cytochrome protein like those of the Methanosarcinales was detected to be encoded by the three genomes and the complete process remains to be uncovered.

Amber codon usage and putative Pyl-containing proteins

Previous studies have shown that the genes coding for methyl:corrinoid methyltransferases B dedicated to methylamines utilization (*mtmB*, *mtbB* and *mttB* for mono-, di- and tri-methylamines, respectively) present in *M. luminyensis*, "*Ca.* M. intestinalis" and "*Ca.* M. alvus" contain an in-frame *amber* Pyl-encoding codon [7,8,25], similarly to what is observed in Methanosarcinaceae and in a few bacteria [77,78], where it encodes the 22nd proteogenic amino acid pyrrolysine (Pyl, O). All the necessary genetic machinery is found in the three Methanomassiliicoccales genomes, including the genes for pyrrolysine synthesis (pylBCD), the *amber* suppressor tRNA^{Pyl} (pylT) and the dedicated amino-acyl tRNA synthetase (pylS) [25]. The presence of decoding *amber* machinery questions the occurrence of Pyl in other proteins than the methyltransferases involved in methylotrophic methanogenesis. This possibility was addressed in the present study by searching all the TAG-interrupted CDS which share the same BLASTP hit with the virtual in-frame translation of the 3' flanking region. These CDS were fused *in silico* as a unique CDS, stopping at the next stop codon and predicted as potentially incorporating Pyl

during the translation process. As a positive control, this strategy identified the above-mentioned methylamines: corrinoid methyltransferases in the three genomes. No putative other Pyl-containing proteins were identified in *M. luminyensis.* One additional *amber*-containing CDS was determined in "*Ca.* M. intestinalis", a putative Fe-S binding protein (AGY50215), which is absent in "*Ca.* M. alvus" and present in *M. luminyensis* but not predicted to incorporate Pyl. "*Ca.* M. alvus" contains the highest number of predicted Pyl-containing proteins, 16 in addition to the methylamines: corrinoid methyltransferases (Table 7, Figure 5). Half of them have homologues in the two other genomes but without in-frame *amber* codons (in bold, Table 7). Among these 16 proteins, several have a hypothetical function and some are highly conserved in methanogens and/or archaea. This is the case of a digeranylgeranylglyceryl phosphate synthase required in the synthesis of archaeal phospholipids and of the putative methyltransferase MtxX (Tables 5 and 7). The CRISPR associated *cas1* gene, although present in the three genomes, is only detected as a Pyl-containing enzyme in "*Ca.* M. alvus". The activity and the effective incorporation of Pyl in such a large range of enzymes of the same organisms remain to be determined experimentally. However, this could reasonably be assumed considering the existence of few functional Pyl-containing proteins (different of methylamines:corrinoid methyltransferases) reported from both Pyl-decoding archaea and bacteria [77,79,80].

Particular genetic signals in the genes containing an in-frame TAG have been proposed to enhance the incorporation of Pyl in the proteins but are not obligatorily requested for that purpose [81]. Two alternative adaptations have been proposed for *Methanosarcina* spp. and the bacteria *Acetohalobium arabaticum* to minimize proteome alteration in consequence of the insertion of Pyl on the stop codons normally intended to stop the translation [77]. In *A. arabaticum* the expression of the Pyl-cassette has been shown to be regulated by substrate (trimethylamine) availability, while in *Methanosarcina* spp. which constitutively express the Pyl-cassette [79,82], the frequency of genes ended by a TAG stop codon is minimized (~4-5% in *Methanosarcina* spp. *vs.* 20-30% in *A. arabaticum* and other Pyl-decoding bacteria, see Additional file 1: Table S12 adapted from [77]). Accordingly, the extremely low frequency of TAG stop codons in "*Ca.* M. alvus" (1.6%) suggests a constitutive expression of the Pyl-cassette and an efficient ability to incorporate Pyl in proteins (Figure 5, Additional file 1: Table S12). In such $tRNA^{Pyl}$ suppressing context, the apparition of an in-frame *amber* codon in a

Table 7 Putative Pyl-containing proteins in "*Ca*. M. alvus"

Accession number	Annotation	Size[a]	Comments
AGI84833.1	hypothetical protein	253	DPM synthase like/GT2 superfamily
AGI85009.1	hypothetical protein	270	digeranylgeranylglyceryl phosphate synthase
AGI85117.2	phosphotransacetylase-like protein	242	putative methyltransferase MtxX
AGI85168.2	filamentation induced by cAMP protein Fic	425	
AGI85186.1	hypothetical protein	149	Rv0623-like transcription factor
AGI85280.1	hypothetical protein	917	glycosyltransferase family 29
AGI85290.1	hypothetical protein	148	
AGI85300.1	hypothetical protein	444	ATPase domain
AGI85437.1	hypothetical protein	536	prophage Lp3 protein 8 (helicase) of *Lactobacillus* spp.
AGI85443.1	hypothetical protein	717	
AGI85449.1	hypothetical protein	262	putative methyltransferase
AGI85596.1	hypothetical protein	162	putative acetyltransferase
AGI85630.1	hypothetical protein	322	CRISPR- associated endonuclease cas1
AGI85862.1	*MMA:corrinoid methyltransferase*	*459*	
AGI85863.1	*MMA:corrinoid methyltransferase*	*461*	
AGI85869.1	*TMA:corrinoid methyltransferase*	*504*	
AGI85870.1	*DMA:corrinoid methyltransferase*	*469*	
AGI86303.1	hypothetical protein	389	Sel-1 domain containing protein
AGI86346.1	transporter family protein	289	bacterial/archaeal transporter family protein
AGI86379.1	uncharacterized protein	187	conserved in archaea (DUF531)

Proteins in bold indicate homologs in the two other members of the Methanomassiliicoccales, devoided of Pyl. Proteins in italics indicate homologs in the two other members of the Methanomassiliicoccales also containing Pyl.
[a]Number of amino acids.

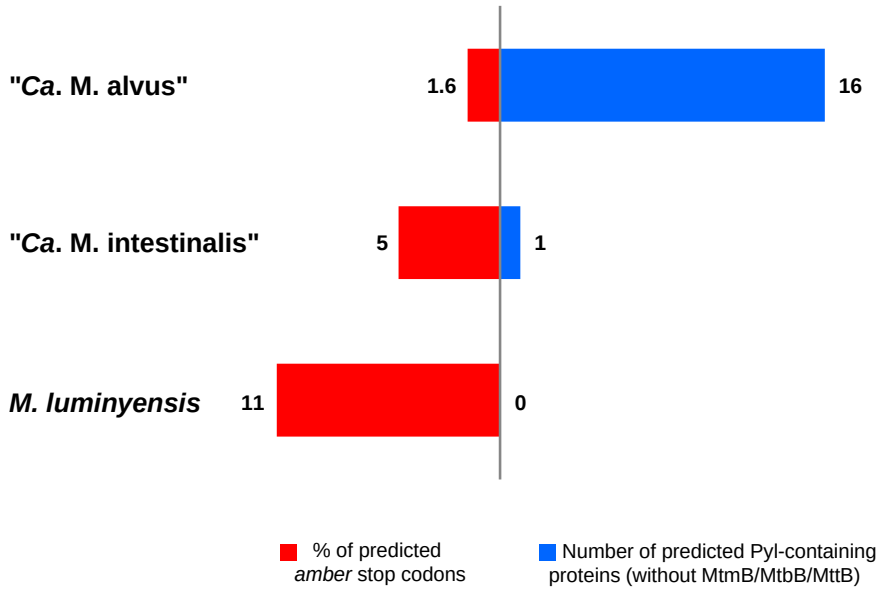


Figure 5 Comparison of the number of putative Pyl-containing proteins (other than MtmB/MtbB/MttB methyltransferases) and of the percentage of *amber* codons used as translational stop signals deduced from the three Methanomassiliicoccales genomes.

CDS would lead to a stable mutation as supported by the high occurrence of genes predicted to encode Pyl-containing proteins in "*Ca.* M. alvus". The phylogenetic position of "*Ca.* M. alvus" among a large cluster of gut methanogens suggests a long evolutionary history in this type of environments where mono- di- and trimethylamine are likely not limiting [17,83,84] and may be obtained through the degradation of glycine betaine, choline and L-carnitine by co-occurring microorganisms [85-87]. This high availability of methylamines during the evolution of "*Ca.* M. alvus", involving a possibly high and constant expression of the Pyl-machinery, could have been a driving factor that has led to this particularly low usage of the triplet TAG in CDSs as termination signals during translation. In addition, the insertion of an *amber* codon in a gene coding for a protein of major function (such as the highly conserved MtxX, Cas1 or the digeranylgeranylglyceryl phosphate synthase in the present case) might have turned the expression of the Pyl cassette and the efficient ability to incorporate Pyl essential for growth. As a feedback this would contribute to tight the association of "*Ca.* M. alvus" cluster methanogens with digestive tract environments. The absence of predicted Pyl-encoding proteins other than MtmB, MtbB and MttB and the high frequency of genes ended by TAG (11.3%) in *M. luminyensis* (Figure 5, Additional file 1: Table S12) argue for a different handling of the Pyl-encoding capacity, possibly through a more important regulation of Pyl-incorporation, and could reflect an adaptation to lower or more variable availability in methylamines [88]. Together with other genomic traits described above, this supports a larger distribution of *M. luminyensis* than digestive tract environments. Following the hypothesis of a methylamine-directed selective pressure on TAG usage in CDSs of the Methanomassiliicoccales, the intermediate TAG usage in the CDSs of "*Ca.* M. intestinalis" (Figure 5, Additional file 1: Table S12) would reflect a more stringent association to digestive tracts compared to *M. luminyensis*.

Conclusions

Several atypical features were identified in the three genomes such as the scattering of the ribosomal RNA genes and the absence of eukaryotic-like histone gene otherwise present in most of Euryarchaeota genomes. The lack of the eukaryotic-like histone gene could represent an ancestral loss of the overall branch composed by Thermoplasmatales and related lineages, replaced by bacterial-type histone in Thermoplasmatales or Alba protein present in all genomes of the branch with the exception of "*Ca.* M. intestinalis" and "*Ca.* M. alvus". Intriguingly, the nature of this protein remains elusive in "*Ca.* M. intestinalis" and "*Ca.* M. alvus".

The absence of a large number of genes otherwise present in all methanogens, but not all restricted to methanogens, was previously reported in *M. luminyensis* and "*Ca.* M. alvus" genomes and is presently extended to "*Ca.* M. intestinalis". The large lack of these genes involved in the CO_2 reduction/methyl-oxidation pathways in other methanogens offers a unique context to redefine the genes encoding enzymes or isoenzymes shared by all and only methanogens. Interestingly, the reevaluation shows that this core is not deeply changed when

Methanomassiliicoccales are considered. In addition to the genes encoding the Mtr complex, only two of these methanogenesis marker genes are absent from the Methanomassiliicoccales genomes. Gathered with *mcrABG* on a small genomic portion in *M. luminyensis* and "*Ca.* M. alvus", core genes encoding uncharacterized proteins could be intimately involved in the functioning of the Mcr complex. The process of energy conservation associated to methanogenesis on methyl-compound reduction with H_2 was analyzed. The original composition of genes presently identified to take part to this process suggests the involvement of a flavin-based electron bifurcation and a membrane associated electron transport chain which are distinctive elements of the two main energy conservation processes defined in other methanogens. However the complete process remains to be uncovered and several components have to be characterized.

While the three Methanomassiliicoccales representatives were cultured from gastrointestinal tract, the analysis of their genome revealed differential adaptations to this environment and possibly contrasted evolutionary history. One of the striking differences among the three species relies on their usage of the TAG codon which could have been shaped by the availability of methylamines as a substrate during their evolution. The long term adaptation of "*Ca.* M. alvus" to GIT environments, suggested by its position among a large cluster of GIT-derived sequences, is supported by its gene composition, along with lateral gene transfer from GIT-associated bacteria. The phylogenetic position of *M. luminyensis* and "*Ca.* M. intestinalis" among soil and sediment methanogens suggests a more recent adaptation or more facultative association to GIT conditions. Consistent with this hypothesis, the *M. luminyensis* genome contains several important genes which are specifically present in soil and sediment methanogens. Although phylogenetically close to *M. luminyensis*, "*Ca.* M. intestinalis" has a reduced genome with a lower [G + C] % and does not share the signatures of soil or sediment adaptations of *M. luminyensis*. These differences could reflect a phenomenon of streamlining in the "*Ca.* M. intestinalis" genome linked with its adaptation to GIT conditions. A similar phenomenon was previously reported from free-living bacteria [29] and with more extreme amplitude, in obligate pathogens [89] as well as in bacterial [90] and archaeal [91] symbionts.

Methods

Gene structure prediction

Complete genome sequences of "*Ca.* M. alvus" [GenBank: NC_020913.1] and "*Ca.* M. intestinalis" [GenBank: NC_021353.1] were obtained from enriched consortia of stool-derived cultures from a 91-year-old woman, with an average genome sequence coverage respectively of 36.9 fold and 42.7 fold [7,8]. Genomic sequences from *Methanomassiliicoccus luminyensis* B10 were retrieved from the Genbank database [GenBank: CAJE01000001-CAJE01000026]. Raw sequences from "*Ca.* M. alvus" Mx1201, "*Ca.* M. intestinalis" Mx1-Issoire and *M. luminyensis* were fed to the RAST Annotation server [92] using Glimmer3 [93] for open-reading frames prediction. The RAST Annotation used the released 59 of FIGfam and no frameshifts fixing parameters. To perform an accurate structural annotation of these genomes, a comparative analysis of the "*Ca.* M. alvus", "*Ca.* M. intestinalis" and *M. luminyensis* annotated proteomes was conducted using the TBLASTN program. To identify genes or distantly related genes, a BLOSUM45 substitution matrix was chosen, and low-complexity filters were suppressed. TBLASTN analyses were manually validated to take into account genes with frame-shifts due to sequencing errors. Translation start codons were then validated through a BLASTP comparative analysis of the three annotated proteomes. Protein sequences from the three proteomes were compared together with the curated SWISS-PROT protein sequences database [94]. Results were filtered using 80% length and 40% identity thresholds and start codons were manually corrected taking into account protein sizes and local alignments. Non-coding RNAs were predicted using the Rfam database [95] with an E-value threshold of 1 and results were manually curated. Additional analyses were performed to detect tRNAs by merging results from tRNAscan [96], TFAM [97], ARAGORN [98] and BLASTN [99]. CRISPRFinder [100] was applied for each of the three genomes to detect CRISPR loci that were compared together using CRISPRcompar [101] and CRISPRmap [102]. Finally, prophages were sought using PHAST [103]. Circular representation of the "*Ca.* M. alvus" and "*Ca.* M. intestinalis" genomes were performed using the CGView Server [104].

Comparative genome analysis and functional annotation

An 'all-*versus*-all' BLASTP comparison of the predicted protein sequences within each of the three genomes was conducted [99]. On the basis of the best BLASTP hits, orthologous relationships were established between the protein sequences of "*Ca.* M. alvus", "*Ca.* M. intestinalis" and *M. luminyensis*. A Venn diagram was then drawn using the Venny web service [105]. Predicted functions provided by the RAST annotation server for each CDS of the three species were kept as functional annotation. Using orthology relationships previously established, a functional annotation transfer was performed. Protein sequences of genes with frame-shift mutations were manually reconstructed. In order to distinguish protein sequences only found within the three genomes and shared protein sequences with closely related species, a BLASTP analysis was conducted. Each protein sequence from the core proteome was compared to i) phylogenetic

neighbors proteomes (*Aciduliprofundum boonei* T469, accession code: NC_013926; *Aciduliprofundum sp.* MAR08-339, accession code: NC_019942; *Ferroplasma acidarmanus* fer1, accession code: CM000428; *Thermoplasma acidophilum* DSM 1728, accession code: NC_002578; *Thermoplasma volcanium* DSS1, accession code: NC_002689 and MG-II, accession code: CM001443), ii) methanogenic archaeon from human gut (*Methanobrevibacter smithii* ATCC 35061, accession code: NC_009515 and *Methanosphaera stadtmanae* DSM 3091, accession code: NC_007681) and iii) the NCBI non-redundant protein sequences database (release 12/2012). Identity threshold was set at 30% with a minimum length coverage of 80%. An arCOG [45] analysis was also performed using the December 2012 release (ftp://ftp.ncbi.nih.gov/pub/wolf/COGs/arCOG/). Each annotated protein sequence from the three genomes was compared to the arCOG database using BLASTP and an E-value threshold equal to $1e^{-3}$. The arCOG profiles of the three genomes and those of the arCOG database were used to identify proteins potentially shared by all and only methanogens, as well as proteins almost specific to methanogens and shared by almost all methanogens. Distribution of each selected protein among sequenced organisms was checked by BLASTP. Conserved domains of the selected proteins were compared to those of the closest results that belong to non-methanogens and phylogenetic three were constructed to verify their monophyly. Additional proteomes from various archaeal orders were also submitted to this comparison: *A. boonei* T469; *Archaeoglobus fulgidus* DSM 4304, accession code: NC_013926; *Archaeoglobus veneficus* SNP6, accession code: NC_015320; *M. smithii* ATCC 35061; *M. stadtmanae* DSM 3091; *Thermoplasma acidophilum* DSM 1728, accession code: NC_002578 and MG-II. In order to detect putative lateral gene transfers, the same BLASTP analysis was performed for the three proteomes using the UniprotKB database [106]. Only best hits were retrieved and classified according to the three domains of life: Archaea, Bacteria or Eukaryota. The genomes of the Methanomassiliicoccales representatives were not included in the subject database. Metabolic pathways reconstruction was performed through the KEGG Automatic Annotation Server (KAAS) [107] using a bi-directional best hit strategy and a custom list of reference organisms. Indeed, based on best BLAST hit results from the three proteomes, 40 species were selected for the KAAS (three-letter organism codes are listed as follows: abi, mac, tac, mba, rci, mig, afu, mpd, tba, mpi, pab, mka, pho, mhu, mja, mla, mth, cdc, amt, drm, mbn, ssg, ele, fnu, mel, mrv, fsv, tsi, lba, ral, sti, msi, sce, eco, ere, aas, eha, sfu, bla, cau). The transportome was determined using the TransportTP server [108] (reference organism: *Escherichia coli*; E-value threshold: 0.1). Results were manually validated and curated using BLASTP analysis using transportDB [109] and taking into account orthology relationships. Signal peptides, transmembrane helices and PFAM domains were predicted through the InterProScan annotation module provided by the BLAST2GO software [110] with default parameters.

Phylogenomic analysis of DNA replication components

Homologs of each major archaeal DNA replication component were retrieved from the reference sequence database at the NCBI using the BLASTP program with different seeds from each archaeal order [99]. The top 100 best hits for each order were then used to create HMM profiles [111] (http://www.hmmer.org) that allowed iteratively searching a local database of 142 complete or nearly complete archaeal genomes including 98 plasmid sequences, as well as in a local database of the available complete archaeal virus genomes (56 total) downloaded from the Viral Genomes database of NCBI (as of June 20th 2013). Absences of a given homolog in a specific genome were verified by performing additional TBLASTN searches [99]. Multiple alignments were done with MUSCLE 3.8.31 [112] and manually inspected using the ED program from the MUST package to remove non-homologous or partial sequences [113]. The alignments were trimmed using the software BMGE [114] with default parameters. Phylogenetic analyses were performed on single protein datasets using Maximum Likelihood and Bayesian methods. Maximum likelihood analyses were performed with RaxML [115]. Mr. Bayes 3.2 [116] was used to perform Bayesian analyses using the mixed amino acid substitution model and four categories of evolutionary rates. Two independent runs were performed for each data set, and runs were stopped when they reached a standard deviation of split frequency below 0.01 or the log likelihood values reached stationarity. The majority-rule consensus trees were obtained after discarding first 25% samples as 'burn-in'.

Data access

The whole genome shotgun projects, the complete genome sequences and annotations have been deposited at DDBJ/EMBL/GenBank for "*Ca.* M. alvus" Mx1201 [GenBank: CP004049] and for "*Ca.* M. intestinalis" Issoire-Mx1 [GenBank: CP005934]. Predicted CDS and protein sequences for *M. luminyensis*, some of which are not annotated in GenBank are provided respectively through Additional file 1: Tables S12 and S13.

Additional files

Additional file 1: Additional tables in a zipped folder containing: Table S1. tRNA and ncRNA contents for the genomes of the three Methanomassiliicoccales representatives. **Table S2.** Codon usage in the three genomes of Methanomassiliicoccales. **Table S3.** CRISPR DR elements found in the three genomes. **Table S4.** Number of best hits

score among the three domains of life. **Table S5.** Genes list of the core genome of the Methanomassiliicoccales, as deduced by a TBLASTN analysis (with reference to CDS of "*Ca*. M. alvus" genome), and their presence or not in phylogenetical neighbors, human gut Methanobacteriales and non-redundant genbank DB. **Table S6.** CDS list of the core genome of the Methanomassiliicoccales, absent in phylogenetical neighbors and the human gut Methanobacteriales. In blue, the 20 CDS not retrieved in genbank database. **Table S7.** arCOG distribution among the Methanomassiliicoccales representative genomes, gut methanogens and some other archaea. **Table S8.** Complete list of transporters detected by TransportDB, in the three genomes of Methanomassiliicoccales. **Table S9.** List of the antioxydant systems in the three genomes of Methanomassiliicoccales. **Table S10.** Genes involved in methanogenesis in "*Ca*. M. alvus", "*Ca*. M. intestinalis" and *M. luminyensis* and accession numbers of the proteins they encode. **Table S11.** Comparative presence of the genes involved in the synthesis of the coM among the seven orders of methanogens. **Table S12.** Numbers of CDS with in-frame TAG, and % of the total CDS in various genomes of microorganisms coding or not pyrrolysine (update information from Prat *et al*. [77]). **Table S13.** CDS list of *M. luminyensis* B10. **Table S14.** Proteome of *M. luminyensis* B10.

Additional file 2: Additional figures in a zipped folder containing: Figure S1. CRISPR Direct Repeats structure. The figure shows the 2D, Minimum Free Energy structure of CRISPR DRs retrieved from the three genomes of the Methanomassiliicoccales (using RNAfold web server [117]) and the sequence alignment of *M. luminyensis* DR with the family 3, motif 27 DRs (using CRISPRmap [34])**. Figure S2.** Chromosome circular maps of (A) "*Candidatus* Methanomethylophilus alvus" Mx1201 and (B) "*Candidatus* Methanomassiliicoccus intestinalis" Mx1-Issoire genomes (generated with CGView [104]). Circles display from outside: 1 and 4, rRNA genes respectively on forward and reverse strand; 2 and 3, CDS on forward and reverse strand; 5, BLASTX results with a maximum expected value of $1e^{-3}$ *versus* the "*Ca*. M. intestinalis" proteome; 6, [G + C] % content deviation from the average [G + C] % content of the genome. Arrows, location and sense of the *orc1/cdc6* genes. **Figure S3.** Phylogeny of Cdc6/Orc1 proteins. **Figure S4.** Phylogenetic trees of NAD-dependent DNA ligase (A) and Choloyglycine hydrolase (B) genes likely transferred from bacteria to "*Ca*. M. alvus". In red, sequences of "*Ca*. M. alvus", in blue sequences from other gut-associated methanogens. **Figure S5.** Metabolic comparison of the three genomes based on KEGG maps. Series of three boxes represent presence or absence of the E.C. numbered enzyme (yellow for "*Ca*. M. alvus", green for "*Ca*. M. intestinalis" and blue for *M. luminyensis*). Green arrows replace complex pathways. Blue boxes, synthetized compounds by the 3 species; Red boxes, compounds not synthetized by the three species. Orange boxes, compounds synthetized by at least 1 species. Question marks show pathways where there is at least one enzyme missing. **Figure S6.** Comparison of the physical map of genes involved in methanogenesis on methyl compounds + H_2 in the three analyzed genomes.

Additional file 3: Additional Data in a zipped MS Word file. Details on lipids, amino acids and purine synthesis, as well as molecular nitrogen fixation deduced from the genomes of the three members of the Methanomassiliicoccales.

Abbreviations

aaRS: Aminoacyl tRNA synthetase; AIR: 5-amino -4-imidazole ribonucleotide; arCOG: Archaeal Cluster of Orthologous Genes; ASAT: Aspartate aminotransferase; BSH: Bile Salt Hydrolase; CAIR: 5-amino-4-imidazole carboxylic acid ribonucleotide; Cdc6: Cell division cycle 6; CDS: Coding DNA sequence; CRISPR: Clustered Regularly Interspaced Short Palindromic Repeat; CS: Cysteate Synthase; DHNOPT: 1,4-dihydroxy-2-naphthoate octaprenyltransferase; DMS: Dimethylsulfide; DPL: DNA polymerase D large subunit; DPM: Dolichol-phosphate mannose; DPS: DNA polymerase D small subunit; FEN-1: Flap EndoNuclease 1; GGPS: (S)-3-O-geranylgeranylglyceryl phosphate synthase; GINS: Go-Ichi-Nii-San protein; GIT: Gastro-intestinal tract; GT: Glycosyl Transferase; H_4MP: Tetrahydromethanopterin; Hec: Unknown hydrogenase, probable energy-converting; HPNQMT: 2-heptaprenyl-1,4-naphthoquinone methyltransferase; IPPK: Isopentenyl phosphate kinase; LGT: Lateral Gene Transfer; LSU: Large subunit; MCM: Minichromosome maintenance protein; MG-II: Uncultured Marine Group II; NCAIR: N5-5-amino -4-imidazole carboxylic acid ribonucleotide; nr: Non-redundant; ORB: Origin recognition box; ORF: Open Reading Frame; Ori: Origin of replication; PCNA: Proliferating Cell Nuclear Antigen; PFOR: Pyruvate:ferredoxin oxydoreductase;
PMDC: Phosphomevalonate decarboxylase; PMK: Phosphomevalonate kinase; PPS: Polyprenyl synthetase; PriL: Primase large subunit; PriS: Primase small subunit; PRPP: Phosphoribosyl pyrophosphate; Pyl: Pyrrolysine; RCC: Rumen cluster C; RFC: Replication Factor C; RNaseH: Ribonuclease H; RPA: Replicative Protein A; SD: Standard Deviation; SSB: Single Strand Binding protein; SSU: small subunit; TMA: Trimethylamine; Topo: Topoisomerase.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

GB, NP, HMBH, EP, NG, WT, PP, PWOT and JFB performed the bioinformatic analyses of these genomes (from assembly to the functional annotations, encompassing general statistics, tRNAs, ncRNA, CRISPRs, transporters,…). OB, GB, NP and JFB determined the general metabolisms. KR and SG identified the core DNA replication genes and performed phylogenetic analyses. GB, NP, PP, EP, PWOT and JFB conceived the study, participated in its design and coordination. GB, NP, PP, EP, PWOT, SG and JFB helped to draft the manuscript. All authors read and approved the final manuscript.

Acknowledgements

This work was supported by three PhD. Scholarship supports, one from the "Direction Générale de l'Armement" (DGA) to N.P., one from the French "Ministère de l'Enseignement Supérieur et de la Recherche" to N.G. and one of the European Union (UE) and the Auvergne Council to W.T. (FEDER). P.W. O.T. was supported by Science Foundation Ireland through a Principal Investigator award, by a CSET award to the Alimentary Pharmabiotic Centre, and by an FHRI award to the ELDERMET project by the Dept. Agriculture, Fisheries and Marine of the Government of Ireland. SG is supported by the Investissement d'Avenir grant "Ancestrome" (ANR-10- BINF-01-01). KR is a scholar from the Pasteur – Paris University (PPU) International PhD program and receives a stipend from the Paul W. Zuccaire Foundation. JFB thanks the "centre hospitalier Paul Ardier" in Issoire, especially Dr Mansoor, Dr Denozi and their staff for their valuable help, and Agnès Mihajlovski for her help in initiating this project.
This article is dedicated to the memory of PB (1937-2009), who was much more than the anecdotally first human known to carry an archaeon from the 7th methanogenic order.

Author details

[1]EA-4678 CIDAM, Clermont Université, Université d'Auvergne, 28 Place Henri Dunant, BP 10448, 63000 Clermont-Ferrand, France. [2]School of Microbiology and Alimentary Pharmabiotic Centre, University College Cork, Cork, Ireland. [3]CNRS, UMR 6023, Université Blaise Pascal, 63000 Clermont-Ferrand, France. [4]GReD, CNRS, UMR 6293, Inserm, UMR 1103, Clermont Université, Université d'Auvergne 28 Place Henri Dunant, BP 10448, 63000 Clermont-Ferrand, France. [5]Département de Microbiologie, Unité de Biologie Moléculaire du Gène chez les Extrêmophiles, Paris 75724 Cedex 15, France. [6]Université Pierre et Marie Curie, Cellule Pasteur UPMC, Paris 75724 Cedex 15, France.

Received: 10 January 2014 Accepted: 18 July 2014
Published: 13 August 2014

References

1. Tajima K, Nagamine T, Matsui H, Nakamura M, Aminov RI: **Phylogenetic analysis of archaeal 16S rRNA libraries from the rumen suggests the existence of a novel group of archaea not associated with known methanogens.** *FEMS Microbiol Lett* 2001, **200**(1):67–72.
2. Wright A-DG, Williams AJ, Winder B, Christophersen CT, Rodgers SL, Smith KD: **Molecular diversity of rumen methanogens from sheep in Western Australia.** *Appl Environ Microb* 2004, **70**(3):1263–1270.
3. Janssen PH, Kirs M: **Structure of the archaeal community of the rumen.** *Appl Environ Microb* 2008, **74**(12):3619–3625.
4. Mihajlovski A, Alric M, Brugère J-F: **A putative new order of methanogenic Archaea inhabiting the human gut, as revealed by molecular analyses of the *mcrA* gene.** *Res Microbiol* 2008, **159**(7):516–521.
5. Mihajlovski A, Doré J, Levenez F, Alric M, Brugère JF: **Molecular evaluation of the human gut methanogenic archaeal microbiota reveals an age-associated increase of the diversity.** *Environ Microbiol Rep* 2010, **2**(2):272–280.

6. Dridi B, Fardeau ML, Ollivier B, Raoult D, Drancourt M: ***Methanomassiliicoccus luminyensis*** **gen. nov., sp. nov., a methanogenic archaeon isolated from human faeces.** *Int J Syst Evol Microbiol* 2012, **62**(Pt 8):1902–1907.
7. Borrel G, Harris HM, Tottey W, Mihajlovski A, Parisot N, Peyretaillade E, Peyret P, Gribaldo S, O'Toole PW, Brugère JF: **Genome sequence of "*Candidatus* Methanomethylophilus alvus" Mx1201, a methanogenic archaeon from the human gut belonging to a seventh order of methanogens.** *J Bacteriol* 2012, **194**(24):6944–6945.
8. Borrel G, Harris HM, Parisot N, Gaci N, Tottey W, Mihajlovski A, Deane J, Gribaldo S, Bardot O, Peyretaillade E: **Genome sequence of "*Candidatus* Methanomassiliicoccus intestinalis" Issoire-Mx1, a third Thermoplasmatales-related methanogenic archaeon from human feces.** *Genome Announc* 2013, **1**(4):e00453–00413.
9. Paul K, Nonoh JO, Mikulski L, Brune A: **"Methanoplasmatales", Thermoplasmatales-related archaea in termite guts and other environments, are the seventh order of methanogens.** *Appl Environ Microb* 2012, **78**(23):8245–8253.
10. Iino T, Tamaki H, Tamazawa S, Ueno Y, Ohkuma M, Suzuki K, Igarashi Y, Haruta S: ***Candidatus*** **Methanogranum caenicola: a novel methanogen from the anaerobic digested sludge, and proposal of methanomassiliicoccaceae fam. nov. and Methanomassiliicoccales ord. nov., for a Methanogenic Lineage of the Class Thermoplasmata.** *Microbes Environ/JSME* 2013, **28**(2):244–250.
11. Hedderich R, Whitman WB: **Physiology and biochemistry of the methane-producing Archaea**. In *The prokaryotes*. New York: Springer; 2006:1050–1079.
12. Oren A, Garrity GM: **List of new names and new combinations previously effectively, but not validly, published.** *Int J Syst Evol Microbiol* 2013, **63**(11):3931–3934.
13. Huang XD, Tan HY, Long R, Liang JB, Wright A-DG: **Comparison of methanogen diversity of yak (*Bos grunniens*) and cattle (*Bos taurus*) from the Qinghai-Tibetan plateau, China.** *BMC Microbiol* 2012, **12**(1):237.
14. Wright A-DG, Auckland CH, Lynn DH: **Molecular diversity of methanogens in feedlot cattle from Ontario and Prince Edward Island, Canada.** *Appl Environ Microb* 2007, **73**(13):4206–4210.
15. Wright A-DG, Toovey AF, Pimm CL: **Molecular identification of methanogenic archaea from sheep in Queensland, Australia reveal more uncultured novel archaea.** *Anaerobe* 2006, **12**(3):134–139.
16. Borrel G, O'Toole PW, Harris HM, Peyret P, Brugère J-F, Gribaldo S: **Phylogenomic data support a seventh order of methylotrophic methanogens and provide insights into the evolution of methanogenesis.** *Genome Biol Evol* 2013, **5**(10):1769–1780.
17. Poulsen M, Schwab C, Jensen BB, Engberg RM, Spang A, Canibe N, Hojberg O, Milinovich G, Fragner L, Schleper C, Weckwerth W, Lund P, Schramm A, Urich T: **Methylotrophic methanogenic Thermoplasmata implicated in reduced methane emissions from bovine rumen.** *Nat Commun* 2013, **4**:1428.
18. Gorlas A, Robert C, Gimenez G, Drancourt M, Raoult D: **Complete genome sequence of *Methanomassiliicoccus luminyensis*, the largest genome of a human-associated Archaea species.** *J Bacteriol* 2012, **194**(17):4745–4745.
19. Gaci N, Borrel G, Tottey W, O'Toole PW, Brugère JF: **Archaea from the human gut: the new beginning of an old story.** *World J Gastroenterol* in press.
20. Brugère JF, Borrel G, Gaci N, Tottey W, O'Toole PW, Malpuech-Brugère C: **Archaebiotics: proposed therapeutic use of archaea to prevent trimethylaminuria and cardiovascular disease.** *Gut Microbes* 2014, **5**(1):6.
21. Mackay RJ, McEntyre CJ, Henderson C, Lever M, George PM: **Trimethylaminuria: causes and diagnosis of a socially distressing condition.** *Clin Biochem Rev* 2011, **32**(1):33.
22. Wang Z, Klipfell E, Bennett BJ, Koeth R, Levison BS, DuGar B, Feldstein AE, Britt EB, Fu X, Chung Y-M: **Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease.** *Nature* 2011, **472**(7341):57–63.
23. Srinivasan G, James CM, Krzycki JA: **Pyrrolysine encoded by UAG in Archaea: charging of a UAG-decoding specialized tRNA.** *Science* 2002, **296**(5572):1459–1462.
24. Krzycki JA: **Function of genetically encoded pyrrolysine in corrinoid-dependent methylamine methyltransferases.** *Curr Opin Chem Biol* 2004, **8**(5):484–491.
25. Borrel G, Gaci N, Peyret P, O'Toole PW, Gribaldo S, Brugère J-F: **Unique characteristics of the pyrrolysine system in the 7th order of methanogens: implications for the evolution of a genetic code expansion cassette.** *Archaea* 2014, **2014**:374146.
26. Sheppard K, Yuan J, Hohn MJ, Jester B, Devine KM, Söll D: **From one amino acid to another: tRNA-dependent amino acid biosynthesis.** *Nucleic Acids Res* 2008, **36**(6):1813–1825.
27. Ree HK, Zimmermann RA: **Organization and expression of the 16S, 23S and 5S ribosomal RNA genes from the archaebacterium *Thermoplasma acidophilum*.** *Nucleic Acids Res* 1990, **18**(15):4471–4478.
28. Ciesielski S, Bulkowska K, Dabrowska D, Kaczmarczyk D, Kowal P, Mozejko J: **Ribosomal intergenic spacer analysis as a tool for monitoring methanogenic archaea changes in an anaerobic digester.** *Curr Microbiol* 2013.
29. Dufresne A, Garczarek L, Partensky F: **Accelerated evolution associated with genome reduction in a free-living prokaryote.** *Genome Biol* 2005, **6**(2):R14.
30. Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, Romero DA, Horvath P: **CRISPR provides acquired resistance against viruses in prokaryotes.** *Science* 2007, **315**(5819):1709–1712.
31. Fischer S, Maier LK, Stoll B, Brendel J, Fischer E, Pfeiffer F, Dyall-Smith M, Marchfelder A: **An archaeal immune system can detect multiple protospacer adjacent motifs (PAMs) to target invader DNA.** *J Biol Chem* 2012, **287**(40):33351–33363.
32. Sorek R, Kunin V, Hugenholtz P: **CRISPR–a widespread system that provides acquired resistance against phages in bacteria and archaea.** *Nat Rev Microbiol* 2008, **6**(3):181–186.
33. Jansen R, Embden JD, Gaastra W, Schouls LM: **Identification of genes that are associated with DNA repeats in prokaryotes.** *Mol Microbiol* 2002, **43**(6):1565–1575.
34. Lange SJ, Alkhnbashi OS, Rose D, Will S, Backofen R: **CRISPRmap: an automated classification of repeat conservation in prokaryotic adaptive immune systems.** *Nucleic Acids Res* 2013, **41**(17):8034–8044.
35. Makarova KS, Haft DH, Barrangou R, Brouns SJ, Charpentier E, Horvath P, Moineau S, Mojica FJ, Wolf YI, Yakunin AF, van der Oost J, Koonin EV: **Evolution and classification of the CRISPR-Cas systems.** *Nat Rev Microbiol* 2011, **9**(6):467–477.
36. Stevenson DM, Weimer PJ: **Dominance of *Prevotella* and low abundance of classical ruminal bacterial species in the bovine rumen revealed by relative quantification real-time PCR.** *Appl Microbiol Biotechnol* 2007, **75**(1):165–174.
37. Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, Fernandes GR, Tap J, Bruls T, Batto JM, Bertalan M, Borruel N, Casellas F, Fernandez L, Gautier L, Hansen T, Hattori M, Hayashi T, Kleerebezem M, Kurokawa K, Leclerc M, Levenez F, Manichanh C, Nielsen HB, Nielsen T, Pons N, Poulain J, Qin J, Sicheritz-Ponten T, Tims S, *et al*: **Enterotypes of the human gut microbiome.** *Nature* 2011, **473**(7346):174–180.
38. Prangishvili D, Forterre P, Garrett RA: **Viruses of the Archaea: a unifying view.** *Nat Rev Microbiol* 2006, **4**(11):837–848.
39. Prangishvili D, Garrett RA, Koonin EV: **Evolutionary genomics of archaeal viruses: unique viral genomes in the third domain of life.** *Virus Res* 2006, **117**(1):52–67.
40. Pelve EA, Martens-Habbena W, Stahl DA, Bernander R: **Mapping of active replication origins *in vivo* in thaum-and euryarchaeal replicons.** *Mol Microbiol* 2013, **90**(3):538–550.
41. Raymann K, Forterre P, Brochier-Armanet C, Gribaldo S: **Global phylogenomic analysis disentangles the complex evolutionary history of DNA replication in Archaea.** *Genome Biol Evol* 2014, **6**(1):192–212.
42. Forterre P, Gribaldo S, Gadelle D, Serre M-C: **Origin and evolution of DNA topoisomerases.** *Biochimie* 2007, **89**(4):427–446.
43. Brochier-Armanet C, Forterre P, Gribaldo S: **Phylogeny and evolution of the Archaea: one hundred genomes later.** *Curr Opin Microbiol* 2011, **14**(3):274–281.
44. White MF, Bell SD: **Holding it together: chromatin in the Archaea.** *Trends Genet* 2002, **18**(12):621–626.
45. Makarova KS, Sorokin AV, Novichkov PS, Wolf YI, Koonin EV: **Clusters of orthologous genes for 41 archaeal genomes and implications for evolutionary genomics of archaea.** *Biol Direct* 2007, **2**:33.
46. Huynen MA, Dandekar T, Bork P: **Variation and evolution of the citric-acid cycle: a genomic perspective.** *Trends Microbiol* 1999, **7**(7):281–291.
47. Brown AM, Hoopes SL, White RH, Sarisky CA: **Purine biosynthesis in archaea: variations on a theme.** *Biol Direct* 2011, **6**(1):63.
48. Sakai S, Takaki Y, Shimamura S, Sekine M, Tajima T, Kosugi H, Ichikawa N, Tasumi E, Hiraki AT, Shimizu A, Kato Y, Nishiko R, Mori K, Fujita N, Imachi H, Takai K: **Genome sequence of a mesophilic hydrogenotrophic methanogen *Methanocella paludicola*, the first cultivated representative of the order Methanocellales.** *PLoS One* 2011, **6**(7):e22898.

49. Erkel C, Kube M, Reinhardt R, Liesack W: **Genome of Rice Cluster I archaea—the key methane producers in the rice rhizosphere.** *Science* 2006, **313**(5785):370–372.
50. Dos Santos PC, Fang Z, Mason SW, Setubal JC, Dixon R: **Distribution of nitrogen fixation and nitrogenase-like sequences amongst microbial genomes.** *BMC genomics* 2012, **13**:162.
51. Lai MC, Hong TY, Gunsalus RP: **Glycine betaine transport in the obligate halophilic archaeon *Methanohalophilus portucalensis*.** *J Bacteriol* 2000, **182**(17):5020–5024.
52. Roessler M, Pfluger K, Flach H, Lienard T, Gottschalk G, Muller V: **Identification of a salt-induced primary transporter for glycine betaine in the methanogen *Methanosarcina mazei* Go1.** *Appl Environ Microbiol* 2002, **68**(5):2133–2139.
53. Fricke WF, Seedorf H, Henne A, Krüer M, Liesegang H, Hedderich R, Gottschalk G, Thauer RK: **The genome sequence of *Methanosphaera stadtmanae* reveals why this human intestinal archaeon is restricted to methanol and H_2 for methane formation and ATP synthesis.** *J Bacteriol* 2006, **188**(2):642–658.
54. Samuel BS, Hansen EE, Manchester JK, Coutinho PM, Henrissat B, Fulton R, Latreille P, Kim K, Wilson RK, Gordon JI: **Genomic and metabolic adaptations of *Methanobrevibacter smithii* to the human gut.** *Proc Natl Acad Sci U S A* 2007, **104**(25):10643–10648.
55. Mittl PR, Schneider-Brachert W: **Sel1-like repeat proteins in signal transduction.** *Cell Signal* 2007, **19**(1):20–31.
56. Tallant TC, Paul L, Krzycki JA: **The MtsA subunit of the methylthiol: coenzyme M methyltransferase of *Methanosarcina barkeri* catalyses both half-reactions of corrinoid-dependent dimethylsulfide: coenzyme M methyl transfer.** *J Biol Chem* 2001, **276**(6):4485–4493.
57. Kaster A-K, Goenrich M, Seedorf H, Liesegang H, Wollherr A, Gottschalk G, Thauer RK: **More than 200 genes required for methane formation from H 2 and CO 2 and energy conservation are present in *Methanothermobacter marburgensis* and *Methanothermobacter thermautotrophicus*.** *Archaea* 2011, **2011**:973848.
58. Rouvière PE, Escalante-Semerena JC, Wolfe RS: **Component A2 of the methylcoenzyme M methylreductase system from *Methanobacterium thermoautotrophicum*.** *J Bacteriol* 1985, **162**(1):61–66.
59. Raymond J, Siefert JL, Staples CR, Blankenship RE: **The natural history of nitrogen fixation.** *Mol Biol Evol* 2004, **21**(3):541–554.
60. Shin DH: **Preliminary structural studies on the MtxX protein from *Methanococcus jannaschii*.** *Acta Crystallogr Sect F: Struct Biol Cryst Commun* 2008, **64**(4):300–303.
61. Graham DE, Taylor SM, Wolf RZ, Namboori SC: **Convergent evolution of coenzyme M biosynthesis in the Methanosarcinales: cysteate synthase evolved from an ancestral threonine synthase.** *Biochem J* 2009, **424**(3):467–478.
62. Graham DE, Graupner M, Xu H, White RH: **Identification of coenzyme M biosynthetic 2-phosphosulfolactate phosphatase: a member of a new class of Mg(2+)-dependent acid phosphatases.** *Eur J Biochem* 2001, **268**(19):5176–5188.
63. Graham DE, Xu H, White RH: **Identification of coenzyme M biosynthetic phosphosulfolactate synthase: a new family of sulfonate-biosynthesizing enzymes.** *J Biol Chem* 2002, **277**(16):13421–13429.
64. Graupner M, Xu H, White RH: **Identification of an archaeal 2-hydroxy acid dehydrogenase catalyzing reactions involved in coenzyme biosynthesis in methanoarchaea.** *J Bacteriol* 2000, **182**(13):3688–3692.
65. Schlegel K, Muller V: **Evolution of Na and H bioenergetics in methanogenic archaea.** *Biochem Soc T* 2013, **41**(1):421–426.
66. Anderson I, Ulrich LE, Lupa B, Susanti D, Porat I, Hooper SD, Lykidis A, Sieprawska-Lupa M, Dharmarajan L, Goltsman E: **Genomic characterization of methanomicrobiales reveals three classes of methanogens.** *PLoS One* 2009, **4**(6):e5797.
67. Kaster A-K, Moll J, Parey K, Thauer RK: **Coupling of ferredoxin and heterodisulfide reduction via electron bifurcation in hydrogenotrophic methanogenic archaea.** *Proc Natl Acad Sci U S A* 2011, **108**(7):2981–2986.
68. Thauer RK, Kaster A-K, Seedorf H, Buckel W, Hedderich R: **Methanogenic archaea: ecologically relevant differences in energy conservation.** *Nat Rev Microbiol* 2008, **6**(8):579–591.
69. Moparthi VK, Hägerhäll C: **The evolution of respiratory chain complex I from a smaller last common ancestor consisting of 11 protein subunits.** *J Mol Evol* 2011, **72**(5–6):484–497.
70. Bäumer S, Ide T, Jacobi C, Johann A, Gottschalk G, Deppenmeier U: **The $F420H_2$ dehydrogenase from *Methanosarcina mazei* is a redox-driven proton pump closely related to NADH dehydrogenases.** *J Biol Chem* 2000, **275**(24):17968–17973.
71. Welte C, Deppenmeier U: **Membrane-bound electron transport in *Methanosaeta thermophila*.** *J Bacteriol* 2011, **193**(11):2868–2870.
72. Tran QH, Bongaerts J, Vlad D, Unden G: **Requirement for the proton-pumping NADH dehydrogenase i of *Escherichia coli* in respiration of NADH to fumarate and its bioenergetic implications.** *Eur J Biochem* 1997, **244**(1):155–160.
73. Welte C, Krätzer C, Deppenmeier U: **Involvement of Ech hydrogenase in energy conservation of *Methanosarcina mazei*.** *FEBS J* 2010, **277**(16):3396–3403.
74. Meuer J, Kuettner HC, Zhang JK, Hedderich R, Metcalf WW: **Genetic analysis of the archaeon *Methanosarcina barkeri* Fusaro reveals a central role for Ech hydrogenase and ferredoxin in methanogenesis and carbon fixation.** *Proc Natl Acad Sci U S A* 2002, **99**(8):5632–5637.
75. Bäumer S, Lentes S, Gottschalk G, Deppenmeier U: **Identification and analysis of proton-translocating pyrophosphatases in the methanogenic archaeon *Methanosarcina mazei*.** *Archaea* 2002, **1**(1):1.
76. Baykov AA, Malinen AM, Luoto HH, Lahti R: **Pyrophosphate-fueled Na + and H + transport in prokaryotes.** *Microbiol Mol Biol R* 2013, **77**(2):267–276.
77. Prat L, Heinemann IU, Aerni HR, Rinehart J, O'Donoghue P, Soll D: **Carbon source-dependent expansion of the genetic code in bacteria.** *Proc Natl Acad Sci U S A* 2012, **109**(51):21070–21075.
78. Gaston MA, Jiang R, Krzycki JA: **Functional context, biosynthesis, and genetic encoding of pyrrolysine.** *Curr Opin Microbiol* 2011, **14**(3):342–349.
79. Heinemann IU, O'Donoghue P, Madinger C, Benner J, Randau L, Noren CJ, Soll D: **The appearance of pyrrolysine in tRNAHis guanylyltransferase by neutral evolution.** *Proc Natl Acad Sci U S A* 2009, **106**(50):21103–21108.
80. Krzycki JA: **Translation of UAG as Pyrrolysine.** In *Recoding: Expansion of Decoding Rules Enriches Gene Expression*. New York: Springer; 2010:53–77.
81. Longstaff DG, Blight SK, Zhang L, Green-Church KB, Krzycki JA: **In vivo contextual requirements for UAG translation as pyrrolysine.** *Mol Microbiol* 2007, **63**(1):229–241.
82. Veit K, Ehlers C, Schmitz RA: **Effects of nitrogen and carbon sources on transcription of soluble methyltransferases in *Methanosarcina mazei* strain Gö1.** *J Bacteriol* 2005, **187**(17):6147–6154.
83. Bailey S, Rycroft A, Elliott J: **Production of amines in equine cecal contents in an *in vitro* model of carbohydrate overload.** *J Anim Sci* 2002, **80**(10):2656–2662.
84. Smith E, Macfarlane G: **Studies on amine production in the human colon: enumeration of amine forming bacteria and physiological effects of carbohydrate and pH.** *Anaerobe* 1996, **2**(5):285–297.
85. Koeth RA, Wang Z, Levison BS, Buffa JA, Org E, Sheehy BT, Britt EB, Fu X, Wu Y, Li L: **Intestinal microbiota metabolism of l-carnitine, a nutrient in red meat, promotes atherosclerosis.** *Nat Med* 2013, **19**(5):576–585.
86. Mitchell AD, Chappell A, Knox K: **Metabolism of betaine in the ruminant.** *J Anim Sci* 1979, **49**(3):764–774.
87. Neill AR, Grime DW, Dawson R: **Conversion of choline methyl groups through trimethylamine into methane in the rumen.** *Biochem J* 1978, **170**:529–535.
88. Benstead J, King G, Williams H: **Methanol promotes atmospheric methane oxidation by methanotrophic cultures and soils.** *Appl Environ Microb* 1998, **64**(3):1091–1098.
89. Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, Fleischmann RD, Bult CJ, Kerlavage AR, Sutton G, Kelley JM: **The minimal gene complement of *Mycoplasma genitalium*.** *Science* 1995, **270**(5235):397–404.
90. Shigenobu S, Watanabe H, Hattori M, Sakaki Y, Ishikawa H: **Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera* sp.** *APS Nat* 2000, **407**(6800):81–86.
91. Waters E, Hohn MJ, Ahel I, Graham DE, Adams MD, Barnstead M, Beeson KY, Bibbs L, Bolanos R, Keller M: **The genome of *Nanoarchaeum equitans*: insights into early archaeal evolution and derived parasitism.** *Proc Natl Acad Sci U S A* 2003, **100**(22):12984–12988.
92. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, Meyer F, Olsen G, Olson R, Osterman A, Overbeek R, McNeil L, Paarmann D, Paczian T, Parrello B, Pusch G, Reich C, Stevens R, Vassieva O, Vonstein V, Wilke A, Zagnitko O: **The RAST Server: rapid annotations using subsystems technology.** *BMC Genomics* 2008, **9**:75.
93. Delcher AL, Bratke KA, Powers EC, Salzberg SL: **Identifying bacterial genes and endosymbiont DNA with Glimmer.** *Bioinformatics* 2007, **23**(6):673–679.
94. Bairoch A, Boeckmann B: **The SWISS-PROT protein sequence data bank.** *Nucleic Acids Res* 1991, **19**(Suppl):2247.

95. Gardner PP, Daub J, Tate JG, Nawrocki EP, Kolbe DL, Lindgreen S, Wilkinson AC, Finn RD, Griffiths-Jones S, Eddy SR: **Rfam: updates to the RNA families database.** *Nucleic Acids Res* 2009, **37**(suppl 1):D136–D140.
96. Schattner P, Brooks AN, Lowe TM: **The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs.** *Nucleic Acids Res* 2005, **33**(Web Server issue):W686–W689.
97. Taquist H, Cui Y, Ardell DH: **TFAM 1.0: an online tRNA function classifier.** *Nucleic Acids Res* 2007, **35**(Web Server issue):W350–W353.
98. Laslett D, Canback B: **ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences.** *Nucleic Acids Res* 2004, **32**(1):11–16.
99. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25**(17):3389–3402.
100. Grissa I, Vergnaud G, Pourcel C: **CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats.** *Nucleic Acids Res* 2007, **35**(Web Server issue):W52–W57.
101. Grissa I, Vergnaud G, Pourcel C: **CRISPRcompar: a website to compare clustered regularly interspaced short palindromic repeats.** *Nucleic Acids Res* 2008, **36**(Web Server issue):W145–W148.
102. Lange SJ, Alkhnbashi OS, Rose D, Will S, Backofen R: **CRISPRmap: an automated classification of repeat conservation in prokaryotic adaptive immune systems.** *Nucleic Acids Res* 2013, **41**(17):8034–8044.
103. Zhou Y, Liang Y, Lynch KH, Dennis JJ, Wishart DS: **PHAST: a fast phage search tool.** *Nucleic Acids Res* 2011, **39**(Web Server issue):W347–W352.
104. Grant JR, Stothard P: **The CGView Server: a comparative genomics tool for circular genomes.** *Nucleic Acids Res* 2008, **36**(suppl 2):W181–W184.
105. Oliveros J: **VENNY: an interactive tool for comparing lists with Venn Diagrams.** 2007, http://bioinfogp.cnb.csic.es/tools/venny/index.html.
106. Magrane M: **UniProt Knowledgebase: a hub of integrated protein data.** *Database* 2011, **2011**:bar009.
107. Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M: **KAAS: an automatic genome annotation and pathway reconstruction server.** *Nucleic Acids Res* 2007, **35**(suppl 2):W182–W185.
108. Li H, Benedito V, Udvardi M, Zhao P: **TransportTP: a two-phase classification approach for membrane transporter prediction and characterization.** *BMC Bioinform* 2009, **10**(1):418.
109. Ren Q, Chen K, Paulsen IT: **TransportDB: a comprehensive database resource for cytoplasmic membrane transport systems and outer membrane channels.** *Nucleic Acids Res* 2007, **35**(suppl 1):D274–D279.
110. Götz S, García-Gómez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, Robles M, Talón M, Dopazo J, Conesa A: **High-throughput functional annotation and data mining with the Blast2GO suite.** *Nucleic Acids Res* 2008, **36**(10):3420–3435.
111. Johnson LS, Eddy S, Portugaly E: **Hidden Markov model speed heuristic and iterative HMM search procedure.** *BMC Bioinform* 2010, **11**(1):431.
112. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput.** *Nucleic Acids Res* 2004, **32**(5):1792–1797.
113. Philippe H: **MUST, a computer package of management utilities for sequences and trees.** *Nucleic Acids Res* 1993, **21**(22):5264–5272.
114. Criscuolo A, Gribaldo S: **BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments.** *BMC Evol Biol* 2010, **10**(1):210.
115. Stamatakis A: **RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models.** *Bioinformatics* 2006, **22**(21):2688–2690.
116. Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP: **MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space.** *Syst Biol* 2012, **61**(3):539–542.
117. Hofacker IL: **Vienna RNA secondary structure server.** *Nucleic Acids Res* 2003, **31**(13):3429–3431.

doi:10.1186/1471-2164-15-679
Cite this article as: Borrel *et al.*: Comparative genomics highlights the unique biology of Methanomassiliicoccales, a Thermoplasmatales-related seventh order of methanogenic archaea that encodes pyrrolysine. *BMC Genomics* 2014 **15**:679.

Collaboration 2

Phylogenetic placement of the first SM-1 Euryarchaeon representative

As mentioned in the Introduction section, the SM1 Euryarchaeon forms an almost single species biofilm in a cold, sulfidic spring and has been for long time uncultured and its phylogenetic affiliation unclear. The group of Christine Moissl-Eichinger at Marburg, Germany, has recently obtained the first enrichment and genomic data for this interesting new archaeal lineage and contacted us to perform a phylogenomic analysis to understand the phylogenetic position of the SM-1 lineage. I performed an extensive phylogenetic analysis of the SM-1 lineage (3 assembled genomes) and the close relative IM-4 (shotgun sequence) based on concatenated ribosomal proteins and my dataset of DNA replication markers. Additionally I annotated and analyzed the DNA replication components in the assembled SM-1 genome. Although its precise placement in the archaeal phylogeny is difficult to assess because of its high evolutionary rates, I could establish that SM-1 is a fast evolving euryarchaeal lineage likely close to Methanococcales. This helped discussing the peculiar characteristics of its unique metabolism potentially derived from a methanogenic ancestor. Altogether, comparative genomics and phylogenetic analysis revealed that SM-1 represents a diverse and widespread novel euryarchaeal order ("Candidatus Altiarchaeales") that can dominate subsurface biotopes.

This analysis has just been resubmitted after revision to Nature Communications.

Collaboration Article 2

Collaboration 3

Phylogenetic analysis of a novel family of DNA topoisomerases

DNA topoisomerases are essential for genome replication and integrity. They are split into two types, type I that introduce single strand breaks into catenated DNA, and type II that introduces double-strand breaks. Of the two types, type II is ubiquitous in all cellular organisms and can be classified in two subfamilies, IIA and IIB based on sequence and structural similarities. In this article a new subfamily of type IIB topoisomerase is characterized, which is present in only a few bacterial genomes, and plasmids, and is called Topoisomerase VIII. Three of these enzymes were characterized and one of them found to exhibit exonuclease activity and relaxation activity, although at low levels. The low activity of these enzymes is likely due to the fact that they are located on decaying mobile elements. These small topoisomerases have a unique evolutionary history, and the discovery of a subfamily of topoisomerases specifically encoded by plasmids confirms that mobile elements are potential reservoirs of novel proteins involved in DNA metabolism. I contributed to this study by performing phylogenomic analysis on these novel topoisomerases. We identified homologs in only three bacterial phyla, the Firmicutes, Bacteroidetes, and Proteobacteria (alpha and gamma), and on one archaeal and two bacterial plasmids. In bacterial genomes Topoisomerase VIII is located within integrated mobile elements likely derived from conjugative plasmids. Those present in evolutionary close genera are grouped together in the phylogenetic tree, and the Topo VIII encoded by plasmids group with the Topo VIII enzymes from species that are closely related to the hosts. Moreover, some species that possess Topo VIII lack the typical Topo IV present in bacterial genomes, suggesting that Topo VIII might compensate for its absence in these species.

These results were published in Nucleic Acids Research July 3, 2014 and recieved a recommendation on F1000 Prime.

Collaboration Article 3

8578–8591 *Nucleic Acids Research, 2014, Vol. 42, No. 13*
doi: 10.1093/nar/gku568

DNA topoisomerase VIII: a novel subfamily of type IIB topoisomerases encoded by free or integrated plasmids in Archaea and Bacteria

Danièle Gadelle[1], Mart Krupovic[2], Kasie Raymann[2], Claudine Mayer[3,4,5] and Patrick Forterre[1,2,*]

[1]Université Paris-Sud, CNRS UMR8621, Institut de Génétique Microbiologie, 91405 Orsay Cedex, France, [2]Institut Pasteur, Unité de Biologie moléculaire du gène chez les extrêmophiles, Département de Microbiologie, F-75015 Paris, France, [3]Institut Pasteur, Unité de Microbiologie structurale, Département de Biologie structurale et Chimie, F-75015 Paris, France, [4]CNRS, UMR3528, F-75015 Paris, France and [5]Université Paris Diderot, Sorbonne Paris Cité, Cellule Pasteur, rue du Dr Roux 75015 Paris, France

Received February 4, 2014; Revised June 10, 2014; Accepted June 11, 2014

ABSTRACT

Type II DNA topoisomerases are divided into two families, IIA and IIB. Types IIA and IIB enzymes share homologous B subunits encompassing the ATP-binding site, but have non-homologous A subunits catalyzing DNA cleavage. Type IIA topoisomerases are ubiquitous in Bacteria and Eukarya, whereas members of the IIB family are mostly present in Archaea and plants. Here, we report the detection of genes encoding type IIB enzymes in which the A and B subunits are fused into a single polypeptide. These proteins are encoded in several bacterial genomes, two bacterial plasmids and one archaeal plasmid. They form a monophyletic group that is very divergent from archaeal and eukaryotic type IIB enzymes (DNA topoisomerase VI). We propose to classify them into a new subfamily, denoted DNA topoisomerase VIII. Bacterial genes encoding a topoisomerase VIII are present within integrated mobile elements, most likely derived from conjugative plasmids. Purified topoisomerase VIII encoded by the plasmid pPPM1a from *Paenibacillus polymyxa* M1 had ATP-dependent relaxation and decatenation activities. In contrast, the enzyme encoded by mobile elements integrated into the genome of *Ammonifex degensii* exhibited DNA cleavage activity producing a full-length linear plasmid and that from *Microscilla marina* exhibited ATP-independent relaxation activity.Topoisomerases VIII, the smallest known type IIB enzymes, could be new promising models for structural and mechanistic studies.

INTRODUCTION

DNA topoisomerases are essential for solving topological problems arising during DNA metabolic processes and are therefore critical for the preservation of genome stability (1–3). They can interconvert different topological forms of DNA and either catenate or decatenate DNA rings by generating transient single- (type I topoisomerases) or double-stranded (type II topoisomerases) DNA breaks in DNA backbones. Type II enzymes are especially important because of their unique capacity to catalyze the transfer of one DNA duplex through another. They are essential for resolving catenanes generated between daughter chromosomes at the end of DNA replication and also play a major role in relaxing positive supercoils generated by tracking processes occurring during transcription and replication. Accordingly, all cellular organisms, without exception, have at least one type II enzyme.

Type II topoisomerases are classified into two families, IIA and IIB, based on sequence and structural similarities (4–7). Archaeal and bacterial types IIA and IIB enzymes are heterotetramers composed of two different subunits (A and B), whereas type IIA enzymes from eukaryotes and their viruses are homodimers, with the B and A moieties fused into a single polypeptide. Several high-resolution structures of both types IIA and IIB enzymes have been solved, highlighting some structural similarities, but also large differences between these two families (8–11). The B subunits are homologous in the two families and contain a similar ATP-binding site located within a protein domain known as the Bergerat fold, which is characteristic of proteins of the GHKL (Gyrase, Hsp90, histidine Kinase, MutL) superfamily (4,12). Moreover, enzymes from the two families share two other functional domains: the Toprim domain displaying the Rossmann fold (involved in magnesium binding) and

*To whom correspondence should be addressed. Tel: +3369156445; Fax: +3369157808; Email: patrick.forterre@pasteur.fr

the Winged Helix Domain (WHD), which contains the active site tyrosine. However, the Toprim domain of type IIA topoisomerases is located within the B subunit, whereas in the IIB family, this domain is located within the A subunit. Furthermore, their WHDs are unrelated, despite the presence of the catalytic tyrosine in both cases (Figure 1). Overall, the A subunits of types IIA and IIB topoisomerases share neither sequence nor structural similarity. This suggests that these two enzyme families originated independently via the association of homologous B subunits with non homologous A subunits (5).

The origin and evolution of topoisomerases are rather puzzling. Ribosomal proteins exist in three distinct versions, corresponding to the three cellular domains, Archaea, Bacteria and Eukarya (13,14); however, each topoisomerase family exists in several versions (hereafter called subfamilies) that do not overlap with the three domains in phylogenomic analyses (5–7,15). In addition, several groups of viruses encode topoisomerases that are not related to those of their hosts but instead form monophyletic groups that branch in between cellular domains in phylogenetic analyses (5,7,15).

Several distinct subfamilies of type IIA enzymes are known, which differ in catalytic activity and/or quaternary structure. These include (i) DNA gyrases, present in all bacteria, in some members of the archaeal phylum Euryarchaeota and in eukaryotes with endosymbionts of cyanobacterial origin; (ii) topoisomerase IV enzymes, specific to bacteria; (iii) eukaryotic type II topoisomerases, present in all eukaryotic species; (iv) viral type II topoisomerases encoded by some Nucleocytoplasmic Large DNA Viruses (NCLDV) that infect eukaryotes; and (v) viral type II topoisomerases encoded by T4-like myoviruses that infect bacteria.

The distribution of type IIB topoisomerases is more limited than that of type IIA. This family includes archaeal type IIB enzymes, also called topoisomerase VI, with close homologs in Archaeplastida (plants, red and green algae), some protists and a few bacteria (7,15). Topoisomerases VI are relaxing enzymes devoid of gyrase activity (16), are the only type II topoisomerases present in most archaea, suggesting that they are critical for chromosome segregation. Topoisomerases VI are also the only enzymes capable of relaxing positive supercoils in two of the three major archaeal phyla (17). Thus, these enzymes are also essential for managing the waves of positive supercoiling generated during DNA replication and transcription.

In Eukarya, type IIA topoisomerases control topological stress generated by chromosome segregation, replication and transcription. However in plants, type IIB topoisomerases (close relatives of archaeal topoisomerase VI) control endoreduplication, a polyploidization process responsible for the enlargement of plant cells, which in turn determines plant size (18).

The inconsistent phylogenomic pattern of topoisomerases in the three domains of life is common to most proteins involved in DNA replication, recombination or repair (for the case of DNA polymerases, see (19)). Lateral gene transfers cannot account for the distribution of these enzymes within the classical tree of life based on the ribosome. Instead, this pattern requires more elaborate evolutionary scenarios. We proposed previously that proteins involved in DNA metabolic activities (including topoisomerases) originated first in the viral world and were later transferred 'randomly' to various cellular lineages (20). This scenario predicts that some contemporary viruses or related mobile elements (plasmids, transposons) still encode unique versions of DNA replication proteins, such as topoisomerases, that were never transferred to the 'cellular world' (except as integrated mobile elements) (7).

Here, we report the *in silico* discovery and preliminary biochemical characterization of a new DNA topoisomerase subfamily that supports the prediction outlined above. The genes encoding these proteins correspond to the fusion of the B and A subunits of type IIB topoisomerases. We identified these genes in 19 bacterial genomes and in three plasmids (two from bacteria and one from an archaeon). In bacterial genomes, these genes are located within integrated mobile elements related to conjugative plasmids. We purified three members of this new subfamily of type IIB enzymes (that we propose to call topoisomerase VIII): one from a mesophilic bacterium, one from a thermophilic bacterium and one from a bacterial plasmid. We show that the enzyme encoded by the bacterial plasmid exhibits ATP-dependent relaxation and decatenation activities, the hallmark of type II topoisomerases, whereas the other two enzymes, encoded by integrated elements, exhibit DNA cleavage activity, producing DNA double-stranded breaks, and/or ATP-independent relaxation activity.

MATERIALS AND METHODS

Plasmids and reagents

Negatively supercoiled plasmid pBR322 DNA was purchased from Fermentas life science and kDNA (kinetoplast DNA) from Topogen. The reverse-gyrase enzyme was kindly given by Prof. Marc Nadal. Primers for site-directed mutagenesis were synthesized by Sigma-Aldrich.

Sequence database search

The fusion of the topoisomerase VI protein subunits A and B was used as a query to carry out psi-BLAST and tBLASTn searches of different databases on the NCBI website (http://blast.ncbi.nlm.nih.gov/Blast.cgi). A multiple sequence alignment was obtained with PipeAlign (21) and manually corrected based on conserved motifs. Secondary structures predictions were performed with PSI-PRED (22).

Phylogenetic analysis

For phylogenetic analysis, 27 identified homologs of topoisomerases VIII were aligned with MUSCLE 3.8.31 (23) and manually inspected with Seaview 4.3.3 (24). The alignment was trimmed with BMGE software (25) and the BLOSUM30 matrix, resulting in 571 amino acid positions. ProtTest 3 (26) was used to determine the best substitution model for further analysis with the Akaike Information Criterion (AICc) criteria. Maximum likelihood analysis was preformed with RaxML 7.4.2 (27) and the combined tree-search/fast bootstrap method ('-f a') under the

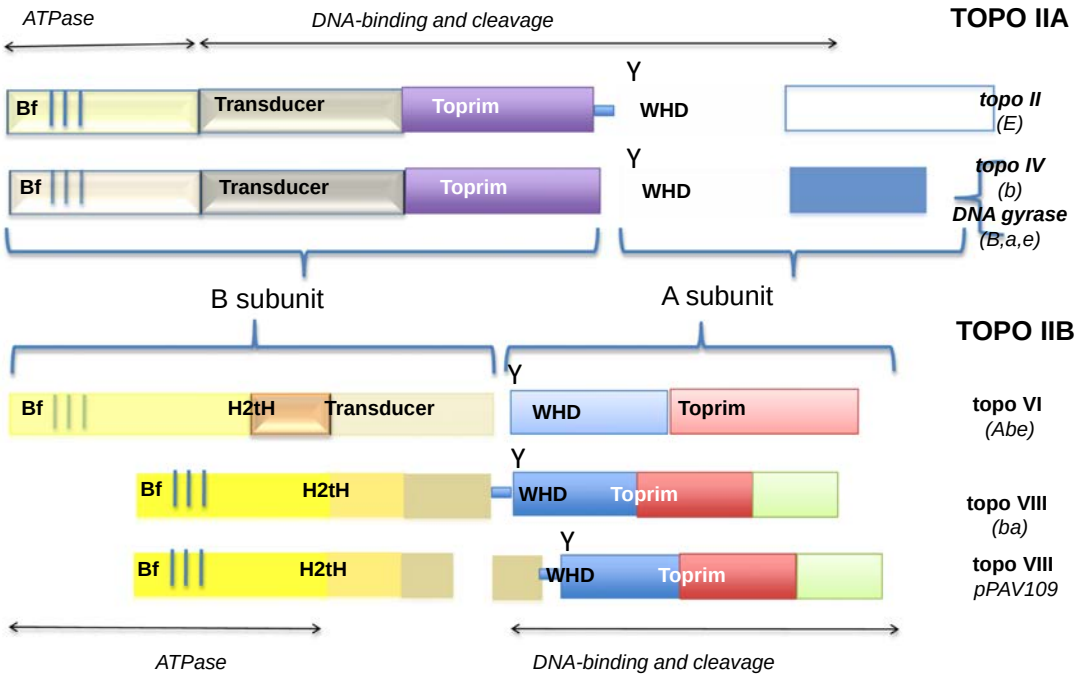


Figure 1. Domain organization of type II topoisomerases of the A and B families. Homologous domains are colored similarly, with color intensities differing according to the extent of sequence similarity. Bf: Bergerat fold (4) corresponding to the ATP-binding site with three conserved amino-acid signatures (small vertical bars), H2TH: Helix-2Turn-Helix domain, WHD: Winged Helix Domain; Y: catalytic tyrosine; triangle: catalytic site. The distributions of various type II topoisomerase subfamilies in the three domains of life, Archaea (A, a), Bacteria (B, b) and Eukarya (E, e) are indicated in brackets.

PROTGAMMALG model (four discrete rate categories) with 1000 fast bootstraps. Bayesian analysis was performed with MrBayes 3.0 (28) and the mixed amino acid model with four discrete rate categories, which supported the WAG model with 100% posterior probability. Four Markov chains starting with a random tree were run simultaneously for 1 million generations, sampling trees at every 100th generation. The first 2500 sampled trees (25%) were discarded as 'burn in'.

Protein structure predictions

Structural predictions were performed with the Phyre2 webserver (29). The N- and C-terminal halves of the three topoisomerase VIII sequences (corresponding to the B and A topoisomerase VI subunits, respectively) were also modeled separately because conserved regions and motifs are scattered all along their alignment. The model was validated by SWISS-MODEL (30). Comparisons between different topoisomerases models were conducted with the PyMOL Molecular Graphics System (http://www.pymol.org).

Analysis of the genetic context and DotPlot analysis

The genomic context of identified topoisomerase VIII genes was analyzed with CLC Genomic Workbench as described previously (31). The topoisomerase VIII gene-containing region was considered to be potentially mobile when two criteria were satisfied: (i) the integrase-coding gene was present in proximity of the topoisomerase VIII gene; and (ii) the region containing both topoisomerase VIII- and integrase-coding genes was flanked by direct repeats (attachment [att] sites). Gepard (32) was used to generate dotplots for the analysis of large mobile genomic regions.

Recombinant protein expression and purification

The gene encoding the topoisomerase VIII from *Ammonifex degensii* (gene ID: 646359367), *Microscilla marina* ATCC 23134 (gene ID: 640218028) and *Paenibacillus polymyxa* M1 (gene ID: 2518132008) were synthesized by GenScript@. The synthetic genes of *A. degensii* and *M. marina* were cloned into a pUC57 vector, and amplified by PCR with primers containing Strep-Tag, NdeI and NotI restriction sites. The PCR products with the Strep-tags either at the 5′ or 3′ ends of the genes were cloned into a pET26b expression vector and the sequences of recombinant topoisomerase VIII clones were confirmed by DNA sequencing. One of each of the recombinant plasmids: MicNStrep.pET26b::topoisomerase, MicCStrep.pET26b::topoisomerase, AdegN-Strep.pET26b::topoisomerase and AdegC-Strep.pET26b::topoisomerase was used to transform *Escherichia coli* BL21 (DE3) strains (Novagen). The synthetic gene of *P. polymyxa* topoisomerase VIII was purchased directly with an N-terminal Strep-Tag. Bacteria were transformed with the various constructs and were grown subsequently in 1–4 l of Luria Broth (LB) medium containing kanamycin (50 μg/ml), at 16°C overnight, after a heat shock at 42°C at the beginning of growth phase ($OD_{600\,nm}$ of 0.1). Induction was carried out with 0.5 mM isopropyl-D-1-thiogalactopyranoside (IPTG) when cell

cultures reached an $OD_{600\,nm}$ of 0.5. Cells were harvested by centrifugation, stored overnight at −80°C, and then suspended in a Tris–HCl 40 mM pH 8.0, NaCl 200–1000 mM, Dirhiotreitol (DTT) 1 mM, acide éthylène diamine tétraacétique (EDTA) 0.1 mM buffer containing protease inhibitors. Cell lysis was completed by sonication. The cell extracts were then centrifuged at 10 000 × *g* for 15 min at 4°C to remove cellular debris and aggregated proteins. Strep-tagged proteins from the soluble fraction were purified by gravity-flow chromatography on a Strep-Tactin column (IBA BioTAGnology) according to the manufacturer's recommendation and the eluted proteins were then run on a gel filtration column (Superdex™ 200 16/600, GE Healthcare) with an FPLC AKTA system. The column was equilibrated with buffer B (buffer A containing 10% (v/v) ethylene glycol or glycerol) and the protein was eluted in the same buffer. Topoisomerase VIII enzymes were detected by checking for the presence of a polypeptide with the expected size on an Sodium dodecyl sulfate (SDS)-polyacrylamide gel. Fractions containing topoisomerase VIII polypeptides were pooled and concentrated with Amicon 30 kDa cutoff concentrators (Millipore). Protein concentration was determined and the fractions were aliquoted and stored at −80°C.

In vitro mutagenesis

Plasmids bearing *P. polymyxa* and *M. marina* topoisomerase VIII mutant genes were generated with the QuikChange site-directed mutagenesis kit (Stratagene). After mutagenesis, plasmids were purified with Macherey Nagel minipreps kit and sequenced to ensure the absence of unwanted mutations.

Positively supercoiled pBR322 preparation

Positively supercoiled pBR322 was prepared by incubation of pBR322 plasmid for 30 min at 90°C in the presence of *Sulfolobus solfataricus* reverse-gyrase (RG1) in buffer containing 50 mM Tris–HCl [pH 8.0], 20 mM $MgCl_2$, 1 mM adenosine triphosphate (ATP) and 1 mM DTT (33). The reaction was stopped with 10 mM EDTA and 200 mM NaCl and the positively supercoiled plasmid was purified with the Macherey-Magel NucleoSpin Gel and PCR Clean-up kit.

DNA relaxation and decatenation assays

Relaxation assays (per 20 μl) were performed with negatively or positively supercoiled pBR322 plasmid or kDNA (200 ng) and the indicated amount of enzyme in buffer containing 50 mM Tris [pH 8.0], 2.5 mM $MgCl_2$, 0.1 mM EDTA and ATP (as indicated in the figure legends) for 40 min at 20–30°C or 10 min at 70°C depending on the topoisomerase host. The reaction also contained a final concentration of 0.30 M NaCl and ≈19% glycerol from the protein storage buffer. In some cases at the end of the reaction, an additional incubation of 30 min at 55°C, was done after addition of 2 μl sodium dodecyl sulfate (SDS) 10% and 2 μl proteinase K (1 mg/ml) to check for stabilized cleavable complexes.

Reactions were stopped with 2 μl 1(10% SDS):1(30% glycerol) stop buffer. The reactions were loaded on a 1% agarose gel in 0.5× Tris-Borate-EDTA (TBE) buffer (45 mM Tris–borate [pH 8.3], 1 mM EDTA). The gels were run at 50 V/cm for 4 h, stained with ethidium bromide (EtBr) and visualized with an imaging system.

RESULTS

Identification of DNA topoisomerase VIII, a new subfamily of type IIB topoisomerase

We routinely screen the NCBI non-redundant protein and environmental sequence databases by psi-BLAST (34) to search for new type IIB topoisomerase-encoding genes. During the course of this work, we used the B subunit of topoisomerase VI from particular archaea (e.g. *Sulfolobus shibatae*) as a query. This led to the detection of a divergent version of type IIB topoisomerases in which the Bergerat fold (ATP-binding site) and the Toprim domain were present in the same polypeptide. We used these unusual proteins as queries to carry out additional Psi-Blast searches. This retrieved a set of closely related proteins of similar sizes corresponding to the topoisomerase VI subunits B and A fused, in that order, into a single polypeptide (Figure 1). After two iterations, we found that genes encoding these atypical type IIB topoisomerases are present in 11 complete and 8 partial bacterial genomes (Supplementary Table S1). In addition, we found two genes encoding these proteins on plasmids (Supplementary Table S1): one in the halophilic archaeon *Halalkalicoccus jeotgali* B3 (plasmid 5) and the other in the firmicute *P. polymyxa* M1 (pPPM1a). We also identified a very similar protein encoded by two genes (separated by 41 bp) located on a plasmid in *Paenibacillus alvei* DSM 29 (pPAV109): one gene encodes a protein homologous to the N-terminal moiety of the topoisomerase VI B subunit, whereas the other encodes a protein homologous to the C-terminal moiety of the B subunit fused to a domain homologous to the topoisomerase VI subunit A (Figure 1). Moreover, we also detected small DNA fragments encoding these proteins in several whole genome shotgun and environmental sequence databases (not shown).

These atypical type IIB topoisomerases, with sizes between 695 and 882 amino acids, are annotated either as topoisomerase VI, hypothetical proteins, ATP-binding proteins, Toprim or topoisomerase (ATP hydrolyzing). They harbor the four conserved regions important for type II topoisomerase function, i.e. the Bergerat fold (ATP-binding site), the transducer domain, the WHD and the Toprim domains, as well as the small H2TH domain specific of the type IIB family (Figures 1 and 2). The putative active site tyrosine was adjacent to either another tyrosine (YY) or a phenylalanine (FY), with the exception of the enzymes from *Roseobacter* sp. AzwK-3b and *Rhizobium gallicum* in which the tyrosine is preceded by a methionine (MY). This can be considered as a hallmark of the type IIB topoisomerase family, because topoisomerase VI and Spo11 proteins always have YY or FY in their active sites. In contrast, this feature is usually not observed in the type IIA family. Moreover, the identified proteins displayed only a remote relationship with topoisomerase VI, and their similarity was restricted to specific amino-acid signatures within the conserved motifs (Figure 2 and Supplementary Figure S1). Overall, the identified group of enzymes is not more

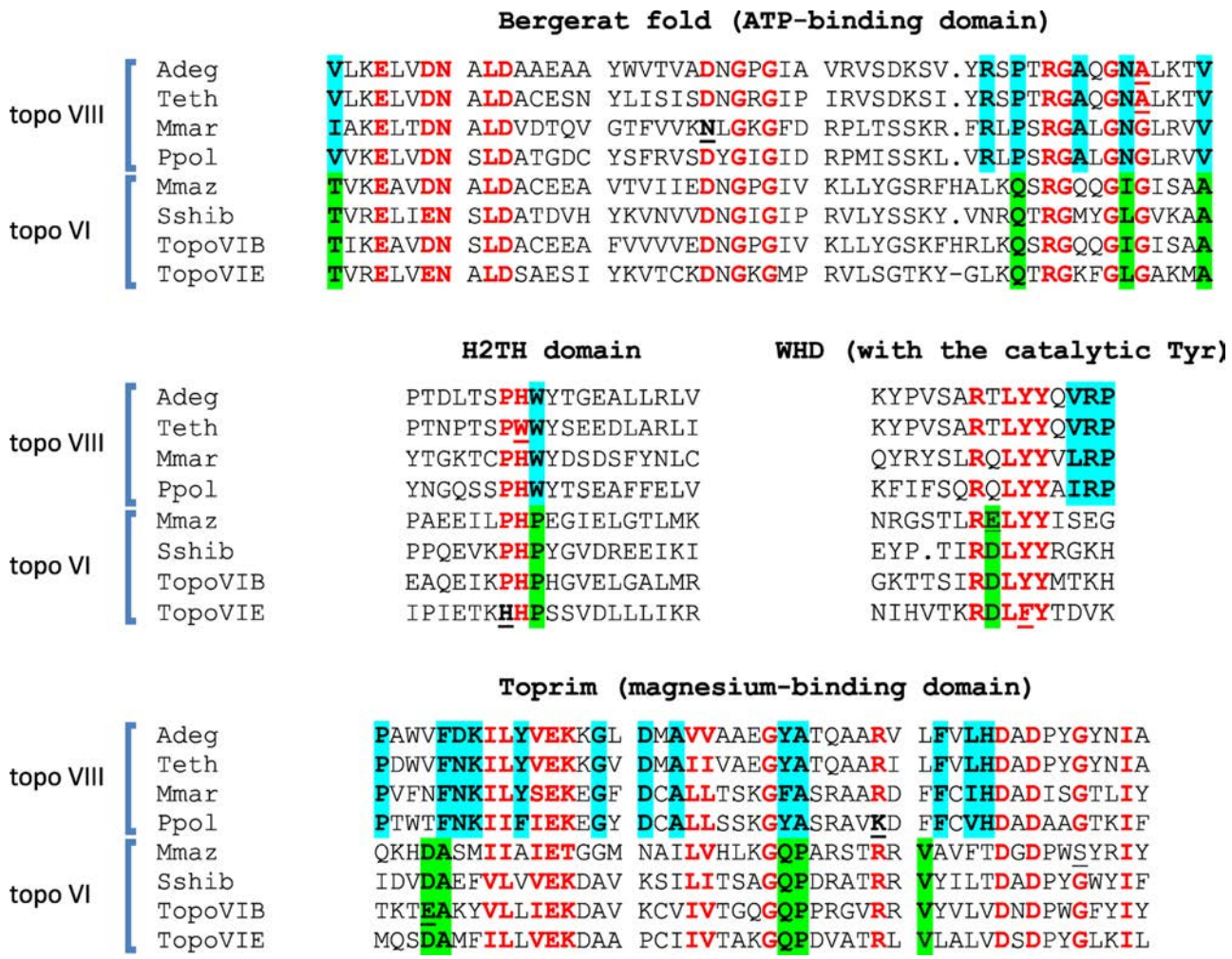


Figure 2. Conserved amino-acid regions shared between topoisomerases VI and VIII enzymes. In red, amino acids shared between the two families; in blue, amino acids specific for the topoisomerase VIII sub-family; in green, amino acids specific for the topoisomerase VI sub-family. Underlined are alternative amino acids rarely found in otherwise strictly conserved motifs. Adeg*: Ammonifex degensii* topoisomerase VIII, Teth: *Thermoanaerobacter ethanolicus* topoisomerase VIII, Mmar*: Microscilla marina* topoisomerase VIII; pPol: *Paenibacillus polymyxa* M1 plasmid pPPM1a topoisomerase VIII; Sshi: *Sulfolobus shibatae* topoisomerase VI; Mmaz: *Methanosarcina mazei* topoisomerase VI; Anae: *Anaeromyxobacter* sp. topoisomerase VI; Atha: *Arabidopsis thaliana* topoisomerase VI. Adeg, Teth, Mmar, Anae *and P. polymyxa* are Bacteria, Sshi and Mmaz are Archaea, Atha is a eukaryote.

closely related to *bona fide* bacterial topoisomerase VI than it is to archaeal or eukaryal topoisomerase VI. In contrast, these atypical type IIB enzymes exhibit extensive sequence similarity with each other, throughout their entire sequence. Furthermore, these proteins share a specific amino-acid signature, 'R V/I E L N A/S M', in their C-terminal region (Supplementary Figure S1), which is not found in topoisomerases from the currently established families and subfamilies.

All these observations strongly suggest the ancient divergence of these enzymes from topoisomerase VI, as opposed to recent independent fusions of topoisomerase VI subunits A and B in various bacterial lineages. Based on these observations as well as the data presented below, we propose that these proteins should be considered as a distinct subfamily of type IIB enzymes named DNA topoisomerase VIII. This classification is consistent with the historical nomenclature of topoisomerases in which subfamilies of type I enzymes have been systematically given odd numbers (topoisomerases I, III and V) and type II enzymes even numbers (topoisomerases II, IV, VI and VIII) (3).

Topoisomerase VIII is present in three bacterial phyla: Firmicutes (Clostridia and Bacilli), Bacteroidetes (Sphingobacteria) and Proteobacteria (alpha and gamma). Notably, those present in evolutionarily close genera are grouped together in phylogenetic analysis, whereas topoisomerase VIII enzymes from various genera of Proteobacteria and Firmicutes sometimes display mixed phylogenetic patterns (Figure 3). The three topoisomerase VIII enzymes encoded by plasmids are grouped with cellular topoisomerase VIII enzymes from species that are closely related to the hosts of these plasmids. For example, topoisomerase VIII enzymes encoded by pPPM1a and pPAV109 from *P. polymyxa* and *P. alvei*, respectively, are grouped with the

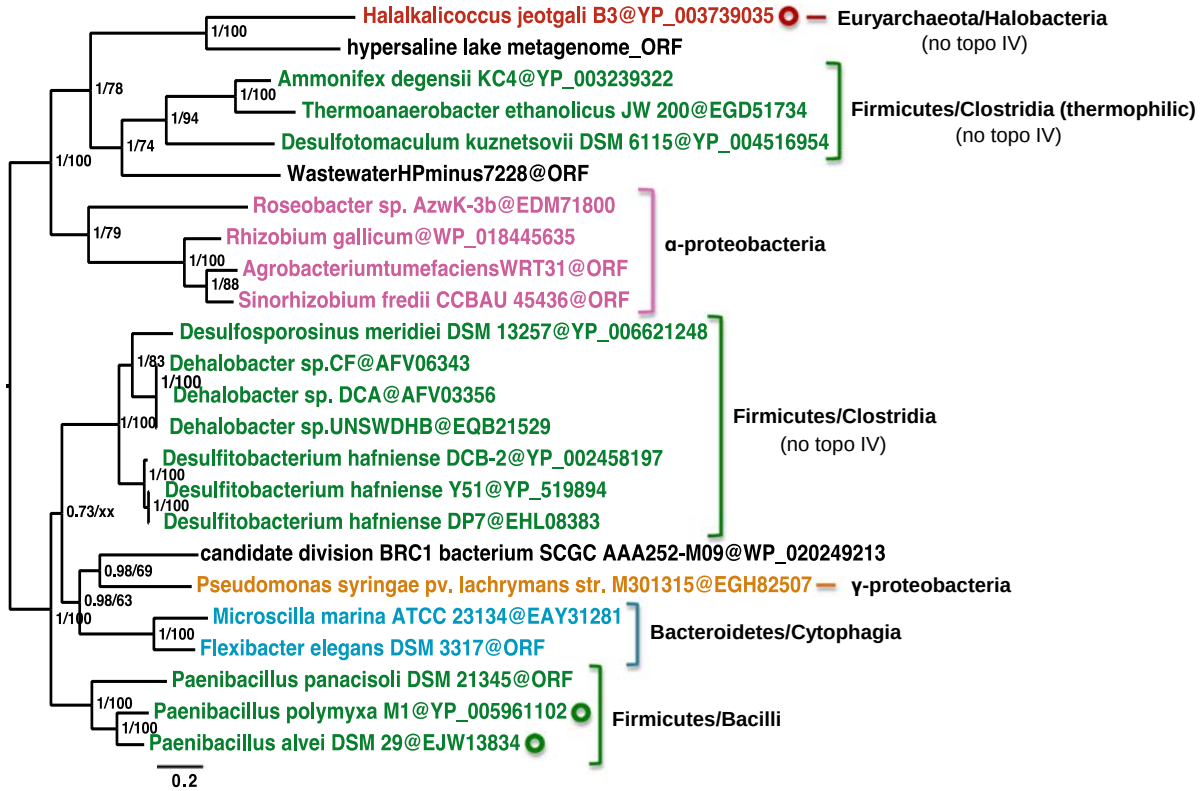


Figure 3. Phylogeny of topoisomerase VIII. Bayesian phylogeny of the 24 topoisomerase VIII homologs listed in Supplementary Table S1 and two topoisomerase VIII enzymes detected in metagenomes (571 amino acid positions). The tree was calculated with MrBayes (MIX model + gamma4). The scale bar shows the average number of substitutions per site. Values at nodes are posterior probabilities and bootstrap values calculated with the rapid bootstrap feature of RAxML (LG + gamma4) from heuristic searches of 1000 resampled datasets, when the same node was recovered. Topoisomerase VIII enzymes encoded by a plasmid are marked by circles.

chromosomally-encoded topoisomerase VIII from *Paenibacillus panacisoli*. Similarly, the topoisomerase VIII encoded by the plasmid from a haloarchaeon clusters with a sequence present in a metagenome from a hypersaline lake (Figure 3). These observations are consistent with previous studies that have shown that mobile elements often co-evolve with their hosts (35,36).

Most bacteria carry two type IIA topoisomerases, DNA gyrase and DNA topoisomerase IV, which have specialized functions (3). However, there are cases (e.g. *Thermotoga maritima* and *Mycobacterium tuberculosis*) where one dual-function enzyme is sufficient (3). Therefore, we considered the possibility that topoisomerase VIII may compensate for the absence of topoisomerase IV in some species. However, we found no correlation between the presence of topoisomerase VIII and the presence/absence of topoisomerase IV in any of the species analyzed (Figure 3).

Topoisomerase VIII genes present in bacterial genomes are located within integrated conjugative plasmids

The presence of topoisomerase VIII-encoding genes on three different plasmids and the sporadic distribution of these genes in taxonomically distant bacteria suggested that the presumably 'cellular' (i.e. encoded on bacterial chromosomes) gene copies are, in fact, be carried by integrated mobile genetic elements (37). Indeed, our previous analysis showed that integrated mobile elements often carry genes encoding proteins involved in DNA metabolism (35,38). We performed a detailed analysis of the genetic context of the 'cellular' topoisomerase VIII-encoding genes to verify this hypothesis. Unfortunately, 12 of the 19 topoisomerase VIII-encoding bacterial genomes were only available as WGS (whole genome shotgun) libraries consisting of genomic contigs of variable lengths, complicating their analysis. Nonetheless, we were able to obtain evidence suggesting that at least 10 of the seemingly cellular topoisomerase VIII-encoding genes are encoded by mobile elements (Supplementary Table S1).

In several cases, the mobile elements could be detected via the identification of their integration sites. During recombination mediated by the element-encoded integrase, the target sequence (attachment site, attB) is duplicated, and flanks the element from both ends as direct repeats (attL and attR). One of the repeats is typically found next to a gene encoding the recombinase. We identified the precise targets of integration for seven of the elements. Two elements, AmmDeg-E1 (in *A. degensii* KC4) and RosAzw-E1 (in *Roseobacter* sp. AzwK-3b), are integrated in intergenic regions, whereas PaePan-E1 (in *P. panacisoli* DSM21345) and SinFre-E1 (in *Sinorhizobium fredii* CCBAU 45436) recombined with 3′-distal regions of protein-coding genes (Supplementary Ta-

ble S1). In addition, elements AgrTum-E1 (in *Agrobacterium tumefaciens* WRT31), DesKuz-E1 (in *Desulfotomaculum kuznetsovii* DSM 6115) and RhiGal-E1 (in *R. gallicum* bv. *gallicum* R602sp) are integrated into tRNA-Thr, tRNA-Pro and tRNA-Cys genes, respectively (Figure 4).

A search for putative attachment sites within topoisomerase VIII-encoding genomes did not result in the identification of additional integrated elements. This may be because some genomic contigs containing the integrated element were incomplete, the elements used another mechanism for integration, or the elements were in an advanced stage of decay, which would render the direct repeats unrecognizable due to mutations. Thus, we used an alternative approach involving the comparison of topoisomerase VIII-encoding and topoisomerase VIII-free closely related bacterial genomes. The integrated elements were expected to disrupt the collinearity of the corresponding genomic loci. DotPlot analysis showed that the genomic regions encoding topoisomerase VIII in *Desulfosporosinus meridiei* DSM 13257, *Desulfitobacterium hafniense* DCB-2 and *D. hafniense* Y51 are probably mobile. Notably, the elements in two *D. hafniense* strains are located in distinct genomic loci (Supplementary Figure S2). In addition, the topoisomerase VIII-encoding contig of *D. hafniense* DP7 is collinear throughout its length with the corresponding mobile region in *D. hafniense* DCB-2 and *D. hafniense* Y51. This suggests that this contig contains a mobile element which is related to the two other *D. hafniense* elements, DesHaf_DCB-2 and DesHaf_Y51.

The predicted integrated elements did not contain any viral signature genes (e.g. for virion components or genome packaging), suggesting that they are unlikely to be of viral origin. In contrast, many of the genes have homologs in bacterial plasmids (Figure 4). For example, 27 of the 42 open reading frames present in AmmDeg-E1 had plasmid homologs, indicating that the topoisomerase VIII-encoding elements are derived from integrative plasmids. We focused on elements for which exact boarders could be unequivocally defined to analyze genetic content in more detail (Figure 4). Elements PaePan-E1, AmmDeg-E1 and DesKuz-E1 and plasmids pPPM1a and pPAV109 encode both proteins typical of conjugative plasmids and integrative and conjugative elements (ICE) (31); e.g. all five elements encode TraG-like conjugal transfer coupling protein homologs. In addition, AmmDeg-E1 and DesKuz-E1 share homologs of ParM, a protein responsible for plasmid segregation upon cell division. Both elements also have several copies of toxin-antitoxin modules (RelE/B-like in AmmDeg-E1 and DesKuz-E1, HicA/HicB in AgrTum-E1 and MazE/F-like in AmmDeg-E1; Figure 4), which often ensure the stable maintenance of plasmids and ICEs (39). Several other proteins involved in DNA metabolism besides topoisomerase VIII are encoded within the predicted elements. AmmDeg-E1 and DesKuz-E1 encode homologs of the replication protein RepE of plasmid F, supporting further a link between the two integrated elements and conjugative plasmids. Interestingly, AmmDeg-E1 carries next to the RepE-like gene a homolog of the bifunctional primase-polymerase, which is typically found in various bacterial and archaeal mobile elements (40–42), whereas DesKuz-E1 contains at the same position a gene corresponding to a unique fusion of DnaG-like primase and DnaB-like helicase (Figure 4). The two smaller elements, RosAzw-E1 and RhiGal-E1, encode a homolog of family I DNA polymerase and a DNA ligase, respectively. Neither RosAzw-E1 nor RhiGal-E1 contains identifiable genes for components of the conjugal apparatus.

The elements harboring topoisomerase VIII-encoding genes appear to be at different stages of deterioration. At one end of the spectrum are the three plasmids, which probably rely on the encoded DNA metabolism proteins, including topoisomerase VIII, for their efficient replication. However, some of the integrated elements appear to be in the process of being inactivated. Analysis of the AmmDeg-E1 revealed that a number of genes, including those encoding superfamily II helicase, DNA methyltransferase, endonuclease, and two transposases, have accumulated mutations resulting in premature stop codons (Figure 4). However, we could identify the attachment sites (perfect direct repeats) flanking this element, suggesting that the insertion of this element into the bacterial chromosome was probably a recent event. At the opposite side of the spectrum are the elements for which the borders could not be defined, as in the case of *M. marina* ATCC 23134 (Supplementary Table S1). Notably, transposase may be involved in the process of element inactivation because topoisomerase VIII-encoding genes in *Dehalobacter* species CP, DCA and UNSWDHB are located next to truncated transposase genes.

Structural analysis of topoisomerases VIII

We performed structural modeling of the three topoisomerase VIII enzymes chosen for biochemical analyses (see below), including those from *A. degensii* KC4 (Adeg, 882 aa), *M. marina* ATCC 23134 (Mmar, 783 aa) and *P. polymyxa* M1 plasmid pPPM1a (pPpol, 751 aa). Structures of topoisomerase VI from *S. shibatae* (pdb code: 2ZBK; (10)) and *Methanosarcina mazei* (pdb code: 2Q2E; (9)) were used as templates. Despite their low sequence similarities, the three full-length topoisomerase VIII enzymes have the same modular organization as topoisomerase VI and contain five domains. Topoisomerase VIII contains the N-terminal Bergerat fold, followed by the S13-like H2TH and the ribosomal S5 domains 2-like fold that is present in the transducer domain of topoisomerase VI. This combination of domains is specific to the subunit B of type IIB topoisomerase. Topoisomerase VIII enzymes also possess a C-terminal WHD and Toprim domains; the organization of the Toprim domain is specific to type IIB topoisomerase subunit A (in type IIA enzymes, Toprim follows the transducer domain).

The ATP-lid in the ATP-binding site of topoisomerase VIII is highly conserved (Supplementary Figure S3). The switch lysine, which interacts with the γ-phosphate of the bound nucleotide is at position 446 in *A. degensii*, 374 in *M. marinus* and 369 in *P. polymyxa* topoisomerase VIII (Supplementary Figure S3). This lysine is conserved in all type IIA enzymes and topoisomerase VI transducer domains (K427 in *S. shibatae* topoisomerase VI). In the modeled structures, this switch lysine points away from the active site because the template used corresponds to the relaxed state conformation (9).

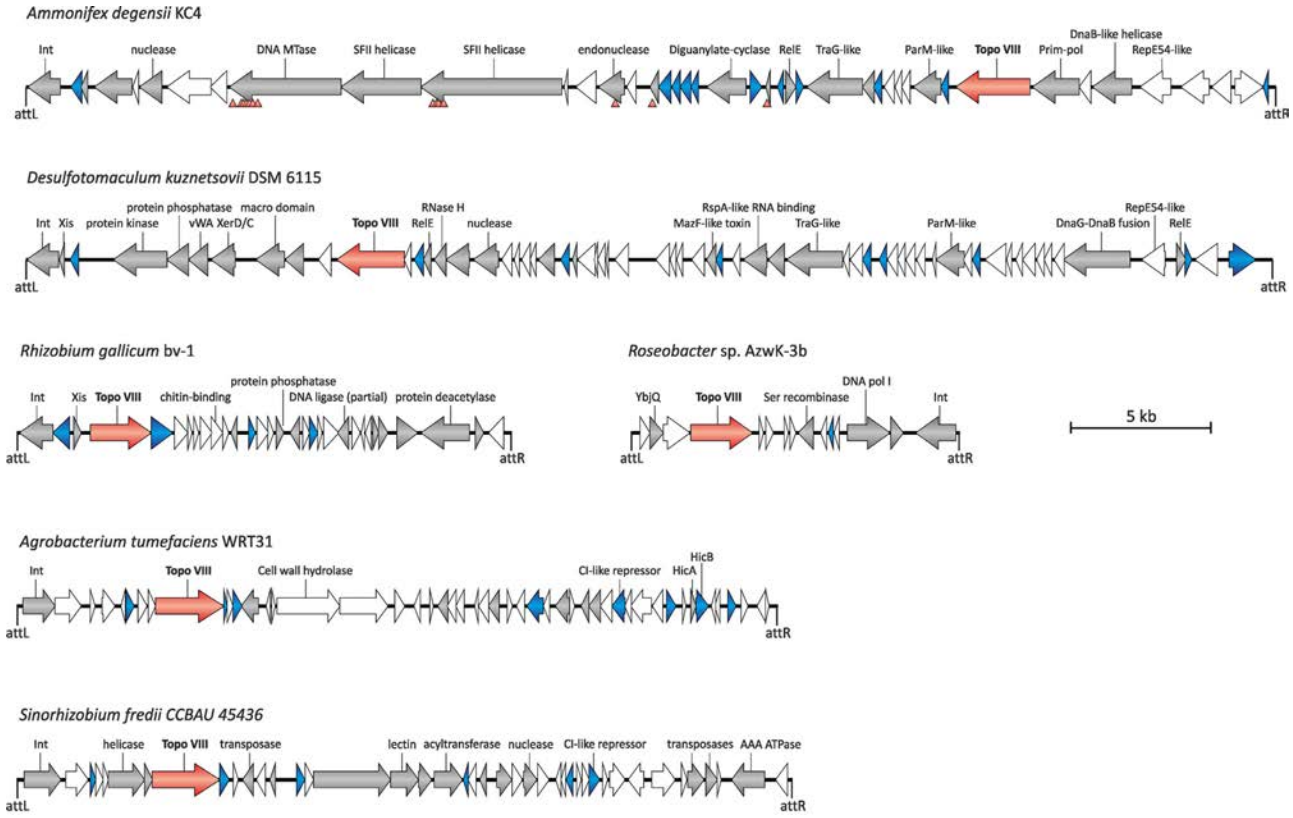


Figure 4. Topoisomerase VIII-coding genes in bacterial genomes are encoded within integrated mobile elements. Genomic organization of the six topoisomerase VIII-encoding integrated mobile elements. Open reading frames (ORFs) are depicted by arrows, and corresponding functions are indicated when possible. Topoisomerase VIII-coding genes are shown in red, genes encoding proteins with homologs in plasmids are depicted in gray, and blue arrows correspond to genes encoding DNA-binding proteins. Note that many of the DNA-binding proteins also have plasmid homologs. Positions of the left and right attachment sites (attL and attR, respectively) are also indicated. The positions of premature stop codons in several genes of the *Ammonifex degensii* KC4 element are depicted with red triangles under the genome map. Abbreviations: Int, integrase; MTase, methyltransferase; prim-pol, bifunctional primase-polymerase.

In regions corresponding to topoisomerase VI subunit B, the major difference between the two type IIB subfamilies is the length of the transducer domain, which is clearly smaller in topoisomerase VIII. The end of the last helix stops at nearly the middle of α11-helix of the *S. shibatae* topoisomerase VI structure (see alignment in Supplementary Figure S1 and Table S2). This corresponds to residue K450 which was identified as a hinge residue involved in helix bending that is crucial for conformational changes during the catalytic cycle (opening-closing of the N-gate) (10). The WHD domain of DNA topoisomerase VIII, a three-helix bundle, is smaller than the equivalent domain of topoisomerase VI (80 residues instead of 160: Supplementary Table S2). The first helix of this domain acts also as a junction between the transducer and the WHD domains (which corresponds to the junction between the B and A subunits). The structure of the Toprim domain is highly conserved and all the residues implicated in magnesium binding (E209, D261 and D263 in the *S. shibatae* structure) are spatially conserved.

Based on the structural similarities of the individual domains, we suggest that the architecture of the topoisomerase VIII homodimer is comparable to that of the topoisomerase VI heterotetramer. However, the C-terminal helix of the transducer and the N-terminal helix of the WHD domain are lacking in topoisomerase VIII; therefore, it is difficult to predict how the short junction characteristic of these enzymes organizes the orientation between these two domains (compare Figure 5 and Supplementary Figure S4). In addition, the β-turn-β in the topoisomerase VI Toprim domain responsible for the interactions between the two A subunits and the formation of a continuous β-sheet is not present in topoisomerase VIII. Therefore, it is difficult to propose a model for the full-length homodimer.

Topoisomerase VIII enzymatic activities

Impairment of the excision of integrated elements occasionally leads to permanent fixation of new functions within cellular lineages, which may even replace ancestral cellular counterparts. This route has been proposed for the origin of a novel nuclear protein in *Dinoflagellates* (43) and bacteriophage T3/T7-like RNA and DNA polymerases in mitochondria (44). With this in mind, we set out to verify experimentally whether plasmid-borne topoisomerase VIII possesses its predicted enzymatic activity and whether homologs encoded by (decaying) integrated elements retain

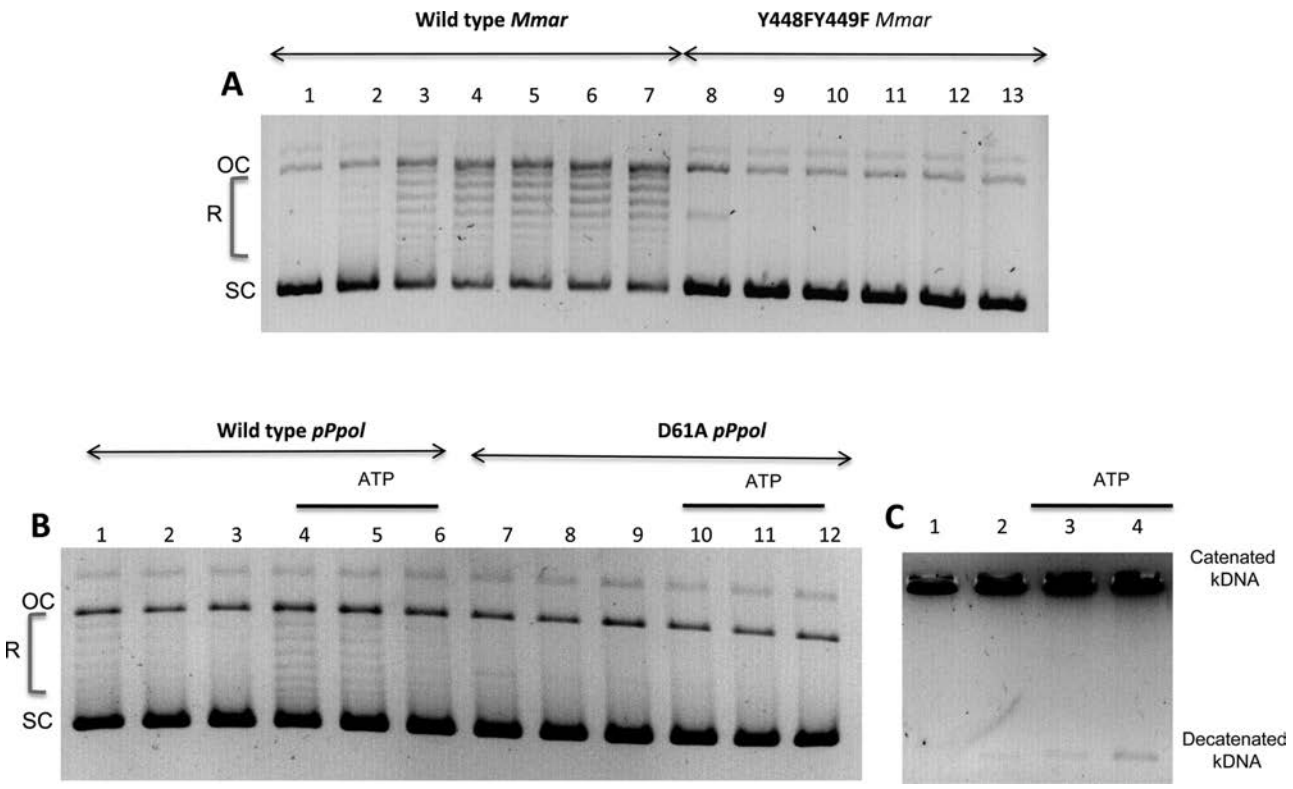


Figure 6. Relaxation and decatenation assays. (**A**) Relaxation assays of wild-type and double mutants of *Microscilla marina* (Mmar) topoisomerases VIII. Reactions were carried out without ATP and 200 ng of pBR322 was used as the substrate. Lane 1: control with no enzyme; lanes 2–7, wild-type (WT) with 1, 1.5, 2, 3, 6 and 12 pmol of enzyme, respectively; lanes 8–13, YY-FF double mutants with 0.5, 0.75, 1, 1.75, 3 and 6 pmol of enzyme, respectively. (**B**) *Paenibacillus polymyxa* (pPol) topoisomerase VIII relaxation assays with the wild-type enzyme; lanes 1–3, without ATP and with 10, 5, 2.5 pmol of enzyme, respectively, and lanes 4–6, with 1mM ATP and 10, 5, 2.5 pmol of enzymes, respectively; and with the D61A mutant lanes 7–9, without ATP, with 10, 5, and 2.5 pmol of enzyme, respectively, and lanes 10–12 with 1 mM ATP and 10, 5 and 2.5 pmol of enzyme, respectively. Supercoiled and relaxed (or open circular) topoisomers are noted as SC and R/OC, respectively. (**C**) *Paenibacillus polymyxa* topoisomerase VIII decatenation assays, with or without ATP; lane 1: no enzyme, lanes 2–4 with 0, 1 and 2 mM ATP, respectively and 20 pmol of enzyme. kDNA: kinetoplastid DNA.

two tyrosines of the active site were replaced with phenylalanines in a recombinant protein (Figure 6A).

Notably, the relaxation activity of the *P. polymyxa* topoisomerase VIII was lower than that of *M. marina* but ATP dependent, as expected of a type II DNA topoisomerase (Figure 6B and Supplementary Figures S8a, lanes 3, S8b, lanes 4 and 7). Substitution of the conserved aspartate in the *P. polymyxa* topoisomerase VIII Bergerat fold (D61), which is essential for ATP binding (D73 in *E. coli* DNA gyrase) abolished the ATP-dependent relaxation activity (Figure 6B). Treatment of the reaction products with proteinase K and SDS only produced linear DNA in the presence of ATP, suggesting the formation of a cleavable complex (Supplementary Figure S8, lane 6). Notably, we showed previously that the formation of a cleavable complex is ATP-dependent for *S. shibatae* DNA topoisomerase VI, whereas it is ATP independent for type IIA enzymes (48).

Some type II topoisomerases relax positively supercoiled DNA more efficiently than negatively supercoiled DNA (49). We thus tested the activity of the *P. polymyxa* DNA topoisomerase VIII on a positively supercoiled plasmid. This enzyme also exhibited low relaxation activity on positively supercoiled DNA (Supplementary Figure S8b). Finally, decatenation assays with kDNA as a substrate for *P. polymyxa* DNA topoisomerase VIII showed that this enzyme had a very weak decatenation activity, which was slightly stimulated by ATP (Figure 8C). The poor stability of the proteins precluded further enzymatic characterization.

DISCUSSION

We have identified a new group of type IIB topoisomerases that we call DNA topoisomerase VIII, which are encoded by either free or integrated plasmids. We partially characterized three proteins of this new subfamily. One of them exhibited only endonuclease activity, and the other two exhibited relaxation activity. In all cases, the relaxation activities detected with purified preparations of recombinant proteins were weak (complete relaxation was never observed). Furthermore, only the topoisomerase VIII encoded by the plasmid of *P. polymyxa* exhibited the expected ATP-dependent relaxation activity of type II topoisomerases. It is possible that the topoisomerase VIII of *A. degensii* and those of *M. marina* do not exhibit this activity because they are located on decaying elements and are no longer fully functional. The enzyme from *P. polymyxa* may exhibit the expected ATP-dependent relaxation activity because its activ-

ity is still required for the maintenance of this large plasmid (366 576 bp). The complete absence of relaxation activity in the case of the *A. degensii* enzyme is somewhat surprising because our *in silico* analysis suggests that the insertion of this gene into the bacterial chromosome is relatively recent. This indicates that the activity of enzymes encoded by mobile elements can be rapidly lost after integration; therefore, caution should be exercised when studying the biochemical properties of 'cellular enzymes' obtained by expressing genes present on cellular chromosome if the origin of these enzymes has not been carefully investigated.

We found that the structures of the Bergerat fold domains of the three topoisomerase VIII enzymes were rather variable (Figure 5). The rmsd values varied from 1.5 Å (*A. degensii* DNA topoisomerase VIII/*S. shibatae* topoisomerase VI) to 2.9 Å (*A. degensii* topoisomerase VIII/*M. marina* topoisomerase VIII and *M. marina* topoisomerase VIII/*S. shibatae* topoisomerase VI). This variability may explain differences in the activities of the three proteins. However, the core fold is highly conserved (especially the ATP-binding pocket, see Supplementary Figure S3). The main difference involves an insertion in the *A. degensii* topoisomerase VIII and *S. shibatae* topoisomerase VI in the C-terminal end of the Bergerat fold domain that is not present in *M. marina* and *P. polymyxa* topoisomerase VIII. At present, we cannot make substantiated conclusions about the relationship between ATP-dependency and structural differences among topoisomerase VIII enzymes from the available information, especially given that two highly conserved enzymes, the DNA gyrases of *M. tuberculosis* and *Mycobacterium leprae*, have different activity spectra (50,51).

The ATP-independent relaxation activity of the *M. marina* topoisomerase VIII is reminiscent of that of bacterial DNA gyrase (52,53). Bates *et al.* suggested that the ancestral role of ATP in type II enzyme reactions was to prevent DNA double-strand breaks (DSBs) by controlling the separation of protein–protein interfaces (54). They suggested that the intersubunit interface is weaker in DNA gyrase than in other type II enzymes, explaining its ATP-independent relaxation activity (54). It is interesting that the three different forms of topoisomerase VIII studied here exhibited low ATP-stimulated relaxation, ATP-independent relaxation or produced DSBs. This suggests that the differences between these three enzymes results from subtle modifications of the protein-protein dimer interface.

The low activity of the *P. polymyxa* topoisomerase VIII in our *in vitro* assays may be due to the absence of potentially important accessory proteins required for optimal activity *in vivo*. For example, plant topoisomerase VI enzymes require two additional proteins, Midget and RLH1, for their activity, at least *in vivo* (55). Moreover, the cell division protein MukB stimulates the activity of *E. coli* topoisomerase IV *in vitro* (56). Other topoisomerases require post-translational modifications to achieve full catalytic strength (57). If additional proteins or processing enzymes are indeed required for the full activity of topoisomerase VIII, these accessory proteins are likely to be encoded by the plasmids bearing the topoisomerase VIII genes.

Origin and evolution of DNA topoisomerase VIII

Several families and subfamilies of DNA topoisomerases are widespread in the living world. Some are almost universal (such as topoisomerase IA) or present in all members of one or two cellular domains (such as DNA gyrases in Bacteria or type IIA topoisomerases in Bacteria and Eukarya) (7). In contrast, topoisomerase VIII enzymes appear to be rare, because they are encoded only in about 0.5% of currently available bacterial genomes and are not present in archaeal or eukaryotic genomes. Moreover, topoisomerase VIII is only present in a handful of representatives of three bacterial phyla. Furthermore, two of them, Proteobacteria and Firmicutes, correspond to the two phyla with by far the highest number of completely sequenced genomes. This very narrow distribution of topoisomerase VIII in the living world resembles the case of topoisomerase V, a unique member of type I topoisomerases (family C) that is only present in the archaeon *Methanopyrus kandleri* (58,59). It also resembles atypical topoisomerases found in the viral world, such as the heterotrimeric type IIA enzymes encoded by T4-like bacteriophages and the homodimeric type IIA enzymes encoded by some Megavirales (7). Strikingly, topoisomerase VIII enzymes are encoded by plasmids (two from bacteria and one from archaea), and in bacterial genomes, they are located within integrated mobile elements. This is therefore another example of a topoisomerase with an unusual phylogenomic distribution and complex evolutionary trajectory.

The fusion of two bacterial topoisomerase VI subunits in one bacterial phylum, followed by sporadic spread to other lineages by lateral gene transfer is one hypothesis that explains the rare and scattered distribution of topoisomerase VIII in Bacteria. However, this hypothesis seems unlikely considering the high divergence of primary sequence between topoisomerase VI and VIII. Notably, *bona fide* topoisomerase VI enzymes present in bacteria cannot be distinguished from their archaeal homologs and branch with archaeal DNA topoisomerase VI enzymes in phylogenetic analyses (38); in contrast, topoisomerase VI and VIII enzymes are so divergent that their amino-acid sequences cannot be reliably aligned for phylogenetic analyses. It is difficult to explain why the fusion protein of the two topoisomerase VI-like subunits (i.e. the ancestor of topoisomerase VIII) would have diverged so rapidly in one particular bacterial lineage but remained conserved during its dispersion in various bacterial lineages.

The divergence between topoisomerases VI and VIII enzymes suggests that the gene duplication at the origin of these two protein subfamilies occurred before the emergence of Archaea and Bacteria (Figure 7). In agreement with this view, topoisomerase VI was probably present in the last archaeal common ancestor (LACA) because this enzyme is now present in all archaea, with the exception of Thermoplasmatales (3). In contrast, the rarity of topoisomerase VIII in bacteria suggests that this protein was not present in the last bacterial common ancestor (LBCA), but was already encoded by plasmids at that time. Modern topoisomerase VIII thus probably originated before the diversification of Bacteria, by a fusion between the A and B subunits, following the duplication that initiated the diver-

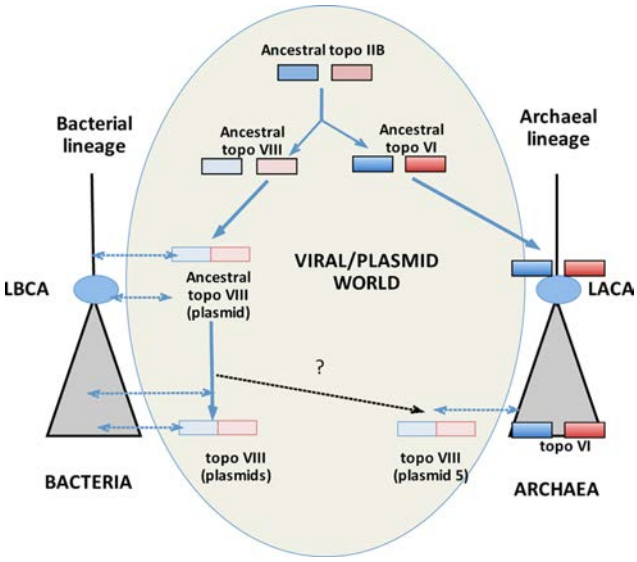


Figure 7. Proposed scenario for the evolution of the type IIB topoisomerase family. LBCA (last bacterial common ancestor), LACA (last archaeal common ancestor). Dotted double arrows indicate co-evolution of cellular and plasmidic lineages, with subsequent integration or loss of plasmidic topoisomerase VIII-encoding genes in cellular genomes. Black dotted arrow and question mark, possible transfer of a bacterial plasmid to the archaeal domain. For simplicity, both the presence of a few topoisomerase VI enzymes in Bacteria (transferred from Archaea) and the secondary split of the topoisomerase VIII gene in the plasmid PAV109 of *Paenibacillus alvei* are not indicated.

gence between topoisomerase VI and VIII enzymes (Figure 7). In this scenario, the sporadic distribution of topoisomerase VIII in the bacterial domain is easily explained by the co-evolution of bacterial plasmids bearing topoisomerase VIII-encoding genes with their hosts and the sporadic integration of these genes into bacterial genomes (37).

It has been suggested that DNA topoisomerases first originated and diverged into various families and subfamilies in the viral world (7). Viruses and plasmids are evolutionarily related and can be considered as the same sequence space within a greater viral world (60–62). Accordingly, it is possible that type IIB topoisomerases first originated and subsequently diversified into two subfamilies (VI and VIII) in this plasmid/viral world. Later on, topoisomerase VI were transferred to the cellular members of the branch of the tree of life leading to LACA, whereas plasmids encoding topoisomerase VIII co-evolved with ancestors of the bacterial domain (Figure 7). The presence of the topoisomerase VIII-encoding gene in the plasmid of a halophilic archaeon can also be explained by the presence of these plasmids in ancient archaeal organisms (LACA and close relatives). However, in this case, it is also possible that these plasmids were transferred from Bacteria to Archaea at the onset of the haloarchaeal lineage, which has experienced the introduction of around 1000 bacterial genes (63). Irrespective of the evolutionary scenario, the discovery of a subfamily of topoisomerases specifically encoded by plasmids confirms that mobile elements are potential reservoirs of novel proteins involved in DNA metabolism (7,36,42,64).

Possible physiological role of topoisomerase VIII

The role of topoisomerase VIII in plasmid physiology remains to be established. Several conjugative plasmids from Proteobacteria and Firmicutes encode enzymes related to bacterial topoisomerase III (a subfamily of type IA enzymes) (65). These plasmidic topoisomerase III enzymes are used either as decatenases to help the faithful segregation of these large plasmids with low copy number, or as swivels during conjugation to facilitate the unwinding of the donor strand that is transferred to the recipient host (65). Alternatively, they may also be involved in the resolution of hemicatenane structures during plasmid replication or recombination (66). Notably, the plasmids pPPM1a of *P. polymyxa* and pAV109 of *P. alvei* encode a DNA topoisomerase III, in addition to a topoisomerase VIII. Topoisomerase III may be involved in the resolution of hemicatenane structures, whereas topoisomerase VIII enzymes could be the decatenase and/or swivelase involved in the segregation of these large plasmids (>100 kb).

Interestingly, the genome of *P. polymyxa* contains nine loci dedicated to the synthesis of non-ribosomal peptides, one of which is located on the plasmid pPPM1a (67). Of note, some plasmids encode proteins containing pentapeptide repeats that interact with cellular topoisomerases, as demonstrated by fluoroquinolone resistance proteins (68). These peptides may serve as important antibiotics in the control of plant pathogens. We can speculate that some of the peptides encoded by the plasmid pPPM1a are also topoisomerase inhibitors. In that case, the plasmid-encoded topoisomerase III and topoisomerase VIII may also be resistant to these antibiotics, and serve as substitutes for antibiotic-sensitive cellular topoisomerases. Thus *P. polymyxa* may be an interesting model to study the role of topoisomerases in plasmid maintenance and transfer. It would also be relevant to screen *Paenibacillus* species for novel anti-topoisomerase inhibitors and/or new proteins interfering with topoisomerases.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

European Research Council (ERC) grant from the European Union's Seventh Framework Programme [FP/2007–2013 to P.F.]; Project EVOMOBIL [ERC Grant Agreement no. 340440 to P.F.]; Paul W. Zuccaire Foundation (to K.R.). Funding for open access charge: Institut Pasteur.
Conflict of interest statement. None declared.

REFERENCES

1. Schoeffler, A.J. and Berger, J.M. Schoeffler, A.J. and Berger, J.M. (2008) Topos: harnessing and constraining energy to govern chromosome topology. *Q. Rev. Biophys.*, **41**, 41–101.
2. Wang, J.C.Wang, J.C. (2009) In: *Untangling the Double Helix*, Cold Spring Harbor Laboratory Press, Harvard University..
3. Forterre, P. Pommier, YEd.Forterre, P.Inedpp.Forterre, P. (2011) Introduction and historical perspective. In: *Topos and Cancer*, Pommier, YEd. Humana press, Springer-Verlag. pp. 1–52.

4. Bergerat, A., de Massy, B., Gadelle, D., Varoutas, P.C., Nicolas, A., and Forterre, P.Bergerat, A., de Massy, B., Gadelle, D., Varoutas, P.C., Nicolas, A., and Forterre, P. (1997) An atypical topoisomerases II from Archaea with implications for meiotic recombination. *Nature*, **386**, 414–417.
5. Gadelle, D., Filée, J., Buhler, C., and Forterre, P.Gadelle, D., Filée, J., Buhler, C., and Forterre, P. (2003) Phylogenomics of type II Topos. *Bioessays*, **25**, 232–242.
6. Corbett, K.D. and Berger, J.M.Corbett, K.D. and Berger, J.M. (2004) Structure, molecular mechanisms, and evolutionary relationships in Topos. *Annu. Rev. Biophys. Biomol. Struct.*, **33**, 95–118.
7. Forterre, P. and Gadelle, D.Forterre, P. and Gadelle, D. (2009) Phylogenomics of Topos: their origin and putative roles in the emergence of modern organisms. *Nucleic Acids Res.*, **37**, 679–692.
8. Schoeffler, A.J. and Berger, J.M.Schoeffler, A.J. and Berger, J.M. (2005) Recent advances in understanding structure-function relationships in the type II topoisomerases mechanism. *Biochem. Soc. Trans.*, **33**, 1465–1470.
9. Corbett, K.D., Benedetti, P., and Berger, J.M.Corbett, K.D., Benedetti, P., and Berger, J.M. (2007) Holoenzyme assembly and ATP-mediated conformational dynamics of topoisomerases VI. *Nat. Struct. Mol. Biol.*, **14**, 611–619.
10. Graille, M., Cladière, L., Durand, D., Lecointe, F., Gadelle, D., Quevillon-Cheruel, S., Vachette, P., Forterre, P., and van Tilbeurgh, H. Graille, M., Cladière, L., Durand, D., Lecointe, F., Gadelle, D., Quevillon-Cheruel, S., Vachette, P., Forterre, P., and van Tilbeurgh, H. (2008) Crystal structure of an intact type II Topos: insights into DNA transfer mechanisms. *Structure*, **16**, 360–370.
11. Laponogov, I., Sohi, M.K., Veselkov, D.A., Pan, X.-S., Sawhney, R., Thompson, A.W., McAuley, K.E., Fisher, L.M., and Sanderson, M.R.Laponogov, I., Sohi, M.K., Veselkov, D.A., Pan, X.-S., Sawhney, R., Thompson, A.W., McAuley, K.E., Fisher, L.M., and Sanderson, M.R. (2009) Structural insight into the quinolone-DNA cleavage complex of type IIA topoisomerases. *Nat. Struct. Mol. Biol.*, **16**, 667–669.
12. Dutta, R. and Inouye, M.Dutta, R. and Inouye, M. (2000) GHKL, an emergent ATPase/kinase superfamily. *Trends Biochem. Sci.*, **25**, 24–28.
13. Woese, C.R., Kandler, O., and Wheelis, M.L.Woese, C.R., Kandler, O., and Wheelis, M.L. (1990) Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl. Acad. Sci. U.S.A.*, **87**, 4576–4579.
14. Lecompte, O., Ripp, R., Thierry, J.-C., Moras, D., and Poch, O.Lecompte, O., Ripp, R., Thierry, J.-C., Moras, D., and Poch, O. (2002) Comparative analysis of ribosomal proteins in complete genomes: an example of reductive evolution at the domain scale. *Nucleic Acids Res.*, **30**, 5382–5390.
15. Forterre, P., Gribaldo, S., Gadelle, D., and Serre, M.-C.Forterre, P., Gribaldo, S., Gadelle, D., and Serre, M.-C. (2007) Origin and evolution of Topos. *Biochimie*, **89**, 427–446.
16. Bergerat, A., Gadelle, D., and Forterre, P.Bergerat, A., Gadelle, D., and Forterre, P. (1994) Purification of a Topos II from the hyperthermophilic archaeon Sulfolobus shibatae. A thermostable enzyme with both bacterial and eucaryal features. *J. Biol. Chem.*, **269**, 27 663–27 669.
17. Brochier-Armanet, C., Gribaldo, S., and Forterre, P.Brochier-Armanet, C., Gribaldo, S., and Forterre, P. (2008) A Topos IB in Thaumarchaeota testifies for the presence of this enzyme in the last common ancestor of Archaea and Eucarya. *Biol. Direct*, **3**, 54.
18. Sugimoto-Shirasu, K., Stacey, N.J., Corsar, J., Roberts, K., and McCann, M.C.Sugimoto-Shirasu, K., Stacey, N.J., Corsar, J., Roberts, K., and McCann, M.C. (2002) Topos VI is essential for endoreduplication in Arabidopsis. *Curr. Biol.*, **12**, 1782–1786.
19. Filée, J., Forterre, P., Sen-Lin, T., and Laurent, J.Filée, J., Forterre, P., Sen-Lin, T., and Laurent, J. (2002) Evolution of DNA polymerase families: evidences for multiple gene exchange between cellular and viral proteins. *J. Mol. Evol.*, **54**, 763–773.
20. Forterre, P.Forterre, P. (2002) The origin of DNA genomes and DNA replication proteins. *Curr. Opin. Microbiol.*, **5**, 525–532.
21. Plewniak, F., Bianchetti, L., Brelivet, Y., Carles, A., Chalmel, F., Lecompte, O., Mochel, T., Moulinier, L., Muller, A., and Muller, J. *et al.*Plewniak, F., Bianchetti, L., Brelivet, Y., Carles, A., Chalmel, F., Lecompte, O., Mochel, T., Moulinier, L., Muller, A., and Muller, J. (2003) PipeAlign: A new toolkit for protein family analysis. *Nucleic Acids Res.*, **31**, 3829–3832.
22. Jones, D.T.Jones, D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, **292**, 195–202.
23. Edgar, R.C.Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
24. Gouy, M., Guindon, S., and Gascuel, O.Gouy, M., Guindon, S., and Gascuel, O. (2010) SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol. Biol. Evol.*, **27**, 221–224.
25. Criscuolo, A. and Gribaldo, S.Criscuolo, A. and Gribaldo, S. (2010) BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evol. Biol.*, **10**, 210.
26. Abascal, F., Zardoya, R., and Posada, D.Abascal, F., Zardoya, R., and Posada, D. (2005) ProtTest: selection of best-fit models of protein evolution. *Bioinformatics*, **21**, 2104–2105.
27. Stamatakis, A.Stamatakis, A. (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, **22**, 2688–2690.
28. Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D.L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M.A., and Huelsenbeck, J.P.Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D.L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M.A., and Huelsenbeck, J.P. (2012) MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.*, **61**, 539–542.
29. Kelley, L.A. and Sternberg, M.J.E.Kelley, L.A. and Sternberg, M.J.E. (2009) Protein structure prediction on the Web: a case study using the Phyre server. *Nat. Protoc.*, **4**, 363–371.
30. Arnold, K., Bordoli, L., Kopp, J., and Schwede, T.Arnold, K., Bordoli, L., Kopp, J., and Schwede, T. (2006) The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics*, **22**, 195–201.
31. Krupovic, M., Forterre, P., and Bamford, D.H.Krupovic, M., Forterre, P., and Bamford, D.H. (2010) Comparative analysis of the mosaic genomes of tailed archaeal viruses and proviruses suggests common themes for virion architecture and assembly with tailed viruses of bacteria. *J. Mol. Biol.*, **397**, 144–160.
32. Krumsiek, J., Arnold, R., and Rattei, T.Krumsiek, J., Arnold, R., and Rattei, T. (2007) Gepard: a rapid and sensitive tool for creating dotplots on genome scale. *Bioinformatics*, **23**, 1026–1028.
33. Bizard, A., Garnier, F., and Nadal, M.Bizard, A., Garnier, F., and Nadal, M. (2011) TopR2, the second reverse gyrase of *Sulfolobus solfataricus*, exhibits unusual properties. *J. Mol. Biol.*, **408**, 839–849.
34. Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J.Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
35. Krupovic, M., Gribaldo, S., Bamford, D.H., and Forterre, P.Krupovic, M., Gribaldo, S., Bamford, D.H., and Forterre, P. (2010) The evolutionary history of archaeal MCM helicases: a case study of vertical evolution combined with hitchhiking of mobile genetic elements. *Mol. Biol. Evol.*, **27**, 2716–2732.
36. Soler, N., Marguet, E., Cortez, D., Desnoues, N., Keller, J., van Tilbeurgh, H., Sezonov, G., and Forterre, P.Soler, N., Marguet, E., Cortez, D., Desnoues, N., Keller, J., van Tilbeurgh, H., Sezonov, G., and Forterre, P. (2010) Two novel families of plasmids from hyperthermophilic archaea encoding new families of replication proteins. *Nucleic Acids Res.*, **38**, 5088–5104.
37. Forterre, P.Forterre, P. (2012) Darwin's goldmine is still open: variation and selection run the world. *Front. Cell Infect. Microbiol.*, **2**, 106.
38. Raymann, K., Forterre, P., Brochier-Armanet, C., and Gribaldo, S.doi:10.1093/gbe/evu004Raymann, K., Forterre, P., Brochier-Armanet, C., and Gribaldo, S. (2014) Genome Biol. Evol.,.
39. Wozniak, R.A.F. and Waldor, M.K.Wozniak, R.A.F. and Waldor, M.K. (2010) Integrative and conjugative elements: mosaic mobile genetic elements enabling dynamic lateral gene flow. *Nat. Rev. Microbiol.*, **8**, 552–563.
40. Lipps, G.Lipps, G. (2004) The replication protein of the Sulfolobus islandicus plasmid pRN1. *Biochem. Soc. Trans.*, **32**, 240–244.

41. Krupovic, M., Gonnet, M., Hania, W.B., Forterre, P., and Erauso, G.Krupovic, M., Gonnet, M., Hania, W.B., Forterre, P., and Erauso, G. (2013) Insights into dynamics of mobile genetic elements in hyperthermophilic environments from five new *Thermococcus* plasmids. *PLoS One*, **8**, e49044.
42. Gill, S., Krupovic, M., Desnoues, N., Béguin, P., Sezonov, G., and Forterre, P.Gill, S., Krupovic, M., Desnoues, N., Béguin, P., Sezonov, G., and Forterre, P. (2014) A highly divergent archaeo-eukaryotic primase from the *Thermococcus nautilus* plasmid, pTN2. *Nucleic Acids Res.*, **42**, 3707–3719.
43. Gornik, S.G., Ford, K.L., Mulhern, T.D., Bacic, A., McFadden, G.I., and Waller, R.F.Gornik, S.G., Ford, K.L., Mulhern, T.D., Bacic, A., McFadden, G.I., and Waller, R.F. (2012) Loss of nucleosomal DNA condensation coincides with appearance of a novel nuclear protein in dinoflagellates. *Curr. Biol.*, **22**, 2303–2312.
44. Filée, J. and Forterre, P.Filée, J. and Forterre, P. (2005) Viral proteins functioning in organelles: a cryptic origin. *Trends Microbiol.*, **13**, 510–513.
45. Lal, S., Romano, S., Chiarini, L., Signorini, A., and Tabacchioni, S.Lal, S., Romano, S., Chiarini, L., Signorini, A., and Tabacchioni, S. (2012) The Paenibacillus polymyxa species is abundant among hydrogen-producing facultative anaerobic bacteria in Lake Averno sediment. *Arch. Microbiol.*, **194**, 345–351.
46. Huber, R., Rossnagel, P., Woese, C.R., Rachel, R., Langworthy, T.A., and Stetter, K.O.Huber, R., Rossnagel, P., Woese, C.R., Rachel, R., Langworthy, T.A., and Stetter, K.O. (1996) Formation of ammonium from nitrate during chemolithoautotrophic growth of the extremely thermophilic bacterium ammonifex degensii gen. nov. sp. nov. *Syst. Appl. Microbiol.*, **19**, 40–49.
47. Lewin, R.A.Lewin, R.A. (1969) A classification of flexibacteria. *J. Gen. Microbiol.*, **58**, 189–206.
48. Buhler, C., Gadelle, D., Forterre, P., Wang, J.C., and Bergerat, A.Buhler, C., Gadelle, D., Forterre, P., Wang, J.C., and Bergerat, A. (1998) Reconstitution of DNA topoisomerase VI of the thermophilic archaeon *Sulfolobus shibatae* from subunits separately overexpressed in Escherichia coli. *Nucleic Acids Res.*, **26**, 5157–62.
49. McClendon, A.K., Rodriguez, A.C., and Osheroff, N.McClendon, A.K., Rodriguez, A.C., and Osheroff, N. (2005) Human topoisomerase IIalpha rapidly relaxes positively supercoiled DNA: implications for enzyme action ahead of replication forks. *J. Biol. Chem.*, **280**, 39 337–45.
50. Aubry, A., Fisher, L.M., Jarlier, V., and Cambau, E.Aubry, A., Fisher, L.M., Jarlier, V., and Cambau, E. (2006) First functional characterization of a singly expressed bacterial type II topoisomerase: the enzyme from *Mycobacterium tuberculosis*. *Biochem. Biophys. Res. Commun.*, **348**, 158–65.
51. Matrat, S., Petrella, S., Cambau, E., Sougakoff, W., Jarlier, V., and Aubry, A.Matrat, S., Petrella, S., Cambau, E., Sougakoff, W., Jarlier, V., and Aubry, A. (2007) Expression and purification of an active form of the *Mycobacterium leprae* DNA gyrase and its inhibition by quinolones. *Antimicrob. Agents Chemother.*, **51**, 643–8.
52. Gellert, M., Mizuuchi, K., O'Dea, M.H., Itoh, T., and Tomizawa, J.I.Gellert, M., Mizuuchi, K., O'Dea, M.H., Itoh, T., and Tomizawa, J.I. (1977) Nalidixic acid resistance: a second genetic character involved in DNA gyrase activity. *Proc. Natl. Acad. Sci. U.S.A.*, **74**, 4772–4776.
53. Sugino, A., Peebles, C.L., Kreuzer, K.N., and Cozzarelli, N.R.Sugino, A., Peebles, C.L., Kreuzer, K.N., and Cozzarelli, N.R. (1977) Mechanism of action of nalidixic acid: purification of Escherichia coli *nalA* gene product and its relationship to DNA gyrase and a novel nicking-closing enzyme. *Proc. Natl. Acad. Sci. U.S.A.*, **74**, 4767–71.
54. Bates, A.D., Berger, J.M., and Maxwell, A.Bates, A.D., Berger, J.M., and Maxwell, A. (2011) The ancestral role of ATP hydrolysis in type II topoisomerases: prevention of DNA double-strand breaks. *Nucleic Acids Res.*, **39**, 6327–6339.
55. Kirik, V., Schrader, A., Uhrig, J.F., and Hulskamp, M.Kirik, V., Schrader, A., Uhrig, J.F., and Hulskamp, M. (2007) MIDGET unravels functions of the *Arabidopsis* topoisomerase VI complex in DNA endoreduplication, chromatin condensation, and transcriptional silencing. *Plant Cell*, **19**, 3100–10.
56. Hayama, R. and Marians, K.J.Hayama, R. and Marians, K.J. (2010) Physical and functional interaction between the condensin MukB and the decatenase topoisomerase IV in *Escherichia coli*. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 18826–18831.
57. Chikamori, K., Grozav, A.G., Kozuki, T., Grabowski, D., Ganapathi, R., and Ganapathi, M.K.Chikamori, K., Grozav, A.G., Kozuki, T., Grabowski, D., Ganapathi, R., and Ganapathi, M.K. (2010) DNA topoisomerase II enzymes as molecular targets for cancer chemotherapy. *Curr. Cancer Drug Targets*, **10**, 758–71.
58. Taneja, B., Patel, A., Slesarev, A., and Mondragón, A.Taneja, B., Patel, A., Slesarev, A., and Mondragón, A. (2006) Structure of the N-terminal fragment of topoisomerases V reveals a new family of topoisomerases. *EMBO J.*, **25**, 398–408.
59. Forterre, P.Forterre, P. (2006) Topos V: a new fold of mysterious origin. *Trends Biotechnol.*, **24**, 245–247.
60. Forterre, P.Forterre, P. (2005) The two ages of the RNA world, and the transition to the DNA world: a story of viruses and cells. *Biochimie*, **87**, 793–803.
61. Koonin, E.V. and Dolja, V.V.Koonin, E.V. and Dolja, V.V. (2013) A virocentric perspective on the evolution of life. *Curr. Opin. Virol.*, **3**, 546–557.
62. Krupovic, M.Krupovic, M. (2013) Networks of evolutionary interactions underlying the polyphyletic origin of ssDNA viruses. *Curr. Opin. Virol.*, **3**, 578–586.
61. Nelson-Sathi, S., Dagan, T., Landan, G., Janssen, A., Steel, M., McInerney, J.O., Deppenmeier, U., and Martin, W.F.Nelson-Sathi, S., Dagan, T., Landan, G., Janssen, A., Steel, M., McInerney, J.O., Deppenmeier, U., and Martin, W.F. (2012) Acquisition of 1,000 eubacterial genes physiologically transformed a methanogen at the origin of Haloarchaea. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 20 537–20 542.
62. Kolkenbrock, S., Naumann, B., Hippler, M., and Fetzner, S.Kolkenbrock, S., Naumann, B., Hippler, M., and Fetzner, S. (2010) A novel replicative enzyme encoded by the linear Arthrobacter plasmid pAL1. *J. Bacteriol.*, **192**, 4935–4943.
65. Li, Z., Hiasa, H., Kumar, U., and DiGate, R.J.Li, Z., Hiasa, H., Kumar, U., and DiGate, R.J. (1997) The traE gene of plasmid RP4 encodes a homologue of Escherichia coli Topos III. *J. Biol. Chem.*, **272**, 19 582–19 587.
66. Lee, S.-H., Siaw, G.E.-L., Willcox, S., Griffith, J.D., and Hsieh, T.-S.Lee, S.-H., Siaw, G.E.-L., Willcox, S., Griffith, J.D., and Hsieh, T.-S. (2013) Synthesis and dissolution of hemicatenanes by type IA Topos. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, E3587–94.
67. Niu, B., Rueckert, C., Blom, J., Wang, Q., and Borriss, R.Niu, B., Rueckert, C., Blom, J., Wang, Q., and Borriss, R. (2011) The genome of the plant growth-promoting rhizobacterium Paenibacillus polymyxa M-1 contains nine sites dedicated to nonribosomal synthesis of lipopeptides and polyketides. *J. Bacteriol.*, **193**, 5862–5863.
68. Vetting, M.W., Hegde, S.S., Wang, M., Jacoby, G.A., Hooper, D.C., and Blanchard, J.S.Vetting, M.W., Hegde, S.S., Wang, M., Jacoby, G.A., Hooper, D.C., and Blanchard, J.S. (2011) Structure of QnrB1, a plasmid-mediated fluoroquinolone resistance factor. *J. Biol. Chem.*, **286**, 25265–25273.

Appendix 3

Supplementary Material Article 1

Figure S1

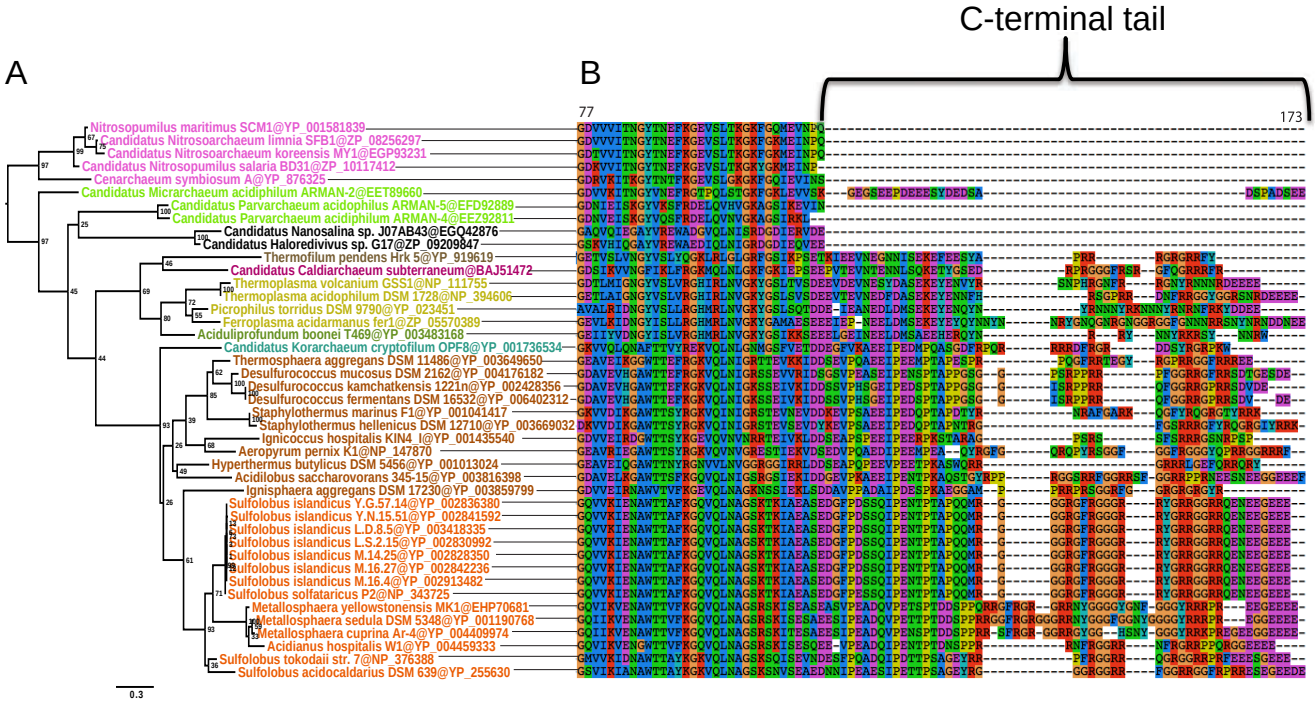


Figure S1. Phylogeny and alignment of archaeal single-stranded binding proteins.
A) Maximum likelihood phylogeny of archaeal SSB proteins (92 postions). The tree was calculated by PhyML (LG model+Γ4). The scale bar represents average substitutions per site. Values at nodes represent bootstrap values based on 100 replicates of the original dataset.
B) C-terminus of the complete SSB alignment (from position 77-173 of the original alignment) displaying the presence/absence of the C-terminal tail. The sequence order of the alignment corresponds to the phylogeny in part A.

Figure S2.

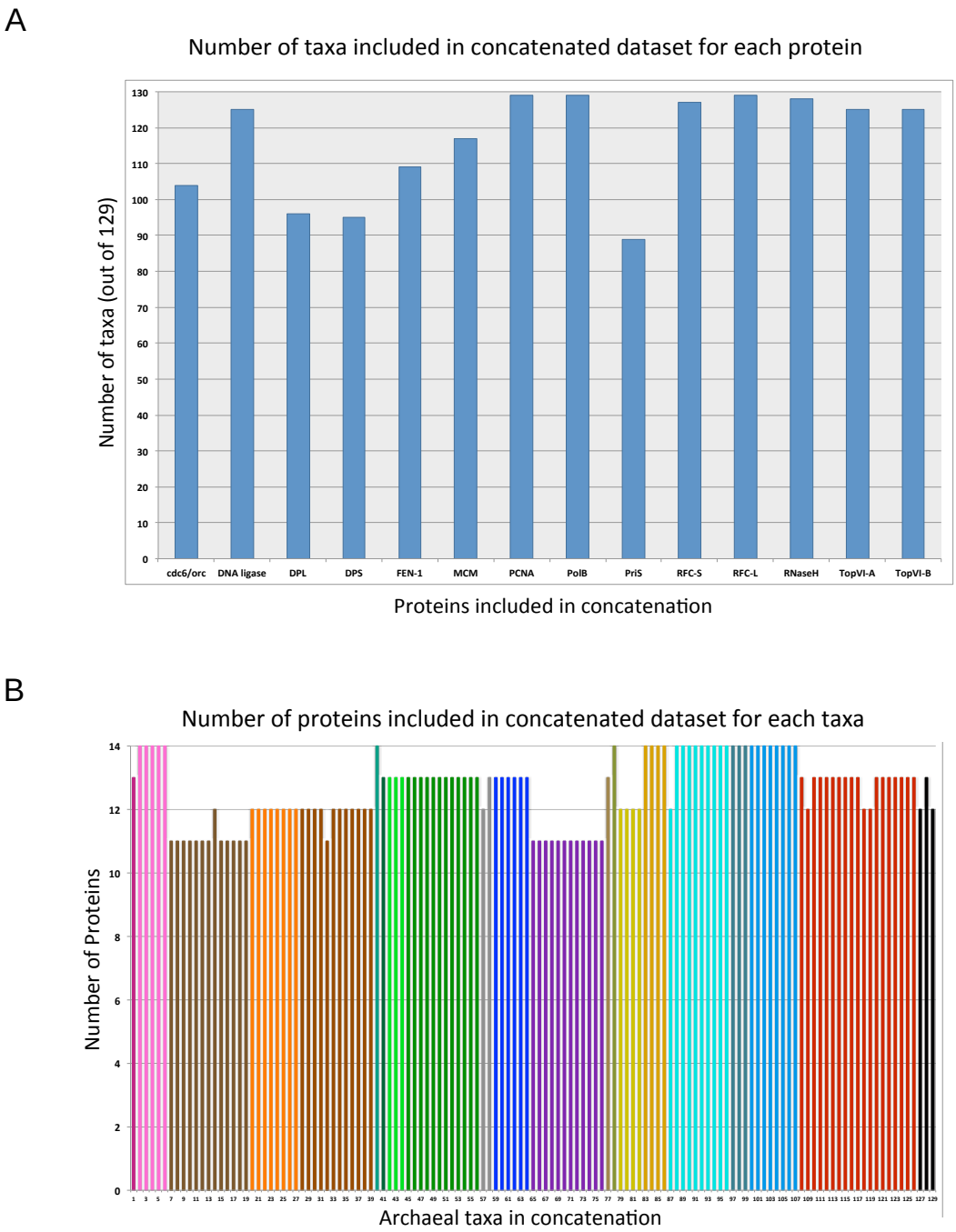


Figure S2. Summary of the DNA replication concatenated dataset
A) Graph displaying the number of archaeal taxa (y-axis) included for each replication protein (x-axis) used to create the concatenated dataset.
B) Graph showing the number of replication proteins out of 14 (y-axis) included for each archaeal taxa (x-axis) in the concatenated dataset. Colors correspond to those in Figure 2.

Table S4

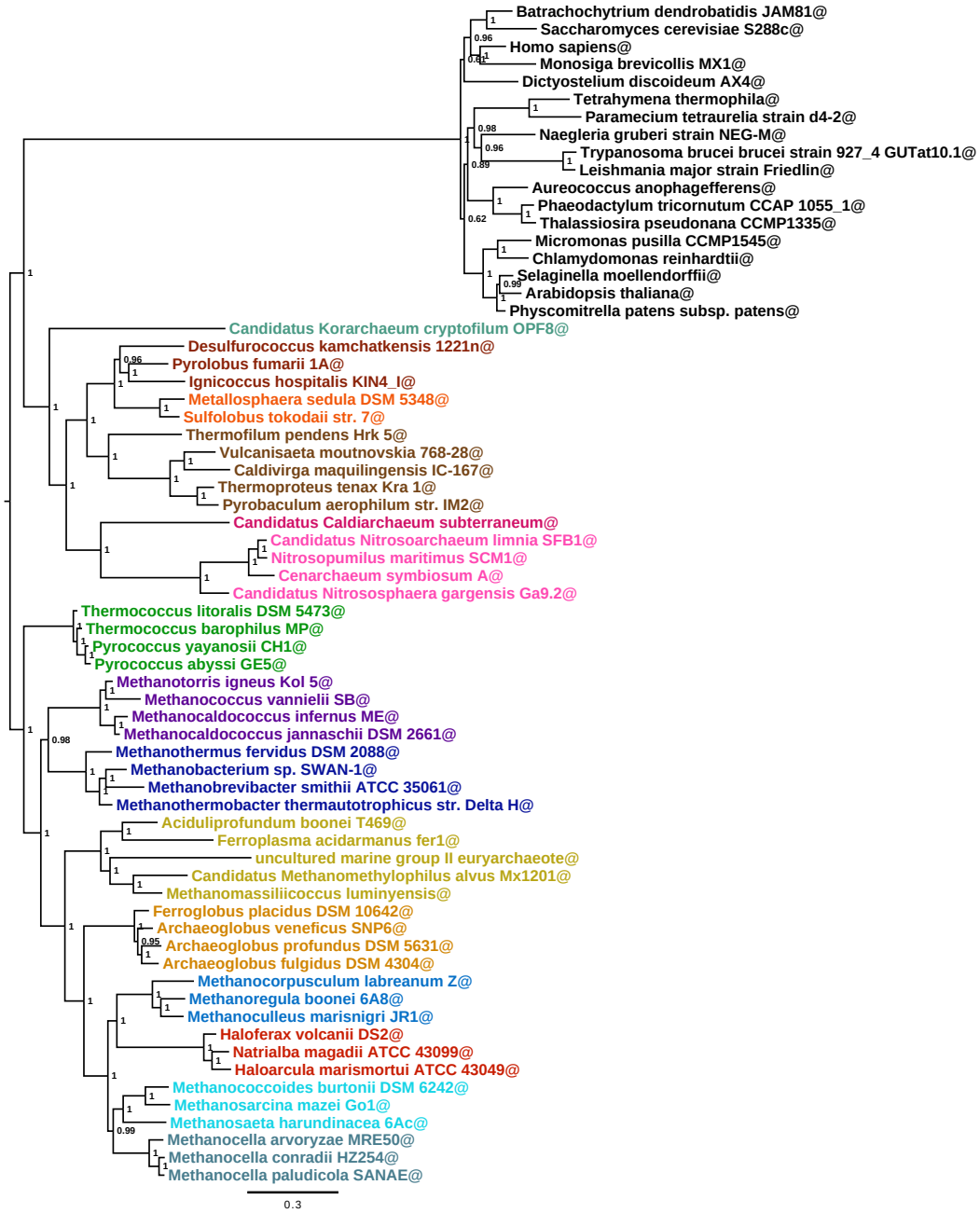
	SOSS B2 (ssb)		SOSS B1 (ssb)
Metazoa			
Pan troglodytes	XP_001167708		XP_001153310
Gorilla gorilla gorilla	XP_004033018		XP_004053414
Homo sapiens	NP_001026886		NP_076973
Rattus norvegicus	NP_001014238		NP_001030111
Mus musculus	NP_082972		NP_081533
Sus scrofa	NP_001231420		XP_001929324
Ovis aries	XP_004004824		XP_004007515
Bos taurus	NP_001091593		NP_001094621
Felis catus	XP_003991031		XP_003988934
Anolis carolinensis	XP_003217474		XP_003216992
Danio rerio	NP_001008643		NP_001025423
Ciona intestinalis	XP_002130978		
Schistosoma mansoni	XP_002574729	XP_002574730	
Drosophila melanogaster	NP_609115		
Anopheles gambiae str. PEST	XP_314101		
Nasonia vitripennis	XP_001601525		
Bombus terrestris	XP_003396637		
Tribolium castaneum	XP_972113		
Daphnia pulex	EFX70646		
Fungi; Microsporidia			
Encephalitozoon intestinalis ATCC 50506	XP_003072300		
Encephalitozoon romaleae SJ-2008	AFN82411		
Encephalitozoon hellem ATCC 50504	XP_003886679		
Encephalitozoon cuniculi GB-M1	XP_965923		
Vavraia culicis 'floridensis'	ELA48161		
Trachipleistophora hominis	ELQ73978		
Nosema bombycis CQ1	EOB14346	EOB11567	
Fungi; Mucoromycotina			
Rhizopus delemar RA 99-880	EIE77053		
Ichthyosporea; Capsaspora			
Capsaspora owczarzaki ATCC 30864	XP_004365757		
Amoebozoa			
Dictyostelium fasciculatum	XP_004352128		
Dictyostelium purpureum	XP_003285700		
Dictyostelium discoideum AX4	XP_641736		
Polysphondylium pallidum PN500	EFA75139	EFA78087	
Archaeplastida			
Arabidopsis thaliana	NP_201174		
Glycine max	NP_001237260		
Medicago truncatula	XP_003598314	XP_003598313	
Solanum lycopersicum	XP_004251872		
Sorghum bicolor	XP_002460865		
Zea mays	NP_001143606	NP_001159037	
Brachypodium distachyon	XP_003560085		
Oryza sativa Japonica Group	BAD31280		
Populus trichocarpa	XP_002320792		
Vitis vinifera	XP_002270604		
Physcomitrella patens subsp. Patens	XP_001773862	XP_001781701	
Selaginella moellendorffii	XP_002978813	XP_002972708	
Chlamydomonas reinhardtii CC3269	XP_001693165		
Volvox carteri f. nagariensis	XP_002950325		
Stramenopiles; Oomycetes			
Albugo laibachii Nc14	CCA17897		
Phytophthora sojae	EGZ16169		
Phytophthora infestans T30-4	XP_002908785		
Stramenopiles; Alveolata			
Paramecium tetraurelia strain d4-2	XP_001449003	XP_001444193	
Tetrahymena thermophila	XP_001021458		
Cryptophyta			
Guillardia theta	EKX42788		

Table S4. Table listing crenarchaeal-like single-stranded binding identified in eukaryotic genomes.

Appendix 4

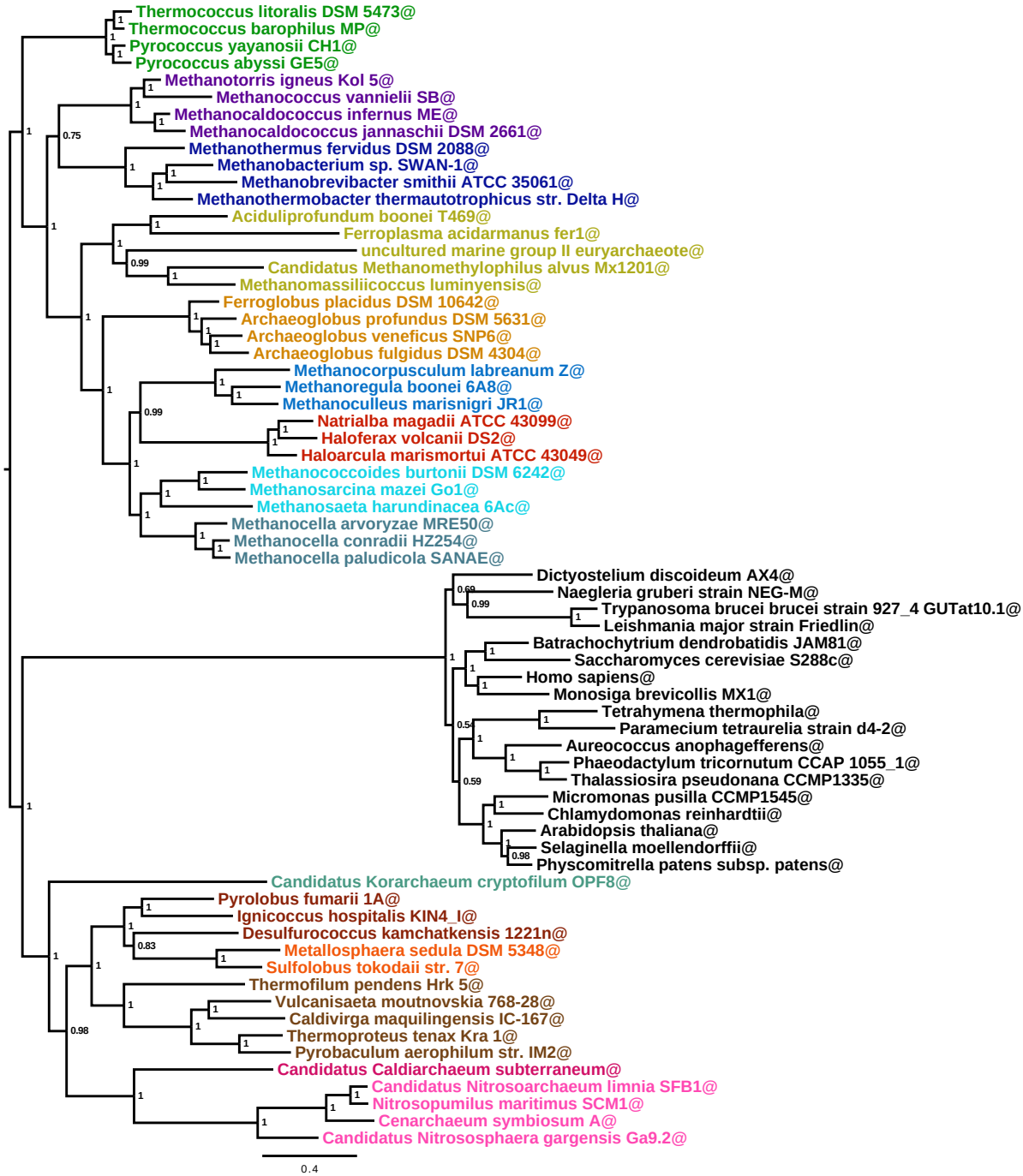
Supplementary Material Article 2

SM Figure 1



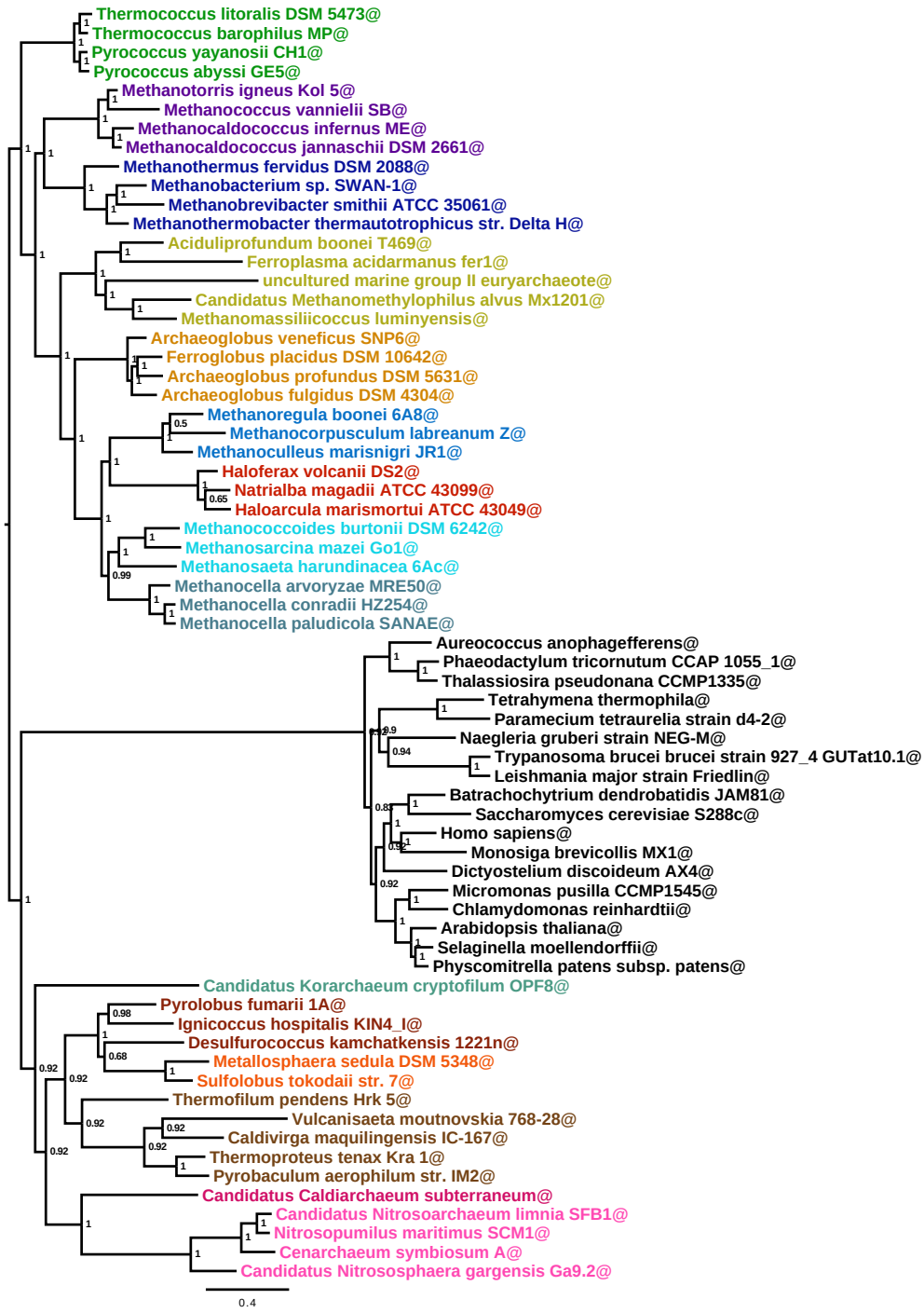
SM Figure 1. Unrooted Bayesian phylogeny of a concatenated data set of 72 proteins shared between archaea and eukaryotes. The tree was calculated by Phylobayes (CAT+Γ4) after recoding according Dayhoff6 amino acid categories. The scale bar represents the average number of substitutions per site. Values at nodes represent posterior probabilities.

SM Figure 2



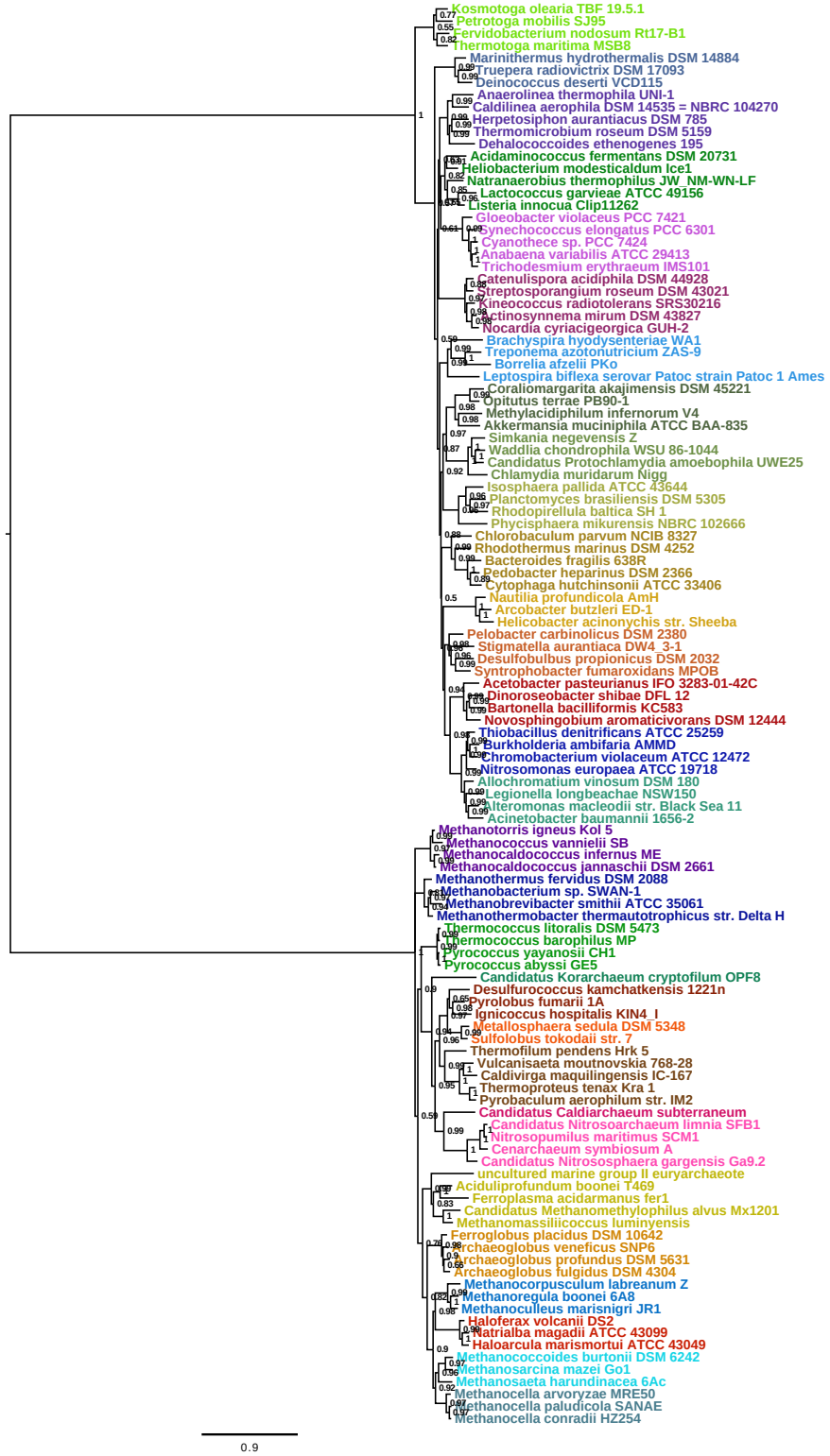
SM Figure 2. Unrooted Bayesian phylogeny of a concatenated data set of 35 proteins uniquely shared between archaea and eukaryotes (6629 amino acid positions). The tree was calculated by Phylobayes (CAT+Γ4). The scale bar represents the average number of substitutions per site. Values at nodes represent posterior probabilities.

SM Figure 3



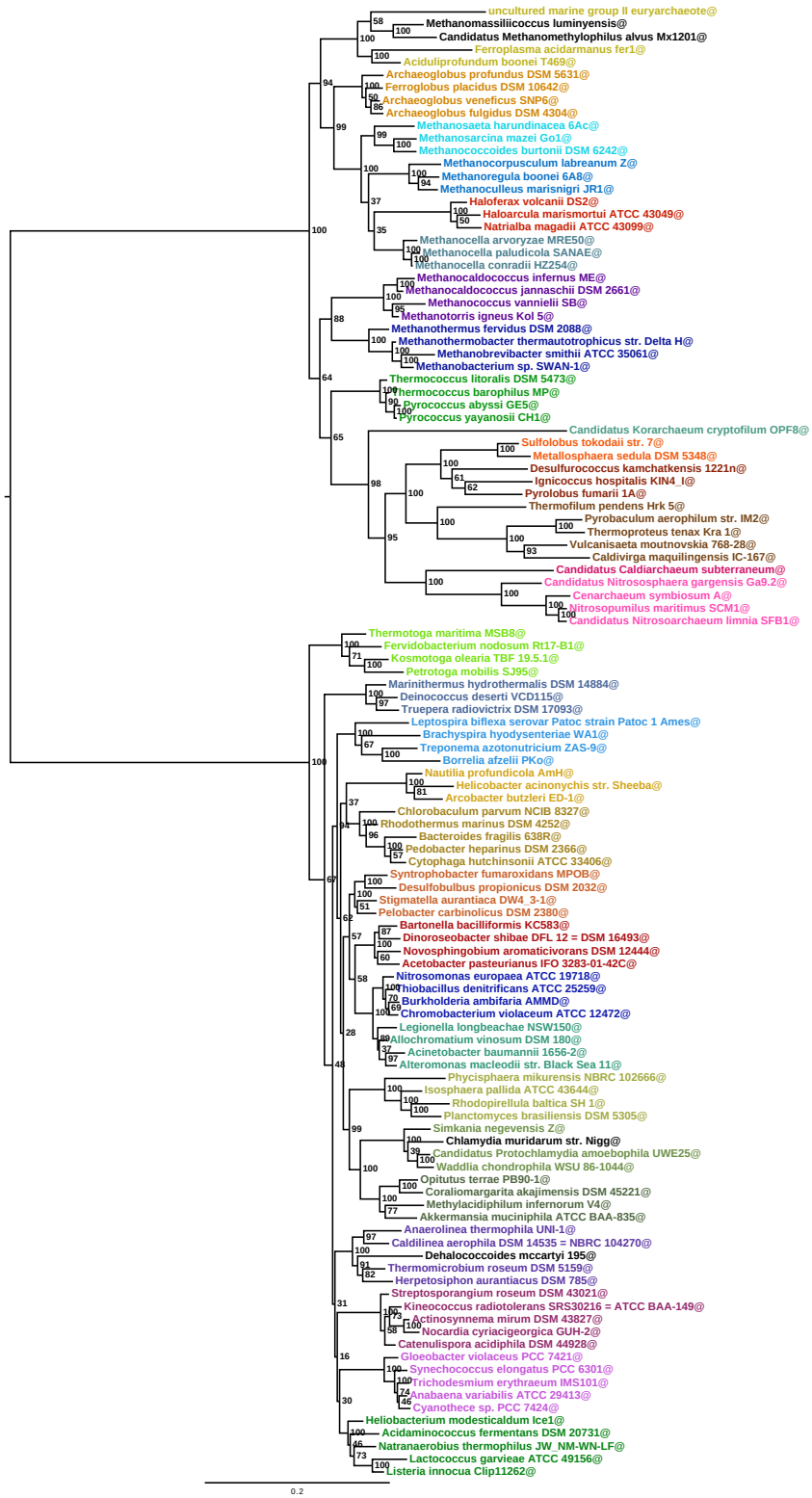
SM Figure 3. Unrooted Bayesian phylogeny of a concatenated data set of 37 universal proteins (10319 amino acid positions). The tree was calculated by Phylobayes (CAT+Γ4). The scale bar represents the average number of substitutions per site. Values at nodes represent posterior probabilities.

SM Figure 4



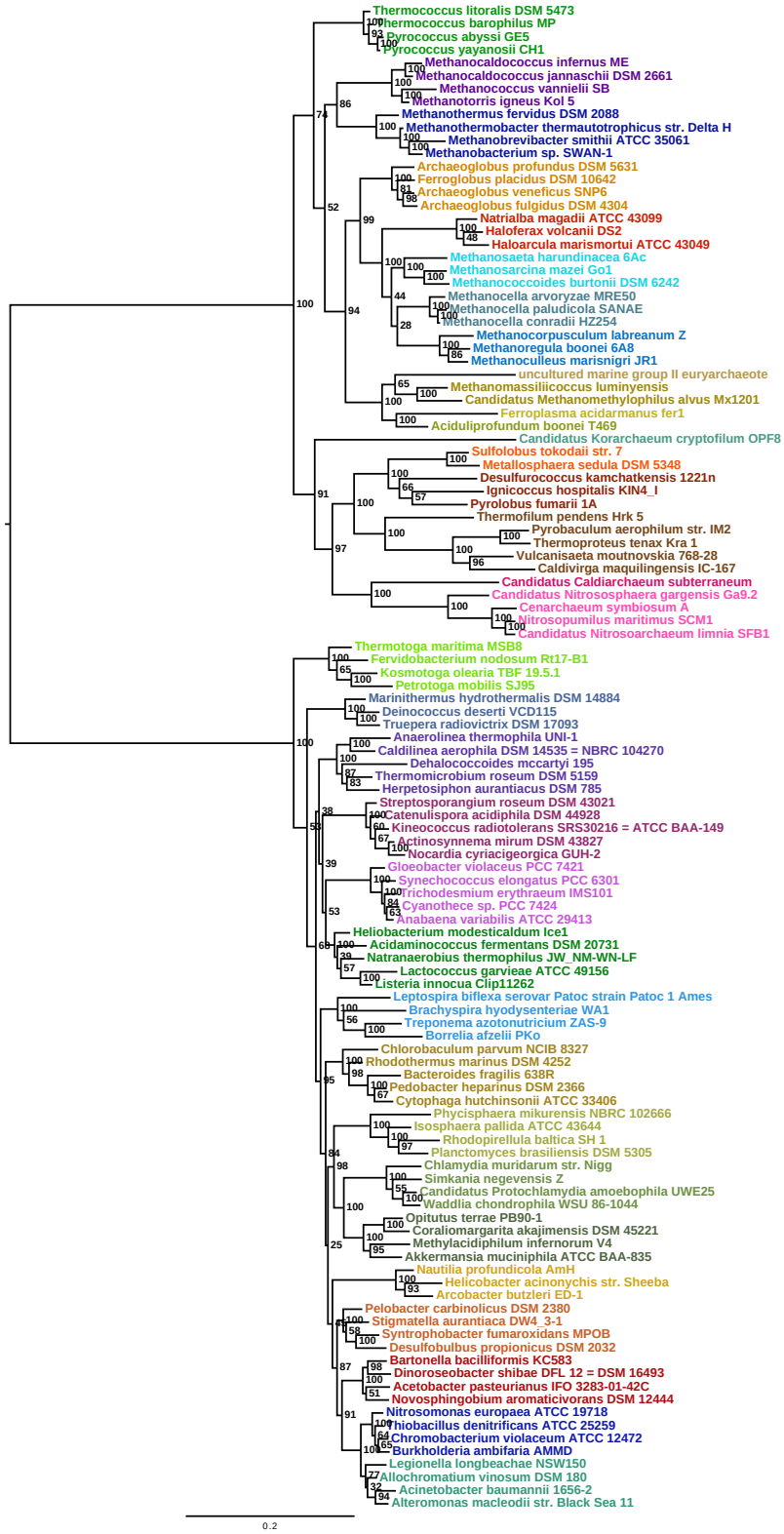
SM Figure 4. Unrooted Bayesian phylogeny of a concatenated data set of 46 proteins shared between archaea and bacteria. The tree was calculated by Phylobayes (CAT+Γ4) after recoding according Dayhoff6 amino acid categories. The scale bar represents the average number of substitutions per site. Values at nodes represent posterior probabilities.

SM Figure 5



SM Figure 5. Unrooted ML phylogeny of the AB SF11 supermatrix (3497 amino acid positions). The tree was calculated by PhyML (LG+Γ4). The scale bar represents the average number of substitutions per site. Values at nodes represent supports obtained by non parametric bootstrapping based on 100 resamplings of the original alignment.

SM Figure 6



SM Figure 6. Unrooted ML phylogeny of the AB SF12 supermatrix (3776 amino acid positions). The tree was calculated by PhyML (LG+Γ4). The scale bar represents the average number of substitutions per site. Values at nodes represent supports obtained by non parametric bootstrapping based on 100 resamplings of the original alignment.

SM Table 1

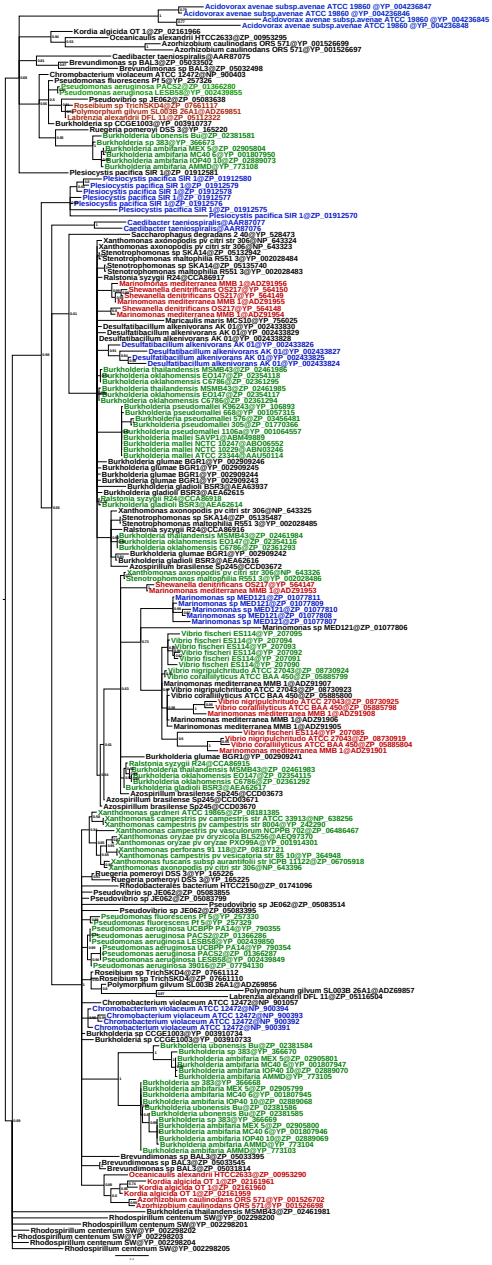
TaxId	Organism	Domain	Phylum	Class	Order	Family
446462	Actinosynnema mirum DSM 43827	Bacteria	Actinobacteria	Actinobacteria	Actinomycetales	Pseudonocardiaceae
479433	Catenulispora acidiphila DSM 44928	Bacteria	Actinobacteria	Actinobacteria	Actinomycetales	Catenulisporaceae
266940	Kineococcus radiotolerans SRS30216	Bacteria	Actinobacteria	Actinobacteria	Actinomycetales	Kineosporiaceae
1127134	Nocardia cyriacigeorgica	Bacteria	Actinobacteria	Actinobacteria	Actinomycetales	Nocardiaceae
479432	Streptosporangium roseum DSM 43021	Bacteria	Actinobacteria	Actinobacteria	Actinomycetales	Streptosporangiaceae
862962	Bacteroides fragilis 638R	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae
269798	Cytophaga hutchinsonii ATCC 33406	Bacteria	Bacteroidetes	Cytophagia	Cytophagales	Cytophagaceae
485917	Pedobacter heparinus DSM 2366	Bacteria	Bacteroidetes	Sphingobacteriia	Sphingobacteriales	Sphingobacteriaceae
518766	Rhodothermus marinus DSM 4252	Bacteria	Bacteroidetes	NA	Bacteroidetes Order II. I	Rhodothermaceae
517417	Chlorobaculum parvum NCIB 8327	Bacteria	Chlorobi	Chlorobia	Chlorobiales	Chlorobiaceae
926569	Anaerolinea thermophila UNI-1	Bacteria	Chloroflexi	Anaerolineae	Anaerolineales	Anaerolineaceae
926550	Caldilinea aerophila DSM 14535 = NBRC 104270	Bacteria	Chloroflexi	Caldilineae	Caldilineales	Caldilineaceae
243164	Dehalococcoides ethenogenes 195	Bacteria	Chloroflexi	Dehalococcoidetes	Dehalococcoidales	Dehalococcoidaceae
316274	Herpetosiphon auranticacus DSM 785	Bacteria	Chloroflexi	Chloroflexi	Herpetosiphonales	Herpetosiphonaceae
309801	Thermomicrobium roseum DSM 5159	Bacteria	Chloroflexi	Thermomicrobia	Thermomicrobiales	Thermomicrobiaceae
240292	Anabaena variabilis ATCC 29413	Bacteria	Cyanobacteria	NA	Nostocales	Nostocaceae
65393	Cyanothece sp. PCC 7424	Bacteria	Cyanobacteria	NA	Chroococcales	NA
251221	Gloeobacter violaceus PCC 7421	Bacteria	Cyanobacteria	Gloeobacteria	Gloeobacterales	NA
269084	Synechococcus elongatus PCC 6301	Bacteria	Cyanobacteria	NA	Chroococcales	NA
203124	Trichodesmium erythraeum IMS101	Bacteria	Cyanobacteria	NA	Oscillatoriales	NA
546414	Deinococcus deserti VCD115	Bacteria	Deinococcus-Thermus	Deinococci	Deinococcales	Deinococcaceae
869210	Marinithermus hydrothermalis DSM 14884	Bacteria	Deinococcus-Thermus	Deinococci	Thermales	Thermaceae
649638	Truepera radiovictrix DSM 17093	Bacteria	Deinococcus-Thermus	Deinococci	Deinococcales	Trueperaceae
591001	Acidaminococcus fermentans DSM 20731	Bacteria	Firmicutes	Negativicutes	Selenomonadales	Acidaminococcaceae
498761	Heliobacterium modesticaldum Ice1	Bacteria	Firmicutes	Clostridia	Clostridiales	Heliobacteriaceae
420889	Lactococcus garvieae ATCC 49156	Bacteria	Firmicutes	Bacilli	Lactobacillales	Streptococcaceae
272626	Listeria innocua Clip11262	Bacteria	Firmicutes	Bacilli	Bacillales	Listeriaceae
457570	Natranaerobius thermophilus JW_NM-WN-LF	Bacteria	Firmicutes	Clostridia	Natranaerobiales	Natranaerobiaceae
264201	Candidatus Protochlamydia amoebophila UWE25	Bacteria	PVC_Chlamydiae	Chlamydiia	Chlamydiales	Parachlamydiaceae
243161	Chlamydia muridarum Nigg	Bacteria	PVC_Chlamydiae	Chlamydiia	Chlamydiales	Chlamydiaceae
331113	Simkania negevensis Z	Bacteria	PVC_Chlamydiae	Chlamydiia	Chlamydiales	Simkaniaceae
716544	Waddlia chondrophila WSU 86-1044	Bacteria	PVC_Chlamydiae	Chlamydiia	Chlamydiales	Waddliaceae
575540	Isosphaera pallida ATCC 43644	Bacteria	PVC_Planctomycetes	Planctomycetia	Planctomycetales	Planctomycetaceae
1142394	Phycisphaera mikurensis NBRC 102666	Bacteria	PVC_Planctomycetes	Phycisphaerae	Phycisphaerales	Phycisphaeraceae
756272	Planctomyces brasiliensis DSM 5305	Bacteria	PVC_Planctomycetes	Planctomycetia	Planctomycetales	Planctomycetaceae
243090	Rhodopirellula baltica SH 1	Bacteria	PVC_Planctomycetes	Planctomycetia	Planctomycetales	Planctomycetaceae
349741	Akkermansia muciniphila ATCC BAA-835	Bacteria	PVC_Verrucomicrobia	Verrucomicrobiae	Verrucomicrobiales	Verrucomicrobiaceae
583355	Coraliomargarita akajimensis DSM 45221	Bacteria	PVC_Verrucomicrobia	Opitutae	Puniceicoccales	Puniceicoccaceae
481448	Methylacidiphilum infernorum V4	Bacteria	PVC_Verrucomicrobia	NA	Methylacidiphiliales	Methylacidiphilaceae
452637	Opitutus terrae PB90-1	Bacteria	PVC_Verrucomicrobia	Opitutae	NA	Opitutaceae
634452	Acetobacter pasteurianus IFO 3283-01	Bacteria	Proteobacteria	Alphaproteobacteria	Rhodospirillales	Acetobacteraceae
360095	Bartonella bacilliformis KC583	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Bartonellaceae
398580	Dinoroseobacter shibae DFL 12	Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae
279238	Novosphingobium aromaticivorans DSM 12444	Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae
339670	Burkholderia ambifaria AMMD	Bacteria	Proteobacteria	Betaproteobacteria	Burkholderiales	Burkholderiaceae
243365	Chromobacterium violaceum ATCC 12472	Bacteria	Proteobacteria	Betaproteobacteria	Neisseriales	Neisseriaceae
228410	Nitrosomonas europaea ATCC 19718	Bacteria	Proteobacteria	Betaproteobacteria	Nitrosomonadales	Nitrosomonadaceae
292415	Thiobacillus denitrificans ATCC 25259	Bacteria	Proteobacteria	Betaproteobacteria	Hydrogenophilales	Hydrogenophilaceae
577650	Desulfobulbus propionicus DSM 2032	Bacteria	Proteobacteria	Deltaproteobacteria	Desulfobacterales	Desulfobulbaceae
338963	Pelobacter carbinolicus DSM 2380	Bacteria	Proteobacteria	Deltaproteobacteria	Desulfuromonadales	Pelobacteraceae
378806	Stigmatella aurantiaca DW4_3-1	Bacteria	Proteobacteria	Deltaproteobacteria	Myxococcales	Cystobacteraceae
335543	Syntrophobacter fumaroxidans MPOB	Bacteria	Proteobacteria	Deltaproteobacteria	Syntrophobacterales	Syntrophobacteraceae
944546	Arcobacter butzleri ED-1	Bacteria	Proteobacteria	Epsilonproteobacteri	Campylobacterales	Campylobacteraceae
382638	Helicobacter acinonychis str. Sheeba	Bacteria	Proteobacteria	Epsilonproteobacteri	Campylobacterales	Helicobacteraceae
598659	Nautilia profundicola AmH	Bacteria	Proteobacteria	Epsilonproteobacteri	Nautiliales	Nautiliaceae
696749	Acinetobacter baumannii 1656-2	Bacteria	Proteobacteria	Gammaproteobacteri	Pseudomonadales	Moraxellaceae
572477	Allochromatium vinosum DSM 180	Bacteria	Proteobacteria	Gammaproteobacteri	Chromatiales	Chromatiaceae
1004785	Alteromonas macleodii str. Black Sea 11	Bacteria	Proteobacteria	Gammaproteobacteri	Alteromonadales	Alteromonadaceae
661367	Legionella longbeachae NSW150	Bacteria	Proteobacteria	Gammaproteobacteri	Legionellales	Legionellaceae
390236	Borrelia afzelii PKo	Bacteria	Spirochaetes	Spirochaetia	Spirochaetales	Spirochaetaceae
565034	Brachyspira hyodysenteriae WA1	Bacteria	Spirochaetes	Spirochaetia	Spirochaetales	Brachyspiraceae
355278	Leptospira biflexa serovar Patoc strain Patoc 1 Ames	Bacteria	Spirochaetes	Spirochaetia	Spirochaetales	Leptospiraceae
545695	Treponema azotonutricium ZAS-9	Bacteria	Spirochaetes	Spirochaetia	Spirochaetales	Spirochaetaceae
381764	Fervidobacterium nodosum Rt17-B1	Bacteria	Thermotogae	Thermotogae	Thermotogales	Thermotogaceae
521045	Kosmotoga olearia TBF 19.5.1	Bacteria	Thermotogae	Thermotogae	Thermotogales	Thermotogaceae
403833	Petrotoga mobilis SJ95	Bacteria	Thermotogae	Thermotogae	Thermotogales	Thermotogaceae
243274	Thermotoga maritima MSB8	Bacteria	Thermotogae	Thermotogae	Thermotogales	Thermotogaceae
490899	Desulfurococcus kamchatkensis 1221n	Archaea	Crenarchaeota	Thermoprotei	Desulfurococcales	Desulfurococcaceae
453591	Ignicoccus hospitalis KIN4_I	Archaea	Crenarchaeota	Thermoprotei	Desulfurococcales	Desulfurococcaceae
694429	Pyrolobus fumarii 1A	Archaea	Crenarchaeota	Thermoprotei	Desulfurococcales	Pyrodictiaceae
399549	Metallosphaera sedula DSM 5348	Archaea	Crenarchaeota	Thermoprotei	Sulfolobales	Sulfolobaceae
273063	Sulfolobus tokodaii str. 7	Archaea	Crenarchaeota	Thermoprotei	Sulfolobales	Sulfolobaceae
368408	Thermofilum pendens Hrk 5	Archaea	Crenarchaeota	Thermoprotei	Thermoproteales	Thermofilaceae
397948	Caldivirga maquilingensis IC-167	Archaea	Crenarchaeota	Thermoprotei	Thermoproteales	Thermoproteaceae
178306	Pyrobaculum aerophilum str. IM2	Archaea	Crenarchaeota	Thermoprotei	Thermoproteales	Thermoproteaceae
768679	Thermoproteus tenax Kra 1	Archaea	Crenarchaeota	Thermoprotei	Thermoproteales	Thermoproteaceae
985053	Vulcanisaeta moutnovskia 768-28	Archaea	Crenarchaeota	Thermoprotei	Thermoproteales	Thermoproteaceae
224325	Archaeoglobus fulgidus DSM 4304	Archaea	Euryarchaeota	Archaeoglobi	Archaeoglobales	Archaeoglobaceae
572546	Archaeoglobus profundus DSM 5631	Archaea	Euryarchaeota	Archaeoglobi	Archaeoglobales	Archaeoglobaceae
693661	Archaeoglobus veneficus SNP6	Archaea	Euryarchaeota	Archaeoglobi	Archaeoglobales	Archaeoglobaceae
589924	Ferroglobus placidus DSM 10642	Archaea	Euryarchaeota	Archaeoglobi	Archaeoglobales	Archaeoglobaceae
272569	Haloarcula marismortui ATCC 43049	Archaea	Euryarchaeota	Halobacteria	Halobacteriales	Halobacteriaceae
309800	Haloferax volcanii DS2	Archaea	Euryarchaeota	Halobacteria	Halobacteriales	Halobacteriaceae
547559	Natrialba magadii ATCC 43099	Archaea	Euryarchaeota	Halobacteria	Halobacteriales	Halobacteriaceae
868131	Methanobacterium sp. SWAN-1	Archaea	Euryarchaeota	Methanobacteria	Methanobacteriales	Methanobacteriaceae
420247	Methanobrevibacter smithii ATCC 35061	Archaea	Euryarchaeota	Methanobacteria	Methanobacteriales	Methanobacteriaceae
187420	Methanothermobacter thermautotrophicus str. Delta H	Archaea	Euryarchaeota	Methanobacteria	Methanobacteriales	Methanobacteriaceae
523846	Methanothermus fervidus DSM 2088	Archaea	Euryarchaeota	Methanobacteria	Methanobacteriales	Methanothermaceae
573063	Methanocaldococcus infernus ME	Archaea	Euryarchaeota	Methanococci	Methanococcales	Methanocaldococcaceae
243232	Methanocaldococcus jannaschii DSM 2661	Archaea	Euryarchaeota	Methanococci	Methanococcales	Methanocaldococcaceae
880724	Methanotorris igneus Kol 5	Archaea	Euryarchaeota	Methanococci	Methanococcales	Methanocaldococcaceae
406327	Methanococcus vannielii SB	Archaea	Euryarchaeota	Methanococci	Methanococcales	Methanococcaceae
351160	Methanocella arvoryzae MRE50	Archaea	Euryarchaeota	Methanomicrobia	Methanocellales	Methanocellaceae
1041930	Methanocella conradii HZ254	Archaea	Euryarchaeota	Methanomicrobia	Methanocellales	Methanocellaceae
304371	Methanocella paludicola SANAE	Archaea	Euryarchaeota	Methanomicrobia	Methanocellales	Methanocellaceae
410358	Methanocorpusculum labreanum Z	Archaea	Euryarchaeota	Methanomicrobia	Methanomicrobiales	Methanocorpusculaceae
368407	Methanoculleus marisnigri JR1	Archaea	Euryarchaeota	Methanomicrobia	Methanomicrobiales	Methanomicrobiaceae
456442	Methanoregula boonei 6A8	Archaea	Euryarchaeota	Methanomicrobia	Methanomicrobiales	Methanoregulaceae
1110509	Methanosaeta harundinacea 6Ac	Archaea	Euryarchaeota	Methanomicrobia	Methanosarcinales	Methanosaetaceae
259564	Methanococcoides burtonii DSM 6242	Archaea	Euryarchaeota	Methanomicrobia	Methanosarcinales	Methanosarcinaceae
192952	Methanosarcina mazei Go1	Archaea	Euryarchaeota	Methanomicrobia	Methanosarcinales	Methanosarcinaceae
1080712	Methanomassiliicoccus luminyensis B10	Archaea	Euryarchaeota	Methanomicrobia	NA	NA
439481	Aciduliprofundum boonei T469	Archaea	Euryarchaeota	NA	NA	NA
1236689	Candidatus Methanomethylophilus alvus Mx1201	Archaea	Euryarchaeota	NA	NA	NA
274854	uncultured marine group II euryarchaeote	Archaea	Euryarchaeota	NA	NA	NA
333146	Ferroplasma acidarmanus fer1	Archaea	Euryarchaeota	Thermoplasmata	Thermoplasmatales	Ferroplasmaceae
272844	Pyrococcus abyssi GE5	Archaea	Euryarchaeota	Thermococci	Thermococcales	Thermococcaceae
529709	Pyrococcus yayanosii CH1	Archaea	Euryarchaeota	Thermococci	Thermococcales	Thermococcaceae
391623	Thermococcus barophilus MP	Archaea	Euryarchaeota	Thermococci	Thermococcales	Thermococcaceae
523849	Thermococcus litoralis DSM 5473	Archaea	Euryarchaeota	Thermococci	Thermococcales	Thermococcaceae
374847	Candidatus Korarchaeum cryptofilum OPF8	Archaea	Korarchaeota	NA	NA	NA
414004	Cenarchaeum symbiosum A	Archaea	Thaumarchaeota	NA	Cenarchaeales	Cenarchaeaceae
311458	Candidatus Caldiarchaeum subterraneum	Archaea	Thaumarchaeota	NA	NA	NA
886738	Candidatus Nitrosoarchaeum limnia SFB1	Archaea	Thaumarchaeota	NA	Nitrosopumilales	Nitrosopumilaceae
436308	Nitrosopumilus maritimus SCM1	Archaea	Thaumarchaeota	NA	Nitrosopumilales	Nitrosopumilaceae
1237085	Candidatus Nitrososphaera gargensis Ga9.2	Archaea	Thaumarchaeota	NA	Nitrososphaerales	Nitrososphaeraceae
412030	Paramecium tetraurelia strain d4-2	Eukaryota	Alveolata	Oligohymenophorea	Peniculida	Paramecilidae
5911	Tetrahymena thermophila SB210	Eukaryota	Alveolata	Oligohymenophorea	Hymenostomatida	Tetrahymenidae
44056	Aureococcus anophagefferens	Eukaryota	Stramenopiles	Pelagophyceae	NA	NA
556484	Phaeodactylum tricornutum CCAP 1055_1	Eukaryota	Stramenopiles	Bacillariophyceae	Naviculales	Phaeodactylaceae
296543	Thalassiosira pseudonana CCMP1335	Eukaryota	Stramenopiles	Coscinodiscophyceae	NA	Thalassiosiraceae
3702	Arabidopsis thaliana	Eukaryota	Viridiplantae	NA	Brassicales	Brassicaceae
3055	Chlamydomonas reinhardtii CC3269	Eukaryota	Viridiplantae	Chlorophyceae	Chlamydomonadales	Chlamydomonadaceae
564608	Micromonas pusilla CCMP1545	Eukaryota	Viridiplantae	Mamiellophyceae	Mamiellales	NA
145481	Physcomitrella patens subsp. patens	Eukaryota	Viridiplantae	Bryopsida	Funariales	Funariaceae
88036	Selaginella moellendorffii	Eukaryota	Viridiplantae	Isoetopsida	Selaginellales	Selaginellaceae
347515	Leishmania major strain Friedlin	Eukaryota	Euglenozoa	NA	Kinetoplastida	Trypanosomatidae
999953	Trypanosoma brucei brucei strain 927_4 GUTat10.1	Eukaryota	Euglenozoa	NA	Kinetoplastida	Trypanosomatidae
744533	Naegleria gruberi strain NEG-M	Eukaryota	Heterolobosea	Heterolobosea	Schizopyrenida	Vahlkampfiidae
352472	Dictyostelium discoideum AX4	Eukaryota	Amoebozoa	NA	Dictyostelida	NA
431895	Monosiga brevicollis MX1	Eukaryota	Choanoflagellida	NA	Choanoflagellida	Codonosigidae
684364	Batrachochytrium dendrobatidis JAM81	Eukaryota	Fungi	Chytridiomycetes	Rhizophydiales	NA
559292	Saccharomyces cerevisiae S288c	Eukaryota	Fungi	Saccharomycetes	Saccharomycetales	Saccharomycetaceae
9606	Homo sapiens	Eukaryota	Metazoa	Mammalia	Primates	Hominidae

SM Table 1. Table showing monophyletic groups (16 bacterial phyla, 12 archaeal orders/phyla, and 4 eukaryotic groups) that we defined for the slow-fast method. Each group contains between 3-5 taxa (with the exception of Korarchaeota that was considered alone). The groups are highlighted alternately in gray and white shading.

Appendix 5

Supplementary Material Article 3

SM Figure 1



Supplementary Figure 1. Phylogeny of Reb homologs. Unrooted Bayesian phylogenetic tree of 203 Reb homologs. The tree was obtained by using Phylobayes 3.3 with the LG model of amino acid substitution and a discrete gamma distribution with four categories to take into account among-site rate variation. Note that the tree is not fully resolved due to the small number of positions that could be used for analysis (73 amino acids). Numbers at nodes represent posterior probabilities. The scale bar represents the average number of substitutions per site. Colored taxa correspond to the rebs represented in Figure 5.

Abstract :

Reconstructing the evolutionary relationships between Archaea and Eukaryotes: a phylogenomic approach

It is widely accepted that there exist an evolutionary relationship between Archaea and Eukaryotes, but the exact nature of this relationship is hotly debated. In this thesis I have taken advantage of the large available genomic data to investigate the issue through two complementary phylogenomic approaches: (i) the analysis of a specific archaeal cellular system with an evolutionary link to eukaryotes, and (ii) a large-scale phylogenomic analysis at the level of the three domains of life.

In the first study, I carried out a detailed analysis of a cellular system with an evolutionary link between Archaea and Eukaryotes, DNA replication. I performed an exhaustive phylogenomic analysis of the components of DNA replication in all complete archaeal genomes. This allowed me to accurately assign them in terms of orthology, paralogy, horizontal gene transfers, and copies originating from mobile elements. My results provide a full picture of the diversity of DNA replication among different lineages, and allowed me to infer the presence of a modern-type DNA replication machinery in the last archaeal common ancestor. I was able to clarify the evolutionary history that shaped this key cellular machinery during archaeal diversification. My study allowed me to highlight a new set of markers that provide information on yet unclear evolutionary relationships within archaea. In addition, I analyzed, for the first time, the phylogenetic signal carried by DNA replication components. This is highly consistent with that harbored by two other key informational machineries, translation and transcription, strengthening the existence of a robust organismal tree for the Archaea. Finally, most of the components inferred to have been present in the archaeal ancestor are shared with eukaryotes, allowing discussion on the evolutionary relationships between Archaea and Eukaryotes. My results provide important and useful information for future functional studies on DNA replication components in the archaea.

In the second study, I have performed a large-scale analysis to study the relationships between archaea and eukaryotes. I applied a novel strategy, which involved extracting and analyzing separately two sets of markers conserved between archaea/eukaryotes on one side, and the archaea/bacteria on the other. I carried out an extensive analysis on all complete archaeal genomes involving the extraction of core protein families, detection of orthologs, and exhaustive searches on bacteria and eukaryal genomes. I identified 72 reliable phylogenetic makers shared between archaea and eukaryotes and 46 shared between bacteria and archaea. By analyzing the archaeal/eukaryote dataset I was able to exclude the origin of eukaryotes from within any of the archaeal orders or phyla. When I analyzed the bacterial/archaeal dataset I could highlight an original and new rooting for the archaeal tree that lies within the Euryarchaeota. My analysis would support an emergence of Eukaryotes from within the Archaea and also dramatically change our view on the phylogeny and evolution of the archaeal domain. Finally, I performed an exhaustive search for homologs of eukaryotic proteins inferred to date back to the Last Eukaryotic Common Ancestor that are not universally present in the Archaea. I identified several new markers that were not reported in previous analyses that belong to a wide range of cellular processes. The origin of these markers could open up new avenues of research for future studies.

My results contribute to expanding our view of the diversity of the Archaea, their phylogeny, and their early evolution, including the nature of the LACA. They also shed light on one of the major questions in Evolutionary Biology: the topology of the universal Tree of Life and the origin of the eukaryotic lineage.