Deep archaeal phylogeny and evolutionary dynamics of DPANNs

Brittany Baker

université PARIS-SACLAY

Deep archaeal phylogeny and the evolutionary dynamics of DPANN

Phylogénie profonde des archées et dynamique évolutive des DPANN

Thèse de doctorat de l'Université Paris-Saclay

École doctorale n° 577 : Structure et Dynamique des Systèmes Vivants (SDSV)
Spécialité de doctorat : Évolution
Graduate School : Sciences de la Vie et Santé
Référent : Faculté des Sciences d'Orsay

Thèse préparée dans l'unité de recherche **Écologie, Systématique, et Évolution (Université Paris-Saclay, CNRS, AgroParisTech)**, sous la direction de **David MOREIRA**, directeur de recherche, la co-direction de **Laura EME**, chargée de recherche, et **Purificación LÓPEZ-GARCÍA**, directrice de recherche

Thèse soutenue à Paris-Saclay, le 01 Mars 2024, par

Brittany Anne BAKER

Composition du Jury

Membres du jury avec voix délibérative

Graeme NICOL Directeur de recherche École Centrale de Lyon	Président
Anja SPANG Directrice de recherche NIOZ	Rapportrice et Examinatrice
Ana BELEN MARTIN-CUADRADO Professeure Universidad de Alicante	Rapportrice et Examinatrice
Tamara BASTA-LE BERRE Maîtresse de conférences I2BC	Examinatrice
Guillaume BORREL Chargé de recherche Institut Pasteur	Examinateur

THÈSE DE DOCTORAT

NNT : 2024UPASL007


cnrs

erc

deem

ESE
Ecologie Systématique Évolution

université PARIS-SACLAY

ÉCOLE DOCTORALE
Structure et dynamique des systèmes vivants (SDSV)

Titre : Phylogénie profonde des archées et dynamique évolutive des DPANN

Mots clés : archées, évolution, phylogénie

Résumé : Les archées constituent l'un des trois grands domaines du vivant, avec les bactéries et les eucaryotes, et sont présentes dans presque tous les environnements terrestres connus. Néanmoins, l'arbre des archées n'est pas encore complètement résolu, limitant notre compréhension de l'évolution de ces organismes. Par exemple, on ne sait pas combien de fois les archées se sont adaptées aux environnements hypersalins. Selon la position phylogénétique des groupes d'halophiles actuellement connus, ce nombre peut aller de un à quatre. De même, les archées DPANN constituent l'un des quatre supergroupes majeurs d'archées avec les TACK, les Asgard et les Euryarchaeota, mais leur monophylie et leur position phylogénétique restent très débattues. Ces archées de taille nanométrique sont généralement considérées comme des épibiontes dont la croissance et la survie sont obligatoirement liées à une autre archée hôte. La position phylogénétique des DPANN a des implications importantes pour comprendre le rythme et le mode d'évolution de ces archées épi symbiotiques mais aussi des autres grands groupes d'archées.

L'objectif principal de ma thèse était de mieux comprendre l'histoire évolutive des archées halophiles et DPANN avec des analyses phylogénétiques approfondies. Mes recherches ont révélé que les archées halophiles se sont adaptées indépendamment aux environnements hypersalins au moins quatre fois. Dans le cadre de ce projet, deux nouvelles familles halophiles extrêmes, les Afararchaeaceae et les Asbonarchaeaceae, ont été identifiées dans des lacs hypersalins de la dépression de Danakil (Éthiopie). Mes analyses ont également clarifié les incohérences phylogénétiques antérieures, soulignant que le contenu très biaisé en acides aminés chez les halophiles conduit à des artefacts phylogénétiques. En minimisant ces biais, j'ai obtenu des placements phylogénétiques plus fiables. J'ai également reconstruit l'histoire évolutive des familles de gènes d'archées en identifiant des événements tels que les duplications, les transferts, les origines et les pertes de gènes à l'aide de méthodes de réconciliation d'arbres. Je me suis concentrée sur les événements spécifiques aux branches menant aux différentes lignées halophiles. Ces résultats suggèrent que la duplication et le transfert horizontal de gènes ont joué un rôle important dans l'adaptation à l'halophilie, par exemple en propageant des gènes clés (tels que ceux codant pour les transporteurs de potassium) dans les différentes lignées halophiles extrêmes. En parallèle, j'ai cherché à élucider l'histoire évolutive des archées DPANN. Cela impliquait une étude complète de leur phylogénie, en utilisant un ensemble étendu de marqueurs protéiques conservés et un échantillonnage complet de taxons qui inclut des représentants des 11 phylums DPANN connus. Avec des méthodes pour atténuer l'impact potentiel des biais de composition et de l'attraction des longues branches, j'ai obtenu un soutien solide pour la monophylie des DPANN et leur placement au sein des Euryarchaeota. En outre, mes recherches ont révélé qu'au sein des DPANN, les Altiarchaeota, des organismes possiblement de vie libre, représentent la branche divergente la plus précoce. L'ensemble de ces résultats a montré que les archées DPANN sont un groupe monophylétique qui a évolué à partir d'un ancêtre Euryarchaeota de vie libre. Alors que les pipelines phylogénétiques automatisés sont capables de résoudre certaines questions phylogénétiques sur les archées, ce travail a montré que des analyses phylogénétiques approfondies sont encore nécessaires pour résoudre les principales branches de l'arbre des archées. Cette recherche a démontré que, malgré la connaissance de longue date des artefacts phylogénétiques tels que les biais de composition, il n'y a pas un seul biais qui puisse expliquer toutes les incohérences observées dans la phylogénie des archées.

Title : Deep archaeal phylogeny and evolutionary dynamics of DPANN

Keywords : archaea, evolution, phylogeny

Abstract : Archaea, one of life's three fundamental domains alongside Bacteria and Eucarya, thrive in nearly every habitat. Nevertheless, the precise structure of the archaeal tree of life remains unclear, clouding our understanding of the evolutionary history of this domain. For instance, it is unknown how many times archaea adapted to hypersaline environments. Depending on the phylogenetic placement of the currently known groups of halophiles, the number could range from one to four times. Similarly, the DPANN archaea are one of the four major archaeal supergroups along with the TACK, Asgard, and Euryarchaeota, yet their monophyly and phylogenetic placement in the archaeal tree remains unresolved. The DPANN archaea are typically classified as nano-sized archaea that grow obligately attached to another archaeon host for their growth and survival. Resolving the phylogenetic position of the DPANN has important implications for understanding the tempo and mode of the evolution of these episymbiotic archaea and other major archaeal clades.

The primary goal of my PhD was to conduct a thorough phylogenomic analysis of halophilic and DPANN archaea to gain deeper insights into their evolutionary history. My research revealed that halophilic archaea independently adapted to hypersaline environments at least four times. As part of this project, two novel family-level lineages of extreme halophiles, Afararchaeaceae and Asbonarchaeaceae, were identified from hypersaline lakes in the Danakil Depression in North-Eastern Ethiopia. My findings also clarified previous phylogenetic inconsistencies, highlighting that unique amino acid compositions in halophiles led to phylogenetic artifacts. By filtering out these biased data points, I achieved more consistent and reliable phylogenetic placements. I also reconstructed the evolutionary history of archaeal gene families by mapping events such as gene duplications, transfers, originations, and losses using gene tree-species tree reconciliation methods. I focused on events specific to the branches leading to the various halophilic lineages. These results suggested that gene duplication and horizontal gene transfer played an important role in the adaptation to halophily, for example, by spreading key genes (such as those encoding potassium transporters) across the various extremely halophilic lineages. In my second project, I aimed to elucidate the evolutionary history of the DPANN archaea. This involved a comprehensive study of their phylogeny by using an extensive set of conserved protein markers and a thorough taxon sampling that included representatives from all 11 known DPANN phyla. By employing various methods to mitigate the potential impact of compositional biases and long-branch attraction (LBA), I obtained robust support for the monophyly of the DPANN and their placement within the Euryarchaeota. Additionally, my research revealed that within the DPANN, the Altiarchaeota, potentially free-living, represent the earliest diverging branch. Together, these results showed that the DPANN archaea are a monophyletic group that evolved from a free-living Euryarchaeota ancestor. While automated phylogenetic pipelines can resolve some archaeal phylogenetic questions, this work has shown that in-depth phylogenomic analyses are still needed to resolve major branches of the archaeal tree. This research has demonstrated that, despite the long-standing awareness of phylogenetic artifacts like compositional sequence biases, there isn't a single bias that can explain all inconsistencies in the archaeal tree.

Acknowledgements

Moving to a new continent is not easy, but to do so during a global pandemic is something entirely different. I would like to thank the following people for making the hard times easier, and for making good times amazing.

First and foremost I want to thank my parents. **Mom** and **Dad**, you have supported me through 13 years of higher education. Moving to San Francisco at 17 years old to start university was not easy, but you always had my back no matter what. I truly don't think I would be where I am today without your never ending words of encouragement and motivation. **Mom**, you are always a shoulder to cry on during my hardest times and the one I want to drink wine with and go on beach hikes with on a Saturday afternoon. You have always supported me no matter what stage of life I am in and that truly has meant the world to me. **Dad**, your hard work and dedication to constantly bettering yourself has always been an inspiration to me since I was little. I don't think there are many dads out there that can help there daughters with their Bayesian homework, DIY projects, french lessons, vacation planning, and surf lessons. Thank you for always inspiring me. I love you guys so much.

Sister, you have been my best friend since I was 5 years old. I call you when I'm excited, when I am sad, when I am angry, and most importantly just because. You are so much wiser than your age and I am so incredibly proud of you. Thank you for always being down to go on vacation adventures and explore new cities and pretty houses. I hope in the future we will live closer to each other but until then, we will meet in the Bahamas :)

I would like to thank my french parent-in-laws, **Cathy and Pascal**. Depuis la première fois que je vous ai rencontrés, vous m'avez accueillie chez vous comme si j'étais votre propre fille. Vous m'avez fait découvrir la culture du Nord et je vous suis éternellement reconnaissante pour votre amour et votre soutien.

I want to thank all my California friends, **Sophie**, **Jill**, **Siena**, **Janae**, and **Cassie**. It has been hard the last 4 years me being so far away especially with the time difference, but everytime we are back together it is like no time has past. I want to especially thank **Sophie**, who has been my best friend since I was 9 years

old. You have been there for me at every stage of life and I am so glad I got to do those stages with you.

This PhD would not be possible without all my friends from the lab. **Jazmin**, **Feriel**, **Pauline**, **Fabian**, **Thomas**, **Kristina**, and **Romain** thank you so much for all your support the last 4 years. A special thank you to **Fabian** and **Jazmin** who were my first friends in France. Without you guys this experience would have not been the same! As I am writing this, I am waiting to meet you for lunch in Paris. Even outside of the lab, you guys are my best friends. To **Pauline**, thank you so much for always be willing to help me with French administration. I will never forget the care you and your mom showed me by going to the Ameli office in Cantal. To **Feriel**, thank you for always grabbing a coffee with me in the morning and listening to my daily problems. You always listen without judgement. To **Thomas**, thank you for always making me laugh and the effort you put in for making a fun work environment. Thank you for all your quizzes, I really loved them! Thank you to **Ana**, who helped me with all of my programming in the first years when I was not able to do it myself. The Halophiles project would not be possible without you. Lastly, thank you to all the previous DEEM members, **Jolien**, **Guillaume R.**, **Guillaume L.**, **Sergio**, **Naoji**, **Julien**, **Luis**, and **Guifre**. DEEM team would not be the same without you all!

A special thank you to my Master's advisor **Jose de la Torre** who encouraged me to pursue a PhD. You have believed in me as a scientist even when I didn't believe in myself. You brought me to international conferences when I was only a Master's student and I truly believe those experiences shaped me into the scientist I am today. Thank you for teaching me how to properly drink whiskey and most importantly for introducing me to archaea. I have no idea where I would be today if it wasn't for your first undergraduate course in microbial genomics.

A huge thank you to my current advisors, **David**, **Puri**, and **Laura**. You have lead me to become a strong and confident scientist. **David**, thank you for your constant dedication to my projects. You are always willing to talk through a problem I am having and your knowledge on the ecology, evolution, and phylogeny of the tree of life is truly inspiring. Thank you for always answering my emails with care and attention and for always giving me timely feedback on my writing. **Puri**, thank you for always pushing me to be a better scientist and for your dedication to the lab. You have taught me that words and details matter in presentations and in writing and I truly thank you for that. **Laura**, thank you for believing in me all those

years ago in Switzerland. That was truly a life changing moment for me. Thank you for your support during covid and your continued support during my PhD. Thank you for teaching me the importance of robust phylogenetics and introducing me to so many new methods. I am truly proud of the work I have done during my PhD and that would not have been possible without the support of my advisors.

I would like to thank my ESE mentor, **Jeanne Ropars**, and all of my collaborators, **Andrew Roger**, **Ed Susko**, **Charley McCarthy**, **Alvaro Rodriguez del Rio**, and **Jaime Huerta-Cepas**. It has been a pleasure to work with all of you.

I would also like to thank the members of my thesis committee, **Andrew Roger**, **Nicolas Lartillot**, and **Guillaume Borrel**, for their invaluable advice over the last four years, and the members of my jury **Anja Spang**, **Ana Belen Martin-Cuadrado**, **Tamara Basta**, **Graeme Nicol**, and **Guillaume Borrel** for taking the time to be apart of my thesis jury.

Last but not least, I would like to thank my partner and best friend **Cedric**. You have truly made my life happier being apart of it. No matter how stressed or overwhelmed I am, you are always there for me. We have gone on so many amazing adventures together, and I truly appreciate all that you do for me. I cannot express enough how happy our life together makes me. I love you.

Index

1. Introduction

1. Introduction

The word archaea is derived from the Greek word *archaios*, meaning “ancient” or “primitive.” Based on biogenic isotopes distinctive of different metabolisms, it is estimated that archaea were present at least 2.5 billion years ago (Gribaldo & Brochier-Armanet, 2006). My thesis is inspired by the question, “How can we understand 2.5 billion years of archaeal evolution with today's methods and data?”.

1.1. Archaea over the last 50 years

The classification of organisms dates back to ancient times, starting with Aristotle in the fourth century BC. He categorized animals based on common morphological characteristics and coined the term "genus" for each group, a term still used today. Until the 1960s, biologists followed Aristotle’s practices, classifying organisms by their phenotypic characteristics. This classification approach grouped all morphologically simple microorganisms (bacteria and archaea), first as the ‘Monera’ by Ernst Haeckel in 1866 and later as ‘Prokaryotes’ by Édouard Chatton in 1937 (Sapp, 2005). However, in the mid-1960s, a groundbreaking shift occurred in our understanding of microbial evolution with the introduction of molecular phylogeny based on comparing the sequences of biological macromolecules (nucleic acids and proteins) (C. R. Woese et al., 1975; Zuckerkandl & Pauling, 1965). These pioneering methods offered initial glimpses into the evolutionary relationships among microorganisms beyond their phenotypic traits and eventually led to the differentiation of two prokaryotic domains, the Archaea and Bacteria (previously Archaebacteria and Eubacteria, respectively; Fig. 1) (C. Woese, 1990).

These initial phylogenetic trees were based on the comparison of ribosomal RNA sequences. Ribosomal RNAs (rRNAs) are non-coding RNAs that are a primary component of the ribosome. The most notable feature of rRNAs is their high abundance in all cellular organisms, making them relatively easy to extract (Van de Peer et al., 1996). rRNA sequences also change slowly over time, allowing the detection of relatedness over very distant species (Woese et al. 1978). The universality of rRNAs allowed for the phylogenetic classification of life into the three primary domains (the Eucarya, Archaea, and Bacteria) (Fig. 1). Additionally, rRNA sequencing also proved to be a compelling approach for identifying novel microbes from the environment that could not be cultured (Pace, 2009). Historically, studying microbes involved the traditional approach of enriching and isolating them individually. However, routine culturing methods often fail to capture a significant portion of microbes, a phenomenon referred to as "The Great Plate Count Anomaly" (Staley and Konopka 1985) and later described as “The Uncultured Majority” (Rappé & Giovannoni, 2003). The molecular analyses of environmental 16S rRNA sequences substantially expanded our knowledge of the diversity of microbial life and introduced the era of culture-independent studies (Pace 2009; Amann et al. 1995; Hugenholtz 2002; Barns et al. 1996; Pace 1997). However, as more rRNA sequences were added to the tree of life, it became apparent that rRNA phylogenies did not provide enough resolution for deep phylogenetic relationships (Brochier-Armanet et al. 2011).

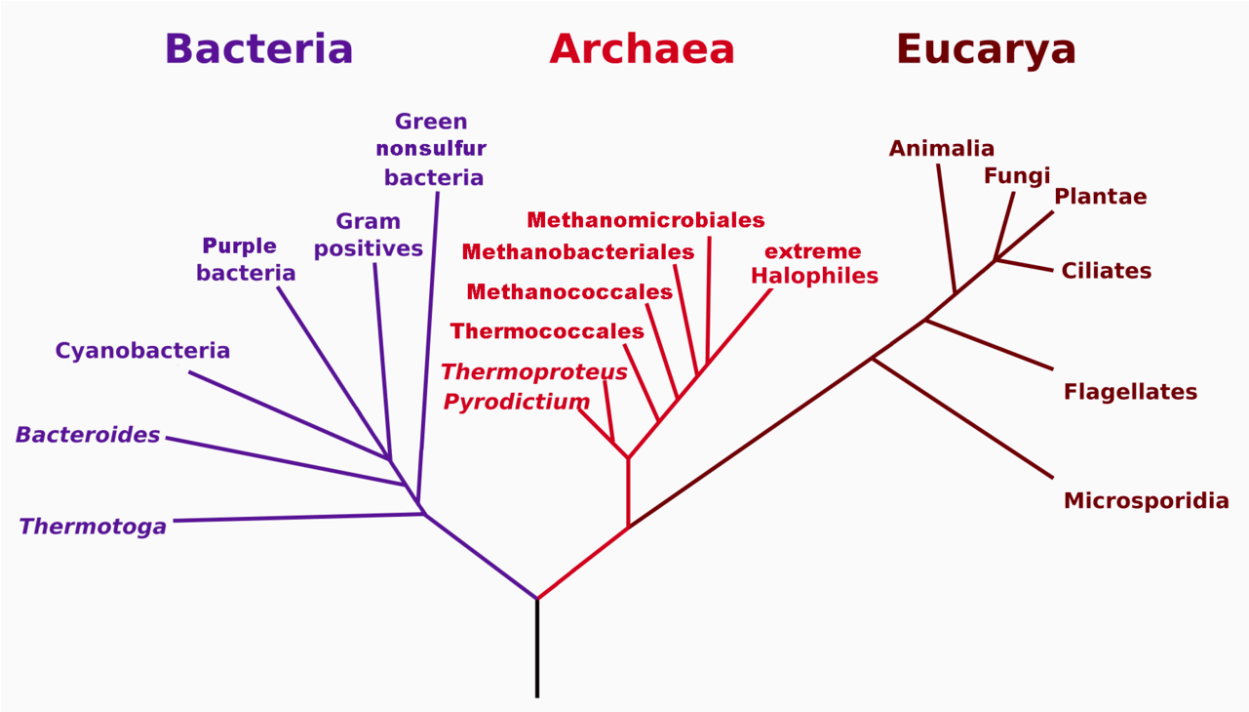


Figure 1 | The first rooted three-domain tree of life adapted from the seminal paper by Carl Woese in 1990. This first-rooted three-domain tree of life marked a pivotal moment in understanding microbial evolution. The tree was built from rRNA sequence comparisons and was rooted between the Bacteria and Archaea. This tree also showed that the Archaea were divided into two main kingdoms, the Euryarchaeota and Crenarchaeota. The Crenarchaeota were considered exclusively thermophiles, while the Euryarchaeota consisted of a more heterogeneous group of methanogens, extreme halophiles, and sulfate reducers.

1.2. Archaea in the Genomics Era

Archaea research entered a new era in 1996 with the whole-genome shotgun sequencing of the first archaeal genome, *Methanococcus* (now *Methanocaldococcus*) *jannaschii* (Bult et al., 1996). *Methanocaldococcus jannaschii* is a methanogenic Euryarchaeota lineage collected from a white chimney smoker 2,600 meters below

sea level (Bult et al., 1996). This study not only provided the initial insight into the coding potential of archaea but also reaffirmed their distinct status as a unique domain of life. In early comparative genomic studies, *Methanocaldococcus jannaschii* was analyzed alongside the bacteria *Escherichia coli* and *Synechocystis sp.* 6803 and the yeast *Saccharomyces cerevisiae* (Rivera et al., 1998). This study confirmed that eukaryotes exhibit a combination of archaeal and bacterial genes, with informational genes involved in translation, transcription, and replication showing a greater similarity to archaeal genes. In contrast, operational genes responsible for amino acid synthesis, cofactor biosynthesis, cell envelope formation, energy metabolism, and phospholipid biosynthesis display a closer relationship to bacterial genes (Rivera et al., 1998).

For over a decade (1990–2002), the Euryarchaeota and Crenarchaeota were the only recognized archaeal phyla (Fig. 2A) (C. Woese, 1990). The Crenarchaeota were considered exclusively thermophiles, while the Euryarchaeota consisted of a more heterogeneous group of methanogens, extreme halophiles, and sulfate reducers. By 2003, 16 complete archaeal genomes had been sequenced, comprising 12 euryarchaeota and four crenarchaeota, with several others nearly complete (Makarova & Koonin, 2003). These early sequencing efforts showed that Archaea have many unique genetic features that are distinct from Bacteria. One notable example is the ability of Archaea to thrive under extreme conditions, such as in the water near hydrothermal vents that are heated to over-boiling temperatures and saturated with hydrogen sulfide or in extreme salinity (Segerer et al., 1993; Stetter, 1999). While bacteria can also be found in these environments, archaea typically dominate these microbial communities (Merino et al., 2019; Oren, 2002; Stetter, 1999). This has largely given rise to a common misconception that all

archaea are extremophiles despite their presence in almost all known moderate environments.

Between 2002 and 2011, advancements in next-generation sequencing, enhanced phylogenetic techniques, and improved genomic assemblies led to the proposal of several new archaeal phyla (Fig. 2B) (Spang et al. 2017). These included the Korarchaeota, a group of thermophiles from terrestrial hot springs (Reigstad et al. 2010; Barns et al. 1996), the Nanoarchaeota, represented by a nano-sized hyperthermophilic archaeon from a submarine hot vent, and the ammonia-oxidizing Thaumarchaeota (DeLong 1992; Brochier-Armanet et al. 2008). Together with the candidate phylum Aigarchaeota (Nunoura et al., 2011), the Thaum-, Cren-, and Korarchaeota were proposed as a new archaeal superphylum known as the "TACK" (also called the "Proteoarchaeota") (Guy and Ettema 2011; Petitjean et al. 2015). While the "TACK" acronym is still used today, it now encompasses a greater diversity of phyla, including the Bathyarchaeota (Barns et al., 1996; Kubo et al., 2012), the Geoarchaeota (Kozubal et al., 2013), and the Verstraetearchaeota (Vanwonterghem et al., 2016). In general, the TACK superphylum contains lineages with diverse metabolisms, such as ammonia-oxidizers, autotrophs, and methanogens, and are important biological members of the global carbon and nitrogen cycles (Adam et al., 2017; Ingalls et al., 2006; Kozubal et al., 2013).

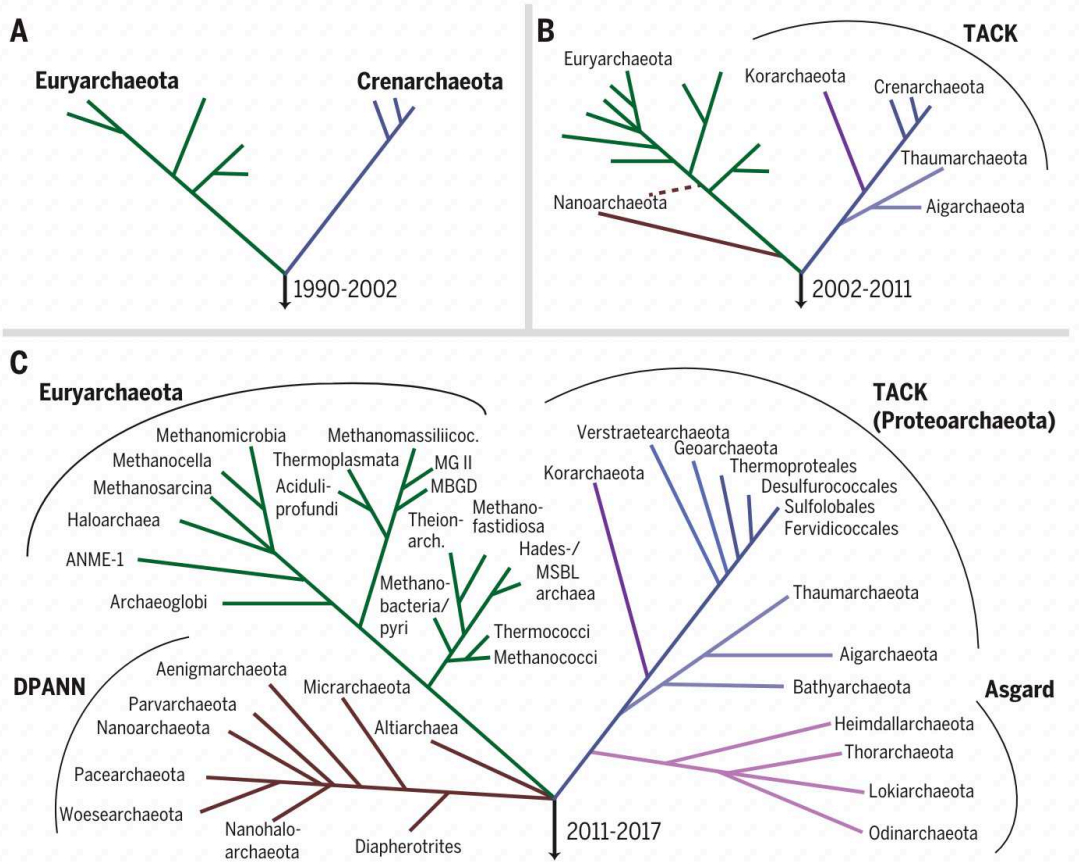


Figure 2 | The expansion of the archaeal tree over the last decades. This figure was adapted from Spang et al. 2017 and illustrates our progressive understanding of the archaeal tree of life over the last 30 years. (A) The initial classification by Carl Woese in 1990 identified the Euryarchaeota and the Crenarchaeota as the first two archaeal phyla. (B) Between 2002 and 2011, advancements in genomic sequencing and environmental 16S rRNA surveys resulted in the classification of the Euryarchaeota and TACK superphyla, along with the identification of a novel phylum, the Nanoarchaeota. (C) The archaeal tree now has four major superphyla—Euryarchaeota, TACK, Asgard, and DPANN. It is important to note that although these four superphyla are commonly used to describe the overall

organization of the archaeal tree, a superphylum is not an officially recognized taxonomic rank.

As of today, the archaeal tree now includes four major superphyla: the TACK, Euryarchaeota, and most recently, the DPANN (Rinke et al., 2013) and Asgard archaea (Zaremba-Niedzwiedzka et al., 2017). The Asgard archaea have garnered significant attention in the last several years owing to their close evolutionary relationship with eukaryotes (Eme et al., 2023; Spang et al., 2015; Zaremba-Niedzwiedzka et al., 2017). The first sequenced member of the Asgard archaea, *Lokiarchaeum*, showed an unexpectedly high number of eukaryotic signature proteins previously thought to be absent in Archaea (Spang et al., 2015). In agreement with these shared genomic characteristics, an updated universal tree showed that the Lokiarchaeota were the closest sister group to eukaryotes (Spang et al., 2015). As more Asgard lineages have been described, phylogenomic analyses also robustly place the TACK as a sister group to the Asgard (Eme et al., 2023; Spang et al., 2019; Zaremba-Niedzwiedzka et al., 2017).

1.3. The DPANN archaea

Along with Asgard, the DPANN is one of the more recently proposed archaeal superphyla. The DPANN acronym represents the first five phyla described for this group (Diapherotrites, Parvarchaeota, Aenigmarchaeota, Nanoarchaeota, and Nanohaloarchaeota) (Rinke et al., 2013). However, the DPANN now encompasses a much broader range of taxonomic diversity (11 phyla according to GTDB r214; (Rinke et al., 2021). DPANN archaea are typically described as nano-sized archaea that encode a limited set of metabolic proteins. Notably, many DPANN genomes lack proteins for amino acid, nucleotide, and phospholipid biosynthesis, which are

typically considered essential for a free-living lifestyle (Castelle et al., 2018). Based on these genomic features and in agreement with the few successful co-culture enrichments, it has been suggested that the defining feature of the DPANN archaea is a symbiotic lifestyle that requires a host for growth and survival (Dombrowski et al., 2019; Rinke et al., 2013). However, to date, this has only been confirmed for the few members that have been successfully co-cultured (Fig. 3).

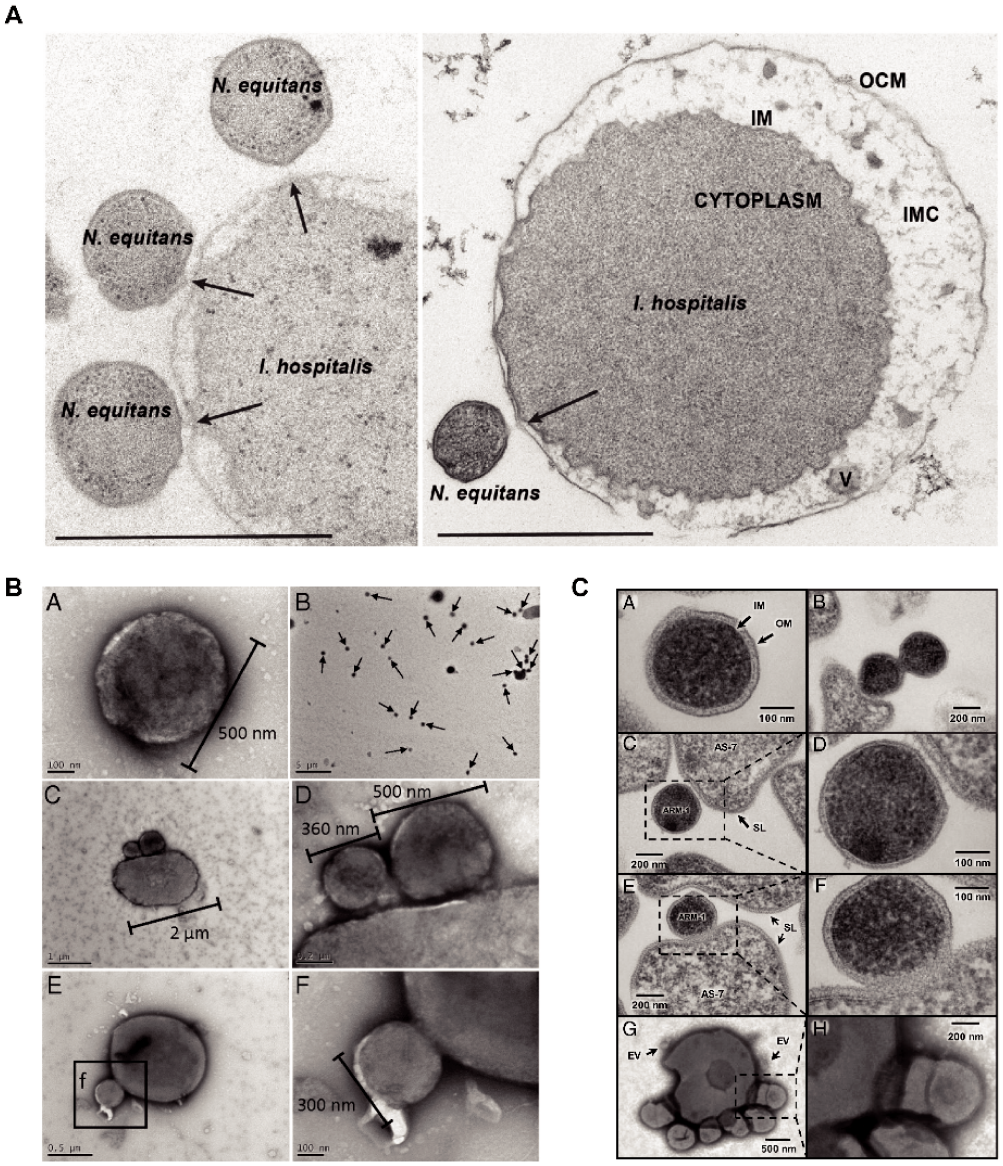


Figure 3 | Examples of DPANN lineages successfully enriched in co-cultures. (A) The first co-cultured enrichment of a DPANN archaeon, *Nanoarchaeum equitans,* with its hyperthermophilic host *Ignicoccus hospitalis* (Huber et al. 2002). (B) The extreme halophilic DPANN archaeon *Nanohaloarchaeum antarcticus* with its host *Halorubrum lacusprofundi* (J. N. Hamm et al., 2019). (C) The acidophilic DPANN

archaeon *Candidatus* Micrarchaeota ARM-1 with its host *Metallosphaera* sp. AS-7 (Sakai et al. 2022).

1.3.1. Nanoarchaeota

The first cultivated member of the DPANN, *Nanoarchaeum equitans*, was discovered in 2002 in a geothermal submarine hot vent (Huber et al., 2002). *N. equitans* was shown to be a hyperthermophilic nanosized archaeon ~400 nm in diameter, whose growth was dependent on an archaeal host from the genus *Ignicoccus* (Huber et al., 2002). Interestingly, when the researchers of this study tried to amplify the 16S rRNA from these two cells using archaeal universal primers, only the rRNA genes from *Ignicoccus* were amplified. The DPANN went undetected for many years in rRNA environmental surveys using 'universal' archaeal primers, initially designed to target highly conserved regions of the 16S rRNA based on available Crenarchaeota and Euryarchaeota sequences (Casanueva et al., 2008; Huber et al., 2003). However, after identifying *N. equitans*, subsequent 16S rRNA surveys with updated archaeal primers detected additional Nanoarchaeota environmental sequences (Hohn et al. 2002). Initial 16S rRNA phylogenetic trees of Nanoarchaeota indicated their status as a novel deep-branching lineage within Archaea, separate from the recognized phyla at the time: the Crenarchaeota, Euryarchaeota, and Korarchaeota (Huber et al., 2002, 2003; Waters et al., 2003). However, depending on the phylogenetic method used, the position of the Nanoarchaeota seemed to vary between a deep-branching position at the base of the archaeal tree or nested within the Euryarchaeota (Brochier et al. 2005; Branciamore et al. 2008). Given its proposed deep-branching position, early researchers questioned whether the small genome of Nanoarchaeota is an

inherent "primitive" feature of archaea or the result of significant genome reduction (Huber et al., 2003; Makarova & Koonin, 2005).

Studies have proposed that the anaerobic and hyperthermophilic lifestyle of Nanoarchaeota supports the idea that its small genome is a primitive rather than a derived feature (Huber et al., 2003; Thomson et al., 2004) given that a similar lifestyle has been proposed for the last archaeal common ancestor (Groussin & Gouy, 2011; Huber et al., 2003; Williams et al., 2017). However, with the discovery of additional DPANN lineages, this group no longer has a preferred ecological niche. DPANN lineages now include mesophiles, thermophiles, aerobes, and anaerobes (Dombrowski et al., 2019).

To date, nine DPANN lineages have now been successfully cultivated, including four Nanoarchaeota (Huber et al., 2002; Podar et al., 2013; St. John et al., 2019; Wurch et al., 2016), three Micrarchaeota (Golyshina et al., 2017; Krause et al., 2017; Sakai et al., 2022), and two Nanohaloarchaeota (J. N. Hamm et al., 2019; La Cono et al., 2020). Based on these cultivated representatives, some DPANN seem to rely on a single host (Hamm et al. 2019), while others can ‘host-switch’ (Sakai et al. 2022; Dombrowski et al. 2020). However, the cultivated DPANN lineages only represent a small fraction of the known DPANN diversity (Dombrowski et al., 2019). For example, different lineages within the Micrarchaeota have demonstrated a broad diversity both in terms of habitat distribution and metabolic coding potential (L.-X. Chen et al., 2018; Golyshina et al., 2017, 2019; Kadnikov et al., 2020).

1.3.2. Micrarchaeota

Previously named Archaeal Richmond Mine Acidophilic Nanoorganisms (ARMAN), the Micrarchaeota were first detected in the acid mine drainage systems of Iron Mountain (CA, USA; (B. J. Baker et al., 2006, 2010) but were later confirmed

to be widely distributed in many low pH environments, as well as in areas with neutral pH, soils, peats, freshwater systems, geothermal lakes, and hypersaline mats (L.-X. Chen et al., 2018; Golyshina et al., 2019; Kadnikov et al., 2020). Despite their small genomes (0.64–1.08 Mb), the Micrarchaeota play an essential role in carbon and nitrogen cycling by breaking down various saccharides and proteins (L.-X. Chen et al., 2018). They have also been shown to generate ATP through aerobic respiration and fermentation but do not possess biosynthetic pathways for amino acids and nucleotides (L.-X. Chen et al., 2018; Golyshina et al., 2019). However, the three cultured representatives of this group are all from acidic environments (Golyshina et al., 2017; Krause et al., 2017; Sakai et al., 2022). Interestingly, one Micrarchaeota lineage (strain Sv326), identified in a freshwater lake, has been proposed to be capable of a free-living lifestyle (Kadnikov et al., 2020). The authors of this study argued that "*Ca.* Micrarchaeota Sv326'' retained several metabolic pathways compared to the acidophilic, host-dependent Micrarchaeota lineages (Golyshina et al., 2019). These metabolic pathways include a complete glycolytic pathway and gluconeogenesis, the ability to generate ATP via phosphorylation, the capacity to utilize some sugars and amino acids as substrates, and pathways for de novo nucleotide biosynthesis (Kadnikov et al., 2020). However, the proposal that strain "*Ca.* Micrarchaeota Sv326' is capable of a free-living lifestyle based solely on genomic data and has yet to be proven experimentally.

1.3.3. Nanohaloarchaeota

The first extreme halophilic archaeon described outside the classic Haloarchaea was the Nanohaloarchaeota. In 2011, two different studies assembled Nanohaloarchaeota metagenome-assembled genomes (MAGs) from disparate hypersaline lakes located in Australia (Narasingarao et al., 2012a) and in Alicante,

Spain (Ghai et al., 2011). Based on 16S rRNA phylogenetic trees, the Nanohaloarchaeota was first described as a new archaeal class within the Euryarchaeota, sister to the Haloarchaea (Narasingarao et al., 2012a). This position suggested a shared ancestry between the Nanohaloarchaeota and Haloarchaea and therefore a single origin of adaptation to extreme halophily in the archaeal tree (Fig. 6). However, subsequent phylogenomic analyses that included a more extensive taxonomic sampling of recently proposed uncultivated archaea, argued that the Nanohaloarchaeota are, in fact, a member of a new archaeal superphylum, the DPANN archaea (see DPANN archaea section above for more details). This revised deeper-branching position of the Nanohaloarchaeota suggested at least two distinct, independent adaptations of archaea to hypersaline environments: one on the branch leading to the Haloarchaea and another on the branch leading to the Nanohaloarchaeota (Fig. 6).

Based on environmental samples, the Nanohaloarchaeota appear as small coccoid cells with an average diameter of ~0.6 μm (Andrade et al., 2015; Narasingarao et al., 2012a) and a genome size of ~1 Mbp (Dombrowski et al., 2019). It was suggested early on that certain Nanohaloarchaeota may be capable of a free-living lifestyle due to the absence of observed associations between nanohaloarchaeal cells and potential host cells in environmental samples (Narasingarao et al., 2012a). However, the first successful co-culture of Nanohaloarchaeota, sampled from hypersaline lakes in Antarctica, revealed that the growth of *Candidatus* Nanohaloarchaeum antarcticus was indeed dependent on the archaeal host *Halorubrum lacusprofundi* for survival (Fig. 2B) (J. N. Hamm et al., 2019). In less than a year, a second co-culture of Nanohaloarchaeota was reported (La Cono et al., 2020) and revealed a distinct Nanohaloarchaeota genus, *Ca.* Nanohalobium was also obligately associated with a haloarchaeal host,

Halomicrobium sp. LC1Hm. These two co-cultures demonstrated that geographically and phylogenetically diverse Nanohaloarchaeota lineages rely on a haloarchaeal host for growth and survival. However, each lineage appears to employ distinct systems for recognizing hosts and conferring host specificity (La Cono et al., 2020). This serves as a notable example of the genetic diversity within DPANN archaea, even within a single family.

1.3.4. Altiarchaeota

Overall, the difficulty with successfully culturing DPANN lineages has made it challenging to define unifying features for this group. While the nine cultured DPANN representatives appear to depend strictly on a host for growth and survival, genomic data suggests that other lineages within may be capable of a more diverse range of lifestyle preferences. An especially intriguing DPANN lineage, likely possessing the capability of a free-living lifestyle, is the Altiarchaeota (Esser et al., 2023; Probst et al., 2014; Schwank et al., 2019). However, the phylogenetic position of the Altiarchaeota in relation to other DPANN members remains a matter of debate (Adam et al., 2017; Bird et al., 2016; Dombrowski et al., 2019, 2020; Schwank et al., 2019).

The Altiarchaeota are strict anaerobes found in cold, sulfidic groundwater and form highly pure biofilms with the filamentous bacterium *Thiothrix* sp. (Probst & Moissl-Eichinger, 2015). Biofilms containing *Candidatus* Altiarchaeum hamiconexum (formerly known as SM1 Euryarchaeon) are primarily composed of Altiarchaeota (95%) with a minimal proportion of diverse sulfate-reducing bacteria (5%) (Henneberger et al., 2006; Probst & Moissl-Eichinger, 2015). The Altiarchaeota also have interesting cellular features where they synthesize long filamentous appendages known as ‘hami’ that establish strong connections with other cells

(Probst et al., 2014). Additionally, they are among the few archaea with a double-cell membrane (Klingl, 2014).

Interestingly, specific members of the Altiarchaeota have been shown to serve as hosts for another DPANN lineage, the Huberarchaeota (Esser et al., 2023; Schwank et al., 2019). Recent findings also indicated that a significant portion of the CRISPR-Cas spacer population in Altiarchaeota targets the genomes of Huberarchaeota (Esser et al., 2023). These findings suggest that CRISPR-Cas systems play an important role in mediating and controlling host-symbiont interactions, warranting further exploration in other host-symbiont systems.

As mentioned previously, however, the exact phylogenetic position of the Altiarchaeota remains unresolved (Adam et al., 2017; Castelle & Banfield, 2018; Dombrowski et al., 2019, 2020; Probst & Moissl-Eichinger, 2015; Spang et al., 2017). Phylogenetic analyses have placed the Altiarchaeota as either a deep-branching Euryarchaeota lineage (Adam et al., 2017), nested within the Euryarchaeota (Probst & Moissl-Eichinger, 2015), or affiliated with the DPANN archaea (Spang et al. 2017; Castelle et al. 2018; Castelle and Banfield 2018). When affiliated with the DPANN, the Altiarchaeota have been placed either sister to all other DPANN (Spang et al. 2017) or sister to the Micrarchaeota and Iainarchaeota (taxonomic names are according to GTDB r207; the Iainarchaetoa were previously named Diapherotrites) (Castelle & Banfield, 2018; Dombrowski et al., 2020; Moody et al., 2022). If the Altiarchaeota are indeed members of the DPANN archaea, this would greatly expand the metabolic and lifestyle potential of this superphylum.

1.3.5. The monophyly of the DPANN archaea

The monophyly of the DPANN and their relationship with other Archaea have been highly debated (Adam et al., 2017; Brochier et al., 2005; Dombrowski et al.,

2019, 2020; Petitjean, Deschamps, López-García, & Moreira, 2015; Rinke et al., 2013; Spang et al., 2017). There are currently four main hypotheses for placing the DPANN archaea (Fig. 4).

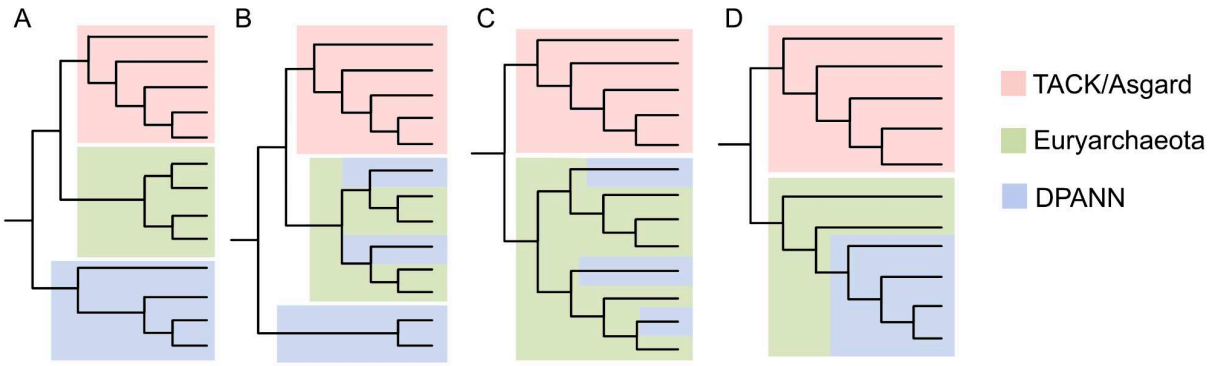


Figure 4 | Schematic representation of the four possible scenarios for the phylogenetic placement of the DPANN archaea. A) DPANN is a monophyletic group at the base of the archaeal tree. B) Some DPANN are monophyletic at the base of the archaeal tree, while other lineages are reduced Euryarchaeota. C) All DPANN lineages are reduced Euryarchaeota. D) DPANN are a monophyletic group that has evolved from a Euryarchaeota-like ancestor.

Rinke et al. initially proposed the DPANN superphylum in 2013. The authors of this paper suggested that the DPANN form a monophyletic group at the base of the archaeal tree (Fig. 4A). This position is consistent with most published trees today that include representatives from all known DPANN groups (Spang et al. 2013; Castelle et al. 2015; Spang et al. 2017; Saw et al. 2015; Williams et al. 2017). However, subsequent studies have attributed the monophyly of the DPANN to potential phylogenetic artifacts such as sequence compositional biases and/or long-branch attraction (LBA) (Aouad et al., 2022; Petitjean, Deschamps,

López-García, & Moreira, 2015). DPANN archaea are typically placed at the end of long branches in the archaeal tree, potentially reflecting a faster rate of genome evolution in symbiotic lineages compared to free-living archaea (Moran & Bennett, 2014; Muñoz-Gómez et al., 2022). In LBA, long branches can be artifactually grouped in a tree even if distantly related (see section Phylogenetic Methods and Limitations below) (Bergsten 2005; Susko and Roger 2021; Felsenstein 1978). This is further supported by the disparity observed between the phylogenetic placement of individual DPANN lineages and their placement in trees when all DPANN lineages are considered (Williams et al., 2017). For instance, Nanoarchaeota, Nanohaloarchaeota, Micrarchaeota, and Altiarchaeota are typically classified within Euryarchaeota when considered individually (Adam et al., 2017; Aouad et al., 2018; Brochier et al., 2005; Feng et al., 2021; Petitjean, Deschamps, López-García, & Moreira, 2015; Schwank et al., 2019). However, they group with other DPANN when a more extensive taxonomic sampling of DPANN is included (Castelle & Banfield, 2018; Dombrowski et al., 2020; Hug et al., 2016; Rinke et al., 2013; Spang et al., 2017; Williams et al., 2017).

Similar to LBA, lineages that share amino acid compositions can also artifactually group together in phylogenetic trees even if they are distantly related (B. A. Baker et al., 2023; Muñoz-Gómez et al., 2022). In bacteria, it has been shown that the genomes of symbiotic lineages tend to be biased towards A and T nucleotides (and their proteins towards F, I, M, N, K, and Y amino acids, as A+T-rich codons encode them) in contrast to free-living lineages (which can be biased towards G+C-rich genomes and G, A, R and P amino acids) (Clark et al., 1999; Gu et al., 1998; Muñoz-Gómez et al., 2019, 2022). Like symbiotic bacteria, many DPANN genomes also tend to be A+T-rich (Dombrowski et al., 2019). However, this does not apply uniformly to all DPANN archaea, particularly those with larger genomes, like

the Micrarchaeota or Nanohaloarchaeota. Shared amino acid compositions could also be why individual DPANN lineages are placed in various parts of the archaeal tree. For example, the Nanohaloarchaeota might be artifactually attracted to their haloarchaeal hosts due to shared amino acid preferences as a mode of adaptation to salinity.

1.4. Halophilic archaea

97% of all aquatic systems on Earth are considered saline environments (Cassardo & Jones, 2011). In these saline environments, halophilic organisms from all three domains of life (Bacteria, Eucarya, and Archaea) can be found (Oren, 2002). In general, halophiles require a minimum of 1 M salt for growth and can thrive across a broad range of salt concentrations (Oren, 2002). Depending on the optimal salt requirement for growth, halophiles are typically classified as either slight (0.34–0.85 M NaCl), moderate (0.85–3.4 M NaCl), or extreme (3.4–5.1 M NaCl) (Dutta & Bandopadhyay, 2022). As salt concentrations increase, the overall taxonomic diversity decreases in high-salt environments (Oren, 2002). In some hypersaline environments, it has been shown that up to 99% of the operational taxonomic units (OTUs) were assigned to diverse groups of halophilic archaea (Belilla et al., 2019, 2021).

For over 30 years, all known extreme halophilic archaea belonged to a single archaeal order, the *Halobacteriales* (also called halobacteria; henceforth: haloarchaea). Yet, most hypersaline environments (e.g., salterns, soda lakes, deep-sea brine pools, and fermented foods) are known to be nutrient-rich ecosystems with high cell densities, making it surprising that only a single archaeal order would thrive in these environments (Oren, 2002). As predicted,

advancements in metagenomics and the capacity to assemble metagenome-assembled genomes (MAGs) have increased the diversity of known groups of extreme halophilic archaea over the past decade. This includes the Nanohaloarchaeota, a group of nano-sized and symbiotic archaea (Ghai et al., 2011; Narasingarao et al., 2012b), the Methanonatronarchaeia, a group of extremely halophilic methanogens (Sorokin et al., 2017), and most recently the Haloplasmatales, a new order within the Thermoplasmatota (Zhou et al., 2022). However, the exact phylogenetic placement of these new lineages and their relationship to each other remains a matter of debate (Aouad et al., 2018, 2019; Feng et al., 2021; Martijn et al., 2020; Petitjean, Deschamps, López-García, Moreira, et al., 2015; Sorokin et al., 2019).

1.4.1. Acidic proteomes of salt-in strategists

The distinguishing characteristic between slight or moderate halophiles and extreme halophiles is their strategy for dealing with osmotic pressure. Slight or moderate halophiles employ what is known as a ‘salt-out’ strategy, whereas extreme halophiles employ a ‘salt-in’ strategy (Andrei, Banciu, and Oren 2012). As the name suggests, ‘salt-out’ strategists actively pump salt out of their cells while simultaneously synthesizing osmoprotects. Osmoprotectants or compatible solutes are small organic molecules like glycine or betaine that, at high concentrations, help a cell maintain osmotic equilibrium (Deole and Hoff 2020). However, this strategy can be energetically expensive for obligate halophiles living in high-salt environments (Gunde-Cimerman, Plemenitaš, and Oren 2018). Conversely, 'salt-in' strategists, unique to extreme halophiles, actively pump salt into their cells, establishing an intracellular osmotic pressure that matches their surrounding environment (Andrei et al., 2012). This means that extreme halophiles can maintain

a high molar salt concentration, usually in the form of potassium ions, in their cytoplasm at any given time (Siglioccolo et al., 2011). For non-halophilic organisms, elevated intracellular salt concentrations can trigger protein aggregation and misfolding (Andrei et al., 2012; Lanyi, 1974; Sun et al., 2016). To deal with high intracellular salt concentrations, extreme halophiles have had to undergo proteome-wide modifications (Deole et al., 2013). Specifically, haloarchaeal proteomes exhibit a massive enrichment in acidic amino acids (aspartic (D) and glutamic (E) acid) and a depletion in basic and large hydrophobic amino acids (such as isoleucine (I) and lysine (K)) (Fukuchi et al., 2003; Lanyi, 1974). The exact mechanistic benefits of an acidic proteome are not fully understood. Still, it is hypothesized that the negative ionic charge of acidic amino acids, through interactions with water molecules, increases the protein's water activity, either directly or cooperatively with hydrated cations (Gunde-Cimerman et al., 2018; Lanyi, 1974; Sun et al., 2016). Using genomic data, these proteome modifications can be detected by calculating the isoelectric point of each protein in a proteome (Ghai et al., 2011; Paul et al., 2008). A protein's isoelectric profile (pI value) represents the distribution of negatively charged amino acids. While non-halophilic archaea have a bi-modal distribution, extreme halophiles have a single spike of around three on a 0 to 14 pI scale (Ghai et al., 2011). While an acidic proteome seems to be a defining feature of salt-in strategists and extreme halophiles, the evolutionary history of these proteome modifications in archaea remains enigmatic until the halophilic species tree is resolved.

1.4.2. Methanogen-to-halophile transition

Since the first archaeal trees, Haloarchaea has been positioned as a sister group to methanogenic Euryarchaeota, specifically Class-II methanogens (C. Woese,

1990). This placement has always been intriguing, given the large-scale genome rearrangement thought to be required to transition from an anaerobic methanogen to a heterotrophic extreme halophile (Martijn et al., 2020). Previous studies suggested that this transition was accompanied by extensive horizontal gene transfer (HGT) events from bacteria (Becker et al., 2014; Groussin & Gouy, 2011; Nelson-Sathi et al., 2012, 2015). However, the extent and timing of these events are still debated. Early studies proposed a single acquisition of over 1,000 bacterial genes in the last common ancestor of Haloarchaea (Nelson-Sathi et al., 2012). These findings indicated that HGT, on a significant scale, played a transformative role, converting a strictly anaerobic, chemolithoautotrophic methanogen into the heterotrophic, oxygen-respiring common ancestor of Haloarchaea. However, upon reevaluating the results of this study, it became apparent that the methodology employed systematically inflated the count of genes acquired at the root of the Haloarchaea and incorrectly assumed bacterial origins for these genes (Groussin et al., 2016). Beyond the mentioned studies, additional research has identified authentic horizontal gene transfer (HGT) events from bacteria to Haloarchaea, particularly with a more extensive taxonomic sampling of Haloarchaea. However, these events occurred on a much smaller scale than previously suggested (Becker et al., 2014; Gadda & McAllister-Wilkins, 2003).

An important discovery for understanding the evolution of Haloarchaea was the Hikarchaeia (Martijn et al., 2020). The Hikarchaeia (originally marine group IV) were first detected by a 16S rRNA gene sequence survey of a deep-sea site located at a 3000 m depth at the Antarctic Polar Front (López-García et al., 2001). However, it took another 20 years before an in-depth phylogenomic analysis reconstructed the first genomes for this group and confirmed their phylogenetic placement as the closest-known relative to the Haloarchaea (Martijn et al., 2020). By incorporating

Hikarchaeia genomes, this study reevaluated the evolution of gene families along the branches connecting methanogenic Euryarchaeota and Haloarchaea. Unlike a single event responsible for this transition, this study proposed an updated scenario in which an aerobic halophilic lifestyle gradually evolved from a methanogenic ancestor through stepwise gene gain and loss events. However, shortly before the publication of this study, a new group of extreme halophiles, known as Methanonatronarchaeia, was identified (Sorokin et al., 2017). This group was significant in understanding the methanogen-to-halophile transition as they were the first described as extreme halophilic methanogens. It is important to note that methanogens had been previously identified in slight or moderate saline environments (Borton et al., 2018; Oren, 2002). However, this was the first instance of an extreme halophilic methanogen. Like other extreme halophiles, the Methanonatronarchaeia were also shown to be salt-in strategists. Initial phylogenetic analysis of the Methanonatronarchaeia placed them as the closest sister group to the Haloarchaea. However, the Hikarchaeia were not included in this analysis (Sorokin et al., 2017). The authors of this study argued that the Methanonatronarchaeia were an evolutionary intermediate between methanogens and extreme halophiles. Subsequent phylogenetic analyses, however, suggested that the Methanonatronarchaeia placed much deeper in the Euryarchaeota tree, at the base of the Methanotecta superclass (Fig. 6), and their placement as a sister group to the Haloarchaea is a result of tree reconstruction artifacts (Aouad et al., 2019; Martijn et al., 2020). If the Methanonatronarchaeia does indeed place at this deeper-branching position, this would suggest an independent adaptation to high salt environments outside the Haloarchaea. However, their exact phylogenetic position is still debated (Feng et al., 2021; Sorokin et al., 2019)

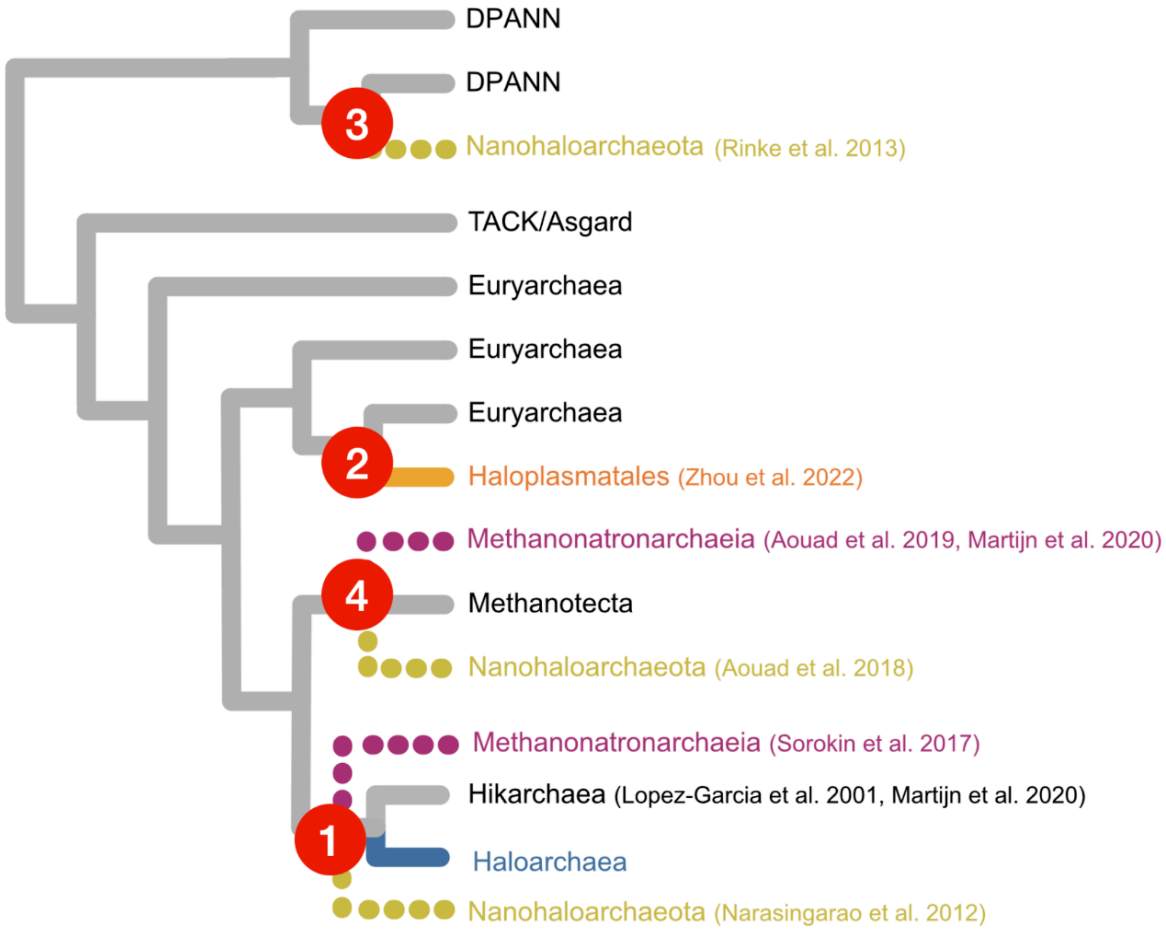


Figure 6 | Schematic representation of the proposed phylogenetic positions of various groups of halophilic archaea. The red circles on this tree indicate proposed instances of archaea adapting to hypersaline environments.

1.5. Archaeal taxonomy

Most proposed archaeal phyla have been identified solely through genomic sequencing and lack isolated cultures or recognized type strains according to the International Code of Nomenclature of Prokaryotes (ICNP) (Pallen, 2021). To manage the surge in genomic data in recent years, the ICNP now assigns '*Candidatus*' as a temporary status to newly described sequence-based taxa.

However, the problem is that the temporary status *Candidatus* has now been applied to over 1,000 taxa (as of 2021) and has been widely adopted by journals and databases (Pallen, 2021). This problem has rapidly been magnified by advancements in metagenomics and metagenome-assembled genomes (MAGs).

Metagenomics enables direct DNA sequencing from an environment, usually achieved through shotgun sequencing (Quince et al., 2017). When analyzing a single genome, shotgun sequencing is often compared to shredding a book into small fragments and then reconstructing it back to its original form by individually matching overlapping words and phrases. If we extend this metaphor to include metagenomic shotgun sequencing, we now have hundreds of books that we must reassemble instead of having only a single book. Scientifically, the reassembled books within a library would be MAGs. The objective is to ensure that the original and reassembled books are identical. However, it is feasible that similar books could contain identical or nearly identical sentences, leading to the misassignment of these sentences. This scenario can also be observed during the assembly of MAGs. When assembling MAGs, the aim is to identify patterns, primarily based on k-mer profiles, to decide whether two contigs belong to the same genome. However, this can quickly become problematic when an environment has high strain heterogeneity (Parks et al., 2015). When working with MAGs, it is essential to consider whether a particular MAG represents a genuine genome or a mosaic of closely related strains/genomes (Setubal, 2021). This issue becomes particularly challenging in regions of high genomic variability, where specific segments are present in some strains but absent in others (data from my Master's thesis, conducted in Dr. Jose de la Torre's Lab). In microbial genomes, these variable regions can contribute important adaptive phenotypes, such as antibiotic resistance or stress response genes (Bellanger et al., 2014; Dobrindt et al., 2004;

López-Pérez et al., 2014). Since these variable regions can be missed in MAG assemblies, it is important to understand what questions can be answered when comparing MAGs.

To address challenges associated with working with MAGs, an early recommendation was to incorporate quality assessments, usually focused on evaluating MAG completeness and contamination (Parks et al., 2015). CheckM (Parks et al. 2015) is the most commonly used software for assessing the completeness and contamination of prokaryotic genomes. CheckM relies on a set of highly conserved marker genes in Archaea or Bacteria and assesses their placement and co-occurrence within a genome (Parks et al., 2015). However, a single standardized set of markers for all archaea or bacteria is insufficient for specific taxonomic groups (Dombrowski et al., 2019). For example, the DPANN archaea are recognized for encoding a limited set of metabolic proteins attributed to their proposed symbiotic lifestyle, which results in a seemingly low completion rate among these lineages, even for genomes that are known to be complete (Dombrowski et al., 2019). However, recent advancements in assessment tools have tried to tackle these challenges (Chklovski et al., 2023).

1.5.1. Genome taxonomy database

The taxonomy of major archaeal groups is inconsistent across studies due to a lack of clear genome-based guidelines for classifying microorganisms. While some genome-based criteria, such as average nucleotide identity (ANI) or average amino acid identity (AAI), exist (Jain et al., 2018), these methods are inadequate for classifications above the genus level. To tackle the observed taxonomic inconsistencies across studies, the Genome Taxonomy Database (GTDB) was introduced in 2018 as a new genome-based microbial taxonomy ranking scheme

(Parks et al., 2018; Rinke et al., 2021). In brief, GTDB first retrieves a set of highly conserved single-copy marker genes from quality-filtered genomes (53 genes for Archaea and 120 for Bacteria). These sequences are then aligned, assessed for taxonomic congruence, concatenated into a supermatrix, and trimmed to approximately 5,000 sites due to computational restrictions. This supermatrix is then fed into an ML framework using IQTREE (Minh et al., 2020) with the LG+C10+F+G model of sequence evolution (see section Phylogenomics for more details). This phylogenetic tree is then used to normalize taxonomic ranks based on relative evolutionary divergence (RED) values between the last common ancestor (set to RED = 0) and all extant taxa (RED = 1). To normalize taxonomic ranks, RED intervals were then defined as the median RED value for taxa at each rank, and internal nodes were linearly interpolated between these values according to lineage-specific rates of evolution (Parks et al., 2018; Rinke et al., 2021). This ranking scheme enables a standardized taxonomic classification of microorganisms using genomic data. However, a notable drawback is that GTDB uses historical taxonomic names to define new taxonomic ranks, posing significant challenges when conducting a historical overview of archaeal taxonomy.

1.5.2. The Archaeal root

One of the major unanswered questions in archaeal phylogenetics is the placement of the archaeal root. To date, three main hypotheses have been proposed for the placement of the archaeal root (Fig. 5).

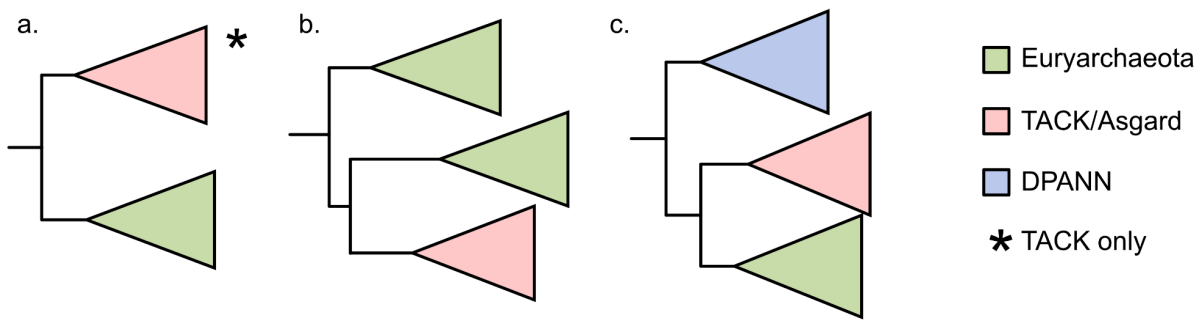


Figure 5 | Schematic representation of the three possible root positions proposed for the archaeal tree of life. Root positions (a,b) were determined using bacterial sequences as an outgroup, whereas root position (c) was identified through a novel approach employing a gene tree-species tree reconciliation model.

The first 16S rRNA-based phylogenies of Archaea suggested a root between the Euryarchaeota and Crenarchaeota (Fig. 5a; (C. Woese, 1990)). Initially, these two groups were considered to have equal taxonomic ranks. However, the Euryarchaeota are now classified as a superphylum, whereas the Crenarchaeota retained their status as a phylum. The discovery of additional archaeal groups, such as the Thaum- (Brochier-Armanet et al. 2008), Aig- (Nunoura et al., 2011), and Korarchaeota (Elkins et al., 2008) (now the TACK superphylum (Guy & Ettema, 2011)), challenged this initial view of a bipartite archaeal tree. Following the incorporation of TACK lineages, two studies sought to reevaluate the archaeal root position by using bacterial sequences as an outgroup (Petitjean et al. 2015; Raymann et al. 2015). However, each study arrived at a distinct rooting position. Mirroring the bipartite archaeal tree initially proposed by Carl Woese (C. Woese, 1990), Petitjean et al. placed the root between the Euryarchaeota and the rest of Archaea, while Raymann et al. placed the root within the Euryarchaeota (Fig. a and b). The latter position suggested that the Euryarchaeota are a paraphyletic group,

and the TACK evolved from an Euryarchaeota ancestor (Fig. 5b; (Raymann et al. 2015)). The main difference between these two analyses stemmed from the variation in taxon sampling and the phylogenetic markers used to reconstruct the phylogenetic trees. The observed impact of these choices on the archaeal root position suggests that the placement of the root may be influenced by a potential phylogenetic artifact rather than a genuine historical signal. Previous studies have suggested that using outgroup sequences to establish the root of a tree can be susceptible to phylogenetic artifacts, such as long-branch attraction (LBA) (Graham et al., 2002; Williams et al., 2017). This is particularly relevant for rooting the archaeal tree with outgroup sequences from bacteria due to the extremely long branch that separates these two domains (Petitjean, Deschamps, López-García, & Moreira, 2015). In LBA, long branches can be artifactually attracted to one another even if they are distantly related (Bergsten, 2005; Felsenstein, 1978; Susko & Roger, 2021).

For this reason, both of these studies chose not to include the DPANN archaea in their main analyses despite many DPANN lineages known at that time (Rinke et al., 2013). The DPANN archaea are especially prone to LBA due to a faster rate of genome evolution in symbiotic lineages compared to free-living archaea (Dombrowski et al., 2019). However, Petitjean et al. did run a secondary phylogenetic tree which included the DPANN. In this analysis, the authors recovered the DPANN as a polyphyletic group, branching at various positions within the Euryarchaeota (Fig. 4C). The authors concluded that based on this analysis, the DPANN should be considered as a bona fide Euryarchaeota species, as opposed to a monophyletic supergroup, as previously suggested (Rinke et al., 2013). Finally, neither of these studies incorporated the Asgardarchaeota due to the unavailability of genomes.

Two years later, the first study included taxon sampling from all four major archaeal superphyla (Williams et al., 2017). This most recent rooting analysis was particularly interesting because it employed a method to root the archaeal tree independent of an outgroup. The authors of this study first combined protein concatenation and a supertree approach using 3,242 single-gene trees to resolve an unrooted archaeal tree. They then inferred a root for that tree using a probabilistic gene tree-species tree reconciliation model (Szöllősi et al., 2013a). In brief, gene tree-species tree reconciliation models describe the evolutionary history of a gene family (also referred to as an orthologous group) along a species tree by mapping gene duplication, transfer, loss, and origination events along the species tree (Boussau and Scornavacca, 2020). The most commonly used software for this analysis is amalgamated likelihood estimation (ALE) (Szöllősi et al., 2013a). The input for ALE requires a rooted species tree to perform the reconciliation. Expanding on this requirement, Williams et al. conducted a reconciliation analysis on several archaeal species trees, each with different root positions. Varied roots on the species tree gave rise to diverse gene family evolution scenarios, which resulted in different likelihoods for gene families within the gene tree-species tree reconciliation model utilized in ALE (Williams et al., 2017). Various rooting hypotheses were then statistically distinguished based on these likelihoods. Applying this approach, Williams et al. determined that rooting the species tree between the DPANN and all other archaea provided the most favorable likelihood.

In summary, the three analyses that focused on resolving the position of the archaeal root have resulted in three different hypotheses (Fig. 5). Although the analysis conducted by Williams et al. stands as the only study incorporating all four major archaeal superphyla, placing the root between the DPANN and all other

archaea carries significant implications for understanding the last archaeal common ancestor (LACA) (see Discussion).

1.6. Phylogenomics

Deciphering the ancient evolutionary relationships within the archaeal tree of life is challenging due to historical artifacts that can potentially introduce noise into the phylogenetic analysis. Considering these artifacts at each analysis stage is essential for producing robust and reliable phylogenetic trees.

1.6.1. Taxon selection

The first step in any phylogenetic analysis is to determine the depth of taxonomic sampling. When analyzing the archaeal tree, the ideal scenario would be to include all known archaeal species; however, this can become computationally demanding and induce biases due to uneven taxonomic sampling (Martinez-Gutierrez & Aylward, 2021). Large taxon sampling alone is not necessarily computationally demanding, but the combination of extensive taxon sampling with many alignment characters can be (Zhu et al., 2019). In a recent study attempting to analyze the phylogenetic relationships among 10,575 taxa of bacteria and archaea, the researchers faced computational constraints. To address this, they compromised by restricting the number of alignment sites to a maximum of 100 sites per gene for the analysis (Zhu et al., 2019). Additionally, the selection of high-quality genomes should be prioritized over the number of genomes to avoid introducing contamination. After completing taxon selection, the next challenge is choosing a set of phylogenetic markers that accurately represents the true phylogenetic relationships among the selected taxa.

1.6.2. Marker selection

Choosing a set of reliable phylogenetic markers aims to pinpoint proteins (or genes) that are genuinely orthologous, display minimal compositional biases, and reflect the average rate of evolution within a given proteome. To provide a general definition, orthologous genes are related via speciation (i.e., vertical descent), whereas paralogs are genes related via duplication (Fitch, 1970). To identify a set of orthologous genes for phylogenetic analysis, studies often use the Basic Local Alignment Search Tool (BLAST) (Altschul et al., 1990). However, BLAST was designed to identify regions of sequence similarity, but these sequences or genes may not necessarily represent the closest phylogenetic neighbors (Dick et al., 2017; Eisen, 2000; Koski & Golding, 2001). There are instances where the best BLAST hit may correspond to a paralog, and the true ortholog might be the second, third, or fourth BLAST hit. This can occur due to variations in evolutionary rates, where two distantly related genes may be identified as best matches if their evolution rates are both slow or two closely related genes may not match well if they have evolved rapidly (Eisen, 2000). A rapidly evolving paralog could match a genuine ortholog in a similarly fast-evolving lineage, even if they are phylogenetically distantly related. There is also the possibility that the best BLAST hit could result from assembly contamination. Without retrieving multiple BLAST hits for each taxon in your dataset, these scenarios will go undetected. It is necessary to reconstruct the phylogeny for each marker family to identify the true ortholog, incorporating the top three or four BLAST hits for each taxon. The goal is then to iteratively remove sequences that are paralogs, contamination, or cases of horizontal gene transfers (HGTs). These sequences can then can be detected through inconsistencies in the phylogenetic tree, divergent patterns in the sequence alignment, or by comparing

the problematic sequences to a larger set of prokaryotic sequences such as ReqSeq's non-redundant database (O'Leary et al., 2016). It is especially important to try to remove all cases of HGTs. Prokaryotic genomes, in particular, have been shown to exhibit high rates of inter- and intra-domain HGT (Doolittle, 1999; Forterre et al., 2002; Hilario & Gogarten, 1993, 1993; Koonin et al., 2001; Makarova et al., 1999). However, detecting HGTs can be challenging in single-gene trees (SGTs) due to a lack of phylogenetic signals in individual proteins (Gogarten & Townsend, 2005). Recent HGTs are generally easier to detect than ancient ones, and transfers across diverse taxonomic groups are more straightforward than among closely related strains. Detecting HGTs can become increasingly more difficult when they have occurred between closely related lineages and/or are ancient transfers (Gogarten & Townsend, 2005; Koonin et al., 2001; Olendzenski et al., 2002). However, HGTs will only be detected if you collect multiple BLAST hits per taxa in your initial marker set curation.

After selecting and curating the phylogenetic markers, the next step involves aligning and trimming each marker before concatenating them into the final supermatrix. However, previous studies have shown that different alignment and trimming programs and settings can produce conflicting phylogenetic trees (Nguyen et al., 2015; Ogden & Rosenberg, 2006; Smythe et al., 2006; Talavera & Castresana, 2007). This highlights a reproducibility challenge in phylogenetics and emphasizes the need to publicly share untrimmed and trimmed alignments (Salomaki et al., 2020).

1.6.3. Phylogenetic methods and models

The final data preparation step in a phylogenomic analysis is concatenating the individually trimmed alignments into a supermatrix. Early studies underscored

the enhanced robustness of supermatrices against phylogenetic artifacts in comparison to single-gene phylogenies (Baldauf et al., 2000; J. R. Brown et al., 2001). This advantage is largely attributed to the increased number of phylogenetically informative sites and the ability to better withstand variations in evolutionary rates among different genes. Once the supermatrix is created, the final step involves running a phylogenetic tree. In modern phylogenetics, two main statistical inference frameworks are used: maximum likelihood and Bayesian. While both frameworks handle nucleotide and protein sequences, the emphasis in this section will be exclusively on protein data.

1.6.3.1. Maximum likelihood

Maximum likelihood (ML) in phylogenetics involves finding the evolutionary tree that produces the highest probability of the observed sequence data given the model of evolution (Felsenstein, 1981). The ML of a tree is found by making small changes to that tree and repeatedly evaluating its likelihood (Felsenstein, 1981). In both ML and Bayesian frameworks, a model of sequence evolution requires at least a substitution model. A substitution model for protein data, also called an amino-acid exchange rate matrix for protein sequences, is a 20x20 matrix, which estimates the instantaneous substitution rate from one amino acid to another. The rate of going to two biochemically different amino acids, for example, going from an arginine (positively charged) to an aspartate (negatively charged), is under negative selection and has a low rate. In contrast, replacements between isoleucine and alanine (both hydrophobic) are frequent and have a high rate (Quang et al., 2008). A reliable rate matrix is needed to accurately estimate the true evolutionary distances (branch lengths) and relationships among species (tree topology) (Whelan & Goldman, 2001). A way to improve amino acid replacement modeling is to use

different matrices depending on the data (Le & Gascuel, 2008). For example, different matrices exist for nuclear, mitochondrial, or viral data. However, if your data does not fall under one of these categories, there are general matrices, such as the WAG (Whelan & Goldman, 2001) or LG (Le & Gascuel, 2008) matrices. When using a substitution model by itself, it assumes that all sites are evolving at the same rate. However, it is known that sites of a protein do not all evolve at the same rate (referred to as site-heterogeneity) (Yang, 1994). For example, some sites evolve slowly due to strong functional or structural constraints, whereas others with low evolutionary pressure evolve rapidly. A discrete gamma model and a rate matrix can account for rate heterogeneity across sites. A discrete gamma model uses several categories of rates to approximate the gamma distribution (Yang, 1994). Within a gamma distribution, each category holds an equal probability, and each site's likelihood is assessed compared to the mean value of each rate category (Susko et al., 2003). The advantage of this method is that you don’t have to decide upfront which category each site falls into. Typically four categories are sufficient, but more categories can also be used (Yang, 1994).

In addition to rate heterogeneity, sequence sites also experience heterogeneity in substitution patterns depending on a protein's function, the secondary and tertiary structure, and solvent exposure (Quang et al., 2008). For example, buried sites within a protein preferentially accept hydrophobic amino acids, whereas residues at an active site are more likely to be charged amino acids (Quang et al., 2008). Standard substitution models accommodate only one amino acid rate matrix, while protein mixture models were introduced to address the different rates (substitutional heterogeneity) of amino acid substitutions for specific sites. In an ML framework, common protein mixture models are the C10, 20, 30, 40, 50, and 60-profile mixture models (Quang et al., 2008). These mixture models

(C10..C60) use 10 to 60 different amino acid substitution rate matrices instead of a standard single matrix. Protein mixture models can help address mutational saturation by allowing different sites in a protein sequence to evolve at different rates. This flexibility allows variations in substitution rates across different protein regions, which can be especially important when certain sites are evolving rapidly while others are more conserved. Protein mixture models can provide a more nuanced representation of the evolutionary dynamics of protein sequences.

1.6.3.2. Bayesian

Another commonly used approach in phylogenetics is Bayesian inference (BI). In a phylogenetic setting, BI is similar to ML in that it uses a likelihood function and a substitution rate matrix. The feature of BI that sets it apart from ML is the ability to include prior information. In theory this is done by giving a prior probability distribution of trees. However, in practice, implementing realistic prior probabilities is challenging so most analyses use a simple prior which assumes that all trees have equal probabilities (Archibald et al., 2003). BI also uses Markov Chain Monte Carlo (MCMC) simulations in combination with the chosen model and data to produce a posterior probability distribution of trees. In a BI phylogenetic software like Phylobayes (Lartillot & Philippe, 2004), the distribution of trees is summarized as a majority rule consensus tree, which is then used to determine the support values (Archibald et al., 2003; Huelsenbeck et al., 2002). The branch support on a Bayesian phylogeny is the posterior probability for that clade, meaning the probability that the clade is “true” given the priors, model, and data (Archibald et al., 2003). It is also common practice in BI to run multiple MCMC chains in parallel and then test whether or not these chains converge to the same tree.

One of the advantages of using BI is the availability of the CAT model (Lartillot & Philippe, 2004). The CAT model is a protein mixture model similar to the C10..C60 mixture model in a ML framework. However, as opposed to the C10..C60 model where the number of profiles/classes are determined beforehand, the CAT model has the substitution rates as parameters, which means the number of mixture model profiles is optimized during the analysis (Lartillot & Philippe, 2004). Unfortunately, the CAT model can pose significant computational demands for large datasets and usually requires the dataset to be subsetted in terms of the number of taxa. Despite the computational challenges associated with the CAT model, it has been shown to be less susceptible to phylogenetic artifacts such as LBA by more accurately modeling amino acid mutational saturations (Lartillot et al., 2007).

1.6.4. Phylogenetic artifacts

It can be difficult to place some archaeal lineages in the archaeal tree for several reasons. One of the major difficulties arises from LBA (Bergsten, 2005; Felsenstein, 1978; Susko & Roger, 2021). In LBA, long branches in a phylogenetic tree can artifactually group even if distantly related. However, many factors can lead to the emergence of long branches in a phylogenetic tree, such as compositional heterogeneity, substitution saturation, or insufficient taxonomic sampling. In the archaeal tree, there are still several major branches that are undersampled, such as the Hadarchaeota (B. J. Baker et al., 2016; Chuvochina et al., 2019), the Hydrothermarchaeota (Chuvochina et al., 2019; Vetriani et al., 1999), and various DPANN phyla, such as the Huberarchaeota (Probst et al., 2018), and several other unnamed DPANN phyla (SpSt-1190, EX4484-52, etc.). These taxa typically

appear at the end of long branches in a phylogenetic tree because there are no other closer relatives to break the long branch.

Long branches can also result from compositional heterogeneity, the unequal distribution of amino acids within a gene or between genes and/or organisms (Collins et al., 1994; Foster & Hickey, 1999). Simple phylogenetic models face challenges distinguishing between amino acids changes that are evidence of shared ancestry (synapomorphies) and instances where multiple changes are a result of convergent evolution (Collins et al., 1994; Foster & Hickey, 1999). One way to deal with this is to remove the most saturated sites (Philippe et al., 2000). However, this can sometimes result in removing more than half of the total number of amino acid sites (Aouad et al., 2019; Martijn et al., 2020; Sorokin et al., 2019).

Current phylogenetic mixture models, as mentioned above, can also account for compositional heterogeneity within genes. However, these models do not account for heterogeneity between organisms (or branches in a tree), a phenomenon referred to as branch heterogeneity (Muñoz-Gómez et al., 2022). Branch-heterogeneity can arise from either lineage-specific evolutionary rates, such as faster-evolving symbiotic lineages, like the DPANN archaea versus typically slower-evolving free-living lineages, a phenomenon known as heterotachy (Lopez et al., 2002), or from selective environmental pressures (B. A. Baker et al., 2023; Muñoz-Gómez et al., 2022).

For example, extreme halophiles have undergone proteome-wide modifications to address high intracellular salt concentrations and prevent protein aggregation (B. A. Baker et al., 2023). These modifications typically include a highly acidic proteome in the form of an over-representation of acidic amino acids, specifically aspartic (D) and glutamic (E) acid, and an under-representation of basic and large hydrophobic amino acids, such as isoleucine (I) and lysine (K),

respectively. In short proteins with few amino acids, such as ribosomal proteins, which are about 2–4 times smaller than average proteins (Reuveni et al., 2017), a few highly biased sites can artifactually group these sequences even if distantly related (B. A. Baker et al., 2023; Eme et al., 2023; Petitjean, Deschamps, López-García, & Moreira, 2015). This has also been proposed as a potential source of phylogenetic artifacts driving the monophyly of the DPANN archaea (Dombrowski et al. 2019). In bacteria, it has been shown that the genomes of symbiotic, especially parasitic lineages tend to be biased towards A and T nucleotides (and their proteins towards amino acids encoded by A+T-rich codons: F, I, M, N, K, and Y) in contrast to free-living groups that have not evolved reductively (which can be biased towards G+C-rich genomes and G, A, R, and P amino acids) (Muñoz-Gómez et al. 2022; Clark, Moran, and Baumann 1999; Muñoz-Gómez et al. 2019; Gu, Hewett-Emmett, and Li 1998). Similar to symbiotic bacteria, DPANN archaea tend to have a higher frequency of F, I, M, N, K, and Y amino acids compared to free-living archaea, which tend to have a higher frequency of G, A, R, and P amino acids.

2. Objectives

2. Objectives

1. Resolve the phylogenetic position of extremely halophilic archaea. Are the Nanohaloarchaeota members of the DPANN archaea?

The genomic features and symbiotic lifestyle of the Nanohaloarchaeota suggest that this lineage belongs to the DPANN archaea. However, since their discovery, their phylogenetic position in the archaeal tree has been unstable for over a decade. Initially, the Nanohaloarchaeota was described as a distinct, deep-branching lineage within the Euryarchaeota, separate from the only other known group of halophilic archaea, the Halobacteriota. However, with the discovery of additional nano-sized archaeal lineages, the Nanohaloarchaeota shifted to a position within the DPANN superphylum. However, studies as recent as 2022 have still questioned the phylogenetic placement of the Nanohaloarchaeota and have argued that they are indeed Euryarchaeota. Considering that the Nanohaloarchaeota are the most frequently discovered group outside of the DPANN in phylogenetic trees, our initial focus was determining their precise phylogenetic position and, by extension, the phylogenetic position of all halophilic archaea. To achieve this, we described two new family-level lineages of halophilic archaea and developed new robust phylogenetic methods to deal with amino acid biases prevalent in all extremely halophilic archaea.

2. Resolve the phylogenetic position of the DPANN archaea. Is the DPANN a monophyletic group? Where does the DPANN place in the archaeal tree of life?

The DPANN archaea are one of the four proposed archaeal superphyla, along with the TACK, Asgard, and Euryarchaeota. The DPANN are typically

characterized by their small cell sizes and limited set of metabolic proteins (e.g., amino acid and lipid biosynthesis). Nonetheless, the monophyletic status of this group is still an unresolved question. Initial phylogenetic analyses of individual DPANN lineages, for example, *Nanoarchaeum equitans* and *Nanosalinarun* sp. J07AB56 suggested these organisms were deep-branching Euryarchaeota. However, as additional DPANN lineages were uncovered, subsequent phylogenetic analyses indicated the formation of a monophyletic group. Advocates of the monophyletic status of the DPANN argue that earlier research lacked sufficient taxonomic sampling.

Conversely, critics argued that adding more taxa could potentially lead to phylogenetic tree reconstruction artifacts, such as long-branch attraction, attributed to the rapid evolutionary rate of these lineages. To address this question, the objective was to investigate whether the unstable position of the DPANN archaea results from taxon sampling in specific lineages or is influenced by phylogenetic artifacts such as long-branch attraction or compositional biases in amino acids within phylogenetic datasets. Subsequently, we will apply comparable phylogenetic methodologies, as outlined in the halophiles paper, to investigate how amino acid biases influence the phylogenetic positioning of DPANN.

3. Materials and Methods

3. Materials and Methods

We used several phylogenomic approaches to infer the phylogenetic position of several branches in the archaeal tree. Our approaches included identifying and eliminating biased amino acid alignment sites unique to our taxonomic groups of interest. To mitigate long-branch attraction, we employed sub-sampling in our datasets and utilized sophisticated models of sequence evolution. Furthermore, in collaboration with a team at Dalhousie University, we implemented a site-and-branch-heterogeneity model tailored specifically to our archaeal groups of interest. Even if our different studies had some specificities, we followed a similar overall methodological approach summarized in this section.

3.1. Generation of backbone datasets for phylogenetic analyses

All archaeal genomes were downloaded from the Genome Taxonomy Database (GTDB) (Rinke et al., 2021) to generate a representative archaeal reference set. We first subsetted the number of genomes using the software Treemmer v0.3 (Menardo et al., 2018). Treemmer selects the leaves in a phylogenetic tree representing the greatest diversity based on a predefined number of taxa. We further subsetted this dataset by selecting at least one representative from each archaeal class according to GTDB. When multiple representatives per class were available, we selected the genomes with the highest completion (>60%) and lowest contamination scores (<5%) based on CheckM (Parks et al., 2015). Once our final taxa were selected, all genomes were run through

Prodigal (Hyatt et al., 2010) to ensure that all open reading frames were uniformly predicted.

3.2. Selection of marker proteins for phylogenetic analyses

The marker proteins used in our phylogenetic analyses were based on a previously curated set of 200 archaeal single-copy markers shown to be highly conserved across the archaeal domain (Petitjean et al. 2015). Throughout the chapters of this thesis, we refer to this dataset as the NM (new markers) dataset. Sequences similar to the NM proteins were identified in the set of archaeal taxa using BLAST (Altschul et al., 1990) with relatively relaxed criteria (>20% sequence identity over 30% query length) to retrieve even divergent homologs, which is especially relevant for fast-evolving lineages like the DPANN archaea. Up to five BLAST hit sequences for each taxon were included for preliminary phylogenetic reconstruction using FastTree2 (Price et al., 2010). These preliminary trees were manually examined to identify the correct orthologue for each taxon and to detect cases of contamination, HGT, or paralogy. These spurious sequences were removed, and the remaining ones were used to reconstruct a new tree. Multiple rounds of manual curation were done this way until all problematic sequences were removed. We also removed some of the initial 200 markers poorly represented in our taxonomic sampling. This resulted in 136 NM markers for the Halophiles project and 126 NM markers for the DPANN project. Once curated, sequences of each of the NM markers were aligned with MAFFT L-INS-i v7.450 (Katoh & Standley, 2013) and trimmed with BMGE v1.12 (-m BLOSUM30 -b 3 -g 0.2 -h 0.5) (Criscuolo & Gribaldo, 2010). We performed a final round of verification of the single gene trees reconstructed using the more sophisticated LG+C60+F+Γ4

model in IQ-TREE (Minh et al., 2020) before concatenating the individually trimmed alignments into the NM supermatrix.

Additionally, we compared the NM dataset and a set of ribosomal proteins (RP), given their extensive use in phylogenetic analyses, serving as a benchmark for evaluating the NM dataset. The RP dataset was manually curated using the methods outlined for the NM dataset, as described above. This resulted in a set of 48 ribosomal proteins used as a second dataset alongside the NM dataset.

3.3. Phylogenetic analyses

All phylogenetic/phylogenomic trees (NM and RP datasets) were reconstructed using maximum likelihood (ML) (Felsenstein, 1981) and Bayesian inference (BI) (Huelsenbeck & Ronquist, 2001). All of our ML analyses were run with IQ-TREE (Minh et al., 2020) under the LG+C60+G model of sequence evolution and PhyloBayes-MPI (Lartillot et al., 2013) for BI analyses under the CAT+GTR model. Due to computational constraints, we could not run the full NM and RP datasets for the BI analyses. Consequently, all BI analyses were conducted on subsets of the NM and RP datasets (see each chapter for more details). All trees were visualized using FigTree (Rambaut et al., 2016) and iTOL (Letunic & Bork, 2007).

3.4. Detecting compositional biases

To obtain an overview of the amino acid composition of our datasets, we computed the frequency of each amino acid for both the individual markers and the proteomes of each taxon. This enabled us to assess the amino acid composition of both the NM and RP datasets and to make comparisons with the overall proteome. While the amino acid composition of the NM dataset closely resembled

that of the proteomes, a distinct contrast was evident in the amino acid profile of the RP dataset compared to the proteomes. We determined what amino acids drove these differences based on principal component analysis (PCA) vectors. We then verified these using a Z-score binning method developed by our collaborators at Dalhousie University (see chapters for more details). Depending on the archaeal group of interest, namely the Halophiles or the DPANN, we identified amino acids that were enriched and depleted in those taxa. This classification resulted in two distinct 'bins' of taxa, which we call the up and down bins. The ‘up-bin’ consists of taxa enriched in the amino acids of interest, while the ‘down-bin’ consists of taxa depleted in those amino acids.

Z-score binning formula:

$$Z = \frac{p1 - p2}{\sqrt{p0(1-p0)\left(\frac{1}{n1}+\frac{1}{n2}\right)}}$$

$$p1 = \frac{X1}{n1},\ p2 = \frac{X2}{n2},\ p0 = \frac{X1+X2}{n1+n2}$$

We then addressed the compositionally biased amino acid sites using two strategies: 1) removing biased sites and 2) modeling biased sites.

3.4.1. Removal of biased sites

After identifying the amino acids responsible for the most compositionally biased alignment sites, we developed a formula to systematically remove the most biased sites from the NM and RP dataset alignments. First, we divided our taxa into two bins depending on the biased sites we focused on (see chapters for more details). We had one bin enriched in a specific set of amino acids and one depleted in those amino acids (P_{down}). We then calculated the frequency of each set of amino

acids of interest (P_{AAset1}, P_{AAset2}) based on the analyses described above. A count constant (α) 0.01 was added to each amino acid frequency of interest (X_i). The summed frequencies for the amino acids of interest were then divided by the total number of amino acids counted (N) plus the count constant multiplied by the total number of possible amino acids ($\alpha d,\ d = 20$). This was calculated for both the up (P_{up}) and down (P_{down}) bins. Lastly, we calculated the total ratio (P_{total}) by taking the logarithmic value of the up bin divided by the down bin.

Biased-sites ranking formula:

$$P_{AAset1} = \frac{X_i + \alpha}{N + \alpha d},\ P_{AAset2} = \frac{X_i + \alpha}{N + \alpha d}$$

$$P_{up} = \frac{P_{FIMNKY}}{P_{GARP}},\ P_{down} = \frac{P_{FIMNKY}}{P_{GARP}}$$

$$P_{total} = log\left(\frac{P_{up}}{P_{down}}\right)$$

The alignment site ratios were then ranked in descending order based on the P_{total} value, and the most biased sites were progressively removed from the NM and RP alignments by 10% increments.

3.4.2. Modeling of biased sites

We also modified the new GFmix model (Muñoz-Gómez et al., 2022) by transforming the b parameter depending on the taxon group of interest (see each chapter for more details). This allowed us to model the compositionally biased sites instead of removing them. The GFmix model was initially designed to represent the GARP/FIMNKY amino acid ratio across all descendant taxa at each branch in a tree.

However, we transformed the b parameter for our analysis to fit the amino acid bins determined in the Z-score binning analysis described above. We then calculated the likelihood of different tree topologies under these variants of the GFmix model with the LG+C60+F+Γ4 model of sequence evolution. Branch length and alpha shape parameters for each tree tested were estimated using IQTREE v2.0.3 (Minh et al., 2020) and then fed into GFmix, specifying the custom enriched and depleted amino acid bins for the taxa of interest.

4. Several independent adaptations of archaea to hypersaline environments

4. Several independent adaptations of archaea to hypersaline environments

4.1. Context

This study started when our team realized we had potentially exciting MAGs assigned to new groups of extreme halophilic archaea. This was particularly interesting to my thesis because one of these groups looked to be a new deep-branching Nanohaloarchaeota lineage. The Nanohaloarchaeota is one of the most phylogenetically unstable DPANN lineages, so adding new taxonomic data to this group might help stabilize its position.

As my Ph.D. project initially focused on the DPANN, our taxonomic sampling favored DPANN lineages, constituting approximately half of the datasets. We were very lucky when we switched our focus to the halophiles because we already had all of the known groups of extreme halophilic archaea in our taxonomic sampling (Nanohaloarchaeota, Halobacteriota, Haloplasmatales, and Methanonatronarchaeia). We were lucky because each time we added new taxa to our dataset, I had to reanalyze the single gene trees (SGTs) for all our phylogenetic protein markers. I spent the first full year of my PhD manually curating our markers. Since the quality of the phylogenetic markers was at the core of all subsequent analyses, we spent a lot of time curating this dataset and ensuring we removed all paralogs, cases of horizontal gene transfers and contamination. In total, I performed four rounds of manual curation on 200 single-gene trees, which means I ended up reviewing over 800 trees.

Ultimately, this project took over three years to complete from start to submission. However, throughout this project, we developed many new tools and

methods that could be applied to various phylogenetic questions. For example, we established a collaboration with a team at Dalhousie University in Canada (Andrew Roger, Ed Susko, and Charley McCarthy) that helped us adapt a new phylogenetic model for dealing with compositional branch-heterogeneity. We also developed a ranking scheme to identify and remove compositionally-biased sites. Last, we performed in-depth comparative analyses between our ribosomal protein and NM datasets. We determined that ribosomal proteins are not robust when placing lineages with strong compositional biases. We later applied these methods to the DPANN project, and our analyses were significantly faster. I also ran a gene tree-species tree reconciliation analysis as implemented through the software ALE, and this was the first time anyone in our lab had done an analysis like this. This analysis had a huge learning curve, but the scripts and methods I developed for this analysis were later shared with other members of my lab, who used them to answer questions regarding the evolution of bacteria and eukaryotes.

We submitted the manuscript of this project to Nature Microbiology in July 2023 and received very positive reviews. As of writing this, we have resubmitted our revisions and await the final decision.

4.2. Results

Our results show that halophilic archaea independently adapted to hypersaline environments at least four times during archaeal evolution. We also robustly show with high confidence that the Nanohaloarchaeota is indeed a member of the DPANN archaea. As part of this project, we also described two new uncultured lineages, Afararchaeaceae and Asbonarchaeaceae, which break the long branches at the base of Haloarchaea and Nanohaloarchaeota, respectively. Our

findings highlight that unique amino acid compositions shared among the halophiles have artifactually placed these lineages together in previous phylogenetic analyses. We achieved more consistent and reliable phylogenetic placements by filtering out these biased data points and modeling these biases with the adapted branch-heterogeneity model. In this project, we also reconstructed the evolutionary history of archaeal gene families by mapping events such as gene duplications, transfers, originations, and losses using gene tree-species tree reconciliation methods. In this analysis, we focused on evolutionary events that impacted the ancestral branches leading to the various groups of halophilic lineages. These results suggested that gene duplication and horizontal gene transfer played an important role in the adaptation to halophily, for example, by spreading key genes (such as those encoding potassium transporters) across the various extreme halophilic lineages.

4.3. Draft manuscript 1

Several independent adaptations of archaea to hypersaline environments

Brittany A. Baker[1], Ana Gutiérrez-Preciado[1], Álvaro Rodríguez del Río[2], Charley G. P. McCarthy[3,4], Purificación López-García[1], Jaime Huerta-Cepas[2], Edward Susko[3,5], Andrew J. Roger[3,4], Laura Eme[1,*], and David Moreira[1,*]

[1]Ecologie Systématique Evolution, CNRS, Université Paris-Saclay, AgroParisTech, Gif-sur-Yvette, France.
[2]Centro de Biotecnología y Genómica de Plantas, Universidad Politécnica de Madrid (UPM) - Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA-CSIC), Madrid, Spain.
[3]Institute for Comparative Genomics, Dalhousie University, Halifax, Canada.
[4]Department of Biochemistry and Molecular Biology, Dalhousie University, Halifax, Canada.
[5]Department of Mathematics and Statistics, Dalhousie University, Halifax, Canada.
*correspondence: david.moreira@universite-paris-saclay.fr, laura.eme@universite-paris-saclay.fr

Abstract

Several extremely halophilic archaeal lineages thrive in saturating salt concentrations. They include Halobacteria (henceforth Haloarchaea), Nanohaloarchaeota, Methanonatronarchaeia, and Halarchaeoplasmatales. They maintain osmotic equilibrium with their environment using a 'salt-in' strategy, which involves pumping molar concentrations of potassium into the cells. This, in turn, has led to extensive proteome-wide accumulation of acidic amino acids to prevent protein aggregation. However, the evolutionary history underlying these adaptations remains poorly understood. In particular, the number of times that these dramatic proteome-sweeping changes occurred is unclear due to the conflicting phylogenetic positions found for several of these lineages. Here, we present a resolved phylogeny of extremely halophilic archaea obtained using improved taxon sampling and state-of-the-art phylogenetic approaches designed to cope with the strong compositional biases of their proteomes. We describe two new uncultured lineages, Afararchaeaceae and Asbonarchaeaceae, which break the long branches at the base of Haloarchaea and Nanohaloarchaeota, respectively. Our extensive phylogenomic analyses show that at least four independent adaptations to extreme halophily occurred during archaeal evolution. Gene-tree/species-tree reconciliation suggests that gene duplication and horizontal gene transfer played an important role in this process, for example, by spreading key genes (such as those encoding potassium transporters) across the various extremely halophilic lineages.

Main

For decades, all known extremely halophilic archaea (growing at salt concentrations >30% w/v) belonged to the Haloarchaea[1,2]. Recently, metagenomics uncovered additional groups, whose phylogenetic positions have been unclear (Extended Data Fig. 1): i) Nanohaloarchaeota[3–5], tiny symbiotic archaea initially thought to be closely related to the Haloarchaea but placed later in the DPANN super-group[6], suggesting an independent adaptation to extreme salinity; ii) Methanonatronarchaeia[7], a class of extremely halophilic methanogens, initially proposed to be an "evolutionary intermediate" between non-halophilic Class II methanogens and Haloarchaea, but placed at the base of Methanotecta in more recent studies[8–11]; and iii) Halarchaeoplasmatales[12], an order robustly placed within Thermoplasmatota. These extremely halophilic archaea have evolved unique strategies to cope with osmotic stress: they pump high levels of potassium into their cells[13] and maintain acidic proteomes, rich in aspartic and glutamic acids and depleted in basic and large hydrophobic amino acids[14–17]. These amino acid usage biases and the higher evolutionary rate at the base of halophilic archaea can lead to long-branch attraction (LBA) and other phylogenetic reconstruction artefacts, resulting in conflicting evolutionary relationships[9,18,19]. Thus, how many times these adaptations evolved remains enigmatic. Here, we introduce two new families of extreme halophiles, Afararchaeaceae and Asbonarchaeaceae. With sophisticated methods and broader taxonomic sampling, we establish a comprehensive phylogeny of halophilic archaea. Our updated scenario highlights at least four independent adaptations to hypersaline environments and emphasizes the adaptive role of horizontal gene transfer (HGT) between different halophilic groups.

Results

Characterization of two new groups of extremely halophilic archaea

The Danakil Depression (Afar region, Ethiopia) contains hypersaline lakes hosting extremely halophilic archaea[20,21]. Among the metagenome-assembled genomes (MAGs) reconstructed from these lakes[22], we identified 13 belonging to two new lineages of extreme halophiles, plus one additional MAG placed deep in the Haloarchaea (Fig. 1a,c, Supplementary Data 1).

The first group – a novel family-level lineage named Afararchaeaceae, after Ethiopia's Afar region – was represented by four moderately GC-rich (53-60%) MAGs with average nucleotide identity (ANI) values between 72 and 74% among them (Supplementary Data 2, Supplementary Fig. 1a,b). Afararchaeaceae branched with maximal support as a sister lineage to the group UBA12382 (or 'hikarchaea'[10])+Haloarchaea (Fig. 1a). Initially described as intermediates between non-halophilic methanogens and Haloarchaea[10], this result suggests that hikarchaea adapted secondarily to low salinity from an extremely halophilic ancestor.

The most complete afararchaeal MAG (DAL-WCL_na_97C3R), formally named *Afararchaeum irisae* gen. nov., sp. nov. (see description below), had a size of ~1.9 Mbp (Supplementary Data 1). KEGG annotation[23] indicates that Afararchaeaceae are likely heterotrophic aerobes that utilize branched-chain amino acids as a carbon source, similar to many known Haloarchaea[24] (Fig. 1b, Supplementary Data 3). They are probably mobile, possessing all genes for the archaeal flagellum (archaellum)[25] and a chemotaxis operon. Additionally, afararchaeal MAGs encode a single type-II sensory rhodopsin for phototaxis[26], but lack bacteriorhodopsin genes, suggesting that these archaea do not use light as an additional energy source like many Haloarchaea[27]. As expected, Afararchaeaceae likely employ a salt-in osmoregulation involving multiple K^+ transporters (eight Trk-like and two Kef-

like), mechanosensitive ion channels (MscS and MscL), and Na^+/Ca^{2+} exchangers (Supplementary Data 2). Consequently, they also exhibit a highly acidic proteome (Fig. 2a,b).

The second group comprises nine MAGs (46-64% GC content) with ANI values between 74 and 79% among them (Supplementary Data 2, Supplementary Fig. 1c,d). They branched as a sister group to the DPANN family Nanosalinaceae (Fig. 1c) and are related to MAGs that were previously classified as 'Nanoanaerosalinaceae' and 'Nanohalalkaliarchaeaceae'[5] (Supplementary Fig. 4). However, these two families have been merged within the family 'JALIDP01' in GTDB[28]. Our MAGs provide substantial coverage of this family, with three related to the former 'Nanoanaerosalinaceae' and six to the single MAG representing the 'Nanohalalkaliarchaeaceae'[5] (Supplementary Fig. 4). Given this taxonomic uncertainty and their presence in both anoxic[5] and oxic (this work) environments, we propose formally naming this family Asbonarchaeaceae, derived from '*asbo*', meaning salt in the Afar language, acknowledging their consistent presence in hypersaline systems.

DPANN genomes, like those in the Asbonarchaeaceae, typically lack certain genes, leading to an underestimated genome completeness, typically maximally ~85%[29,30]. We thus likely obtained a nearly complete asbonarchaeal MAG (DAL-WCL_45_84C1R, 84% complete) representing the type species for this family, *Asbonarchaeum danakilense* gen. nov., sp. nov. (see description below), with a genome size of 1.2 Mbp, similar to other DPANNs[29] (Supplementary Data 1). Asbonarchaeaceae lack crucial biosynthetic pathways (lipid, nucleotide, and amino acid biosynthesis), suggesting they live symbiotically, relying on a host like other DPANN groups[31–33] (Fig. 1d and Supplementary Data 4). They lack a canonical electron transport chain but possess all essential subunits of a V/A-type ATP synthase (Fig. 1d)[31]. We again predict that Asbonarchaeaceae employ salt-in osmoregulation with multiple K^+ transporters (Supplementary Data 4) and a highly acidic proteome (Fig. 2a,b). Despite their phylogenetic relationship with the Nanosalinaceae, they display a distinct amino acid composition (Fig. 2a), confirming their status as a new group within the Nanohaloarchaeota.

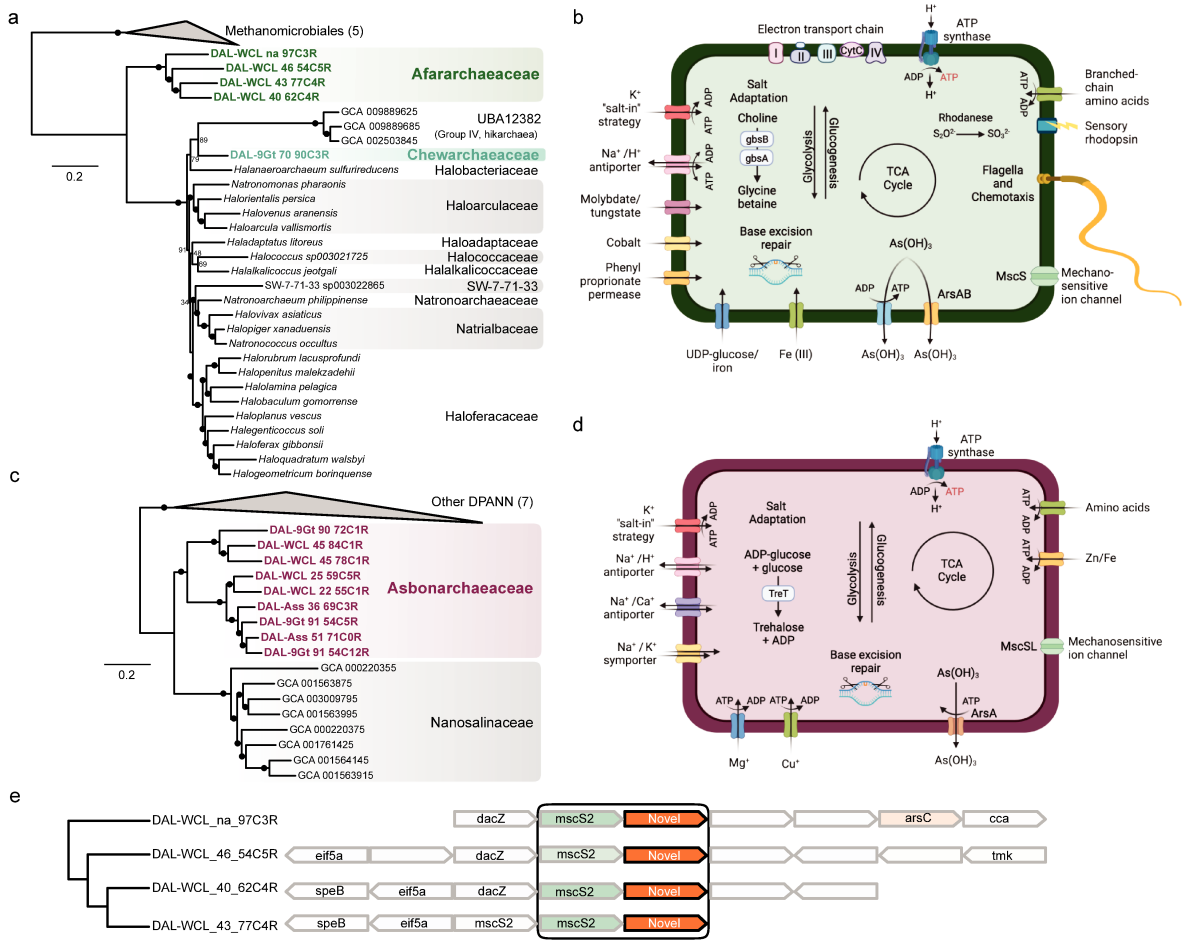


Fig. 1 | Phylogenetic position and metabolic potential of the new families Afararchaeaceae and Asbonarchaeaceae. (a) Maximum likelihood phylogenetic tree of 35 euryarchaea, including the four new Afararchaeaceae MAGs (highlighted in green), based on the concatenation of 122 single-copy proteins obtained from the Genome Taxonomy Database (GTDB). The tree was inferred via IQ-TREE with the LG+C60+F+Γ4 model of sequence evolution. The statistical support for branches, with filled circles representing values equal to or larger than 99% support, corresponds to 1,000 ultra-fast bootstrap replicates. The scale bar indicates the expected average number of substitutions per site. All taxonomic ranks shown are based on the GTDB r207 family-level classification. See Supplementary Fig. 2 for the uncollapsed tree. **(b)** Non-exhaustive metabolic scheme based on the predicted gene content of the most complete afararchaeal MAG (DAL-WCL_na_97C3R). A detailed table of the predicted gene content can be found in Supplementary Data 3. **(c)** Maximum likelihood phylogenetic tree of 24 DPANN archaea, including the nine new Asbonarchaeaceae MAGs (highlighted in wine), based on the concatenation of 99 single-copy proteins obtained from GTDB. The tree was inferred by IQ-TREE with the LG+C60+F+Γ4 model of sequence evolution. The statistical support for branches corresponds to 1,000 ultra-fast bootstrap replicates. The scale bar indicates the expected average number of substitutions per site. All taxonomic ranks are based on the GTDB r207 family-level classification. See Supplementary Fig. 3 for the uncollapsed tree. **(d)** Non-exhaustive metabolic scheme based on the predicted gene content of the most complete asbonarcharchaeal MAG (DAL-WCL_45_84C1R). A detailed table of the predicted gene content can be found in Supplementary Data 4. **(e)** Gene maps showing a novel gene family (orange) linked to a conserved mechanosensitive ion channel (mscS2) in

the afararchaeal MAGs. Gene abbreviations are as follows: agmatinase (speB), eukaryotic initiation factor 5A (eif5a), di-adenylate cyclase (dacZ), arsenate reductase (arsC), tRNA nucleotidyltransferase (cca), thymidylate kinase (tmk).

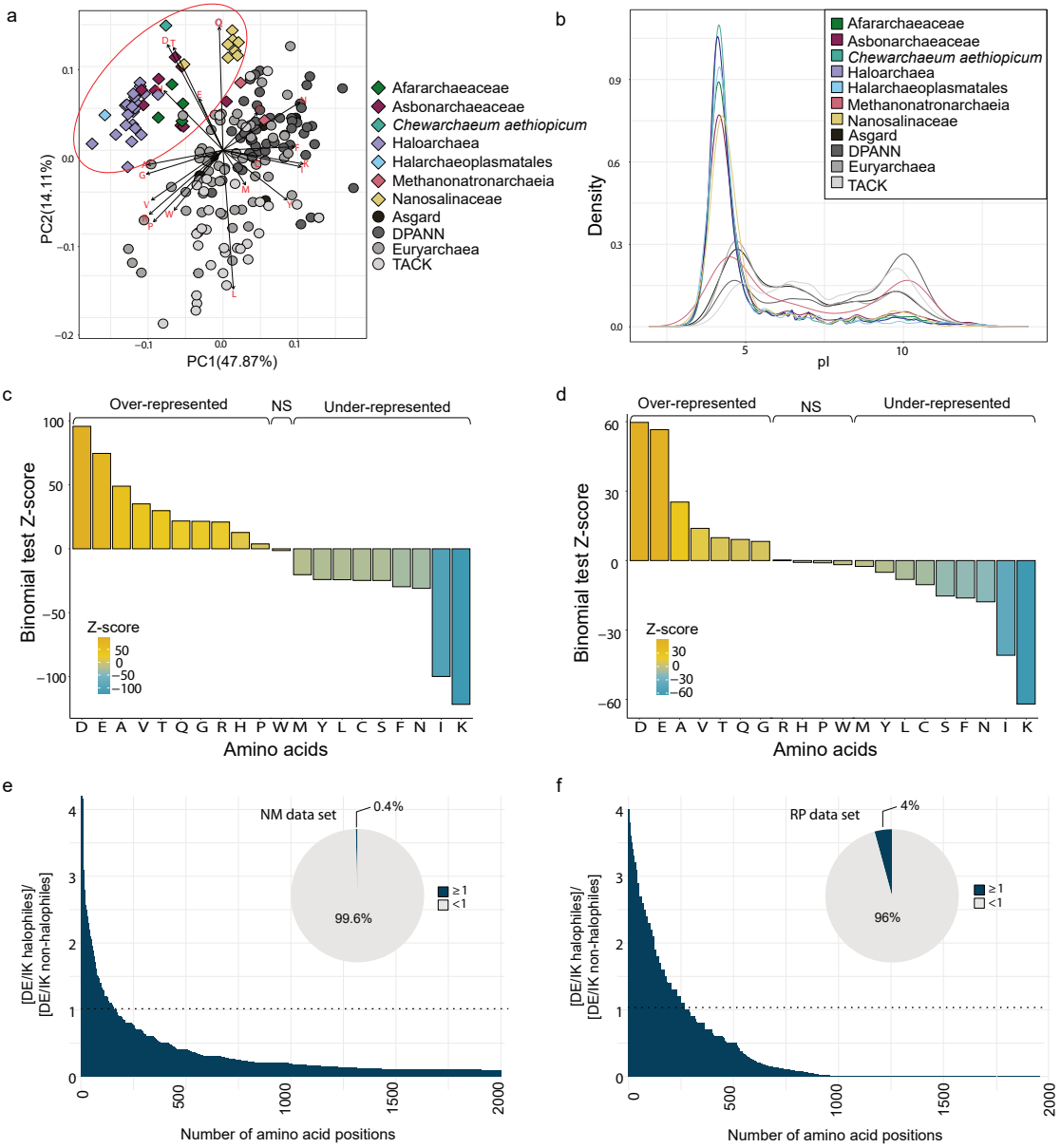


Fig. 2 | Protein amino acid compositional biases in extremely halophilic archaeal lineages. **(a)** PCA plot of 192 archaeal proteomes based on amino acid frequencies. The red ellipse indicates the clustering of all extreme halophiles (colored diamonds), including the newly identified families Afararchaeaceae (green color) and Asbonarchaeaceae (wine color). **(b)** Isoelectric point (pI) distribution of 192 archaeal proteomes. Non-halophilic archaea (grey lines) display a bimodal distribution of pI values, while extreme halophiles (colored lines) exhibit a single spike at pI ~4, indicating a highly acidic proteome. **(c,d)** D+E/I+K site-by-site bias (defined as the ratio [D+E/I+K for halophiles]/[D+E/I+K for non-halophiles]) for the 2,000 most biased sites of the **(c)** NM dataset (39,385 amino acid positions) and **(d)** RP dataset (6,792 amino acid positions). Inset pie charts depict the proportion of amino acids with a ratio greater than or equal to 1 (dark blue) versus less than 1 (grey). **(e,f)** Binomial tests for the **(e)**

NM and **(f)** RP datasets compare the proportions of all 20 amino acids between extreme and non-halophiles. Z-scores were calculated relative to extreme halophiles, with |Z| > 1.96 indicating significant enrichment of a given amino acid in extreme halophile sequences ("Over-represented"), |Z| < -1.96 indicating significant depletion of a given amino acid in extreme halophile sequences ("Under-represented"), and some amino acids showing no significant bias ("NS").

Novel gene families in Afararchaeaceae and Asbonarchaeaceae

We identified novel gene families using a two-step approach. First, we searched for genes in Afararchaeaceae and Asbonarchaeaceae genomes with no detectable homologs in sequence databases of cultured organisms (RefSeq[34], Pfam[35], and EggNOG[36]), revealing a significant number of potentially novel genes (10-30% of their total genes; Extended Data Fig. 2a). Second, we compared these genes against a vast collection of 169,529 prokaryotic genomes[37], confirming that only 14% (Asbonarchaeaceae) and 17.1% (Afararchaeaceae) have related genes in other uncultured species, highlighting many unknown lineage-specific genes (Supplementary Data 4 and 5). Notably, these novel genes encode proteins with an acidic pH isoelectric point, aligning with adaptation to hypersaline environments[38] (Extended Data Fig. 2b). A considerable percentage of these proteins contain transmembrane domains or signal peptides, likely targeting them to the membrane or extracellular space, directly interacting with the external high salt concentrations. We analyzed their genomic context to predict their functions. Approximately 5% (Afararchaeaceae) and 18% (Asbonarchaeaceae) of them have conserved synteny and co-localize with genes with known functions, indicating roles related to those of their neighboring genes (Supplementary Data 5 and 6). For example, we found a novel protein in Afararchaeaceae next to a mechanosensitive ion channel (Fig. 1e), suggesting a potential role in osmotic regulation[39].

Identifying a conserved core of archaeal phylogenetic markers

Previous attempts to determine the phylogenetic placement of extreme halophiles mainly relied on limited datasets like single proteins[3,9] or concatenated ribosomal proteins[7,10]. However, these small datasets contain few sites, and provide limited phylogenetic information[18,40]. Moreover, ribosomal proteins may have compositional biases that differ from the rest of the proteome due to their complex protein-protein and protein-RNA interactions[18,41]. To address these issues and accurately determine the positions of extreme halophilic archaea, we conducted a comprehensive phylogenomic analyses using a dataset of 136 new marker proteins (NM dataset; 39,385 positions) highly conserved among archaea[18]. These proteins serve various functions (Supplementary Data 7), reducing potential biases linked to co-evolution patterns. Based on individual phylogenetic trees, we manually curated our NM marker set to exclude possible HGT or hidden paralogy (see Methods). Additionally, we curated a set of 48 ribosomal proteins (RP dataset, 6,792 positions) to compare their phylogenetic signal with that of the NM dataset.

Testing the influence of taxon sampling

Extreme halophilic archaea often display long branches, potentially yielding artefactual placements due to LBA[42,43]. To address this, we employed different datasets and approaches. In addition to the full dataset (Fig. 3a, Extended Data Figs. 3 and 4), we used smaller taxon samplings, focusing on specific archaeal groups such as Euryarchaeota only (including Afararchaeaceae) and Euryarchaeota+Nanohaloarchaeota (Supplementary Figs. 5-8,

Supplementary Data 8). The corresponding phylogenies revealed congruent placements for all extreme halophiles except Methanonatronarchaeia. NM-based maximum likelihood (ML) trees grouped them with Methanotecta (i.e., Haloarchaea, 'hikarchaea', Class II methanogens, Methanopagales, ANME-1, Synthrophoarchaeales, and Archaeoglobales) or with the Afararchaeaceae+'hikarchaea'+Haloarchaea (AHH) clade, while RP-based ML trees placed them as sisters to the AHH-clade. The two topologies were significantly different based on an approximately unbiased (AU) test[44] since the NM topology was rejected based on the RP alignment (P-value=0.0000431) and the RP topology was rejected based on the NM alignment (P-value=0.0000165). Bayesian analyses, with four Markov chain Monte Carlo (MCMC) chains each and applying the complex CAT+GTR model, showed similar conflicting placements (Supplementary Figs. 9-12), highlighting how different taxon samplings, models, and phylogenetic frameworks can showcase conflicting signals in phylogenetic analyses of Methanonatronarchaeia. These results underscore the challenges in placing extreme halophiles accurately, most likely because of their unique compositional biases linked to their 'salt-in' osmoregulation strategy[14–17], which are not properly modeled by standard substitution models[45].

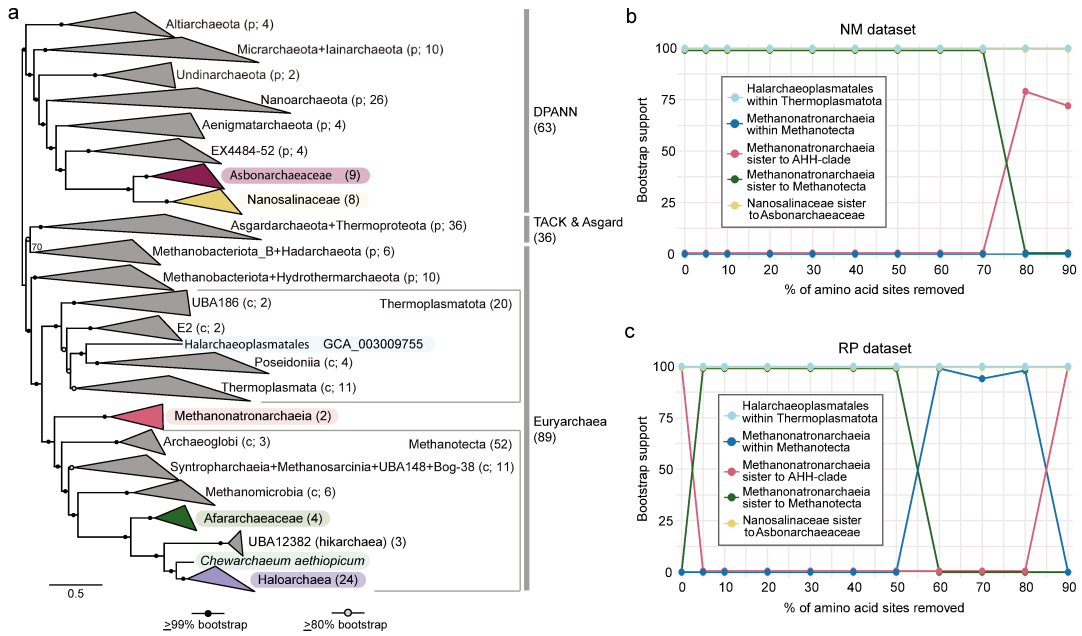


Fig. 3 | Maximum likelihood phylogeny of archaea, including the new groups Afararchaeaceae and Asbonarchaeaceae. (a) Phylogenetic tree based on the concatenation of 136 conserved markers (NM dataset) across 192 taxa (39,385 sites) via IQ-TREE under the LG+C60+F+Γ4 model of evolution. Statistical support indicated on the branches corresponds to 1,000 ultra-fast bootstrap replicates. The scale bar indicates the number of substitutions per site. Colors indicate the currently known groups of extremely halophilic archaea. The size of collapsed clades is indicated in parentheses; see Extended Data Fig. 3 for the uncollapsed tree. **(b,c)** Impact of the progressive removal (in steps of 10%) of the most compositionally biased sites from the **(b)** 192-NM (39,385 amino acid positions) and **(c)** 192-RP (6,792 amino acid positions) datasets. Lines show the statistical support values for the position of each of the halophilic clades of interest. These support values were estimated using the ultrafast bootstrap approximation from the ML tree reconstruction (LG+C60+F+Γ4 model) for each site-removal step.

Addressing the effect of compositional biases

Model misspecification induced by compositional bias is a known source of phylogenetic error. To reduce potential LBA artifacts affecting extreme halophiles, previous studies either recoded data into four character states[10,46] or removed the fastest-evolving sites[8,10,46]. However, the latter method resulted in the loss of up to 50% of alignment sites, which is problematic for small datasets like the RP-based ones[11]. Therefore, we explored two alternative approaches to address halophile-specific compositional biases while preserving substantial phylogenetic information.

First, we identified amino acids significantly over or under-represented in extreme halophiles compared to non-halophiles in the 192 taxa NM and RP datasets. D+E and I+K were the most over-and under-represented amino acids, respectively (Fig. 2c,d). We then applied the GFmix model[45] to cope with these specific compositional biases. GFmix is a site-heterogeneous mixture model that adjusts amino acid frequencies for each class of the mixture model in a branch-specific manner to accommodate shifts in amino acid composition over the branch. Amino acids were categorized into three groups: those that increased, decreased, or remained unchanged in frequency on the branch. We used the LG+C60+F+Γ4 model with GFmix (GFmix-DE/IK model), where [D+E]/[I+K] compositional ratio varied over branches. Despite improvements in likelihood values under this model, the RP and NM datasets remained incongruent regarding the position of Methanonatronarchaeia (Extended Data Fig. 5). We also explored a GFmix variant with larger groups of significantly over and under-represented amino acids (Fig. 2c,d). Although it further improved the likelihood, the relative preferences of topologies for each dataset remained unchanged (Extended Data Fig. 5).

Our second approach involved the gradual removal of highly compositionally biased alignment sites. We calculated the D+E/I+K ratio for halophilic versus non-halophilic lineages, ranked the sites accordingly, and then progressively removed the most biased sites. For the 192 taxa NM dataset, the position of Methanonatronarchaeia remained unchanged until 80% of sites were removed, after which they branched as the sister group of the AHH-clade with weak support (Fig. 3b). By contrast, for the 192 taxa RP dataset, Methanonatronarchaeia shifted to a fully supported sister position to Methanotecta with only 5% of the most biased sites removed (Fig. 3c). This indicates that while the NM dataset does contain sites with biased D+E/I+K ratios (Extended Data Fig. 6), their impact is very minor compared to the RP dataset, which has a higher proportion of highly biased sites (0.4% versus 4% of positions with a ratio ≥1, respectively; Fig. 2e,f).

We examined the ribosomal proteins with the most biased sites (e.g., L1, L12e, S6, and S15) and found they were located on the outer surface of the ribosomal complex, in close interaction with the K^+-rich cytoplasm (Supplementary Fig. 13 and Supplementary Video 1). To confirm the impact of the D+E/I+K bias on the RP-based phylogeny, we inferred an ML tree using a concatenation of the 18 most biased ribosomal proteins, which resulted in all extremely halophilic groups clustering with 100% support (Supplementary Fig. 14). We also reconstructed Bayesian phylogenies with 20% of the most biased alignment sites removed for both the 104-NM and 104-RP datasets. Contrary to trees constructed with the untreated datasets (see above), all MCMC chains for both datasets supported the deeper-branching position of Methanonatronarchaeia sister to Methanotecta (Supplementary Figs. 15 and 16).

A recent study suggested that, given their slow evolutionary rate and their belonging to a single complex, ATP synthase subunits A and B are less susceptible to phylogenetic artifacts[9]. A phylogeny based on the concatenation of both subunits supported the Nanohaloarchaeota

sister to Haloarchaea[9,47]. However, when we removed 15% of the highest D+E/I+K ratio sites from this dataset, Nanohaloarchaeota branched deeper (Extended Data Fig. 7), indicating that a few highly biased sites artificially drove their position close to Haloarchaea.

In conclusion, our phylogenomic analyses, especially those mitigating the strong convergent compositional bias shared by the halophilic lineages, robustly support at least four independent adaptations to extreme halophily in archaea: in the AHH-clade, Methanonatronarchaeia, Halarchaeoplasmatales, and Nanosalinaceae+Asbonarchaeaceae.

Tree-aware reconstruction of gene content evolution in archaeal extreme halophiles

We used the amalgamated likelihood estimation (ALE) method to examine gene content evolution in the 192 taxa dataset. By reconciling individual gene trees with the species tree (Fig. 3a), we estimated gene duplications, transfers, originations, losses, and copy numbers at all ancestral nodes. This approach included Methanonatronarchaeia, previously excluded from similar analyses due to their unresolved phylogenetic position[10]. Gene transfer and loss appear to be the primary drivers of gene content evolution in archaea, including halophilic groups (Fig. 4, Extended Data Figs. 8-9). Haloarchaea, with some of the largest genome sizes among archaea[48], also experienced significant gene originations and duplications during their early evolution. This expansion involved key inorganic ion transporters (Trk and Kef-type K^+ transporters, Mg^{2+} transporters, SSF Na^+/solute symporters, NhaP-type K^+/H^+ antiporters, Ca^+/Na^+ and Na^+/H^+ antiporters) crucial for osmotic regulation (Supplementary Figs. 17-24, Extended Data Fig. 10a), and molecular chaperones like GrpE (Supplementary Fig. 25), which prevents protein aggregation during response to hyperosmotic stress[49]. Amino acid transporters, vital for species of these groups thriving on amino acids[24], also exhibited duplications (Extended Data Fig. 9). Presence probabilities estimated by ALE at key halophilic ancestors are reported for each of these proteins in Supplementary Data 9.

Halarchaeoplasmatales also had numerous gene duplications, spanning metabolism and informational processes like transcription, DNA replication, and repair (Extended Data Fig. 9). In Nanosalinaceae and Asbonarchaeaceae, gene transfer was dominant but less pronounced due to constraints in these small-sized archaea to maintain compact genomes[50]. By contrast, the 'hikarchaea' displayed extensive gene loss, which supports the hypothesis that these marine archaea evolved from extremely halophilic ancestors (the Hik-Haloarchaea ancestor with 1,323 inferred protein-coding genes, Fig. 4) and adapted to nutrient-poor deep-sea environments through gene loss, typical of many streamlined marine prokaryotes[51,52]. Nevertheless, this adaptation also included duplications of specific genes linked to energy production, conversion, and carbohydrate and amino acid transport and metabolism (Extended Data Fig. 9). Notably, we observed multiple copies of aerobic-type carbon monoxide dehydrogenase, found in other microorganisms adapted to the same nutrient-poor environments[53] (Supplementary Fig. 26).

Massive HGT from bacteria has likely played a significant role in the evolution of Haloarchaea, although its extent and timing are still debated[54–57]. Several transfers happened before the split between Afararchaeaceae and Haloarchaea, facilitating the adaptation of their common ancestor to extreme halophily. For instance, the choline dehydrogenase BetA, involved in glycine-betaine osmoprotectant synthesis[58], was acquired through HGT (Extended Data Fig. 10b). Notably, this gene is absent in hickarchaea, reinforcing the idea of gene loss during their secondary adaptation to low-salt environments. Another example is a BCCT family transporter involved in osmoprotectant uptake, such as glycine and betaine[58], which Methanonatronarchaeia acquired from bacteria (Supplementary Fig. 27).

HGT between Haloarchaea and other halophilic archaea has also played a role in their convergent adaptations to extreme salinity. Examples include the chaperone GrpE and various multi-copy transporters like K^+ (Trk- and Kef-type) and Mg^{2+} transporters, and K^+/H^+, Ca^{2+}/Na^+, and Na^+/H^+ antiporters. Additionally, other inorganic molecule transporters have been transferred among halophilic archaeal groups, such as SNF-family Na^+-dependent transporters, ZupT- and FieF-type metal transporters, sulfur transporters, Na^+/H^+ antiporters, and Na^+/phosphate symporters (Supplementary Figs. 28-34).

HGT of organic molecule transporters is also observed, such as a transporter of Krebs cycle intermediates shared by Haloarchaea and Asbonarchaeaceae (Supplementary Fig. 35). We also identified genes of bacterial origin encoding various transporters subsequently transferred between different halophilic archaeal groups. These include an AmiS/UreI urea transporter transferred between Haloarchaea and Nanohaloarchaea and a TauE/SafE sulfite exporter transferred between Haloarchaea and Methanonatronarchaeia (Supplementary Figs. 36-37), consistent with previous reports of inter-domain HGT followed by intra-domain HGT[59].

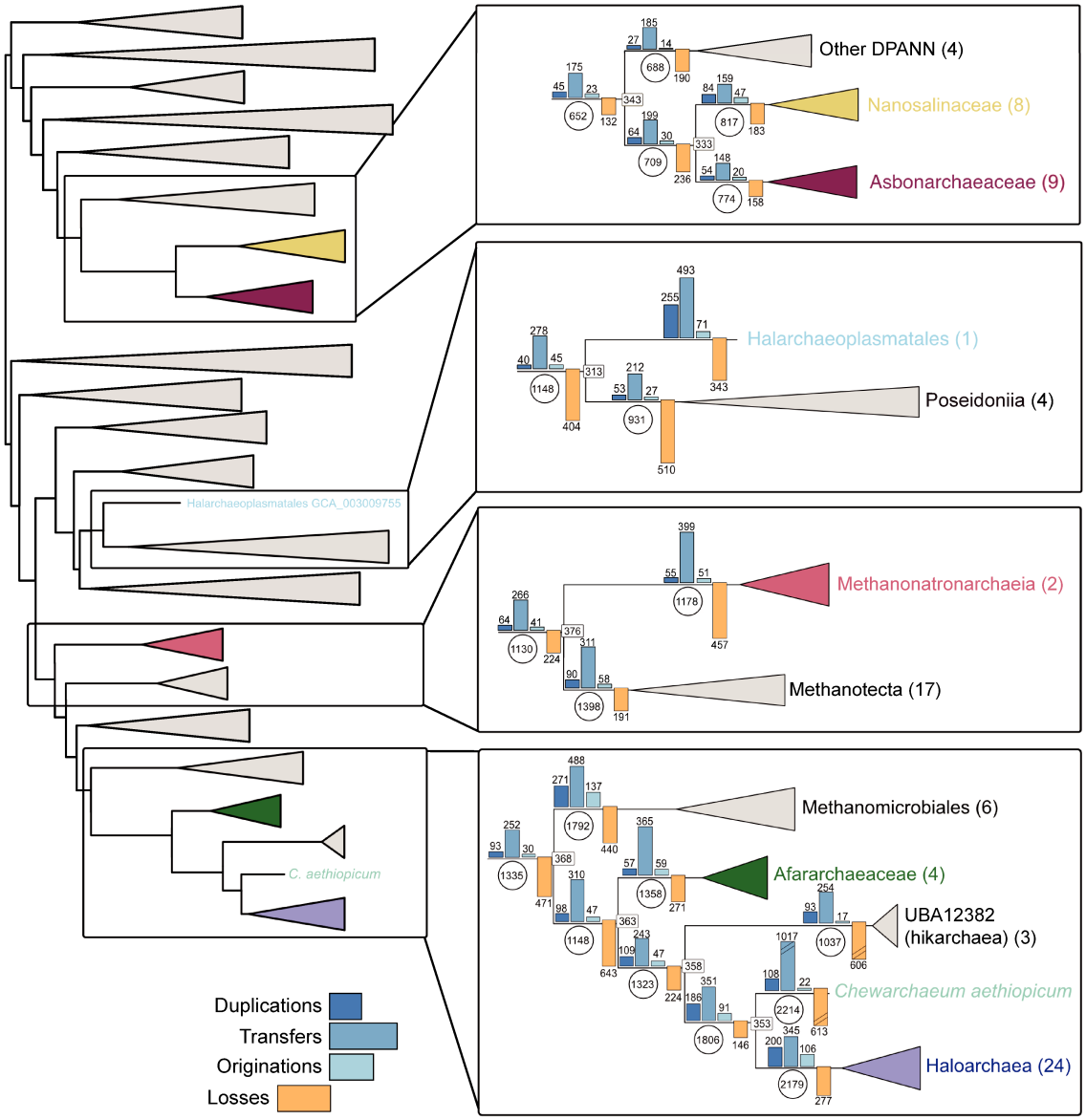


Fig. 4 | Schematic representation of the tree reconciliation analysis based on the NM species tree. The full archaeal tree is shown on the left; boxes on the right highlight the details for the four main groups of halophilic archaea: Nanosalinaceae+Asbonarchaeaceae, Halarchaeoplasmatales, Methanonatronarchaeia, and Afararchaeaceae+Haloarchaea. The bar plots on the branches represent the number of gene duplications, transfers, originations, and losses, and the circles indicate the number of predicted ancestral gene copy numbers. The number of taxa in each collapsed clade is indicated by the number in parentheses next to the clade name. The complete version of this tree with the events for all archaeal nodes can be found in Extended Data Fig. 8.

Conclusions

Our study yields a robust archaeal phylogeny, including two newly discovered halophilic lineages, Asbonarchaeaceae (closely related to Nanosalinaceae within the DPANN) and Afararchaeaceae (closely related to the Haloarchaea+'hikarchaea' group). The position of Afararchaeaceae challenges the previous notion of 'hikarchaea' being intermediates between methanogens and haloarchaea[10], as they instead adapted secondarily to low salinity from extremely halophilic ancestors. Our phylogenomic analyses also position Methanonatronarchaeia as sister to Methanotecta, not as intermediates between Class II methanogens and haloarchaea[7]. Thus, we identify four independent adaptations to extreme halophily in archaea: in Haloarchaea+Afararchaeaceae, Methanonatronarchaeia, Halarchaeoplasmatales, and Nanosalinaceae+Asbonarchaeaceae. All these adaptations involve a salt-in strategy with convergent independent extensive proteome acidifications. In addition, HGT played a crucial role in spreading key genes, such as ion transporters, among these halophilic lineages. This prompts the question of whether the initial adaptations to extreme halophily occurred as a singular event in one group, spreading through HGT to the other groups, and which lineage of extreme halophiles emerged first. Answering these intriguing questions will require further investigation of adaptive genes and their distribution in known and potential novel halophilic archaea.

Taxonomic descriptions

All new names have been described under the SeqCode[60] as follows:

Description of *Afararchaeum* gen. nov.

Afararchaeum (A.far.ar.chae'um. N.L. neut. n. archaeum, an archaeon; N.L. neut. n. *Afararchaeum*, an archaeon from the Afar region). Type species: *Afararchaeum irisae*.

Description of *Afararchaeum irisae* sp. nov.

Afararchaeum irisae (i.ri'sae. N.L. gen. n. irisae, named after the Iris Foundation (France), which supports the study and preservation of endangered ecosystems including those in the Afar region. This archaeon lives in oxic hypersaline waters. It encodes genes for aerobic respiration and likely uses amino acids for organoheterotrophic growth. Its genome is around 1.9 Mbp (GC content: 55%). It is known from environmental sequencing only. DAL-WCL_na_97C3R is the designated type MAG.

Description of Afararchaeaceae fam. nov.

Afarachaeaceae (A.far.ar.chae.a.ce'ae. N.L. neut. n. *Afararchaeum*, a genus name; -aceae, ending to denote a family; N.L. fem. pl. n. Afararchaeaceae, the *Afararchaeum* family).

Description of *Asbonarchaeum* gen. nov.

Asbonarchaeum (As.bon.ar.chae'um. asbo, salt in the Afar language; N.L. neut. n. archaeum, an archaeon; N.L. neut. n. *Asbonarchaeum*, a salt archaeon). Type species: *Asbonarchaeum danakilense*.

***Description of* Asbonarchaeum danakilense sp. nov.**

Asbonarchaeum danakilense (da.na.kil.en'se. N.L. neut. adj. danakilense, pertaining to the Danakil Depression). This halophilic archaeon lives in oxic hypersaline waters of the Danakil Depression. It has a ~1.2 Mb streamlined genome (GC content: 61%). It lacks most biosynthetic pathways, most likely growing as a symbiont of an unknown host. It is known from environmental sequencing only. DAL-WCL_45_84C1R is the designated type MAG.

Description of Asbonarchaeaceae fam. nov.

Asbonarchaeaceae: (As.bon.ar.chae.a.ce'ae. N.L. neut. n. *Asbonarchaeum*, a genus name; -aceae, ending to denote a family; N.L. fem. pl. n. Asbonarchaeaceae, the *Asbonarchaeum* family).

Description of *Chewarchaeum* gen. nov.

Chewarchaeum (Chew.ar.chae'um. chew, salt in the Amharic language; N.L. neut. n. archaeum, an archaeon; N.L. neut. n. *Chewarchaeum*, a salt archaeon). Type species: *Chewarchaeum aethiopicum*.

***Description of Chewarchaeum aethiopicum* sp. nov.**

Chewarchaeum aethiopicum (ae.thi.o'pi.cum. L. neut. adj. aethiopicum, Ethiopian). This halophilic archaeon lives in oxic hypersaline waters of the Danakil Depression. It encodes genes for aerobic respiration and likely uses amino acids for organoheterotrophic growth. Its genome is around 2.9 Mb (GC content: 61%). It is known from environmental sequencing only. DAL-9Gt_70_90C3R is the designated type MAG.

Description of *Chewarchaeaceae* fam. nov.

Chewarchaeaceae (Chew.ar.chae.a.ce'ae. N.L. neut. n. *Chewarchaeum*, a genus name; -aceae, ending to denote a family; N.L. fem. pl. n. Chewarchaeaceae, the *Chewarchaeum* family).

Methods

Selection of metagenome-assembled genomes

We searched for MAGs related to known groups of extremely halophilic archaea in the Danakil Depression dataset obtained by Gutiérrez-Preciado et al[22]. For this, we included 61 Danakil MAGs in a preliminary phylogenetic tree containing 488 representatives of archaeal diversity and constructed a phylogenetic tree using 49 concatenated ribosomal proteins with IQ-TREE v1.6.10[61] (Supplementary Fig. 38). The tree was built using the LG+C20+F+Γ4 model of sequence evolution and support at branches was estimated from 1,000 ultrafast bootstrap replicates. From this analysis, we selected 14 high-quality MAGs (>50% completeness, ≤5% redundancy) representing potential new groups of extremely halophilic archaea based on their position compared to other known halophilic archaea. These 14 MAGs were taxonomically classified using GTDB-TK[28] (version 2.3.0, r207; April 1st, 2022) and assigned to novel families within three GTDB orders: four MAGs were assigned to a novel family belonging

to the order 'JAHENH01', for which we have named Afararchaeaceae; nine MAGs were assigned to another novel family belonging to the order Nanosalinales, for which we propose to name Asbonarchaeaceae; and one MAG belonged to a third novel family in the order Halobacteriales, for which we have named Chewarchaeaceae (see taxonomic description above for more details). The pairwise ANI values for the four Afararchaeaceae MAGs (Supplementary Fig. 1a) and nine Asbonarchaeaceae MAGs (Supplementary Fig. 1c) were calculated using FastANI[62]. The pairwise AAI values for the four Afararchaeaceae MAGs (Supplementary Fig. 1c). and nine Asbonarchaeaceae MAGs (Supplementary Fig. 1d) were calculated using an online calculator[63]. This AAI calculator estimates the AAI using the reciprocal best hits (two-way AAI) between two genomic datasets of proteins.

Metagenome-assembled genome annotation

Coding DNA sequences (CDSs) were predicted with Prodigal v2.6.3[64] and subjected to Pfam[35] and COG[65] functional annotations inside the Anvi'o v5 pipeline[66]. Genes were also annotated with KofamKOALA[67] and eggNOG-mapper v2.1.5[36]. Additional manual curation was done for the two most complete Afararchaeaceae and Asbonarchaeaceae MAGs (DAL-WCL_na_97C3R and DAL-WCL_45_84C1R, respectively). Further information on gene annotations and functional predictions can be found in Supplementary Data 3 and 4.

Detecting novel protein families in Afararchaeaceae and Asbonarchaeaceae

We computed family clusters of the proteins predicted for the MAGs of the new archaeal families Afararchaeaceae and Asbonarchaeaceae using Mmseqs2[68] with relaxed thresholds: minimum percentage of amino acids identity of 30%, e-value <1e-3, and a minimum sequence coverage of 50% (--min-seq-id 0.3 -c 0.5 --cov-mode 2 --cluster-mode 0). To detect families with no homologs in reference databases, we mapped i) the protein sequences encoded in the MAGs against EggNOG using eggNOG-mapper v2[36] (hits with an e-value <1e-3 were considered as significant) ii) the protein sequences encoded in the MAGs against PfamA domains using HMMER[69] (hits with an e-value <1e-5 were considered as significant), iii) the protein sequences encoded in the MAGs against PfamB domains using HMMER[69] (hits with an e-value < 1e-5 were considered as significant) and iv) the CDS sequences of the MAGs against RefSeq using Diamond BLASTx[70] ('sensitive' flag, hits with an e-value <1e-3 and query coverage >50% were considered as significant). We only considered novel families those with no detectable homologs in these databases. To address the taxonomic breadth of the novel families, we used Diamond BLASTp[70] ('sensitive' flag, hits with an e-value <1e-3 and query coverage >50% were considered as significant) to map the longest sequence of each family against the proteins encoded in a collection of 169,484 genomes spanning the prokaryotic tree of life and including non-cultured species coming from diverse sequencing efforts: the Genomic Catalog of Earth's Microbiomes (GEM)[71], the Global Microbial Gene Catalog (GMGC)[72], the Unified Human Gastrointestinal Genome collection (UHGG)[73], and the Ocean Microbiomics Database (OMD)[74]. We then expanded each protein family with the hits from this database. If, after expanding, a family incorporated genes with homologs in EggNOG, that family was then discarded from the novel family set. We predicted signal peptides and transmembrane domains on the gene families using SignalP[75] and TMHMM[76]. Protein families were considered as transmembrane or exported if >80% of their members had a predicted transmembrane domain or a signal peptide, respectively.

Phylogenomic analyses

We collected the proteomes of 192 taxa spanning all major archaeal super-groups (including the new Afararchaeaceae and Asbonarchaeaceae). We reconstructed two phylogenomic datasets consisting of 48 ribosomal proteins (RP) and 136 new markers (NM) widely distributed in archaea (Supplementary Fig. 39). The 136 NM dataset was based on curating a set of 200 markers previously shown to be highly conserved across the archaeal domain[18]. To ensure standardized protein-coding gene predictions, all 192 genomes were first run through Prodigal[64]. Next, sequences similar to the RP and NM proteins were identified using BLAST[77] with relatively relaxed criteria (>20% sequence identity over 30% query length) to retrieve even divergent homologs, such as those found in fast-evolving lineages like the DPANN archaea. For each of the 192 taxa, up to five of the best BLAST hit sequences were kept and included in a single file for phylogenetic reconstruction. Preliminary trees inferred with FastTree2[78] were manually examined to identify the correct orthologue for each taxon and to detect cases of contamination, HGT, or paralogy. These spurious sequences were removed and the remaining ones used for reconstruction of a new tree. Multiple rounds of manual curation were done in this way until all problematic sequences were removed. Once curated, each orthologous group was aligned with MAFFT L-INS-i v7.450[79] and trimmed with BMGE v1.12[80] (-m BLOSUM30 -b 3 -g 0.2 -h 0.5). We performed a final round of verification of the single gene trees reconstructed using the more sophisticated LG+C60+F+Γ4 model in IQ-TREE before concatenating the individually trimmed alignments into two supermatrices (RP and NM). The 192-RP and 192-NM alignments were then subsampled to generate two additional alignments consisting of 87 taxa containing only Euryarchaeota (87-RP and 87-NM) and 104 taxa, including the 87 Euryarchaeota plus 8 Nanosalinaceae and 9 Asbonarchaeaceae (104-RP and 104-NM). These six alignments were then used for maximum likelihood (ML) phylogenetic reconstruction under the LG+C60+F+Γ4 sequence evolution model (with 1,000 ultra-fast bootstrap replicates) using IQTREE v2.0.3[81]. For four of the six alignments (87-RP, 104-RP, 87-NM, and 104-NM), Bayesian phylogenetic reconstructions were also run using the CAT+GTR model as implemented in PhyloBayes v1.8[82]. Four MCMC chains were run in parallel for each alignment. Although convergence was not reached after 8 months of calculation, a sufficient effective sample size was reached (effsize >300) while using a burnin of 3,000 cycles and sampling every 50 generations after the burn-in.

Amino acid composition analysis

We used an in-house Python script (https://github.com/bbaker567/phylogenetics) to estimate the frequency of each amino acid in our selection of 192 archaeal taxa for the whole predicted proteomes, as well as for the RP and NM datasets. These frequencies were analyzed using principal component analysis with ggplot2[83].

In addition, for each amino acid, the compositional bias between halophiles and non-halophiles was measured for the RP and NM datasets with the Z-score from a binomial test of two proportions:

$$Z = \frac{p1 - p2}{\sqrt{p0(1-p0)(\frac{1}{n1} + \frac{1}{n2})}}$$

$$p1 = \frac{X1}{n1}, p2 = \frac{X2}{n2}, p0 = \frac{X1 + X2}{n1 + n2}$$

where X1 and X2 are the total numbers of that amino acid, and n1 and n2 are the total numbers of all 20 amino acids across halophiles and non-halophiles, respectively. Calculating Z-scores in this way assumes that the proportions of an amino acid across halophiles and non-halophiles are approximately normal, with the null hypothesis that p1=p2. |Z| >1.96 indicates rejection of the null hypothesis at a significance level of $p < 0.05$. Amino acids with |Z| >1.96 were considered significantly enriched in halophiles relative to non-halophiles, whereas amino acids with |Z| <-1.96 were considered significantly depleted in halophiles relative to non-halophiles. Amino acids were divided into 'Over-represented' (|Z| >1.96), 'Under-represented' (|Z| <-1.96), and 'Not significant' (|Z| not statistically significant).

We also implemented the new GFmix-DE/IK model by transforming the b parameter of the GFmix model[45] (originally designed to represent the ratio of GARP/FYMINK amino acids across all descendant taxa at each branch in a tree) to accommodate amino acid groupings other than GARP/FYMINK, in our case those identified to be biased in extreme halophiles. We then calculated the likelihood of different tree topologies under these variants of the GFmix model with LG+C60+F+Γ4[45]. Branch length and alpha shape parameters for each tree tested were estimated using IQTREE v2.0.3[81] and then fed into GFmix, specifying the custom enriched and depleted amino acid bins for halophiles versus non-halophiles. We used this approach to calculate the likelihood of four different tree topologies: i) Nanosalinaceae+Asbonarchaeaceae within DPANN and Methanonatronarchaeia sister to the AHH-clade; ii) Nanosalinaceae+Asbonarchaeaceae within DPANN and Methanonatronarchaeia deep within Euryarchaeota; iii) monophyly of the AHH-clade, Methanonatronarchaeia, and Nanosalinaceae+Asbonarchaeaceae, with Methanonatronarchaeia as the deepest branch, and iv) monophyly of the AHH-clade, Methanonatronarchaeia, and Nanosalinaceae+Asbonarchaeaceae, with Nanosalinaceae+Asbonarchaeaceae as the deepest branch (Extended Data Fig. 5).

Progressive removal of compositionally biased sites

To remove the most compositionally biased sites from the sequence datasets, we split the sequence alignments in two based on whether the taxa were classified as extreme halophiles or non-halophiles. We then calculated the ratio of D+E divided by I+K for each alignment site for both the halophiles and non-halophiles sub-alignments. We then divided the D+E/I+K ratio for each halophile sub-alignment site by the corresponding ratio in the non-halophile sub-alignment. When the denominator of one of the ratios was equal to zero, we substituted '0' for '0.1' in order to still consider the alignment position. Alignment sites were then ranked from the highest to the lowest ratio, using the highest ratio as a proxy for the most biased alignment site. Next, we progressively removed alignment sites in increments of 1%, 5%, 10%, 20%, 30%, and up to 90%. This resulted in 11 alignments for both the RP and NM datasets. These 11 alignments were then used for ML phylogenetic reconstruction under the LG+C60+F+Γ4 model (with 1,000 ultra-fast bootstraps).

In the case of ribosomal proteins, we mapped the acidic amino acid positions on the large ribosomal subunit structures of the extremely halophilic haloarchaeon *Haloarcula marismortui* (PDB accession number 1S72[84]) and the non-halophilic methanogen *Methanothermobacter thermautotrophicus* (PDB accession number 4ADX[85]). We located these positions on their respective structures using ChimeraX[86], which was also used to produce a video showing them (Supplementary Video 1).

Orthologous groups and single-gene trees

Orthologous groups (OGs) were identified for all the proteins of the species included in the 192 taxa dataset using OrthoFinder v2.5.1[87] with Diamond BLAST (--ultra-sensitive, --query-cover 50%, and --id 30%) and an inflation parameter of 1.1. This resulted in 17,827 OGs, which were aligned using MAFFT --auto v7.450[79] with default settings and trimmed using trimAl[88] (-automated1 -resoverlap 0.75 -seqoverlap 75). To avoid poorly resolved single gene trees due to little phylogenetic information, we removed OGs that presented a trimmed alignment length of less than 60 amino acids. This resulted in 17,288 OGs, which were used to reconstruct individual trees with IQTREE v2.0.3[81]. For computational time reasons, the trees of the 200 OGs containing the largest number of sequences were inferred under the LG+C20+F+Γ4 model of sequence evolution, while the remaining phylogenies were run under LG+C60+F+Γ4. Statistical support at branches was estimated using 1,000 ultrafast bootstrap replicates. Finally, for OGs containing only two or three sequences, "bootstrap" samples were artificially generated for subsequent analysis in ALE[89], corresponding to the single possible unrooted tree topology.

Gene tree-aware ancestral gene content reconstruction

The 17,288 single-gene trees were reconciled with the species tree inferred from the 192-NM dataset using the ALEml_undated algorithm of the ALE suite v0.4[89]. ALE infers, for each gene family, duplications, losses, transfers, and originations events along a species tree[89]. The raw relative reconciliation frequencies outputted by ALE were summed for all events. These relative frequency values support an evolutionary event occurring at a given node by incorporating the uncertainty of the reconstructed individual gene tree, as represented by the bootstrap replicates. A few gene families were manually selected based on their patterns of presence/absence and/or HGTs in halophilic groups. The presence probability for the various nodes of interest for each of these gene families mentioned in the text can be found in Supplementary Data 9. ALE also predicts the ancestral copy number for each node in the species tree. Phylogenetic trees were visualized using Figtree v.1.4.4 (http://tree.bio.ed.ac.uk/software/figtree), iTOL[90], and the ETE3 Toolkit v.3.1.2[91].

To detect possible genes of bacterial origin in halophilic archaea, we carried out Blast[77] searches of the proteins considered by ALE as 'originations' in these archaea against the RefSeq[34] database. Proteins with similar sequences in bacteria were aligned using MAFFT --auto v7.450[79] with default settings and trimmed using trimAl[88] (-automated1). Maximum likelihood trees were then reconstructed with IQTREE v2.0.3[61] under the LG+C60+F+Γ4 model of sequence evolution. Statistical support at branches was estimated using 1,000 ultrafast bootstrap replicates. Phylogenetic trees were visualized using Figtree v.1.4.4 (http://tree.bio.ed.ac.uk/software/figtree).

Data availability

The MAGs reported in this study have been deposited in GenBank under BioProject number PRJNA901412. All raw data underlying phylogenomic analyses (raw and processed alignments and corresponding phylogenetic trees) and all predicted proteomes have been deposited into Figshare (https://figshare.com/s/353259800b42a4e190eb).

Code availability

Custom code used for data analysis is available at GitHub: (https://github.com/bbaker567/phylogenetics).

References

1. Oren, A. Diversity of halophilic microorganisms: Environments, phylogeny, physiology, and applications. *J. Ind. Microbiol. Biotechnol.* **28**, 56–63 (2002).
2. Oren, A. Molecular ecology of extremely halophilic Archaea and Bacteria. *FEMS Microbiol. Ecol.* **39**, 1–7 (2002).
3. Narasingarao, P. et al. De novo metagenomic assembly reveals abundant novel major lineage of Archaea in hypersaline microbial communities. *ISME J.* **6**, 81–93 (2012).
4. Ghai, R. et al. New abundant microbial groups in aquatic hypersaline environments. *Sci. Rep.* **1**, 135 (2011).
5. Zhao, D. et al. Comparative genomic insights into the evolution of Halobacteria-associated "*Candidatus* Nanohaloarchaeota". *mSystems* **7**, e0066922 (2022).
6. Rinke, C. et al. Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**, 431–437 (2013).
7. Sorokin, D. Y. et al. Discovery of extremely halophilic, methyl-reducing euryarchaea provides insights into the evolutionary origin of methanogenesis. *Nat. Microbiol.* **2**, 17081 (2017).
8. Aouad, M., Borrel, G., Brochier-Armanet, C. & Gribaldo, S. Evolutionary placement of Methanonatronarchaeia. *Nat. Microbiol.* **4**, 558–559 (2019).
9. Feng, Y. et al. The evolutionary origins of extreme halophilic archaeal lineages. *Genome Biol. Evol.* **13**, evab166 (2021).
10. Martijn, J. et al. Hikarchaeia demonstrate an intermediate stage in the methanogen-to-halophile transition. *Nat. Commun.* **11**, 5490 (2020).
11. Sorokin, D. Y. et al. Reply to 'Evolutionary placement of Methanonatronarchaeia'. *Nat. Microbiol.* **4**, 560–561 (2019).
12. Zhou, H. et al. Metagenomic insights into the environmental adaptation and metabolism of *Candidatus* Haloplasmatales, one archaeal order thriving in saline lakes. *Environ. Microbiol.* **24**, 2239–2258 (2022).
13. Oren, A. Microbial life at high salt concentrations: phylogenetic and metabolic diversity. *Saline Syst.* **4**, 2 (2008).
14. Fukuchi, S., Yoshimune, K., Wakayama, M., Moriguchi, M. & Nishikawa, K. Unique amino acid composition of proteins in halophilic bacteria. *J. Mol. Biol.* **327**, 347–357 (2003).
15. Lanyi, J. K. Salt-dependent properties of proteins from extremely halophilic bacteria. *Bacteriol. Rev.* **38**, 272–290 (1974).
16. Madern, D., Ebel, C. & Zaccai, G. Halophilic adaptation of enzymes. *Extremophiles* **4**, 91–98 (2000).
17. Tadeo, X. et al. Structural basis for the aminoacid composition of proteins from halophilic archea. *PLOS Biol.* **7**, e1000257 (2009).
18. Petitjean, C., Deschamps, P., López-García, P., Moreira, D. & Brochier-Armanet, C. Extending the conserved phylogenetic core of Archaea disentangles the evolution of the third Domain of Life. *Mol. Biol. Evol.* **32**, 1242–1254 (2015).
19. Dombrowski, N., Lee, J.-H., Williams, T. A., Offre, P. & Spang, A. Genomic diversity, lifestyles and evolutionary origins of DPANN archaea. *FEMS Microbiol. Lett.* **366**, (2019).
20. Belilla, J. et al. Archaeal overdominance close to life-limiting conditions in geothermally influenced hypersaline lakes at the Danakil Depression, Ethiopia. *Environ. Microbiol.* **23**, 7168–7182 (2021).

21. Belilla, J. et al. Hyperdiverse archaea near life limits at the polyextreme geothermal Dallol area. *Nat. Ecol. Evol.* **3**, 1552–1561 (2019).
22. Gutiérrez-Preciado A., Moreira D., Baker B., Eme L., Deschamps P., López-García P. Extremely acidic proteomes and diversification of archaea convergently adapted to increasingly chaotropic brines. (In prep.).
23. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **44**, D457–D462 (2016).
24. Falb, M. et al*.* Metabolism of halophilic archaea. *Extremophiles* **12**, 177–196 (2008).
25. Albers, S.-V. & Jarrell, K. F. The archaellum: how Archaea swim. *Front. Microbiol.* **6**, (2015).
26. Sasaki, J. & Spudich, J. L. Signal transfer in haloarchaeal sensory rhodopsin– transducer complexes. *Photochem. Photobiol.* **84**, 863–868 (2008).
27. Dassarma, S. et al*.* Genomic perspective on the photobiology of *Halobacterium* species NRC-1, a phototrophic, phototactic, and UV-tolerant haloarchaeon. *Photosynth. Res.* **70**, 3–17 (2001).
28. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* **36**, 1925–1927 (2020).
29. Dombrowski, N., Lee, J.-H., Williams, T. A., Offre, P. & Spang, A. Genomic diversity, lifestyles and evolutionary origins of DPANN archaea. *FEMS Microbiol. Lett.* **366**, (2019).
30. Chklovski, A., Parks, D. H., Woodcroft, B. J. & Tyson, G. W. CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. *Nat. Methods* **20**, 1203–1212 (2023).
31. Castelle, C. J. et al*.* Biosynthetic capacity, metabolic variety and unusual biology in the CPR and DPANN radiations. *Nat. Rev. Microbiol.* **16**, 629–645 (2018).
32. Hamm, J. N. *et al.* Unexpected host dependency of Antarctic Nanohaloarchaeota. *Proc. Natl. Acad. Sci. USA* **116**, 14661–14670 (2019).
33. La Cono, V. *et al.* Symbiosis between nanohaloarchaeon and haloarchaeon is based on utilization of different polysaccharides. *Proc. Natl. Acad. Sci. USA* **117**, 20223–20234 (2020).
34. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **35**, D61–D65 (2007).
35. Mistry, J. et al*.* Pfam: The protein families database in 2021. *Nucleic Acids Res.* **49**, D412–D419 (2021).
36. Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: Functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol. Biol. Evol.* **38**, 5825–5829 (2021).
37. Rodríguez del Río, Á. et al*.* Functional and evolutionary significance of unknown genes from uncultivated taxa. Preprint at https://doi.org/10.1101/2022.01.26.477801 (2022).
38. Cabello-Yeves, P. J. & Rodriguez-Valera, F. Marine-freshwater prokaryotic transitions require extensive changes in the predicted proteome. *Microbiome* **7**, 117 (2019).
39. Rasmussen, T. How do mechanosensitive channels sense membrane tension? *Biochem. Soc. Trans.* **44**, 1019–1025 (2016).
40. Petitjean, C., Deschamps, P., López-García, P. & Moreira, D. Rooting the Domain Archaea by phylogenomic analysis supports the foundation of the new Kingdom Proteoarchaeota. *Genome Biol. Evol.* **7**, 191–204 (2015).

41. Eme, L. et al. Inference and reconstruction of the heimdallarchaeial ancestry of eukaryotes. *Nature* **618**, 992–999 (2023).
42. Bergsten, J. A review of long-branch attraction. *Cladistics* **21**, 163–193 (2005).
43. Susko, E. & Roger, A. J. Long branch attraction biases in phylogenetics. *Syst. Biol.* **70**, 838–843 (2021).
44. Shimodaira, H. An approximately unbiased test of phylogenetic tree selection. *Syst. Biol.* **51**, 492–508 (2002).
45. Muñoz-Gómez, S. A. et al. Site-and-branch-heterogeneous analyses of an expanded dataset favour mitochondria as sister to known Alphaproteobacteria. *Nat. Ecol. Evol.* **6**, 253–262 (2022).
46. Aouad, M. et al. Extreme halophilic archaea derive from two distinct methanogen Class II lineages. *Mol. Phylogenet. Evol.* **127**, 46–54 (2018).
47. Mahendrarajah, T. A. et al. ATP synthase evolution on a cross-braced dated tree of life. Preprint at https://doi.org/10.1101/2023.04.11.536006 (2023).
48. Kellner, S. et al. Genome size evolution in the Archaea. *Emerg. Top. Life Sci.* ETLS20180021 (2018) doi:10.1042/ETLS20180021.
49. Brehmer, D., Gässler, C., Rist, W., Mayer, M. P. & Bukau, B. Influence of GrpE on DnaK-substrate interactions. *J. Biol. Chem.* **279**, 27957–27964 (2004).
50. Williams, T. A. et al. Integrative modeling of gene and genome evolution roots the archaeal tree of life. *Proc. Natl. Acad. Sci. USA.* **114**, E4602–E4611 (2017).
51. Giovannoni, S. J. *et al.* Genome streamlining in a cosmopolitan oceanic bacterium. *Science* **309**, 1242–1245 (2005).
52. Swan, B. K. et al. Prevalent genome streamlining and latitudinal divergence of planktonic bacteria in the surface ocean. *Proc. Natl. Acad. Sci. USA.* **110**, (2013).
53. Martin-Cuadrado, A.-B., Ghai, R., Gonzaga, A. & Rodriguez-Valera, F. CO Dehydrogenase genes found in metagenomic fosmid clones from the deep Mediterranean Sea. *Appl. Environ. Microbiol.* **75**, 7436–7444 (2009).
54. Becker, E. A. et al. Phylogenetically driven sequencing of extremely halophilic archaea reveals strategies for static and dynamic osmo-response. *PLOS Genet.* **10**, e1004784 (2014).
55. Groussin, M. et al. Gene acquisitions from Bacteria at the origins of major archaeal clades are vastly overestimated. *Mol. Biol. Evol.* **33**, 305–310 (2016).
56. Nelson-Sathi, S. et al. Acquisition of 1,000 eubacterial genes physiologically transformed a methanogen at the origin of Haloarchaea. *Proc. Natl. Acad. Sci. USA.* **109**, 20537–20542 (2012).
57. Nelson-Sathi, S. et al. Origins of major archaeal clades correspond to gene acquisitions from bacteria. *Nature* **517**, 77–80 (2015).
58. Gadda, G. & McAllister-Wilkins, E. E. Cloning, expression, and purification of choline dehydrogenase from the moderate halophile *Halomonas elongata*. *Appl. Environ. Microbiol.* **69**, 2126–2132 (2003).
59. Deschamps, P., Zivanovic, Y., Moreira, D., Rodriguez-Valera, F. & López-García, P. Pangenome evidence for extensive interdomain horizontal transfer affecting lineage core and shell Genes in uncultured planktonic Thaumarchaeota and Euryarchaeota. *Genome Biol. Evol.* **6**, 1549–1563 (2014).
60. Hedlund, B.P., Chuvochina, M., Hugenholtz, P., Konstantinidis, K.T., Murray, A.E., Palmer, M., Parks, D.H., Probst, A.J., Reysenbach, A.L., Rodriguez-R, L.M., Rossello-Mora, R.,

754 Sutcliffe, I.C., Venter, S.N. & Whitman, W.B. SeqCode: a nomenclatural code for
755 prokaryotes described from sequence data. *Nat Microbiol*. 7, 1702-1708 (2022).
756 61. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: A fast and
757 effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol.*
758 *Evol.* **32**, 268–274 (2015).
759 62. Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High
760 throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries.
761 *Nat. Commun.* **9**, 1–8 (2018).
762 63. Rodriguez-R, L. M. & Konstantinidis, K. T. Bypassing cultivation to identify bacterial
763 species: Culture-independent genomic approaches identify credibly distinct clusters,
764 avoid cultivation bias, and provide true insights into microbial species. *Microbe Mag.* **9**,
765 111–118 (2014).
766 64. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site
767 identification. *BMC Bioinformatics* **11**, 119 (2010).
768 65. Tatusov, R. L., Galperin, M. Y., Natale, D. A. & Koonin, E. V. The COG database: a tool for
769 genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* **28**, 33–36
770 (2000).
771 66. Eren, A. M. et al. Anvi'o: an advanced analysis and visualization platform for 'omics data.
772 *PeerJ* **3**, e1319 (2015).
773 67. Aramaki, T. et al. KofamKOALA: KEGG Ortholog assignment based on profile HMM and
774 adaptive score threshold. *Bioinformatics* **36**, 2251–2252 (2020).
775 68. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for
776 the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).
777 69. Eddy, S. R. Accelerated Profile HMM Searches. *PLOS Comput. Biol.* **7**, e1002195 (2011).
778 70. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND.
779 *Nat. Methods* **12**, 59–60 (2015).
780 71. Nayfach, S. et al. A genomic catalog of Earth's microbiomes. *Nat. Biotechnol.* **39**, 499–
781 509 (2021).
782 72. Coelho, L. P. et al. Towards the biogeography of prokaryotic genes. *Nature* **601**, 252–256
783 (2022).
784 73. Almeida, A. et al. A unified catalog of 204,938 reference genomes from the human gut
785 microbiome. *Nat. Biotechnol.* **39**, 105–114 (2021).
786 74. Paoli, L. et al. Uncharted biosynthetic potential of the ocean microbiome. Preprint at
787 https://doi.org/10.1101/2021.03.24.436479 (2021).
788 75. Almagro Armenteros, J. J. et al. SignalP 5.0 improves signal peptide predictions using
789 deep neural networks. *Nat. Biotechnol.* **37**, 420–423 (2019).
790 76. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. L. Predicting transmembrane
791 protein topology with a hidden markov model: application to complete genomes. *J. Mol.*
792 *Biol.* **305**, 567–580 (2001).
793 77. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment
794 search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
795 78. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 – approximately maximum-likelihood
796 trees for large alignments. *PLOS ONE* **5**, e9490 (2010).
797 79. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7:
798 improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).

80. 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.* **10**, 210 (2010).
81. Minh, B. Q. *et al.* IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
82. Lartillot, N. PhyloBayes: Bayesian phylogenetics using site-heterogeneous models. in *Phylogenetics in the Genomic Era* (eds. Scornavacca, C., Delsuc, F. & Galtier, N.) 1.5:1-1.5:16 (No commercial publisher | Authors open access book, 2020).
83. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. (Springer, 2009).
84. Klein, D. J., Moore, P. B. & Steitz, T. A. The roles of ribosomal proteins in the structure assembly, and evolution of the large ribosomal subunit. *J. Mol. Biol.* **340**, 141–177 (2004).
85. Greber, B. J. et al. Cryo-EM structure of the archaeal 50S ribosomal subunit in complex with initiation factor 6 and implications for ribosome evolution. *J. Mol. Biol.* **418**, 145–160 (2012).
86. Pettersen, E. F. et al. UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Sci. Publ. Protein Soc.* **30**, 70–82 (2021).
87. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
88. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
89. Szöllősi, G. J., Rosikiewicz, W., Boussau, B., Tannier, E. & Daubin, V. Efficient exploration of the space of reconciled gene trees. *Syst. Biol.* **62**, 901–912 (2013).
90. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* **49**, W293–W296 (2021).
91. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: Reconstruction, analysis, and visualization of phylogenomic data. *Mol. Biol. Evol.* **33**, 1635–1638 (2016).

Acknowledgments

D.M. and L.E were supported by grants from the European Research Council (ERC Advanced grant 787904 and ERC Starting grant 803151, respectively). This work was also supported by the Moore-Simons Project Call on the Origin of the Eukaryotic Cell, Simons Foundation 812811 (A.J.R, E.S., and L.E.), and Moore Foundation GBMF9739 (P.L.G.). A.R.R. was supported by "la Caixa" Foundation (ID 100010434, fellowship code LCF/BQ/DI18/11660009, the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 713673) and by an EMBO Scientific Exchange Grant. We thank P. Deschamps for help in managing our bioinformatic cluster, A. Oren for his advice on taxonomic descriptions, and two anonymous reviewers for constructive suggestions. We are grateful to the Iris Foundation for the continuous support of our work on the microbial diversity of the Danakil Depression.

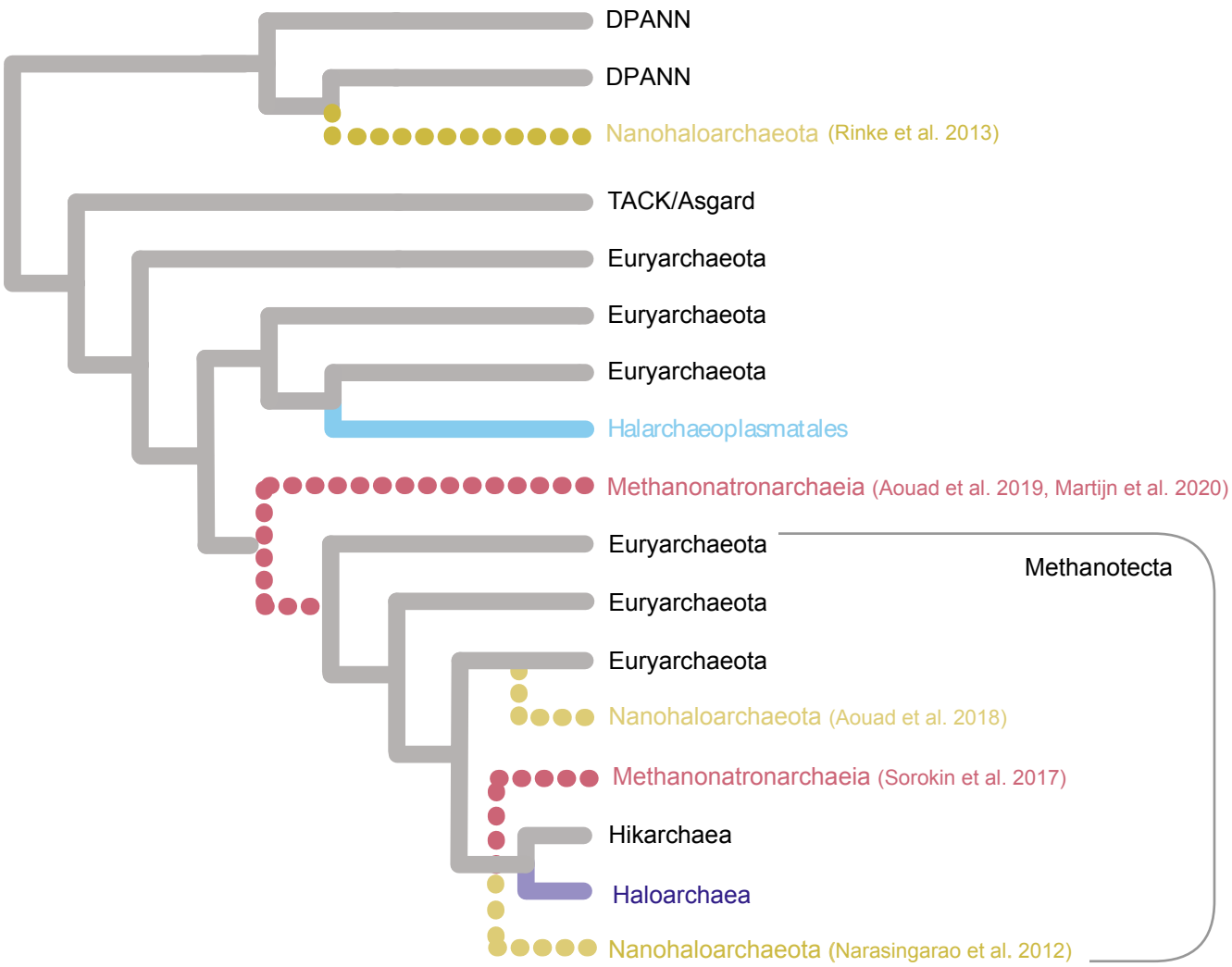
Author contributions

D.M., P.L.G., and L.E designed the study. A.G.P. and B.B. annotated the new archaeal MAGs. A.R.R., B.B., and J.H.C. studied the new protein families. C.G.P.MC., A.J.R., and E.S. conceived of the binomial methods to identify significant shifts in amino acid composition, and E.S. implemented the new features of the GFmix model in the GFmix software. B.B., L.E., D.M.,

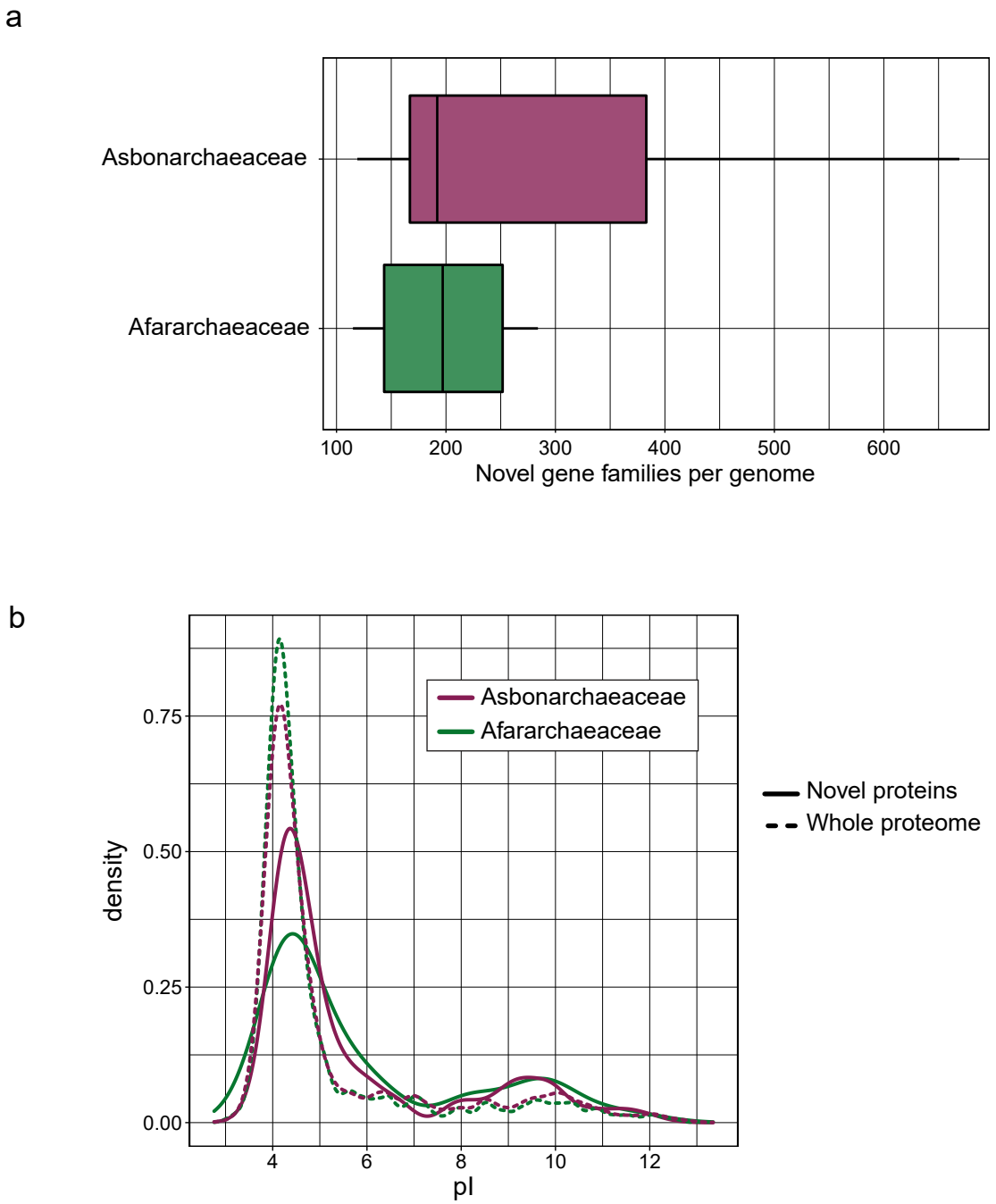
C.G.P.M., A.J.R., and E.S. carried out phylogenetic analyses. B.B., L.E., P.L.G., and D.M. wrote the paper with contributions from all authors.

Competing interests

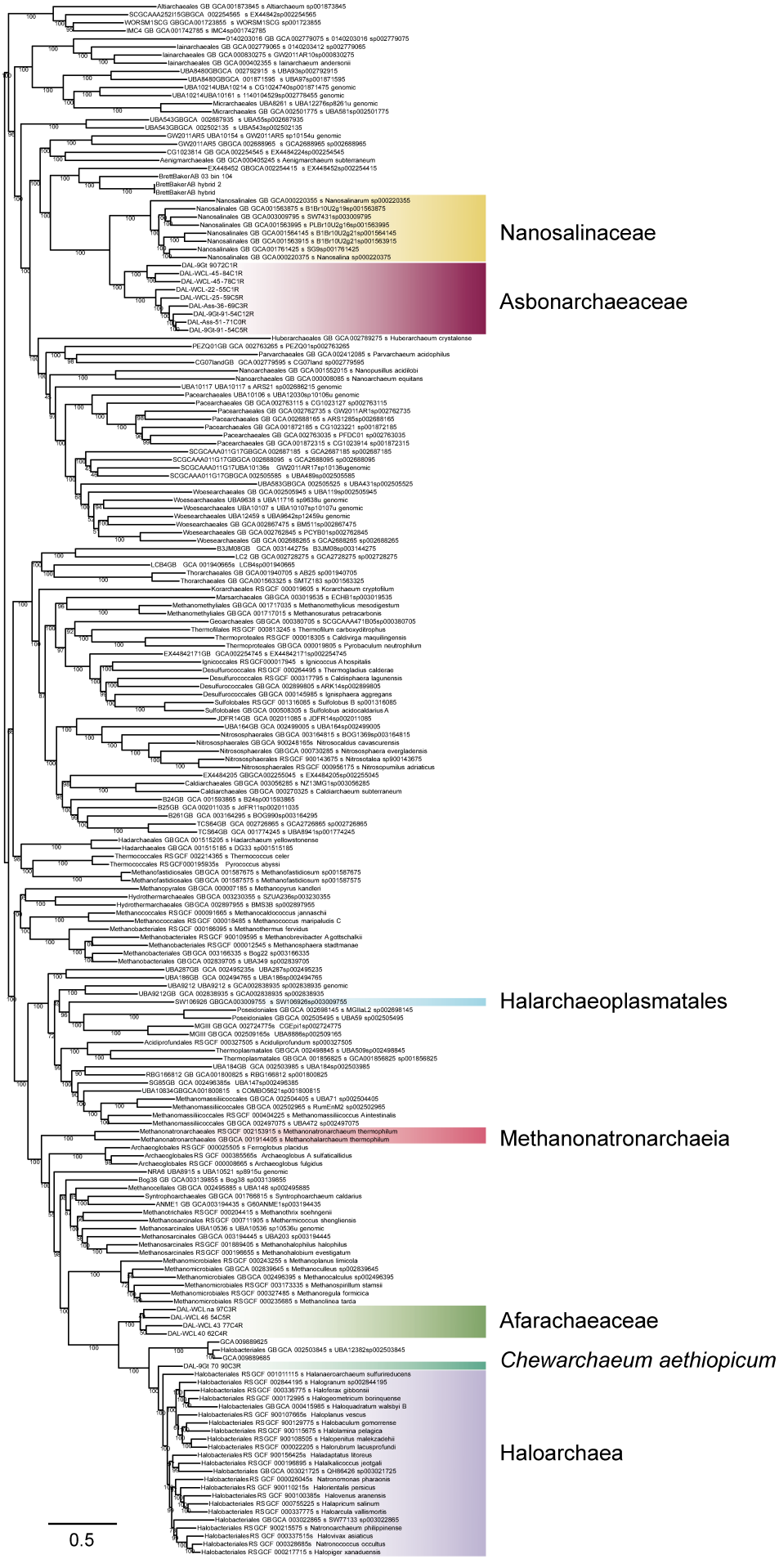
The authors declare no competing interests.



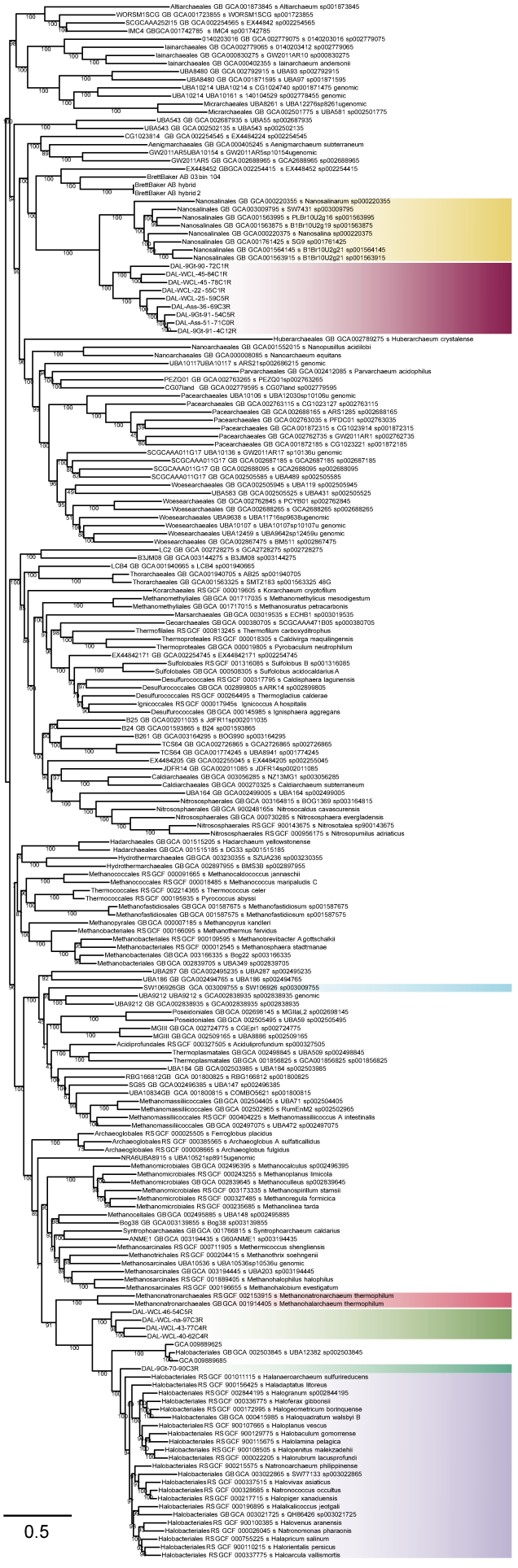
Extended Data Fig. 1 | Schematic tree showing the phylogenetic position of extremely halophilic archaeal groups (colored branches) proposed in previous articles. Branches that have been found at different places in the tree of archaea are indicated with dashed lines (Narasingarao et al. 2012, Rinke et al. 2013, Sorokin et al. 2017, Aouad et al. 2018, Aouad et al. 2019, Martijn et al. 2020).



Extended Data Fig. 2 | Number and isoelectric point of novel gene families identified in the Asbonarchaeaceae and Afararchaeaceae MAGs. (a) The average number of novel genes in the nine asbonarchaeal and four afarachaeal MAGs described in this study (see Methods). **(b)** The isoelectric point of these novel proteins (solid lines) compared to the average isoelectric point of the whole proteomes (dashed lines).

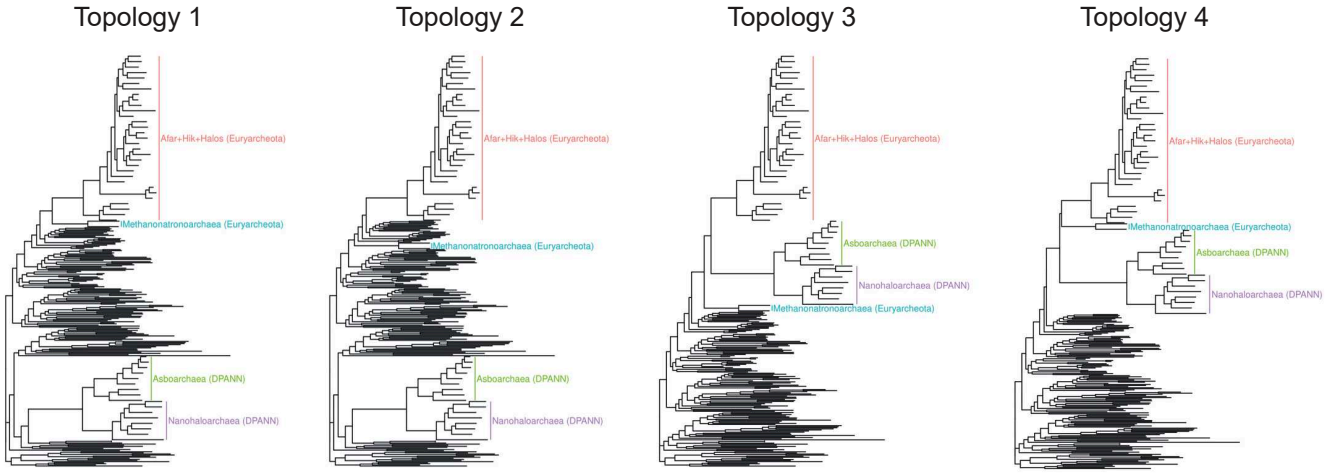


Extended Data Fig. 3 | Maximum likelihood phylogeny of 192 archaea based on the NM dataset. The ML tree was inferred with the LG+C60+F+Γ4 model of sequence evolution with 1,000 ultrafast bootstraps as implemented via IQ-TREE. The scale bar indicates the expected average number of substitutions per site. Extremely halophilic archaea are indicated in color.



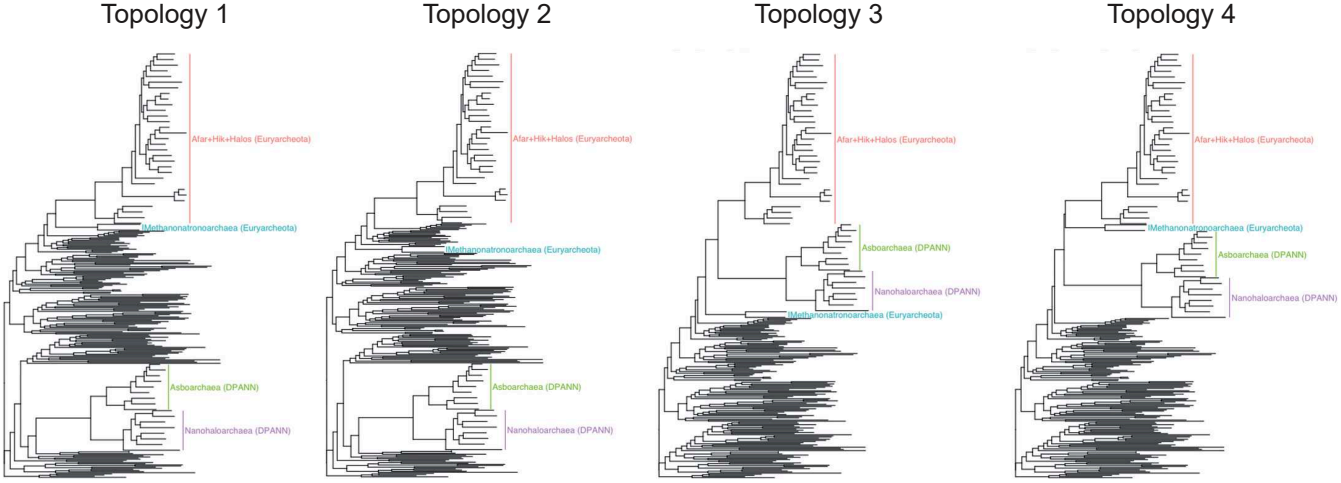
Extended Data Fig. 4 | Maximum likelihood phylogeny of 192 archaea based on the RP dataset. The ML tree was inferred with the LG+C60+F+Γ4 model of sequence evolution with 1,000 ultrafast bootstraps as implemented via IQ-TREE. The scale bar indicates the expected average number of substitutions per site. Extremely halophilic archaea are indicated in color.

a



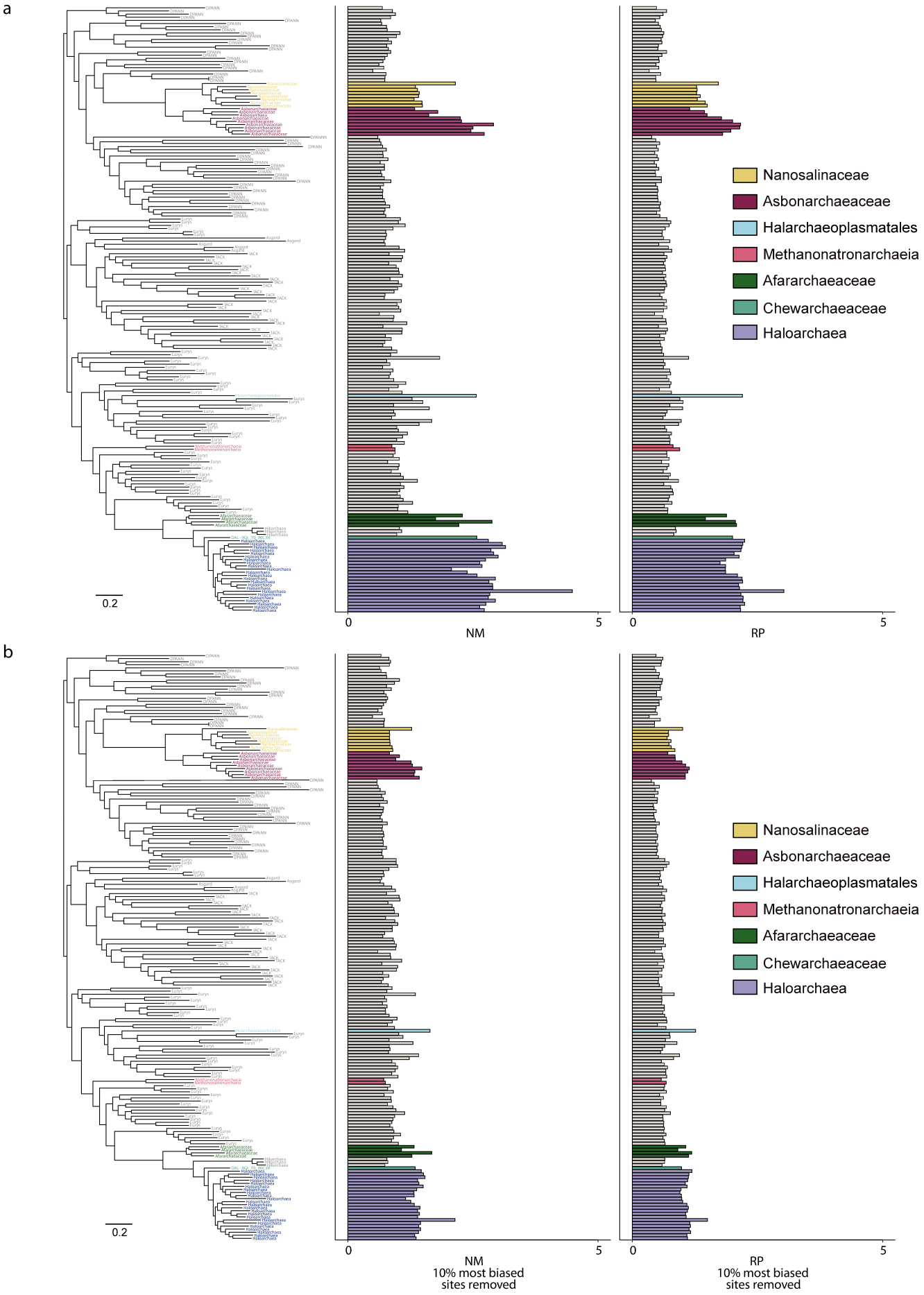
192-taxa RP dataset						
Model	**Bin 1**	**Bin 2**	Topology 1	**Topology 2**	**Topology 3**	**Topology 4**
LG+C60+G+F	None	None	Likelihood=-1317976,09 (Total di ff.=0)	Likelihood= -1318026,107 (Total diff.= -50,078)	Likelihood= -1318569,945 (Total diff.= -593,916)	Likelihood= -1318599,398 (Total diff.= -623,369)
LG+C60+G+F+GFmix	DE	IK	Likelihood=-1314937,111 (Total di ff.=0)	Likelihood= -1314977,65 (Total diff.= -40,539)	Likelihood= -1315670,918 (Total diff.= -733,808)	Likelihood= -1315685,576 (Total diff.= -748,465)
LG+C60+G+F+GFmix	DEQTAVG	KILCMFYWS	Likelihood=-1315018,218 (Total di ff.=0)	Likelihood= -1315065,115 (Total diff.= -46,898)	Likelihood= -1315704,916 (Total diff.= --686,698)	Likelihood= -1315727,147 (Total diff.= -708,93)

b

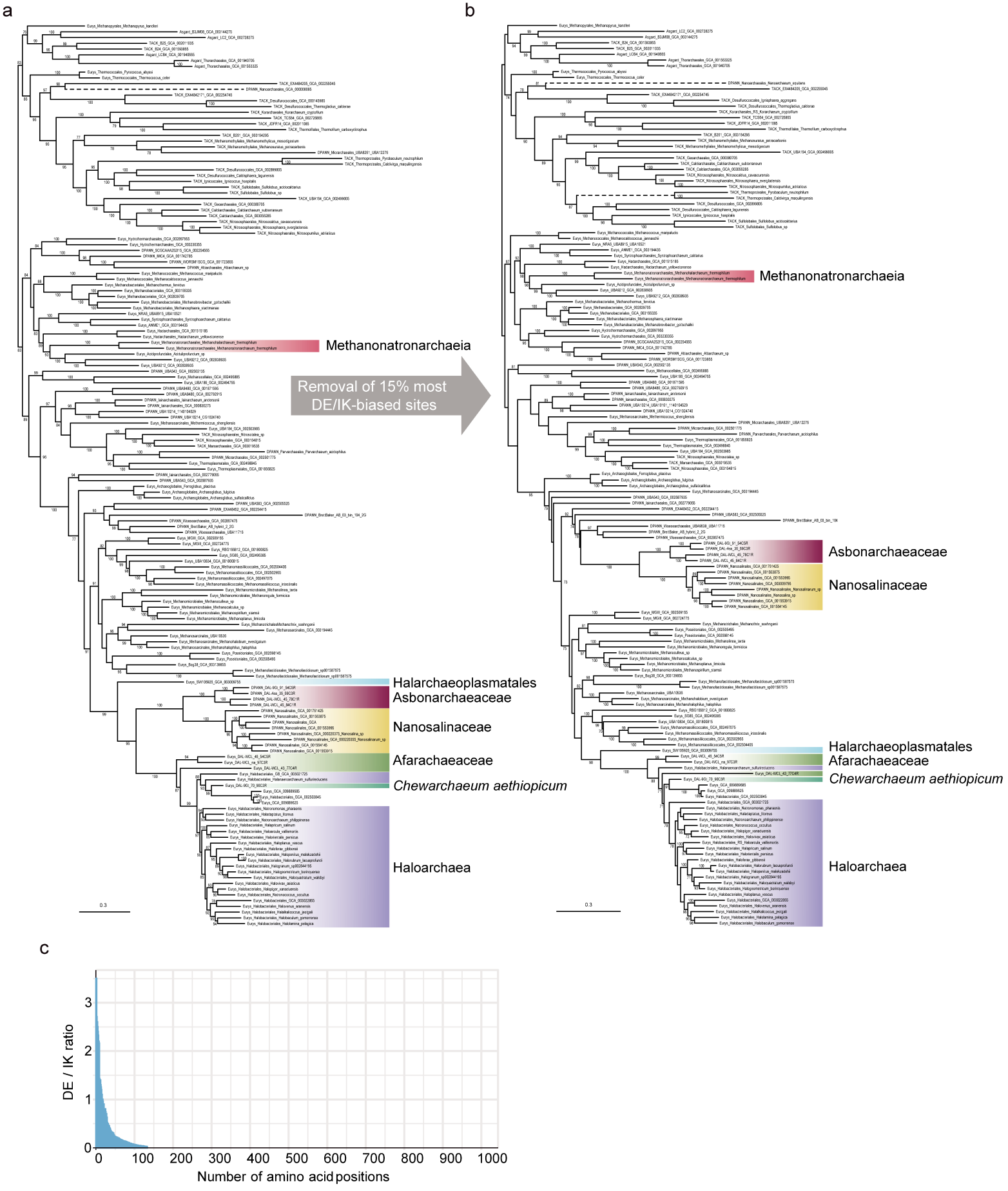


192-taxa NM dataset						
Model	**Bin 1**	**Bin 2**	**Topology 1**	Topology 2	**Topology 3**	**Topology 4**
LG+C60+G+F	None	None	Likelihood= -6557969,994 (Total diff.= -330,418)	Likelihood=-6557639,577 (Total di ff.=0)	Likelihood= -6561223,122 (Total diff.= -3583,545)	Likelihood= -6561620,115 (Total diff.= -3980,538)
LG+C60+G+F+GFmix	DE	IK	Likelihood= -6547076,576 (Total diff.= -364,896)	Likelihood=-6546711,68 (Total di ff.=0)	Likelihood= -6550656,783 (Total diff.= -3945,103)	Likelihood= -6550979,577 (Total diff.= -4267,898)
LG+C60+G+F+GFmix	DEAVTQGRHP	KINFSCLYM	Likelihood= -6541714,646 (Total diff.= -309,156)	Likelihood=-6541405,489 (Total di ff.=0)	Likelihood= -6545093,421 (Total diff.= -3687,931)	Likelihood= -6545459,741 (Total diff.= -4054,252)

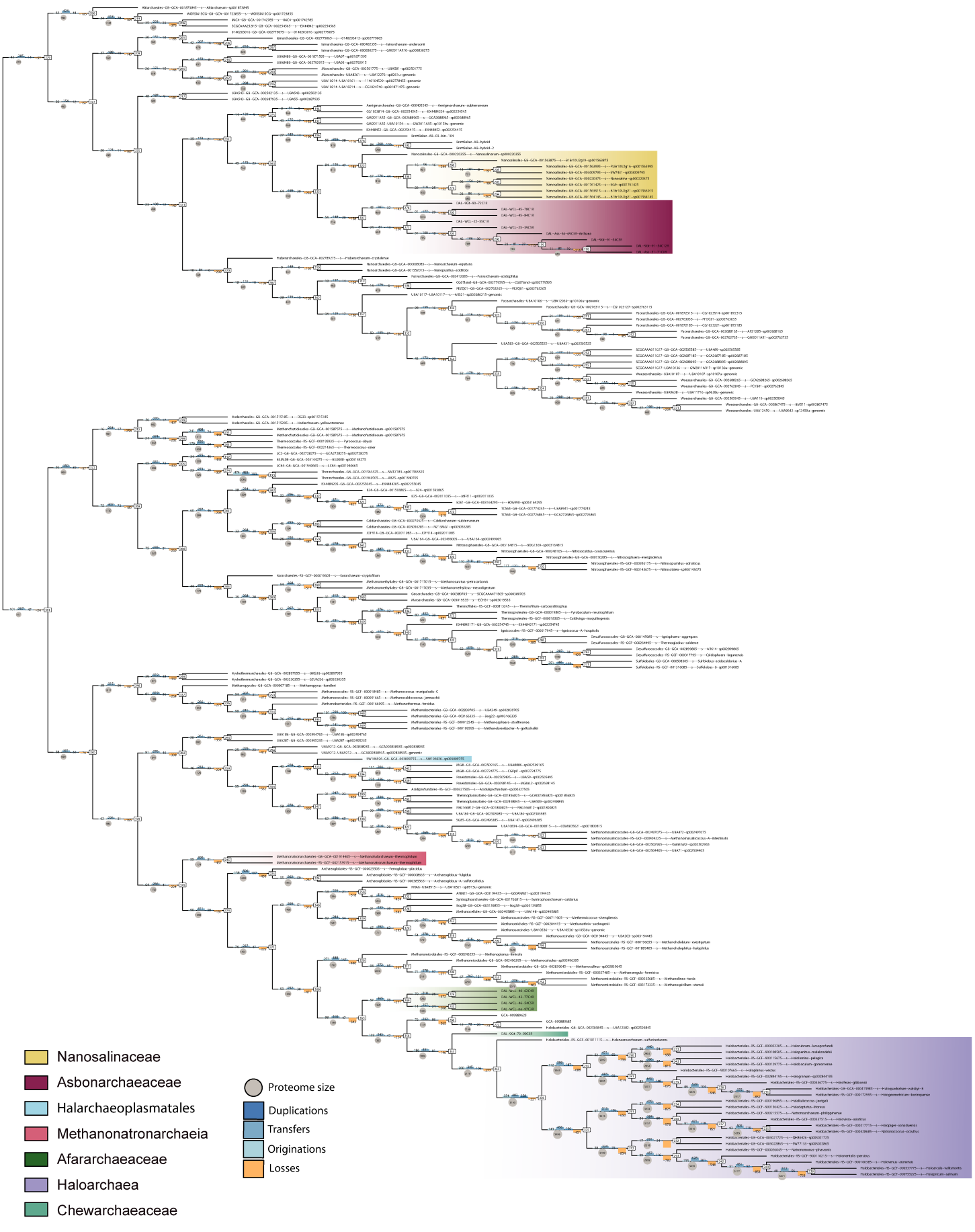
Extended Data Fig. 5 | Likelihood values for alternative positions of the extremely halophilic archaeal lineages. Likelihoods are calculated using IQ-TREE with the LG+C60+F+Γ4 model alone or combined with the new GFmix model (taking into account all significantly enriched (Bin 1) or depleted (Bin 2) amino acids in halophiles or only the most extremely biased ones (D+E and I+K). The highest-scoring topology is indicated with a red rectangle for the **(a)** RP and **(b)** NM datasets. Likelihood differences between a given topology and the highest-scoring topology per model are given in parentheses.



Extended Data Fig. 6 | Halophilic-specific amino acid compositional biases along the phylogeny of 192 archaeal taxa. **(a)** The ratio of [D+E/I+K] amino acids of 192 archaeal taxa was calculated along the untreated NM and RP alignments (39,385 and 6,792 amino acid positions, respectively). **(b)** 10% of the most biased sites (i.e., those with the highest ratio) were removed from the NM and RP alignments. Distinct halophilic clades are indicated in color, including the Nanosalinaceae (sand), Asbonarchaeaceae (wine), Halarchaeoplasmatales (cyan), Methanonatronarchaeia (rose), Afarachaeaceae (green), and Haloarchaea (indigo). The scale bar indicates the expected average number of substitutions per site.

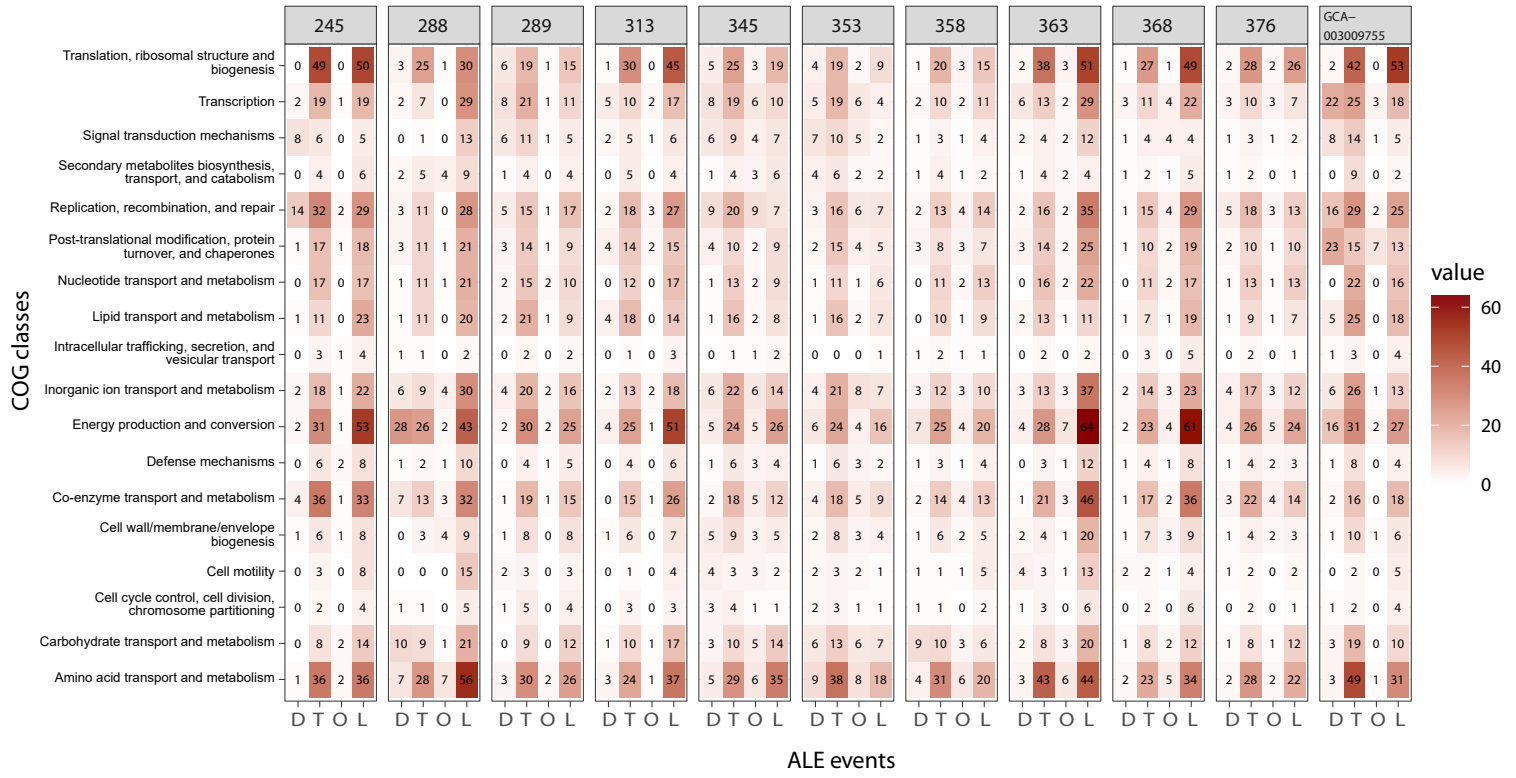


Extended Data Fig. 7 | Impact of compositional bias on the phylogeny of archaeal ATP synthase. Maximum likelihood phylogenetic trees based on the concatenation of ATP synthase subunits A and B **(a)** before and **(b)** after removal of 15% of sites with the highest D+E/I+K ratio. Notice the shift in the position of the Nanosalinaceae+Asbonarchaeaceae group. The trees were reconstructed using the LG+C60+F+Γ4 model of sequence evolution. Numbers at branches indicate 1,000 ultrafast bootstrap support values. Only values >70% are indicated. The scale bar indicates the expected average number of substitutions per site. **(c)** D+E/I+K ratio for all sites in the ATP synthase subunits A and B dataset ordered from highest to lowest values.

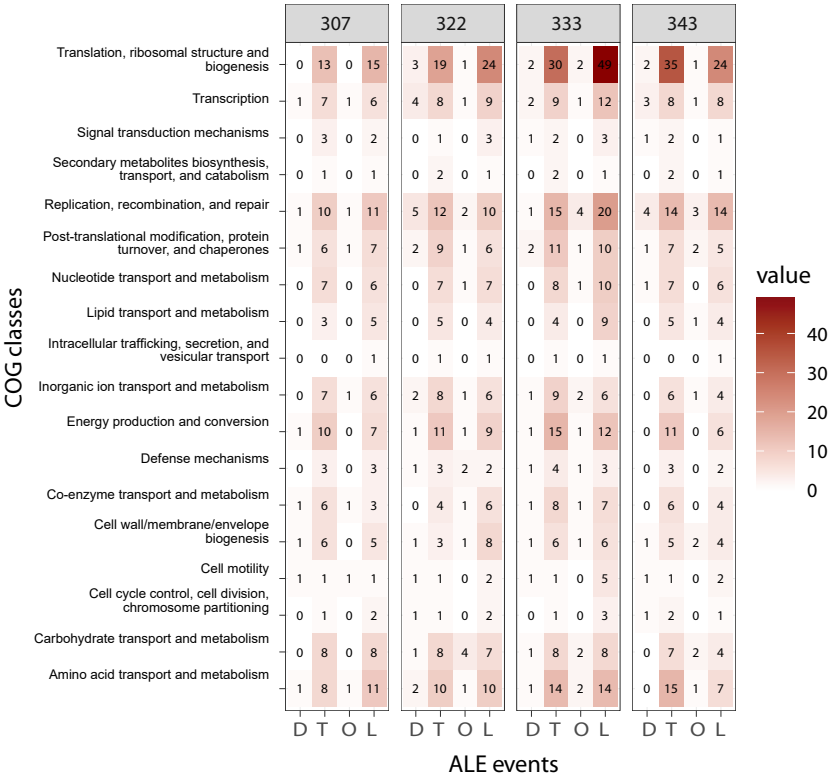


Extended Data Fig. 8 | Ancestral proteome sizes and numbers of gene loss, duplication, and gain inferred from reconciliation analyses across 192-NM archaeal taxa. Barplots at each branch indicate duplication, transfer, and origination events (blue bars; see legend) and loss events (orange bars) (see Methods). Each grey circle indicates the inferred proteome size (i.e., the number of protein-coding gene copies) for the ancestor to the branch's right. Numbers in rounded rectangles correspond to the node identifiers. Halophilic clades are colored.

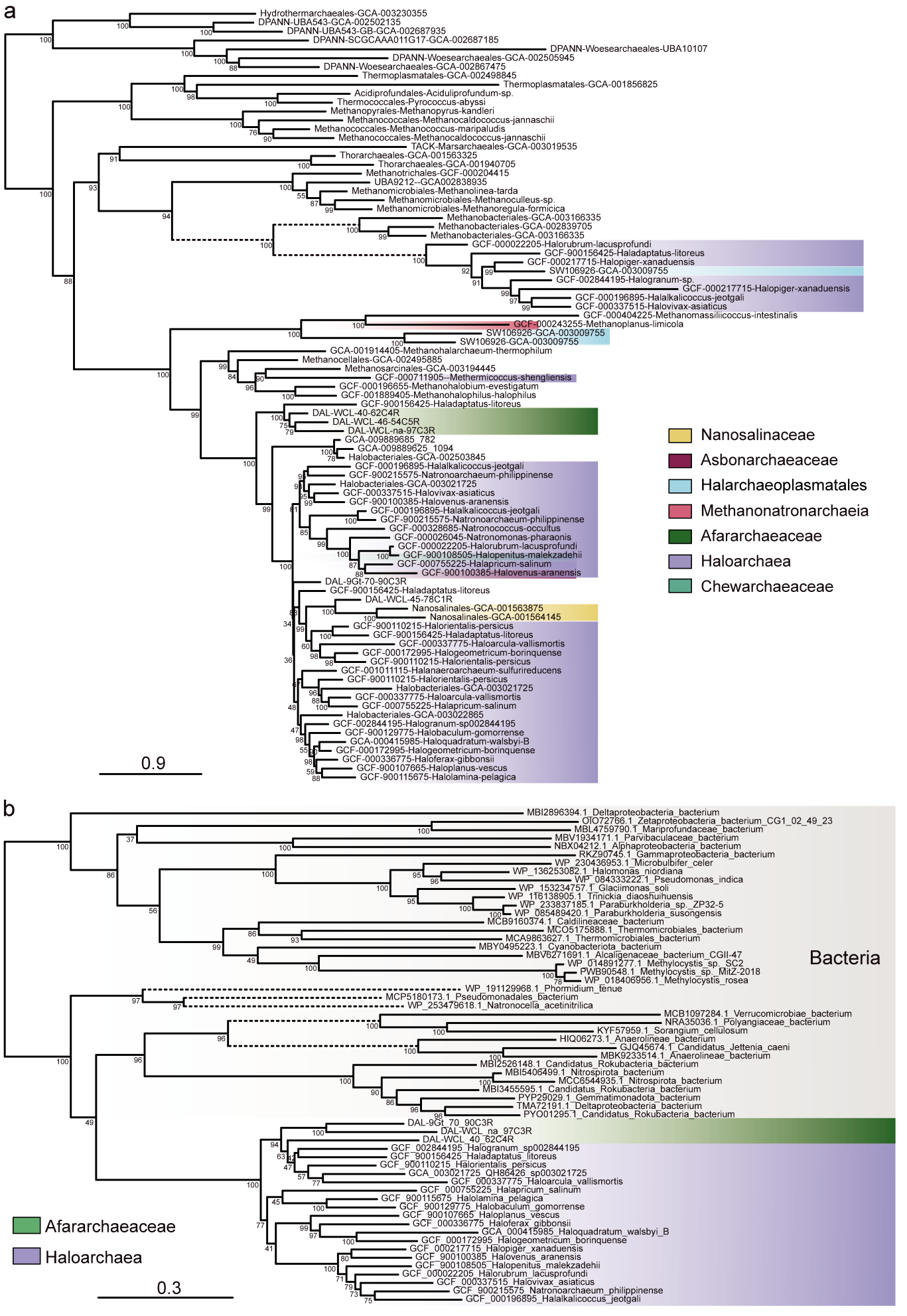
a



b



Extended Data Fig. 9 | Heat map of the number of gene duplications, transfers, originations, and losses in various archaeal halophilic lineages according to their COG classification. The counts were obtained using the amalgamated likelihood estimation (ALE) tree reconciliation method on the set of 17,288 orthologous genes present in the 192-taxa genomic dataset for several nodes within the **(a)** Euryarchaeota and the **(b)** DPANN archaea (see Methods). Node numbers correspond to the nodes in the complete tree shown in Extended Data Fig. 8.



Extended Data Fig. 10 | Maximum likelihood trees showing cases of horizontal gene transfer involving archaeal halophilic lineages. (a) NhaP-type Na+/H+ and K+/H+ antiporters. **(b)** choline dehydrogenase BetA. The trees were constructed with the LG+C60+F+Γ4 model. Dashed branches have been shortened to half of their actual length. The scale bar indicates the expected average number of substitutions per site.

5. Phylogenomic analysis of DPANN archaea reveals their monophyly and evolutionary origins

5. Phylogenomic analysis of DPANN archaea reveals their monophyly and evolutionary origins

5.1. Context

The phylogenetic position of the DPANN archaea is one of the major unresolved questions in archaeal phylogenetics. However, since we invested substantial time in curating our phylogenetic markers and developing tools and methods during the project focused on the phylogeny of halophilic archaea, the project about the phylogeny of DPANN archaea progressed relatively smoothly. Throughout the halophiles project, it became evident that our ribosomal protein phylogenies were affected by phylogenetic artifacts, primarily stemming from amino acid compositional biases. Consequently, ribosomal protein markers proved unreliable when resolving our phylogenetic questions. Because of this prior knowledge from the first project, we prioritized the results of the NM phylogenetic markers for the DPANN project. Additionally, many of the DPANN branches of interest had extremely low statistical branch support in the RP-based trees compared to the NM-based ones. A few analyses for this project are still finishing, but clear trends can be observed in the available results.

If I had more time on this project and in my PhD in general, I would have wanted to address the position of the root of the archaeal tree. However, since this is still an open question, we had to decide how to root our trees to compare the topologies across datasets. We decided to root all trees on the TACK/Asgard branch for several reasons. One, when we rooted the trees on the DPANN branch, we

routinely recovered the Euryarchaeota to be paraphyletic. While the paraphyly of the Euryarchaeota has been found before (Adam et al., 2017a; Raymann et al., 2015), these analyses did not include any DPANN lineages. Based on our data, we looked for any evidence that could support or refute it. We examined the orthologous groups (OGs) shared among the archaeal superphyla as an initial analysis. Our rationale was that if the Euryarchaeota is indeed paraphyletic, the group that places sister to the TACK/Asgard (Euryarchaeota 1) might share more OGs with this branch than with the other Euryarchaeota (Euryarchaeota 2). However, we found the opposite: Euryarchaeota 2 shared more OGs with the TACK/Asgard than Euryarchaeota 1.

Additionally, in previous rooting analyses that place the root between DPANN and all other archaea, the Euryarchaeota were monophyletic (Williams et al., 2017). Surprisingly, no archaeal rooting analyses currently include the DPANN archaea and the Altiarchaeota. As both of these lineages have been proposed as the deepest-branching archaeal lineages, their inclusion could significantly influence the placement of the root. Together with the fact that the root has been previously found between the TACK (Crenarchaeota) and all other archaea (Petitjean, Deschamps, López-García, & Moreira, 2015; C. Woese, 1990), we thought these factors were supportive evidence for placing the root on the TACK/Asgard branch.

5.2. Results

In this project, we aimed to evaluate the monophyly of the DPANN archaea and their position within the archaeal tree. We robustly show that the DPANN archaea are a monophyletic group that branches well nested within the Euryarchaeota when the tree is rooted on the TACK/Asgard branch. The placement

of the DPANN nested within the Euryarchaeota suggests that the DPANN evolved from a Euryarchaeota-like free-living ancestor. This ancestral trait is further reinforced by the position of the free-living Altiarchaeota at the base of the DPANN clade. If all other DPANN lineages are truly obligate episymbionts, this would suggest that the transition from a free-living ancestor to a host-dependent episymbiont occurred only once at the base of the DPANN clade after the divergence from the Altiarchaeota. Yet, the ubiquity of a host-dependent lifestyle remains a major open question that needs to be confirmed experimentally for most known DPANN phyla. We also observed ancient evolutionary links between the DPANN archaea and diverse groups of bacteria, specifically the Patescibacteria and Omnitrophota, which have also been shown to have a similar host-dependent lifestyle (López-García & Moreira, 2021; Seymour et al., 2023). HGT appears to have played a pivotal role in shaping the convergent evolution of these distinct lineages in Bacteria and Archaea towards similar host-dependent lifestyles (Aouad et al., 2018; Castelle et al., 2021; Castelle & Banfield, 2018; Dombrowski et al., 2020; Jaffe et al., 2019; Narasingarao et al., 2012a; Podar et al., 2008; Rinke et al., 2013). Altogether our results provide an updated scenario for the evolution of the DPANN archaea where a single evolutionary event leading to the transition to a symbiotic lifestyle from a free-living ancestor occurred within the Euryarchaeota, most likely accompanied by several HGT events from bacteria.

5.3. Draft manuscript 2

A unique origin of DPANN archaea from free-living euryarchaeal-like ancestors

Brittany A. Baker[1], Charley G. P. McCarthy[2,3], Edward Susko[3,4], Andrew J. Roger[2,3], Purificación López-García[1], Laura Eme[1], David Moreira[1]

[1]Ecologie Systématique Evolution, CNRS, Université Paris-Saclay, AgroParisTech, Gif-sur-Yvette, France.
[2]Institute for Comparative Genomics, Dalhousie University, Halifax, Canada.
[3]Department of Biochemistry and Molecular Biology, Dalhousie University, Halifax, Canada.
[4]Department of Mathematics and Statistics, Dalhousie University, Halifax, Canada.

Abstract

The DPANN archaea have been proposed as one of the four major archaeal superphyla, alongside the TACK, Asgard, and Euryarchaeota. However, the monophyly of this group and its placement within the archaeal tree remain open questions. Reconstructing the phylogeny of the DPANN archaea is difficult and has been shown to be influenced by the choice of markers, taxon sampling, and phylogenetic approaches. This has been particularly apparent in studies that tried to place individual DPANN lineages within the broader archaeal tree. This instability suggests a non-negligible effect of phylogenetic artefacts. In this study, we use a set of 126 highly conserved protein markers, extensive taxon sampling including representatives of the 11 known phyla, and in-depth phylogenomic analyses to reevaluate the monophyly of DPANN archaea and their relationship to other archaea. Our analyses, designed to alleviate possible long-branch attraction (LBA) artefacts derived from differences in evolution and sequence compositional biases, robustly support the monophyly of DPANN archaea and place them well nested within the Euryarchaeota. Notably, we also find the likely free-living Altiarchaeota emerging as the earliest diverging branch within the DPANN archaea, providing new insights into the early stages of the symbiotic lifestyle of the DPANN archaea. Further, we identify 14 proteins that appear exclusive to and shared across the DPANN radiation. Phylogenetic analyses reveal that DPANN archaea likely acquired several of them through ancient horizontal gene transfer events from different bacterial donors. These donors notably include Patescibacteria and Omnitrophota, two bacterial phyla that exhibit an episymbiotic lifestyle similar to DPANN archaea. Our results support that the DPANN archaea are monophyletic and evolved from a free-living euryarchaeal ancestor and suggest that proteins of bacterial origin played a role in establishing their ancestrally symbiotic lifestyle.

Introduction

Nanoarchaeum equitans, which grows as an obligate episymbiont of the hyperthermophilic archaeon *Ignicoccus hospitalis*, was the first nano-sized archaeon to be described (Huber et al. 2002). This discovery was important not only as the initial characterization of an episymbiotic archaeon but also for unveiling a novel, deep-branching archaeal phylum, the Nanoarchaeota (Huber et al. 2002; 2003), which joined the three other archaeal phyla recognized at the time, the Euryarchaeota, Crenarchaeota, and Korarchaeota (Barns et al. 1996; Woese 1990). Subsequent environmental exploration and advancements in metagenomic and single-cell-based methods revealed several additional phyla characterized by small-cell sizes and reduced genomes (Dombrowski et al. 2019; Rinke et al. 2013; Castelle and Banfield 2018; Hug et al. 2016; Castelle et al. 2015). Initial phylogenetic analyses of these newly

identified archaea suggested that these lineages constituted a new monophyletic superphylum, collectively known as the DPANN archaea (an acronym that refers to the first five phyla described for this group; Diapherotrites, Parvarchaeota, Aenigmarchaeota, Nanoarchaeota, and Nanohaloarchaeota) (Rinke et al. 2013). Currently, the DPANN superphylum encompasses a much broader diversity, including 11 phyla (according to the Genome Taxonomy Database, GTDB r207 (Rinke et al. 2021)), which represent a significant fraction of the known diversity of archaea. In GTDB r207, the archaeal tree, along with the 11 DPANN phyla, comprises 8 additional phyla, including Thermoproteota (historically the TACK superphylum), Asgardarchaeota (historically the Asgard superphylum), and 6 phyla historically associated with the Euryarchaeota superphylum (Halobacteriota, Hydrothermarchaeota, Methanobacteriota, Methanobacteriota_B, Thermoplasmatota, and Hadarchaeota).

DPANN genomes consistently lack genes coding for amino acid, nucleotide, and phospholipid biosynthesis pathways, typically considered essential for a free-living lifestyle (Castelle et al. 2018). Based on these genomic characteristics, it has been suggested that the defining feature of the DPANN archaea is a symbiotic lifestyle that requires a host for growth and survival (Dombrowski et al. 2019; Rinke et al. 2013). However, to date, this has only been confirmed for the few species that have been successfully co-cultured, which include four Nanoarchaeota (Huber et al. 2002; Podar et al. 2013; St. John et al. 2019; Wurch et al. 2016), three Micrarchaetoa (Kram et al. 2017; Golyshina et al. 2017; Sakai et al. 2022), and two Nanohaloarchaeota (Hamm et al. 2019; La Cono et al. 2020). Based on these cultivated representatives, some DPANN seems to rely on a single host (Hamm et al. 2019), while others can ‘host-switch’ (Sakai et al. 2022; Dombrowski et al. 2020). However, these cultivated lineages only represent a small fraction of the known DPANN diversity (Dombrowski et al. 2019). For example, different lineages within the Micrarchaeota exhibit a broad diversity both in terms of habitat distribution and metabolic potential (Golyshina et al. 2017; Kadnikov et al. 2020; Golyshina et al. 2019; Chen et al. 2018). Although all three cultured representatives of the Micrarchaeota are from acidic environments (Krause et al. 2017; Golyshina et al. 2017; Sakai et al. 2022), this phylum has also been detected in environments with neutral pH, including soils, freshwater systems, geothermal lakes, and hypersaline mats (Kadnikov et al. 2020; Golyshina et al. 2019; Chen et al. 2018). Interestingly, the freshwater lineage *Candidatus* ‘Fermentimicrarchaeum limneticum*’* has retained several metabolic pathways lost in the acidophilic, host-dependent Micrarchaeota lineages (Golyshina et al. 2019) and has been proposed to be capable of a free-living lifestyle (Kadnikov et al. 2020). However, this is based solely on genomic data and has yet to be proven experimentally.

Since the proposal of the DPANN superphylum, its monophyletic status has been questioned based on potential phylogenetic artefacts due to sequence compositional biases and/or high evolutionary rates (Petitjean et al. 2015; Aouad et al. 2018; Feng et

al. 2021; Brochier et al. 2005). Lineages with similar protein amino acid compositions can artefactually group in phylogenetic trees even if they are distantly related (B. A. Baker et al. 2023; Muñoz-Gómez et al. 2022). Like many parasitic lineages (Rocha and Danchin 2002; Bohlin et al. 2020), many DPANN genomes tend to be A+T-rich (Dombrowski et al. 2019), and consequently, their proteomes are enriched in amino acids encoded by A+T-rich codons. In addition, DPANN generally has a faster evolutionary rate of genome evolution typical of symbiotic lineages, in contrast to free-living ones (Moran and Bennett 2014; Muñoz-Gómez et al. 2022). This tendency often places them at the end of long branches in phylogenetic trees, often leading to long-branch attraction (LBA); this artefact erroneously groups together lineages with long branches in a phylogenetic tree, regardless of their actual evolutionary distances (Bergsten 2005; Susko and Roger 2021; Felsenstein 1978). Further support of the DPANN being susceptible to LBA arises from the discrepancy between the phylogenetic placement of individual DPANN phyla and their placement in trees when all DPANN lineages are considered (Williams et al. 2017). This has been routinely shown for the Nanoarchaeota (Brochier et al. 2005), Nanohaloarchaeota (Feng et al. 2021; Aouad et al. 2018), Micrarchaeota (Petitjean et al. 2015), and Altiarchaeota (Adam et al. 2017; Dombrowski et al. 2020; Schwank et al. 2019).

The phylogenetic placement of the Altiarchaeota is a specifically interesting case. As for the Micrarchaeota commented above, Altiarchaeota has also been proposed to have a free-living lifestyle (Schwank et al. 2019; Probst and Moissl-Eichinger 2015). More recently, it has even been shown that certain members of the Altiarchaeota serve as hosts for members of another DPANN lineage, the Huberarchaeota (Schwank et al. 2019; Esser et al. 2023). It is still unclear whether the Altiarchaeota are truly members of the DPANN archaea. Previous phylogenetic analyses have either placed them as a very deep-branching lineage within the Euryarchaeota (Adam et al. 2017), nested within the Euryarchaeota (Probst and Moissl-Eichinger 2015), sister to all other DPANN (Spang, Caceres, and Ettema 2017) or within the DPANN as a sister group to the Micrarchaeota and Iainarchaeota (previously named Diapherotrites) (Moody et al. 2022; Dombrowski et al. 2020; Castelle and Banfield 2018). If the Altiarchaeota are indeed members of the DPANN archaea, this would greatly expand the metabolic and lifestyle potential of this archaeal superphylum.

In this work, we studied the phylogeny of DPANN archaea using a large set of conserved protein markers and a comprehensive taxon sampling, including representatives of the 11 known phyla. Applying different methods to minimize the potential impact of compositional biases and LBA, we found strong support for the monophyly of DPANN and their placement within the Euryarchaeota. Among the DPANN, the potentially free-living Altiarchaeota is the first branch to diverge. These results have important implications for understanding the tempo and mode of the evolution of the symbiotic lifestyle of DPANN.

Results and Discussion

To investigate the monophyletic nature of the DPANN and their evolutionary relationship with other archaea, we performed a comprehensive range of phylogenomic analyses. For these analyses, we used a carefully selected set of 126 conserved proteins (NM126 dataset; see Methods; Suppl. Data 1) that are present in all four archaeal supergroups (TACK, Asgard, Euryarchaeota, and DPANN) (Suppl. Fig. 1). Our taxonomic sampling comprised 321 taxa, spanning all 19 archaeal phyla. To maintain uniformity, we utilized the taxonomic nomenclature and archaeal classification system suggested by GTDB (release 207; Rinke et al. 2021), along with the informal archaeal supergroups (TACK, Asgard, Euryarchaeota, and DPANN) (Suppl. Data 2). To capture the diversity of the known DPANN lineages, 158 out of the 321 taxa were assigned to the 11 known DPANN phyla. Initial maximum-likelihood (ML) phylogenetic inference based on the NM126 dataset confirmed the monophyly of the 11 DPANN phyla, of the TACK (Thermoproteota sensu GTDB) and Asgard (Asgardarchaeota sensu GTDB), as well as of six distinct Euryarchaeota phyla (Fig. 1). The NM126 tree, rooted on the TACK/Asgard branch, placed the DPANN as a monophyletic group (100% bootstrap support), sister to a clade containing the phyla Halobacteriota, Thermoplasmatota, Methanobacteriota_A, and Hydrothermarchaeota (98% bootstrap support) (Fig. 1 and Suppl. Fig. 2). However, given the recurrent uncertainty regarding the monophyly of the DPANN (Petitjean et al. 2015; Feng et al. 2021; Aouad et al. 2018; Brochier et al. 2005; Dombrowski et al. 2019; B. J. Baker et al. 2020), we conducted in-depth phylogenomic analyses of the NM126 dataset to ascertain the reliability of the DPANN both in terms of their monophyly and relationship to other archaea.

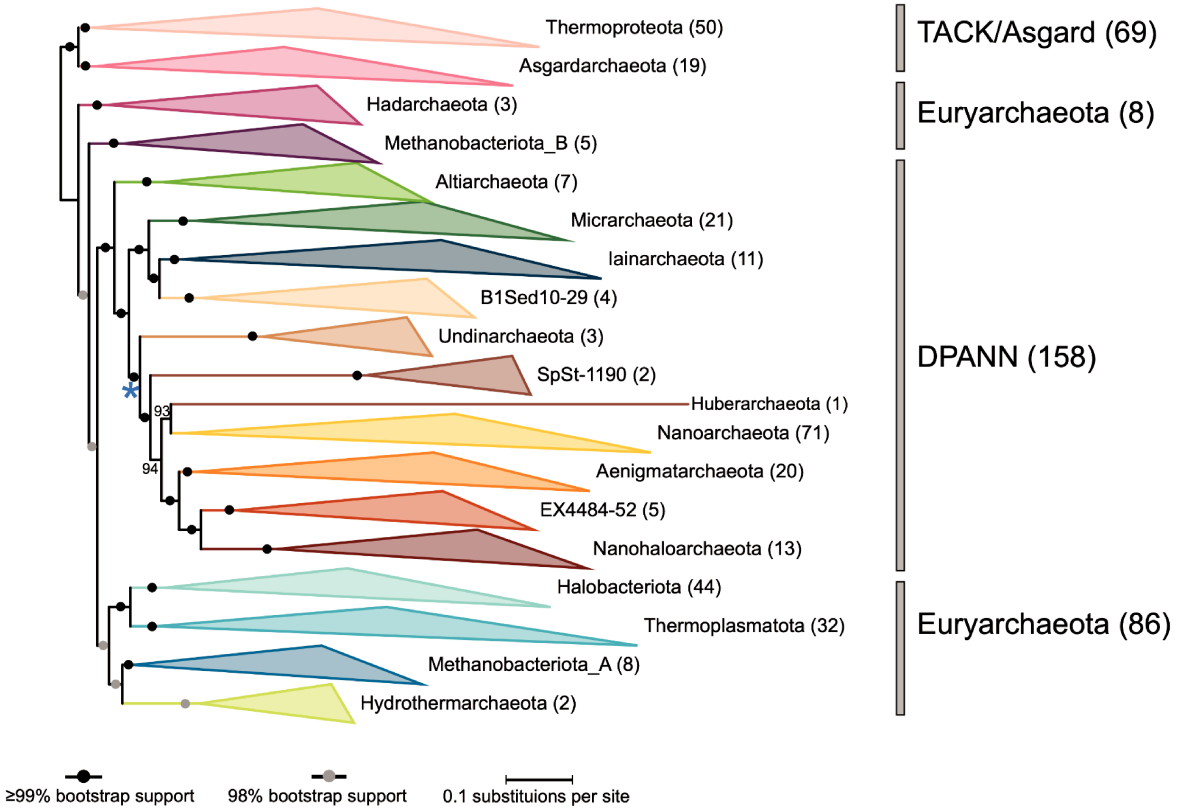


Figure 1 | Maximum likelihood phylogenetic tree spanning all four major archaeal supergroups. The phylogenetic tree is based on the concatenated NM126 dataset (321 taxa, 34,495 amino acid sites). The tree was inferred via IQ-TREE with the LG+C60+Γ4 model of sequence evolution. The statistical support for branches corresponds to 1,000 ultrafast bootstrap replicates. The scale bar indicates the expected average number of substitutions per site. All taxonomic names shown are based on the GTDB r207 phylum-level classification (Rinke et al. 2021). The blue asterisk indicates the group of DPANN phyla that encodes a fused DNA primase (PriS+PriL) instead of two distinct subunits. The expanded tree can be found in Suppl. Fig. 2.

The monophyly of the DPANN archaea

Previous analyses have shown that differences in evolutionary rate, and in particular, long branches, can lead to LBA. To limit potential LBA derived from strong differences in evolutionary rates, we used a strategy consisting of the generation of chimeric datasets that compile the shortest branching sequence for each NM126 marker for each DPANN phylum (“one-chimera” dataset; see Methods section for more details). Using this dataset, we recovered the DPANN as a monophyletic group in an ML tree, with maximum statistical support (Fig. 2a), consistent with the full all taxa dataset.

As expected, we consistently observed shorter branch lengths within the DPANN part of the tree in the one-chimera dataset compared to the full dataset (Fig. 2b). This approach yielded the additional advantage of substantially improving marker coverage for each DPANN chimeric lineage (Fig. 2c). However, an exception to the general trend of branch shortening arose with the phylum SpSt-1190, which displayed a longer root-to-tip branch in the one-chimera tree compared to the full tree (Fig. 2c). This deviation can be attributed to a shift in the position of this group, from a deeper one in the full tree to a more apical one in the one-chimera tree (Fig. 1 and 2a). This shift provides support for the hypothesis that the placement of this phylum in the full tree may have been influenced by an LBA artefact, which is known to misplace long branches towards the base of the phylogenetic trees (Philippe et al. 2000). The chimera analysis appears, therefore, to effectively alleviate potential LBA artefacts. Recently, a member of the SpSt-1190 was found to have a genome size of 4 Mb, which encodes a nearly complete pathway for methanogenesis and several other methane-related metabolisms, such as the ribulose monophosphate pathway and methanofuran biosynthesis (Zhang et al. 2023). To our knowledge, this is the first instance of a DPANN member encoding pathways related to methanogenesis. Determining the precise phylogenetic position of this group is important to understand the evolution of these metabolic features better.

A potential negative effect of the chimera approach is that it reduces the number of taxa, and thus can increase the distances between internal nodes of the DPANN tree. To test if this affected our results, we subsequently included a second chimera for each DPANN lineage by selecting the two shortest branches from each DPANN phylum in the NM126 single-gene trees (“two-chimeras” dataset; see Methods for more details). The resulting ML tree again recovered the DPANN as a monophyletic group with maximum statistical support (Fig. 2d), consistent with the one-chimera and all-taxa datasets (Fig. 1 and 2a). Additionally, since only one Huberarchaeota representative was present in the full dataset, we could not generate a chimera for this lineage. To evaluate the impact of the long branch of Huberarchaeota on our analyses, we ran the one-chimera dataset without the Huberarchaeota (Suppl. Fig. 3). The new tree still retrieved the monophyly of the DPANN with full support and the same relationships among the DPANN branches as in the previous analyses.

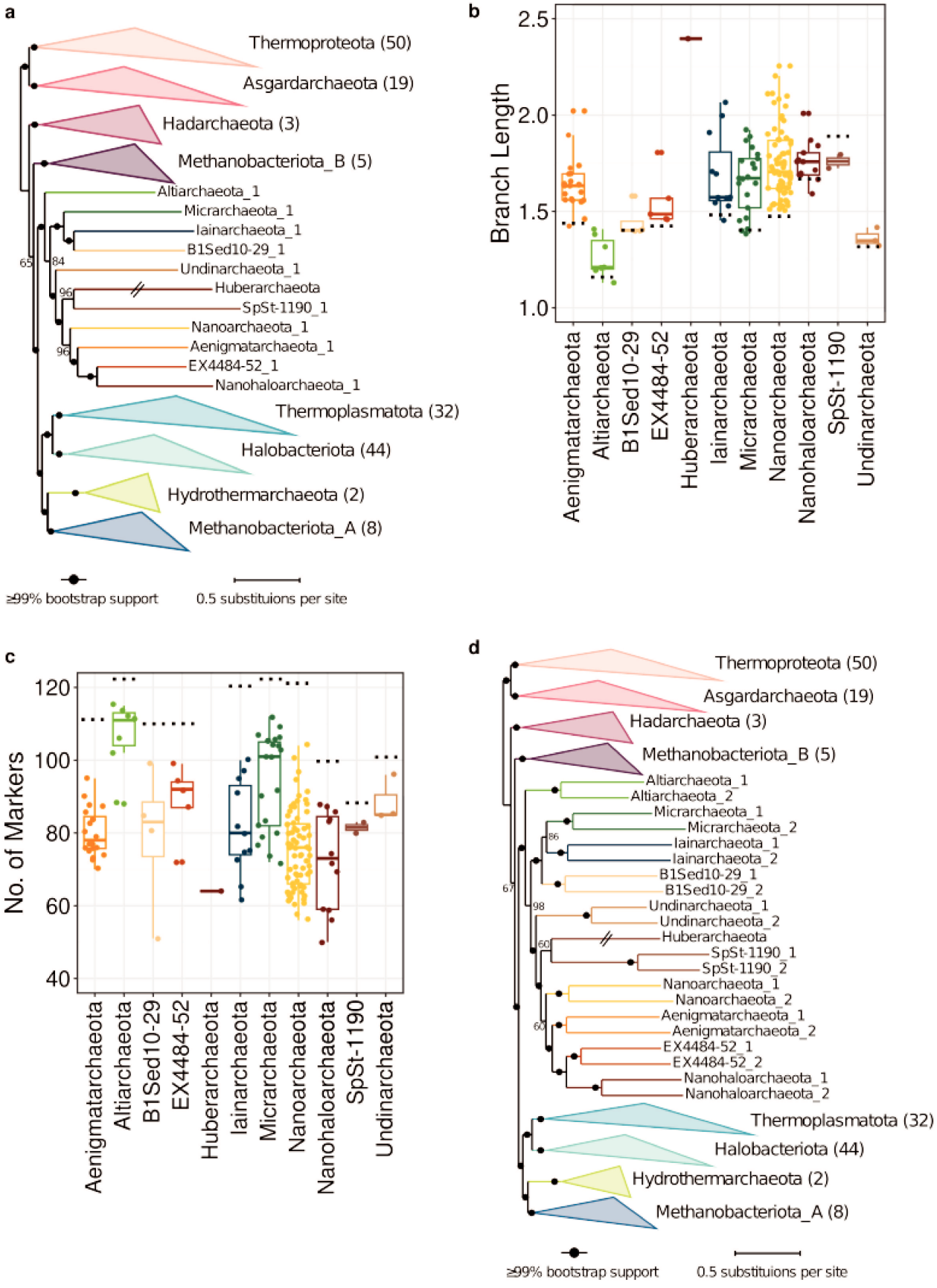

a
Thermoproteota (50)
Asgardarchaeota (19)
Hadarchaeota (3)
Methanobacteriota_B (5)
Altiarchaeota_1
Micrarchaeota_1
Iainarchaeota_1
B1Sed10-29_1
Undinarchaeota_1
Huberarchaeota
SpSt-1190_1
Nanoarchaeota_1
Aenigmatarchaeota_1
EX4484-52_1
Nanohaloarchaeota_1
Thermoplasmatota (32)
Halobacteriota (44)
Hydrothermarchaeota (2)
Methanobacteriota_A (8)
65
84
96
96
≥99% bootstrap support
0.5 substituions per site
b
Branch Length
2.5
2.0
1.5
1.0
Aenigmatarchaeota
Altiarchaeota
B1Sed10-29
EX4484-52
Huberarchaeota
Iainarchaeota
Micrarchaeota
Nanoarchaeota
Nanohaloarchaeota
SpSt-1190
Undinarchaeota
c
No. of Markers
120
100
80
60
40
Aenigmatarchaeota
Altiarchaeota
B1Sed10-29
EX4484-52
Huberarchaeota
Iainarchaeota
Micrarchaeota
Nanoarchaeota
Nanohaloarchaeota
SpSt-1190
Undinarchaeota
d
Thermoproteota (50)
Asgardarchaeota (19)
Hadarchaeota (3)
Methanobacteriota_B (5)
Altiarchaeota_1
Altiarchaeota_2
Micrarchaeota_1
Micrarchaeota_2
Iainarchaeota_1
Iainarchaeota_2
B1Sed10-29_1
B1Sed10-29_2
Undinarchaeota_1
Undinarchaeota_2
Huberarchaeota
SpSt-1190_1
SpSt-1190_2
Nanoarchaeota_1
Nanoarchaeota_2
Aenigmatarchaeota_1
Aenigmatarchaeota_2
EX4484-52_1
EX4484-52_2
Nanohaloarchaeota_1
Nanohaloarchaeota_2
Thermoplasmatota (32)
Halobacteriota (44)
Hydrothermarchaeota (2)
Methanobacteriota_A (8)
67
98
60
60
86
≥99% bootstrap support
0.5 substituions per site

Figure 2 | Phylogenetic analysis of the DPANN using chimera datasets. (a) Maximum likelihood (ML) phylogenetic tree (LG+C60+G model) of the one-chimera dataset. The dataset comprised 11 DPANN phyla, each represented by a single chimera and 163 other Archaea. The statistical support for branches corresponds to 1,000 ultrafast bootstrap replicates. The scale bar indicates the expected average number of substitutions per site. All taxonomic names shown are based on the GTDB r207 phylum-level classification. (b) Root-to-tip branch lengths for the 158 DPANN taxa of the full NM126 dataset. The branch lengths (y-axis) are the root-to-tip number of substitutions per site. Dashed lines represent the branch lengths for the 11 DPANN in the one-chimera dataset. (c) The number of NM126 markers present in each of the 158 DPANN taxa. Black dotted lines represent the number of markers present in the DPANN chimeras of the one-chimera dataset. (d) ML phylogenetic tree (LG+C60+G model) of the two-chimera dataset, consisting of two chimeras for each of the 11 DPANN phyla and 163 other Archaea. The statistical support for branches corresponds to 1,000 ultrafast bootstrap replicates. The scale bar indicates the expected average number of substitutions per site. All taxonomic names shown are based on the GTDB r207 phylum-level classification. The branch leading to Huberarchaeota was shortened by half in both trees and did not represent a chimeric sequence.

As with LBA, lineages with similar protein amino acid compositions can artefactually group in a tree even if they are distantly related (B. A. Baker et al. 2023; Muñoz-Gómez et al. 2022). Therefore, we tested the possible impact of shared amino acid preferences in symbiotic lineages as a potential source of phylogenetic artefacts driving the monophyly of the DPANN (Dombrowski et al. 2019). In bacteria, it has been shown that the genomes of symbiotic, especially parasitic lineages tend to be biased towards A and T nucleotides (and their proteins towards amino acids encoded by A+T-rich codons: F, I, M, N, K, and Y) in contrast to free-living groups that have not evolved reductively (which can be biased towards G+C-rich genomes, and G, A, R, and P amino acids) (Muñoz-Gómez et al. 2022; Clark, Moran, and Baumann 1999; Muñoz-Gómez et al. 2019; Gu, Hewett-Emmett, and Li 1998). Similar to symbiotic bacteria, we observed that the DPANN archaea tend to have a higher frequency of F, I, M, N, K, and Y amino acids compared to free-living archaea, which tend to have a higher frequency of G, A, R, and P amino acids (Suppl. Fig. 4). Most models of sequence evolution, such as the C10-C60 mixture model in an ML framework (Quang, Gascuel, and Lartillot 2008) or the CAT model in a Bayesian framework (Lartillot and Philippe 2004), are built to deal with compositional differences between sites in a protein dataset (site-heterogeneity) but not with amino acid compositional differences between different branches of a tree (branch-heterogeneity) (Muñoz-Gómez et al. 2022; Groussin, Boussau, and Gouy 2013).

To mitigate potential phylogenetic artefacts arising from shared amino acid biases among the DPANN taxa, we excluded the most biased DPANN taxa from the NM126 dataset. These were identified as the lineages displaying the most pronounced compositional biases based on a principal component analysis (PCA) of the amino acid composition of the sequences of the different taxa (Suppl. Fig. 5; see Methods for more details). Using this method, we again observed full statistical support for the monophyly of the DPANN (Suppl. Fig. 6).

We also examined the impact of removing the most biased alignment sites as an alternative to removing the most biased taxa. For our 321-taxa dataset, both a PCA analysis and a binomial test of two proportions revealed that FIMNKY and GARP are the two sets of amino acids that exhibited the highest over- and under-representation, respectively, in DPANN in comparison to other archaeal lineages (see Methods for more details; Suppl. Figs. 4 and 7). We therefore developed a ranking scheme for progressively removing the sites exhibiting the greatest compositional biases associated with FIMNKY and GARP amino acids. We first calculated the FIMNKY/GARP ratio for DPANN versus non-DPANN lineages, ranked the sites accordingly, and then removed 15% of the most biased sites. A phylogenetic tree based on the resulting alignment again retrieved the monophyly of the DPANN with full statistical support (Suppl. Fig. 8). However, as all DPANN taxa did not show uniform amino acid preferences (Suppl Figs. 4 and 9), we also calculated the site-by-site FIMNKY/GARP ratio between two taxon sets selected based on their proteome-wide amino acid compositions. One set comprised taxa whose overall proteome is enriched in FIMNKY amino acids, while the second set included taxa whose proteome is enriched in GARP amino acids, irrespective of their phylogenetic affiliation (Suppl. Data 3). After removing 15% of the most biased sites identified in this way, we again recovered the DPANN as a monophyletic group with full statistical support (Suppl. Fig. 10).

The phylogenetic relationship between the DPANN and other Archaea

Based on the initial ML phylogenetic inference of the NM126 dataset rooted on the TACK/Asgard branch, the DPANN archaea placed as a monophyletic group within the Euryarchaeota, sister to the clade encompassing the phyla Halobacteriota, Thermoplasmatota, Methanobacteriota_A, and Hydrothermarchaeota (Fig. 1). While individual DPANN lineages have been previously found within the Euryarchaeota (Petitjean, Deschamps, López-García, and Moreira 2015; Aouad et al. 2018; Feng et al. 2021; Brochier et al. 2005), to our knowledge, this is the first time all DPANN lineages were recovered as a monophyletic group within the Euryarchaeota. To verify this position, we again tested whether this placement could be attributed to potential phylogenetic artefacts such as sequence compositional biases and/or LBA.

First, to test the impact of LBA on the position of the DPANN, we analyzed the positions of individual DPANN lineages (Micrarchaeota, Nanoarchaea, Nanohaloarchaeota, and Altiarchaeota) known for their variable placements (Petitjean et al. 2015; Narasingarao et al. 2012; Aouad et al. 2018; B. J. Baker et al. 2006; Brochier et al. 2005). Each of these studies used a different taxonomic sampling and set of phylogenetic markers. However, without any artefact, all the constituent DPANN lineages should be individually placed at the same point in the archaeal tree when using similar markers (Williams et al. 2017). In our analyses, always rooted on the TACK/Asgard branch, the Micrarchaeota, and Nanoarchaeota were placed individually at the base of the Euryarchaeota with maximum statistical support (Suppl. Figs. 11 and 12). While this placement has been previously reported for a subset of Nanoarchaeota lineages (Huber et al. 2002; Brochier et al. 2005; Williams et al. 2017), this is the first instance where the Micrarchaeota placed at the base of the Euryarchaeota when other DPANN phyla were not considered.

Contrary to the Micrachaeota and Nanoarchaeota, the Altiarchaeota and Nanohaloarchaeota were nested within the Euryarchaeota with full statistical support, albeit at different positions (Suppl. Fig. 13 and 14). Unlike previous studies which have found the Nanohaloarchaeota as either sister to the Halobacteriota (Narasingarao et al. 2012; Petitjean et al. 2015; Feng et al. 2021) or sister to Methanocellales (Aouad et al. 2018), we found the Nanohaloarchaeaota sister to the clade formed by Halobacteriota and Thermoplasmatota (Suppl. Fig. 14). However, the position of the Nanohaloarchaeota has been previously shown to be sensitive to tree reconstruction artefacts due to the convergent evolution of amino acid preferences with other halophilic archaea (B. A. Baker et al. 2023). This suggests that the Nanohaloarchaeota might be artefactually drawn to the Halobacteriota when other DPANN lineages are absent.

Furthermore, like the Nanohaloarchaeota, the Altiarchaeota have also been found in various positions within the Euryarchaeota (Probst et al. 2014; Adam et al. 2017). In our analyses, we routinely recover the Altiarchaeota sister to the Halobacteriota, Thermoplasmatota, Methanobacteriota_A, and Hydrothermarchaeota, agreeing with the position of all DPANN in the full NM126 tree (Fig. 1 and Suppl. Fig. 13). Interestingly, in our full NM126 tree, we recover the Altiarchaeota as a sister group of all other DPANN (Fig. 1). This position differs from recent phylogenetic analyses of the DPANN, which places the Altiarchaeota sister to the Micrarchaeota and Iainarchaeota (Dombrowski et al. 2020; Moody et al. 2022; Castelle and Banfield 2018). These two conflicting positions for the Altiarchaeota significantly impact our interpretation of the evolutionary history of the DPANN archaea. Altiarchaeota, which exhibit the most gene-rich genomes among DPANN archaea, have been proposed to be free-living organisms (Probst et al. 2014). Recent studies have also suggested that certain members of the Altiarchaeota likely function as a host for another DPANN group, the Huberarchaeota (Esser et al. 2023; Schwank et al. 2019). If the Altiarchaeota do

place as a sister group of the other DPANN lineages, this would suggest that they diverged before the other DPANN lineages adapted to a symbiotic lifestyle. We conducted an approximately unbiased (AU) test (Shimodaira 2002) to compare these two tree topologies. We found that the NM126 alignment strongly rejects the placement of the Altiarchaeota sister to the Micrarchaeota and Iainarchaeota (*P*-value = 0.00546).

To test the possible effect of LBA on the analysis of individual DPANN phyla, we applied the same approach of branch shortening based on the construction of two chimeric sequences for each DPANN subset, as described above (Suppl. Figs. 15-18). Interestingly, the Nanoarchaeota and Micrarchaeota trees with the two-chimera sequences agreed with the full NM126 dataset, placing them with full support sister to the group formed by the Halobacteriota, Thermoplasmatota, Methanobacteriota_A, and Hydrothermarchaeota (Fig. 1 and Suppl. Figs. 15 and 16). In contrast, the Altiarchaeota moved to a deeper-branching position sister to the Hydrothermarchaeota, albeit with low statistical support (68% ultrafast bootstrap) (Suppl. Fig. 17), while the Nanohaloarchaeota again placed sister to the Thermoplasmatota and Halobacteriota (Suppl. Fig. 18). The consistent placing of the Nanohaloarchaota sister to the Thermoplasmatota and Haloarchaeota suggests that compositional biases, as opposed to LBA, drive the signal in this dataset. The observed conflicting placements of the Micrarchaeota and Nanoarchaeota lineages in the full subset analyses seem to be influenced, at least in part, by LBA.

The phylogeny of DPANN archaea based on ribosomal proteins

Ribosomal proteins are one of the most commonly used sets of proteins for reconstructing the phylogeny of archaea (Hug et al. 2016; Dombrowski et al. 2020; Castelle and Banfield 2018; Matte-Tailliez et al. 2002; Ramulu et al. 2014; Brochier et al. 2005; Pace 1997). Despite their prevalence in phylogenomic analyses, ribosomal proteins have been suggested to contribute to phylogenetic artefacts due to inherent sequence compositional biases (Eme et al. 2023; Petitjean et al. 2015; B. A. Baker et al. 2023). Indeed, PCA analyses of our comprehensive archaeal taxon dataset corresponding to the amino acid composition of a set of 48 ribosomal proteins (RP48 dataset) and the NM126 dataset showed remarkable differences between the two (Fig. 3a and b). Nevertheless, due to their historical significance, we ran all aforementioned phylogenetic analyses using the RP48 dataset. For all of the RP48 datasets, we recovered the DPANN as a monophyletic group with full statistical support (Fig. 3c; Suppl. Figs. 19-31). Additionally, when rooted on the TACK/Asgard branch, we again found the DPANN nested within the Euryarchaeota, consistent with the NM dataset, but in a varied position, sister to a group consisting of the Halobacteriota and Thermoplasmatota, albeit with low statistical support (Fig. 3c; Suppl. Figs. 19-31).

Furthermore, the Altiarchaeota routinely placed sister to the Micrarchaeota and Iainarchaeota in all the RP trees (Fig. 3c; Suppl. Figs. 19-25), similar to what has been

observed in previous analyses (Moody et al. 2022; Dombrowski et al. 2020; Castelle and Banfield 2018). We conducted an AU test (Shimodaira 2002) to compare the Altiarchaeota NM and RP topologies. We found that the NM alignment strongly rejected the RP topology as mentioned above (*P*-value = 0.00546), but the RP alignment did not reject the NM topology (*P*-value = 0.0771). The Altiarchaeota also placed as a sister group to the Micrarchaeota and Iainarchaeota in the one- and two-chimeras RP datasets with full statistical support (Suppl. Figs. 20-22). This supports the idea that this placement is not a result of LBA, indicating the presence of a conflicting signal in the NM and RP datasets (Fig. 3a and b).

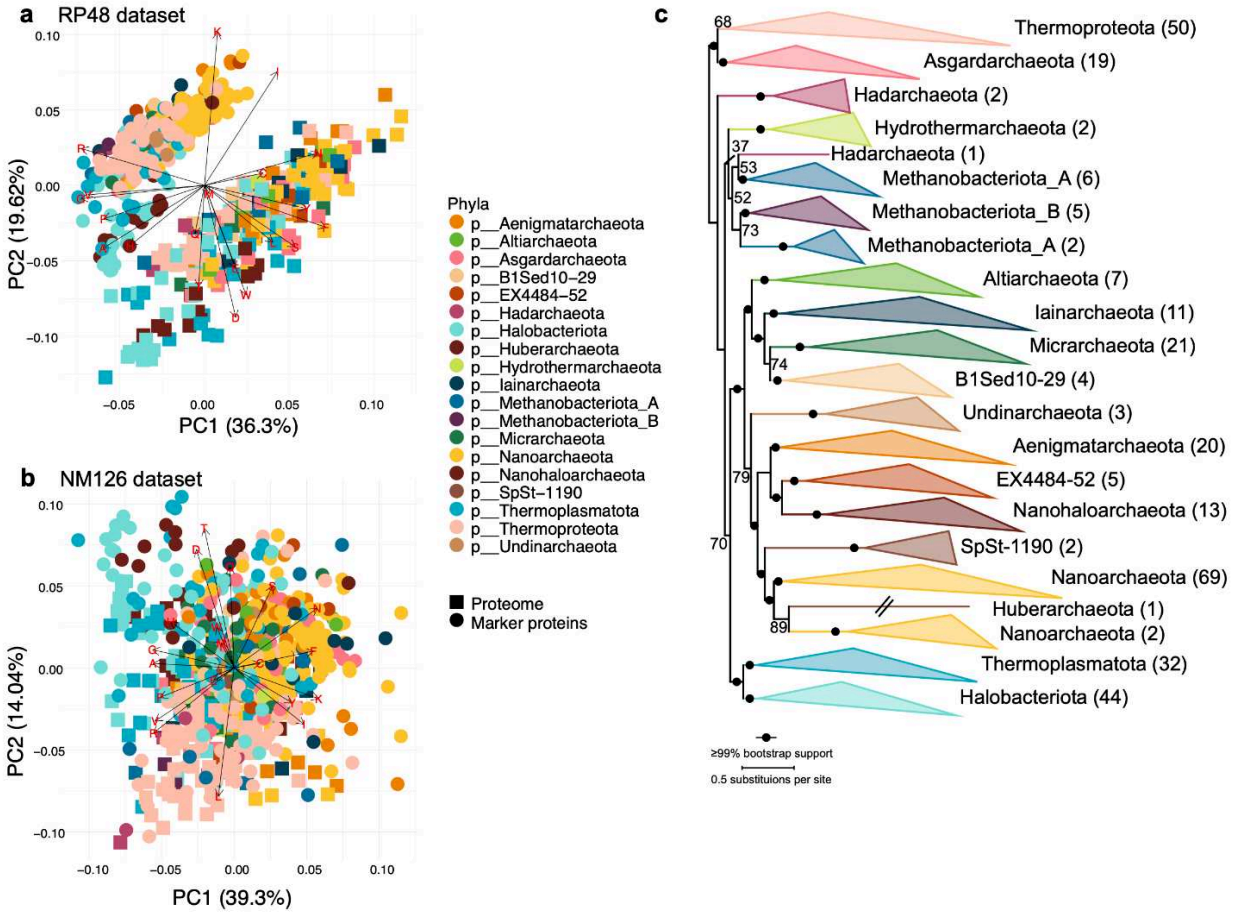


Figure 3 | Compositional and phylogenetic analyses of ribosomal proteins. (a) Principal component analysis (PCA) based on the amino acid composition of 48 ribosomal proteins (RP48 dataset, circles) versus the complete proteomes (squares) for 321 archaeal species. (a) PCA based on the amino acid composition of 126 non-ribosomal proteins (NM126 dataset, circles) versus the complete proteomes (squares) for 321 archaeal taxa. (c) Maximum likelihood phylogenetic tree (LG+C60+G model) based on the ribosomal RP48 dataset (321 taxa, 5,998 sites). The statistical

support for branches corresponds to 1,000 ultrafast bootstrap replicates, with filled circles representing values equal to or larger than 99% support. The scale bar indicates the expected average number of substitutions per site. All taxonomic names shown are based on the GTDB r207 phylum-level classification. The expanded tree can be found in Suppl. Fig. 19.

Identification of DPANN signature proteins

To identify proteins specific to the DPANN archaea that may provide additional evidence for the monophyly of this group, we grouped the 531,483 proteins encoded in the 321 archaeal genomes used for phylogenetic analyses into 25,730 clusters of orthologous genes (OGs). None of these OGs was present across all the 11 DPANN phyla and absent in other archaea. However, several DPANN genomes are very small and encode very reduced sets of proteins (e.g., the 518 kb genome of *Huberarchaeum crystalense* only codes for 541 proteins (Schwank et al. 2019), which limits the possibility of finding characteristic DPANN signature proteins. Therefore, we used a less strict filter to identify proteins shared by at least seven DPANN phyla but absent in other archaea. Despite these relaxed criteria, we only identified 14 proteins (Suppl. Fig. 32 and Suppl. Data 4), which are likely ancient proteins that have been conserved throughout the evolutionary history of the DPANN superphylum. An additional OG (OG0889) was also identified as DPANN-unique. However, it corresponds to the fusion of the two primase subunits, PriS and PriL. Although both subunits are universal in archaea, this fused version is only found in several DPANN lineages (Adam et al. 2017). Our results confirm that the PriS+PriL fusion is present in the Undinarchaeota, SpSt-1190, Huberarchaeota, Nanoarchaeota, Aengmatarchaeota, EX4484-52, and Nanohaloarchaeota (Fig. 1), providing a synapomorphy that strongly supports the monophyly of this DPANN subgroup (Dombrowski et al. 2020).

Most DPANN archaea are thought to have a symbiotic lifestyle, given their reduced genomes and metabolic repertoires (Castelle et al. 2018; Dombrowski et al. 2019; Castelle and Banfield 2018). It is tempting to speculate that some of the widespread ancient DPANN-exclusive proteins may have been involved in the evolution of this lifestyle that depends on close physical interactions with host cells. In agreement with this idea, transmembrane and secreted proteins were overrepresented among these proteins (10 out of 14, see Suppl. Data 4). Three of them share a similar architecture, consisting of an N-terminal transmembrane domain that anchors the protein to the cell membrane and a C-terminal segment predicted to be exposed outside the cells and containing a conserved sequence domain involved in adhesion. This is the case for OG2138, a large protein (between 600 and 1400 amino acids) with a C-terminal type A von Willebrand factor (vWA, InterPro IPR002035) domain, and for OG1262 and OG1760, both with C-terminal bacterial-like immunoglobulin fold (Ig, InterPro IPR013783) domains. vWA and Ig domains have been found in many proteins

involved in cell adhesion (Whittaker and Hynes 2002; Chatterjee et al. 2021) and have been proposed to participate in cell-cell interactions in archaea, including DPANN (Qu et al. 2023).

It can be speculated that three other OGs (OG1509, OG2458, and OG2966) that have the same architecture may also be involved in cell recognition and adhesion. However, they do not contain any recognizable protein domain in their C-terminal regions. Although these OGs are absent in non-DPANN archaea, OG2458 is also present in diverse bacteria. Phylogenetic analysis of this OG shows that most DPANN sequences form a well-supported group (100% ultrafast bootstrap) that branches within a large group of bacterial sequences (Suppl. Fig. 33). Interestingly, these sequences belong to very diverse Patescibacteria. These bacteria, previously known as Candidate Phyla Radiation (CPR; (Hug et al. 2016)), form a bacterial phylum that exhibits several similarities with DPANN archaea (López-García and Moreira 2021), as it contains species with small cells, reduced genomes and metabolic repertoires, and presumed to be obligate symbionts of other bacteria (Brown et al. 2015; Castelle and Banfield 2018). The phylogeny of OG2458 suggests that this gene was acquired by DPANN archaea from a patescibacterial donor by an ancient horizontal gene transfer (HGT) event. In addition, several bacterial sequences intermixed with the DPANN ones most likely reflect more recent cases of HGT.

We observed similar cases of HGT from bacteria at the origin of DPANN OGs OG2142 and OG2651, which correspond to proteins with a predicted cytosolic localization (Suppl. Data 4 and Suppl. Figs. 34 and 35). In the case of OG2651, an HD superfamily phosphohydrolase, the DPANN sequences are also closely related to patescibacterial sequences (Suppl. Fig. 35), whereas for OG2142, an HxsB His-Xaa-Ser system radical SAM maturase, the DPANN sequences are closely related to bacterial sequences from the phylum Omnitrophota (Suppl. Fig. 34). As Patescibacteria, Omnitrophota seems to correspond to small cells with reduced genomes and a gene content suggesting a symbiotic lifestyle (Seymour et al. 2023). Our results support that DPANN archaea obtained some of their most ancient and widespread genes by HGT from these two groups of host-associated bacteria. Interestingly, a member of the Omnitrophota, '*Candidatus* Velamenicoccous archaeavorus', can parasitize methanogenic archaea of the genus *Methanosaeta* (Kizina et al. 2022). This opens the intriguing possibility that some Omnitrophota and DPANN archaea may share similar archaeal hosts, which could have facilitated the exchange of genes between these two distant lineages.

Concluding remarks

Our in-depth phylogenomic analyses support that the monophyly of the DPANN archaea does not result from phylogenetic artefacts, such as LBA and/or sequence

compositional biases. These analyses robustly place the DPANN within the Euryarchaeota in trees rooted on the TACK/Asgard branch. Given the ongoing uncertainty surrounding the position of the root of the archaeal tree, we cannot definitively exclude the possibility that the root is positioned between the DPANN and the rest of the Archaea. However, according to our phylogenetic analyses, this root placement would imply that the Euryarchaeota are paraphyletic and that the TACK/Asgard group evolved from an Euryarchaeota-like ancestor. While this root position has been previously proposed (Adam et al. 2017; Raymann et al. 2015), neither of these studies included DPANN lineages in their phylogenetic analysis. Further, a more recent study that used an outgroup-free rooting approach identified the root between the DPANN and all other archaea, but contrary to our analysis, recovered the Euryarchaeota as monophyletic when using a supertree approach (Williams et al. 2017). These studies highlight the ongoing uncertainty surrounding the position of the archaeal root.

The placement of the DPANN nested within the Euryarchaeota suggests that the DPANN evolved from a Euryarchaeota-like free-living ancestor. This ancestral trait is further reinforced by the position of the free-living Altiarchaeota at the base of the DPANN clade. If all other DPANN lineages are truly obligate episymbionts, this would suggest that the transition from a free-living ancestor to a host-dependent episymbiont occurred only once at the base of the DPANN clade after the divergence from the Altiarchaeota. Yet, the ubiquity of a host-dependent lifestyle remains a major open question that needs to be confirmed experimentally for most known DPANN phyla. Intriguingly, lineages within the Micrarchaeota and Iainarchaeota have been proposed to be capable of a free-living lifestyle based on the analysis of their gene content (Youssef et al. 2015; Kadnikov et al. 2020). Since these lineages have phylogenetic positions well nested within the DPANN radiation, it was tempting to speculate that they reverted to a free-living lifestyle from host-associated ancestors. However, Kadnikov et al. argued that the potentially free-living *Candidatus* ‘Fermentimicrarchaeum limneticum’ represents the ancestral state rather than a derived feature of phylum Micrarchaeota (Kadnikov et al. 2020).

Conversely, for *Candidatus* ‘Iainarchaeum andersonii’ (Iainarchaeota), the authors argued that this parasitic-to-free-living evolutionary transition was shaped by extensive HGT-mediated acquisition of key metabolic proteins from diverse bacterial donors (Youssef et al. 2015). Interestingly, we also observe ancient evolutionary links between the DPANN archaea and diverse groups of bacteria, specifically the Patescibacteria and Omnitrophota, which have also been shown to have a similar host-dependent lifestyle (López-García and Moreira 2021; Seymour et al. 2023). HGT appears to have played a pivotal role in shaping the convergent evolution of these distinct lineages in Bacteria and Archaea towards similar host-dependent lifestyles (Podar et al. 2008; Dombrowski et al. 2020; Aouad et al. 2018; Jaffe et al. 2019;

Narasingarao et al. 2012; Castelle and Banfield 2018; Rinke et al. 2013; Castelle et al. 2021). Altogether our results provide an updated scenario for the evolution of the DPANN archaea where a single evolutionary event leading to the transition to a symbiotic lifestyle from a free-living ancestor occurred within the Euryarchaeota, most likely accompanied by several HGT events from bacteria that triggered the DPANN radiation.

Methods

Selection of archaeal taxa

To ensure we had a representative set of archaeal taxa for our phylogenetic analyses, we downloaded all archaeal proteomes from the Genome Taxonomy Database (r207) (Rinke et al. 2021) and selected genomes from all archaeal classes using Treemmer v0.3 (Menardo et al. 2018). Treemmer selects the leaves in a phylogenetic tree representing the greatest diversity based on a predefined number of taxa. After automatically selecting lineages using Treemmer, we manually refined the selection based on the genome completeness and contamination (>75% and <5%, respectively) for each archaeal class according to GTDB (r207). Our final selection included 321 archaeal taxa (158 DPANN, 94 Euryarchaeota, and 69 TACK/Asgard).

Curation of phylogenetic markers

The NM126 dataset was based on curating a set of 136 markers previously shown to be highly conserved across the archaeal domain (Petitjean et al. 2015; B. A. Baker et al. 2023). Sequences similar to the NM proteins were identified in the set of 321 archaeal taxa using BLAST (Altschul et al. 1990) with relatively relaxed criteria (>20% sequence identity over 30% query length) to retrieve even divergent homologs, which is especially important in the case of fast-evolving lineages like the DPANN archaea. For each 321 taxa, up to three of the best BLAST hit sequences were included for preliminary phylogenetic reconstruction using FastTree2 (Price, Dehal, and Arkin 2010). These preliminary trees were manually examined to identify the correct orthologue for each taxon and to detect cases of contamination, HGT, or paralogy. These spurious sequences were removed, and the remaining ones were used to reconstruct a new tree. Multiple rounds of manual curation were done this way until all problematic sequences were removed. This also removed 10 of the 136 markers because they were too poorly represented in DPANN. Once curated, sequences of each of the 126 remaining markers were aligned with MAFFT L-INS-i v7.450 (Katoh and Standley 2013) and trimmed with BMGE v1.12 (-m BLOSUM30 -b 3 -g 0.2 -h 0.5) (Criscuolo and Gribaldo 2010). We performed a final round of verification of the single

gene trees reconstructed using the more sophisticated LG+C60+F+Γ4 model in IQ-TREE (Minh et al. 2020) before concatenating the individually trimmed alignments into the NM126 supermatrix (126 proteins, 34,495 alignment sites). Finally, the RP48 dataset was based on a previously curated set of 48 ribosomal proteins (B. A. Baker et al. 2023; Petitjean et al. 2015). The selection and curation process for the RP48 dataset mirrored that of the NM126 dataset described above. The final RP48 supermatrix consisted of 48 proteins and 5,998 alignment sites.

Short-branching chimera datasets

To construct the one-chimera dataset, we first selected the shortest branch for each of the 11 DPANN phyla from the NM126 and RP48 single gene trees (see above). Once the shortest branch was selected, all other sequences from the corresponding DPANN phylum were removed from the marker alignment. The new marker datasets generated in this way were then realigned with MAFFT L-INS-i v7.450 (Katoh and Standley 2013) and trimmed with BMGE v1.12 (-m BLOSUM30 -b 3 -g 0.2 -h 0.5) (Criscuolo and Gribaldo 2010). The trimmed alignments were concatenated into the one-chimera supermatrices for the NM and RP markers. We ran this analysis again to generate the two-chimera datasets, including the two shortest branches for each of the 11 DPANN phyla. To ensure we did not have overlapping taxa in the two chimeras, we first subdivided each of the 11 phyla into two groups based on the full NM126 tree (Suppl. Fig. 2 chimera subsets). We then selected the shortest branch from each of the two subgroups. Again, these new marker datasets were realigned with MAFFT L-INS-i v7.450 (Katoh and Standley 2013) and trimmed with BMGE v1.12 (-m BLOSUM30 -b 3 -g 0.2 -h 0.5) (Criscuolo and Gribaldo 2010). The trimmed alignments were concatenated into the two-chimera supermatrices for the NM and RP datasets.

Sequence composition-based datasets

To remove the DPANN taxa with the most biased amino acid compositions from the NM and RP datasets, we built a principal component analysis (PCA) based on the amino acid frequency of the 158 DPANN proteomes. We then divided the principal component 1 (PC1) at 2.5 for both the positive and negative directions (Suppl. Fig. 5. This allowed us to keep all DPANN phyla except the Huberarchaeota. We opted to remove taxa based on PC1, given its larger contribution to the data (44.1%) compared to PC2 (10.53%). Taxa that fell outside this boundary were removed from the individual marker alignments. These datasets were then realigned with MAFFT L-INS-i v7.450 (Katoh and Standley 2013) and trimmed with BMGE v1.12 (-m BLOSUM30 -b 3 -g 0.2 -h 0.5) (Criscuolo and Gribaldo 2010). The trimmed alignments were concatenated into a supermatrix for each NM and RP dataset.

Removal of compositionally biased sites

We devised a ranking scheme to remove the most {FIMNKY/GARP}-biased amino acid (AA) alignment sites from the NM and RP datasets. This involved calculating the ratio of {FIMNKY} versus {GARP} AAs for one set of taxa in relation to a second set of taxa. For this analysis, we divided the taxa into two bins (P_{up} and P_{down}), using two different partitioning schemes. One scheme consisted of binning the taxa by DPANN (P_{up}) versus all other Archaea(P_{down}), and the second consisted of binning the taxa with a {FIMNKY/GARP} ratio above the global average for all Archaea (P_{up}) versus taxa below the global average (P_{down}). We then calculated the frequency of the {FIMNKY} (P_{FIMNKY}) versus {GARP} (P_{GARP}) AAs for each alignment site (X_i) for each bin. To adjust for a possible denominator of 0, a pseudocount (α) of 0.01 was added to the frequency of each AA considered (i.e., 0.6 for FIMNKY and 0.4 for GARP). The adjusted ratios were then divided by the total number of alignment sites counted (N) (i.e., 6 for FIMNKY and 4 for GARP) plus the corrected pseudocount for all 20 amino acids ($\alpha d = 0.01 \times 20$). To calculate the ratio for each of the bins (P_{up} and P_{down}), we then divided the {FIMNKY} ratio by the {GARP} ratio. We then calculated P_{total} as the logarithm of the ratio of the up bin versus the down bin. The alignment site ratios were then ranked in descending order based on the P_{total} value, and the top 15% of the most biased sites were removed from the NM and RP datasets.

Biased-sites ranking formula:

$$P_{FIMNKY} = \frac{X_i+\alpha}{N+\alpha d}, P_{GARP} = \frac{X_i+\alpha}{N+\alpha d}$$

$$P_{up} = \frac{P_{FIMNKY}}{P_{GARP}}, P_{down} = \frac{P_{FIMNKY}}{P_{GARP}}$$

$$P_{total} = log\left(\frac{P_{up}}{P_{down}}\right)$$

Identification of DPANN signature proteins

Orthologous groups (OGs) were identified for all species in our list of 321 archaeal taxa using OrthoFinder v2.5.1 with Diamond BLAST set to (--ultra-sensitive, --query-cover 50%, and --id 30%) and an inflation parameter of 1.1 (Emms and Kelly 2019). We then analyzed the presence and absence of taxa within the DPANN, specifically focusing on OGs present in a minimum of 7 out of the 11 DPANN phyla. Since some absences could be due to genome sequence incompleteness or specific

losses in some taxa, we deemed a phylum present if at least one species of that phylum was identified in the OG. This led to the identification of 14 OGs that were present in at least seven out of the 11 DPANN phyla and absent in the rest of Archaea. Conserved protein domains were identified using InterproScan v5.31-70.0 (Jones et al. 2014), and cellular localization was predicted with DeepTMHMM v1.0.24 (Hallgren et al. 2022). We looked for potential bacterial homologs of these OGs using BLAST (Altschul et al. 1990) searches against the GTDB r207 database (Rinke et al. 2021) with an e-value threshold of 1e-03 and keeping a maximum of 500 hits. The sequences of OGs containing both DPANN and bacterial sequences were aligned with MAFFT L-INS-i v7.450 (Katoh and Standley 2013) and trimmed with BMGE v1.12 (-m BLOSUM30 -b 3 -g 0.2 -h 0.5) (Criscuolo and Gribaldo 2010). Maximum likelihood phylogenetic trees were then reconstructed using the LG+C20+F+Γ4 model of sequence evolution in IQ-TREE (Minh et al. 2020).

References

Adam, Panagiotis S., Guillaume Borrel, Céline Brochier-Armanet, and Simonetta Gribaldo. 2017. “The Growing Tree of Archaea: New Perspectives on Their Diversity, Evolution and Ecology.” *The ISME Journal* 11 (11): 2407–25. https://doi.org/10.1038/ismej.2017.122.

Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. 1990. “Basic Local Alignment Search Tool.” *Journal of Molecular Biology* 215 (3): 403–10. https://doi.org/10.1016/S0022-2836(05)80360-2.

Aouad, Monique, Najwa Taib, Anne Oudart, Michel Lecocq, Manolo Gouy, and Céline Brochier-Armanet. 2018. “Extreme Halophilic Archaea Derive from Two Distinct Methanogen Class II Lineages.” *Molecular Phylogenetics and Evolution* 127 (October): 46–54. https://doi.org/10.1016/j.ympev.2018.04.011.

Baker, Brett J., Valerie De Anda, Kiley W. Seitz, Nina Dombrowski, Alyson E. Santoro, and Karen G. Lloyd. 2020. “Diversity, Ecology and Evolution of Archaea.” *Nature Microbiology* 5 (7): 887–900. https://doi.org/10.1038/s41564-020-0715-z.

Baker, Brett J., Gene W. Tyson, Richard I. Webb, Judith Flanagan, Philip Hugenholtz, Eric E. Allen, and Jillian F. Banfield. 2006. “Lineages of Acidophilic Archaea Revealed by Community Genomic Analysis.” *Science* 314 (5807): 1933–35. https://doi.org/10.1126/science.1132690.

Baker, Brittany A., Ana Gutierrez-Preciado, Alvaro Rodriguez del Rio, Charley McCarthy, Purificacion Lopez-Garcia, Jaime Huerta-Cepas, Edward Susko, Andrew J. Roger, Laura Eme, and David Moreira. 2023. “Several Independent Adaptations of Archaea to Hypersaline Environments.” bioRxiv. https://doi.org/10.1101/2023.07.03.547478.

Barns, S M, C F Delwiche, J D Palmer, and N R Pace. 1996. “Perspectives on Archaeal Diversity, Thermophily and Monophyly from Environmental rRNA Sequences.” *Proceedings of the National Academy of Sciences* 93 (17): 9188–93. https://doi.org/10.1073/pnas.93.17.9188.

Bergsten, Johannes. 2005. “A Review of Long-Branch Attraction.” *Cladistics* 21 (2): 163–93. https://doi.org/10.1111/j.1096-0031.2005.00059.x.

Bohlin, Jon, Brittany Rose, Ola Brynildsrud, and Birgitte Freiesleben De Blasio. 2020. “A Simple

Stochastic Model Describing Genomic Evolution over Time of GC Content in Microbial Symbionts." *Journal of Theoretical Biology* 503 (October): 110389. https://doi.org/10.1016/j.jtbi.2020.110389.

Brochier, Celine, Simonetta Gribaldo, Yvan Zivanovic, Fabrice Confalonieri, and Patrick Forterre. 2005. "Nanoarchaea: Representatives of a Novel Archaeal Phylum or a Fast-Evolving Euryarchaeal Lineage Related to Thermococcales?" *Genome Biology* 6 (5): R42. https://doi.org/10.1186/gb-2005-6-5-r42.

Brown, Christopher T., Laura A. Hug, Brian C. Thomas, Itai Sharon, Cindy J. Castelle, Andrea Singh, Michael J. Wilkins, Kelly C. Wrighton, Kenneth H. Williams, and Jillian F. Banfield. 2015. "Unusual Biology across a Group Comprising More than 15% of Domain Bacteria." *Nature* 523 (7559): 208–11. https://doi.org/10.1038/nature14486.

Castelle, Cindy J., and Jillian F. Banfield. 2018. "Major New Microbial Groups Expand Diversity and Alter Our Understanding of the Tree of Life." *Cell* 172 (6): 1181–97. https://doi.org/10.1016/j.cell.2018.02.016.

Castelle, Cindy J., Christopher T. Brown, Karthik Anantharaman, Alexander J. Probst, Raven H. Huang, and Jillian F. Banfield. 2018. "Biosynthetic Capacity, Metabolic Variety and Unusual Biology in the CPR and DPANN Radiations." *Nature Reviews Microbiology* 16 (10): 629–45. https://doi.org/10.1038/s41579-018-0076-2.

Castelle, Cindy J., Raphaël Méheust, Alexander L. Jaffe, Kiley Seitz, Xianzhe Gong, Brett J. Baker, and Jillian F. Banfield. 2021. "Protein Family Content Uncovers Lineage Relationships and Bacterial Pathway Maintenance Mechanisms in DPANN Archaea." *Frontiers in Microbiology* 12. https://doi.org/10.3389/fmicb.2021.660052.

Castelle, Cindy J., Kelly C. Wrighton, Brian C. Thomas, Laura A. Hug, Christopher T. Brown, Michael J. Wilkins, Kyle R. Frischkorn, et al. 2015. "Genomic Expansion of Domain Archaea Highlights Roles for Organisms from New Phyla in Anaerobic Carbon Cycling." *Current Biology* 25 (6): 690–701. https://doi.org/10.1016/j.cub.2015.01.014.

Chatterjee, Shruti, Aditya J Basak, Asha V Nair, Kheerthana Duraivelan, and Dibyendu Samanta. 2021. "Immunoglobulin-Fold Containing Bacterial Adhesins: Molecular and Structural Perspectives in Host Tissue Colonization and Infection." *FEMS Microbiology Letters* 368 (2): fnaa220. https://doi.org/10.1093/femsle/fnaa220.

Chen, Lin-Xing, Celia Méndez-García, Nina Dombrowski, Luis E. Servín-Garcidueñas, Emiley A. Eloe-Fadrosh, Bao-Zhu Fang, Zhen-Hao Luo, et al. 2018. "Metabolic Versatility of Small Archaea Micrarchaeota and Parvarchaeota." *The ISME Journal* 12 (3): 756–75. https://doi.org/10.1038/s41396-017-0002-z.

Criscuolo, Alexis, and Simonetta Gribaldo. 2010. "BMGE (Block Mapping and Gathering with Entropy): A New Software for Selection of Phylogenetic Informative Regions from Multiple Sequence Alignments." *BMC Evolutionary Biology* 10 (1): 210. https://doi.org/10.1186/1471-2148-10-210.

Dombrowski, Nina, Jun-Hoe Lee, Tom A. Williams, Pierre Offre, and Anja Spang. 2019. "Genomic Diversity, Lifestyles and Evolutionary Origins of DPANN Archaea." *FEMS Microbiology Letters* 366 (2). https://doi.org/10.1093/femsle/fnz008.

Dombrowski, Nina, Tom A. Williams, Jiarui Sun, Benjamin J. Woodcroft, Jun-Hoe Lee, Bui Quang Minh, Christian Rinke, and Anja Spang. 2020. "Undinarchaeota Illuminate DPANN Phylogeny and the Impact of Gene Transfer on Archaeal Evolution." *Nature Communications* 11 (1): 3939. https://doi.org/10.1038/s41467-020-17408-w.

Eme, Laura, Daniel Tamarit, Eva F. Caceres, Courtney W. Stairs, Valerie De Anda, Max E. Schön, Kiley W. Seitz, et al. 2023. "Inference and Reconstruction of the Heimdallarchaeial Ancestry of Eukaryotes." *Nature* 618 (7967): 992–99. https://doi.org/10.1038/s41586-023-06186-2.

Emms, David M., and Steven Kelly. 2019. "OrthoFinder: Phylogenetic Orthology Inference for Comparative Genomics." *Genome Biology* 20 (1): 238.

https://doi.org/10.1186/s13059-019-1832-y.
Esser, Sarah P., Janina Rahlff, Weishu Zhao, Michael Predl, Julia Plewka, Katharina Sures, Franziska Wimmer, et al. 2023. “A Predicted CRISPR-Mediated Symbiosis between Uncultivated Archaea.” *Nature Microbiology* 8 (9): 1619–33. https://doi.org/10.1038/s41564-023-01439-2.
Felsenstein, Joseph. 1978. “Cases in Which Parsimony or Compatibility Methods Will Be Positively Misleading.” *Systematic Zoology* 27 (4): 401–10. https://doi.org/10.2307/2412923.
Feng, Yutian, Uri Neri, Sean Gosselin, Artemis S Louyakis, R Thane Papke, Uri Gophna, and Johann Peter Gogarten. 2021. “The Evolutionary Origins of Extreme Halophilic Archaeal Lineages.” *Genome Biology and Evolution* 13 (8): evab166. https://doi.org/10.1093/gbe/evab166.
Golyshina, Olga V., Rafael Bargiela, Stepan V. Toshchakov, Nikolay A. Chernyh, Soshila Ramayah, Aleksei A. Korzhenkov, Ilya V. Kublanov, and Peter N. Golyshin. 2019. “Diversity of ‘Ca. Micrarchaeota’ in Two Distinct Types of Acidic Environments and Their Associations with Thermoplasmatales.” *Genes* 10 (6). https://doi.org/10.3390/genes10060461.
Golyshina, Olga V., Stepan V. Toshchakov, Kira S. Makarova, Sergey N. Gavrilov, Aleksei A. Korzhenkov, Violetta La Cono, Erika Arcadi, et al. 2017a. “‘ARMAN’ Archaea Depend on Association with Euryarchaeal Host in Culture and in Situ.” *Nature Communications* 8 (July). https://doi.org/10.1038/s41467-017-00104-7.
2017. “‘ARMAN’ Archaea Depend on Association with Euryarchaeal Host in Culture and in Situ.” *Nature Communications* 8 (July): 60. https://doi.org/10.1038/s41467-017-00104-7.
Hallgren, Jeppe, Konstantinos D. Tsirigos, Mads Damgaard Pedersen, José Juan Almagro Armenteros, Paolo Marcatili, Henrik Nielsen, Anders Krogh, and Ole Winther. 2022. “DeepTMHMM Predicts Alpha and Beta Transmembrane Proteins Using Deep Neural Networks.” bioRxiv. https://doi.org/10.1101/2022.04.08.487609.
Hamm, Joshua N., Susanne Erdmann, Emiley A. Eloe-Fadrosh, Allegra Angeloni, Ling Zhong, Christopher Brownlee, Timothy J. Williams, et al. 2019. “Unexpected Host Dependency of Antarctic Nanohaloarchaeota.” *Proceedings of the National Academy of Sciences* 116 (29): 14661–70.
Huber, Harald, Michael J. Hohn, Reinhard Rachel, Tanja Fuchs, Verena C. Wimmer, and Karl O. Stetter. 2002. “A New Phylum of Archaea Represented by a Nanosized Hyperthermophilic Symbiont.” *Nature* 417 (6884): 63–67. https://doi.org/10.1038/417063a.
Huber, Harald, Michael J Hohn, Karl O Stetter, and Reinhard Rachel. 2003. “The Phylum Nanoarchaeota: Present Knowledge and Future Perspectives of a Unique Form of Life.” *Research in Microbiology* 154 (3): 165–71. https://doi.org/10.1016/S0923-2508(03)00035-4.
Hug, Laura A., Brett J. Baker, Karthik Anantharaman, Christopher T. Brown, Alexander J. Probst, Cindy J. Castelle, Cristina N. Butterfield, et al. 2016. “A New View of the Tree of Life.” *Nature Microbiology* 1 (5): 1–6. https://doi.org/10.1038/nmicrobiol.2016.48.
Jaffe, Alexander L., Cindy J. Castelle, Christopher L. Dupont, and Jillian F. Banfield. 2019. “Lateral Gene Transfer Shapes the Distribution of RuBisCO among Candidate Phyla Radiation Bacteria and DPANN Archaea.” *Molecular Biology and Evolution* 36 (3): 435–46. https://doi.org/10.1093/molbev/msy234.
Jones, Philip, David Binns, Hsin-Yu Chang, Matthew Fraser, Weizhong Li, Craig McAnulla, Hamish McWilliam, et al. 2014. “InterProScan 5: Genome-Scale Protein Function Classification.” *Bioinformatics* 30 (9): 1236–40. https://doi.org/10.1093/bioinformatics/btu031.

Kadnikov, Vitaly V., Alexander S. Savvichev, Andrey V. Mardanov, Alexey V. Beletsky, Artem V. Chupakov, Natalia M. Kokryatskaya, Nikolay V. Pimenov, and Nikolai V. Ravin. 2020. “Metabolic Diversity and Evolutionary History of the Archaeal Phylum ‘Candidatus Micrarchaeota’ Uncovered from a Freshwater Lake Metagenome.” *Applied and Environmental Microbiology* 86 (23). https://doi.org/10.1128/AEM.02199-20.

Katoh, Kazutaka, and Daron M. Standley. 2013. “MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability.” *Molecular Biology and Evolution* 30 (4): 772–80. https://doi.org/10.1093/molbev/mst010.

Kizina, Jana, Sebastian F. A. Jordan, Gerrit Alexander Martens, Almud Lonsing, Christina Probian, Androniki Kolovou, Rachel Santarella-Mellwig, et al. 2022. “Methanosaeta and ‘Candidatus Velamenicoccus Archaeovorus.’” *Applied and Environmental Microbiology* 88 (7): e02407-21. https://doi.org/10.1128/aem.02407-21.

Kram, Karin E., Christopher Geiger, Wazim Mohammed Ismail, Heewook Lee, Haixu Tang, Patricia L. Foster, and Steven E. Finkel. 2017. “Adaptation of Escherichia Coli to Long-Term Serial Passage in Complex Medium: Evidence of Parallel Evolution.” *mSystems* 2 (2): e00192-16. https://doi.org/10.1128/mSystems.00192-16.

Krause, Susanne, Andreas Bremges, Philipp C. Münch, Alice C. McHardy, and Johannes Gescher. 2017. “Characterisation of a Stable Laboratory Co-Culture of Acidophilic Nanoorganisms.” *Scientific Reports* 7 (1): 3289. https://doi.org/10.1038/s41598-017-03315-6.

La Cono, Violetta, Enzo Messina, Manfred Rohde, Erika Arcadi, Sergio Ciordia, Francesca Crisafi, Renata Denaro, et al. 2020. “Symbiosis between Nanohaloarchaeon and Haloarchaeon Is Based on Utilization of Different Polysaccharides.” *Proceedings of the National Academy of Sciences* 117 (33): 20223–34. https://doi.org/10.1073/pnas.2007232117.

López-García, Purificación, and David Moreira. 2021. “Physical Connections: Prokaryotes Parasitizing Their Kin.” *Environmental Microbiology Reports* 13 (1): 54–61. https://doi.org/10.1111/1758-2229.12910.

Matte-Tailliez, Oriane, Céline Brochier, Patrick Forterre, and Hervé Philippe. 2002. “Archaeal Phylogeny Based on Ribosomal Proteins.” *Molecular Biology and Evolution* 19 (5): 631–39. https://doi.org/10.1093/oxfordjournals.molbev.a004122.

Menardo, Fabrizio, Chloé Loiseau, Daniela Brites, Mireia Coscolla, Sebastian M. Gygli, Liliana K. Rutaihwa, Andrej Trauner, Christian Beisel, Sonia Borrell, and Sebastien Gagneux. 2018. “Treemmer: A Tool to Reduce Large Phylogenetic Datasets with Minimal Loss of Diversity.” *BMC Bioinformatics* 19 (1): 164. https://doi.org/10.1186/s12859-018-2164-8.

Minh, Bui Quang, Heiko A Schmidt, Olga Chernomor, Dominik Schrempf, Michael D Woodhams, Arndt von Haeseler, and Robert Lanfear. 2020. “IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era.” *Molecular Biology and Evolution* 37 (5): 1530–34. https://doi.org/10.1093/molbev/msaa015.

Moody, Edmund RR, Tara A Mahendrarajah, Nina Dombrowski, James W Clark, Celine Petitjean, Pierre Offre, Gergely J Szöllősi, Anja Spang, and Tom A Williams. 2022. “An Estimate of the Deepest Branches of the Tree of Life from Ancient Vertically-Evolving Genes.” Edited by George H Perry. *eLife* 11 (February): e66695. https://doi.org/10.7554/eLife.66695.

Moran, Nancy A., and Gordon M. Bennett. 2014. “The Tiniest Tiny Genomes.” *Annual Review of Microbiology* 68: 195–215. https://doi.org/10.1146/annurev-micro-091213-112901.

Muñoz-Gómez, Sergio A., Edward Susko, Kelsey Williamson, Laura Eme, Claudio H. Slamovits, David Moreira, Purificación López-García, and Andrew J. Roger. 2022. “Site-and-Branch-Heterogeneous Analyses of an Expanded Dataset Favour Mitochondria as Sister to Known Alphaproteobacteria.” *Nature Ecology & Evolution* 6 (3): 253–62. https://doi.org/10.1038/s41559-021-01638-2.

Narasingarao, Priya, Sheila Podell, Juan A Ugalde, Céline Brochier-Armanet, Joanne B Emerson, Jochen J Brocks, Karla B Heidelberg, Jillian F Banfield, and Eric E Allen. 2012. “De Novo Metagenomic Assembly Reveals Abundant Novel Major Lineage of Archaea in Hypersaline Microbial Communities.” *The ISME Journal* 6 (1): 81–93. https://doi.org/10.1038/ismej.2011.78.

Pace, Norman R. 1997. “A Molecular View of Microbial Diversity and the Biosphere.” *Science* 276 (5313): 734–40. https://doi.org/10.1126/science.276.5313.734.

Petitjean, Céline, Philippe Deschamps, Purificación López-García, and David Moreira. 2015. “Rooting the Domain Archaea by Phylogenomic Analysis Supports the Foundation of the New Kingdom Proteoarchaeota.” *Genome Biology and Evolution* 7 (1): 191–204. https://doi.org/10.1093/gbe/evu274.

Petitjean, Céline, Philippe Deschamps, Purificación López-García, David Moreira, and Céline Brochier-Armanet. 2015. “Extending the Conserved Phylogenetic Core of Archaea Disentangles the Evolution of the Third Domain of Life.” *Molecular Biology and Evolution* 32 (5): 1242–54. https://doi.org/10.1093/molbev/msv015.

Philippe, H, P Lopez, H Brinkmann, K Budin, A Germot, J Laurent, D Moreira, M Müller, and H Le Guyader. 2000. “Early-Branching or Fast-Evolving Eukaryotes? An Answer Based on Slowly Evolving Positions.” *Proceedings of the Royal Society B: Biological Sciences* 267 (1449): 1213–21.

Podar, Mircea, Iain Anderson, Kira S. Makarova, James G. Elkins, Natalia Ivanova, Mark A. Wall, Athanasios Lykidis, et al. 2008. “A Genomic Analysis of the Archaeal System Ignicoccus Hospitalis-Nanoarchaeum Equitans.” *Genome Biology* 9 (11): R158. https://doi.org/10.1186/gb-2008-9-11-r158.

Podar, Mircea, Kira S. Makarova, David E. Graham, Yuri I. Wolf, Eugene V. Koonin, and Anna-Louise Reysenbach. 2013. “Insights into Archaeal Evolution and Symbiosis from the Genomes of a Nanoarchaeon and Its Inferred Crenarchaeal Host from Obsidian Pool, Yellowstone National Park.” *Biology Direct* 8 (1): 9. https://doi.org/10.1186/1745-6150-8-9.

Price, Morgan N., Paramvir S. Dehal, and Adam P. Arkin. 2010. “FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments.” *PLOS ONE* 5 (3): e9490. https://doi.org/10.1371/journal.pone.0009490.

Probst, Alexander J., and Christine Moissl-Eichinger. 2015. “‘Altiarchaeales’: Uncultivated Archaea from the Subsurface.” *Life* 5 (2): 1381–95. https://doi.org/10.3390/life5021381.

Probst, Alexander J., Thomas Weinmaier, Kasie Raymann, Alexandra Perras, Joanne B. Emerson, Thomas Rattei, Gerhard Wanner, et al. 2014. “Biology of a Widespread Uncultivated Archaeon That Contributes to Carbon Fixation in the Subsurface.” *Nature Communications* 5 (1): 5497. https://doi.org/10.1038/ncomms6497.

Qu, Yan-Ni, Yang-Zhi Rao, Yan-Ling Qi, Yu-Xian Li, Andrew Li, Marike Palmer, Brian P. Hedlund, et al. 2023. “Panguiarchaeum Symbiosum, a Potential Hyperthermophilic Symbiont in the TACK Superphylum.” *Cell Reports* 42 (3): 112158. https://doi.org/10.1016/j.celrep.2023.112158.

Ramulu, Hemalatha Golaconda, Mathieu Groussin, Emmanuel Talla, Remi Planel, Vincent Daubin, and Céline Brochier-Armanet. 2014. “Ribosomal Proteins: Toward a next Generation Standard for Prokaryotic Systematics?” *Molecular Phylogenetics and Evolution* 75 (June): 103–17. https://doi.org/10.1016/j.ympev.2014.02.013.

Raymann, Kasie, Céline Brochier-Armanet, and Simonetta Gribaldo. 2015. “The Two-Domain Tree of Life Is Linked to a New Root for the Archaea.” *Proceedings of the National Academy of Sciences* 112 (21): 6670–75. https://doi.org/10.1073/pnas.1420858112.

Rinke, Christian, Maria Chuvochina, Aaron J. Mussig, Pierre-Alain Chaumeil, Adrián A. Davín, David W. Waite, William B. Whitman, Donovan H. Parks, and Philip Hugenholtz. 2021. “A Standardized Archaeal Taxonomy for the Genome Taxonomy Database.” *Nature*

Microbiology 6 (7): 946–59. https://doi.org/10.1038/s41564-021-00918-8.
Rinke, Christian, Patrick Schwientek, Alexander Sczyrba, Natalia N. Ivanova, Iain J. Anderson, Jan-Fang Cheng, Aaron Darling, et al. 2013. “Insights into the Phylogeny and Coding Potential of Microbial Dark Matter.” *Nature* 499 (7459): 431–37. https://doi.org/10.1038/nature12352.
Rocha, Eduardo P. C., and Antoine Danchin. 2002. “Base Composition Bias Might Result from Competition for Metabolic Resources.” *Trends in Genetics* 18 (6): 291–94. https://doi.org/10.1016/S0168-9525(02)02690-2.
Sakai, Hiroyuki D., Naswandi Nur, Shingo Kato, Masahiro Yuki, Michiru Shimizu, Takashi Itoh, Moriya Ohkuma, Antonius Suwanto, and Norio Kurosawa. 2022. “Insight into the Symbiotic Lifestyle of DPANN Archaea Revealed by Cultivation and Genome Analyses.” *Proceedings of the National Academy of Sciences* 119 (3). https://doi.org/10.1073/pnas.2115449119.
Schwank, Katrin, Till L. V. Bornemann, Nina Dombrowski, Anja Spang, Jillian F. Banfield, and Alexander J. Probst. 2019. “An Archaeal Symbiont-Host Association from the Deep Terrestrial Subsurface.” *The ISME Journal* 13 (8): 2135–39. https://doi.org/10.1038/s41396-019-0421-0.
Seymour, Cale O., Marike Palmer, Eric D. Becraft, Ramunas Stepanauskas, Ariel D. Friel, Frederik Schulz, Tanja Woyke, et al. 2023. “Hyperactive Nanobacteria with Host-Dependent Traits Pervade Omnitrophota.” *Nature Microbiology* 8 (4): 727–44. https://doi.org/10.1038/s41564-022-01319-1.
Shimodaira, Hidetoshi. 2002. “An Approximately Unbiased Test of Phylogenetic Tree Selection.” *Systematic Biology* 51 (3): 492–508. https://doi.org/10.1080/10635150290069913.
Spang, Anja, Eva F. Caceres, and Thijs J. G. Ettema. 2017. “Genomic Exploration of the Diversity, Ecology, and Evolution of the Archaeal Domain of Life.” *Science* 357 (6351). https://doi.org/10.1126/science.aaf3883.
St. John, Emily, Yitai Liu, Mircea Podar, Matthew B. Stott, Jennifer Meneghin, Zhiqiang Chen, Kirill Lagutin, Kevin Mitchell, and Anna-Louise Reysenbach. 2019. “A New Symbiotic Nanoarchaeote (Candidatus Nanoclepta Minutus) and Its Host (Zestosphaera Tikiterensis Gen. Nov., Sp. Nov.) from a New Zealand Hot Spring.” *Systematic and Applied Microbiology*, Taxonomy of uncultivated Bacteria and Archaea, 42 (1): 94–106. https://doi.org/10.1016/j.syapm.2018.08.005.
Susko, Edward, and Andrew J Roger. 2021. “Long Branch Attraction Biases in Phylogenetics.” *Systematic Biology* 70 (4): 838–43. https://doi.org/10.1093/sysbio/syab001.
Whittaker, Charles A., and Richard O. Hynes. 2002. “Distribution and Evolution of von Willebrand/Integrin A Domains: Widely Dispersed Domains with Roles in Cell Adhesion and Elsewhere.” *Molecular Biology of the Cell* 13 (10): 3369–87. https://doi.org/10.1091/mbc.e02-05-0259.
Williams, Tom A., Gergely J. Szöllősi, Anja Spang, Peter G. Foster, Sarah E. Heaps, Bastien Boussau, Thijs J. G. Ettema, and T. Martin Embley. 2017. “Integrative Modeling of Gene and Genome Evolution Roots the Archaeal Tree of Life.” *Proceedings of the National Academy of Sciences of the United States of America* 114 (23): E4602–11. https://doi.org/10.1073/pnas.1618463114.
Woese, Carl. 1990. “Towards a Natural System of Organisms: Proposal for the Domains Archaea, Bacteria, and Eucarya.” 1990. https://doi.org/10.1073/pnas.87.12.4576.
Wurch, Louie, Richard J. Giannone, Bernard S. Belisle, Carolyn Swift, Sagar Utturkar, Robert L. Hettich, Anna-Louise Reysenbach, and Mircea Podar. 2016. “Genomics-Informed Isolation and Characterization of a Symbiotic Nanoarchaeota System from a Terrestrial Geothermal Environment.” *Nature Communications* 7 (1): 12115. https://doi.org/10.1038/ncomms12115.
Youssef, Noha H, Christian Rinke, Ramunas Stepanauskas, Ibrahim Farag, Tanja Woyke, and

Mostafa S Elshahed. 2015. “Insights into the Metabolism, Lifestyle and Putative Evolutionary History of the Novel Archaeal Phylum ‘Diapherotrites’.” *The ISME Journal* 9 (2): 447–60. https://doi.org/10.1038/ismej.2014.141.

Zhang, Irene H., Benedict Borer, Rui Zhao, Steven Wilbert, Dianne K. Newman, and Andrew R. Babbin. 2023. “Uncultivated DPANN Archaea Are Ubiquitous Inhabitants of Global Oxygen Deficient Zones with Diverse Metabolic Potential.” bioRxiv. https://doi.org/10.1101/2023.10.30.564641.

6. Discussion and perspectives

6. Discussion and perspectives

6.1. The origins of extreme halophily

For over 30 years, extreme halophilic archaea were thought to belong to a single archaeal order, the Halobacteriales (henceforth Haloarchaea) (Andrei et al., 2012; Lanyi, 1974; Oren, 2002). This implied a singular origin of adaptation to hypersaline environments within the archaeal tree. Along with the adaptation to extreme halophily, the proteomes of Haloarchaea were also shown to be enriched in acidic amino acids and depleted in large hydrophobic and basic ones to decrease the hydrophobicity of their proteins and prevent protein aggregation in the face of high intracellular salt concentrations resulting from a 'salt-in' osmotic strategy (Christian & Waltho, 1962; Dennis & Shimmin, 1997; Lanyi, 1974). Since high concentrations of salts in an environment are generally thought to be deleterious to proteins and other macromolecules (Dennis & Shimmin, 1997), it was reasonable to conclude that the adaptation to extreme halophily was an evolutionarily challenging transition, particularly since it only occurred once. However, the discovery of additional extreme halophilic archaeal taxa over the last decade has fundamentally altered this viewpoint (Aouad et al., 2018; Feng et al., 2021; Martijn et al., 2020; Narasingarao et al., 2012a; Sorokin et al., 2017; Zhou et al., n.d.).

Nevertheless, as evident from my thesis's introduction and initial chapter, establishing the evolutionary relationships among these groups has proven challenging. This complexity arises from phylogenetic artifacts, including long-branch attraction and/or shared biases in sequence compositions among extreme halophilic archaea. Additionally, cases of horizontal gene transfers among these halophilic lineages further contribute to the complexities of their evolutionary history (B. A. Baker et al., 2023; Feng et al., 2021; Mongodin et al., 2005). However,

our in-depth phylogenomic analyses of the various extreme halophilic archaeal lineages, plus the addition of two new family-level lineages of extreme halophilic archaea, the Afararchaeaceae and Asbonarchaeaceae, robustly showed that adaptation to extreme halophily in the archaeal tree occurred at least four times independently, suggesting that this transition was not as evolutionarily unique as previously thought. Moreover, with the addition of the new Afararchaeaceae family and their phylogenetic placement as the closest sister group to the Hikarchaeia (Martijn et al., 2020) and Haloarchaea, our analysis also showed that the Hikarchaeia could transition back to a moderate saline environment from an extreme halophilic ancestor. This transition was also marked by the Hikarchaeia returning to a non-acidic proteome. This is somewhat surprising given the fact that extreme halophilic archaea are obligate halophiles because their proteins have been shown to denature and unfold in low salt media due to charge repulsion (Dennis & Shimmin, 1997; Lanyi, 1974; Madern et al., 2000; Müller-Santos et al., 2009). However, one important exception is the Methanonatronarchaeia (Sorokin et al., 2017). While they have experimentally been shown to be 'salt-in' strategists and contain molar concentrations of intracellular salt, their proteomes are noticeably less acidic than the other groups of known salt-in strategists. It is tempting to speculate that the Methanonatronarchaeia might have adopted a previously unknown osmoprotective strategy to prevent protein aggregation in the face of high intracellular salt concentrations. Only five genomes are currently available for the Methanonatronarchaeia class, according to the Genome Taxonomy Database (GTDB) release214 (Rinke et al., 2021). Considering the current genomic data, it is challenging to determine whether the non-acidic proteome observed in this group is unique to extreme halophilic methanogens or if these potentially new adaptive strategies might be present in other archaeal lineages.

6.1.1. Gene tree-species tree reconciliation and the evolution of gene content in halophilic archaea

For many decades it was thought that methanogenic Euryarchaeota were the closest relatives to the Haloarchaea (Adam et al., 2017; Forterre et al., 2002; Petitjean, Deschamps, López-García, Moreira, et al., 2015; Raymann et al., 2015; C. Woese, 1990). It was theorized that a massive genome rearrangement would have been necessary to achieve this evolutionary transition and was thought early on to be facilitated by a single and substantial episode of massive HGT from Bacteria (Kennedy et al., 2001; Nelson-Sathi et al., 2012). However, with enhanced methods and broader taxonomic sampling, current evidence now suggests that this transition was accompanied by multiple cases of HGT that occurred gradually (Groussin et al., 2016; Martijn et al., 2020). In our analysis, we also observed a comparable trend when incorporating a more comprehensive taxonomic sampling of all currently known groups of extreme halophiles (i.e., Haloarchaea, Nanohaloarchaea, Methanonatronarchaea, Haloplasmatales, Afararchaea, and Asbonarchaea).

We employed a recently developed method to analyze gene family evolution, known as a gene tree-species tree reconciliation, as implemented in the amalgamated likelihood estimation (ALE) software (Szöllősi et al., 2013b). In brief, ALE reconciles individual gene trees onto a species tree to estimate the probability that a gene has experienced duplication, transfer, origination, or loss at each ancestral node in that species tree. Unlike previous parsimony-based reconciliation methods, such as NOTUNG (K. Chen et al., 2000) or ecceTerra (Jacox et al., 2016) that require the costs of such events to be specified a priori, one advantage of ALE is that it estimates the rates directly from the data (Szöllősi et al., 2013a). However,

in an ALE analysis, various decisions regarding parameter settings must be made at every data preparation stage, and these choices can greatly affect the final results.

One of the first steps of a gene tree-species tree reconciliation analysis involves building reliable orthologous groups (OGs). This is a challenging task, especially at the scale of the archaeal domain. Several bioinformatic pipelines have been developed to cluster OGs for a given set of taxa, including but not exclusively eggNOG-mapper (Cantalapiedra et al., 2021), OrthoMCL (Li et al., 2003), and OrthoFinder (Emms & Kelly, 2019). However, each software employs a unique algorithm for OG clustering, and many user-defined parameters can significantly change the composition of the OGs. In our analysis, we used OrthoFinder (Emms & Kelly, 2019). We had to make many prior decisions regarding the DIAMOND (Buchfink et al., 2015) parameters like e-value, sequence identity, and query coverage, along with the sensitivity parameters (--very-sensitive and --ultra-sensitive), the inflation parameter (I=1.1, 1.3, or 1.5) and lastly, the software used for reconstructing the single-gene trees used in refining the OGs (Price et al., 2010). When OGs are excessively split, the ALE analysis may deduce more originations for specific branches.

Conversely, if OGs are not split adequately, the analysis might indicate a complex history for that family involving multiple HGTs and losses across the tree. Though these decisions are not unique to OrthoFinder, they exemplify the multitude of a priori choices that can influence the outcome of the ALE results. In our analysis, the broad taxonomic sampling across the entire archaeal domain made it challenging to identify parameters that would suit diverse taxonomic groups with varying rates of genome evolution. For example, we wanted to identify fast-evolving genes common to the DPANN archaea. However, we needed to be

careful not to be too relaxed in our clustering criteria to avoid clustering distantly related paralogs from other archaeal taxa.

Once the OGs are defined, the next data-preparation step involves reconstructing phylogenetic trees for each OG. This is arguably the most important data-preparation step for an ALE analysis but can be extremely computationally demanding. In our dataset, which comprises around 200 archaeal taxa, our analysis yielded approximately 17,000 OGs, which translated to 17,000 single-gene trees (SGTs). ALE recommends using a subset of trees generated from a Bayesian Markov Chain Monte Carlo (MCMC) program, such as PhyloBayes (Lartillot et al., 2009) or mrBayes (Huelsenbeck & Ronquist, 2001). However, for many analyses, this approach is computationally impractical. ALE states that the sample of gene trees for each OG may also be obtained using bootstrap replicates from an ML analysis, but it is technically less correct (https://github.com/ssolo/ALE). ML phylogenetic analyses are generally less computationally demanding than Bayesian analyses. However, complex mixture models in an ML framework, like C10-C60, can extend the computational time for the largest OGs, for example, those containing over 2,000 sequences, to several weeks or months. Although this can vary greatly depending on your access to computational resources, this is an important consideration when deciding whether to do an ALE analysis. In our analysis, completing SGTs for 17,000 OGs took approximately six months. While this is not inherently a limitation to the analysis, it does constrain the ability to incorporate new taxa if important lineages to an analysis become available.

Moreover, even when the most sophisticated ML models of sequence evolution are used to reconstruct SGTs, it has long been shown that the phylogenetic signal of those trees can be unreliable due to the minimal phylogenetic signal contained in single proteins (Forterre et al., 2002; Gogarten &

Townsend, 2005). ALE is currently the most sophisticated method available to gain insight into the broad evolutionary patterns of gene families. However, it is important to remember that each step of the data preparation can influence the outcome of the final results, and the results are only as reliable as your reconstructed SGTs.

6.2. The origins of the DPANN archaea

The DPANN archaea are one of the four proposed archaeal superphyla, along with the TACK, Asgard, and Euryarchaeota. Yet, their monophyly and position within the archaeal tree remain a matter of debate (Adam et al., 2017; Brochier et al., 2005; Dombrowski et al., 2019, 2020; Feng et al., 2021; Huber et al., 2002; Moody et al., 2022; Petitjean, Deschamps, López-García, & Moreira, 2015; Rinke et al., 2013; Waters et al., 2003). The uncertainty surrounding the position of the DPANN stems from the instability of phylogenies that only include individual DPANN lineages, such as the Nanoarchaeota (Brochier et al., 2005; Huber et al., 2002), Nanohaloarchaeota (Narasingarao et al., 2012a; Petitjean, Deschamps, López-García, & Moreira, 2015), Micrarchaeota (Petitjean, Deschamps, López-García, & Moreira, 2015), and Altiarchaeota (Adam et al., 2017; Probst & Moissl-Eichinger, 2015). Initial phylogenetic studies suggested that the instability observed for individual DPANN lineages might be addressed by introducing additional genomic data (Castelle et al., 2015; Rinke et al., 2013), which has been shown to help resolve unstable evolutionary relationships (Graybeal, 1998; Hedtke et al., 2006; Hillis et al., 2003). However, as more lineages were described, subsequent studies argued that the introduction of additional taxa (also long-branching and compositionally biased) could compound the impact of

phylogenetic artifacts among these lineages (Aouad et al., 2018; Petitjean, Deschamps, López-García, & Moreira, 2015).

Identifying potential phylogenetic artifacts influencing the phylogeny of the DPANN can be challenging due to the potential co-occurrence of multiple artifacts arising simultaneously. For example, when constructing a principal component analysis (PCA) on the amino acid frequencies of all archaea, unique patterns emerge among various DPANN lineages compared to other archaea (see DPANN thesis chapter for more details). However, these patterns are not conserved across all DPANN lineages. Notably, the Nanohaloarchaeota exhibit a bias towards acidic amino acids in response to strong environmental pressures (see Halophiles thesis chapter for more details). In contrast, other DPANN lineages display a preference for F, I, M, N, K, and Y amino acids, attributed to genome reduction and enrichment in A+T nucleotides (Clark et al., 1999; Gu et al., 1998; Muñoz-Gómez et al., 2019, 2022).

Since amino acid biases are not universally conserved across all DPANN, a general alignment treatment, like removing the fastest-evolving sites, is suboptimal. Nevertheless, removing the fastest-evolving sites has been suggested as a potential solution to mitigate artifacts caused by LBA (Dombrowski et al., 2020; Williams et al., 2017). However, as amino acid preferences, the rate of genome evolution varies among DPANN lineages, with lineages such as the Nanohaloarchaeota and Huberarchaeota showing a faster rate of genome evolution compared to the Micrarchaeota and Iainarchaeota based on relative branch lengths. Lastly, it is also challenging to select a reliable set of phylogenetic markers for placing the DPANN in the archaeal tree due to high rates of horizontal gene transfers among DPANN lineages and their hosts, as well as with other diverse archaea and bacteria (Dombrowski et al., 2020; Rinke et al., 2013; Waters et al., 2003).

In our phylogenomic analysis of the DPANN, we systematically addressed each of the aforementioned phylogenetic artifacts. We performed multiple rounds of manual curation on our set of phylogenetic markers to eliminate any potential HGT, paralogs, and/or contamination cases. To deal with compositionally biased sites, we implemented a ranking scheme to identify and remove the most biased sites influenced by GARP/FIMNKY preferences. We also implemented the GF-mix model (Muñoz-Gómez et al., 2022) to model the GARP/FIMNKY preferences instead of removing them. GFmix is a site-heterogeneous mixture model that adjusts amino acid frequencies for each class of the mixture model in a branch-specific manner to accommodate shifts in amino acid composition over the branch. This model refines the likelihood of existing phylogenetic trees rather than constructing new ones. Although branch-heterogeneity models hold promise in phylogenetics, current computational constraints hinder their full implementation. We also selected the "least-biased" DPANN taxa based on our PCA analyses of amino acid frequencies. To mitigate potential artifacts from LBA, we conducted a chimera analysis by selecting the slowest-evolving marker genes for each DPANN phylum. These were considered chimeric sequences because markers from different DPANN taxa within a single phylum were concatenated into the chimera supermatrix. Last, we placed individual DPANN lineages, such as the Nanoarchaeota, Micrarchaeota, Nanohaloarchaeota, and Altiarchaeota, within the archaeal tree and also conducted the chimera analysis on these specific groups. Together, these results robustly support the monophyly of the DPANN archaea and place them well nested within the Euryarchaeota when rooted on the TACK/Asgard branch. Recent studies that recover the archaeal root between DPANN and other archaea (Castelle et al., 2015; Dombrowski et al., 2020; Saw et al., 2015; Spang et al., 2013; Williams et al., 2017), also routinely recover the Euryarchaeota as monophyletic. When we root our trees

on the DPANN branch, we always recover the Euryarchaeota as paraphyletic. Previous analyses that do recover the Euryarchaeota as paraphyletic do not include the DPANN in their taxonomic sampling and suggest that the archaeal root lies within the Euryarchaeota (Adam et al., 2017; Raymann et al., 2015)

Our results suggest that the DPANN archaea evolved through genome-reduction from a free-living Euryarchaeota-like ancestor. This ancestral trait is further reinforced by the position of the free-living Altiarchaeota at the base of the DPANN clade. Previous phylogenetic analyses have either placed the Altiarchaeota as a very deep-branching lineage within the Euryarchaeota (Adam et al., 2017), nested within the Euryarchaeota (Probst & Moissl-Eichinger, 2015), sister to all other DPANN (Spang et al., 2017) or within the DPANN as a sister group to the Micrarchaeota and Iainarchaeota (previously named Diapherotrites) (Castelle & Banfield, 2018; Dombrowski et al., 2020; Moody et al., 2022). If all other DPANN lineages are truly obligate episymbionts, the position of the Altiarchaeota at the base of the DPANN would suggest that the transition from a free-living ancestor to a host-dependent episymbiont occurred only once at the base of the DPANN clade after the divergence from the Altiarchaeota. However, the ubiquity of a host-dependent lifestyle remains a central open question that needs to be confirmed experimentally for most known DPANN phyla.

Most of what is known about the DPANN is based solely on genomic data due to the difficulty of establishing cultures for various DPANN lineages (Dombrowski et al., 2019). One interesting example is the presence of the ATP synthase in some DPANN lineages based on genomic data but the absence of an electron transport chain (ETC) (Mahendrarajah et al., 2023). Without an ETC generating a proton motive force, it is unclear how these ATP synthases can produce ATP (Castelle et al., 2018). It is still an open question whether or not these

ATP synthases produce ATP or if these lineages have adapted this complex for another use. Interestingly, a comparable pattern of ATP synthase presence and absence (Castelle et al., 2018; Mahendrarajah et al., 2023) has also been observed in the bacterial superphylum known as the Candidate Phyla Radiation (CPR; (Hug et al., 2016)) or the Patescibacteria (Rinke et al., 2013).

6.3. Convergent evolution to a symbiotic lifestyle in the DPANN archaea and CPR bacteria?

In the bacterial domain, the Patescibacteria exhibits several similarities with the DPANN archaea (López-García & Moreira, 2021), as it contains species with small cells, reduced genomes and metabolic repertoires, and presumed to be obligate symbionts of other bacteria (C. T. Brown et al., 2015; Castelle & Banfield, 2018). Like the DPANN archaea, the phylogenetic position of the Patescibacteria has also been a matter of debate. Initial phylogenetic analyses placed them at the base of all other bacteria (Bokhari et al., 2020; C. T. Brown et al., 2015; Castelle & Banfield, 2018; Hug et al., 2016; Méheust et al., 2019; Zhu et al., 2019), while more recent analyses suggest that they are nested within the Terrabacteria, as a sister clade to Chloroflexota and Dormibacterota (Coleman et al., 2021; Martinez-Gutierrez & Aylward, 2021; Moody et al., 2022; Taib et al., 2020). The position of the Patescibacteria nested within the Terrabacteria suggests that they evolved reductively from a free-living ancestor (Coleman et al., 2021). The phylogenomic placement of the DPANN nested within the Euryarchaeota suggests convergent evolution towards genome reduction in both the bacterial and archaeal domains. Interestingly, our examination of protein families exclusive to DPANN and absent in other archaea showed that three of the eleven identified proteins are also

found in diverse bacteria. Specifically, the phylogeny of two of these families (OG2458 and OG2651) suggested that they were acquired by DPANN from a patescibacterial donor by an ancient HGT event. In addition, there were also several bacterial sequences intermixed with the DPANN ones, which most likely reflect more recent cases of HGT. Additionally, the DPANN sequences of a third protein family (OG2142) were shown to be closely related to bacterial sequences from the phylum Omnitrophota. Like the Patescibacteria, Omnitrophota seems to correspond to small cells with reduced genomes and a gene content suggesting a symbiotic lifestyle (Seymour et al., 2023). These results support the idea that DPANN archaea obtained some of their most ancient and widespread genes by HGT from these two groups of host-associated bacteria. Interestingly, a member of the Omnitrophota, '*Candidatus* Velamenicoccous archaeavorus', can parasitize methanogenic archaea of the genus *Methanosaeta* (Kizina et al., 2022). This opens the intriguing possibility that some Omnitrophota and DPANN archaea may share similar archaeal hosts, which could have facilitated the exchange of genes between these two distant lineages. Interestingly, two recent papers have identified giant proteins (≥30,000 amino acids) in several genomes of Omnitrophota (Kizina et al., 2022; West-Roberts et al., 2023). A more in-depth *in silico* analysis of these giant proteins suggests they may enable prey adhesion and cell wall digestion during bacterial predation (West-Roberts et al., 2023). Interestingly, large proteins have also been suggested to mediate the interactions between DPANN archaea and their hosts (J. N. Hamm et al., 2019). These large proteins in DPANN are considerably smaller (~8,000 amino acids) than the giant proteins in Omnitrophota (≥30,000 amino acids) but are the second largest protein described for Archaea (J. N. Hamm et al., 2019). Giant proteins have yet to be identified in the genomes of Patescibacteria. However, it has been suggested that these giant proteins might be

more common than previously thought but may be obscured by assembly fragmentation (West-Roberts et al., 2023). It is tempting to speculate that giant proteins have independently evolved several times within bacterial and archaeal symbiotic lineages.

6.4. Rooting the archaeal tree

The placement of the archaeal root remains a major unanswered question. Currently, there are three main hypotheses for its placement: 1) between Euryarchaeota and the rest of the archaea (Petitjean, Deschamps, López-García, & Moreira, 2015; C. Woese, 1990), 2) within the Euryarchaeota (Adam et al., 2017; Raymann et al., 2015), and 3) between DPANN and the rest of archaea (Castelle et al., 2015; Rinke et al., 2013; Spang et al., 2017; Williams et al., 2017). The position of the archaeal root is particularly pertinent for understanding the evolution of the DPANN, as well as to characterize the last archaeal common ancestor (LACA). If the DPANN place at the base of the archaeal tree, this could mean that some of these symbiotic lineages evolved before their archaeal host. Nevertheless, if this holds, there is also the possibility that bacteria could act as hosts for DPANN. However, to date, no such cases have yet been reported.

Generally, two main methods have been used to place the archaeal root. One is to use bacterial sequences as an outgroup (Adam et al., 2017; Castelle et al., 2015; Petitjean, Deschamps, López-García, & Moreira, 2015; Raymann et al., 2015; Rinke et al., 2013; Spang et al., 2017; C. Woese, 1990), and a second is to use a gene tree-species tree reconciliation approach (Williams et al., 2017). While outgroup rooting has been the predominant approach, this method can be plagued with phylogenetic artifacts, such as LBA (Bergsten, 2005, 2005; Philippe & Laurent, 1998;

Shavit et al., 2007) due to the large evolutionary distances between the archaeal and bacterial domains (Petitjean, Deschamps, López-García, & Moreira, 2015; Williams et al., 2017). This is particularly problematic when the DPANN are included in these analyses due to the frequent placement of DPANN lineages at the end of long branches compared to other neighboring taxa (Dombrowski et al., 2019). Previous studies have suggested that the basal position of the DPANN is a result of LBA when the trees are rooted with bacterial sequences (Adam et al., 2017; Petitjean, Deschamps, López-García, & Moreira, 2015; Raymann et al., 2015). An analogous observation in the bacterial domain further reinforces this idea. When archaeal sequences were used to root the bacterial domain, the Patescibacteria, with similarly long branches, were placed at the base of the tree (Bokhari et al., 2020; C. T. Brown et al., 2015; Castelle & Banfield, 2018; Hug et al., 2016; Méheust et al., 2019; Zhu et al., 2019). However, when an outgroup-free rooting method (gene tree-species tree reconciliation) was used, the Patescibacteria moved to a nested position within the Terrabacteria (Coleman et al., 2021). Surprisingly, when a gene tree-species tree reconciliation approach was applied to the archaeal domain, the root placement between DPANN and all other Archaea recovered the highest likelihood from the DTL modeling (Williams et al., 2017). However, while a gene tree-species tree reconciliation approach avoids potential LBA artifacts that can result from using an outgroup rooting approach, as mentioned previously (see Discussion section Gene tree-species tree reconciliation and the evolution of gene content in halophilic archaea), these analyses can induce their own artifacts during the data preparation stage (i.e., while clustering OGs and building single-gene trees). Unfortunately, the archaeal reconciliation rooting analysis did not include the Altiarchaeota (Williams et al., 2017). Considering their location at the base of the

DPANN and their proposed free-living lifestyle, their absence could significantly influence the root position and the inference of the genomic content of LACA.

6.5. A standardized system for taxonomy and in-depth phylogenomic analyses

One of the most impactful changes in archaeal taxonomy in the last five years has been the introduction of the Genome Taxonomy Database (GTDB) (Rinke et al., 2021). GTDB is a standardized taxonomy framework that uses normalized ranks based on relative evolutionary divergence (RED values) to delineate higher-rank taxa and average nucleotide identity (ANI) to delineate species clusters (Parks et al., 2022). GTDB was developed in response to the deluge of new sequence data deposited into public databases (Hugenholtz et al., 2021). Before GTDB, the rules for naming new taxa were based on the guidelines of the International Code of Nomenclature of Prokaryotes (ICNP) (Parker et al., 2019). However, the ICNP did not consider uncultured microorganisms or recognize the ranks of phylum and superphylum. Under the ICNP, uncultured Archaea and Bacteria could be provisionally named using the *Candidatus* status (Muray and Stackebrandt 1995). However, these names have no formal standing in nomenclature and are subject to change. Considering that a significant portion of newly described microbial taxa stems from uncultured lineages, there was a pressing need for a revamped system in microbial taxonomy.

However, despite significant improvements, there are still some imperfections with GTDB. For example, to reduce computational requirements, the phylogenetic markers that are used to build the base phylogenetic tree are trimmed by randomly selecting 42 amino acids from each marker, resulting in a

total of 5,124 alignment sites (Rinke et al., 2021). In our phylogenetic analyses of 48 ribosomal proteins with an initial alignment length of ~6,000 amino acids, the use of strict trimming parameters, resulting in an alignment of ~5,500 amino acids, resulted in the movement of deep branches (those that are relevant for the definition of phyla, for example). This suggests that an alignment of 5,000 amino acids may lack sufficient phylogenetic signals to place all archaeal lineages robustly. In these short alignments, we also showed that a few highly compositionally biased amino acid sites had a strong impact compared to longer alignments (see Halophiles chapter for more details).

Additionally, GTDB now uses an updated marker set of 53 proteins (Dombrowski et al., 2020), of which 38 are ribosomal proteins. In both chapters of my thesis, we routinely show that ribosomal protein markers are especially prone to phylogenetic artifacts, specifically compositional biases. As individual ribosomal proteins co-evolve to form a single complex, namely the ribosome, any compositional biases—such as a preference for acidic amino acids in extreme halophilic archaea—become magnified across multiple ribosomal proteins. Due to the generally shorter length of ribosomal proteins than other proteins (approximately 2–4 times smaller than average proteins (Reuveni et al., 2017)), these biases have a greater impact on the stability of the phylogenetic tree.

All things considered, I am not sure if there is a better solution other than potentially reevaluating the phylogenetic markers used. GTDB is an essential resource for microbial taxonomy in the age of big data genomic studies. I believe in-depth phylogenomic studies are still essential and can be complemented by a standardized taxonomic system. However, I'm unsure how to integrate the two systems. For example, our phylogenomic study of the DPANN changes their position significantly within the archaeal tree. However, this shift will not be

reflected in their taxonomy. The challenge is compounded by the community's uncertainty about the reliability of specific in-depth phylogenomic analyses, which can change based on taxonomic sampling and the chosen phylogenetic methods. In conclusion, GTDB has significantly progressed microbial taxonomy, and ongoing improvements will further refine this system.

6.6. Perspectives

- **The position of the archaeal root**

The position of the archaeal root is still a major unanswered question in archaeal phylogenetics. In previous analyses, there have been two main methods to root the archaeal tree: 1) using bacterial sequences as an outgroup and 2) using a gene tree-species tree reconciliation approach. Unfortunately, phylogenetic artifacts, such as sequence compositional biases and LBA, can directly affect both methods. While a gene tree-species tree reconciliation avoids the potential LBA artifacts induced by outgroup rooting, it is still susceptible to phylogenetic artifacts when reconstructing the single gene trees needed for the reconciliation. A third new approach will need to be developed to answer the question of the placement of the archaeal root or a more in-depth phylogenomic analysis. The position of the archaeal root also depends on the taxonomic sampling available at the time of the analysis. This question will need to be continuously re-evaluated if new deep-branching lineages are described.

- **Phylogenetic position of deep-branching euryarchaea**

In our phylogenomic analyses, two euryarchaeal phyla, the Hadarchaeota and Methanobacteriota_B, had unstable positions in some phylogenetic trees. The Hadarchaeota and Methanobacteriota_B were the two lineages that also grouped

with the TACK/Asgard when the trees were rooted on the DPANN branch. Both of these lineages contain thermophilic representatives, which have been shown to have sequence compositional biases to increase the thermostability of their proteins (Barns et al., 1996; Singer & Hickey, 2003). An in-depth phylogenomic analysis of these lineages is needed to know if these compositional biases are responsible for their phylogenetic instability. Their position is also highly relevant to the position of the archaeal root.

- **In-depth phylogenomic analyses of the Huberarchaeota and SpSt-1190 DPANN phyla**

At the time of our DPANN taxa data collection, only a single Huberarchaeota genome was available. Since then, additional Huberarchaeota genomes have been deposited (Esser et al., 2023). In our analyses, the Huberarchaeota had significantly longer branches than the other DPANN lineages. It would be interesting to investigate whether these long branches were due to taxonomic sampling issues or to a faster rate of genome evolution compared to other DPANN. The Huberarchaeota representative in our dataset had one of the smallest DPANN genomes (0.5 Mbp). Additional work is needed to examine the evolutionary history of the Huberarchaeota. This is also true for the unnamed DPANN lineage SpSt-1190. Recently, a member of the SpSt-1190 was found to have a genome size of 4 Mb, which encodes a nearly complete pathway for methanogenesis and several other methane-related metabolisms, such as the ribulose monophosphate pathway and methanofuran biosynthesis (Zhang et al., 2023). To our knowledge, this is the first instance of a DPANN member encoding pathways related to methanogenesis. Determining this group's precise phylogenetic position is important to better

understand the evolution of these metabolic features and to understand if their larger genome is ancestral or derived.

- **Additional DPANN cultures**

Discovering and cultivating new DPANN lineages will reshape our understanding of this group. Although genomic data has advanced our knowledge, many predicted proteins in DPANN proteomes still have unknown functions. Many questions remain elusive for this group. For instance, the cellular mechanisms used by these lineages to interact with their host are poorly understood. Are all DPANN episymbionts? Or can some enter their host cell, as recently suggested? (J. Hamm et al., 2023). Can DPANN only be symbiotic partners of other archaea? Or can their hosts also include bacteria and eukaryotes? Are there free-living DPANN lineages? Are their symbiotic relationship strictly commensal? Mutualistic? Parasitic? Despite significant progress in understanding this group over the past decade, many of these questions can only be resolved by cultivating additional DPANN lineages.

- **An in-depth analysis of the shared gene content between the DPANN archaea and symbiotic bacteria**

In our DPANN paper, we identified three protein families shared between DPANN and other symbiotic bacteria. This finding was intriguing as it indicated that these symbiotic groups evolved through convergent evolution. However, this analysis was not an exhaustive search. We ran the proteomes using default settings through OrthoFinder (Emms and Kelly 2019) to identify these proteins. However, based on previous experience with OrthoFinder, sometimes a more tailored approach is needed for identifying shared orthologous groups (OGs) in the DPANN. Detecting conserved OGs in DPANN can be difficult due to their accelerated rate of genome

evolution compared to other archaea, resulting in highly divergent protein sequences. Moreover, due to their small genome size, several proteins that could have been present in the ancestral branch leading to the DPANN may have been lost in extant lineages. To account for these possible losses, we defined an OG as being ancestral if it was present in at least 7 out of the 11 DPANN phyla.

7. Conclusions

7. Conclusions

In this work, I performed in-depth phylogenomic analyses on two significant archaeal groups—extreme halophilic archaea and DPANN. I introduced novel methods to mitigate phylogenetic artifacts, including a ranking scheme for identifying and excluding compositionally biased amino acid sites. Additionally, I employed advanced models of genome and sequence evolution, such as gene tree-species tree reconciliation and the GF-mix model (Muñoz-Gómez et al., 2022), to address branch-heterogeneity. In summary, these analyses revealed that compositional sequence biases, LBA, and factors such as taxon sampling and choice of phylogenetic markers contribute significantly to the instability or stability of the phylogenetic position of archaeal lineages. In these analyses, I demonstrated that ribosomal proteins are not robust phylogenetic markers for deep-branching lineages due to biased amino acid sites and relatively short sequence lengths. Moreover, unique amino acid preferences within specific archaeal groups influence their phylogenetic stability, presenting a challenge for a universal phylogenomic approach to resolve the archaeal tree.

Key outcomes of my work:

1) Independent Adaptations in Extreme Halophilic Archaea: Our study reveals that extreme halophilic archaea have independently adapted to hypersaline environments at least four times within the archaeal tree.

2) Amino Acid Sequence Biases in Ribosomal Proteins: The phylogenetic instability of Methanonatronarchaeia is driven by strong amino acid sequence biases in ribosomal proteins specific to extreme halophilic archaea.

3) Novel Uncultured Lineages: Two novel uncultured lineages, Afararchaeaceae and Asbonarchaeaceae, disrupt long branches at the base of Haloarchaea and Nanohaloarchaeota, respectively. Additionally, Chewarchaeaceae emerges as the deepest-branching Haloarchaea.

4) Secondary Adaptation of Hikarchaeia: The position of Afararchaeaceae, sister to Haloarchaea and Hikarchaeia, suggests a secondary adaptation of Hikarchaeia from a hypersaline ancestor to a moderate saline environment.

5) Role of Gene Duplication and HGT: Gene tree-species tree reconciliation highlights the significant role of gene duplication and HGT in archaeal adaptation to hypersaline environments, including the spread of key genes like those encoding potassium transporters.

6) Horizontal Gene Transfers from Bacteria: Several HGTs from bacteria occurred before the split between Afararchaeaceae and Haloarchaea, facilitating the adaptation of their common ancestor to extreme halophily.

7) Origin of DPANN Archaea: DPANN archaea form a monophyletic group nested within the Euryarchaeota, suggesting their evolution from a free-living euryarchaeota-like ancestor.

8) Role of Altiarchaeota in DPANN Evolution: Altiarchaeota's position at the base of all other DPANN further supports the evolution of DPANN from a free-living ancestor. The transition to host-dependent episymbiotic lifestyle likely occurred once at the base of the DPANN clade after the divergence from Altiarchaeota.

9) Conserved Orthologous Groups: Fourteen orthologous groups found in at least seven of the 11 DPANN phyla suggest these are ancient proteins conserved throughout the DPANN. Ten of these 14 orthologous groups were predicted to be secreted or contain transmembrane domains, suggesting that they may be involved

in the interaction with their host and therefore their adaptation to an episymbiotic lifestyle.

10) Horizontal Gene Transfer from Episymbiotic Bacteria: Three of the 14 DPANN-specific orthologous groups are also present in diverse episymbiotic bacteria (Patescibacteria and Omnitrophota), indicating the pivotal role of HGT in shaping convergent evolution across Bacteria and Archaea towards similar host-dependent lifestyles.

8. Résumé en français

Phylogénie profonde des archées et dynamique évolutive des DPANN

1.1. Introduction et objectifs

La classification des organismes remonte à l'Antiquité, avec les travaux d'Aristote au quatrième siècle avant Jésus-Christ. Les animaux étaient classés en fonction de caractéristiques morphologiques communes et le terme "genre" a été proposé pour chaque groupe. Ce terme est encore utilisé de nos jours. Jusqu'aux années 1960, les biologistes ont suivi les pratiques d'Aristote, classant les organismes en fonction de leurs phénotypes. Cette méthode de classification a permis de regrouper tous les micro-organismes morphologiquement simples (bactéries et archées), d'abord sous le nom de "Monera" par Ernst Haeckel en 1866, puis sous le nom de "Prokaryotes" par Édouard Chatton en 1937 (Sapp 2005). Cependant, au milieu des années 1960, un changement radical s'est produit dans notre compréhension de l'évolution microbienne avec l'introduction de la phylogénie moléculaire basée sur la comparaison des séquences de macromolécules biologiques (acides nucléiques et protéines) (Carl R. Woese et al. 1975; Zuckerkandl and Pauling 1965). Ces méthodes pionnières ont donné un premier aperçu des relations évolutives entre les micro-organismes au-delà de leurs caractéristiques phénotypiques et ont finalement conduit à la différenciation de deux domaines Procaryotes, les Archaea et les Bacteria (anciennement Archaebacteria et Eubacteria, respectivement ; Fig. 1) (C. Woese 1990).

Ces premiers arbres phylogénétiques étaient basés sur la comparaison des séquences d'ARN ribosomal. Les ARN ribosomiques (ARNr) sont des ARN non

codants qui constituent un composant primaire du ribosome. La caractéristique la plus notable des ARNr est leur grande abondance dans tous les organismes cellulaires, ce qui les rend relativement faciles à extraire (Van de Peer, Chapelle, and De Wachter 1996). Les séquences d'ARNr évoluent également lentement dans le temps, ce qui permet de détecter des liens de parenté entre des espèces très éloignées (Woese et al. 1978). L'universalité des ARNr a permis la classification phylogénétique de la vie en trois domaines primaires (Eucarya, Archaea et Bacteria) (Fig. 1). En outre, le séquençage de l'ARNr s'est avéré être une approche convaincante pour identifier de nouveaux microbes dans l'environnement qui ne micro-organismes impliquait l'approche traditionnelle consistant à les enrichir et à les isoler individuellement. Les analyses moléculaires des séquences d'ARNr 16S de l'environnement ont considérablement élargi notre connaissance de la diversité de la vie microbienne et ont introduit l'ère des études indépendantes de la culture (Pace 2009; Amann et al. 1995; Hugenholtz 2002; Barns et al. 1996; Pace 1997). Cependant, au fur et à mesure que des séquences d'ARNr étaient ajoutées à l'arbre de la vie, il est apparu que les phylogénies d'ARNr n'offraient pas une résolution suffisante pour établir des relations phylogénétiques profondes (Brochier-Armanet et al. 2011).

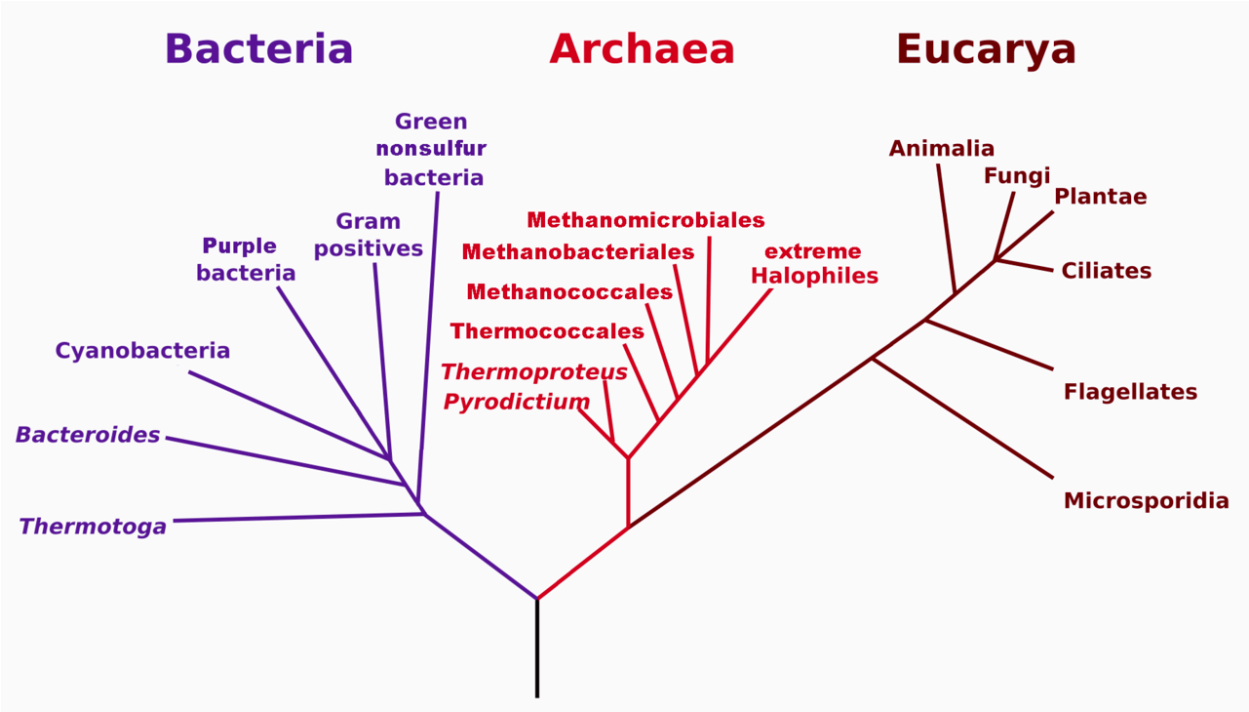


Figure 1 | Le premier arbre de vie à trois domaines enraciné, adapté de l'article fondateur de Carl Woese en 1990. Ce premier arbre de vie à trois domaines a marqué un tournant dans la compréhension de l'évolution microbienne. L'arbre a été construit à partir de comparaisons de séquences d'ARNr et a pris racine entre les bactéries et les archées. Cet arbre a également montré que les Archaea étaient divisés en deux royaumes principaux, les Euryarchaeota et les Crenarchaeota. Les Crenarchaeota sont considérés comme exclusivement thermophiles, tandis que les Euryarchaeota se composent d'un groupe plus hétérogène de méthanogènes, d'halophiles extrêmes et de réducteurs de sulfate.

Pendant plus d'une décennie (1990-2002), les Euryarchaeota et les Crenarchaeota ont été les seuls groupes d'archées reconnus (Fig. 2A) (C. Woese 1990). Les Crenarchaeota étaient considérés comme exclusivement thermophiles, tandis que les Euryarchaeota se composaient d'un groupe plus hétérogène de

méthanogènes, d'halophiles extrêmes et de réducteurs de sulfate. En 2003, 16 génomes complets d'archées avaient été séquencés, dont 12 euryarchaeota et quatre crenarchaeota, plusieurs autres étant presque complets (Makarova and Koonin 2003). Ces premiers efforts de séquençage ont montré que les archées possèdent de nombreuses caractéristiques génétiques uniques, distinctes de celles des bactéries. Un exemple notable est la capacité des Archaea à se développer dans des conditions extrêmes, comme dans l'eau à proximité des cheminées hydrothermales chauffée à des températures supérieures à l'ébullition et saturée de sulfure d'hydrogène, ou dans des conditions de salinité extrême (Stetter 1999; Segerer et al. 1993). Bien que l'on puisse également trouver des bactéries dans ces environnements, les archées dominent généralement ces communautés microbiennes (Stetter 1999; Oren 2002; Merino et al. 2019). Cela a largement donné lieu à une idée fausse selon laquelle toutes les archées sont des extrêmophiles, bien qu'elles soient présentes dans presque tous les environnements modérés connus.

Entre 2002 et 2011, les progrès du séquençage de nouvelle génération, l'amélioration des techniques phylogénétiques et des assemblages génomiques ont conduit à la proposition de plusieurs nouveaux embranchements d'archées (Fig. 2B) (Spang et al. 2017). Il s'agit notamment des Korarchaeota, un groupe de thermophiles provenant de sources chaudes terrestres (Reigstad et al. 2010; Barns et al. 1996), les Nanoarchaeota, représentés par une archée hyperthermophile de taille nanométrique provenante d'une cheminée sous-marine, et les Thaumarchaeota oxydant l'ammoniac (DeLong 1992; Brochier-Armanet et al. 2008). Avec le phylum candidat Aigarchaeota (Nunoura et al. 2011), les Thaum-, Cren- et Korarchaeota ont été proposés comme un nouveau superphylum d’archées connu sous le nom de "TACK" (également appelé "Proteoarchaeota") (Guy and Ettema

2011; Petitjean et al. 2015). Si l'acronyme "TACK" est toujours utilisé aujourd'hui, il englobe désormais une plus grande diversité de phyla, y compris les Bathyarchaeota (Barns et al. 1996; Kubo et al. 2012), les Geoarchaeota (Kozubal et al. 2013), et les Verstraetearchaeota (Vanwonterghem et al. 2016). En général, le superphylum TACK contient des lignées aux métabolismes divers, tels que les oxydants d'ammoniac, les autotrophes et les méthanogènes, et sont des membres biologiques importants des cycles globaux du carbone et de l'azote (Kozubal et al. 2013; Ingalls et al. 2006; Adam et al. 2017).

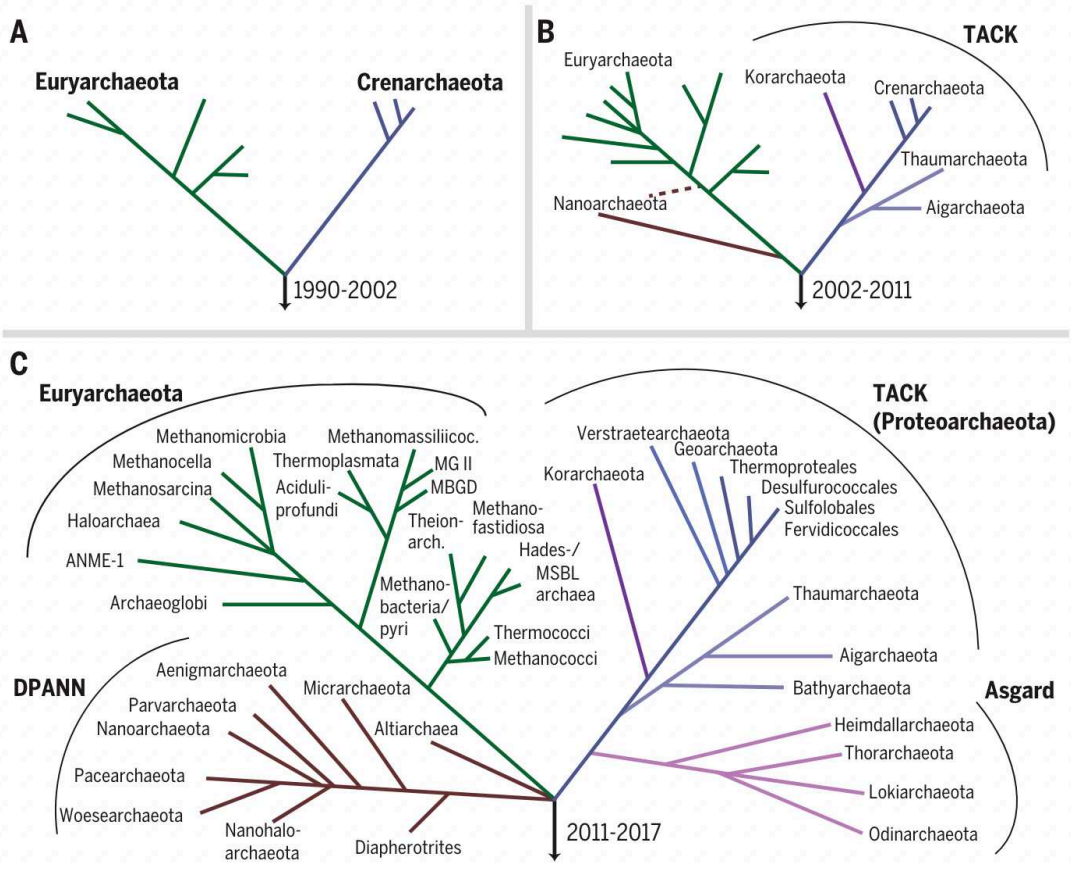


Figure 2 | L'expansion de l'arbre des archées au cours des dernières décennies. Cette figure a été adaptée de Spang et al. 2017 et illustre notre compréhension progressive de l'arbre des archées au cours des 30 dernières

années. (A) La classification initiale de Carl Woese en 1990 identifiait les Euryarchaeota et les Crenarchaeota comme les deux premiers phylums d'archées. (B) Entre 2002 et 2011, les progrès du séquençage génomique et les études environnementales sur l'ARNr 16S ont permis de classer les Euryarchaeota et les superphyla TACK, ainsi que d'identifier un nouveau phylum, les Nanoarchaeota. (C) L'arbre des archées compte désormais quatre superphylums majeurs : Euryarchaeota, TACK, Asgard et DPANN.

1.2. Objectifs

Dans ce contexte général, les objectifs spécifiques de mon doctorat sont les suivants :

2. **Résoudre la position phylogénétique des archées extrêmement halophiles. Les Nanohaloarchaeota appartiennent-ils aux archées DPANN ?**

Les caractéristiques génomiques et le mode de vie symbiotique des Nanohaloarchaeota suggèrent que cette lignée appartient aux archées DPANN. Cependant, depuis leur découverte il y a plus d'une décennie, leur position phylogénétique dans l'arbre des archées est instable. Au départ, les Nanohaloarchaeota ont été décrits comme une lignée distincte, aux ramifications profondes, au sein des Euryarchaeota, séparée du seul autre groupe connu d'archées halophiles, les Halobacteriota. Avec la découverte d'autres lignées d'archées de taille nanométrique, les Nanohaloarchaeota ont été placées au sein du superphylum DPANN. Or, en 2022 de nouvelles études ont mis une fois de plus en question la position phylogénétique des Nanohaloarchaeota, les classant dans les Euryarchaeota. Étant donné que les Nanohaloarchaeota sont le groupe le plus

fréquemment découvert en dehors des DPANN dans les arbres phylogénétiques, notre objectif initial était de déterminer leur position phylogénétique précise et, par extension, la position phylogénétique de toutes les archées halophiles. Pour ce faire, nous avons décrit deux nouvelles lignées familiales d'archées halophiles et développé de nouvelles méthodes phylogénétiques robustes pour traiter les biais d'acides aminés prévalant dans toutes les archées extrêmement halophiles.

2. Déterminer la position phylogénétique des archées DPANN. Les DPANN constituent-ils un groupe monophylétique ? Quelle est la place des DPANN dans l'arbre de vie des archées ?

Les archées DPANN constituent l'un des quatre superphylums d'archées proposés, avec les TACK, les Asgard et les Euryarchaeota. Les DPANN se caractérisent typiquement par la petite taille de leurs cellules et par un ensemble limité de protéines métaboliques (par exemple, biosynthèse d'acides aminés et de lipides). Néanmoins, le statut monophylétique de ce groupe reste une question non résolue. Les premières analyses phylogénétiques des lignées DPANN individuelles, par exemple Nanoarchaeum equitans et Nanosalinarun sp. J07AB56, ont suggéré que ces organismes étaient des Euryarchaeota à ramification profonde. Cependant, à mesure que d'autres lignées DPANN ont été découvertes, les analyses phylogénétiques ultérieures ont indiqué la formation d'un groupe monophylétique. Les défenseurs du statut monophylétique des DPANN soutiennent que les recherches antérieures n'ont pas été suffisamment échantillonnées sur le plan taxonomique.

À l'inverse, les détracteurs de l'étude soutiennent que l'ajout de taxons supplémentaires pourrait entraîner des artefacts dans la reconstruction de l'arbre phylogénétique, tels que l'attraction des longues branches, attribuée au taux

d'évolution rapide de ces lignées. Pour répondre à cette question, l'objectif était de déterminer si la position instable des archées DPANN résulte de l'échantillonnage de taxons dans des lignées spécifiques ou si elle est influencée par des artefacts phylogénétiques tels que l'attraction des longues branches ou des biais de composition en acides aminés dans les ensembles de données phylogénétiques. Par la suite, nous appliquerons des méthodologies phylogénétiques comparables, telles que décrites dans l'article sur les halophiles, afin d'étudier comment les biais liés aux acides aminés influencent le positionnement phylogénétique de DPANN.

2.1. Résultats

Manuscrit 1 : Plusieurs adaptations indépendantes des archées aux environnements hypersalins

Nos résultats montrent que les archées halophiles se sont adaptées indépendamment aux environnements hypersalins au moins quatre fois au cours de l'évolution des archées. Nous avons également démontré avec une grande confiance que le Nanohaloarchaeota est bien un membre des archées DPANN. Dans le cadre de ce projet, nous avons également décrit deux nouvelles lignées non cultivées, Afararchaeaeae et Asbonarchaeaeae, qui brisent les longues branches à la base des Haloarchaea et des Nanohaloarchaeota, respectivement. Nos résultats soulignent que des compositions uniques d'acides aminés partagées par les halophiles ont artificiellement placé ces lignées ensemble dans des analyses phylogénétiques antérieures. Nous avons obtenu des placements phylogénétiques plus cohérents et plus fiables en filtrant ces points de données biaisés et en modélisant ces biais à l'aide du modèle d'hétérogénéité de branche adapté. Dans ce projet, nous avons également reconstruit l'histoire évolutive des familles de

gènes d'archées en cartographiant des événements tels que les duplications, les transferts, les origines et les pertes de gènes à l'aide de méthodes de réconciliation arbre génique-arbre d'espèce. Dans cette analyse, nous nous sommes concentrés sur les événements évolutifs qui ont eu un impact sur les branches ancestrales menant aux différents groupes de lignées halophiles. Ces résultats suggèrent que la duplication et le transfert horizontal de gènes ont joué un rôle important dans l'adaptation à l'halophilie, par exemple en diffusant des gènes clés (tels que ceux codant pour les transporteurs de potassium) dans les différentes lignées halophiles extrêmes.

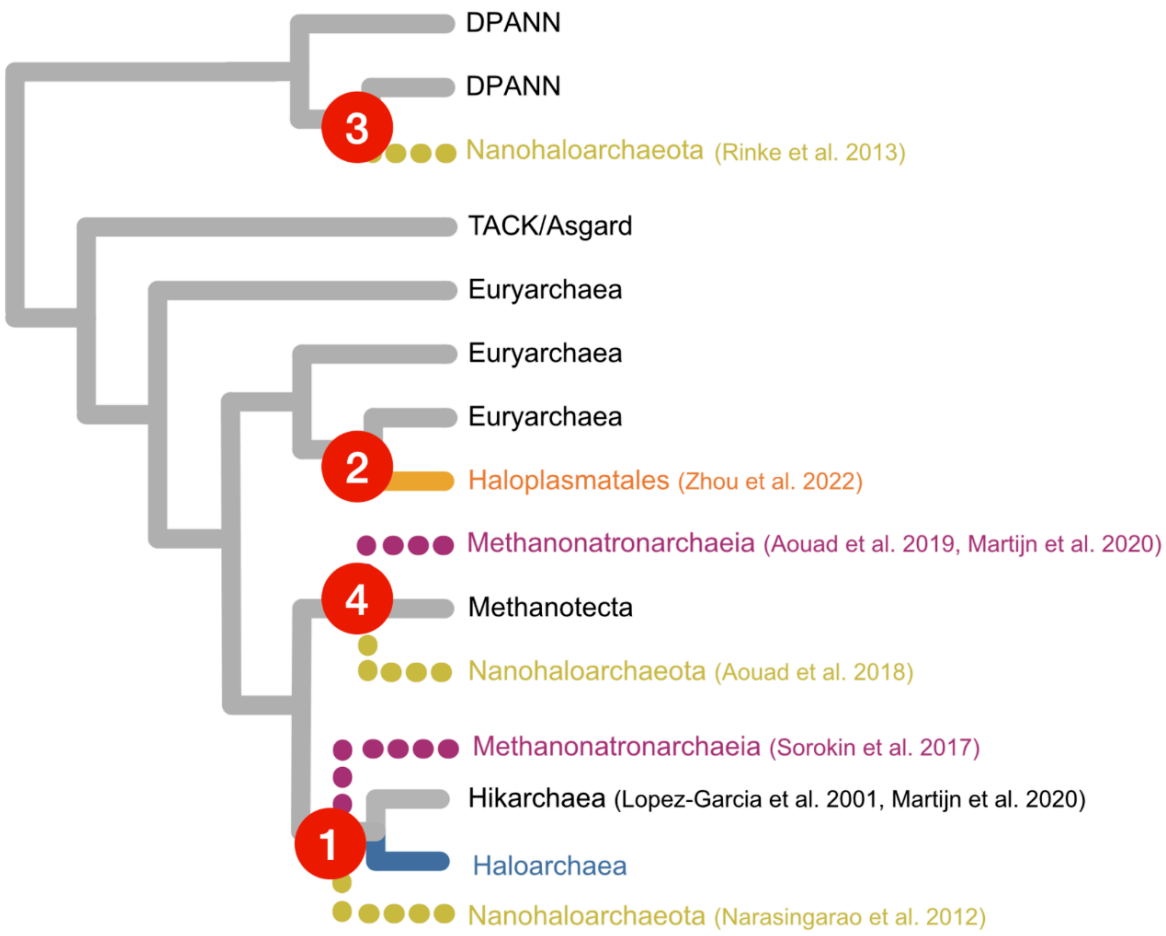


Figure 3 | Représentation schématique des positions phylogénétiques proposées pour divers groupes d'archées halophiles. Les cercles rouges sur cet arbre indiquent les cas proposés d'adaptation des archées aux environnements hypersalins.

Manuscrit 2 : L'analyse phylogénomique des archées DPANN révèle leur monophilie et leurs origines évolutives

Dans ce projet, nous avons cherché à évaluer la monophylie des archées DPANN et leur position dans l'arbre des archées. Nous montrons de manière robuste que les archées DPANN sont un groupe monophylétique qui se ramifie bien à l'intérieur des Euryarchaeota lorsque l'arbre est enraciné sur la branche TACK/Asgard. Le positionnement des DPANN au sein des Euryarchaeota suggère que les DPANN ont évolué à partir d'un ancêtre vivant librement, semblable aux Euryarchaeota. Cette caractéristique ancestrale est encore renforcée par la position des Altiarchaeota vivant librement à la base du clade DPANN. Si toutes les autres lignées DPANN sont réellement des épisymbiontes obligatoires, cela suggère que la transition d'un ancêtre vivant librement à un épisymbionte dépendant de l'hôte ne s'est produite qu'une seule fois à la base du clade DPANN après la divergence avec les Altiarchaeota. Cependant, l'ubiquité d'un mode de vie dépendant de l'hôte reste une question ouverte majeure qui doit être confirmée expérimentalement pour la plupart des phyla DPANN connus. Nous avons également observé d'anciens liens évolutifs entre les archées DPANN et divers groupes de bactéries, en particulier les Patescibacteria et les Omnitrophota, qui ont également montré qu'ils avaient un mode de vie similaire dépendant de l'hôte (Seymour et al. 2023; López-García and Moreira 2021). L'HGT semble avoir joué un rôle essentiel dans l'évolution convergente de ces lignées distinctes de bactéries et d'archées vers des modes de

vie similaires dépendant de l'hôte (Podar et al. 2008; Dombrowski et al. 2020; Aouad et al. 2018; Jaffe et al. 2019; Narasingarao et al. 2012a; Castelle and Banfield 2018; Rinke et al. 2013; Castelle et al. 2021). Dans l'ensemble, nos résultats fournissent un scénario actualisé pour l'évolution des archées DPANN, dans lequel un événement évolutif unique conduisant à la transition vers un mode de vie symbiotique à partir d'un ancêtre vivant librement s'est produit au sein des Euryarchaeota, très probablement accompagné de plusieurs événements HGT à partir de bactéries.

2.2. Conclusions

Dans ce travail, j'ai effectué des analyses phylogénomiques approfondies sur deux groupes d'archées importants - les archées halophiles extrêmes et les DPANN. J'ai introduit de nouvelles méthodes pour atténuer les artefacts phylogénétiques, y compris un schéma de classement pour identifier et exclure les sites d'acides aminés biaisés par la composition. En outre, j'ai utilisé des modèles avancés d'évolution des génomes et des séquences, tels que la réconciliation arbre génique-arbre d'espèce et le modèle GF-mix (Muñoz-Gómez et al. 2022), pour traiter l'hétérogénéité des branches. En résumé, ces analyses ont révélé que les biais de composition des séquences, l'ACL et des facteurs tels que l'échantillonnage des taxons et le choix des marqueurs phylogénétiques contribuent de manière significative à l'instabilité ou à la stabilité de la position phylogénétique des lignées d'archées. Dans ces analyses, j'ai démontré que les protéines ribosomiques ne sont pas des marqueurs phylogénétiques robustes pour les lignées à ramification profonde en raison de sites d'acides aminés biaisés et de longueurs de séquences relativement courtes. En outre, des préférences uniques en matière d'acides

aminés au sein de groupes d'archées spécifiques influencent leur stabilité phylogénétique, ce qui représente un défi pour une approche phylogénomique universelle visant à résoudre l'arbre des archées.

9. References

Adam, P. S., Borrel, G., Brochier-Armanet, C., & Gribaldo, S. (2017). The growing tree of Archaea: New perspectives on their diversity, evolution and ecology. *The ISME Journal*, *11*(11), Article 11. https://doi.org/10.1038/ismej.2017.122

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Amann, R. I., Ludwig, W., & Schleifer, K. H. (1995). Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiological Reviews*, *59*(1), 143–169. https://doi.org/10.1128/mr.59.1.143-169.1995

Andrade, K., Logemann, J., Heidelberg, K. B., Emerson, J. B., Comolli, L. R., Hug, L. A., Probst, A. J., Keillar, A., Thomas, B. C., Miller, C. S., Allen, E. E., Moreau, J. W., Brocks, J. J., & Banfield, J. F. (2015). Metagenomic and lipid analyses reveal a diel cycle in a hypersaline microbial ecosystem. *The ISME Journal*, *9*(12), Article 12. https://doi.org/10.1038/ismej.2015.66

Andrei, A.-Ş., Banciu, H. L., & Oren, A. (2012). Living with salt: Metabolic and phylogenetic diversity of archaea inhabiting saline ecosystems. *FEMS Microbiology Letters*, *330*(1), 1–9. https://doi.org/10.1111/j.1574-6968.2012.02526.x

Aouad, M., Borrel, G., Brochier-Armanet, C., & Gribaldo, S. (2019). Evolutionary placement of Methanonatronarchaeia. *Nature Microbiology*, *4*(4), Article 4. https://doi.org/10.1038/s41564-019-0359-z

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

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

Archibald, J. K., Mort, M. E., & Crawford, D. J. (2003). Bayesian Inference of Phylogeny: A Non-Technical Primer. *Taxon*, *52*(2), 187–191.

https://doi.org/10.2307/3647388

Baker, B. A., Gutierrez-Preciado, A., Rio, A. R. del, McCarthy, C., Lopez-Garcia, P., Huerta-Cepas, J., Susko, E., Roger, A. J., Eme, L., & Moreira, D. (2023). *Several independent adaptations of archaea to hypersaline environments* (p. 2023.07.03.547478). bioRxiv. https://doi.org/10.1101/2023.07.03.547478

Baker, B. J., Comolli, L. R., Dick, G. J., Hauser, L. J., Hyatt, D., Dill, B. D., Land, M. L., VerBerkmoes, N. C., Hettich, R. L., & Banfield, J. F. (2010). Enigmatic, ultrasmall, uncultivated Archaea. *Proceedings of the National Academy of Sciences*, *107*(19), 8806–8811. https://doi.org/10.1073/pnas.0914470107

Baker, B. J., Saw, J. H., Lind, A. E., Lazar, C. S., Hinrichs, K.-U., Teske, A. P., & Ettema, T. J. G. (2016). Genomic inference of the metabolism of cosmopolitan subsurface Archaea, Hadesarchaea. *Nature Microbiology*, *1*(3), Article 3. https://doi.org/10.1038/nmicrobiol.2016.2

Baker, B. J., Tyson, G. W., Webb, R. I., Flanagan, J., Hugenholtz, P., Allen, E. E., & Banfield, J. F. (2006). Lineages of Acidophilic Archaea Revealed by Community Genomic Analysis. *Science*, *314*(5807), 1933–1935. https://doi.org/10.1126/science.1132690

Baldauf, S. L., Roger, A. J., Wenk-Siefert, I., & Doolittle, W. F. (2000). A Kingdom-Level Phylogeny of Eukaryotes Based on Combined Protein Data. *Science*, *290*(5493), 972–977. https://doi.org/10.1126/science.290.5493.972

Barns, S. M., Delwiche, C. F., Palmer, J. D., & Pace, N. R. (1996). Perspectives on archaeal diversity, thermophily and monophyly from environmental rRNA sequences. *Proceedings of the National Academy of Sciences of the United States of America*, *93*(17), 9188–9193.

Becker, E. A., Seitzer, P. M., Tritt, A., Larsen, D., Krusor, M., Yao, A. I., Wu, D., Madern, D., Eisen, J. A., Darling, A. E., & Facciotti, M. T. (2014). Phylogenetically Driven Sequencing of Extremely Halophilic Archaea Reveals Strategies for Static and Dynamic Osmo-response. *PLOS Genetics*, *10*(11), e1004784. https://doi.org/10.1371/journal.pgen.1004784

Belilla, J., Iniesto, M., Moreira, D., Benzerara, K., López-García, J. M., López-Archilla, A. I., Reboul, G., Deschamps, P., Gérard, E., & López-García, P. (2021). Archaeal overdominance close to life-limiting conditions in geothermally influenced hypersaline lakes at the Danakil Depression, Ethiopia.

Environmental Microbiology, *23*(11), 7168–7182. https://doi.org/10.1111/1462-2920.15771

Belilla, J., Moreira, D., Jardillier, L., Reboul, G., Benzerara, K., López-García, J. M., Bertolino, P., López-Archilla, A. I., & López-García, P. (2019). Hyperdiverse archaea near life limits at the polyextreme geothermal Dallol area. *Nature Ecology & Evolution*, *3*(11), Article 11. https://doi.org/10.1038/s41559-019-1005-0

Bellanger, X., Payot, S., Leblond-Bourget, N., & Guédon, G. (2014). Conjugative and mobilizable genomic islands in bacteria: Evolution and diversity. *FEMS Microbiology Reviews*, *38*(4), 720–760. https://doi.org/10.1111/1574-6976.12058

Bergsten, J. (2005). A review of long-branch attraction. *Cladistics*, *21*(2), 163–193. https://doi.org/10.1111/j.1096-0031.2005.00059.x

Bird, J. T., Baker, B. J., Probst, A. J., Podar, M., & Lloyd, K. G. (2016). Culture Independent Genomic Comparisons Reveal Environmental Adaptations for Altiarchaeales. *Frontiers in Microbiology*, *7*. https://doi.org/10.3389/fmicb.2016.01221

Bokhari, R. H., Amirjan, N., Jeong, H., Kim, K. M., Caetano-Anollés, G., & Nasir, A. (2020). Bacterial Origin and Reductive Evolution of the CPR Group. *Genome Biology and Evolution*, *12*(3), 103–121. https://doi.org/10.1093/gbe/evaa024

Borton, M. A., Daly, R. A., O'Banion, B., Hoyt, D. W., Marcus, D. N., Welch, S., Hastings, S. S., Meulia, T., Wolfe, R. A., Booker, A. E., Sharma, S., Cole, D. R., Wunch, K., Moore, J. D., Darrah, T. H., Wilkins, M. J., & Wrighton, K. C. (2018). Comparative genomics and physiology of the genus Methanohalophilus, a prevalent methanogen in hydraulically fractured shale. *Environmental Microbiology*, *20*(12), 4596–4611. https://doi.org/10.1111/1462-2920.14467

Boussau, B., & Scornavacca, C. (n.d.). *Reconciling Gene trees with Species Trees*.

Branciamore, S., Gallori, E., & Di Giulio, M. (2008). The basal phylogenetic position of Nanoarchaeum equitans (Nanoarchaeota). *Frontiers in Bioscience: A Journal and Virtual Library*, *13*, 6886–6892. https://doi.org/10.2741/3196

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 Biology*, *6*(5), R42.

https://doi.org/10.1186/gb-2005-6-5-r42

Brochier-Armanet, C., Boussau, B., Gribaldo, S., & Forterre, P. (2008). Mesophilic crenarchaeota: Proposal for a third archaeal phylum, the Thaumarchaeota. *Nature Reviews Microbiology*, *6*(3), 245–252. https://doi.org/10.1038/nrmicro1852

Brochier-Armanet, C., Forterre, P., & Gribaldo, S. (2011). Phylogeny and evolution of the Archaea: One hundred genomes later. *Current Opinion in Microbiology*, *14*(3), 274–281. https://doi.org/10.1016/j.mib.2011.04.015

Brown, C. T., Hug, L. A., Thomas, B. C., Sharon, I., Castelle, C. J., Singh, A., Wilkins, M. J., Wrighton, K. C., Williams, K. H., & Banfield, J. F. (2015). Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature*, *523*(7559), 208–211. https://doi.org/10.1038/nature14486

Brown, J. R., Douady, C. J., Italia, M. J., Marshall, W. E., & Stanhope, M. J. (2001). Universal trees based on large combined protein sequence data sets. *Nature Genetics*, *28*(3), Article 3. https://doi.org/10.1038/90129

Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, *12*(1), 59–60. https://doi.org/10.1038/nmeth.3176

Bult, C. J., White, O., Olsen, G. J., Zhou, L., Fleischmann, R. D., Sutton, G. G., Blake, J. A., FitzGerald, L. M., Clayton, R. A., Gocayne, J. D., Kerlavage, A. R., Dougherty, B. A., Tomb, J.-F., Adams, M. D., Reich, C. I., Overbeek, R., Kirkness, E. F., Weinstock, K. G., Merrick, J. M., … Venter, J. C. (1996). Complete Genome Sequence of the Methanogenic Archaeon, Methanococcus jannaschii. *Science*, *273*(5278), 1058–1073. https://doi.org/10.1126/science.273.5278.1058

Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P., & Huerta-Cepas, J. (2021). eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Molecular Biology and Evolution*, *38*(12), 5825–5829. https://doi.org/10.1093/molbev/msab293

Casanueva, A., Galada, N., Baker, G., Grant, W., Heaphy, S., Jones, B., Yanhe, M., Ventosa, A., Blamey, J., & Cowan, D. (2008). Nanoarchaeal 16S rRNA gene sequences are widely dispersed in hyperthermophilic and mesophilic halophilic environments. *Extremophiles : Life under Extreme Conditions*, *12*, 651–656. https://doi.org/10.1007/s00792-008-0170-x

Cassardo, C., & Jones, J. A. A. (2011). Managing Water in a Changing World. *Water*, *3*(2), Article 2. https://doi.org/10.3390/w3020618

Castelle, C. J., & Banfield, J. F. (2018). Major New Microbial Groups Expand Diversity and Alter our Understanding of the Tree of Life. *Cell*, *172*(6), 1181–1197. https://doi.org/10.1016/j.cell.2018.02.016

Castelle, C. J., Brown, C. T., Anantharaman, K., Probst, A. J., Huang, R. H., & Banfield, J. F. (2018). Biosynthetic capacity, metabolic variety and unusual biology in the CPR and DPANN radiations. *Nature Reviews Microbiology*, *16*(10), Article 10. https://doi.org/10.1038/s41579-018-0076-2

Castelle, C. J., Méheust, R., Jaffe, A. L., Seitz, K., Gong, X., Baker, B. J., & Banfield, J. F. (2021). Protein Family Content Uncovers Lineage Relationships and Bacterial Pathway Maintenance Mechanisms in DPANN Archaea. *Frontiers in Microbiology*, *12*. https://doi.org/10.3389/fmicb.2021.660052

Castelle, C. J., Wrighton, K. C., Thomas, B. C., Hug, L. A., Brown, C. T., Wilkins, M. J., Frischkorn, K. R., Tringe, S. G., Singh, A., Markillie, L. M., Taylor, R. C., Williams, K. H., & Banfield, J. F. (2015). Genomic Expansion of Domain Archaea Highlights Roles for Organisms from New Phyla in Anaerobic Carbon Cycling. *Current Biology*, *25*(6), 690–701. https://doi.org/10.1016/j.cub.2015.01.014

Chen, K., Durand, D., & Farach-Colton, M. (2000). NOTUNG: A Program for Dating Gene Duplications and Optimizing Gene Family Trees. *Journal of Computational Biology*, *7*(3–4), 429–447. https://doi.org/10.1089/106652700750050871

Chen, L.-X., Méndez-García, C., Dombrowski, N., Servín-Garcidueñas, L. E., Eloe-Fadrosh, E. A., Fang, B.-Z., Luo, Z.-H., Tan, S., Zhi, X.-Y., Hua, Z.-S., Martinez-Romero, E., Woyke, T., Huang, L.-N., Sánchez, J., Peláez, A. I., Ferrer, M., Baker, B. J., & Shu, W.-S. (2018). Metabolic versatility of small archaea Micrarchaeota and Parvarchaeota. *The ISME Journal*, *12*(3), Article 3. https://doi.org/10.1038/s41396-017-0002-z

Chklovski, A., Parks, D. H., Woodcroft, B. J., & Tyson, G. W. (2023). CheckM2: A rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. *Nature Methods*, *20*(8), Article 8. https://doi.org/10.1038/s41592-023-01940-w

Christian, J. H. B., & Waltho, J. A. (1962). Solute concentrations within cells of

halophilic and non-halophilic bacteria. *Biochimica et Biophysica Acta*, *65*(3), 506–508. https://doi.org/10.1016/0006-3002(62)90453-5

Chuvochina, M., Rinke, C., Parks, D. H., Rappé, M. S., Tyson, G. W., Yilmaz, P., Whitman, W. B., & Hugenholtz, P. (2019). The importance of designating type material for uncultured taxa. *Systematic and Applied Microbiology*, *42*(1), 15–21. https://doi.org/10.1016/j.syapm.2018.07.003

Clark, M. A., Moran, N. A., & Baumann, P. (1999). Sequence evolution in bacterial endosymbionts having extreme base compositions. *Molecular Biology and Evolution*, *16*(11), 1586–1598. https://doi.org/10.1093/oxfordjournals.molbev.a026071

Coleman, G. A., Davín, A. A., Mahendrarajah, T. A., Szánthó, L. L., Spang, A., Hugenholtz, P., Szöllősi, G. J., & Williams, T. A. (2021). A rooted phylogeny resolves early bacterial evolution. *Science*, *372*(6542), eabe0511. https://doi.org/10.1126/science.abe0511

Collins, T. M., Wimberger, P. H., & Naylor, G. J. P. (1994). Compositional Bias, Character-State Bias, and Character-State Reconstruction Using Parsimony. *Systematic Biology*, *43*(4), 482–496. https://doi.org/10.1093/sysbio/43.4.482

Criscuolo, A., & Gribaldo, S. (2010). BMGE (Block Mapping and Gathering with Entropy): A new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evolutionary Biology*, *10*(1), 210. https://doi.org/10.1186/1471-2148-10-210

DeLong, E. F. (1992). Archaea in coastal marine environments. *Proceedings of the National Academy of Sciences of the United States of America*, *89*(12), 5685–5689.

Dennis, P. P., & Shimmin, L. C. (1997). Evolutionary divergence and salinity-mediated selection in halophilic archaea. *Microbiology and Molecular Biology Reviews: MMBR*, *61*(1), 90–104. https://doi.org/10.1128/mmbr.61.1.90-104.1997

Deole, R., Challacombe, J., Raiford, D. W., & Hoff, W. D. (2013). An Extremely Halophilic Proteobacterium Combines a Highly Acidic Proteome with a Low Cytoplasmic Potassium Content. *Journal of Biological Chemistry*, *288*(1), 581–588. https://doi.org/10.1074/jbc.M112.420505

Dick, A. A., Harlow, T. J., & Gogarten, J. P. (2017). Short branches lead to systematic artifacts when BLAST searches are used as surrogate for phylogenetic

reconstruction. *Molecular Phylogenetics and Evolution*, *107*, 338–344. https://doi.org/10.1016/j.ympev.2016.11.016

Dobrindt, U., Hochhut, B., Hentschel, U., & Hacker, J. (2004). Genomic islands in pathogenic and environmental microorganisms. *Nature Reviews Microbiology*, *2*(5), 414–424. https://doi.org/10.1038/nrmicro884

Dombrowski, N., Lee, J.-H., Williams, T. A., Offre, P., & Spang, A. (2019). Genomic diversity, lifestyles and evolutionary origins of DPANN archaea. *FEMS Microbiology Letters*, *366*(2). https://doi.org/10.1093/femsle/fnz008

Dombrowski, N., Williams, T. A., Sun, J., Woodcroft, B. J., Lee, J.-H., Minh, B. Q., Rinke, C., & Spang, A. (2020). Undinarchaeota illuminate DPANN phylogeny and the impact of gene transfer on archaeal evolution. *Nature Communications*, *11*(1), Article 1. https://doi.org/10.1038/s41467-020-17408-w

Doolittle, W. F. (1999). Phylogenetic Classification and the Universal Tree. *Science*, *284*(5423), 2124–2128. https://doi.org/10.1126/science.284.5423.2124

Dutta, B., & Bandopadhyay, R. (2022). Biotechnological potentials of halophilic microorganisms and their impact on mankind. *Beni-Suef University Journal of Basic and Applied Sciences*, *11*(1), 75. https://doi.org/10.1186/s43088-022-00252-w

Eisen, J. A. (2000). Horizontal gene transfer among microbial genomes: New insights from complete genome analysis. *Current Opinion in Genetics & Development*, *10*(6), 606–611. https://doi.org/10.1016/S0959-437X(00)00143-X

Elkins, J. G., Podar, M., Graham, D. E., Makarova, K. S., Wolf, Y., Randau, L., Hedlund, B. P., Brochier-Armanet, C., Kunin, V., Anderson, I., Lapidus, A., Goltsman, E., Barry, K., Koonin, E. V., Hugenholtz, P., Kyrpides, N., Wanner, G., Richardson, P., Keller, M., & Stetter, K. O. (2008). A korarchaeal genome reveals insights into the evolution of the Archaea. *Proceedings of the National Academy of Sciences*, *105*(23), 8102–8107. https://doi.org/10.1073/pnas.0801980105

Eme, L., Tamarit, D., Caceres, E. F., Stairs, C. W., De Anda, V., Schön, M. E., Seitz, K. W., Dombrowski, N., Lewis, W. H., Homa, F., Saw, J. H., Lombard, J., Nunoura, T., Li, W.-J., Hua, Z.-S., Chen, L.-X., Banfield, J. F., John, E. S., Reysenbach, A.-L., … Ettema, T. J. G. (2023). Inference and reconstruction of the heimdallarchaeial ancestry of eukaryotes. *Nature*, *618*(7967), Article 7967. https://doi.org/10.1038/s41586-023-06186-2

Emms, D. M., & Kelly, S. (2019). OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biology*, *20*(1), 238. https://doi.org/10.1186/s13059-019-1832-y

Esser, S. P., Rahlff, J., Zhao, W., Predl, M., Plewka, J., Sures, K., Wimmer, F., Lee, J., Adam, P. S., McGonigle, J., Turzynski, V., Banas, I., Schwank, K., Krupovic, M., Bornemann, T. L. V., Figueroa-Gonzalez, P. A., Jarett, J., Rattei, T., Amano, Y., … Probst, A. J. (2023). A predicted CRISPR-mediated symbiosis between uncultivated archaea. *Nature Microbiology*, *8*(9), Article 9. https://doi.org/10.1038/s41564-023-01439-2

Felsenstein, J. (1978). Cases in which Parsimony or Compatibility Methods Will be Positively Misleading. *Systematic Zoology*, *27*(4), 401–410. https://doi.org/10.2307/2412923

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, *17*(6), 368–376. https://doi.org/10.1007/BF01734359

Feng, Y., Neri, U., Gosselin, S., Louyakis, A. S., Papke, R. T., Gophna, U., & Gogarten, J. P. (2021). The Evolutionary Origins of Extreme Halophilic Archaeal Lineages. *Genome Biology and Evolution*, *13*(8), evab166. https://doi.org/10.1093/gbe/evab166

Fitch, W. M. (1970). Distinguishing Homologous from Analogous Proteins. *Systematic Biology*, *19*(2), 99–113. https://doi.org/10.2307/2412448

Forterre, P., Brochier, C., & Philippe, H. (2002). Evolution of the Archaea. *Theoretical Population Biology*, *61*(4), 409–422. https://doi.org/10.1006/tpbi.2002.1592

Foster, P. G., & Hickey, D. A. (1999). Compositional Bias May Affect Both DNA-Based and Protein-Based Phylogenetic Reconstructions. *Journal of Molecular Evolution*, *48*(3), 284–290. https://doi.org/10.1007/PL00006471

Fukuchi, S., Yoshimune, K., Wakayama, M., Moriguchi, M., & Nishikawa, K. (2003). Unique Amino Acid Composition of Proteins in Halophilic Bacteria. *Journal of Molecular Biology*, *327*(2), 347–357. https://doi.org/10.1016/S0022-2836(03)00150-5

Gadda, G., & McAllister-Wilkins, E. E. (2003). Cloning, Expression, and Purification of Choline Dehydrogenase from the Moderate Halophile Halomonas elongata. *Applied and Environmental Microbiology*, *69*(4), 2126–2132.

https://doi.org/10.1128/AEM.69.4.2126-2132.2003

Ghai, R., Pašić, L., Fernández, A. B., Martin-Cuadrado, A.-B., Mizuno, C. M., McMahon, K. D., Papke, R. T., Stepanauskas, R., Rodriguez-Brito, B., Rohwer, F., Sánchez-Porro, C., Ventosa, A., & Rodríguez-Valera, F. (2011). New Abundant Microbial Groups in Aquatic Hypersaline Environments. *Scientific Reports*, *1*, 135. https://doi.org/10.1038/srep00135

Gogarten, J. P., & Townsend, J. P. (2005). Horizontal gene transfer, genome innovation and evolution. *Nature Reviews Microbiology*, *3*(9), 679–687. https://doi.org/10.1038/nrmicro1204

Golyshina, O. V., Bargiela, R., Toshchakov, S. V., Chernyh, N. A., Ramayah, S., Korzhenkov, A. A., Kublanov, I. V., & Golyshin, P. N. (2019). Diversity of "Ca. Micrarchaeota" in Two Distinct Types of Acidic Environments and Their Associations with Thermoplasmatales. *Genes*, *10*(6). https://doi.org/10.3390/genes10060461

Golyshina, O. V., Toshchakov, S. V., Makarova, K. S., Gavrilov, S. N., Korzhenkov, A. A., La Cono, V., Arcadi, E., Nechitaylo, T. Y., Ferrer, M., Kublanov, I. V., Wolf, Y. I., Yakimov, M. M., & Golyshin, P. N. (2017). 'ARMAN' archaea depend on association with euryarchaeal host in culture and in situ. *Nature Communications*, *8*. https://doi.org/10.1038/s41467-017-00104-7

Graham, S. W., Olmstead, R. G., & Barrett, S. C. H. (2002). Rooting Phylogenetic Trees with Distant Outgroups: A Case Study from the Commelinoid Monocots. *Molecular Biology and Evolution*, *19*(10), 1769–1781. https://doi.org/10.1093/oxfordjournals.molbev.a003999

Graybeal, A. (1998). Is It Better to Add Taxa or Characters to a Difficult Phylogenetic Problem? *Systematic Biology*, *47*(1), 9–17. https://doi.org/10.1080/106351598260996

Gribaldo, S., & Brochier-Armanet, C. (2006). The origin and evolution of Archaea: A state of the art. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *361*(1470), 1007–1022. https://doi.org/10.1098/rstb.2006.1841

Groussin, M., Boussau, B., Szöllõsi, G., Eme, L., Gouy, M., Brochier-Armanet, C., & Daubin, V. (2016). Gene Acquisitions from Bacteria at the Origins of Major Archaeal Clades Are Vastly Overestimated. *Molecular Biology and Evolution*, *33*(2), 305–310. https://doi.org/10.1093/molbev/msv249

Groussin, M., & Gouy, M. (2011). Adaptation to Environmental Temperature Is a Major Determinant of Molecular Evolutionary Rates in Archaea. *Molecular Biology and Evolution*, *28*(9), 2661–2674. https://doi.org/10.1093/molbev/msr098

Gu, X., Hewett-Emmett, D., & Li, W.-H. (1998). Directional mutational pressure affects the amino acid composition and hydrophobicity of proteins in bacteria. In R. C. Woodruff & J. N. Thompson (Eds.), *Mutation and Evolution* (pp. 383–391). Springer Netherlands. https://doi.org/10.1007/978-94-011-5210-5_31

Gunde-Cimerman, N., Plemenitaš, A., & Oren, A. (2018). Strategies of adaptation of microorganisms of the three domains of life to high salt concentrations. *FEMS Microbiology Reviews*, *42*(3), 353–375. https://doi.org/10.1093/femsre/fuy009

Guy, L., & Ettema, T. J. G. (2011). The archaeal 'TACK' superphylum and the origin of eukaryotes. *Trends in Microbiology*, *19*(12), 580–587. https://doi.org/10.1016/j.tim.2011.09.002

Hamm, J. N., Erdmann, S., Eloe-Fadrosh, E. A., Angeloni, A., Zhong, L., Brownlee, C., Williams, T. J., Barton, K., Carswell, S., Smith, M. A., Brazendale, S., Hancock, A. M., Allen, M. A., Raftery, M. J., & Cavicchioli, R. (2019). Unexpected host dependency of Antarctic Nanohaloarchaeota. *Proceedings of the National Academy of Sciences*, *116*(29), 14661–14670.

Hamm, J., Y, L., Av, K., N, D., E, L., C, B., Emv, J., Rm, W., Mab, B., B, B., Tam, B., Ig, D., A, S., & R, C. (2023). *The parasitic lifestyle of an archaeal symbiont*. https://doi.org/10.1101/2023.02.24.529834

Hedtke, S. M., Townsend, T. M., & Hillis, D. M. (2006). Resolution of Phylogenetic Conflict in Large Data Sets by Increased Taxon Sampling. *Systematic Biology*, *55*(3), 522–529. https://doi.org/10.1080/10635150600697358

Henneberger, R., Moissl, C., Amann, T., Rudolph, C., & Huber, R. (2006). 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*(1), 192–199. https://doi.org/10.1128/AEM.72.1.192-199.2006

Hilario, E., & Gogarten, J. P. (1993). Horizontal transfer of ATPase genes—The tree of life becomes a net of life. *Bio Systems*, *31*(2–3), 111–119. https://doi.org/10.1016/0303-2647(93)90038-e

Hillis, D. M., Pollock, D. D., McGuire, J. A., & Zwickl, D. J. (2003). Is Sparse Taxon Sampling a Problem for Phylogenetic Inference? *Systematic Biology*, *52*(1), 124–126.

Hohn, M. J., Hedlund, B. P., & Huber, H. (2002). Detection of 16S rDNA Sequences Representing the Novel Phylum "Nanoarchaeota": Indication for a Wide Distribution in High Temperature Biotopes. *Systematic and Applied Microbiology*, *25*(4), 551–554. https://doi.org/10.1078/07232020260517698

Huber, H., Hohn, M. J., Rachel, R., Fuchs, T., Wimmer, V. C., & Stetter, K. O. (2002). A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont. *Nature*, *417*(6884), 63–67. https://doi.org/10.1038/417063a

Huber, H., Hohn, M. J., Stetter, K. O., & Rachel, R. (2003). The phylum Nanoarchaeota: Present knowledge and future perspectives of a unique form of life. *Research in Microbiology*, *154*(3), 165–171. https://doi.org/10.1016/S0923-2508(03)00035-4

Huelsenbeck, J. P., Larget, B., Miller, R. E., & Ronquist, F. (2002). Potential Applications and Pitfalls of Bayesian Inference of Phylogeny. *Systematic Biology*, *51*(5), 673–688. https://doi.org/10.1080/10635150290102366

Huelsenbeck, J. P., & Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, *17*(8), 754–755. https://doi.org/10.1093/bioinformatics/17.8.754

Hug, L. A., Baker, B. J., Anantharaman, K., Brown, C. T., Probst, A. J., Castelle, C. J., Butterfield, C. N., Hernsdorf, A. W., Amano, Y., Ise, K., Suzuki, Y., Dudek, N., Relman, D. A., Finstad, K. M., Amundson, R., Thomas, B. C., & Banfield, J. F. (2016). A new view of the tree of life. *Nature Microbiology*, *1*(5), Article 5. https://doi.org/10.1038/nmicrobiol.2016.48

Hugenholtz, P. (2002). Exploring prokaryotic diversity in the genomic era. *Genome Biology*, *3*(2), Article 2. https://doi.org/10.1186/gb-2002-3-2-reviews0003

Hugenholtz, P., Chuvochina, M., Oren, A., Parks, D. H., & Soo, R. M. (2021). Prokaryotic taxonomy and nomenclature in the age of big sequence data. *The ISME Journal*, *15*(7), Article 7. https://doi.org/10.1038/s41396-021-00941-x

Hyatt, D., Chen, G.-L., LoCascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010). Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, *11*(1), 119.

https://doi.org/10.1186/1471-2105-11-119

Ingalls, A. E., Shah, S. R., Hansman, R. L., Aluwihare, L. I., Santos, G. M., Druffel, E. R. M., & Pearson, A. (2006). Quantifying archaeal community autotrophy in the mesopelagic ocean using natural radiocarbon. *Proceedings of the National Academy of Sciences*, *103*(17), 6442–6447. https://doi.org/10.1073/pnas.0510157103

Jacox, E., Chauve, C., Szöllősi, G. J., Ponty, Y., & Scornavacca, C. (2016). ecceTERA: Comprehensive gene tree-species tree reconciliation using parsimony. *Bioinformatics*, *32*(13), 2056–2058. https://doi.org/10.1093/bioinformatics/btw105

Jaffe, A. L., Castelle, C. J., Dupont, C. L., & Banfield, J. F. (2019). Lateral Gene Transfer Shapes the Distribution of RuBisCO among Candidate Phyla Radiation Bacteria and DPANN Archaea. *Molecular Biology and Evolution*, *36*(3), 435–446. https://doi.org/10.1093/molbev/msy234

Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T., & Aluru, S. (2018). High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nature Communications*, *9*(1), 1–8. https://doi.org/10.1038/s41467-018-07641-9

James T. Staley & Allan Konopka. (1985). MEASUREMENT OF IN SITU ACTIVITIES OF NONPHOTOSYNTHETIC MICROORGANISMS IN AQUATIC AND TERRESTRIAL HABITATS. *Annual Review of Microbiology*, *Vol. 39:321-346*.

Kadnikov, V. V., Savvichev, A. S., Mardanov, A. V., Beletsky, A. V., Chupakov, A. V., Kokryatskaya, N. M., Pimenov, N. V., & Ravin, N. V. (2020). Metabolic Diversity and Evolutionary History of the Archaeal Phylum “Candidatus Micrarchaeota” Uncovered from a Freshwater Lake Metagenome. *Applied and Environmental Microbiology*, *86*(23). https://doi.org/10.1128/AEM.02199-20

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, *30*(4), 772–780. https://doi.org/10.1093/molbev/mst010

Kennedy, S. P., Ng, W. V., Salzberg, S. L., Hood, L., & DasSarma, S. (2001). Understanding the Adaptation of Halobacterium Species NRC-1 to Its Extreme Environment through Computational Analysis of Its Genome Sequence. *Genome Research*, *11*(10), 1641–1650.

https://doi.org/10.1101/gr.190201

Kizina, J., Jordan, S. F. A., Martens, G. A., Lonsing, A., Probian, C., Kolovou, A., Santarella-Mellwig, R., Rhiel, E., Littmann, S., Markert, S., Stüber, K., Richter, M., Schweder, T., & Harder, J. (2022). Methanosaeta and "Candidatus Velamenicoccus archaeovorus." *Applied and Environmental Microbiology*, *88*(7), e02407-21. https://doi.org/10.1128/aem.02407-21

Klingl, A. (2014). S-layer and cytoplasmic membrane – exceptions from the typical archaeal cell wall with a focus on double membranes. *Frontiers in Microbiology*, *5*. https://www.frontiersin.org/articles/10.3389/fmicb.2014.00624

Koonin, E. V., Makarova, K. S., & Aravind, L. (2001). Horizontal gene transfer in prokaryotes: Quantification and classification. *Annual Review of Microbiology*, *55*, 709–742. https://doi.org/10.1146/annurev.micro.55.1.709

Koski, L. B., & Golding, G. B. (2001). The Closest BLAST Hit Is Often Not the Nearest Neighbor. *Journal of Molecular Evolution*, *52*(6), 540–542. https://doi.org/10.1007/s002390010184

Kozubal, M. A., Romine, M., Jennings, R. deM, Jay, Z. J., Tringe, S. G., Rusch, D. B., Beam, J. P., McCue, L. A., & Inskeep, W. P. (2013). Geoarchaeota: A new candidate phylum in the Archaea from high-temperature acidic iron mats in Yellowstone National Park. *The ISME Journal*, *7*(3), 622–634. https://doi.org/10.1038/ismej.2012.132

Krause, S., Bremges, A., Münch, P. C., McHardy, A. C., & Gescher, J. (2017). Characterisation of a stable laboratory co-culture of acidophilic nanoorganisms. *Scientific Reports*, *7*(1), Article 1. https://doi.org/10.1038/s41598-017-03315-6

Kubo, K., Lloyd, K. G., F Biddle, J., Amann, R., Teske, A., & Knittel, K. (2012). Archaea of the Miscellaneous Crenarchaeotal Group are abundant, diverse and widespread in marine sediments. *The ISME Journal*, *6*(10), Article 10. https://doi.org/10.1038/ismej.2012.37

La Cono, V., Messina, E., Rohde, M., Arcadi, E., Ciordia, S., Crisafi, F., Denaro, R., Ferrer, M., Giuliano, L., Golyshin, P. N., Golyshina, O. V., Hallsworth, J. E., La Spada, G., Mena, M. C., Merkel, A. Y., Shevchenko, M. A., Smedile, F., Sorokin, D. Y., Toshchakov, S. V., & Yakimov, M. M. (2020). Symbiosis between

nanohaloarchaeon and haloarchaeon is based on utilization of different polysaccharides. *Proceedings of the National Academy of Sciences*, *117*(33), 20223–20234. https://doi.org/10.1073/pnas.2007232117

Lanyi, J. K. (1974). Salt-dependent properties of proteins from extremely halophilic bacteria. *Bacteriological Reviews*, *38*(3), 272–290.

Lartillot, N., Brinkmann, H., & Philippe, H. (2007). Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC Evolutionary Biology*, *7*(Suppl 1), S4. https://doi.org/10.1186/1471-2148-7-S1-S4

Lartillot, N., Lepage, T., & Blanquart, S. (2009). PhyloBayes 3: A Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics*, *25*(17), 2286–2288. https://doi.org/10.1093/bioinformatics/btp368

Lartillot, N., & Philippe, H. (2004). A Bayesian Mixture Model for Across-Site Heterogeneities in the Amino-Acid Replacement Process. *Molecular Biology and Evolution*, *21*(6), 1095–1109. https://doi.org/10.1093/molbev/msh112

Lartillot, N., Rodrigue, N., Stubbs, D., & Richer, J. (2013). PhyloBayes MPI: Phylogenetic Reconstruction with Infinite Mixtures of Profiles in a Parallel Environment. *Systematic Biology*, *62*(4), 611–615. https://doi.org/10.1093/sysbio/syt022

Le, S. Q., & Gascuel, O. (2008). An Improved General Amino Acid Replacement Matrix. *Molecular Biology and Evolution*, *25*(7), 1307–1320. https://doi.org/10.1093/molbev/msn067

Letunic, I., & Bork, P. (2007). Interactive Tree Of Life (iTOL): An online tool for phylogenetic tree display and annotation. *Bioinformatics*, *23*(1), 127–128. https://doi.org/10.1093/bioinformatics/btl529

Li, L., Stoeckert, C. J., & Roos, D. S. (2003). OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. *Genome Research*, *13*(9), 2178–2189. https://doi.org/10.1101/gr.1224503

Lopez, P., Casane, D., & Philippe, H. (2002). Heterotachy, an Important Process of Protein Evolution. *Molecular Biology and Evolution*, *19*(1), 1–7. https://doi.org/10.1093/oxfordjournals.molbev.a003973

López-García, P., & Moreira, D. (2021). Physical connections: Prokaryotes

parasitizing their kin. *Environmental Microbiology Reports*, *13*(1), 54–61. https://doi.org/10.1111/1758-2229.12910

López-García, P., Moreira, D., López-López, A., & Rodríguez-Valera, F. (2001). A novel haloarchaeal-related lineage is widely distributed in deep oceanic regions. *Environmental Microbiology*, *3*(1), 72–78. https://doi.org/10.1046/j.1462-2920.2001.00162.x

López-Pérez, M., Martin-Cuadrado, A.-B., & Rodriguez-Valera, F. (2014). Homologous recombination is involved in the diversity of replacement flexible genomic islands in aquatic prokaryotes. *Frontiers in Genetics*, *5*. https://doi.org/10.3389/fgene.2014.00147

Madern, D., Ebel, C., & Zaccai, G. (2000). Halophilic adaptation of enzymes. *Extremophiles*, *4*(2), 91–98. https://doi.org/10.1007/s007920050142

Mahendrarajah, T. A., Moody, E. R. R., Schrempf, D., Szánthó, L. L., Dombrowski, N., Davín, A. A., Pisani, D., Donoghue, P. C. J., Szöllősi, G. J., Williams, T. A., & Spang, A. (2023). *ATP synthase evolution on a cross-braced dated tree of life* (p. 2023.04.11.536006). bioRxiv. https://doi.org/10.1101/2023.04.11.536006

Makarova, K. S., Aravind, L., Galperin, M. Y., Grishin, N. V., Tatusov, R. L., Wolf, Y. I., & Koonin, E. V. (1999). Comparative genomics of the Archaea (Euryarchaeota): Evolution of conserved protein families, the stable core, and the variable shell. *Genome Research*, *9*(7), 608–628.

Makarova, K. S., & Koonin, E. V. (2003). Comparative genomics of archaea: How much have we learned in six years, and what's next? *Genome Biology*, *4*(8), 115. https://doi.org/10.1186/gb-2003-4-8-115

Makarova, K. S., & Koonin, E. V. (2005). Evolutionary and functional genomics of the Archaea. *Current Opinion in Microbiology*, *8*(5), 586–594. https://doi.org/10.1016/j.mib.2005.08.003

Martijn, J., Schön, M. E., Lind, A. E., Vosseberg, J., Williams, T. A., Spang, A., & Ettema, T. J. G. (2020). Hikarchaeia demonstrate an intermediate stage in the methanogen-to-halophile transition. *Nature Communications*, *11*(1), Article 1. https://doi.org/10.1038/s41467-020-19200-2

Martinez-Gutierrez, C. A., & Aylward, F. O. (2021). Phylogenetic Signal, Congruence, and Uncertainty across Bacteria and Archaea. *Molecular Biology and Evolution*, *38*(12), 5514–5527. https://doi.org/10.1093/molbev/msab254

Méheust, R., Burstein, D., Castelle, C. J., & Banfield, J. F. (2019). The distinction of CPR bacteria from other bacteria based on protein family content. *Nature Communications*, *10*(1), Article 1. https://doi.org/10.1038/s41467-019-12171-z

Menardo, F., Loiseau, C., Brites, D., Coscolla, M., Gygli, S. M., Rutaihwa, L. K., Trauner, A., Beisel, C., Borrell, S., & Gagneux, S. (2018). Treemmer: A tool to reduce large phylogenetic datasets with minimal loss of diversity. *BMC Bioinformatics*, *19*(1), 164. https://doi.org/10.1186/s12859-018-2164-8

Merino, N., Aronson, H. S., Bojanova, D. P., Feyhl-Buska, J., Wong, M. L., Zhang, S., & Giovannelli, D. (2019). Living at the Extremes: Extremophiles and the Limits of Life in a Planetary Context. *Frontiers in Microbiology*, *10*. https://doi.org/10.3389/fmicb.2019.00780

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., & Lanfear, R. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution*, *37*(5), 1530–1534. https://doi.org/10.1093/molbev/msaa015

Mongodin, E. F., Nelson, K. E., Daugherty, S., DeBoy, R. T., Wister, J., Khouri, H., Weidman, J., Walsh, D. A., Papke, R. T., Sanchez Perez, G., Sharma, A. K., Nesbø, C. L., MacLeod, D., Bapteste, E., Doolittle, W. F., Charlebois, R. L., Legault, B., & Rodriguez-Valera, F. (2005). The genome of Salinibacter ruber: Convergence and gene exchange among hyperhalophilic bacteria and archaea. *Proceedings of the National Academy of Sciences*, *102*(50), 18147–18152. https://doi.org/10.1073/pnas.0509073102

Moody, E. R., Mahendrarajah, T. A., Dombrowski, N., Clark, J. W., Petitjean, C., Offre, P., Szöllősi, G. J., Spang, A., & Williams, T. A. (2022). An estimate of the deepest branches of the tree of life from ancient vertically-evolving genes. *eLife*, *11*, e66695. https://doi.org/10.7554/eLife.66695

Moran, N. A., & Bennett, G. M. (2014). The tiniest tiny genomes. *Annual Review of Microbiology*, *68*, 195–215. https://doi.org/10.1146/annurev-micro-091213-112901

Müller-Santos, M., de Souza, E. M., Pedrosa, F. de O., Mitchell, D. A., Longhi, S., Carrière, F., Canaan, S., & Krieger, N. (2009). First evidence for the salt-dependent folding and activity of an esterase from the halophilic archaea Haloarcula marismortui. *Biochimica et Biophysica Acta (BBA) - Molecular and*

Cell Biology of Lipids, *1791*(8), 719–729. https://doi.org/10.1016/j.bbalip.2009.03.006

Muñoz-Gómez, S. A., Hess, S., Burger, G., Lang, B. F., Susko, E., Slamovits, C. H., & Roger, A. J. (2019). An updated phylogeny of the Alphaproteobacteria reveals that the parasitic Rickettsiales and Holosporales have independent origins. *eLife*, *8*, e42535. https://doi.org/10.7554/eLife.42535

Muñoz-Gómez, S. A., Susko, E., Williamson, K., Eme, L., Slamovits, C. H., Moreira, D., López-García, P., & Roger, A. J. (2022). Site-and-branch-heterogeneous analyses of an expanded dataset favour mitochondria as sister to known Alphaproteobacteria. *Nature Ecology & Evolution*, *6*(3), Article 3. https://doi.org/10.1038/s41559-021-01638-2

MURRAY, R. G. E., & STACKEBRANDT, E. (1995). Taxonomic Note: Implementation of the Provisional Status Candidatus for Incompletely Described Procaryotes. *International Journal of Systematic and Evolutionary Microbiology*, *45*(1), 186–187. https://doi.org/10.1099/00207713-45-1-186

Narasingarao, P., Podell, S., Ugalde, J. A., Brochier-Armanet, C., Emerson, J. B., Brocks, J. J., Heidelberg, K. B., Banfield, J. F., & Allen, E. E. (2012a). De novo metagenomic assembly reveals abundant novel major lineage of Archaea in hypersaline microbial communities. *The ISME Journal*, *6*(1), 81–93. https://doi.org/10.1038/ismej.2011.78

Narasingarao, P., Podell, S., Ugalde, J. A., Brochier-Armanet, C., Emerson, J. B., Brocks, J. J., Heidelberg, K. B., Banfield, J. F., & Allen, E. E. (2012b). De novo metagenomic assembly reveals abundant novel major lineage of Archaea in hypersaline microbial communities. *The ISME Journal*, *6*(1), 81–93. https://doi.org/10.1038/ismej.2011.78

Nelson-Sathi, S., Dagan, T., Landan, G., Janssen, A., Steel, M., McInerney, J. O., Deppenmeier, U., & Martin, W. F. (2012). Acquisition of 1,000 eubacterial genes physiologically transformed a methanogen at the origin of Haloarchaea. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(50), 20537–20542. https://doi.org/10.1073/pnas.1209119109

Nelson-Sathi, S., Sousa, F. L., Roettger, M., Lozada-Chávez, N., Thiergart, T., Janssen, A., Bryant, D., Landan, G., Schönheit, P., Siebers, B., McInerney, J. O., &

Martin, W. F. (2015). Origins of major archaeal clades correspond to gene acquisitions from bacteria. *Nature*, *517*(7532), Article 7532. https://doi.org/10.1038/nature13805

Nguyen, N. D., Mirarab, S., Kumar, K., & Warnow, T. (2015). Ultra-large alignments using phylogeny-aware profiles. *Genome Biology*, *16*(1), Article 1. https://doi.org/10.1186/s13059-015-0688-z

Nunoura, T., Takaki, Y., Kakuta, J., Nishi, S., Sugahara, J., Kazama, H., Chee, G.-J., Hattori, M., Kanai, A., Atomi, H., Takai, K., & Takami, H. (2011). Insights into the evolution of Archaea and eukaryotic protein modifier systems revealed by the genome of a novel archaeal group. *Nucleic Acids Research*, *39*(8), 3204–3223. https://doi.org/10.1093/nar/gkq1228

Ogden, T. H., & Rosenberg, M. S. (2006). Multiple Sequence Alignment Accuracy and Phylogenetic Inference. *Systematic Biology*, *55*(2), 314–328. https://doi.org/10.1080/10635150500541730

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., … Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, *44*(D1), D733-745. https://doi.org/10.1093/nar/gkv1189

Olendzenski, L., Zhaxybayeva, O., & Gogarten, J. P. (2002). What's in a Tree? In J. Seckbach (Ed.), *Symbiosis: Mechanisms and Model Systems* (pp. 65–79). Springer Netherlands. https://doi.org/10.1007/0-306-48173-1_4

Oren, A. (2002). Diversity of halophilic microorganisms: Environments, phylogeny, physiology, and applications. *Journal of Industrial Microbiology and Biotechnology*, *28*(1), 56–63. https://doi.org/10.1038/sj/jim/7000176

Pace, N. R. (1997). A Molecular View of Microbial Diversity and the Biosphere. *Science*, *276*(5313), 734–740. https://doi.org/10.1126/science.276.5313.734

Pace, N. R. (2009). Mapping the Tree of Life: Progress and Prospects. *Microbiology and Molecular Biology Reviews : MMBR*, *73*(4), 565–576. https://doi.org/10.1128/MMBR.00033-09

Pallen, M. J. (2021). The status Candidatus for uncultured taxa of Bacteria and Archaea: SWOT analysis. *International Journal of Systematic and Evolutionary*

Microbiology, *71*(9), 005000. https://doi.org/10.1099/ijsem.0.005000

Parker, C. T., Tindall, B. J., & Garrity, G., George. (2019). *International Code of Nomenclature of Prokaryotes | Microbiology Society*. https://www.microbiologyresearch.org/content/journal/ijsem/10.1099/ijsem.0.000778

Parks, D. H., Chuvochina, M., Rinke, C., Mussig, A. J., Chaumeil, P.-A., & Hugenholtz, P. (2022). GTDB: An ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Research*, *50*(D1), D785–D794. https://doi.org/10.1093/nar/gkab776

Parks, D. H., Chuvochina, M., Waite, D. W., Rinke, C., Skarshewski, A., Chaumeil, P.-A., & Hugenholtz, P. (2018). A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nature Biotechnology*, *36*(10), 996–1004. https://doi.org/10.1038/nbt.4229

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., & Tyson, G. W. (2015). CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research*, *25*(7), 1043–1055. https://doi.org/10.1101/gr.186072.114

Paul, S., Bag, S. K., Das, S., Harvill, E. T., & Dutta, C. (2008). Molecular signature of hypersaline adaptation: Insights from genome and proteome composition of halophilic prokaryotes. *Genome Biology*, *9*(4), R70. https://doi.org/10.1186/gb-2008-9-4-r70

Petitjean, C., Deschamps, P., López-García, P., & Moreira, D. (2015). Rooting the Domain Archaea by Phylogenomic Analysis Supports the Foundation of the New Kingdom Proteoarchaeota. *Genome Biology and Evolution*, *7*(1), 191–204. https://doi.org/10.1093/gbe/evu274

Petitjean, C., Deschamps, P., López-García, P., Moreira, D., & Brochier-Armanet, C. (2015). Extending the conserved phylogenetic core of archaea disentangles the evolution of the third domain of life. *Molecular Biology and Evolution*, *32*(5), 1242–1254. https://doi.org/10.1093/molbev/msv015

Philippe, H., & Laurent, J. (1998). How good are deep phylogenetic trees? *Current Opinion in Genetics & Development*, *8*(6), 616–623. https://doi.org/10.1016/S0959-437X(98)80028-2

Philippe, H., Lopez, P., Brinkmann, H., Budin, K., Germot, A., Laurent, J., Moreira, D., Müller, M., & Le Guyader, H. (2000). Early-branching or fast-evolving eukaryotes? An answer based on slowly evolving positions. *Proceedings of the Royal Society B: Biological Sciences*, *267*(1449), 1213–1221.

Podar, M., Anderson, I., Makarova, K. S., Elkins, J. G., Ivanova, N., Wall, M. A., Lykidis, A., Mavromatis, K., Sun, H., Hudson, M. E., Chen, W., Deciu, C., Hutchison, D., Eads, J. R., Anderson, A., Fernandes, F., Szeto, E., Lapidus, A., Kyrpides, N. C., … Stetter, K. O. (2008). A genomic analysis of the archaeal system Ignicoccus hospitalis-Nanoarchaeum equitans. *Genome Biology*, *9*(11), R158. https://doi.org/10.1186/gb-2008-9-11-r158

Podar, M., Makarova, K. S., Graham, D. E., Wolf, Y. I., Koonin, E. V., & Reysenbach, A.-L. (2013). Insights into archaeal evolution and symbiosis from the genomes of a nanoarchaeon and its inferred crenarchaeal host from Obsidian Pool, Yellowstone National Park. *Biology Direct*, *8*(1), 9. https://doi.org/10.1186/1745-6150-8-9

Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLOS ONE*, *5*(3), e9490. https://doi.org/10.1371/journal.pone.0009490

Probst, A. J., Ladd, B., Jarett, J. K., Geller-McGrath, D. E., Sieber, C. M. K., Emerson, J. B., Anantharaman, K., Thomas, B. C., Malmstrom, R. R., Stieglmeier, M., Klingl, A., Woyke, T., Ryan, M. C., & Banfield, J. F. (2018). Differential depth distribution of microbial function and putative symbionts through sediment-hosted aquifers in the deep terrestrial subsurface. *Nature Microbiology*, *3*(3), Article 3. https://doi.org/10.1038/s41564-017-0098-y

Probst, A. J., & Moissl-Eichinger, C. (2015). "Altiarchaeales": Uncultivated Archaea from the Subsurface. *Life*, *5*(2), 1381–1395. https://doi.org/10.3390/life5021381

Probst, A. J., Weinmaier, T., Raymann, K., Perras, A., Emerson, J. B., Rattei, T., Wanner, G., Klingl, A., Berg, I. A., Yoshinaga, M., Viehweger, B., Hinrichs, K.-U., Thomas, B. C., Meck, S., Auerbach, A. K., Heise, M., Schintlmeister, A., Schmid, M., Wagner, M., … Moissl-Eichinger, C. (2014). Biology of a widespread uncultivated archaeon that contributes to carbon fixation in the subsurface. *Nature Communications*, *5*(1), Article 1. https://doi.org/10.1038/ncomms6497

Quang, L. S., Gascuel, O., & Lartillot, N. (2008). Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics (Oxford, England)*, *24*(20), 2317–2323. https://doi.org/10.1093/bioinformatics/btn445

Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J., & Segata, N. (2017). Shotgun metagenomics, from sampling to analysis. *Nature Biotechnology*, *35*(9), Article 9. https://doi.org/10.1038/nbt.3935

Rambaut, A., Lam, T. T., Max Carvalho, L., & Pybus, O. G. (2016). Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evolution*, *2*(1), vew007. https://doi.org/10.1093/ve/vew007

Rappé, M. S., & Giovannoni, S. J. (2003). The uncultured microbial majority. *Annual Review of Microbiology*, *57*, 369–394. https://doi.org/10.1146/annurev.micro.57.030502.090759

Raymann, K., Brochier-Armanet, C., & Gribaldo, S. (2015). The two-domain tree of life is linked to a new root for the Archaea. *Proceedings of the National Academy of Sciences*, *112*(21), 6670–6675. https://doi.org/10.1073/pnas.1420858112

Reigstad, L. J., Jorgensen, S. L., & Schleper, C. (2010). Diversity and abundance of Korarchaeota in terrestrial hot springs of Iceland and Kamchatka. *The ISME Journal*, *4*(3), Article 3. https://doi.org/10.1038/ismej.2009.126

Reuveni, S., Ehrenberg, M., & Paulsson, J. (2017). Ribosomes are optimized for autocatalytic production. *Nature*, *547*(7663), 293–297. https://doi.org/10.1038/nature22998

Rinke, C., Chuvochina, M., Mussig, A. J., Chaumeil, P.-A., Davín, A. A., Waite, D. W., Whitman, W. B., Parks, D. H., & Hugenholtz, P. (2021). A standardized archaeal taxonomy for the Genome Taxonomy Database. *Nature Microbiology*, *6*(7), Article 7. https://doi.org/10.1038/s41564-021-00918-8

Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N. N., Anderson, I. J., Cheng, J.-F., Darling, A., Malfatti, S., Swan, B. K., Gies, E. A., Dodsworth, J. A., Hedlund, B. P., Tsiamis, G., Sievert, S. M., Liu, W.-T., Eisen, J. A., Hallam, S. J., Kyrpides, N. C., Stepanauskas, R., … Woyke, T. (2013). Insights into the phylogeny and coding potential of microbial dark matter. *Nature*, *499*(7459), 431–437. https://doi.org/10.1038/nature12352

Rivera, M. C., Jain, R., Moore, J. E., & Lake, J. A. (1998). Genomic evidence for two

functionally distinct gene classes. *Proceedings of the National Academy of Sciences*, *95*(11), 6239–6244. https://doi.org/10.1073/pnas.95.11.6239

Sakai, H. D., Nur, N., Kato, S., Yuki, M., Shimizu, M., Itoh, T., Ohkuma, M., Suwanto, A., & Kurosawa, N. (2022). Insight into the symbiotic lifestyle of DPANN archaea revealed by cultivation and genome analyses. *Proceedings of the National Academy of Sciences*, *119*(3). https://doi.org/10.1073/pnas.2115449119

Salomaki, E. D., Eme, L., Brown, M. W., & Kolisko, M. (2020). Releasing uncurated datasets is essential for reproducible phylogenomics. *Nature Ecology & Evolution*, *4*(11), Article 11. https://doi.org/10.1038/s41559-020-01296-w

Sapp, J. (2005). The Prokaryote-Eukaryote Dichotomy: Meanings and Mythology. *Microbiology and Molecular Biology Reviews*, *69*(2), 292–305. https://doi.org/10.1128/MMBR.69.2.292-305.2005

Saw, J. H., Spang, A., Zaremba-Niedzwiedzka, K., Juzokaite, L., Dodsworth, J. A., Murugapiran, S. K., Colman, D. R., Takacs-Vesbach, C., Hedlund, B. P., Guy, L., & Ettema, T. J. G. (2015). Exploring microbial dark matter to resolve the deep archaeal ancestry of eukaryotes. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *370*(1678), 20140328. https://doi.org/10.1098/rstb.2014.0328

Schwank, K., Bornemann, T. L. V., Dombrowski, N., Spang, A., Banfield, J. F., & Probst, A. J. (2019). An archaeal symbiont-host association from the deep terrestrial subsurface. *The ISME Journal*, *13*(8), Article 8. https://doi.org/10.1038/s41396-019-0421-0

Segerer, A. H., Burggraf, S., Fiala, G., Huber, G., Huber, R., Pley, U., & Stetter, K. O. (1993). Life in hot springs and hydrothermal vents. *Origins of Life and Evolution of the Biosphere*, *23*(1), 77–90. https://doi.org/10.1007/BF01581992

Setubal, J. C. (2021). Metagenome-assembled genomes: Concepts, analogies, and challenges. *Biophysical Reviews*, *13*(6), 905–909. https://doi.org/10.1007/s12551-021-00865-y

Seymour, C. O., Palmer, M., Becraft, E. D., Stepanauskas, R., Friel, A. D., Schulz, F., Woyke, T., Eloe-Fadrosh, E., Lai, D., Jiao, J. Y., Hua, Z. S., Liu, L., Lian, Z. H., Li, W. J., Chuvochina, M., Finley, B. K., Koch, B. J., Schwartz, E., Dijkstra, P., … Hedlund, B. P. (2023). Hyperactive nanobacteria with host-dependent traits

pervade Omnitrophota. *Nature Microbiology*, *8*(4), 727–744. https://doi.org/10.1038/s41564-022-01319-1

Shavit, L., Penny, D., Hendy, M. D., & Holland, B. R. (2007). The Problem of Rooting Rapid Radiations. *Molecular Biology and Evolution*, *24*(11), 2400–2411. https://doi.org/10.1093/molbev/msm178

Siglioccolo, A., Paiardini, A., Piscitelli, M., & Pascarella, S. (2011). Structural adaptation of extreme halophilic proteins through decrease of conserved hydrophobic contact surface. *BMC Structural Biology*, *11*(1), 50. https://doi.org/10.1186/1472-6807-11-50

Singer, G. A. C., & Hickey, D. A. (2003). Thermophilic prokaryotes have characteristic patterns of codon usage, amino acid composition and nucleotide content. *Gene*, *317*(1–2), 39–47. https://doi.org/10.1016/s0378-1119(03)00660-7

Smythe, A. B., Sanderson, M. J., & Nadler, S. A. (2006). Nematode Small Subunit Phylogeny Correlates with Alignment Parameters. *Systematic Biology*, *55*(6), 972–992. https://doi.org/10.1080/10635150601089001

Sorokin, D. Y., Makarova, K. S., Abbas, B., Ferrer, M., Golyshin, P. N., Galinski, E. A., Ciorda, S., Mena, M. C., Merkel, A. Y., Wolf, Y. I., van Loosdrecht, M. C. M., & Koonin, E. V. (2019). Reply to 'Evolutionary placement of Methanonatronarchaeia.' *Nature Microbiology*, *4*(4), 560–561. https://doi.org/10.1038/s41564-019-0358-0

Sorokin, D. Y., Makarova, K. S., Abbas, B., Ferrer, M., Golyshin, P. N., Galinski, E. A., Ciordia, S., Mena, M. C., Merkel, A. Y., Wolf, Y. I., van Loosdrecht, M. C. M., & Koonin, E. V. (2017). Discovery of extremely halophilic, methyl-reducing euryarchaea provides insights into the evolutionary origin of methanogenesis. *Nature Microbiology*, *2*(8), 17081. https://doi.org/10.1038/nmicrobiol.2017.81

Spang, A., Caceres, E. F., & Ettema, T. J. G. (2017). Genomic exploration of the diversity, ecology, and evolution of the archaeal domain of life. *Science*, *357*(6351). https://doi.org/10.1126/science.aaf3883

Spang, A., Martijn, J., Saw, J. H., Lind, A. E., Guy, L., & Ettema, T. J. G. (2013, November 19). *Close Encounters of the Third Domain: The Emerging Genomic View of Archaeal Diversity and Evolution* [Review Article]. Archaea; Hindawi. https://doi.org/10.1155/2013/202358

Spang, A., Saw, J. H., Jørgensen, S. L., Zaremba-Niedzwiedzka, K., Martijn, J., Lind, A. E., van Eijk, R., Schleper, C., Guy, L., & Ettema, T. J. G. (2015). Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature*, *521*(7551), Article 7551. https://doi.org/10.1038/nature14447

Spang, A., Stairs, C. W., Dombrowski, N., Eme, L., Lombard, J., Caceres, E. F., Greening, C., Baker, B. J., & Ettema, T. J. G. (2019). Proposal of the reverse flow model for the origin of the eukaryotic cell based on comparative analyses of Asgard archaeal metabolism. *Nature Microbiology*, *4*(7), 1138–1148. https://doi.org/10.1038/s41564-019-0406-9

St. John, E., Liu, Y., Podar, M., Stott, M. B., Meneghin, J., Chen, Z., Lagutin, K., Mitchell, K., & Reysenbach, A.-L. (2019). A new symbiotic nanoarchaeote (Candidatus Nanoclepta minutus) and its host (Zestosphaera tikiterensis gen. Nov., sp. Nov.) from a New Zealand hot spring. *Systematic and Applied Microbiology*, *42*(1), 94–106. https://doi.org/10.1016/j.syapm.2018.08.005

Stetter, K. O. (1999). Extremophiles and their adaptation to hot environments. *FEBS Letters*, *452*(1), 22–25. https://doi.org/10.1016/S0014-5793(99)00663-8

Sun, W., Jiao, C., Xiao, Y., Wang, L., Yu, C., Liu, J., Yu, Y., & Wang, L. (2016). Salt-Dependent Aggregation and Assembly of E coli-Expressed Ferritin. *Dose-Response*, *14*(1), 1559325816632102. https://doi.org/10.1177/1559325816632102

Susko, E., Field, C., Blouin, C., & Roger, A. J. (2003). Estimation of Rates-Across-Sites Distributions in Phylogenetic Substitution Models. *Systematic Biology*, *52*(5), 594–603. https://doi.org/10.1080/10635150390235395

Susko, E., & Roger, A. J. (2021). Long Branch Attraction Biases in Phylogenetics. *Systematic Biology*, *70*(4), 838–843. https://doi.org/10.1093/sysbio/syab001

Szöllősi, G. J., Rosikiewicz, W., Boussau, B., Tannier, E., & Daubin, V. (2013a). Efficient Exploration of the Space of Reconciled Gene Trees. *Systematic Biology*, *62*(6), 901–912. https://doi.org/10.1093/sysbio/syt054

Szöllősi, G. J., Rosikiewicz, W., Boussau, B., Tannier, E., & Daubin, V. (2013b). Efficient Exploration of the Space of Reconciled Gene Trees. *Systematic Biology*, *62*(6), 901–912. https://doi.org/10.1093/sysbio/syt054

Taib, N., Megrian, D., Witwinowski, J., Adam, P., Poppleton, D., Borrel, G., Beloin, C., & Gribaldo, S. (2020). Genome-wide analysis of the Firmicutes illuminates the

diderm/monoderm transition. *Nature Ecology & Evolution*, *4*(12), Article 12. https://doi.org/10.1038/s41559-020-01299-7

Talavera, G., & Castresana, J. (2007). Improvement of Phylogenies after Removing Divergent and Ambiguously Aligned Blocks from Protein Sequence Alignments. *Systematic Biology*, *56*(4), 564–577. https://doi.org/10.1080/10635150701472164

Thomson, N. R., Sebaihia, M., Cerdeño-Tárraga, A. M., Holden, M. T. G., & Parkhill, J. (2004). Shrinking genomics. *Nature Reviews Microbiology*, *2*(1), Article 1. https://doi.org/10.1038/nrmicro800

Van de Peer, Y., Chapelle, S., & De Wachter, R. (1996). A Quantitative Map of Nucleotide Substitution Rates in Bacterial rRNA. *Nucleic Acids Research*, *24*(17), 3381–3391. https://doi.org/10.1093/nar/24.17.3381

Vanwonterghem, I., Evans, P. N., Parks, D. H., Jensen, P. D., Woodcroft, B. J., Hugenholtz, P., & Tyson, G. W. (2016). Methylotrophic methanogenesis discovered in the archaeal phylum Verstraetearchaeota. *Nature Microbiology*, *1*(12), Article 12. https://doi.org/10.1038/nmicrobiol.2016.170

Vetriani, C., Jannasch, H. W., MacGregor, B. J., Stahl, D. A., & Reysenbach, A.-L. (1999). Population Structure and Phylogenetic Characterization of Marine Benthic Archaea in Deep-Sea Sediments. *Applied and Environmental Microbiology*, *65*(10), 4375–4384. https://doi.org/10.1128/AEM.65.10.4375-4384.1999

Waters, E., Hohn, M. J., Ahel, I., Graham, D. E., Adams, M. D., Barnstead, M., Beeson, K. Y., Bibbs, L., Bolanos, R., Keller, M., Kretz, K., Lin, X., Mathur, E., Ni, J., Podar, M., Richardson, T., Sutton, G. G., Simon, M., Söll, D., … Noordewier, M. (2003). The genome of Nanoarchaeum equitans: Insights into early archaeal evolution and derived parasitism. *Proceedings of the National Academy of Sciences*, *100*(22), 12984–12988. https://doi.org/10.1073/pnas.1735403100

West-Roberts, J., Valentin-Alvarado, L., Mullen, S., Sachdeva, R., Smith, J., Hug, L. A., Gregoire, D. S., Liu, W., Lin, T.-Y., Husain, G., Amano, Y., Ly, L., & Banfield, J. F. (2023). *Giant genes are rare but implicated in cell wall degradation by predatory bacteria* (p. 2023.11.21.568195). bioRxiv. https://doi.org/10.1101/2023.11.21.568195

Whelan, S., & Goldman, N. (2001). A General Empirical Model of Protein Evolution

Derived from Multiple Protein Families Using a Maximum-Likelihood Approach. *Molecular Biology and Evolution*, *18*(5), 691–699. https://doi.org/10.1093/oxfordjournals.molbev.a003851

Williams, T. A., Szöllősi, G. J., Spang, A., Foster, P. G., Heaps, S. E., Boussau, B., Ettema, T. J. G., & Embley, T. M. (2017). Integrative modeling of gene and genome evolution roots the archaeal tree of life. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(23), E4602–E4611. https://doi.org/10.1073/pnas.1618463114

Woese, C. (1990). *Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria, and Eucarya.* https://doi.org/10.1073/pnas.87.12.4576

Woese, C. R., Fox, G. E., Zablen, L., Uchida, T., Bonen, L., Pechman, K., Lewis, B. J., & Stahl, D. (1975). Conservation of primary structure in 16S ribosomal RNA. *Nature*, *254*(5495), Article 5495. https://doi.org/10.1038/254083a0

Woese, C. R., Magrum, L. J., & Fox, G. E. (1978). Archaebacteria. *Journal of Molecular Evolution*, *11*(3), 245–252. https://doi.org/10.1007/BF01734485

Wurch, L., Giannone, R. J., Belisle, B. S., Swift, C., Utturkar, S., Hettich, R. L., Reysenbach, A.-L., & Podar, M. (2016). Genomics-informed isolation and characterization of a symbiotic Nanoarchaeota system from a terrestrial geothermal environment. *Nature Communications*, *7*(1), Article 1. https://doi.org/10.1038/ncomms12115

Yang, Z. (1994). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *Journal of Molecular Evolution*, *39*(3), 306–314. https://doi.org/10.1007/BF00160154

Zaremba-Niedzwiedzka, K., Caceres, E. F., Saw, J. H., Bäckström, D., Juzokaite, L., Vancaester, E., Seitz, K. W., Anantharaman, K., Starnawski, P., Kjeldsen, K. U., Stott, M. B., Nunoura, T., Banfield, J. F., Schramm, A., Baker, B. J., Spang, A., & Ettema, T. J. G. (2017). Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature*, *541*(7637), Article 7637. https://doi.org/10.1038/nature21031

Zhang, I. H., Borer, B., Zhao, R., Wilbert, S., Newman, D. K., & Babbin, A. R. (2023). *Uncultivated DPANN archaea are ubiquitous inhabitants of global oxygen deficient zones with diverse metabolic potential* (p. 2023.10.30.564641). bioRxiv. https://doi.org/10.1101/2023.10.30.564641

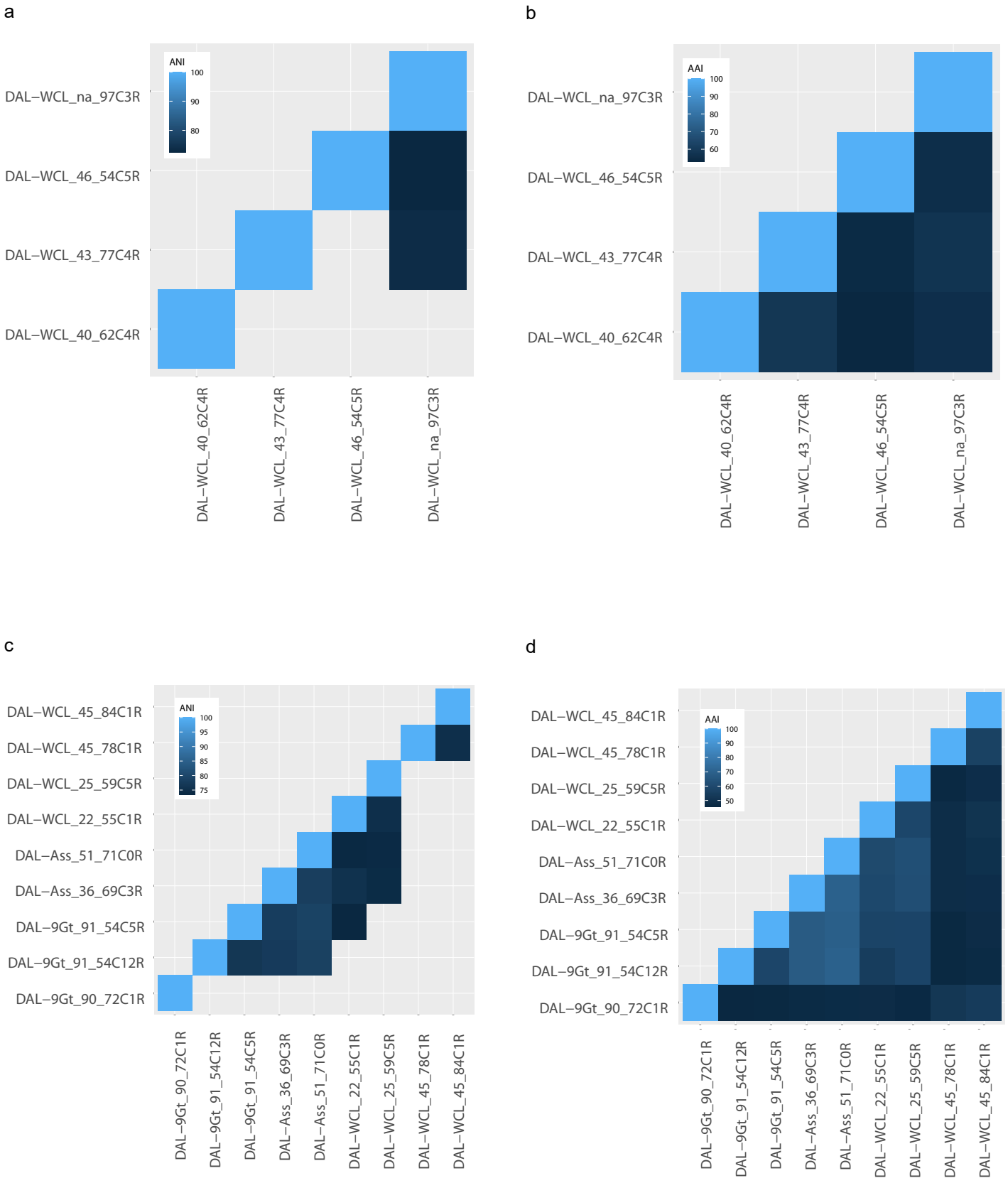
Zhou, H., Zhao, D., Zhang, S., Xue, Q., Zhang, M., Yu, H., Zhou, J., Li, M., Kumar, S., & Xiang, H. (n.d.). Metagenomic insights into the environmental adaptation and metabolism of Candidatus Haloplasmatales, one archaeal order thriving in saline lakes. *Environmental Microbiology*, *n/a*(n/a). https://doi.org/10.1111/1462-2920.15899

Zhou, H., Zhao, D., Zhang, S., Xue, Q., Zhang, M., Yu, H., Zhou, J., Li, M., Kumar, S., & Xiang, H. (2022). Metagenomic insights into the environmental adaptation and metabolism of *Candidatus* Haloplasmatales, one archaeal order thriving in saline lakes. *Environmental Microbiology*, *24*(5), 2239–2258. https://doi.org/10.1111/1462-2920.15899

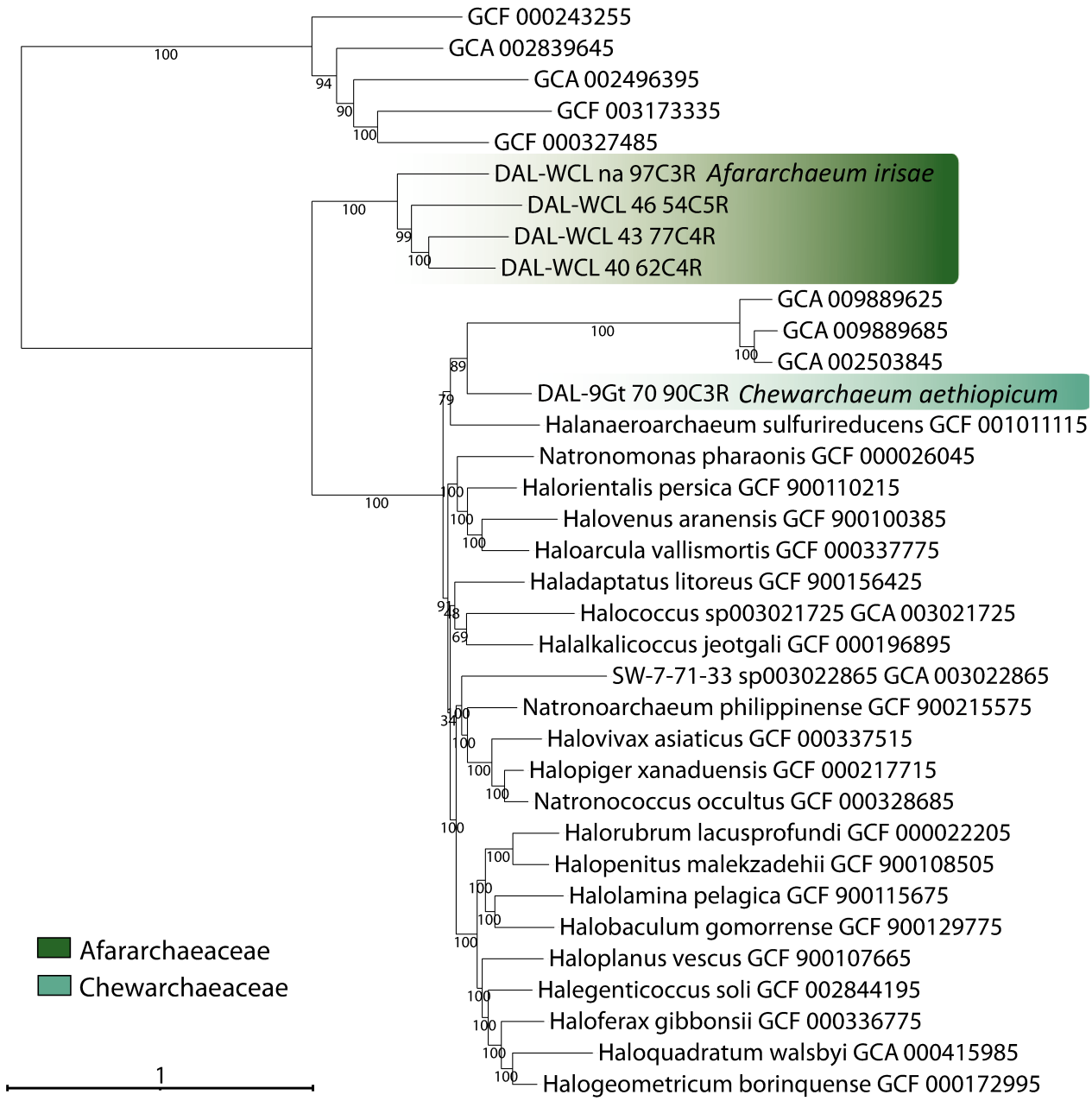
Zhu, Q., Mai, U., Pfeiffer, W., Janssen, S., Asnicar, F., Sanders, J. G., Belda-Ferre, P., Al-Ghalith, G. A., Kopylova, E., McDonald, D., Kosciolek, T., Yin, J. B., Huang, S., Salam, N., Jiao, J.-Y., Wu, Z., Xu, Z. Z., Cantrell, K., Yang, Y., … Knight, R. (2019). Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea. *Nature Communications*, *10*(1), Article 1. https://doi.org/10.1038/s41467-019-13443-4

Zuckerkandl, E., & Pauling, L. (1965). Molecules as documents of evolutionary history. *Journal of Theoretical Biology*, *8*(2), 357–366. https://doi.org/10.1016/0022-5193(65)90083-4

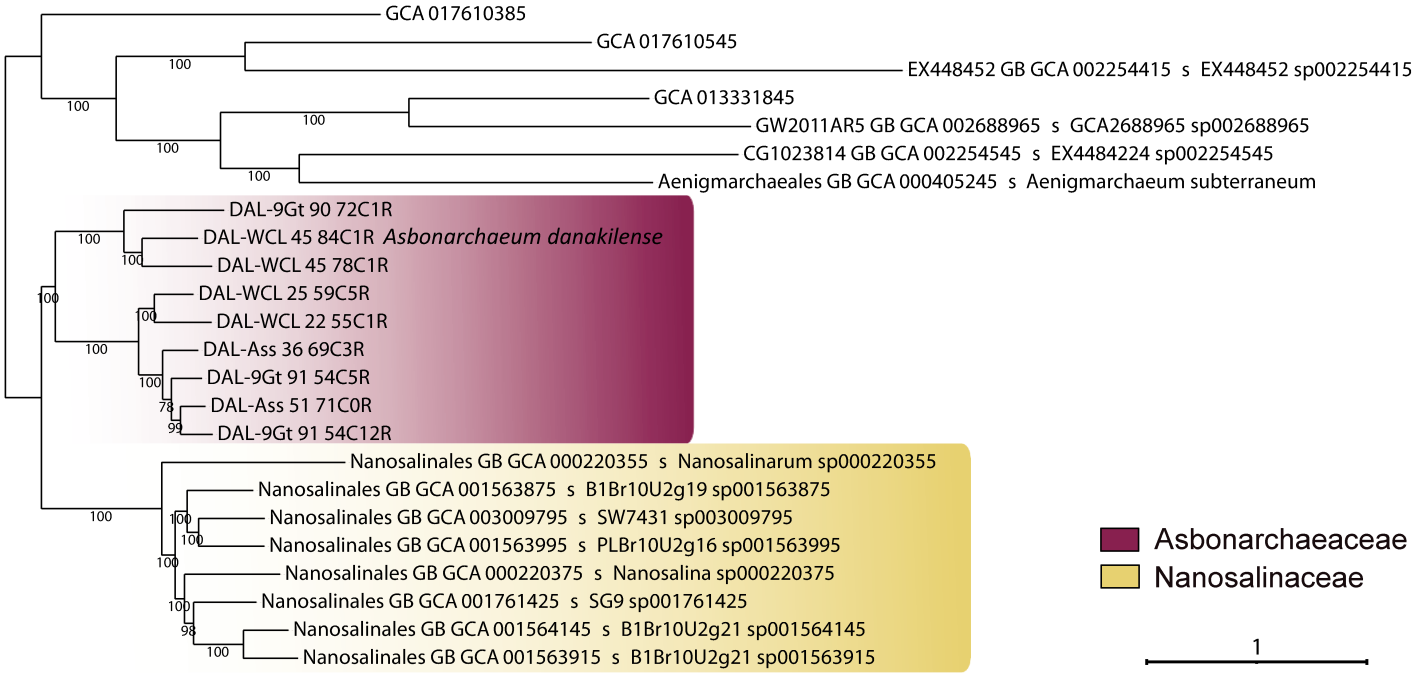
10. Supplementary Material of Manuscript 1



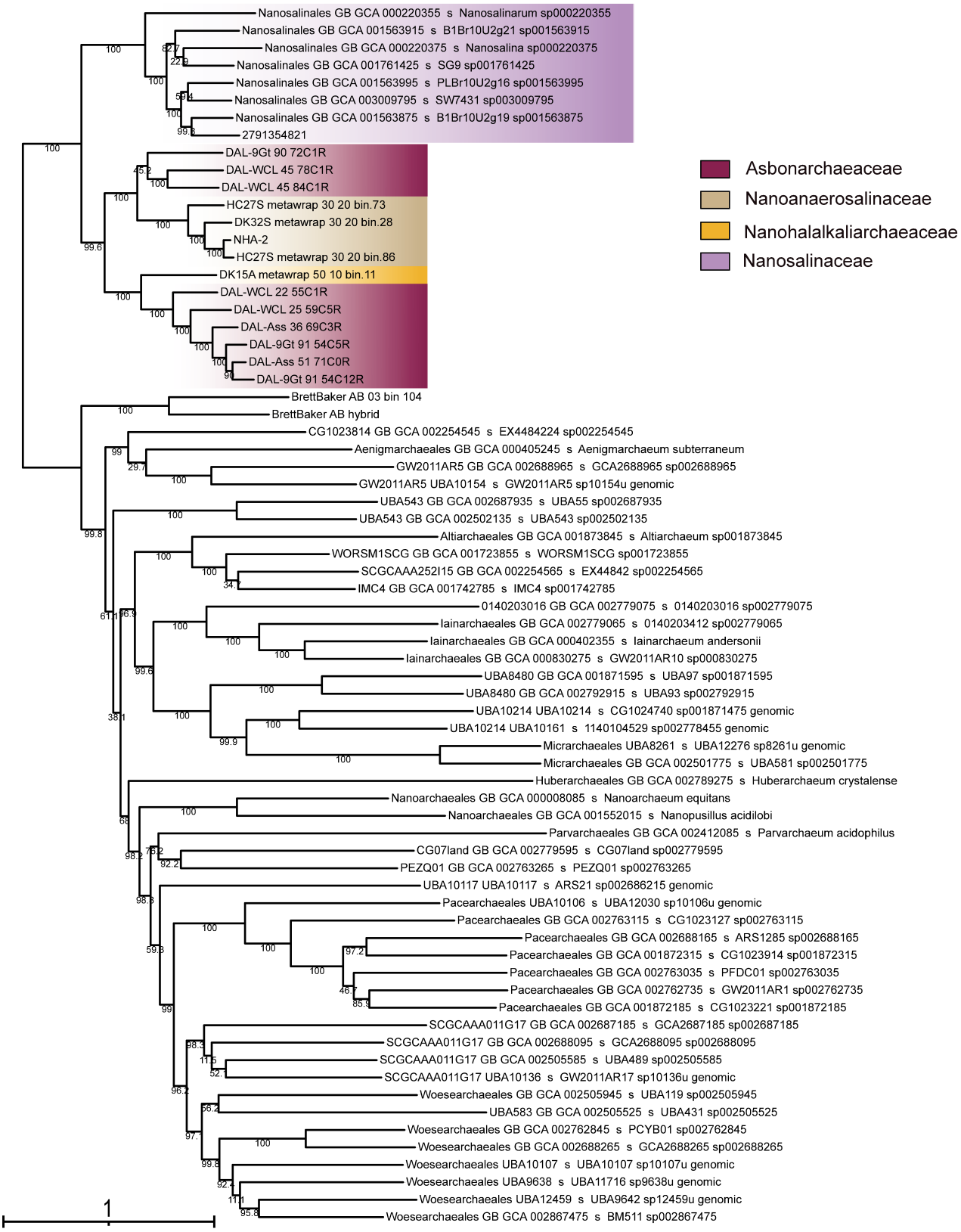
Supplementary Figure 1 | Average nucleotide identities (ANI) and average amino acid identities (AAI) among the Afararchaeaceae and Asbonarchaeaceae MAGs. (a,c) Pairwise ANI for the four Afararchaeaceae and nine Asbonarchaeaceae MAGs. (b,d) Pairwise AAI for the four Afararchaeaceae and nine Asbonarchaeaceae MAGs. ANI are only reported for values above 70% identty.



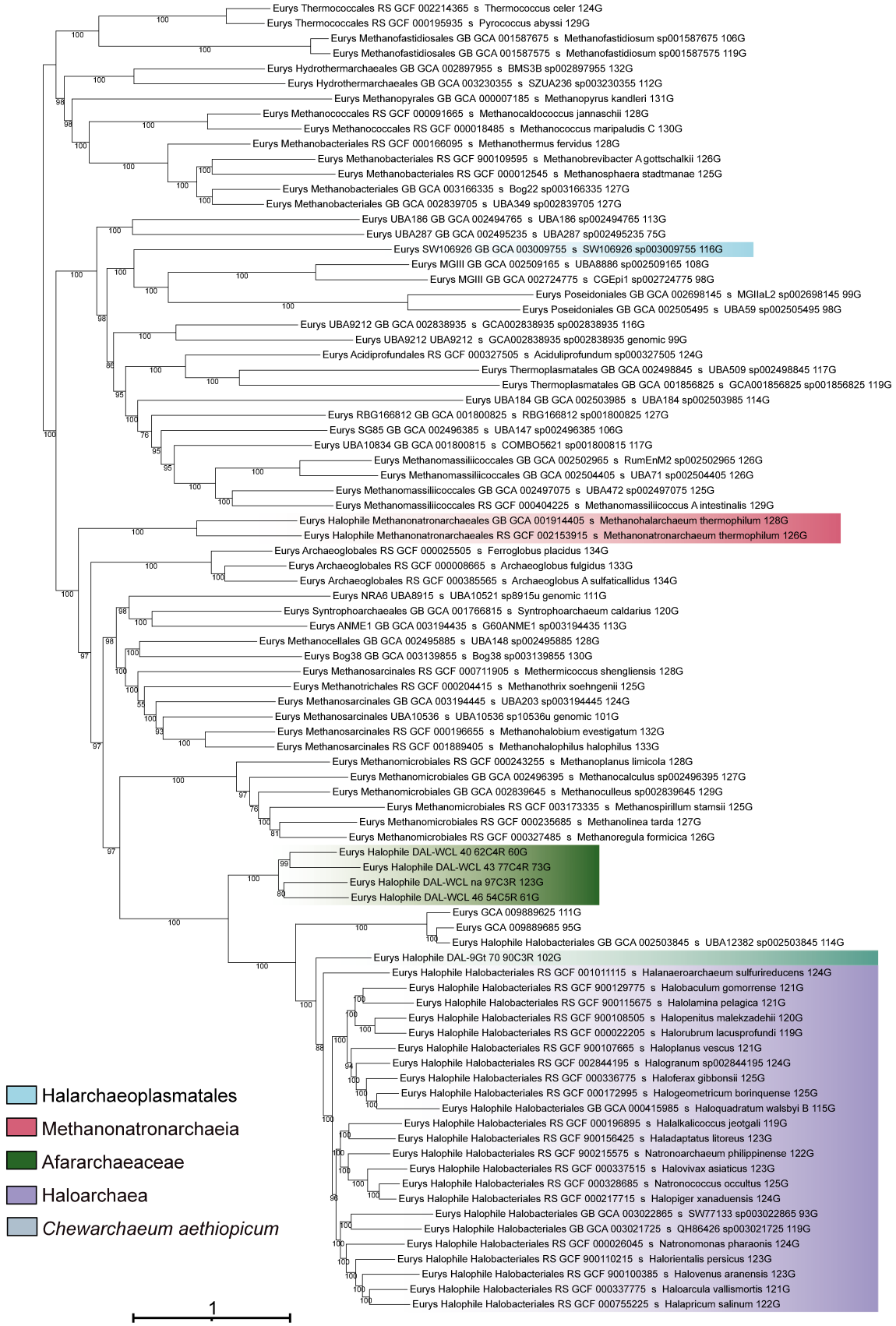
Supplementary Figure 2 | Maximum likelihood phylogeny of 35 archaeal taxa based on the concatenated alignment of 122 single-copy proteins obtained from the Genome Taxonomy Database (GTDB). The ML tree was constructed using the LG+C60+ F+Γ4 model of evolution. Branch support was assessed using 1,000 ultrafast bootstraps via IQ-TREE implementation. The four new afararchaeal MAGs are highlighted in green, and the new deep-branching Haloarchaea MAG '*Chewarchaeum aethiopicum*' is highlighted in teal. The scale bar represents the estimated number of substitutions per site. Each tip contains a GTDB identification label and the species name when available.



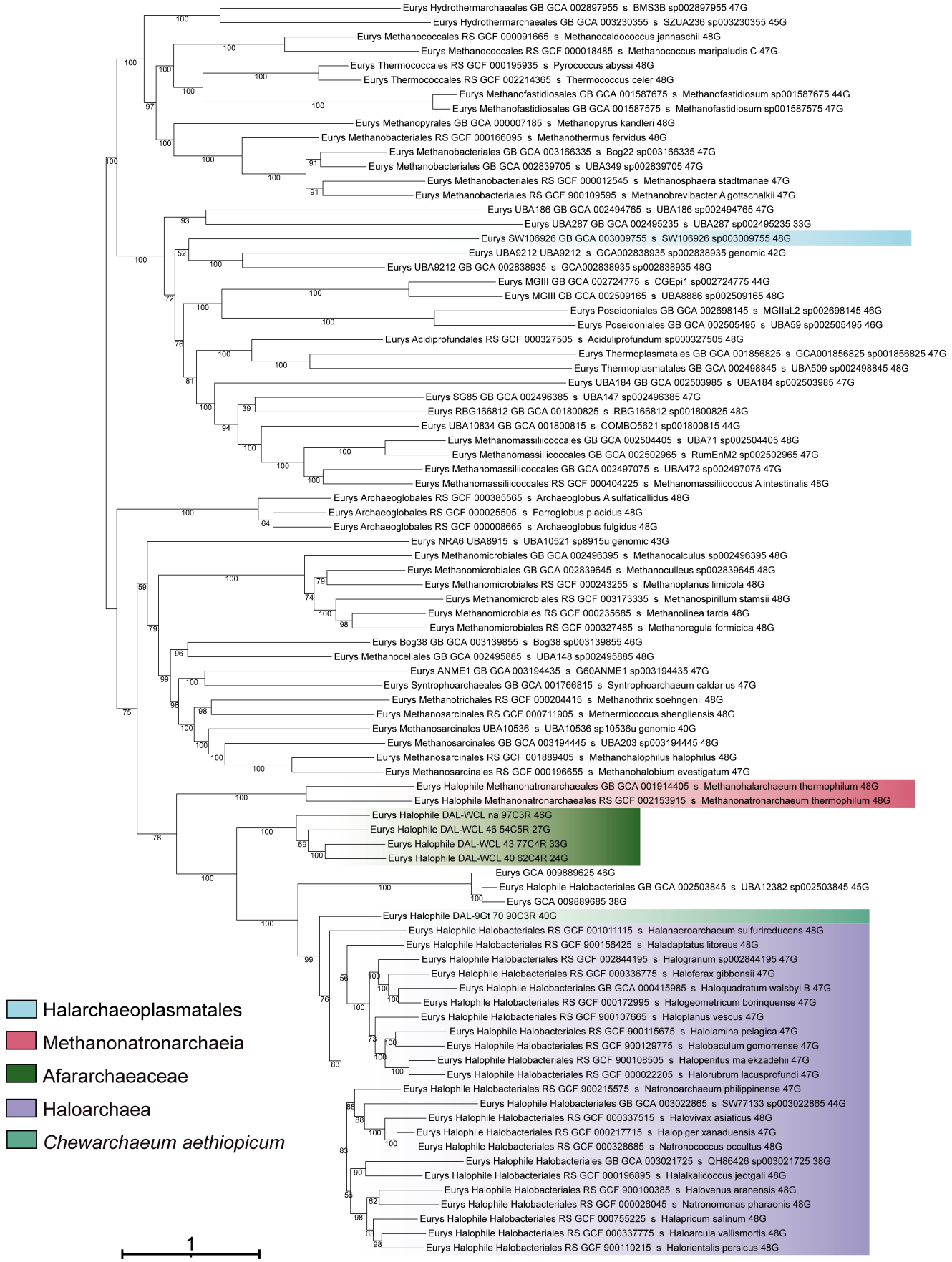
Supplementary Figure 3 | Maximum likelihood phylogeny of 24 DPANN archaea based on the concatenated alignment of 99 single-copy proteins obtained from the Genome Taxonomy Database (GTDB). The ML tree was constructed using the LG+C60+F+Γ4 model of evolution. Branch support was assessed using 1,000 ultrafast bootstraps via IQ-TREE implementation. The scale bar represents the estimated number of substitutions per site. Each tip contains a GTDB identification label, the GTDB order, and the species name when available.



Supplementary Figure 4 | Maximum likelihood phylogeny of 70 DPANN archaea based on the concatenated alignment of 24 large subunit ribosomal proteins. The ML tree was constructed using the LG+C20+Γ4 model of evolution. Branch support was assessed using 1,000 ultrafast bootstraps via IQ-TREE implementation. The scale bar represents the estimated number of substitutions per site. Each tip contains a GTDB identification label, the GTDB order, and the species name when available.



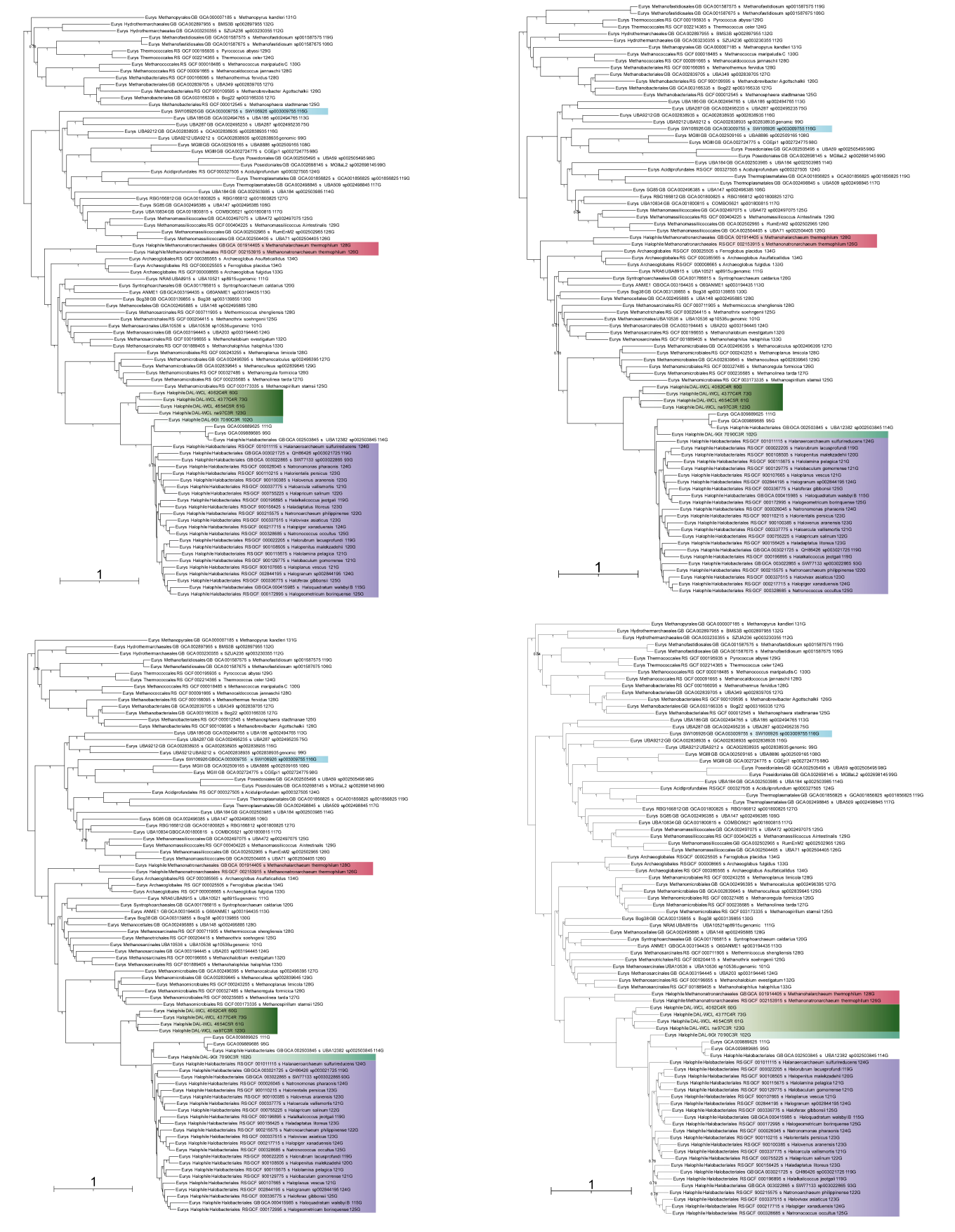
Supplementary Figure 5 | Maximum likelihood phylogeny of 87 archaeal taxa based on the concatenated alignment of 136 new marker (NM) protein dataset. The ML tree was constructed using the LG+C60+F+Γ4 model of evolution. Branch support was assessed using 1,000 ultrafast bootstraps via IQ-TREE implementation. The scale bar represents the estimated number of substitutions per site. Each tip label provides information about the archaeal supergroup, taxonomic order based on GTDB, GTDB accession number, species identification, and the number of markers identified for each taxon out of the 136 NM proteins.



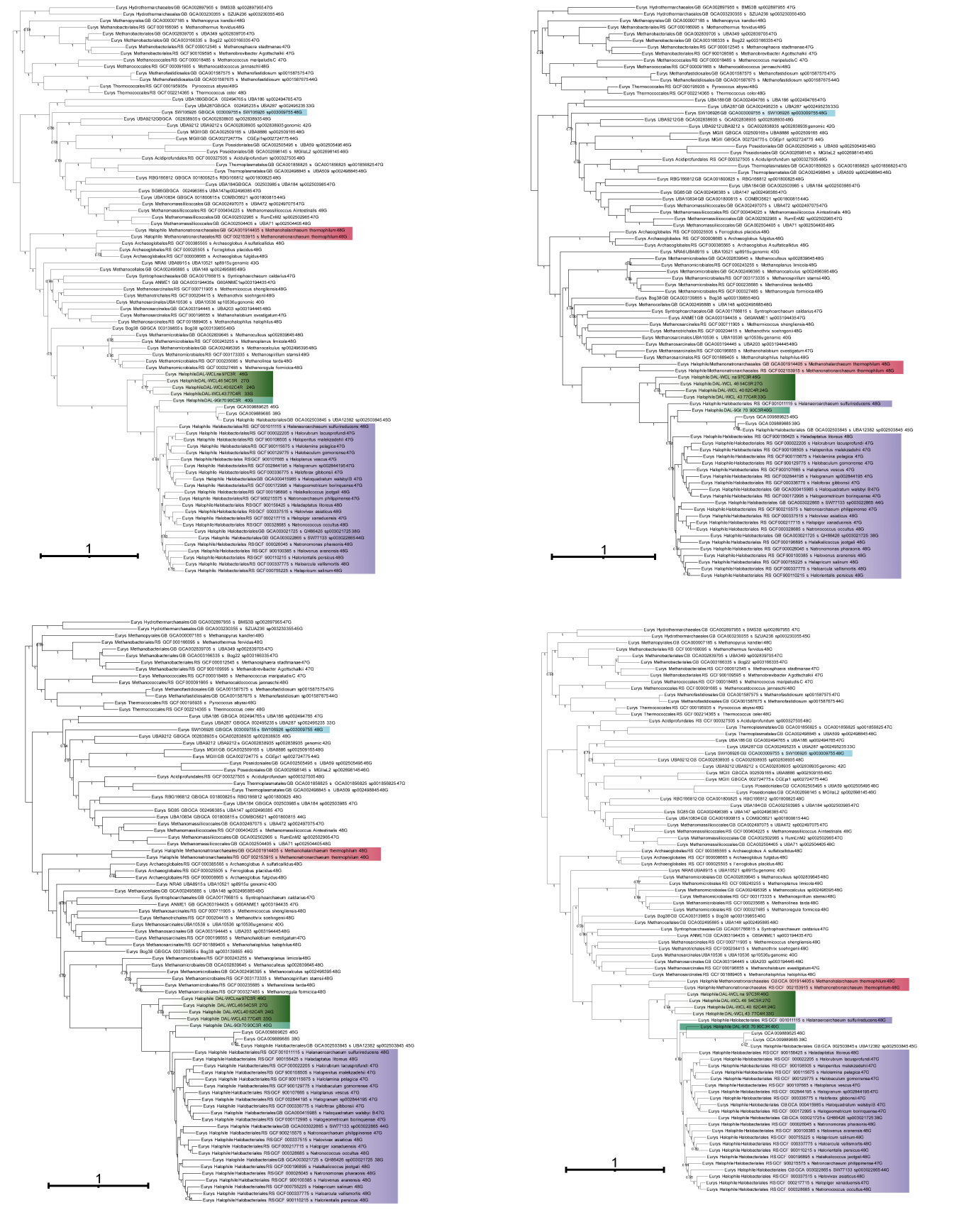
Supplementary Figure 6 | Maximum likelihood phylogeny of 87 archaeal taxa based on the concatenated alignment of 48 ribosomal protein (RP) dataset. The ML tree was constructed using the LG+C60+F+Γ4 model of evolution. Branch support was assessed using 1000 ultrafast bootstraps via IQ-TREE implementation. The scale bar represents the estimated number of substitutions per site. Each tip label provides information about the archaeal super-group, taxonomic order based on GTDB, GTDB accession number, species identification, and the number of markers identified for each taxon out of the 48 RP proteins.

Supplementary Figure 7 | Maximum likelihood phylogeny of 104 archaeal taxa based on the concatenated alignment of 136 new marker (NM) protein dataset. The ML tree was constructed using the LG+C60+F+Γ4 model of evolution. Branch support was assessed using 1000 ultrafast bootstraps via IQ-TREE implementation. The scale bar represents the estimated number of substitutions per site. Each tip label provides information about the archaeal supergroup, taxonomic order based on GTDB, GTDB accession number, species identification, and the number of markers identified for each taxon out of the 136 NM proteins.

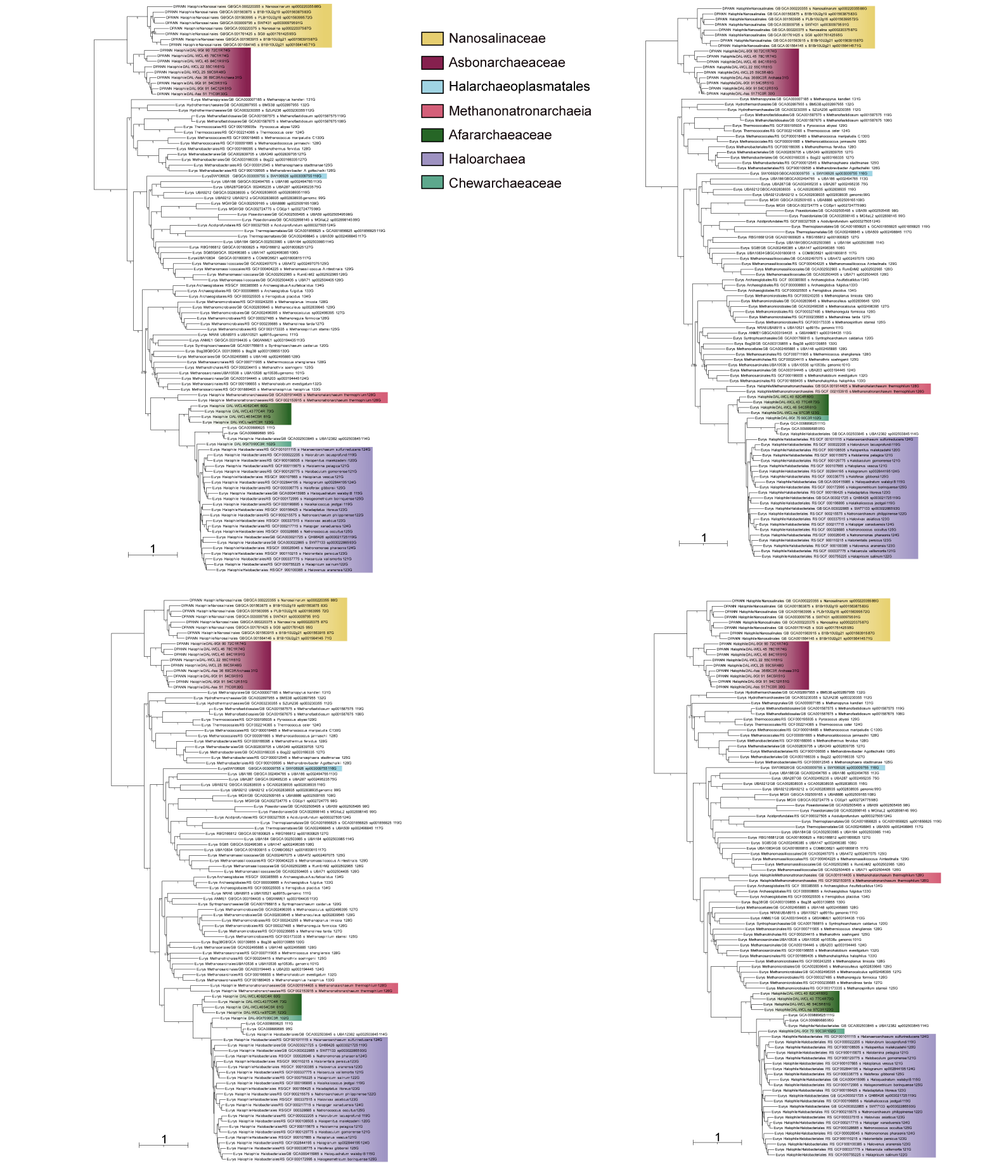
Supplementary Figure 8 | Maximum likelihood phylogeny of 104 archaeal taxa based on the concatenated alignment of 48 ribosomal protein (RP) dataset. The ML tree was constructed using the LG+C60+F+Γ4 model of evolution. Branch support was assessed using 1000 ultrafast bootstraps via IQ- TREE implementation. The scale bar represents the estimated number of substitutions per site. Each tip label provides information about the archaeal super-group, taxonomic order based on GTDB, GTDB accession number, species identification, and the number of markers identified for each taxon out of the 48 RP proteins.



Supplementary Figure 9 | Consensus tree for each of the four MCMC chains for the 87-NM dataset. Four independent Markov chain Monte Carlo (MCMC) chains were inferred using PhyloBayes (CAT+GTR, 15,000 generations with a burn-in of 3,000). Support at branches corresponds to posterior probabilities estimated post-burnin. The scale bar represents the estimated number of substitutions per site. Different halophilic clades are visually represented with distinct colors: halarchaeoaplasmatales (cyan), methanonatronarchaea (rose), afararchaea (green), chewarchaea (teal), and haloarchaea (violet). Each tip label provides information about the archaeal supergroup, taxonomic order based on GTDB, GTDB accession number, species identification, and the number of markers identified for each taxon out of the 136 NM proteins.

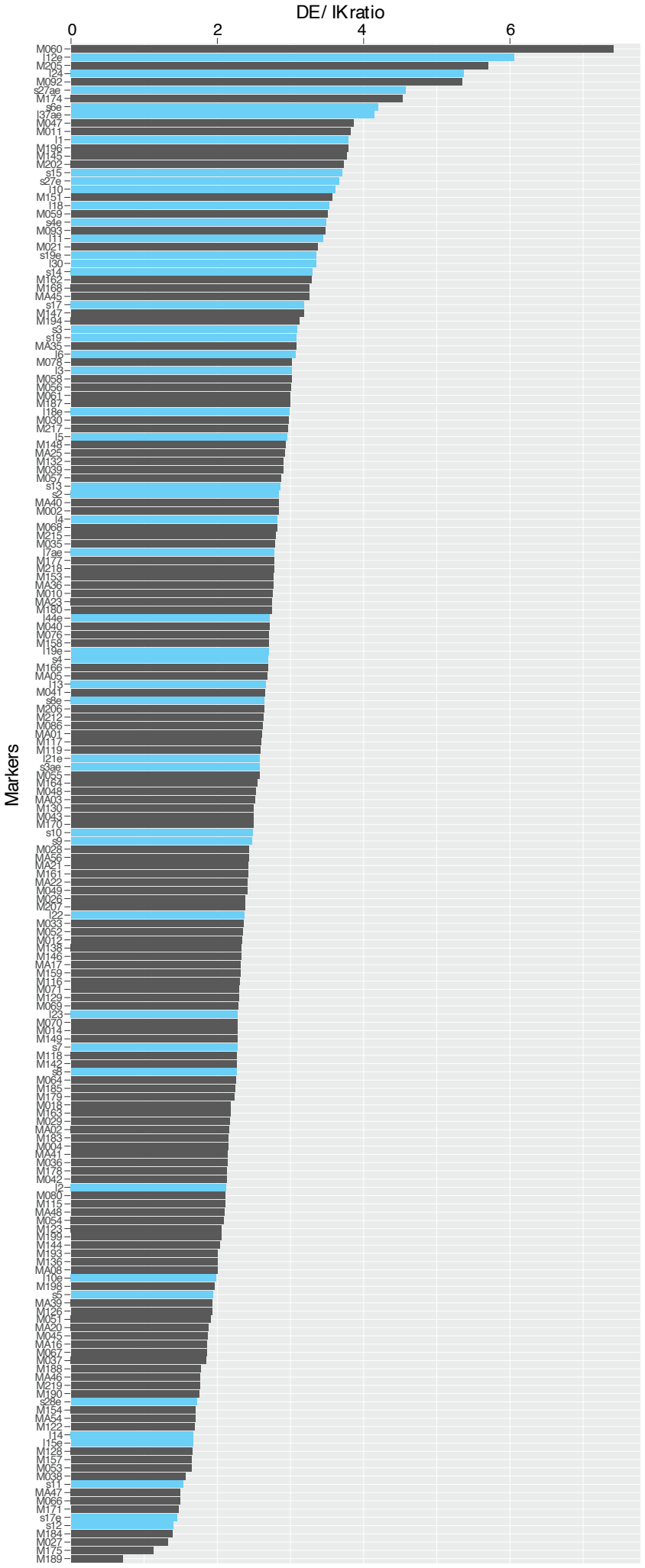


Supplementary Figure 10 | Consensus tree for each of the four MCMC chains for the 87-RP dataset. Four independent Markov chain Monte Carlo (MCMC) chains were inferred using PhyloBayes (CAT+GTR, 15,000 generations with a burn-in of 3,000). Support at branches corresponds to posterior probabilities estimated post-burnin. The scale bar represents the estimated number of substitutions per site. Different halophilic clades are visually represented with distinct colors: halarchaeoaplasmatales (cyan), methanonatronarchaea (rose), afararchaea (green), chewarchaea (teal), and haloarchaea (violet). Each tip label provides information about the archaeal super group, taxonomic order based on GTDB, GTDB accession number, species identification, and the number of markers identified for each taxon out of the 48 RP proteins.

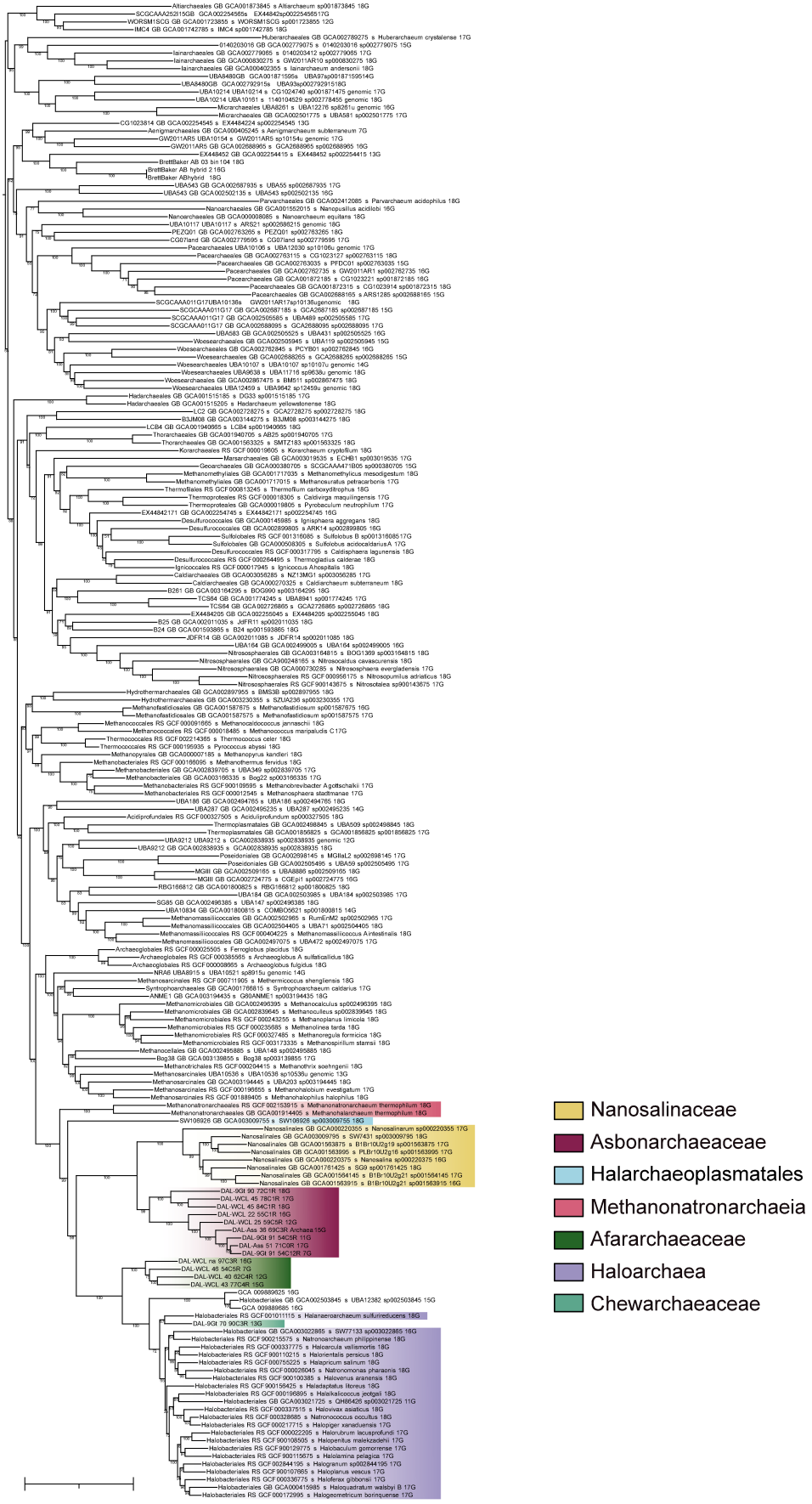


Supplementary Figure 11 | Consensus tree for each of the four MCMC chains for the 104-NM dataset. Four independent Markov chain Monte Carlo (MCMC) chains were inferred using PhyloBayes (CAT+GTR, 15,000 generations with a burn-in of 3,000). Support at branches corresponds to posterior probabilities estimated post-burnin. The scale bar represents the estimated number of substitutions per site. Different halophilic clades are visually represented with distinct colors. Each tip label provides information about the archaeal supergroup, taxonomic order based on GTDB, GTDB accession number, species identification, and the number of markers identified for each taxon out of the 136 NM proteins.

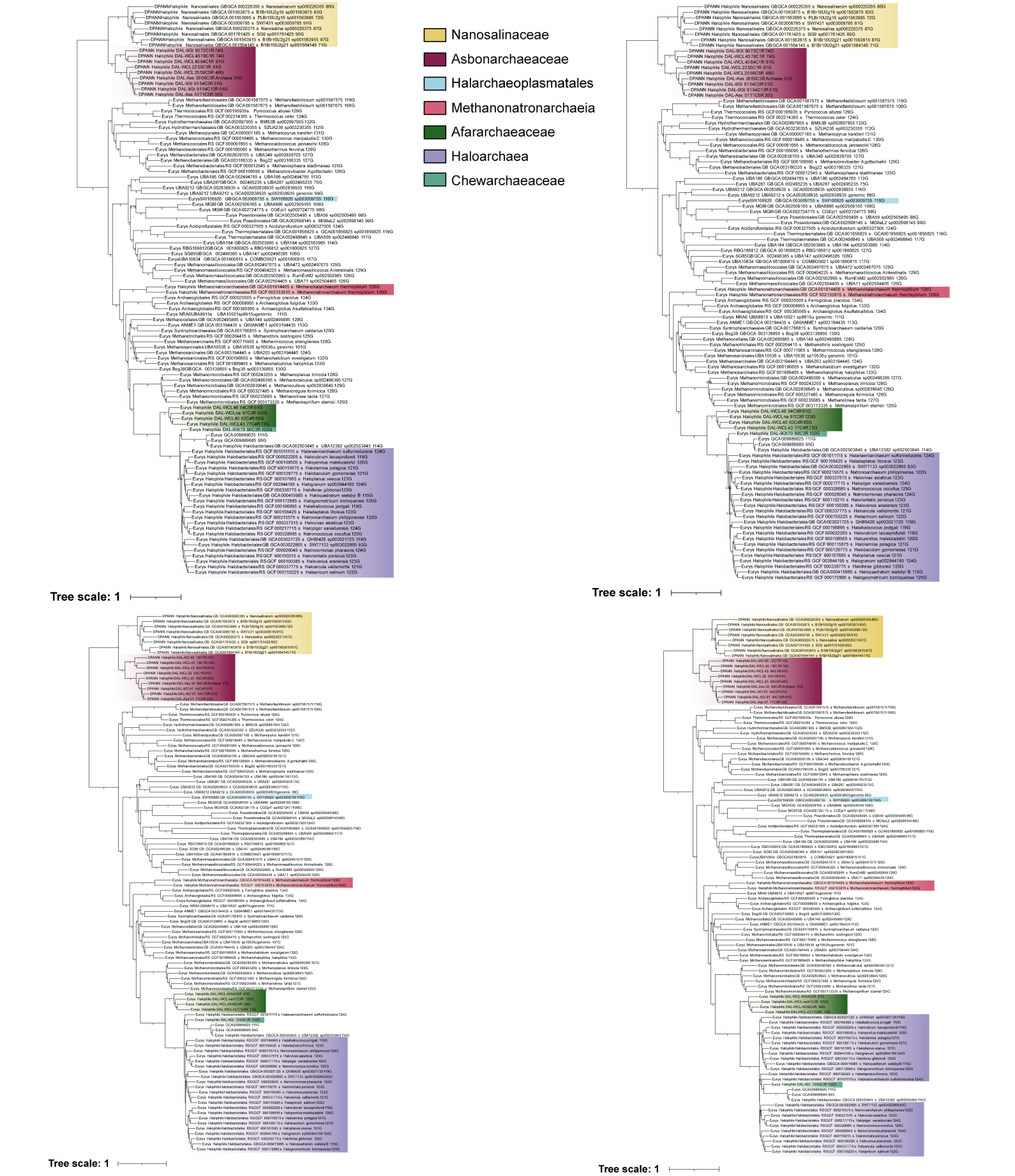
Supplementary Figure 12 | Consensus tree for each of the four MCMC chains for the 104-RP dataset. Four independent Markov chain Monte Carlo (MCMC) chains were inferred using PhyloBayes (CAT+GTR, 15,000 generations with a burn-in of 3,000). Support at branches corresponds to posterior probabilities estimated post-burnin. The scale bar represents the estimated number of substitutions per site. Different halophilic clades are visually represented with distinct colors. Each tip label provides information about the archaeal super-group, taxonomic order based on GTDB, GTDB accession number, species identification, and the number of markers identified for each taxon out of the 48 RP proteins.



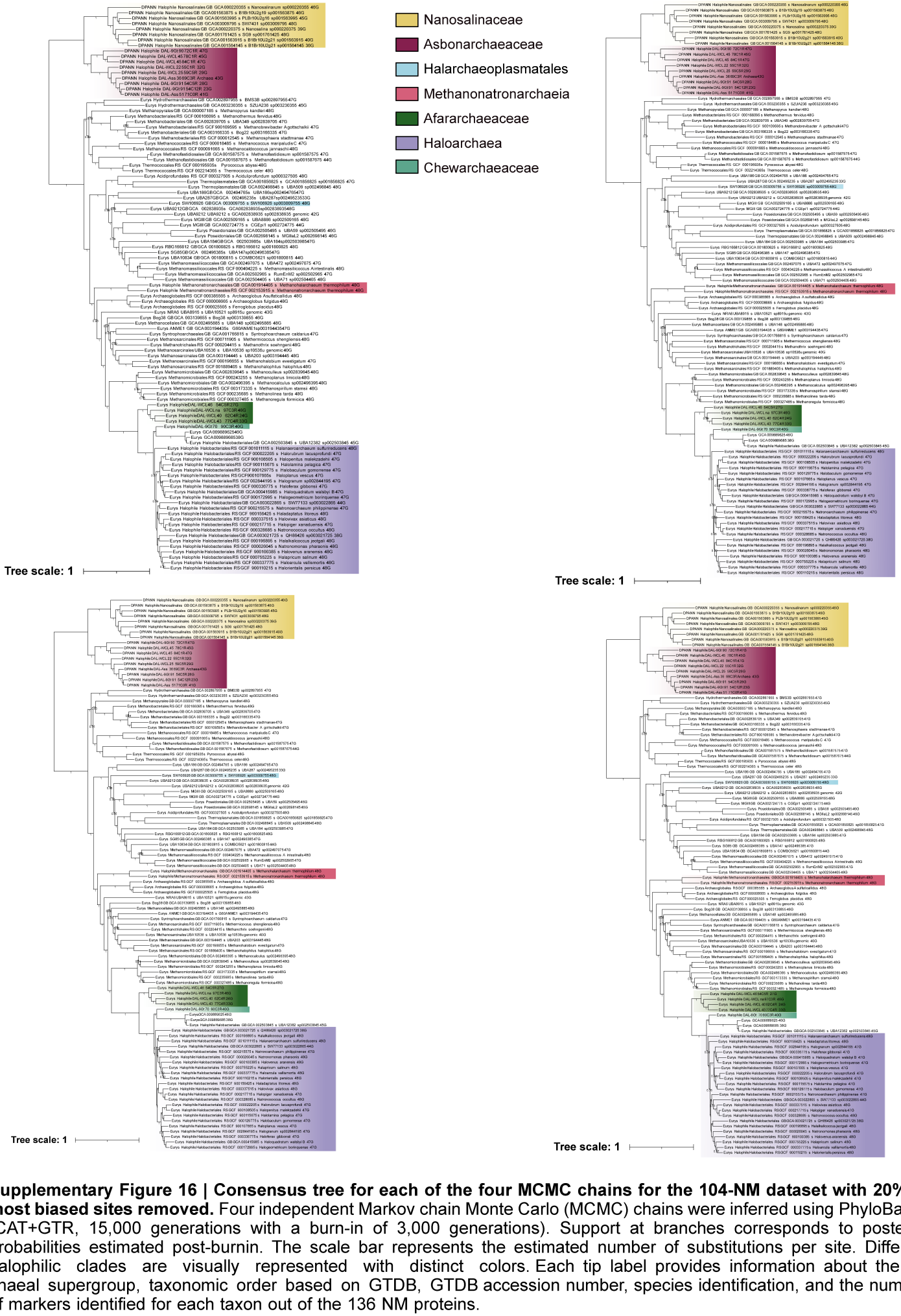
Supplementary Figure 13 | RP (blue) and NM (grey) markers ranked from the most to least biased based on the calculated DE/IK ratio of halophiles versus the DE/IK ratio of non-halophiles.



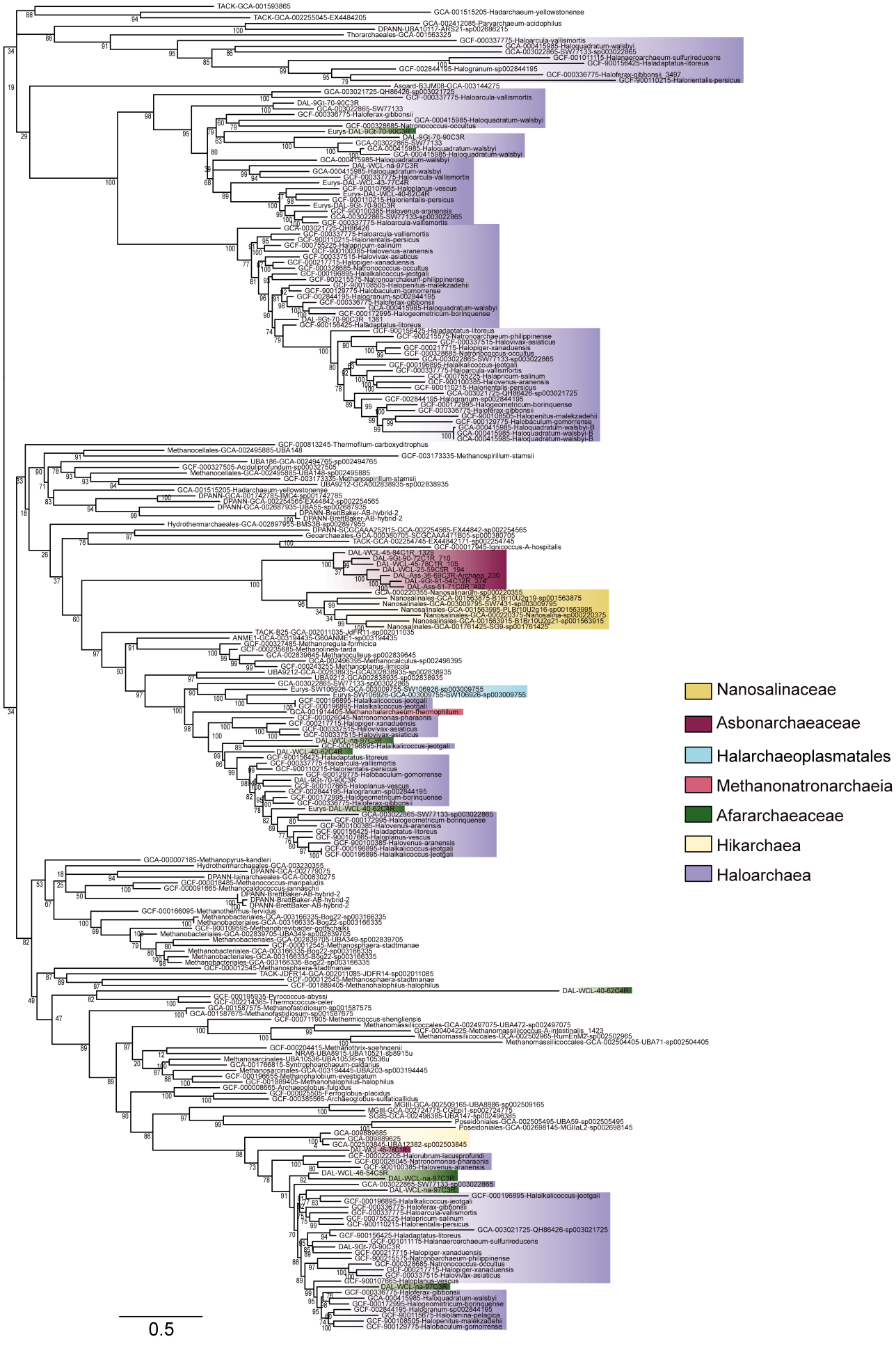
Supplementary Figure 14 | Maximum likelihood phylogeny based on concatenating the 18 most biased ribosomal proteins. The ML tree was constructed using the LG+C60+F+Γ4 model of evolution. Branch support was assessed using 1,000 ultrafast bootstraps via IQ-TREE implementation. Different halophilic clades are visually represented with distinct colors. The scale bar represents the estimated number of substitutions per site. Each tip label provides information about the archaeal super-group, taxonomic order based on GTDB, GTDB accession number, species identification, and the number of markers identified for each taxon out of the 18 RP proteins.



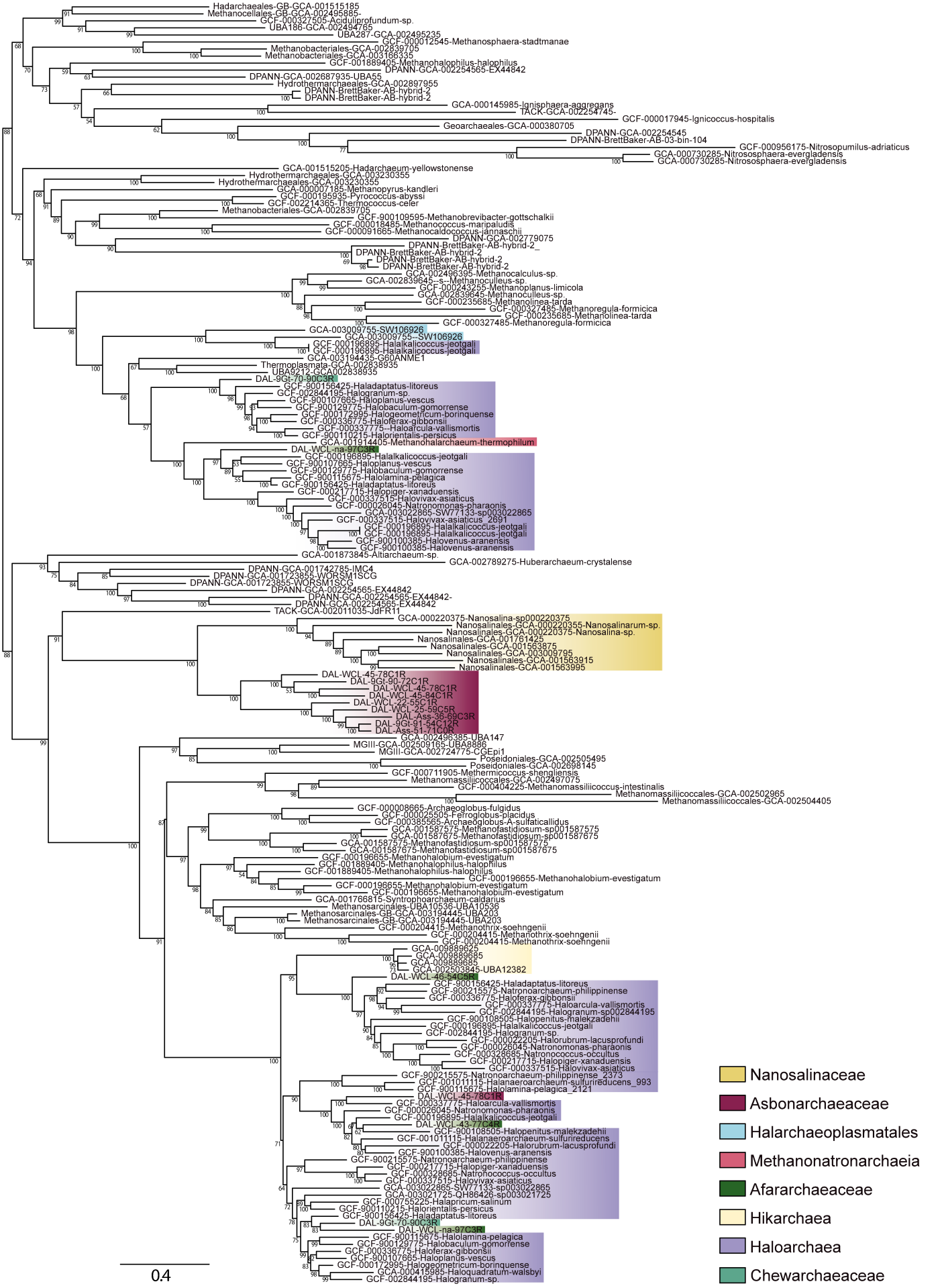
Supplementary Figure 15 | Consensus tree for each of the four MCMC chains for the 104-RP dataset with 20% of most biased sites removed. Four independent Markov chain Monte Carlo (MCMC) chains were inferred using PhyloBayes (CAT+GTR, 15,000 generations with a burn-in of 3,000). Support at branches corresponds to posterior probabilities estimated post-burnin. The scale bar represents the estimated number of substitutions per site. Different halophilic clades are visually represented with distinct colors. Each tip label provides information about the archaeal super group, taxonomic order based on GTDB, GTDB accession number, species identification, and the number of markers identified for each taxon out of the 48 RP proteins.



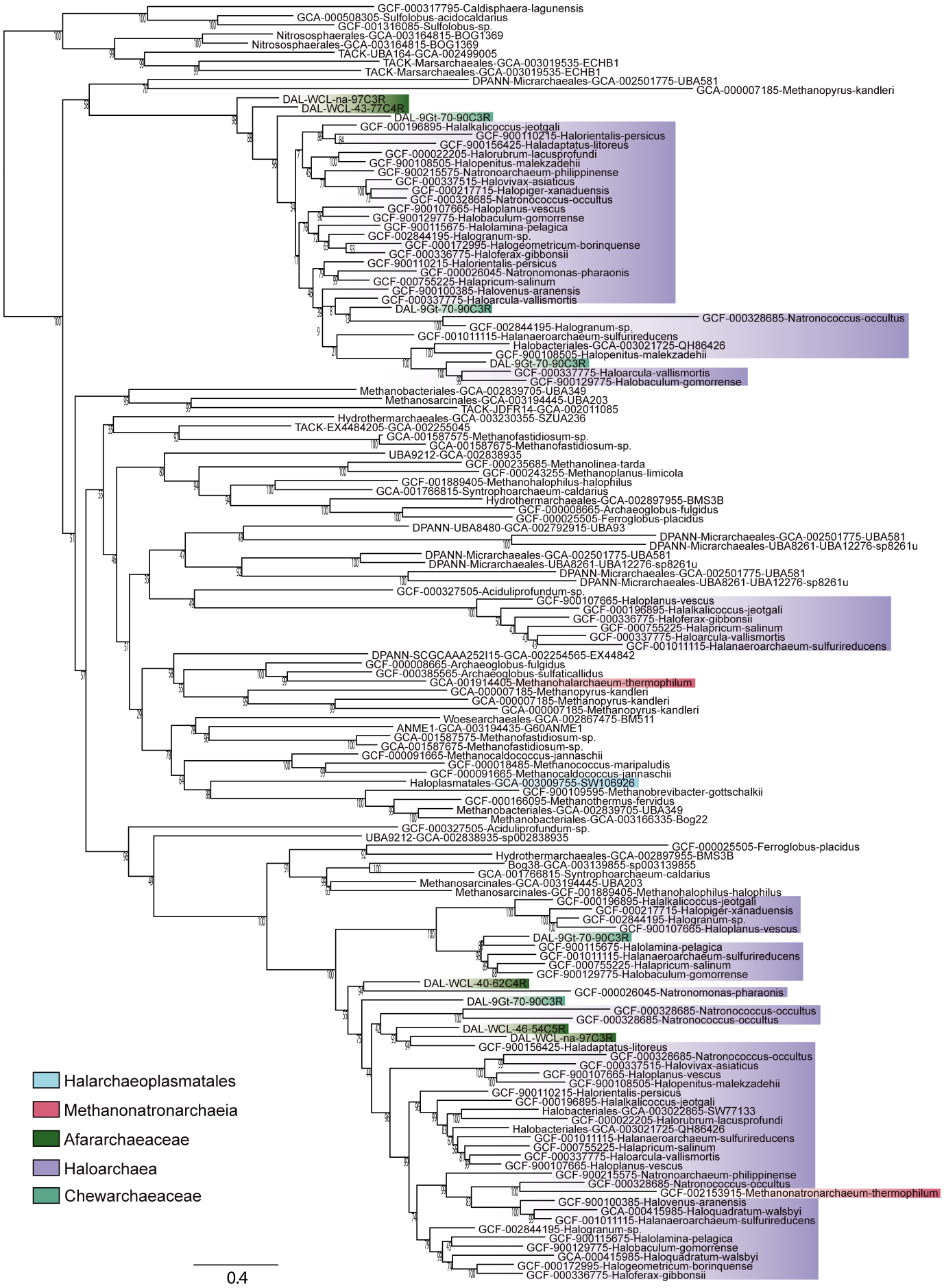
Supplementary Figure 16 | Consensus tree for each of the four MCMC chains for the 104-NM dataset with 20% of most biased sites removed. Four independent Markov chain Monte Carlo (MCMC) chains were inferred using PhyloBayes (CAT+GTR, 15,000 generations with a burn-in of 3,000 generations). Support at branches corresponds to posterior probabilities estimated post-burnin. The scale bar represents the estimated number of substitutions per site. Different halophilic clades are visually represented with distinct colors. Each tip label provides information about the archaeal supergroup, taxonomic order based on GTDB, GTDB accession number, species identification, and the number of markers identified for each taxon out of the 136 NM proteins.



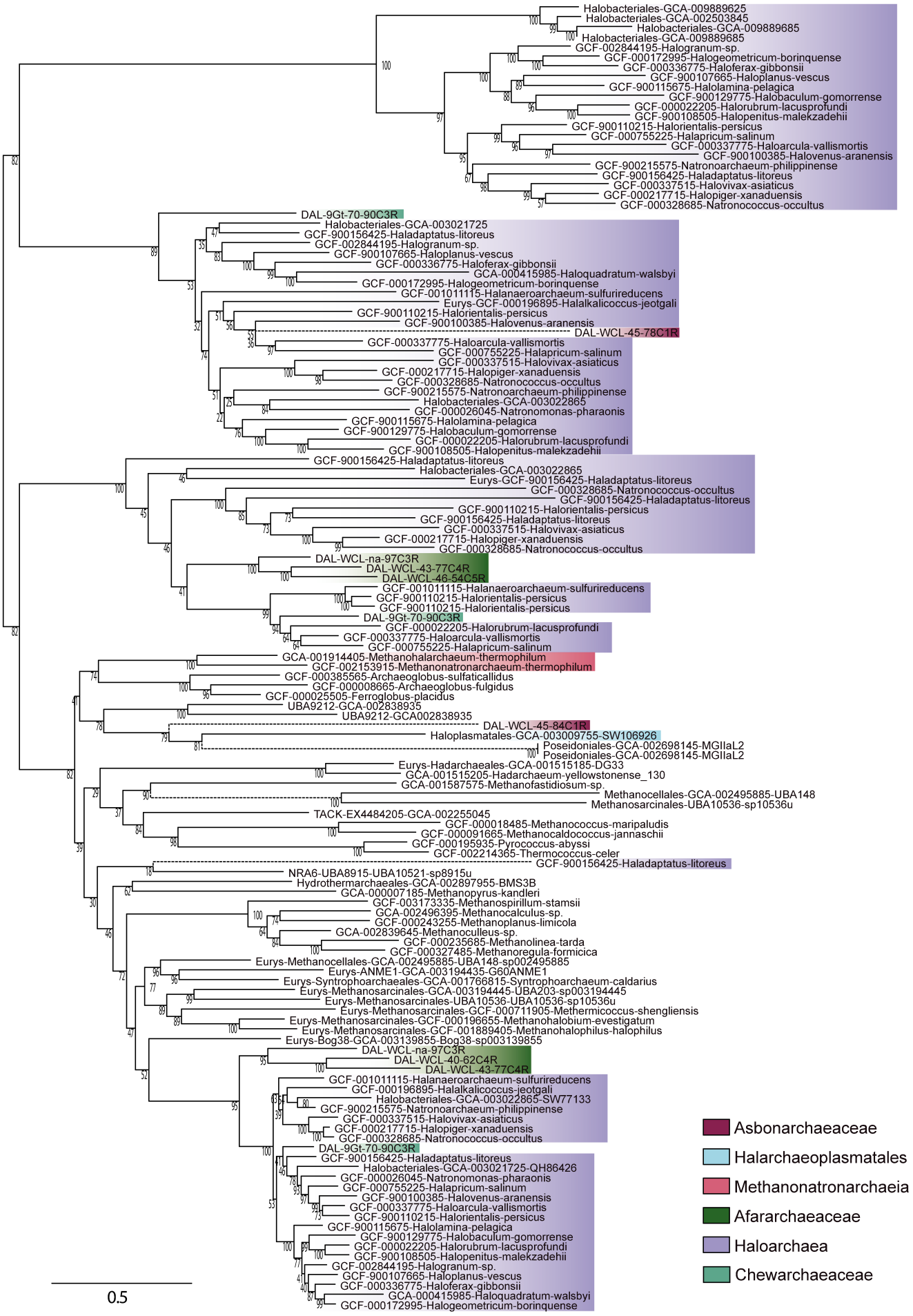
Supplementary Fig. 17 | Maximum likelihood phylogenetic tree of TrkA K+ transporter. The tree was constructed with the LG+C60+F+Γ4 model of sequence evolution. Branch support was assessed using 1,000 ultrafast bootstraps. The scale bar represents the number of substitutions per position.



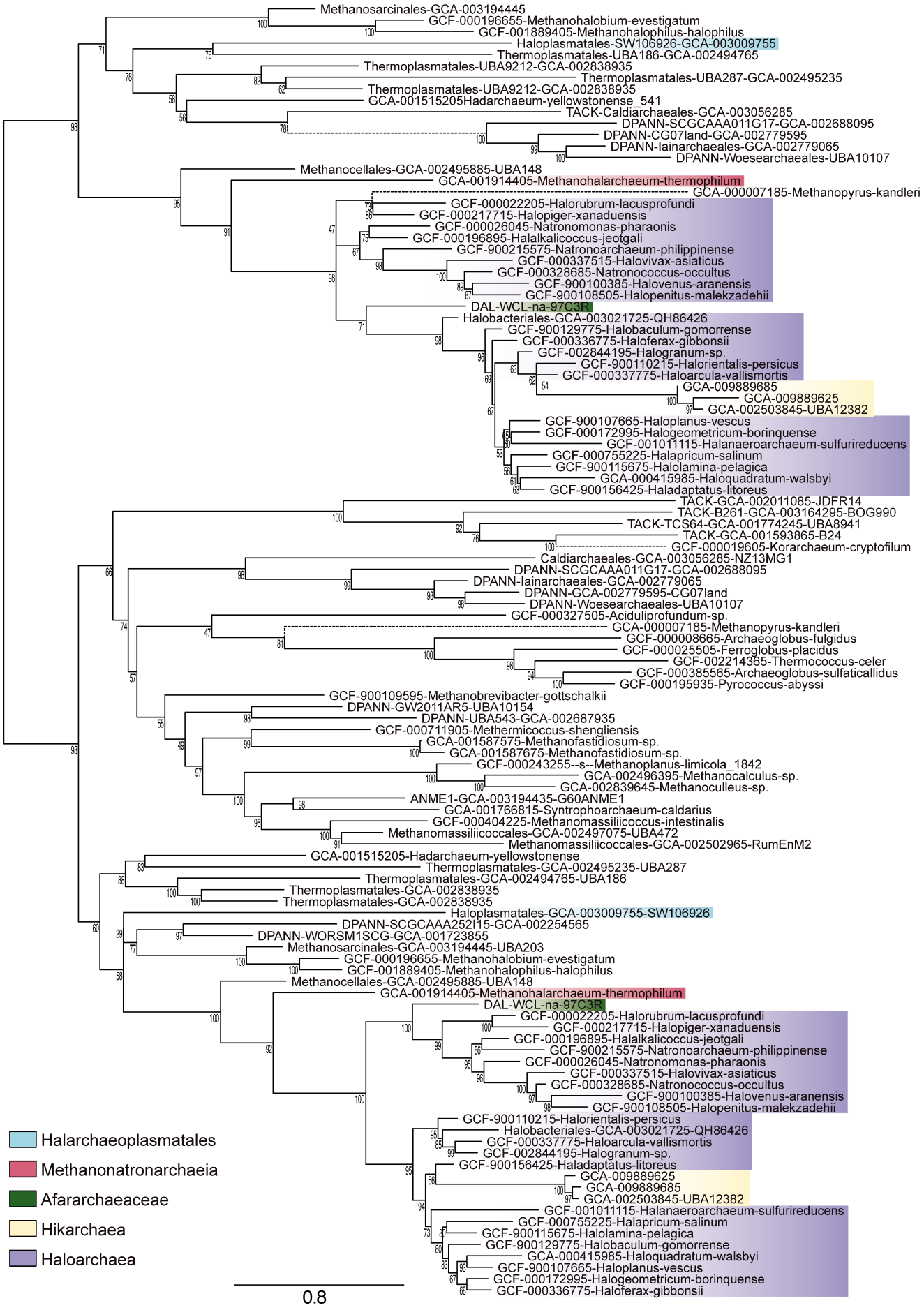
Supplementary Fig. 18 | Maximum likelihood phylogenetic tree of TrkH K+ transporter. The tree was constructed with the LG+C60+F+Γ4 model of sequence evolution. Branch support was assessed using 1,000 ultrafast bootstraps. The scale bar represents the number of substitutions per position.



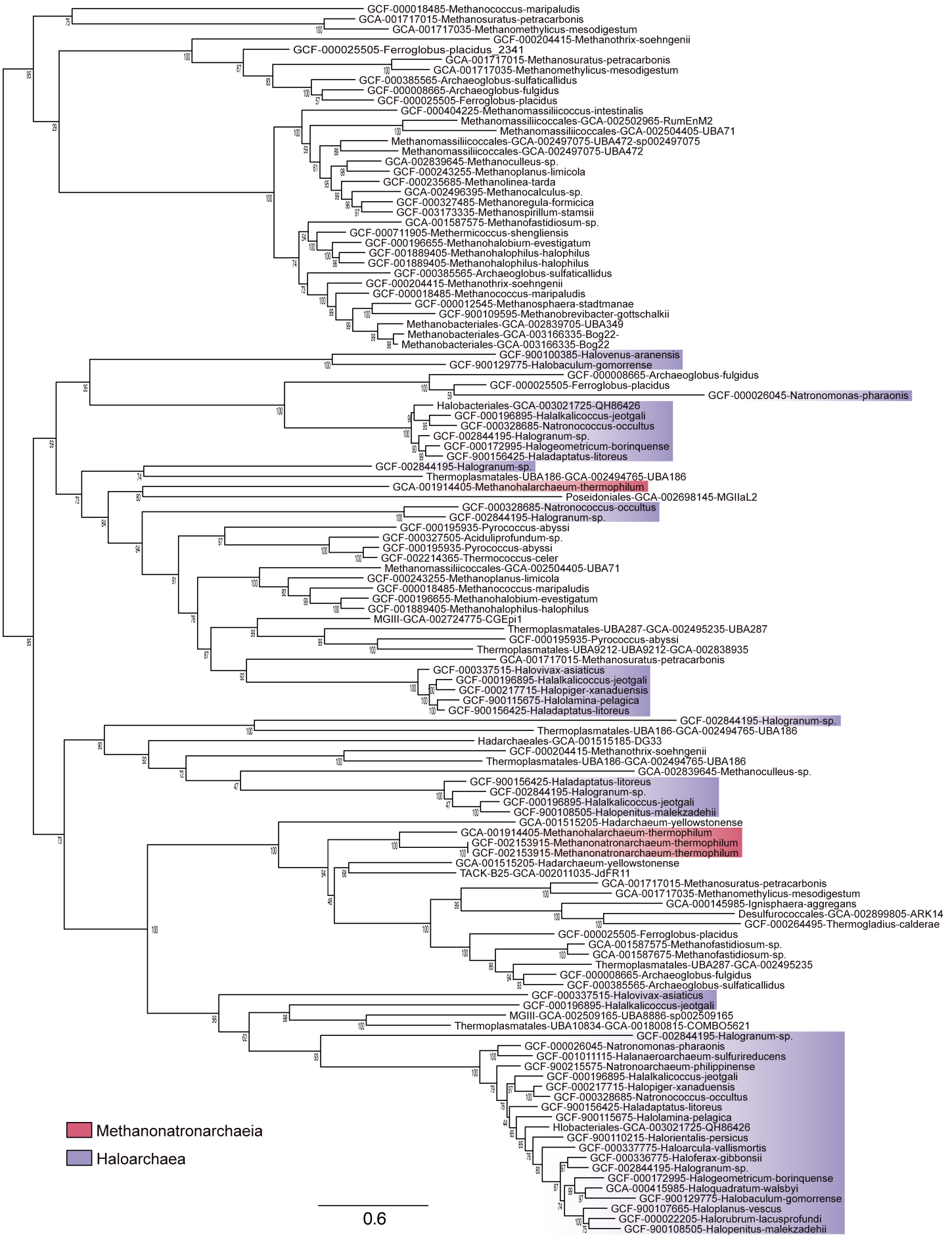
Supplementary Fig. 19 | Maximum likelihood phylogenetic tree of TrkA-N K+ transporter. The tree was constructed with the LG+C60+F+Γ4 model of sequence evolution. Branch support was assessed using 1,000 ultrafast bootstraps. The scale bar represents the number of substitutions per position.



Supplementary Fig. 20 | Maximum likelihood phylogenetic tree of Kef K+ transporters. The tree was constructed with the LG+C60+F+Γ4 model of sequence evolution. Branch support was assessed using 1,000 ultrafast bootstraps. The scale bar represents the number of substitutions per position.

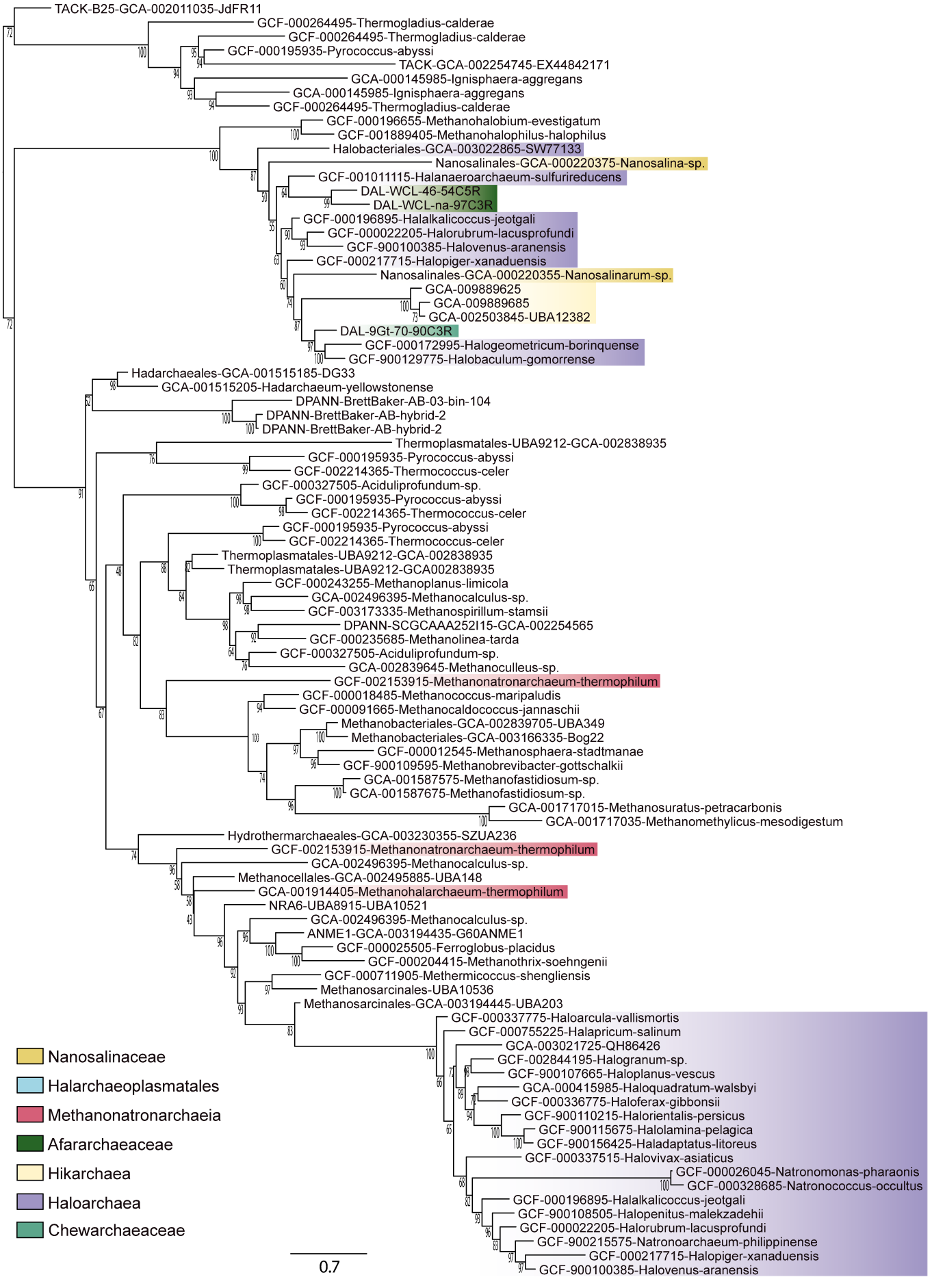


Supplementary Fig. 21 | Maximum likelihood phylogenetic tree of Mg2+ transporters. The tree was constructed with the LG+C60+F+Γ4 model of sequence evolution. Branch support was assessed using 1,000 ultrafast bootstraps. The scale bar represents the number of substitutions per position.



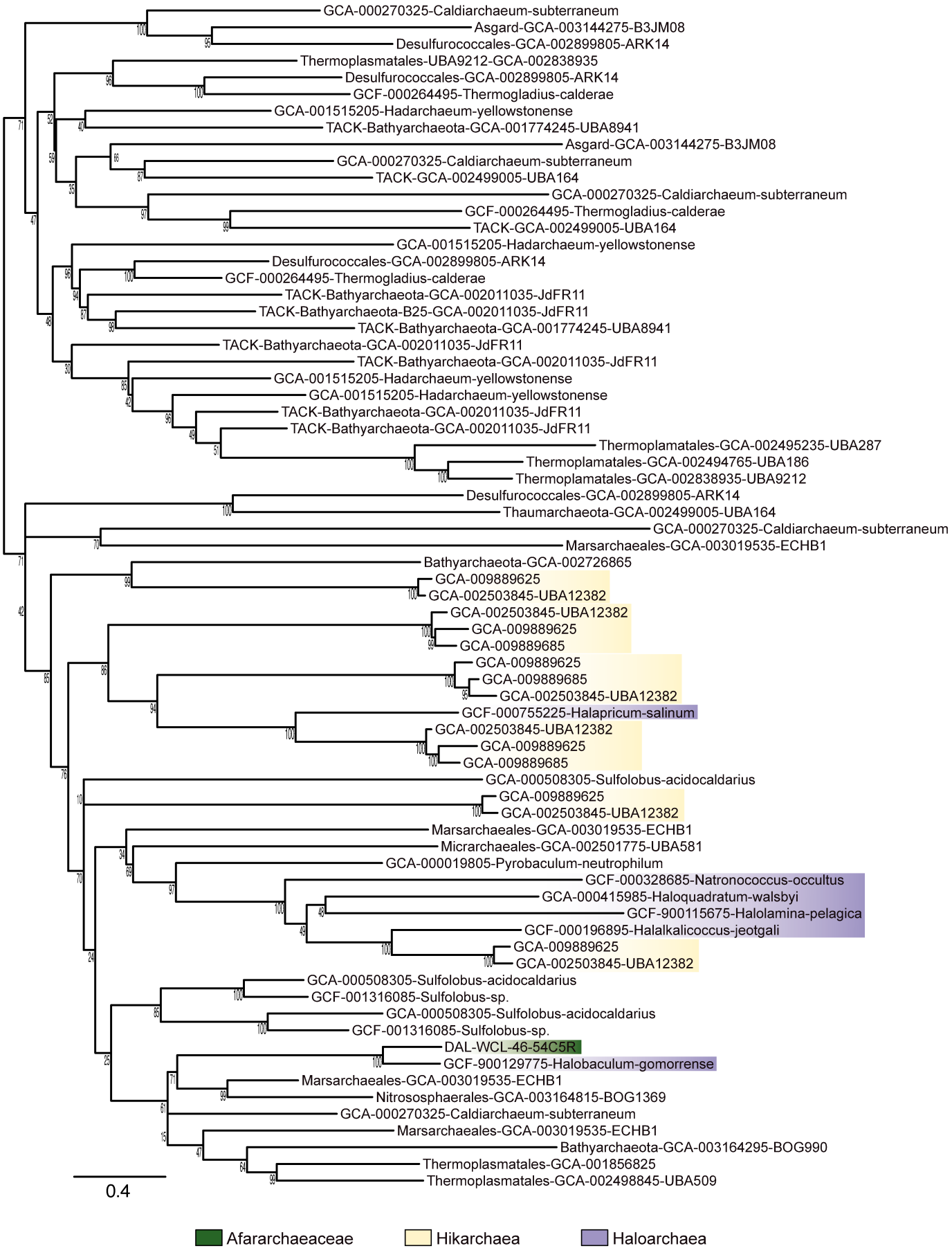
Supplementary Fig. 22 | Maximum likelihood phylogenetic tree of SSF Na+/solute symporters. The tree was constructed with the LG+C60+F+Γ4 model of sequence evolution. Branch support was assessed using 1,000 ultrafast bootstraps. The scale bar represents the number of substitutions per position.

Supplementary Fig. 23 | Maximum likelihood phylogenetic tree of Ca+/Na+ antiporters. The tree was constructed with the LG+C60+F+Γ4 model of sequence evolution. Branch support was assessed using 1,000 ultrafast bootstraps. The scale bar represents the number of substitutions per position.

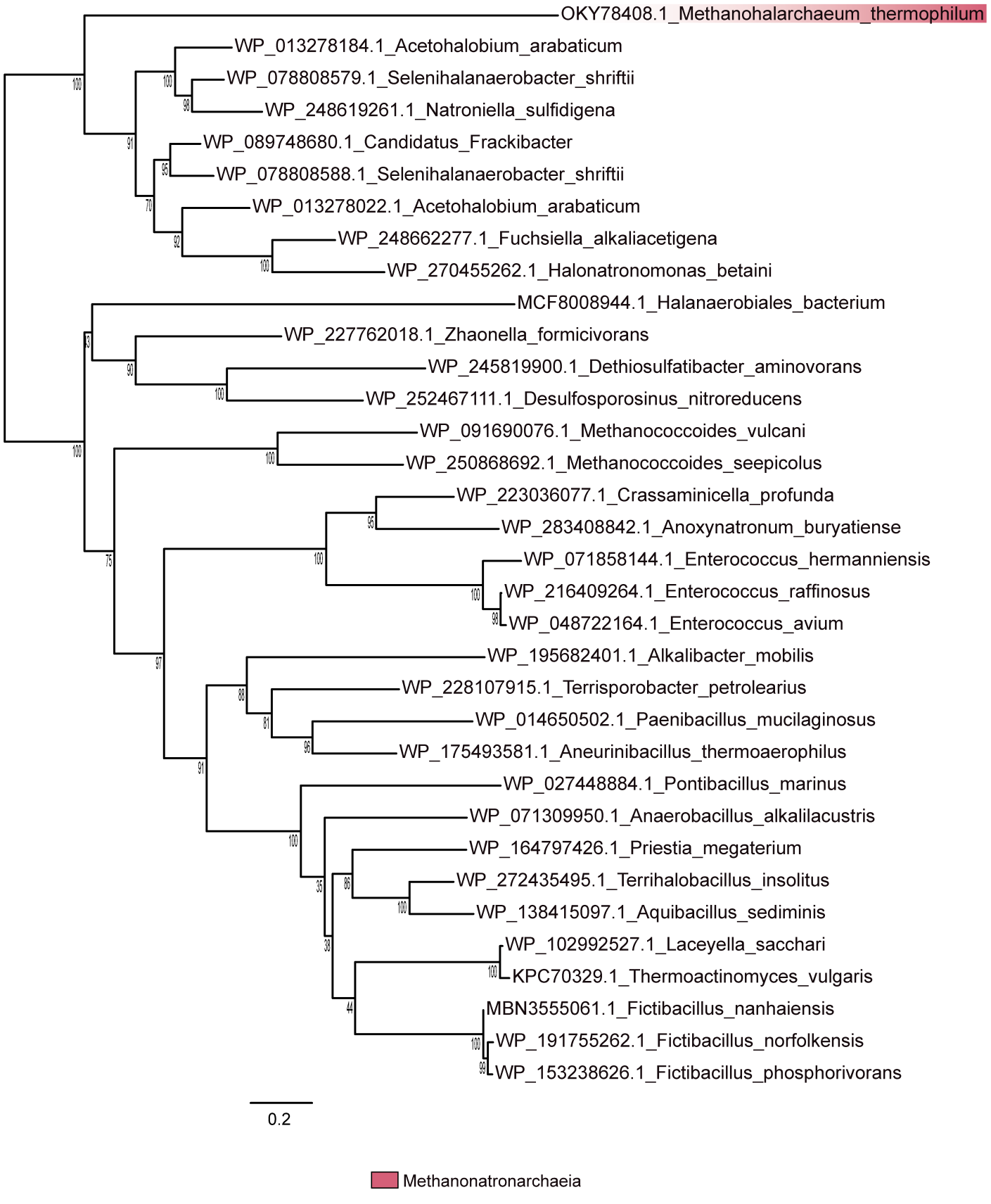


Supplementary Fig. 24 | Maximum likelihood phylogenetic tree of Na+/H+ antiporters. The tree was constructed with the LG+C60+F+Γ4 model of sequence evolution. Branch support was assessed using 1,000 ultrafast bootstraps. The scale bar represents the number of substitutions per position.

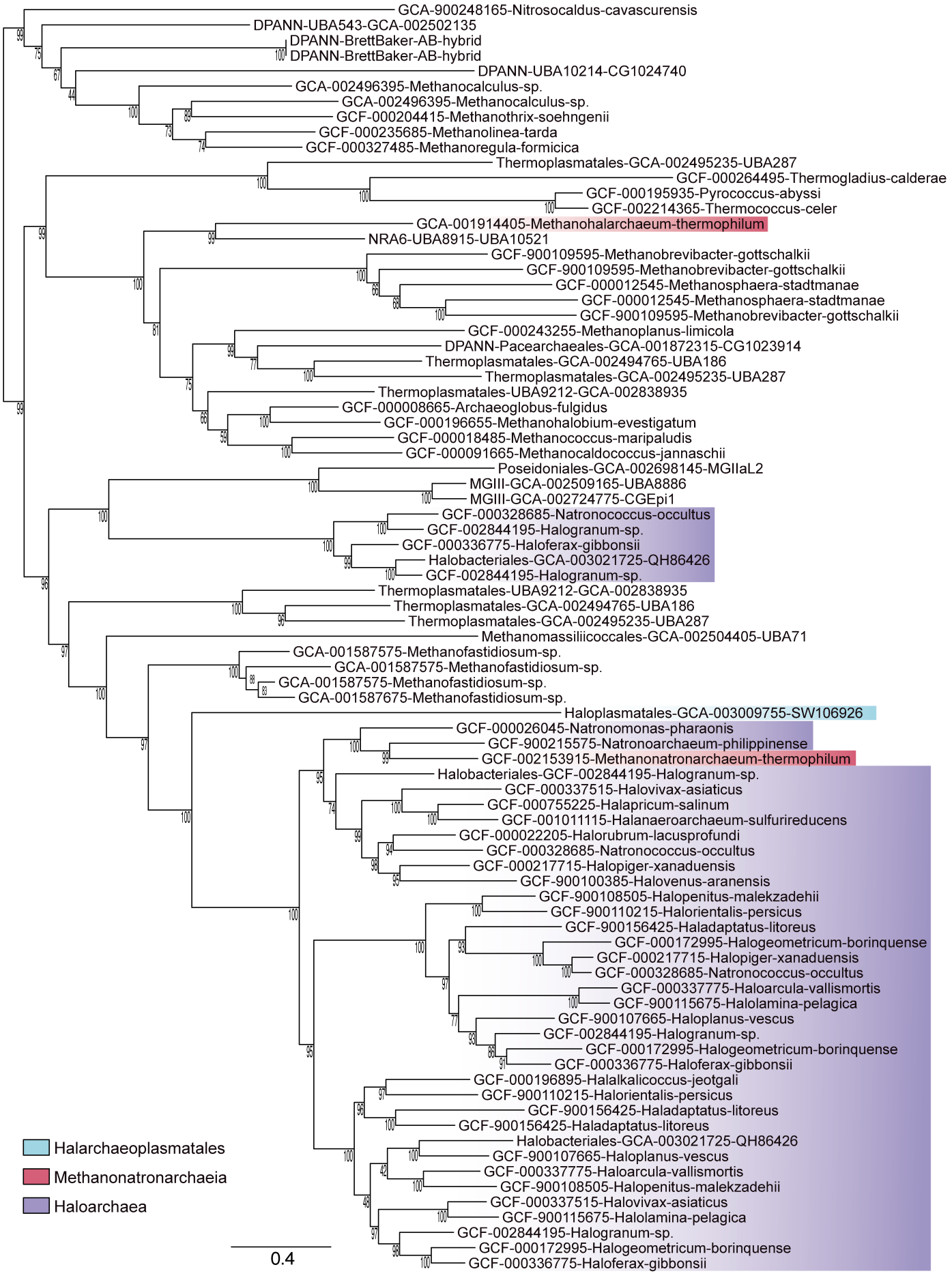
Supplementary Fig. 25 | Maximum likelihood phylogenetic tree of the chaperone GrpE. The tree was constructed with the LG+C60+F+Γ4 model of sequence evolution. Branch support was assessed using 1,000 ultrafast bootstraps. The scale bar represents the number of substitutions per position.



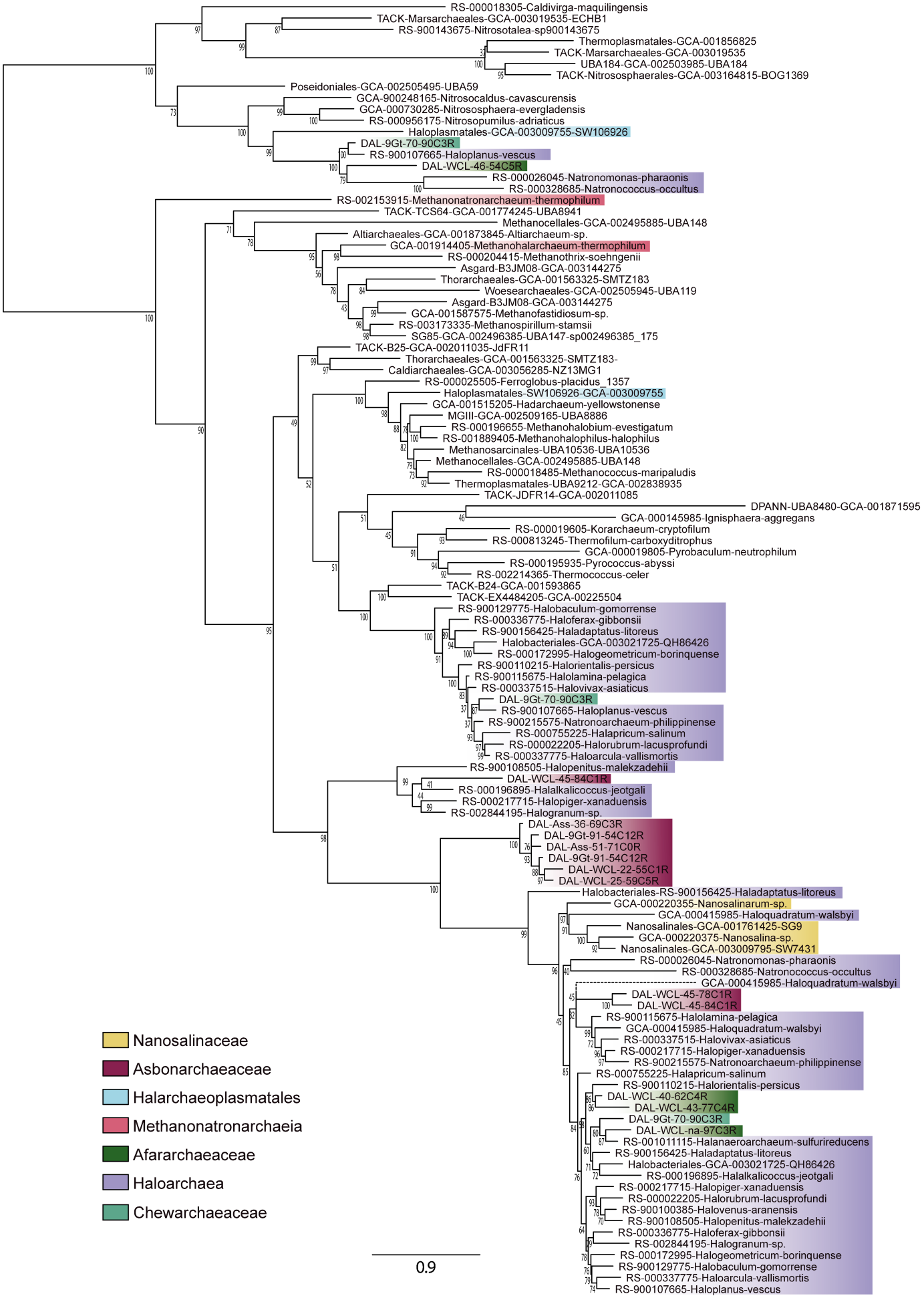
Supplementary Fig. 26 | Maximum likelihood phylogenetic tree of the aerobic-type carbon monoxide dehydrogenase. The tree was constructed with the LG+C60+F+Γ4 model of sequence evolution. Branch support was assessed using 1,000 ultrafast bootstraps. The scale bar represents the number of substitutions per position.



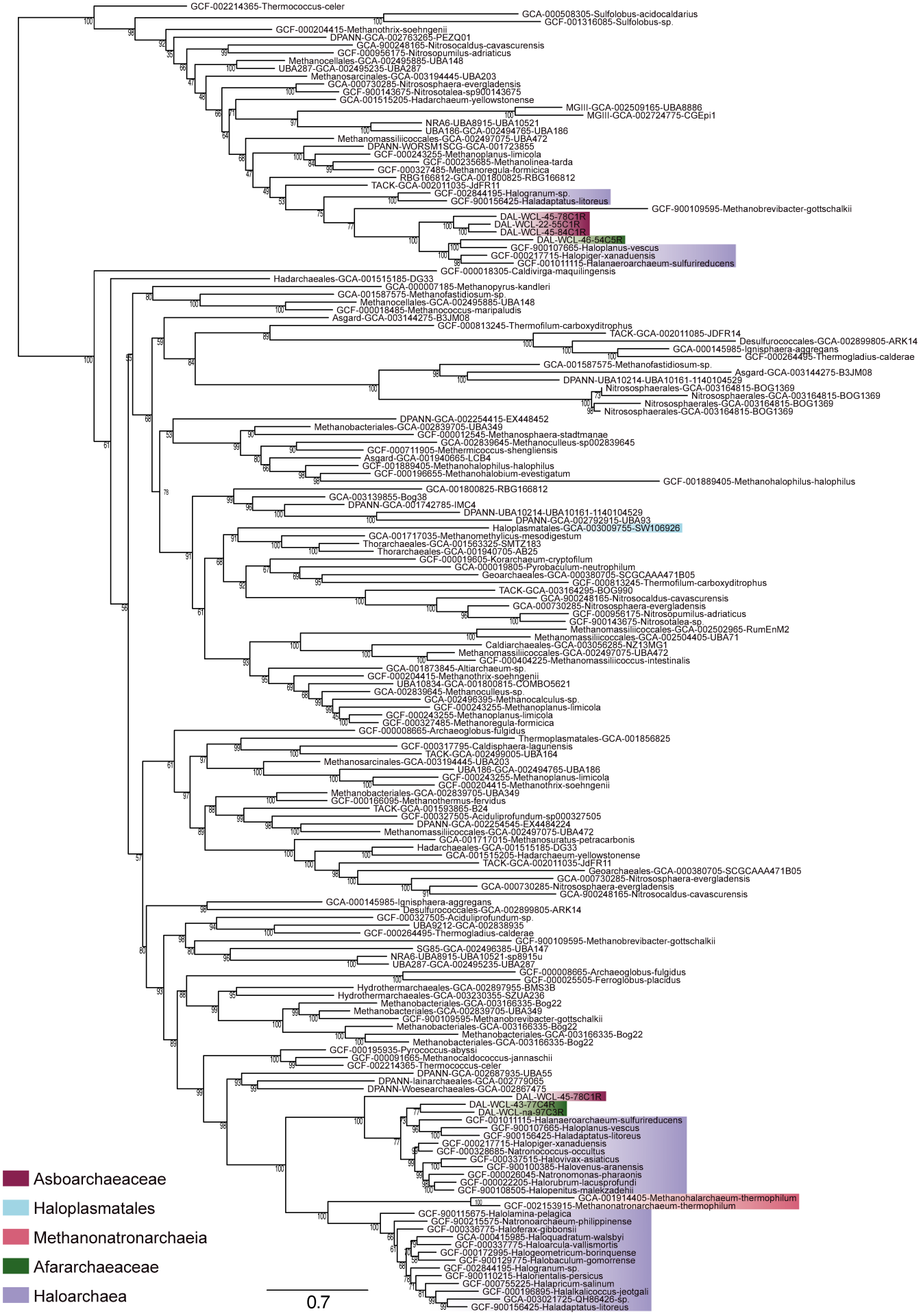
Supplementary Fig. 27 | Maximum likelihood phylogenetic tree of BCCT transporter. The tree was constructed with the LG+C60+F+Γ4 model of sequence evolution. Branch support was assessed using 1,000 ultrafast bootstraps. The scale bar represents the number of substitutions per position.



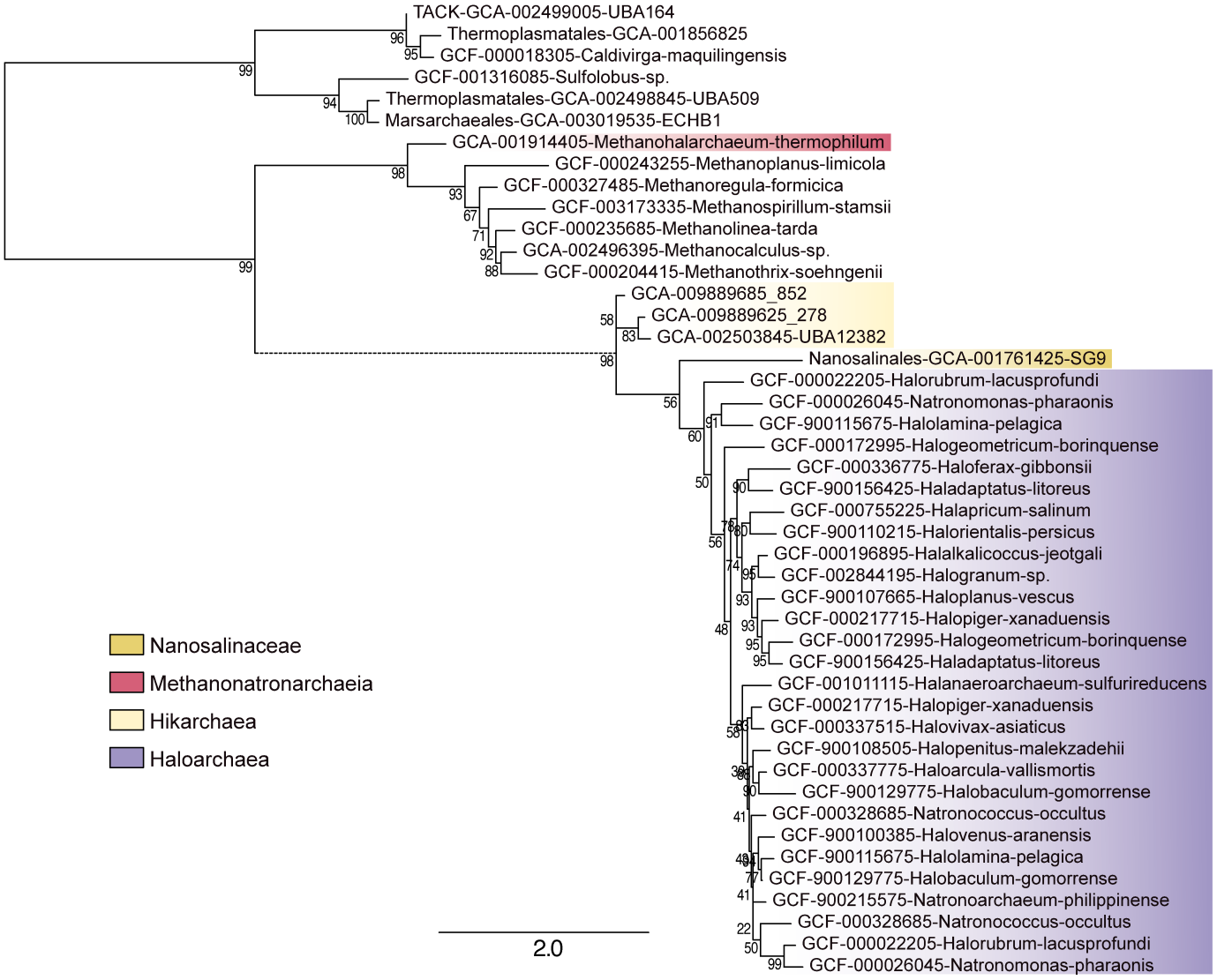
Supplementary Fig. 28 | Maximum likelihood phylogenetic tree of SNF-family Na+-dependent transporters. The tree was constructed with the LG+C60+F+Γ4 model of sequence evolution. Branch support was assessed using 1,000 ultrafast bootstraps. The scale bar represents the number of substitutions per position.



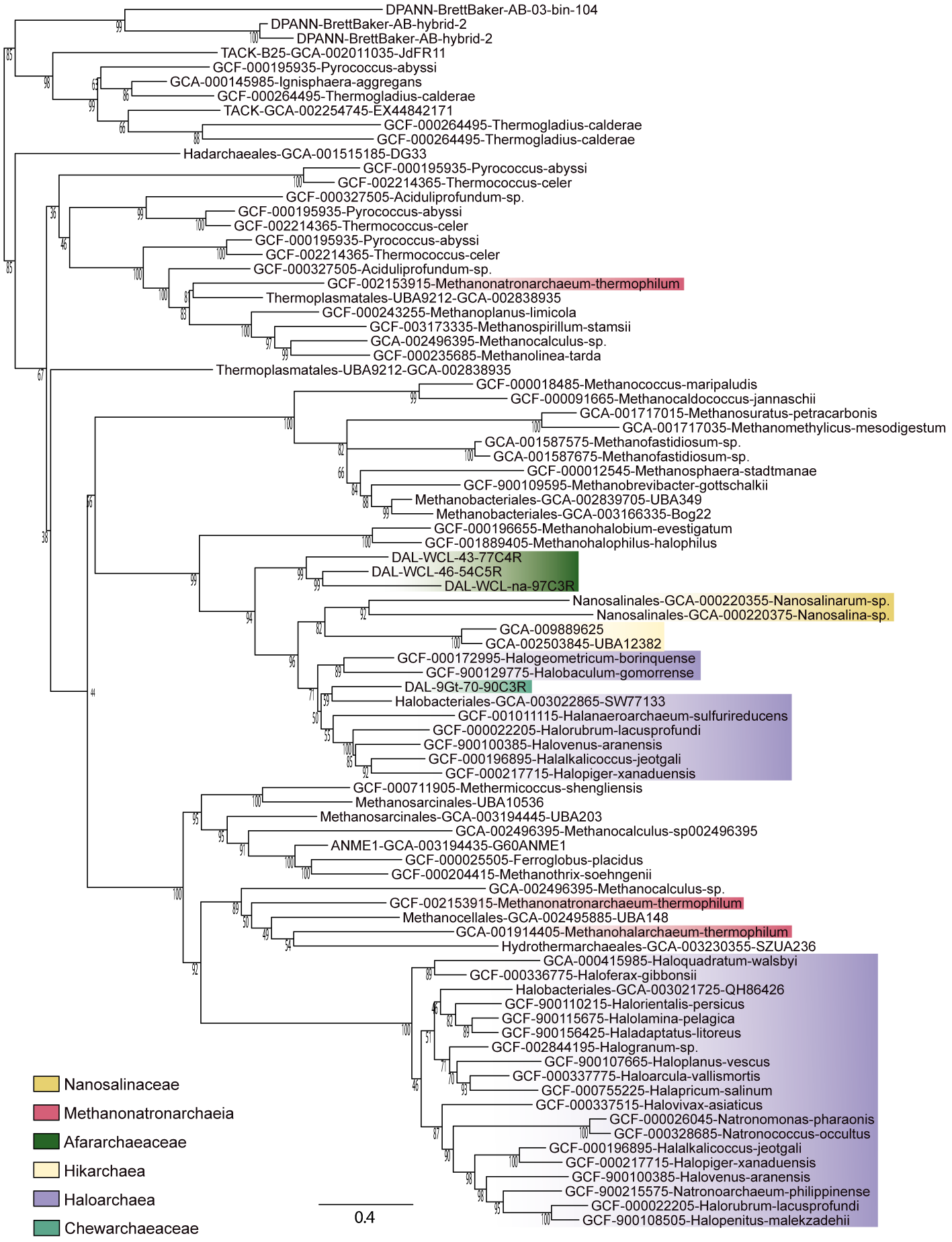
Supplementary Fig. 29 | Maximum likelihood phylogenetic tree of ZupT metal transporters. The tree was constructed with the LG+C60+F+Γ4 model of sequence evolution. Branch support was assessed using 1,000 ultrafast bootstraps. The scale bar represents the number of substitutions per position.



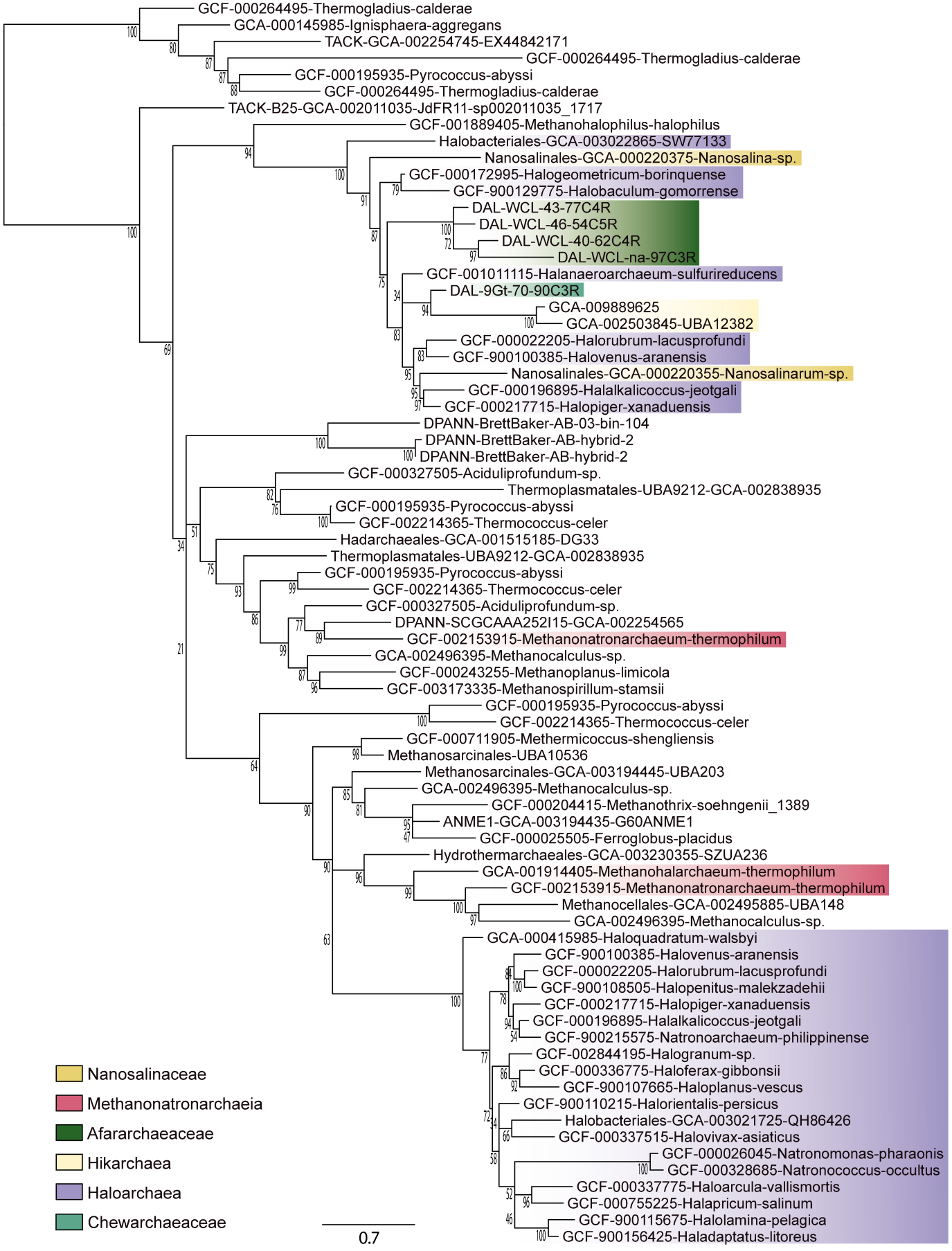
Supplementary Fig. 30 | Maximum likelihood phylogenetic tree of FieF metal transporters. The tree was constructed with the LG+C60+F+Γ4 model of sequence evolution. Branch support was assessed using 1,000 ultrafast bootstraps. The scale bar represents the number of substitutions per position.



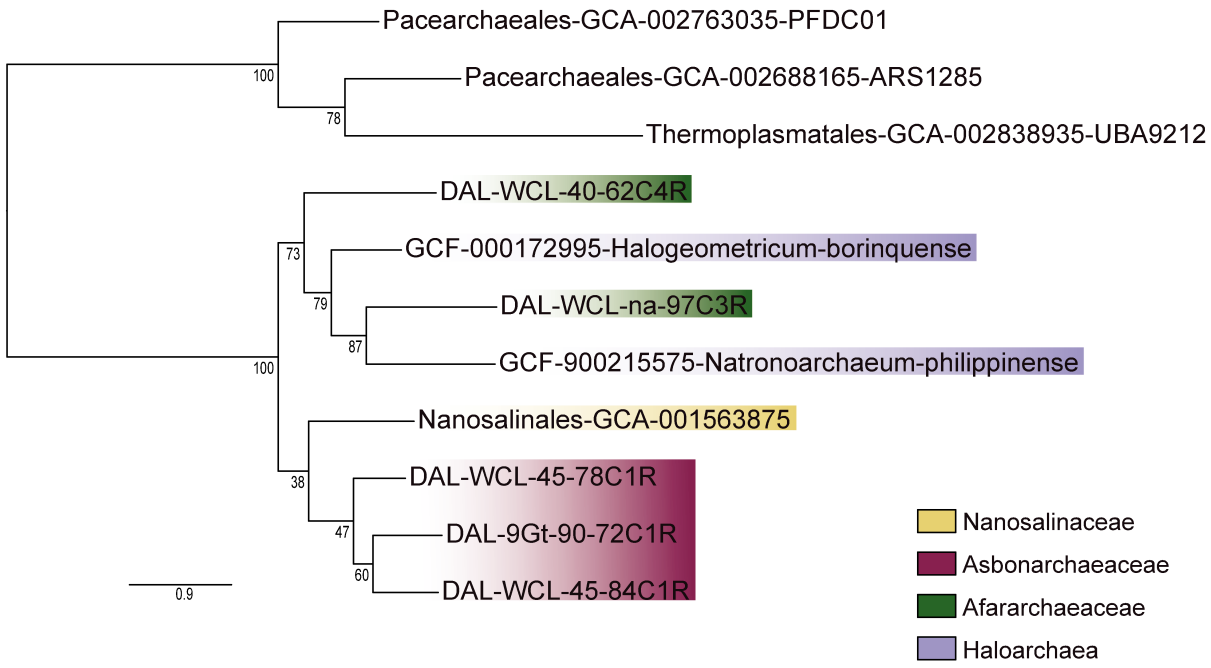
Supplementary Fig. 31 | Maximum likelihood phylogenetic tree of sulfur transporters. The tree was constructed with the LG+C60+F+Γ4 model of sequence evolution. Branch support was assessed using 1,000 ultrafast bootstraps. The scale bar represents the number of substitutions per position.



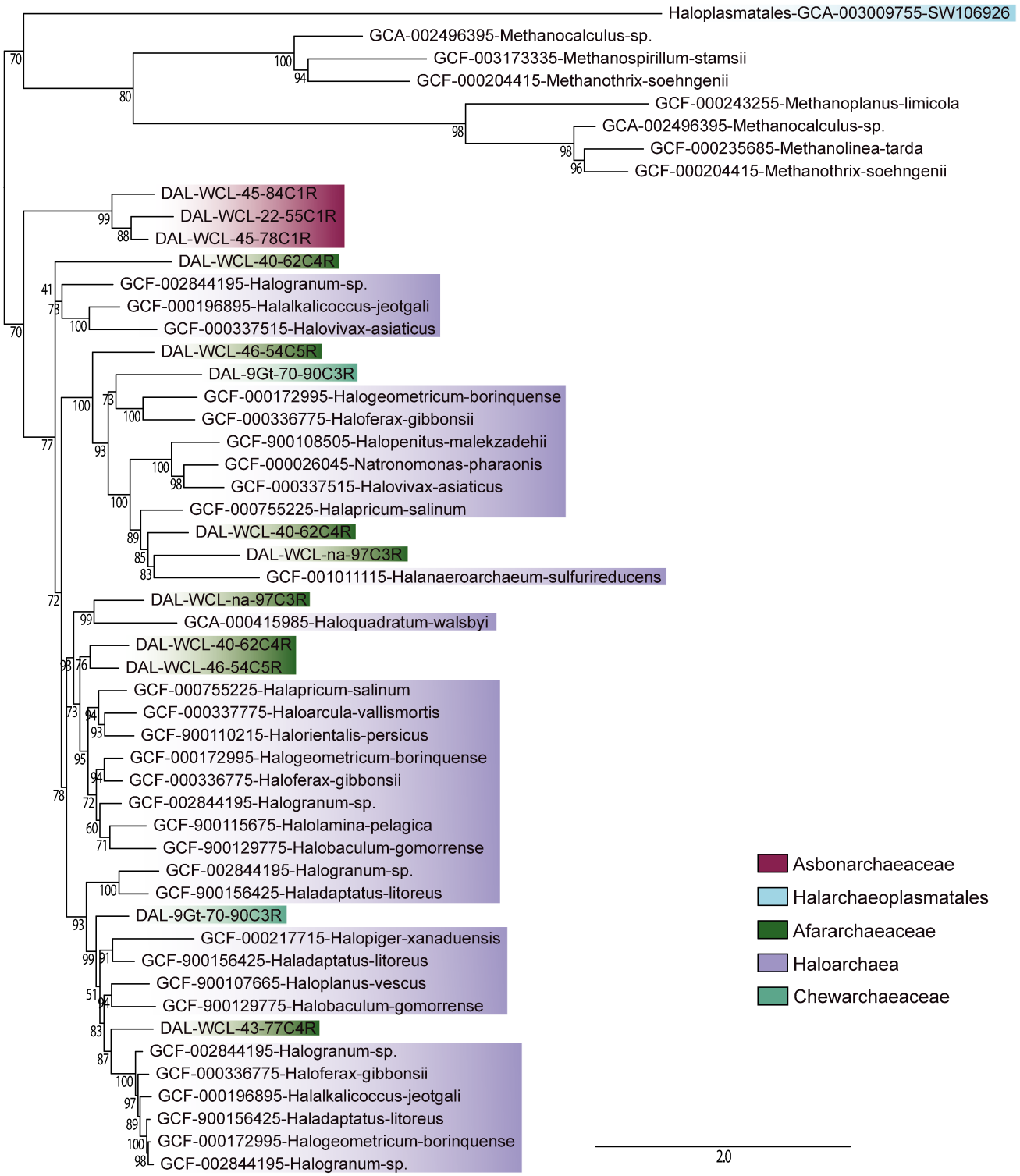
Supplementary Fig. 32 | Maximum likelihood phylogenetic tree of Na+/H+ antiporters. The tree was constructed with the LG+C60+F+Γ4 model of sequence evolution. Branch support was assessed using 1,000 ultrafast bootstraps. The scale bar represents the number of substitutions per position.



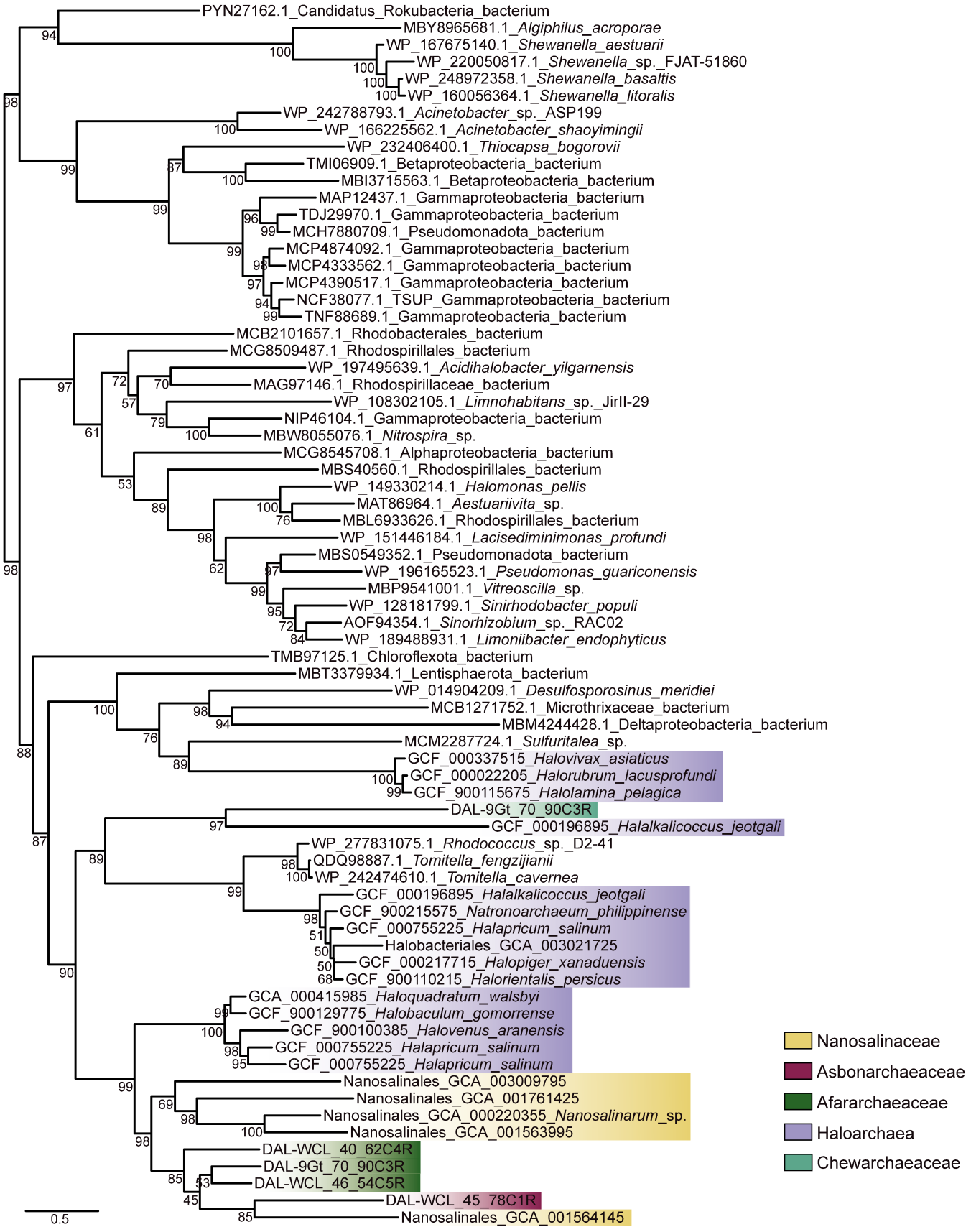
Supplementary Fig. 33 | Maximum likelihood phylogenetic tree of Na+/H+ antiporters. The tree was constructed with the LG+C60+F+Γ4 model of sequence evolution. Branch support was assessed using 1,000 ultrafast bootstraps. The scale bar represents the number of substitutions per position.



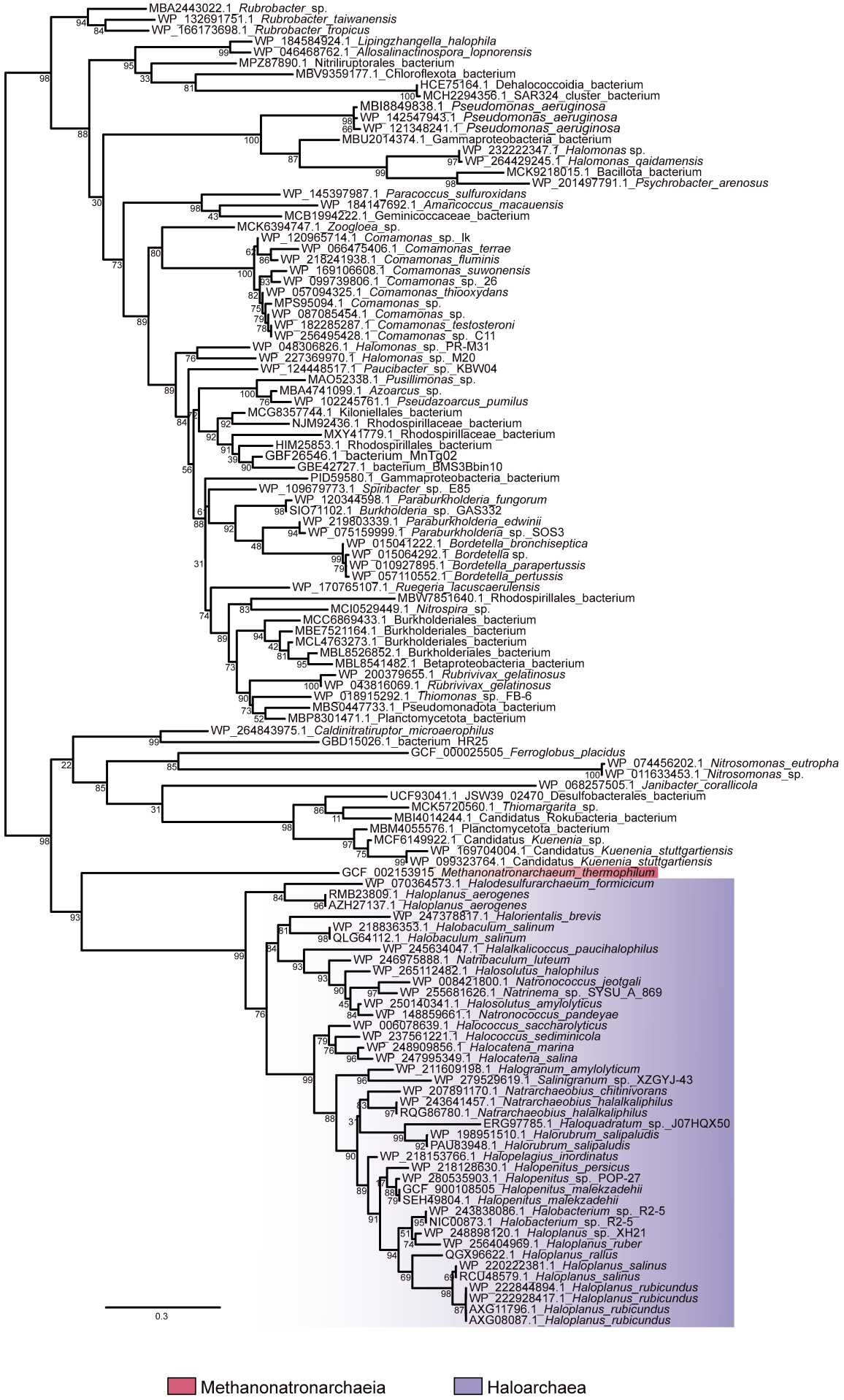
Supplementary Fig. 34 | Maximum likelihood phylogenetic tree of Na+/phosphate symporters. The tree was constructed with the LG+C60+F+Γ4 model of sequence evolution. Branch support was assessed using 1,000 ultrafast bootstraps. The scale bar represents the number of substitutions per position.



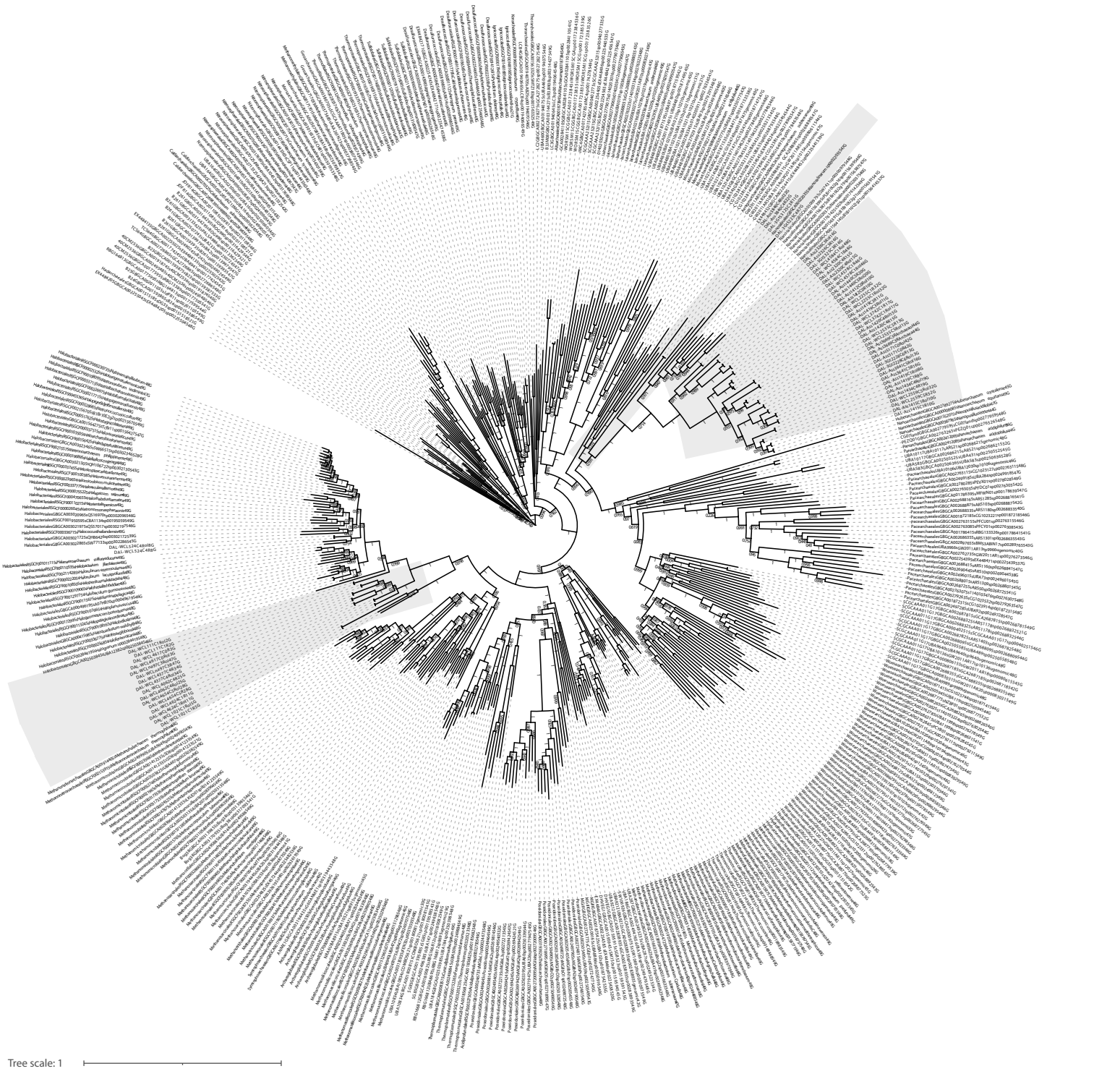
Supplementary Fig. 35 | Maximum likelihood phylogenetic tree of transporters of di- and tricarboxylate Krebs cycle intermediates. The tree was constructed with the LG+C60+F+Γ4 model of sequence evolution. Branch support was assessed using 1,000 ultrafast bootstraps. The scale bar represents the number of substitutions per position.



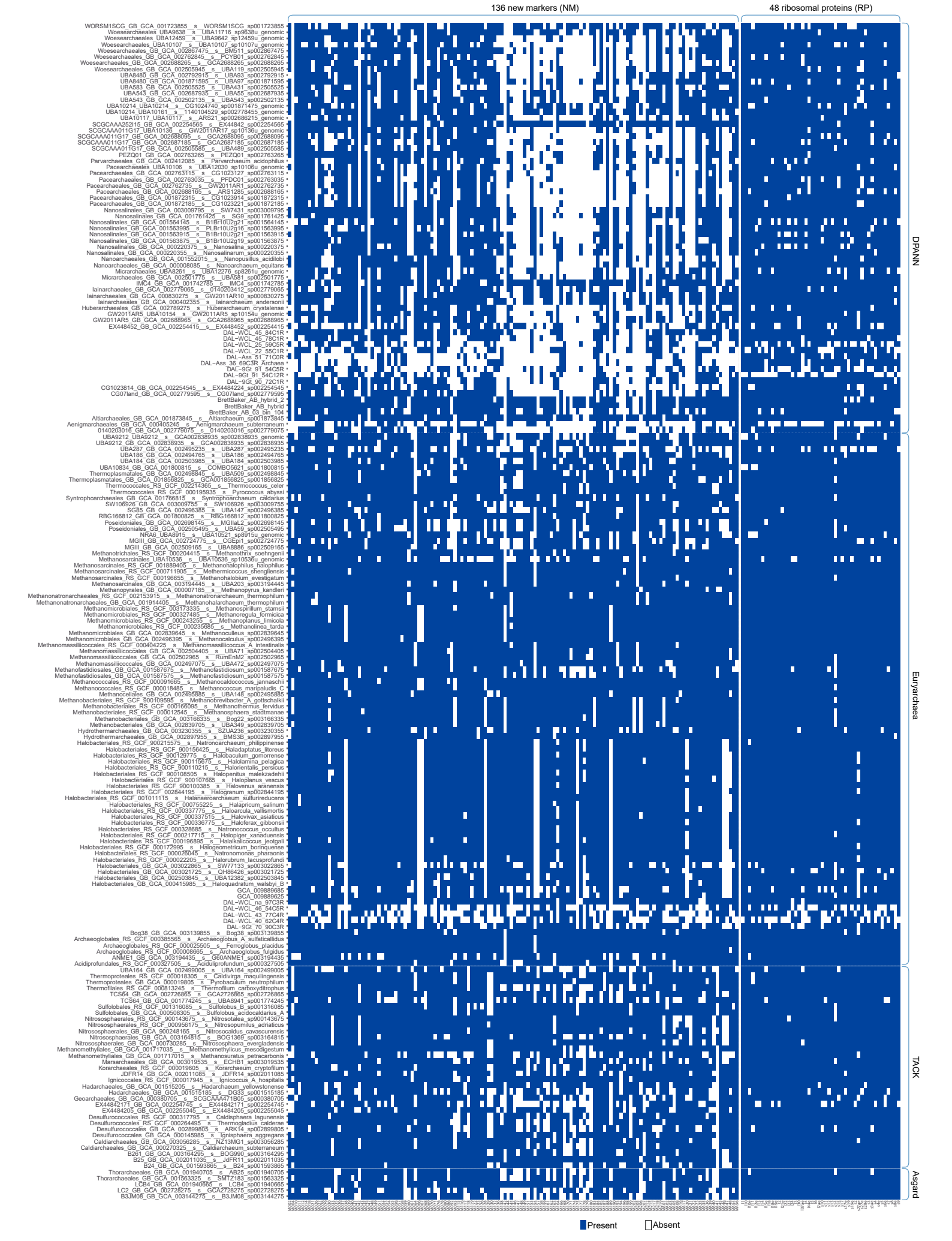
Supplementary Fig. 36 | Maximum likelihood phylogenetic tree of the AmiS/UreI urea transporter. The tree was constructed with the LG+C60+F+Γ4 model of sequence evolution. Branch support was assessed using 1,000 ultrafast bootstraps. The scale bar represents the number of substitutions per position.



Supplementary Fig. 37 | Maximum likelihood phylogenetic tree of TauE/SafE sulfite exporter. The tree was constructed with the LG+C60+F+Γ4 model of sequence evolution. Branch support was assessed using 1,000 ultrafast bootstraps. The scale bar represents the number of substitutions per position.

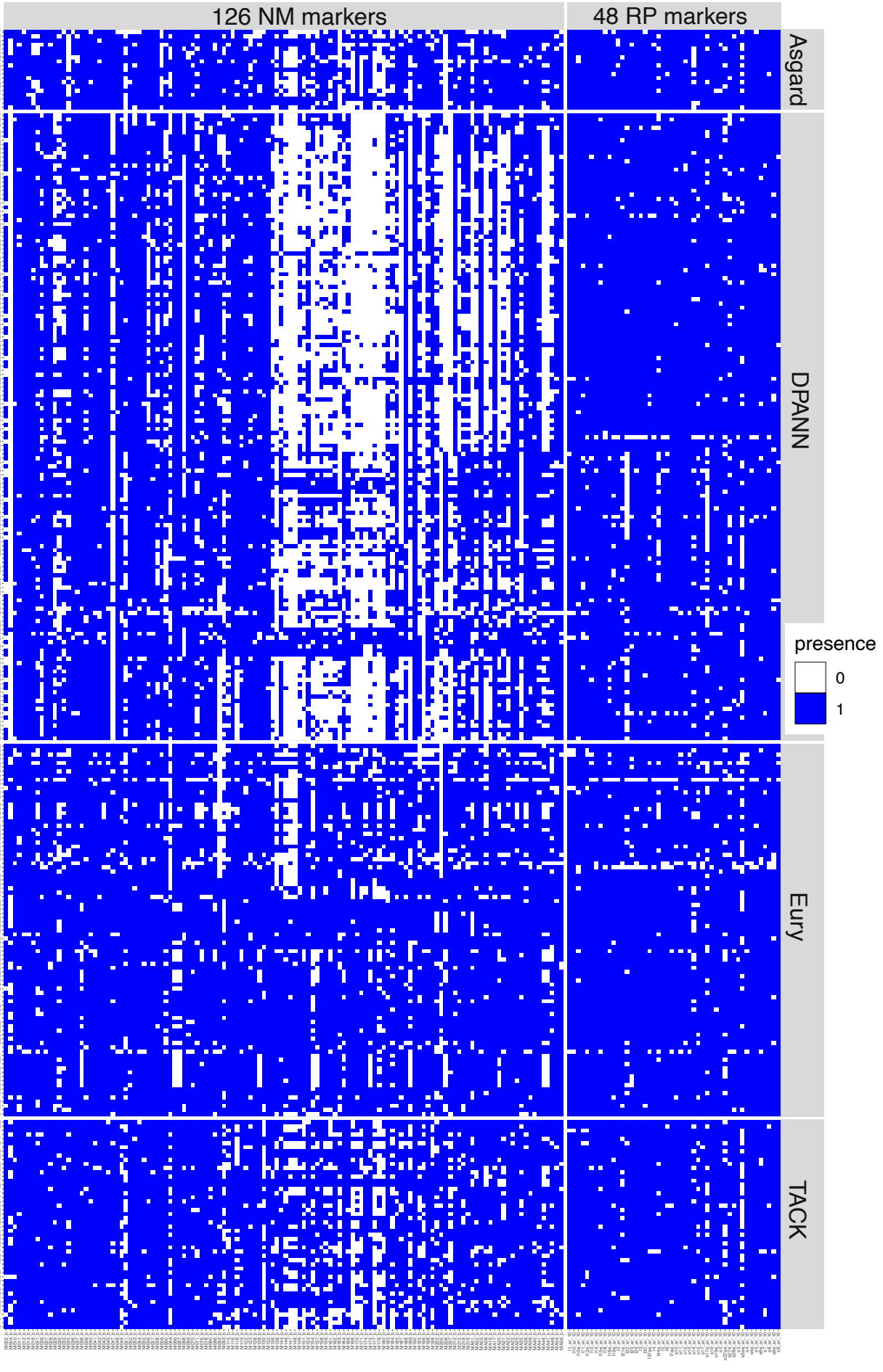


Supplementary Fig. 38 | Maximum likelihood phylogeny of 427 archaeal taxa based on the concatenated alignment of 49 ribosomal proteins. he ML tree was constructed using the LG+C20+F+Γ4 model of evolution. Branch support was assessed using 1,000 ultrafast bootstraps via IQ-TREE implementation. The scale bar represents the estimated number of substitutions per site. Each tip label provides information about the archaeal supergroup, taxonomic order based on GTDB, GTDB accession number, species identification, and the number of markers identified for each taxon out of the 49 ribosomal proteins**.**

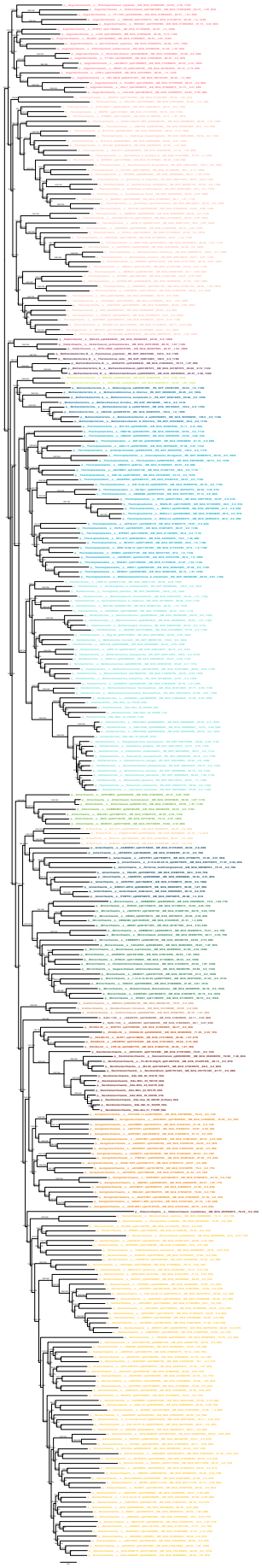


Supplementary Fig. 39 | Distribution of NM and RP markers in the 192 archaeal taxa used in this study.

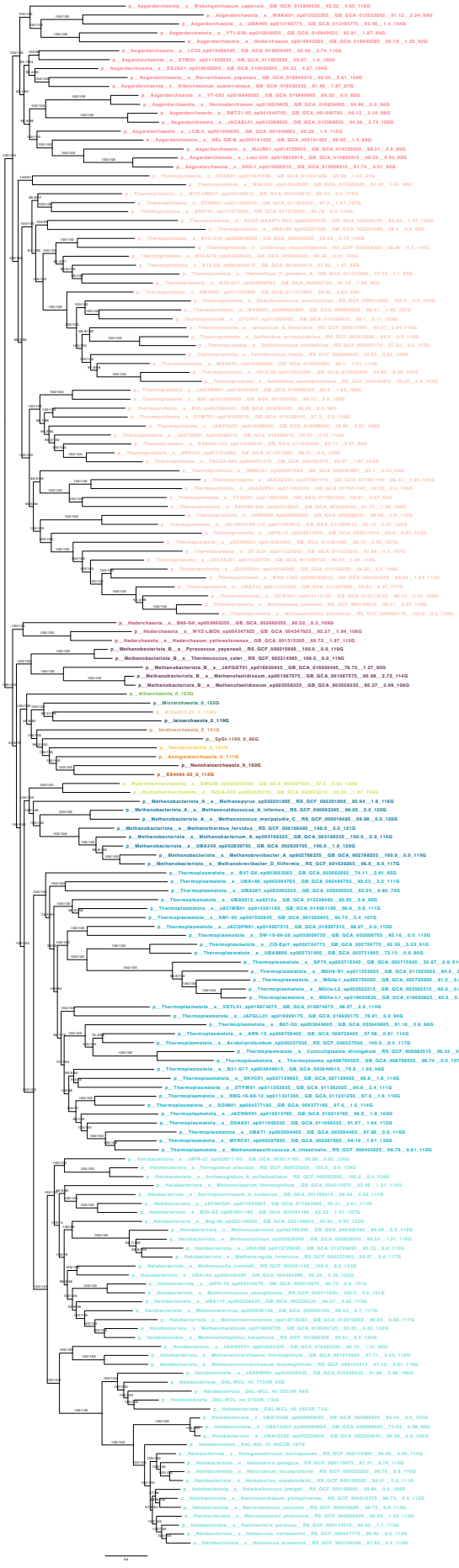
11. Supplementary Material of Manuscript 2



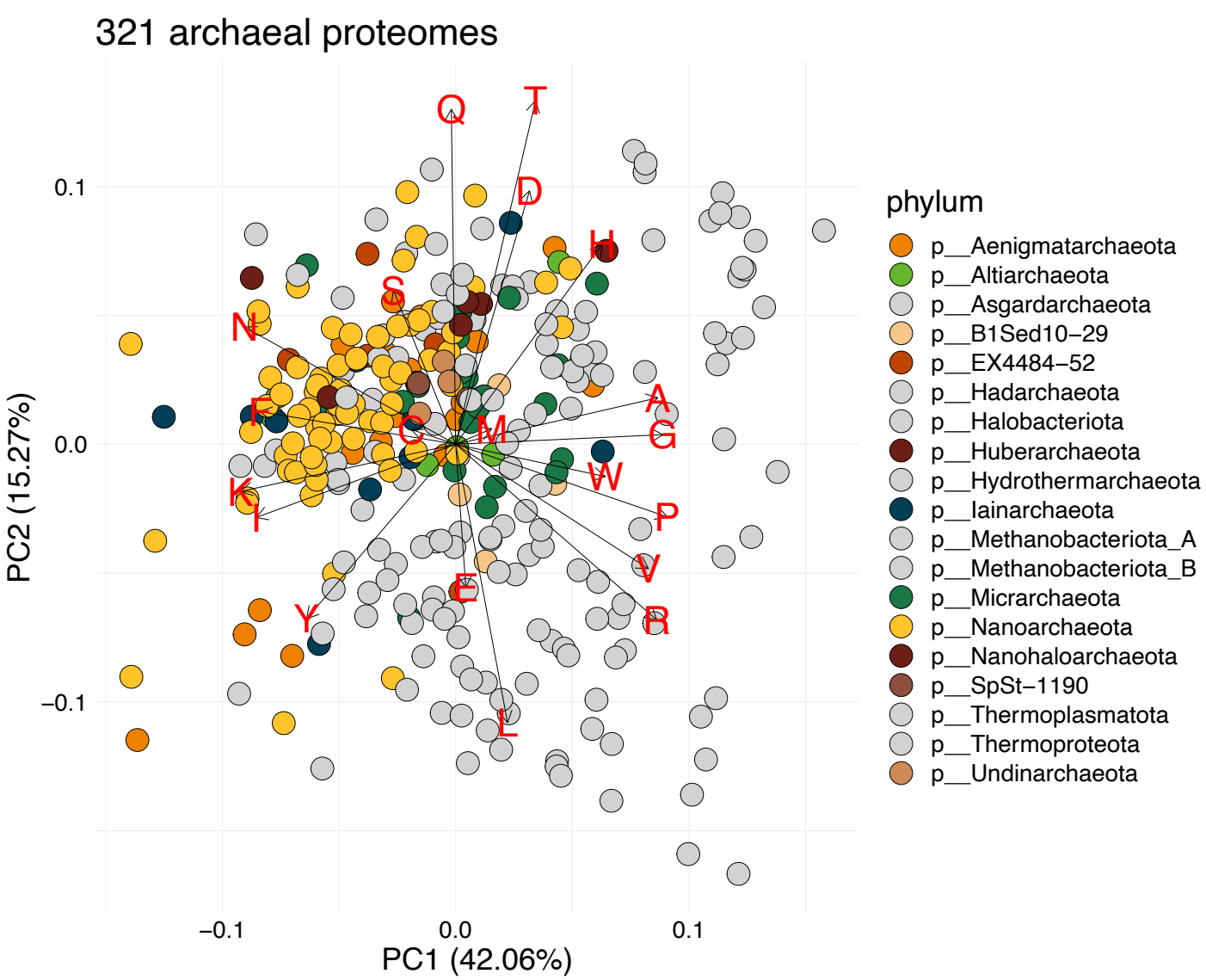
Supplementary Figure 1 | Presence and absence heatmap of the NM126 and RP48 marker datasets. The x-axis contains the 321 archaeal taxa, and the y-axis contains the individual NM and RP markers. The blue squares indicate the presence, and the absence is white.



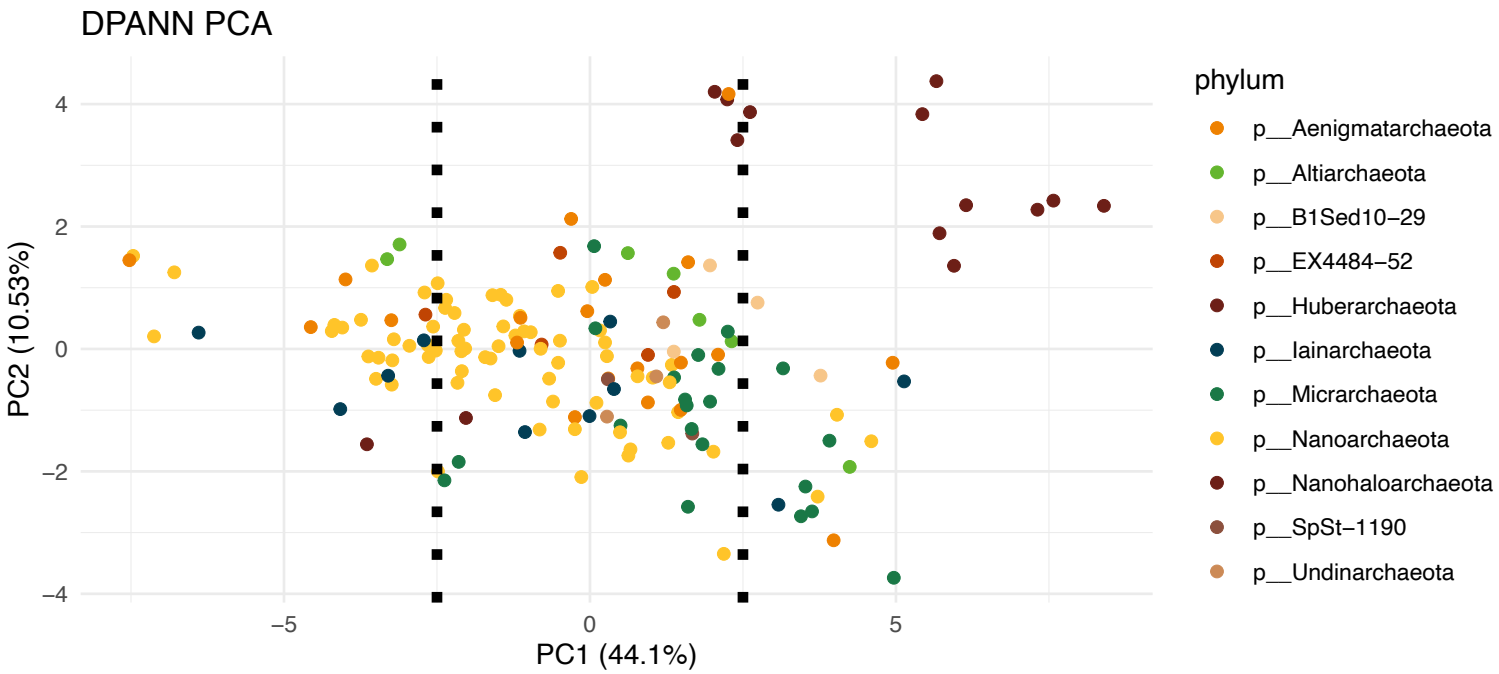
Supplementary Figure 2 | Maximum likelihood phylogeny of 321 archaeal taxa based on the NM126 dataset. The ML tree was reconstructed using the LG+C60+Γ4 model. SH-aLRT/ultrafast bootstrap support is shown above each branch via IQ-TREE implementation. Each color represents one of the 19 archaeal phyla, according to GTDB r207. The scale bar represents the estimated number of substitutions per site. Each tip label contains the GTDB phylum, species, accession number, completeness, contamination, and the total number of markers present in that taxa out of the NM126 dataset.



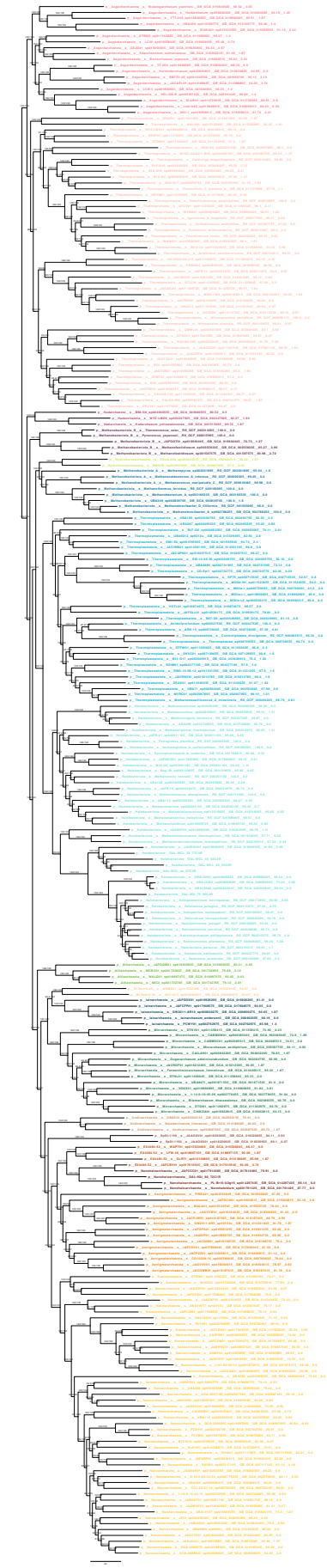
Supplementary Figure 3 | Maximum likelihood phylogeny of NM126 one-chimera tree with no Huberarchaeota. The ML tree was reconstructed using the LG+C60+Γ4 model. SH-aLRT/ultrafast bootstrap support is shown above each branch via IQ-TREE implementation. Each color represents one of the 19 archaeal phyla, according to GTDB r207. The scale bar represents the estimated number of substitutions per site. Each tip label contains the GTDB phyla, species, accession number, completeness, contamination, and the total number of markers present in that taxa out of the NM126 dataset. The chimera taxa are indicated by the DPANN GTDB phylum, followed by a _0, and the total number of markers per chimera.



Supplementary Figure 4 | PCA plot of 321 archaeal proteomes based on amino acid frequencies. Colored circles represent the 11 DPANN phyla according to GTDB r207, and gray circles represent the other eight phyla.

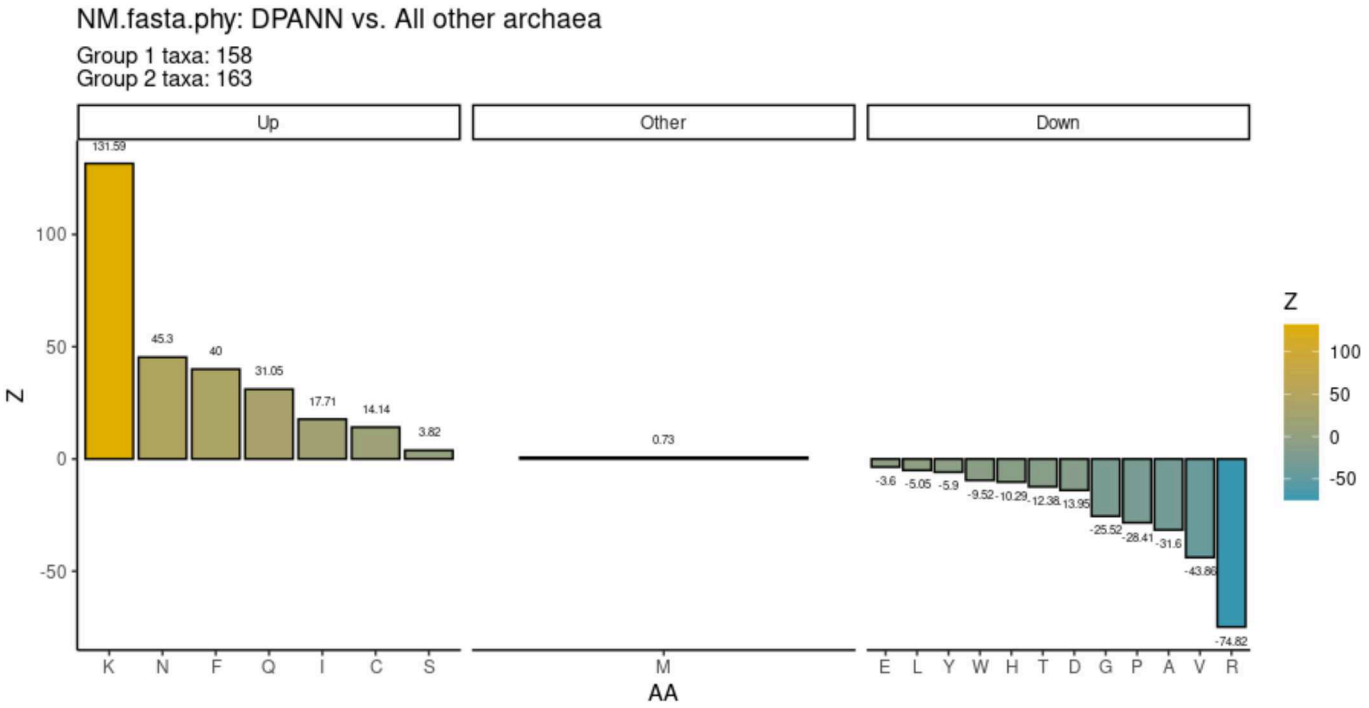


Supplementary Figure 5 | PCA plot of 158 DPANN proteomes based on amino acid frequencies. Each color represents one of the 11 DPANN phyla. The black dashed line indicates the cut-off for taxa removal. Taxa outside these dashed lines (less than -2.5 or greater than 2.5) were removed from the NM126 dataset.

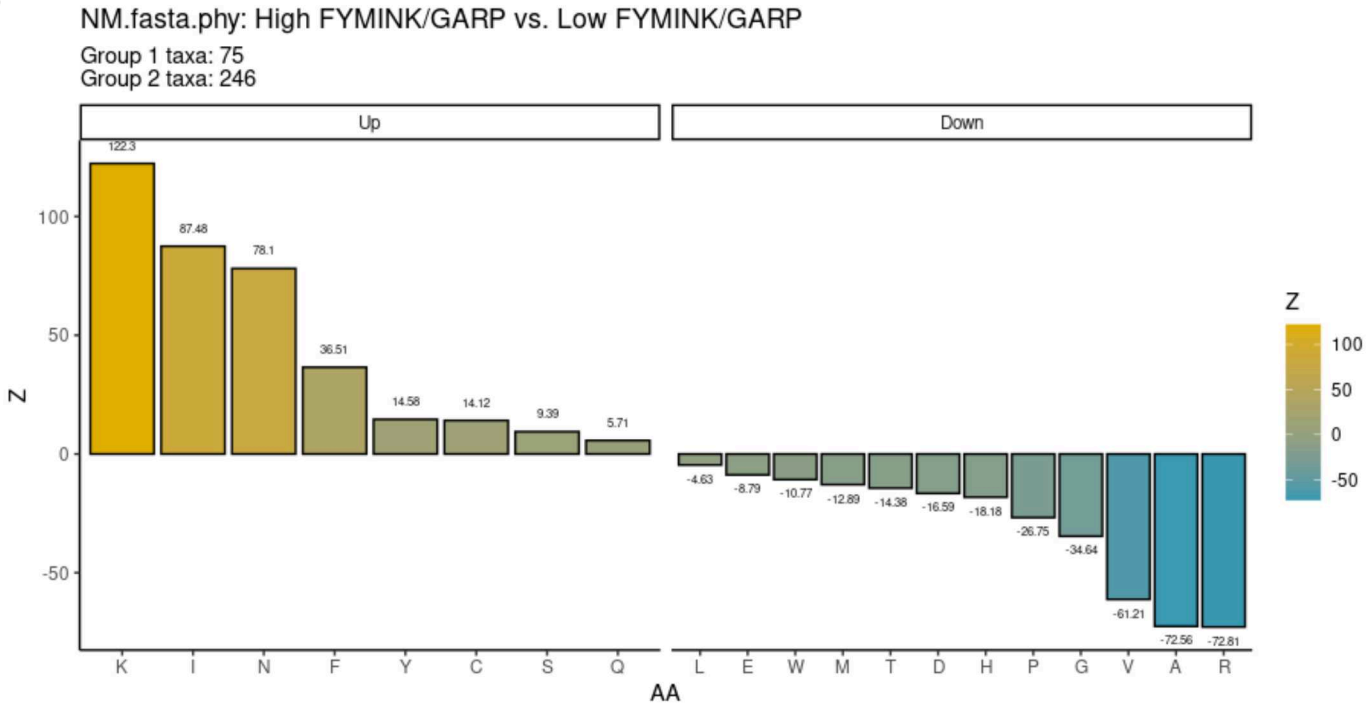


Supplementary Figure 6 | Maximum likelihood phylogeny of NM126 PCA tree with “most biased” DPANN taxa removed. The ML tree was reconstructed using the LG+C60+Γ4 model. SH-aLRT/ultrafast bootstrap support is shown above each branch via IQ-TREE implementation. Each color represents one of the 19 archaeal phyla, according to GTDB r207. The scale bar represents the estimated number of substitutions per site. Each tip label contains the GTDB phylum, species, accession number, completeness, contamination, and the total number of markers present in that taxa out of the NM126 dataset.

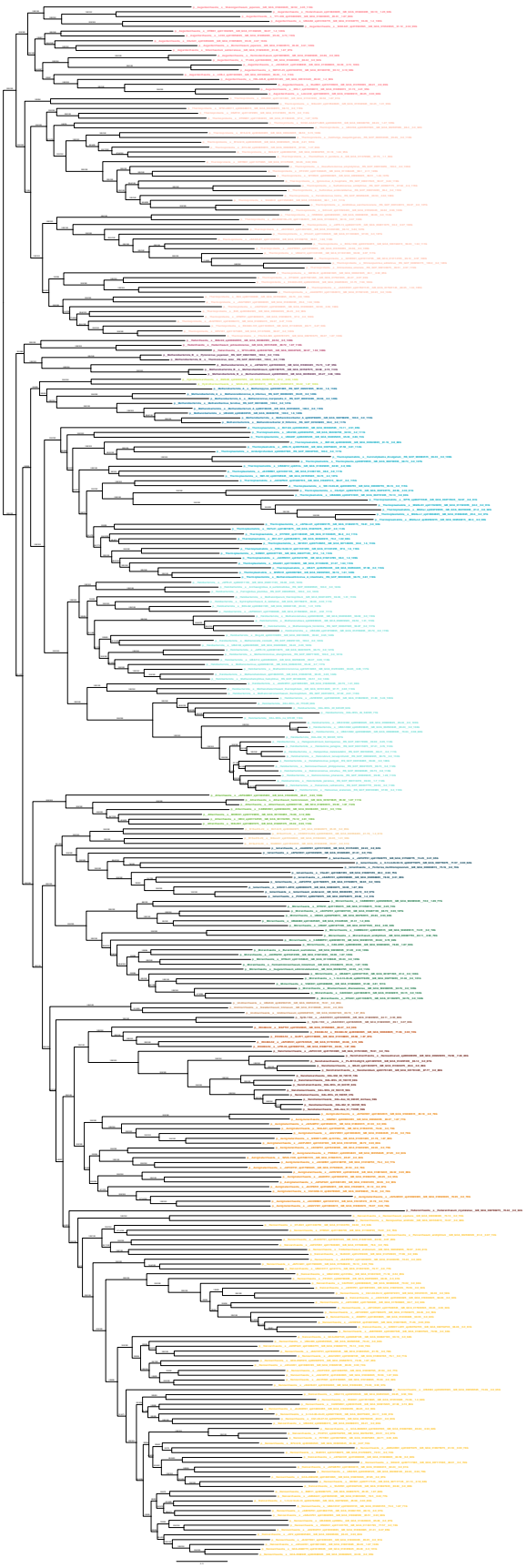
a



b

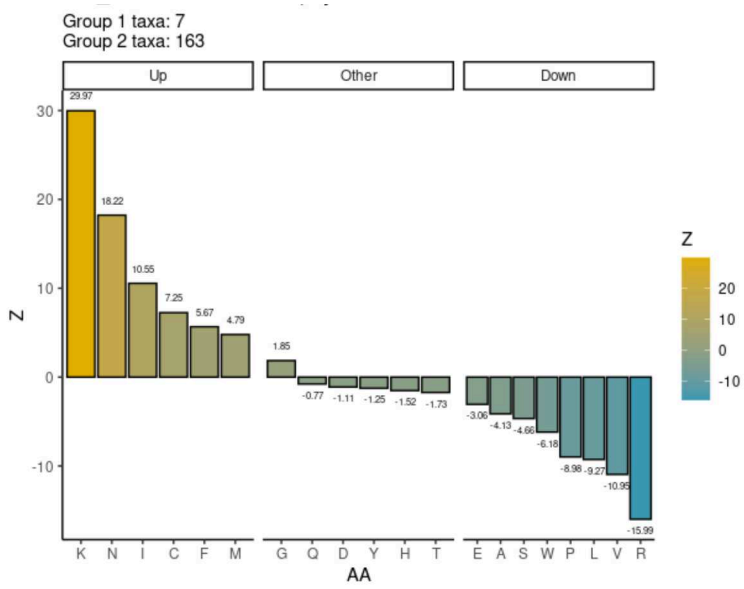


Supplementary Figure 7 | Z-score from a binomial test of two proportions for the full-NM126 dataset. Z-scores were calculated relative to the (a) DPANN and (b) FIMNKY-enriched taxa, with |Z| > 1.96 indicating significant enrichment of a given amino acid in (a) DPANN or (b) FIMNKY-enriched taxa sequences (“Up”), |Z| < -1.96 indicating significant depletion of a given amino acid in (a) DPANN and (b) FIMNKY-enriched taxa (“Down”), and some amino acids showing no significant bias (“Other”).

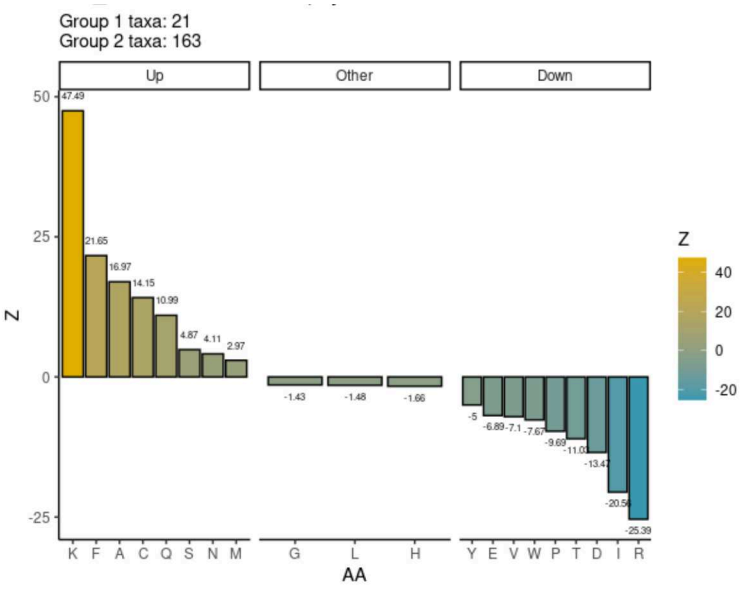


Supplementary Figure 8 | Maximum likelihood phylogeny of NM126 with 15% most GARP/FIMNKY-biased sites removed binned DPANN versus non-DPANN lineages. The ML tree was reconstructed using the LG+C60+Γ4 model. SH-aLRT/ultrafast bootstrap support is shown above each branch via IQ-TREE implementation. Each color represents one of the 19 archaeal phyla, according to GTDB r207. The scale bar represents the estimated number of substitutions per site. Each tip label contains the GTDB phylum, species, accession number, completeness, contamination, and the total number of markers present in that taxa out of the NM126 dataset.

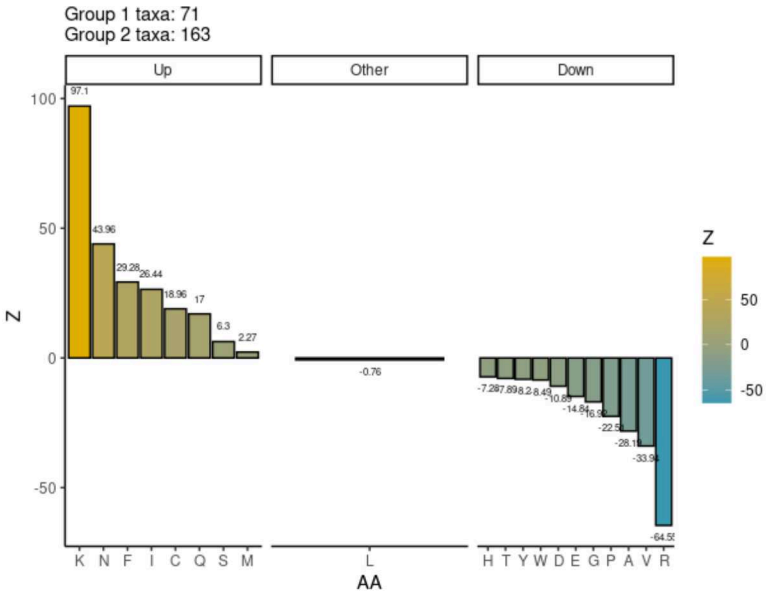
a NM126 - Altiarchaeota



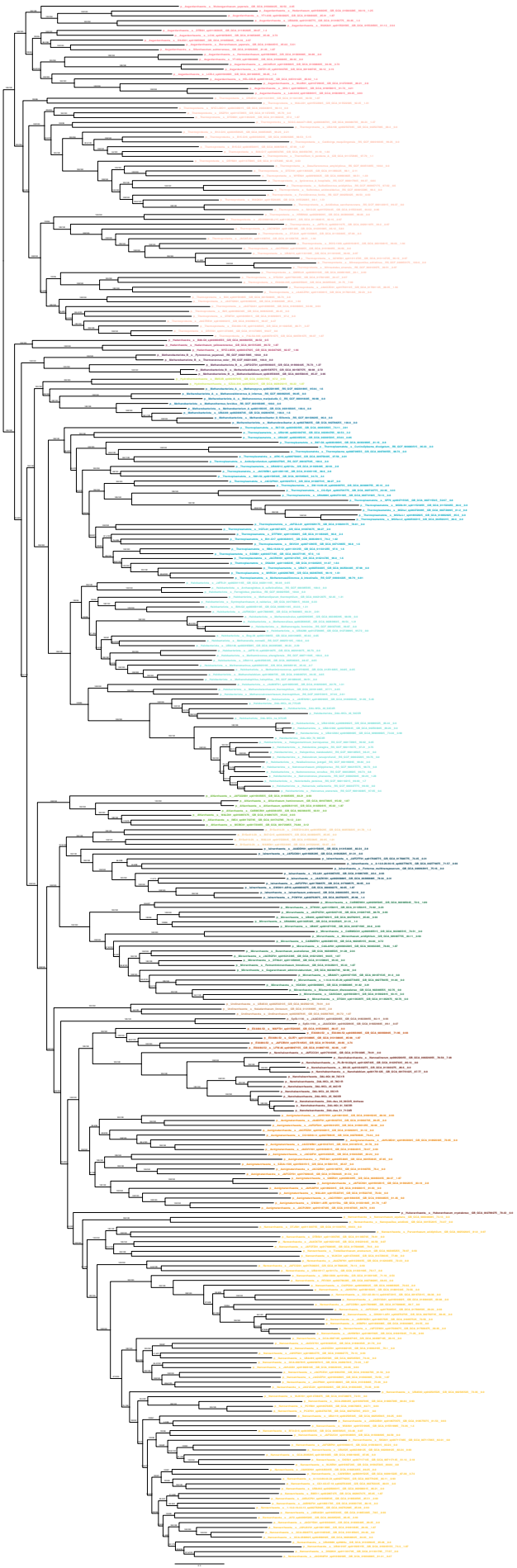
b NM126 - Micrarchaeota



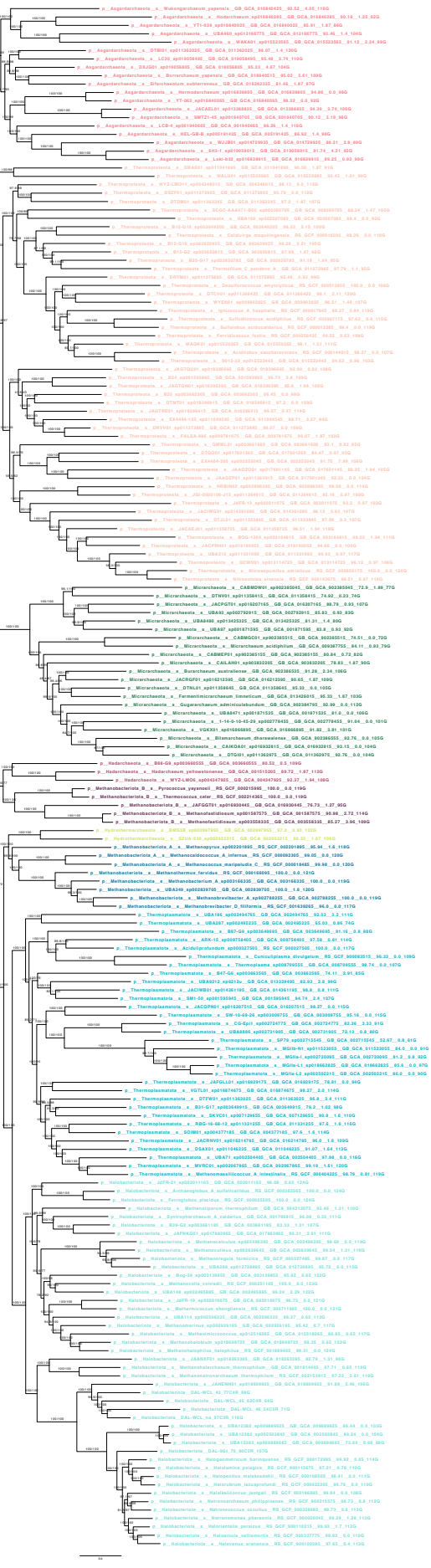
c NM126 - Nanoarchaeota



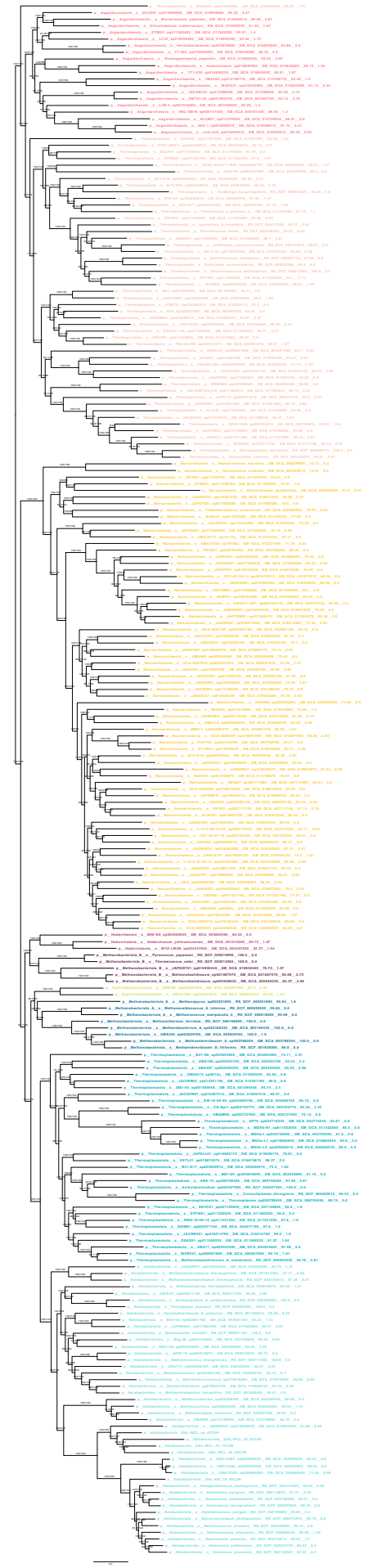
Supplementary Figure 9 | Z-score from a binomial test of two proportions for NM126 subset datasets. Z-scores were calculated relative to the (a) Altiarchaeota, (b) Micrarchaeota, or (c) Nanoarchaeota, with |Z| > 1.96 indicating significant enrichment of a given amino acid in (a) Altiarchaeota, (b) Micrarchaeota, or (c) Nanoarchaeota sequences (“Up”), |Z| < -1.96 indicating significant depletion of a given amino acid in (a) Altiarchaeota, (b) Micrarchaeota, or (c) Nanoarchaeota (“Down”), and some amino acids showing no significant bias (“Other”).



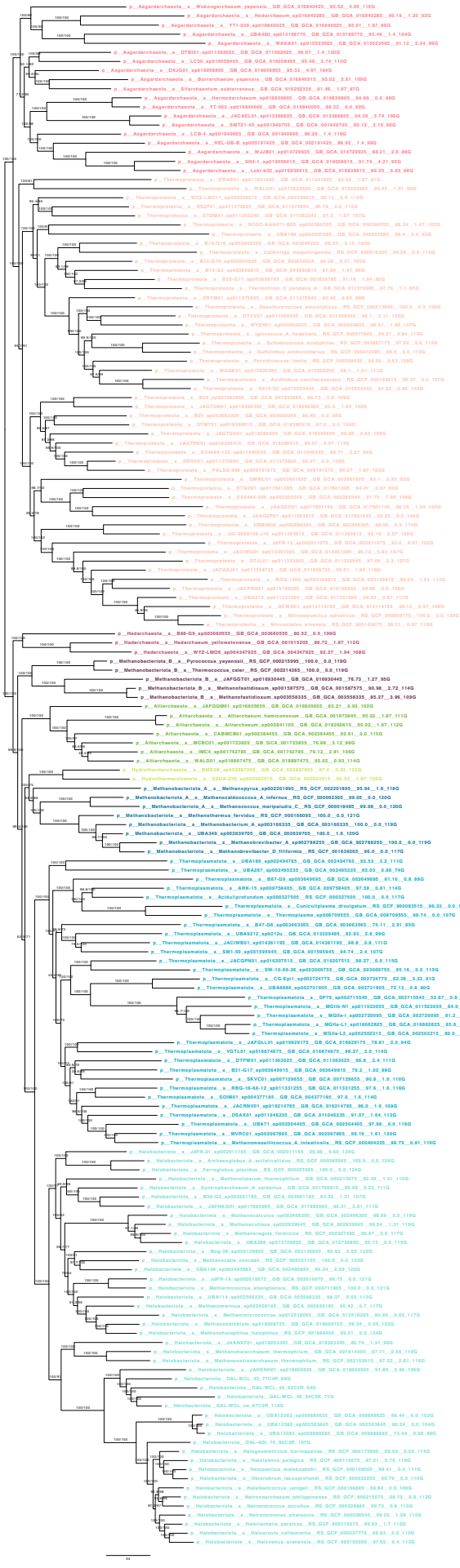
Supplementary Figure 10 | Maximum likelihood phylogeny of NM126 with 15% most GARP/FIMNKY-biased sites removed binned FIMNKY-enriched taxa versus GARP-enriched taxa. The ML tree was reconstructed using the LG+C60+Γ4 model. SH-aLRT/ultrafast bootstrap support is shown above each branch via IQ-TREE implementation. Each color represents one of the 19 archaeal phyla, according to GTDB r207. The scale bar represents the estimated number of substitutions per site. Each tip label contains the GTDB phylum, species, accession number, completeness, contamination, and the total number of markers present in that taxa out of the NM126 dataset.



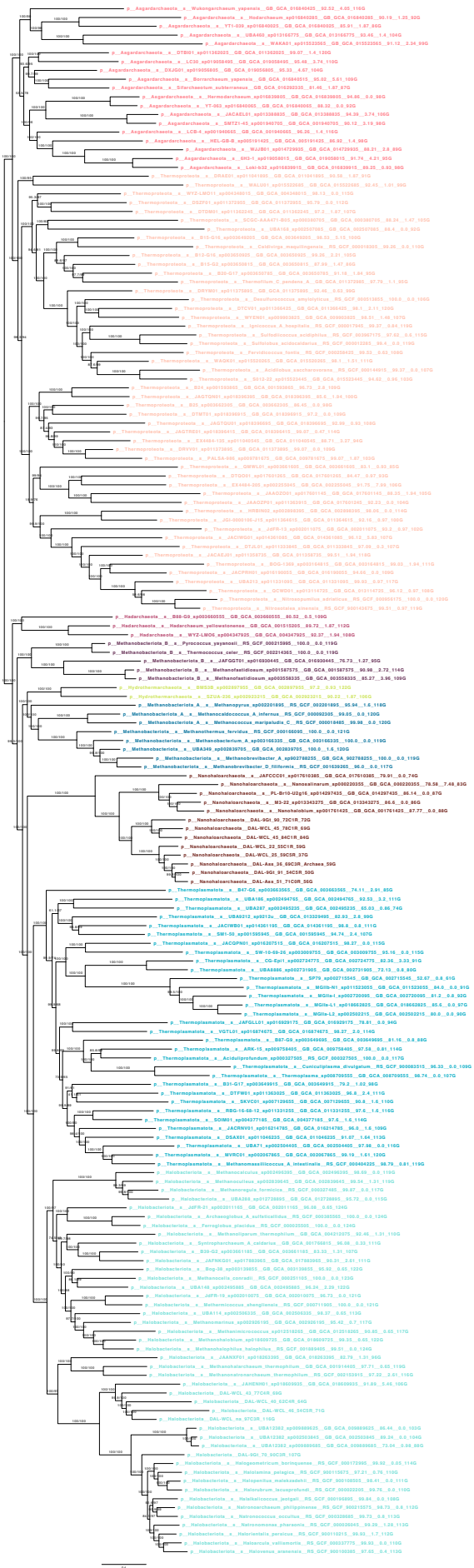
Supplementary Figure 11 | Maximum likelihood phylogeny of NM126 Micrarchaeota only tree. The ML tree was reconstructed using the LG+C60+Γ4 model. SH-aLRT/ultrafast bootstrap support is shown above each branch via IQ-TREE implementation. Each color represents one of the 19 archaeal phyla according to GTDB r207. The scale bar represents the estimated number of substitutions per site. Each tip label contains the GTDB phylum, species, accession number, completeness, contamination, and the total number of markers present in that taxa out of the NM126 dataset.



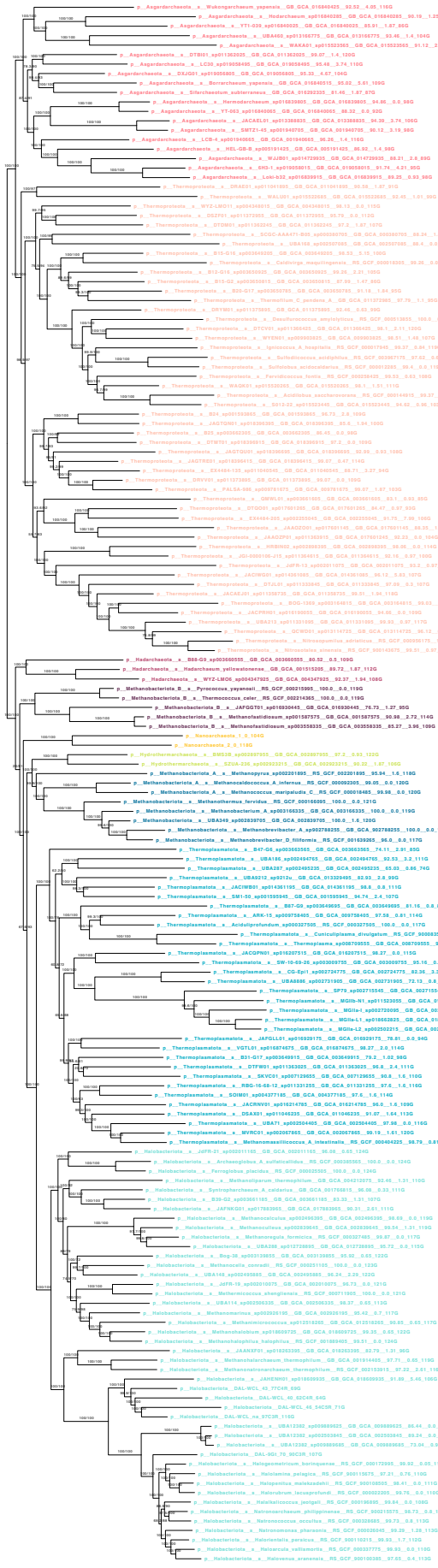
Supplementary Figure 12 | Maximum likelihood phylogeny of NM126 Nanoarchaeota only tree. The ML tree was reconstructed using the LG+C60+Γ4 model. SH-aLRT/ultrafast bootstrap support is shown above each branch via IQ-TREE implementation. Each color represents one of the 19 archaeal phyla, according to GTDB r207. The scale bar represents the estimated number of substitutions per site. Each tip label contains the GTDB phylum, species, accession number, completeness, contamination, and the total number of markers present in that taxa out of the NM126 dataset.



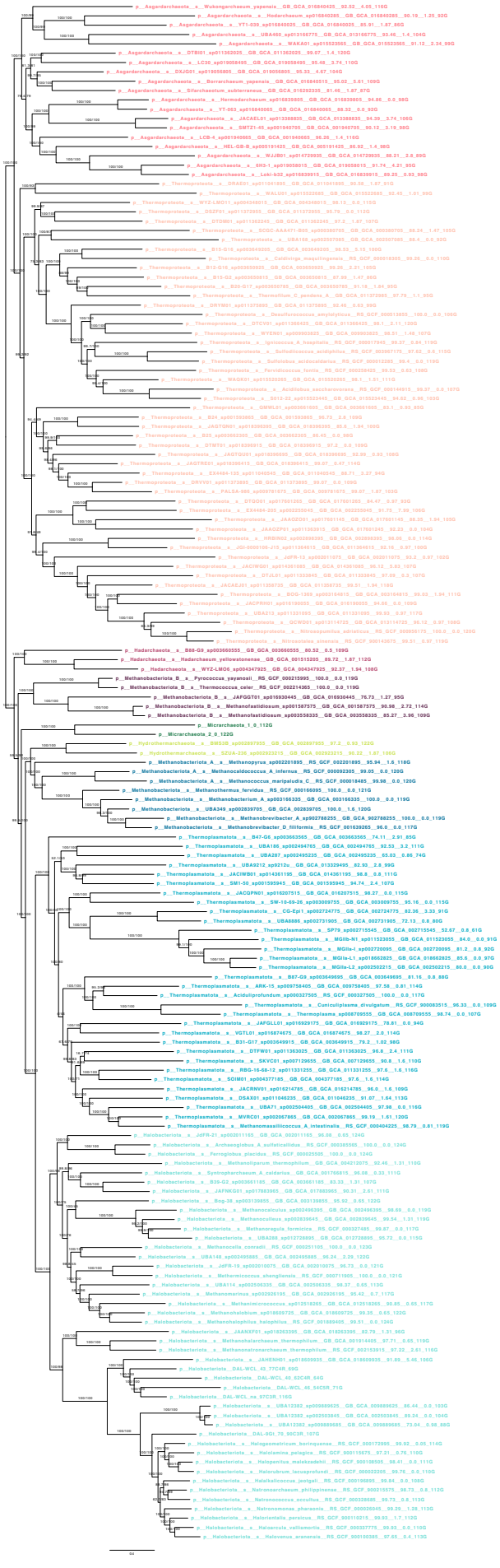
Supplementary Figure 13 | Maximum likelihood phylogeny of NM126 Altiarchaeota only tree. The ML tree was reconstructed using the LG+C60+Γ4 model. SH-aLRT/ultrafast bootstrap support is shown above each branch via IQ-TREE implementation. Each color represents one of the 19 archaeal phyla, according to GTDB r207. The scale bar represents the estimated number of substitutions per site. Each tip label contains the GTDB phylum, species, accession number, completeness, contamination, and the total number of markers present in that taxa out of the NM126 dataset.



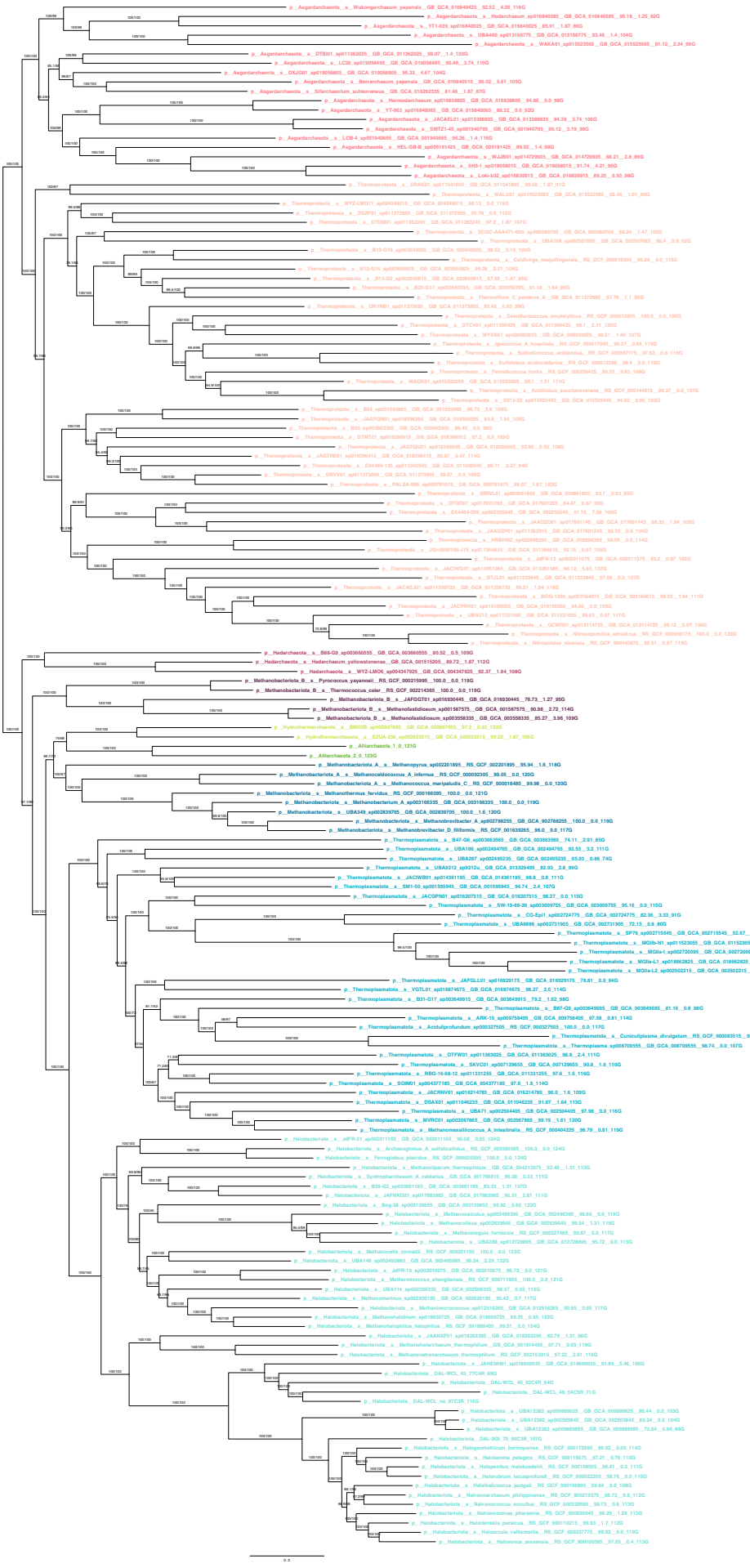
Supplementary Figure 14 | Maximum likelihood phylogeny of NM126 Nanohaloarchaeota only tree. The ML tree was reconstructed using the LG+C60+Γ4 model. SH-aLRT/ultrafast bootstrap support is shown above each branch via IQ-TREE implementation. Each color represents one of the 19 archaeal phyla, according to GTDB r207. The scale bar represents the estimated number of substitutions per site. Each tip label contains the GTDB phylum, species, accession number, completeness, contamination, and the total number of markers present in that taxa out of the NM126 dataset.



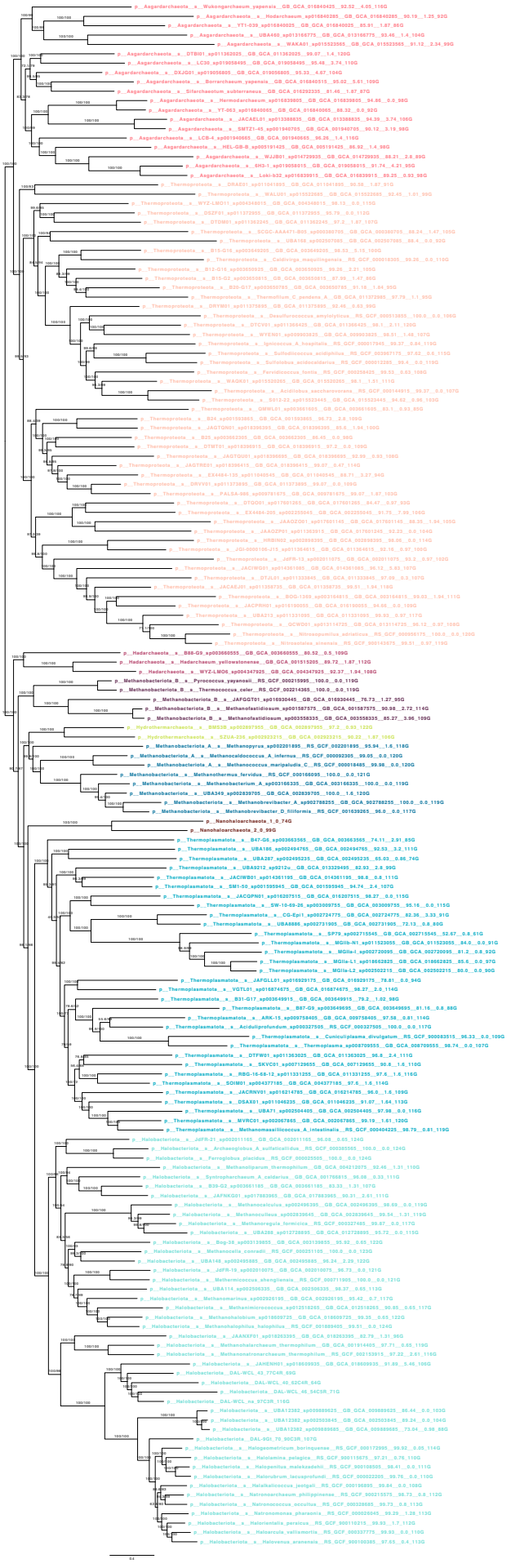
Supplementary Figure 15 | Maximum likelihood phylogeny of NM126 Nanoarchaeota only chimera tree. The ML tree was reconstructed using the LG+C60+Γ4 model. SH-aLRT/ultrafast bootstrap support is shown above each branch via IQ-TREE implementation. Each color represents one of the 19 archaeal phyla, according to GTDB r207. The scale bar represents the estimated number of substitutions per site. Each tip label contains the GTDB phylum, species, accession number, completeness, contamination, and the total number of markers present in that taxa out of the NM126 dataset.



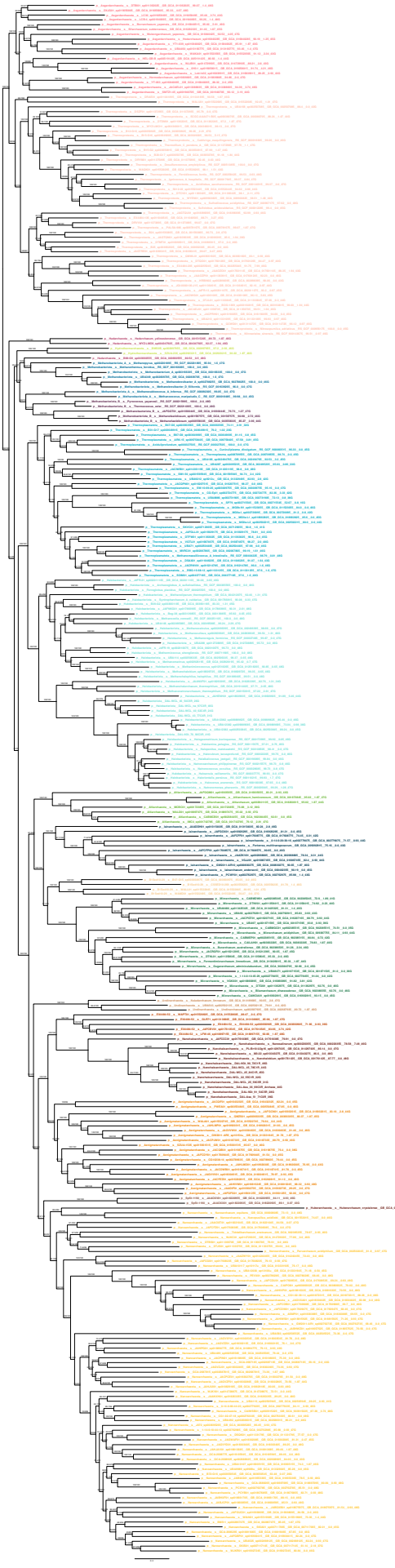
Supplementary Figure 16 | Maximum likelihood phylogeny of NM126 Micrarchaeota only chimera tree. The ML tree was reconstructed using the LG+C60+Γ4 model. SH-aLRT/ultrafast bootstrap support is shown above each branch via IQ-TREE implementation. Each color represents one of the 19 archaeal phyla, according to GTDB r207. The scale bar represents the estimated number of substitutions per site. Each tip label contains the GTDB phylum, species, accession number, completeness, contamination, and the total number of markers present in that taxa out of the NM126 dataset.



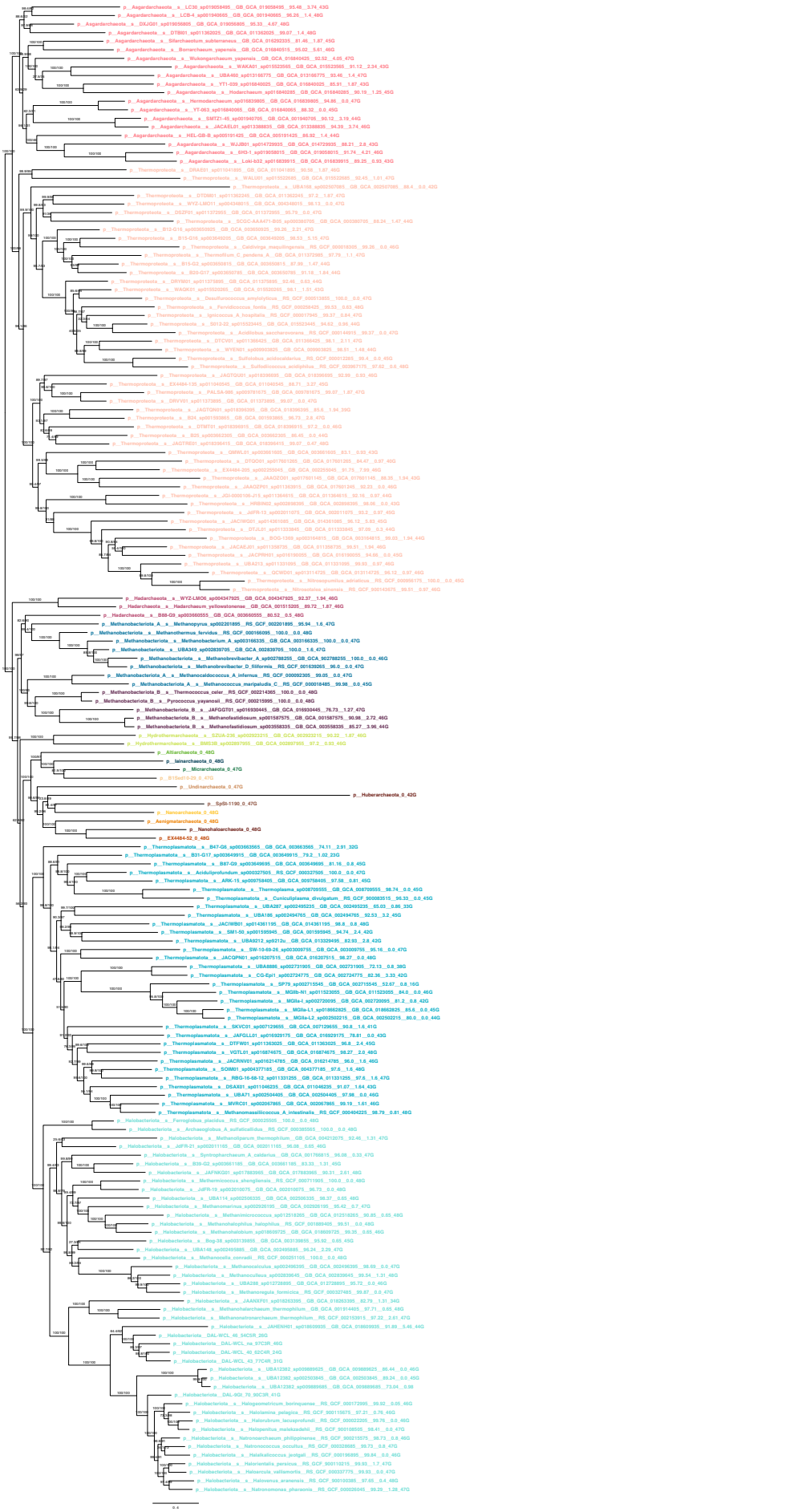
Supplementary Figure 17 | Maximum likelihood phylogeny of NM126 Altiarchaeota only chimera tree. The ML tree was reconstructed using the LG+C60+Γ4 model. SH-aLRT/ultrafast bootstrap support is shown above each branch via IQ-TREE implementation. Each color represents one of the 19 archaeal phyla, according to GTDB r207. The scale bar represents the estimated number of substitutions per site. Each tip label contains the GTDB phylum, species, accession number, completeness, contamination, and the total number of markers present in that taxa out of the NM126 dataset.



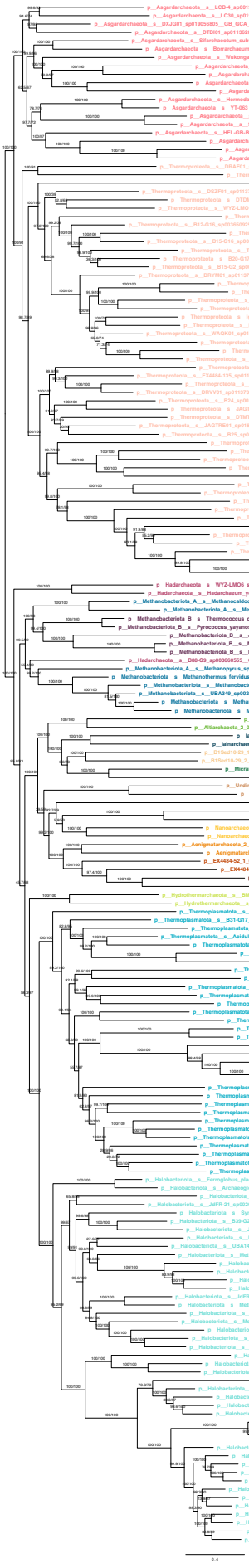
Supplementary Figure 18 | Maximum likelihood phylogeny of NM126 Nanohaloarchaeota only chimera tree. The ML tree was reconstructed using the LG+C60+Γ4 model. SH-aLRT/ultrafast bootstrap support is shown above each branch via IQ-TREE implementation. Each color represents one of the 19 archaeal phyla, according to GTDB r207. The scale bar represents the estimated number of substitutions per site. Each tip label contains the GTDB phylum, species, accession number, completeness, contamination, and the total number of markers present in that taxa out of the NM126 dataset.



Supplementary Figure 19 | Maximum likelihood phylogeny of 321 archaeal taxa based on the RP48 dataset. The ML tree was reconstructed using the LG+C60+Γ4 model. SH-aLRT/ultrafast bootstrap support is shown above each branch via IQ-TREE implementation. Each color represents one of the 19 archaeal phyla, according to GTDB r207. The scale bar represents the estimated number of substitutions per site. Each tip label contains the GTDB phylum, species, accession number, completeness, contamination, and the total number of markers present in that taxa out of the RP48 dataset.

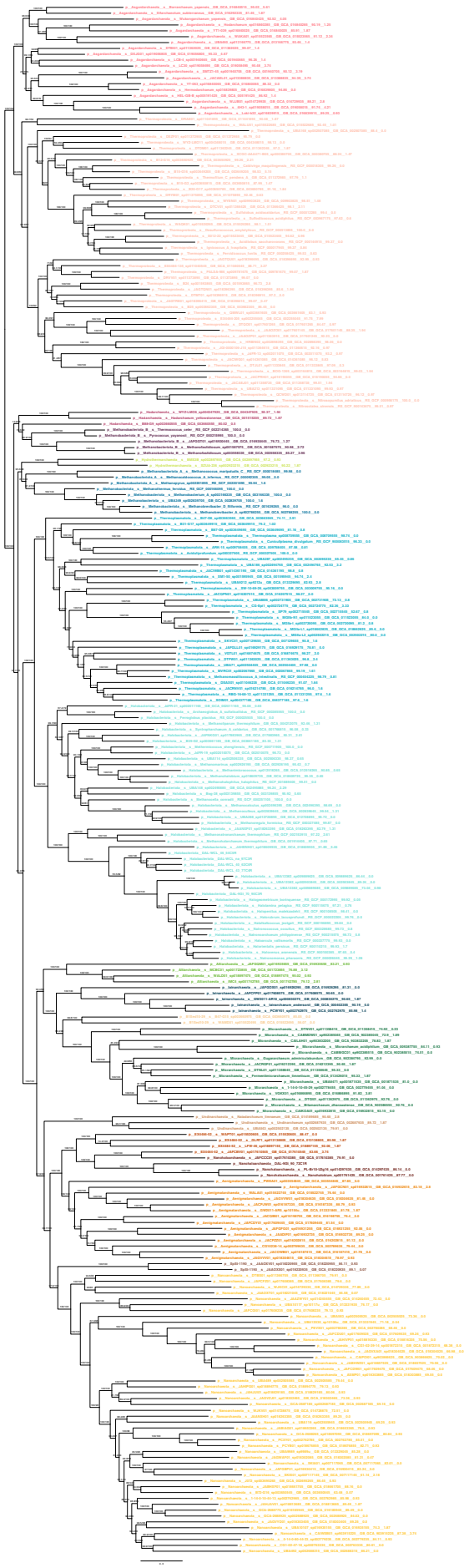


Supplementary Figure 20 | Maximum likelihood phylogeny of RP48 one-chimera tree. The ML tree was reconstructed using the LG+C60+Γ4 model. SH-aLRT/ultrafast bootstrap support is shown above each branch via IQ-TREE implementation. Each color represents one of the 19 archaeal phyla, according to GTDB r207. The scale bar represents the estimated number of substitutions per site. Each tip label contains the GTDB phyla, species, accession number, completeness, contamination, and the total number of markers present in that taxa out of the NM126 dataset. The chimera taxa are indicated by the DPANN GTDB phylum, followed by a _0, and the total number of markers per chimera.

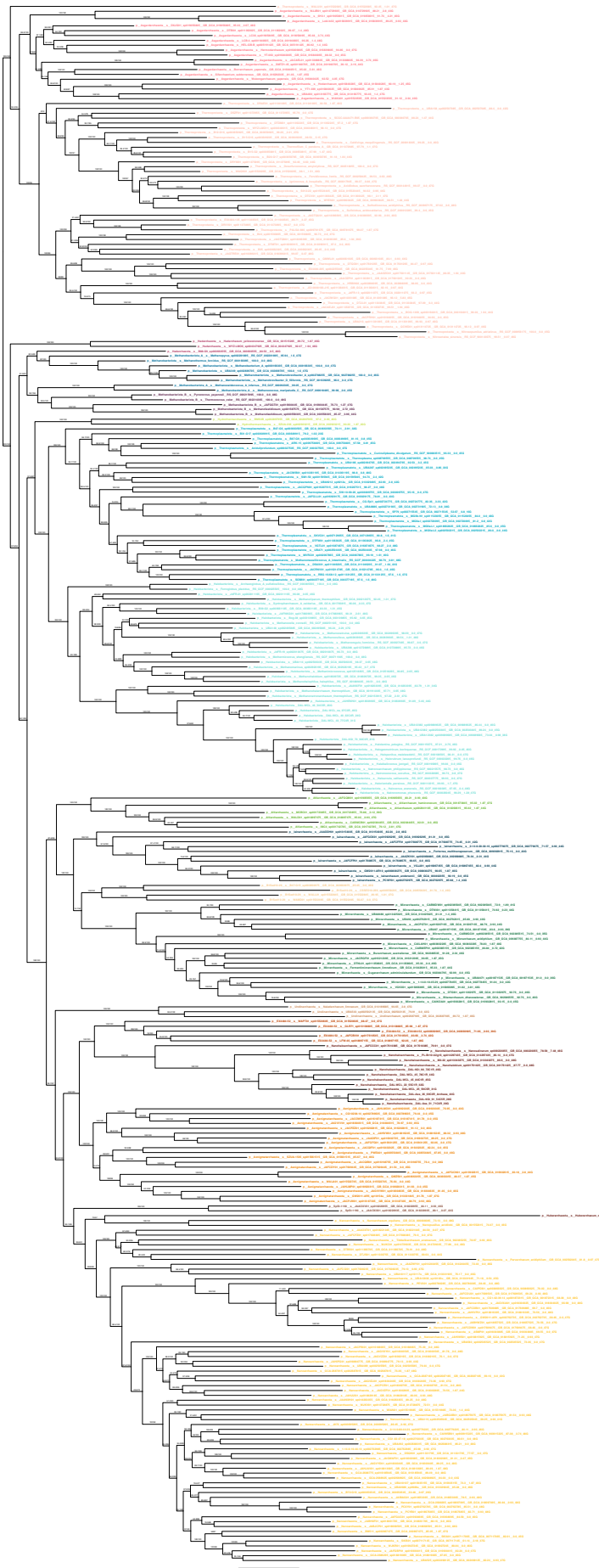


Supplementary Figure 21 | Maximum likelihood phylogeny of RP48 two-chimera tree. The ML tree was reconstructed using the LG+C60+Γ4 model. SH-aLRT/ultrafast bootstrap support is shown above each branch via IQ-TREE implementation. Each color represents one of the 19 archaeal phyla, according to GTDB r207. The scale bar represents the estimated number of substitutions per site. Each tip label contains the GTDB phyla, species, accession number, completeness, contamination, and the total number of markers present in that taxa out of the NM126 dataset. The chimera taxa are indicated by the DPANN GTDB phylum, followed by a _1_0 or _2_0, followed by the total number of markers per chimera.

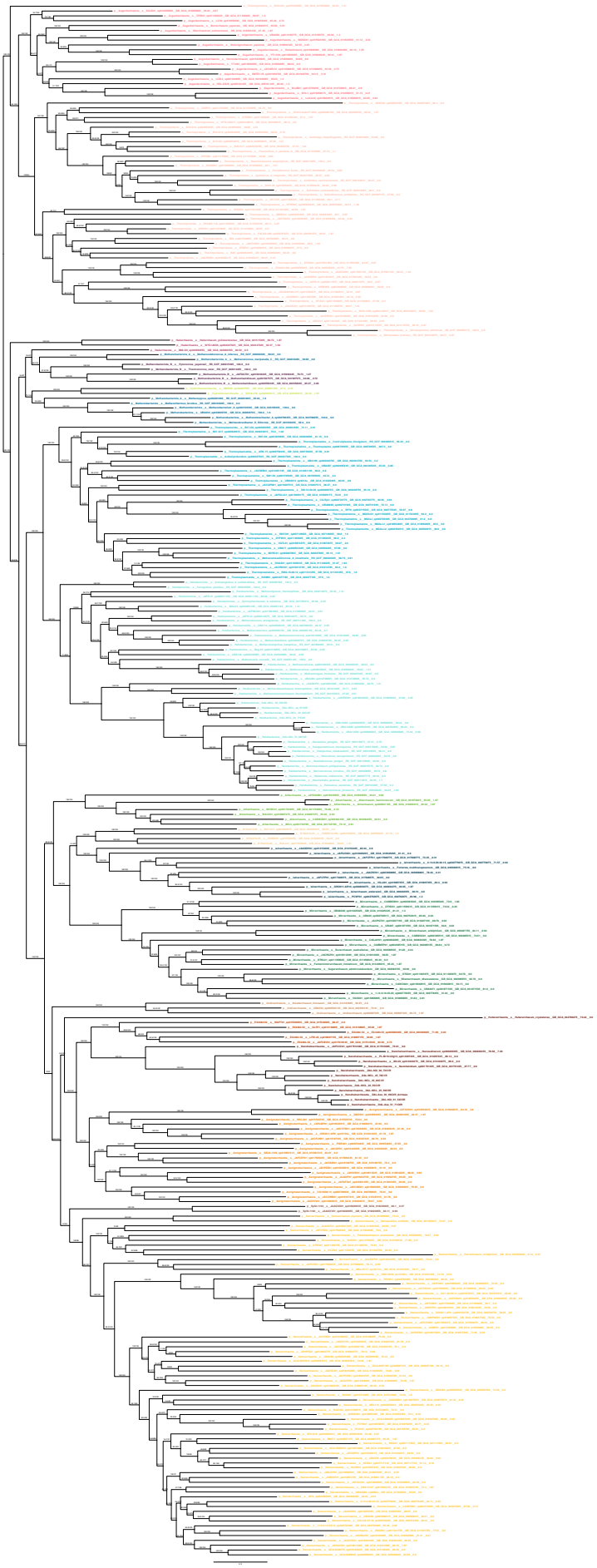
Supplementary Figure 22 | Maximum likelihood phylogeny of RP48 one-chimera tree with no Huberarchaeota. The ML tree was reconstructed using the LG+C60+Γ4 model. SH-aLRT/ultrafast bootstrap support is shown above each branch via IQ-TREE implementation. Each color represents one of the 19 archaeal phyla, according to GTDB r207. The scale bar represents the estimated number of substitutions per site. Each tip label contains the GTDB phyla, species, accession number, completeness, contamination, and the total number of markers present in that taxa out of the RP48 dataset. The chimera taxa are indicated by the DPANN GTDB phylum, followed by a _0, followed by the total number of markers per chimera.



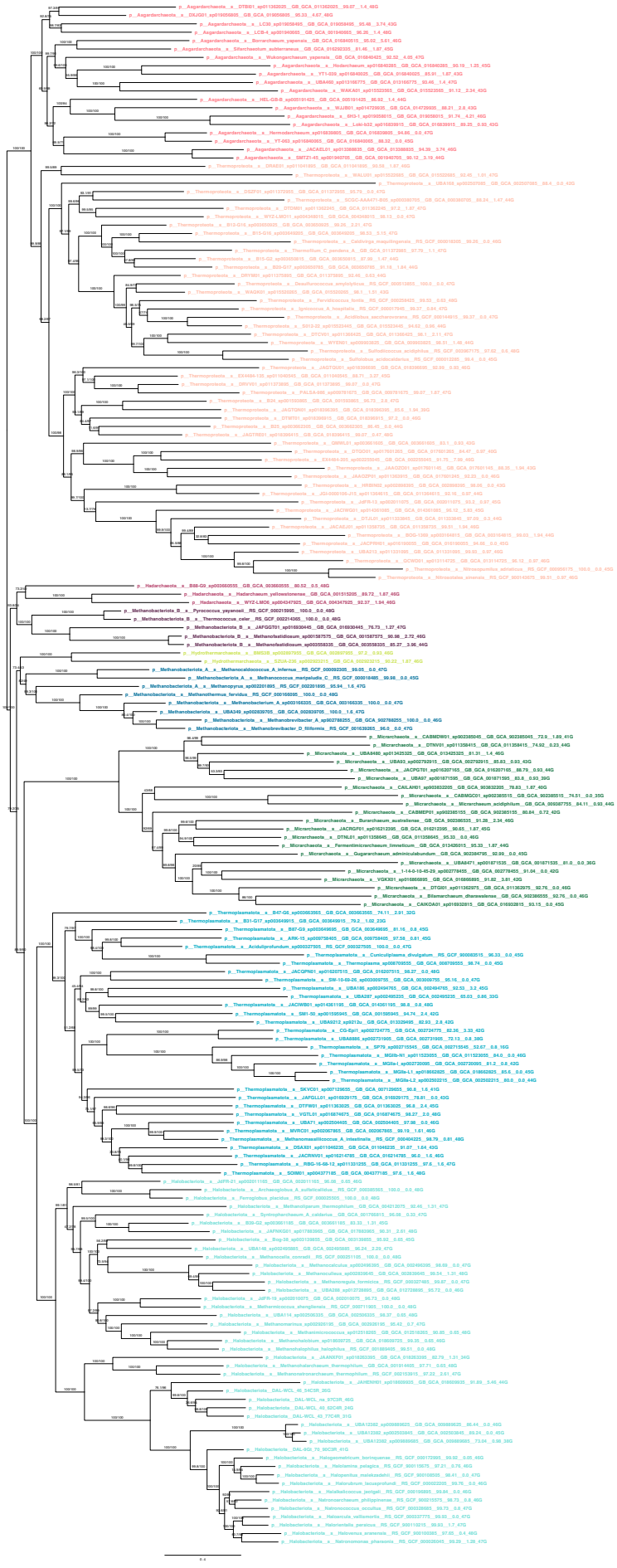
Supplementary Figure 23 | Maximum likelihood phylogeny of RP48 PCA tree with “most biased” DPANN taxa removed. The ML tree was reconstructed using the LG+C60+Γ4 model. SH-aLRT/ ultrafast bootstrap support is shown above each branch via IQ-TREE implementation. Each color represents one of the 19 archaeal phyla, according to GTDB r207. The scale bar represents the estimated number of substitutions per site. Each tip label contains the GTDB phylum, species, accession number, completeness, contamination, and the total number of markers present in that taxa out of the RP48 dataset.



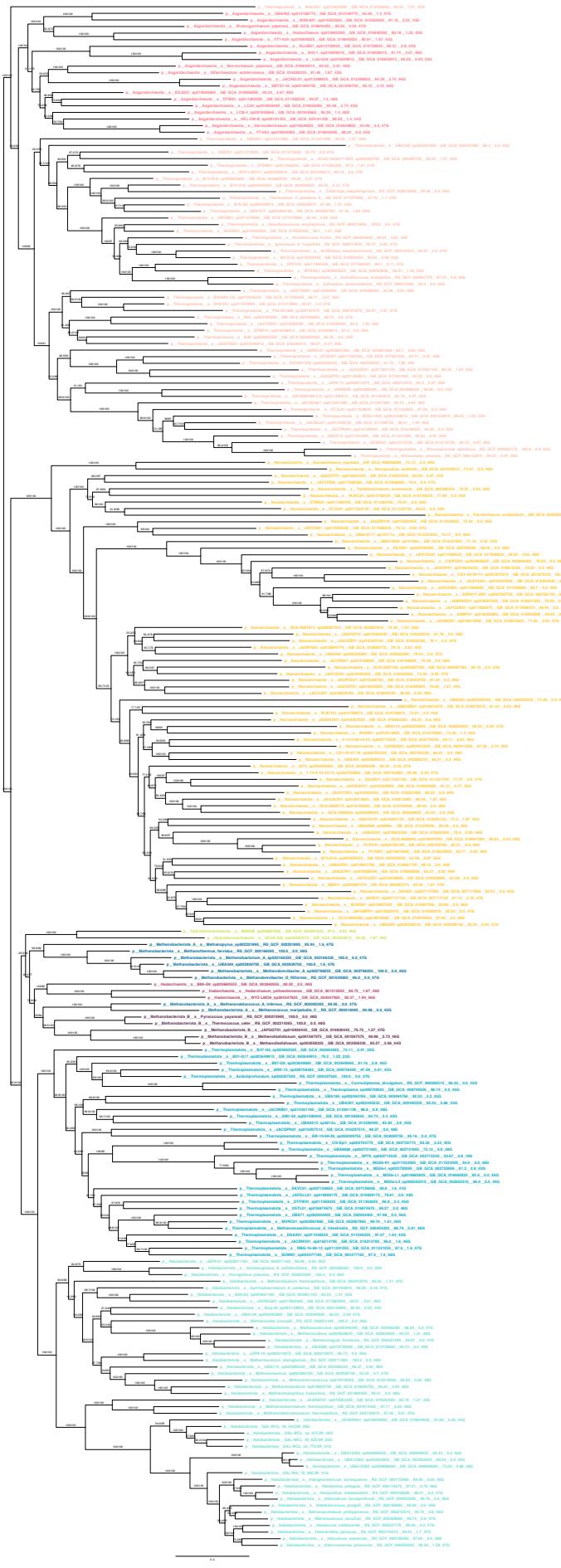
Supplementary Figure 24 | Maximum likelihood phylogeny of RP48 with 15% most GARP/FIMNKY-biased sites removed binned DPANN versus non-DPANN lineages. The ML tree was reconstructed using the LG+C60+Γ4 model. SH-aLRT/ultrafast bootstrap support is shown above each branch via IQ-TREE implementation. Each color represents one of the 19 archaeal phyla, according to GTDB r207. The scale bar represents the estimated number of substitutions per site. Each tip label contains the GTDB phylum, species, accession number, completeness, contamination, and the total number of markers present in that taxa out of the RP48 dataset.



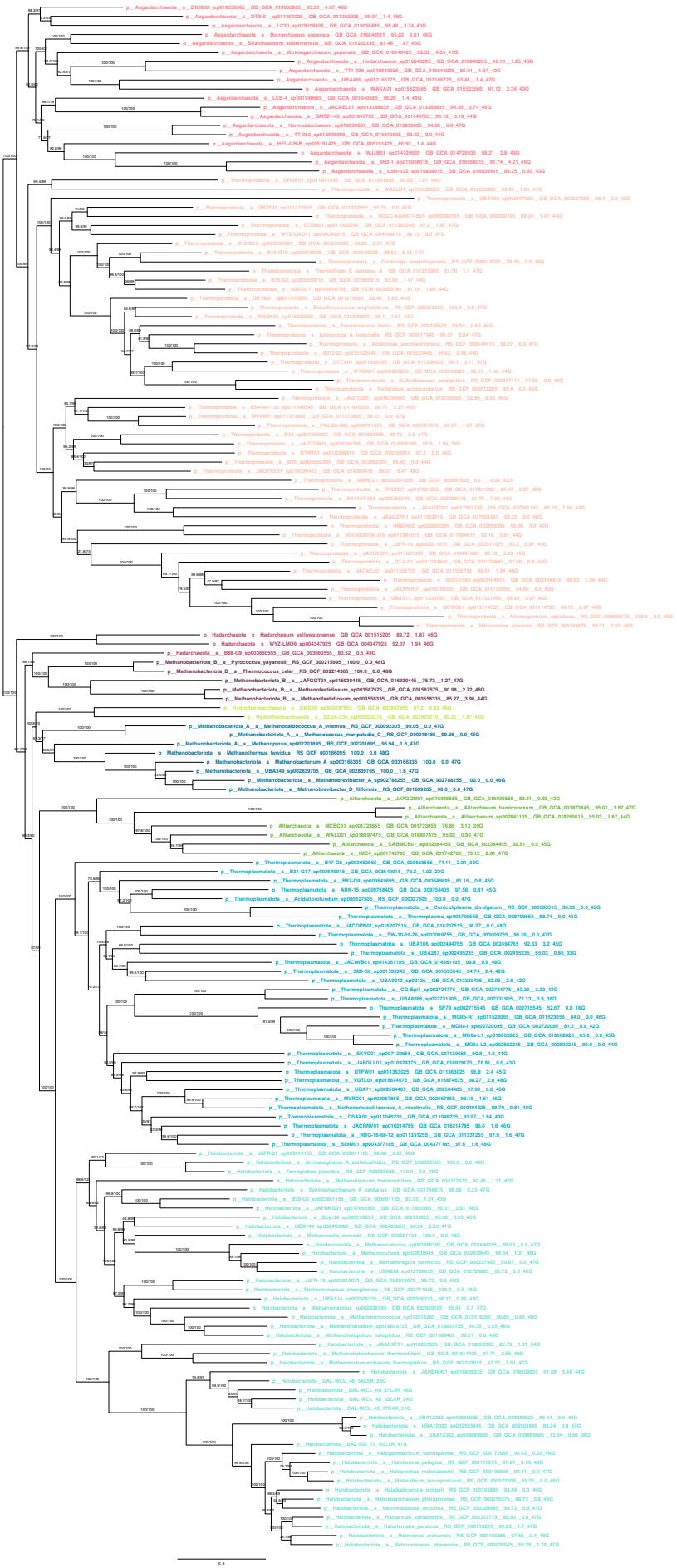
Supplementary Figure 25 | Maximum likelihood phylogeny of RP48 with 15% most GARP/FIMNKY-biased sites removed binned FIMNKY-enriched taxa versus GARP-enriched taxa. The ML tree was reconstructed using the LG+C60+Γ4 model. SH-aLRT/ultrafast bootstrap support is shown above each branch via IQ-TREE implementation. Each color represents one of the 19 archaeal phyla, according to GTDB r207. The scale bar represents the estimated number of substitutions per site. Each tip label contains the GTDB phylum, species, accession number, completeness, contamination, and the total number of markers present in that taxa out of the RP48 dataset.



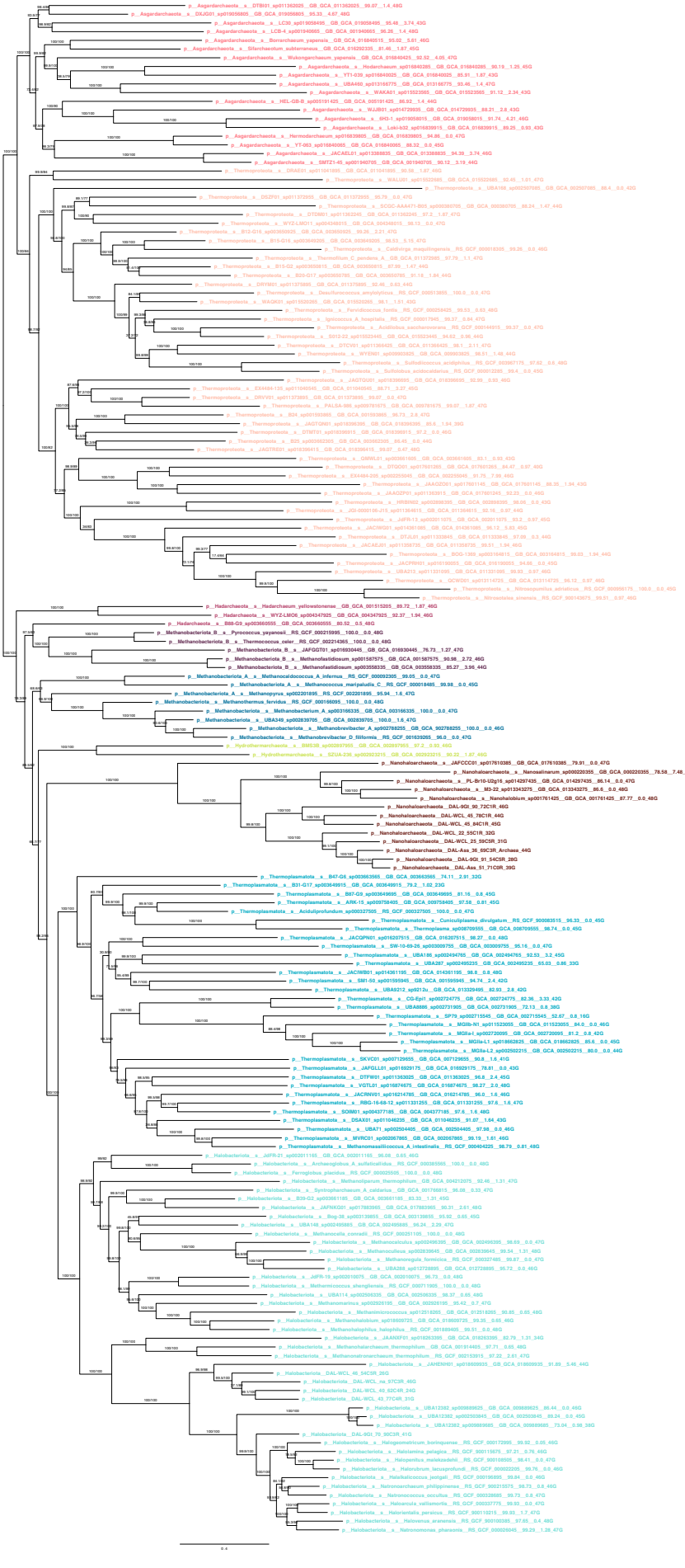
Supplementary Figure 26 | Maximum likelihood phylogeny of RP48 Micrarchaeota only tree. The ML tree was reconstructed using the LG+C60+Γ4 model. SH-aLRT/ultrafast bootstrap support is shown above each branch via IQ-TREE implementation. Each color represents one of the 19 archaeal phyla, according to GTDB r207. The scale bar represents the estimated number of substitutions per site. Each tip label contains the GTDB phylum, species, accession number, completeness, contamination, and the total number of markers present in that taxa out of the RP48 dataset.



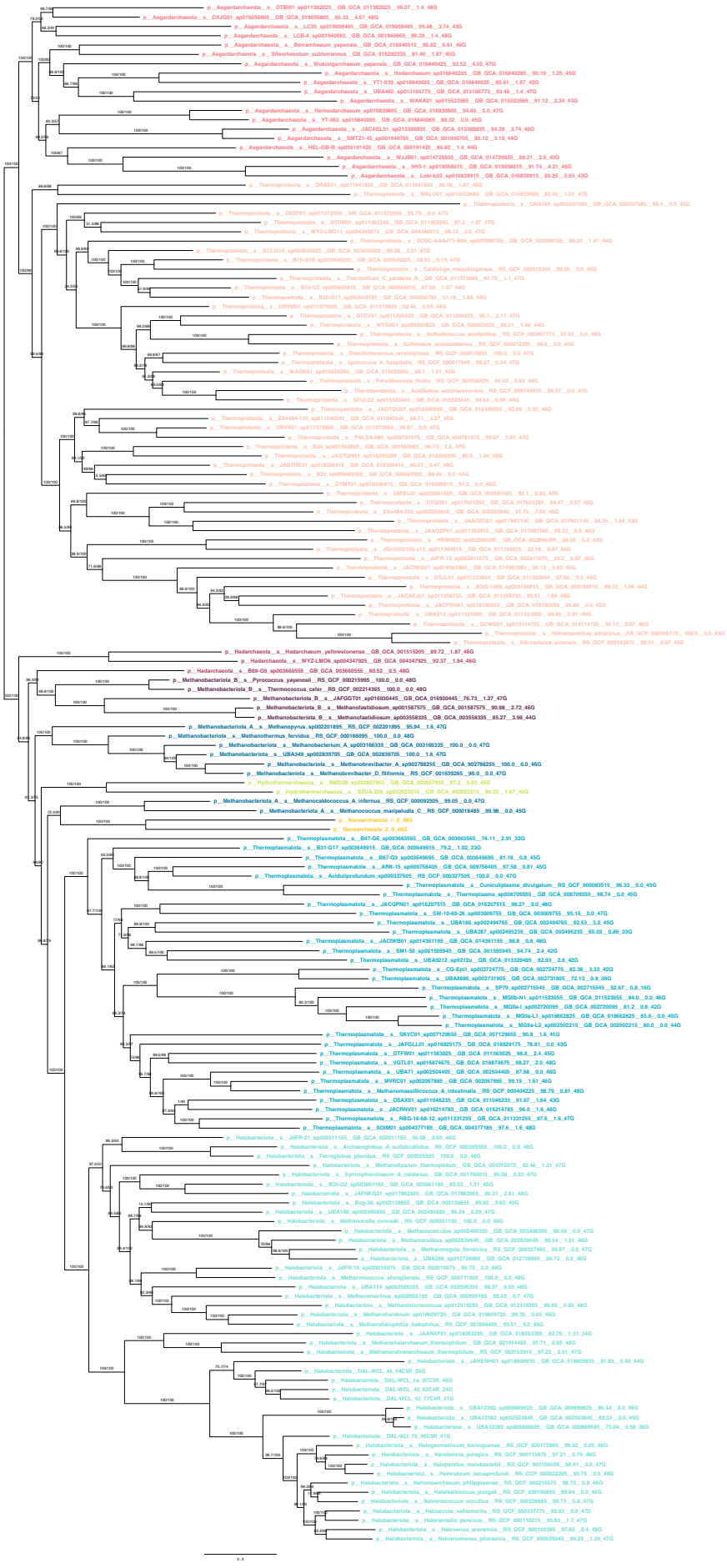
Supplementary Figure 27 | Maximum likelihood phylogeny of RP48 Nanoarchaeota only tree. The ML tree was reconstructed using the LG+C60+Γ4 model. SH-aLRT/ultrafast bootstrap support is shown above each branch via IQ-TREE implementation. Each color represents one of the 19 archaeal phyla, according to GTDB r207. The scale bar represents the estimated number of substitutions per site. Each tip label contains the GTDB phylum, species, accession number, completeness, contamination, and the total number of markers present in that taxa out of the RP48 dataset.



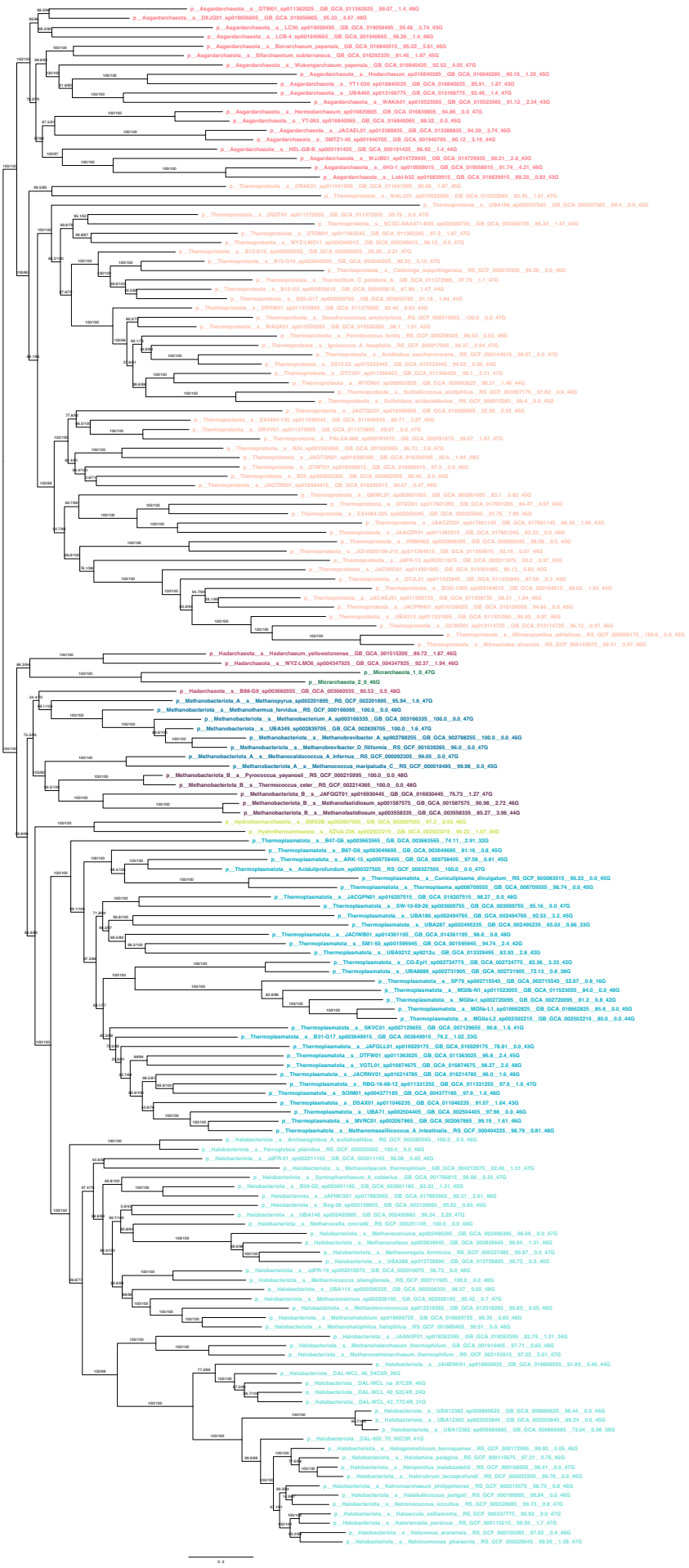
Supplementary Figure 28 | Maximum likelihood phylogeny of RP48 Altiarchaeota only tree. The ML tree was reconstructed using the LG+C60+Γ4 model. SH-aLRT/ultrafast bootstrap support is shown above each branch via IQ-TREE implementation. Each color represents one of the 19 archaeal phyla, according to GTDB r207. The scale bar represents the estimated number of substitutions per site. Each tip label contains the GTDB phylum, species, accession number, completeness, contamination, and the total number of markers present in that taxa out of the RP48 dataset.



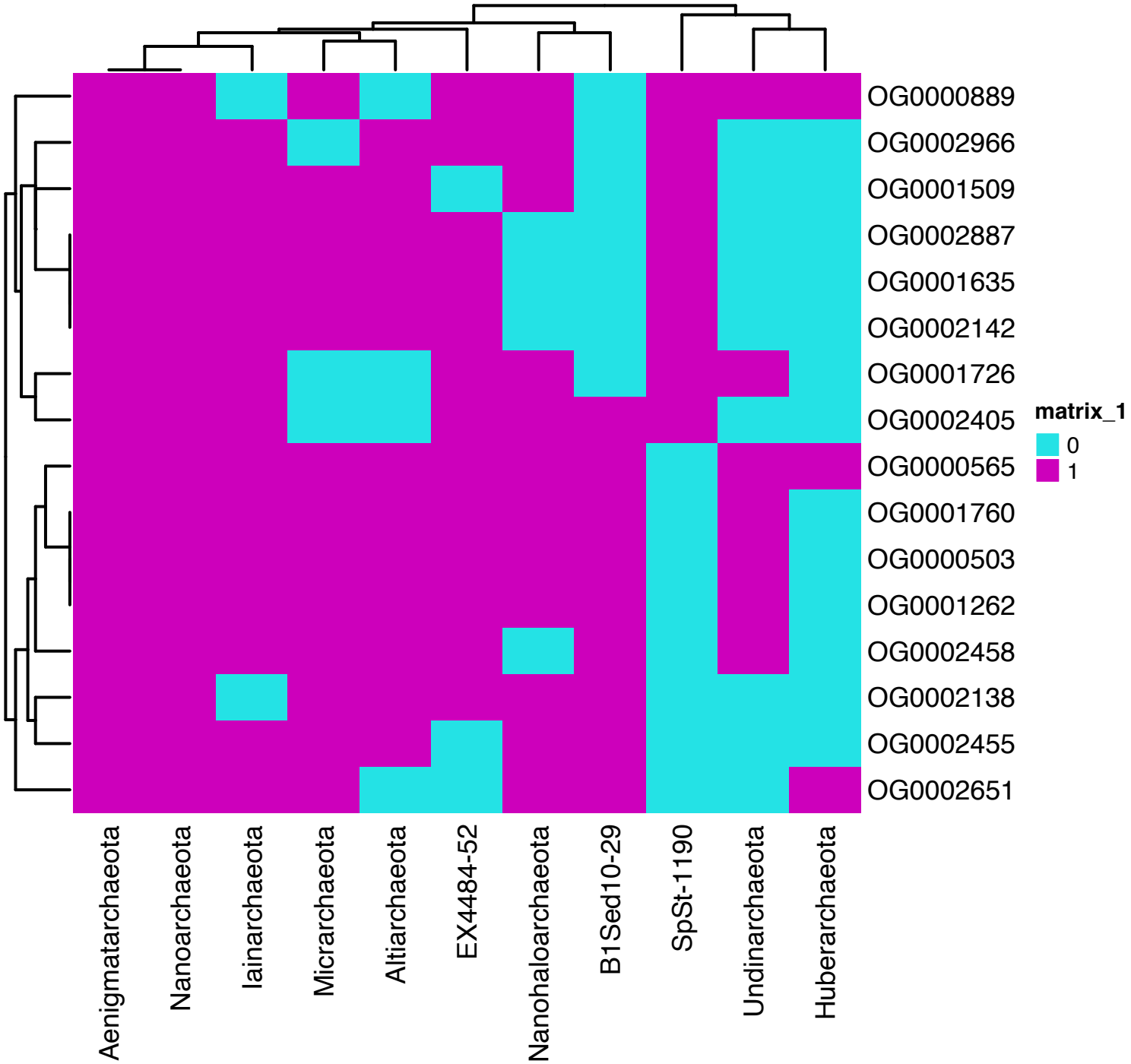
Supplementary Figure 29 | Maximum likelihood phylogeny of RP48 Nanohaloarchaeota only tree. The ML tree was reconstructed using the LG+C60+Γ4 model. SH-aLRT/ultrafast bootstrap support is shown above each branch via IQ-TREE implementation. Each color represents one of the 19 archaeal phyla, according to GTDB r207. The scale bar represents the estimated number of substitutions per site. Each tip label contains the GTDB phylum, species, accession number, completeness, contamination, and the total number of markers present in that taxa out of the RP48 dataset.



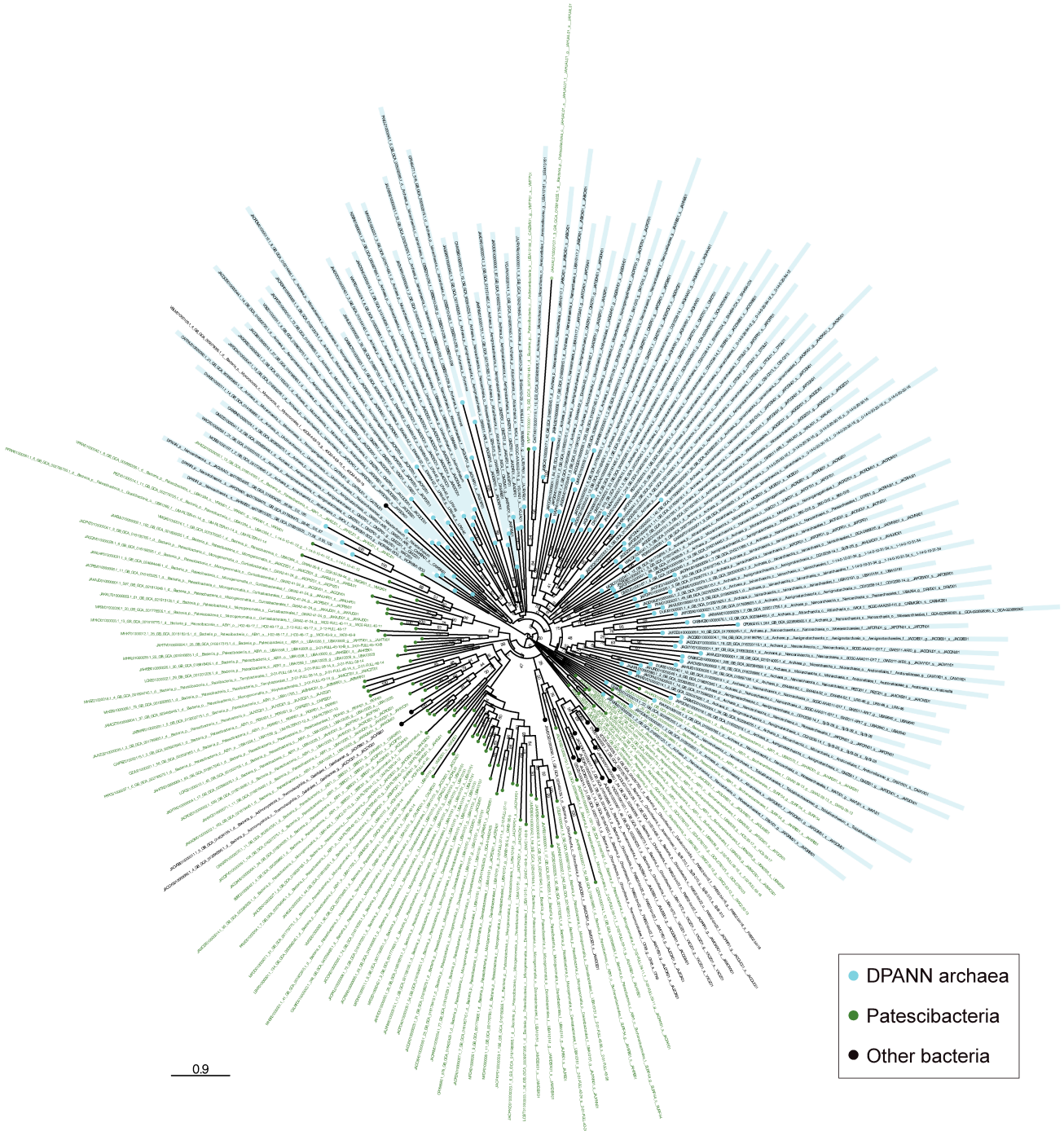
Supplementary Figure 30 | Maximum likelihood phylogeny of RP48 Nanoarchaeota only chimera tree. The ML tree was reconstructed using the LG+C60+Γ4 model. SH-aLRT/ultrafast bootstrap support is shown above each branch via IQ-TREE implementation. Each color represents one of the 19 archaeal phyla, according to GTDB r207. The scale bar represents the estimated number of substitutions per site. Each tip label contains the GTDB phylum, species, accession number, completeness, contamination, and the total number of markers present in that taxa out of the RP48 dataset.



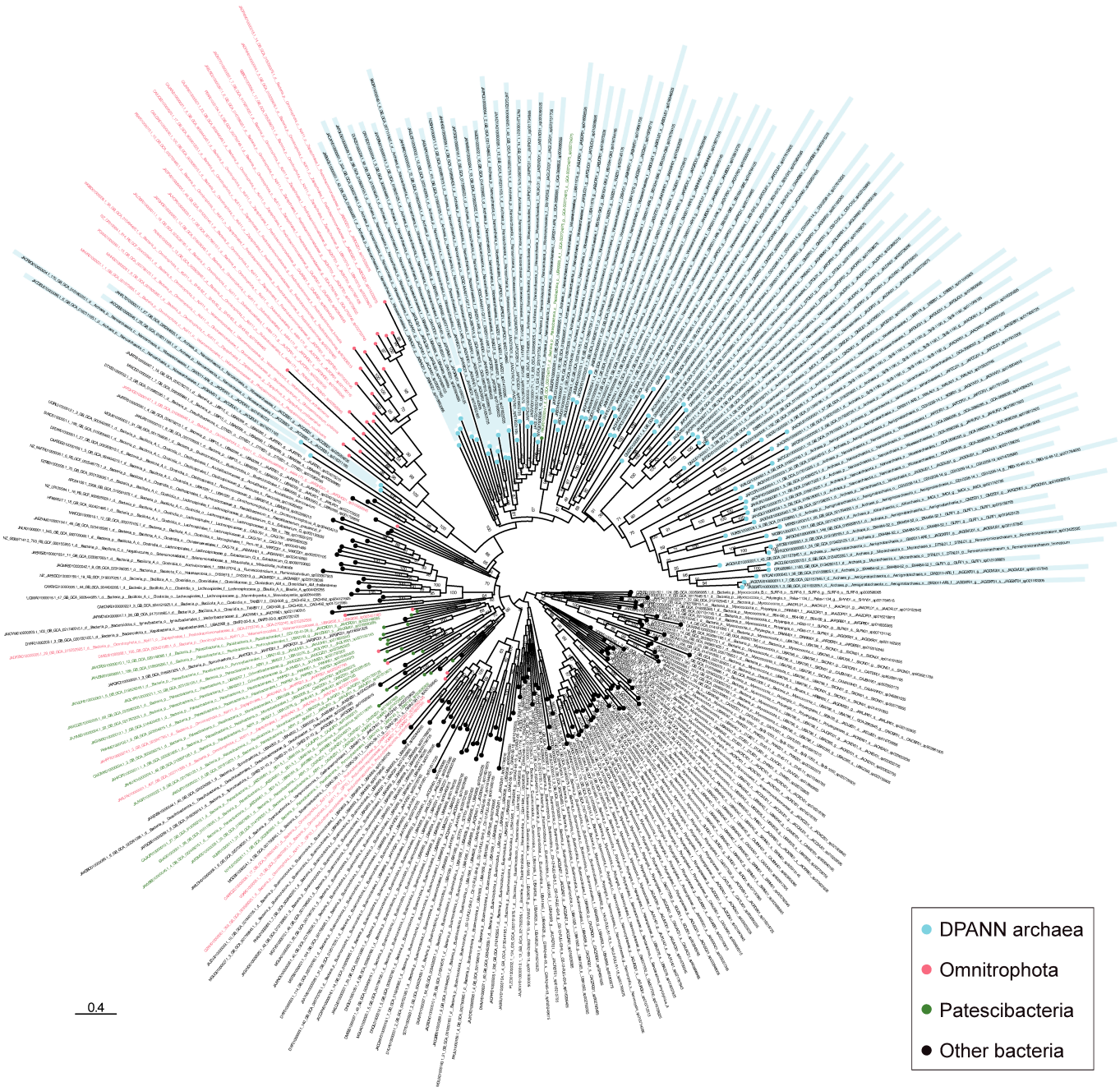
Supplementary Figure 31 | Maximum likelihood phylogeny of RP48 Micrarchaeota only chimera tree. The ML tree was reconstructed using the LG+C60+Γ4 model. SH-aLRT/ultrafast bootstrap support is shown above each branch via IQ-TREE implementation. Each color represents one of the 19 archaeal phyla, according to GTDB r207. The scale bar represents the estimated number of substitutions per site. Each tip label contains the GTDB phylum, species, accession number, completeness, contamination, and the total number of markers present in that taxa out of the RP48 dataset.



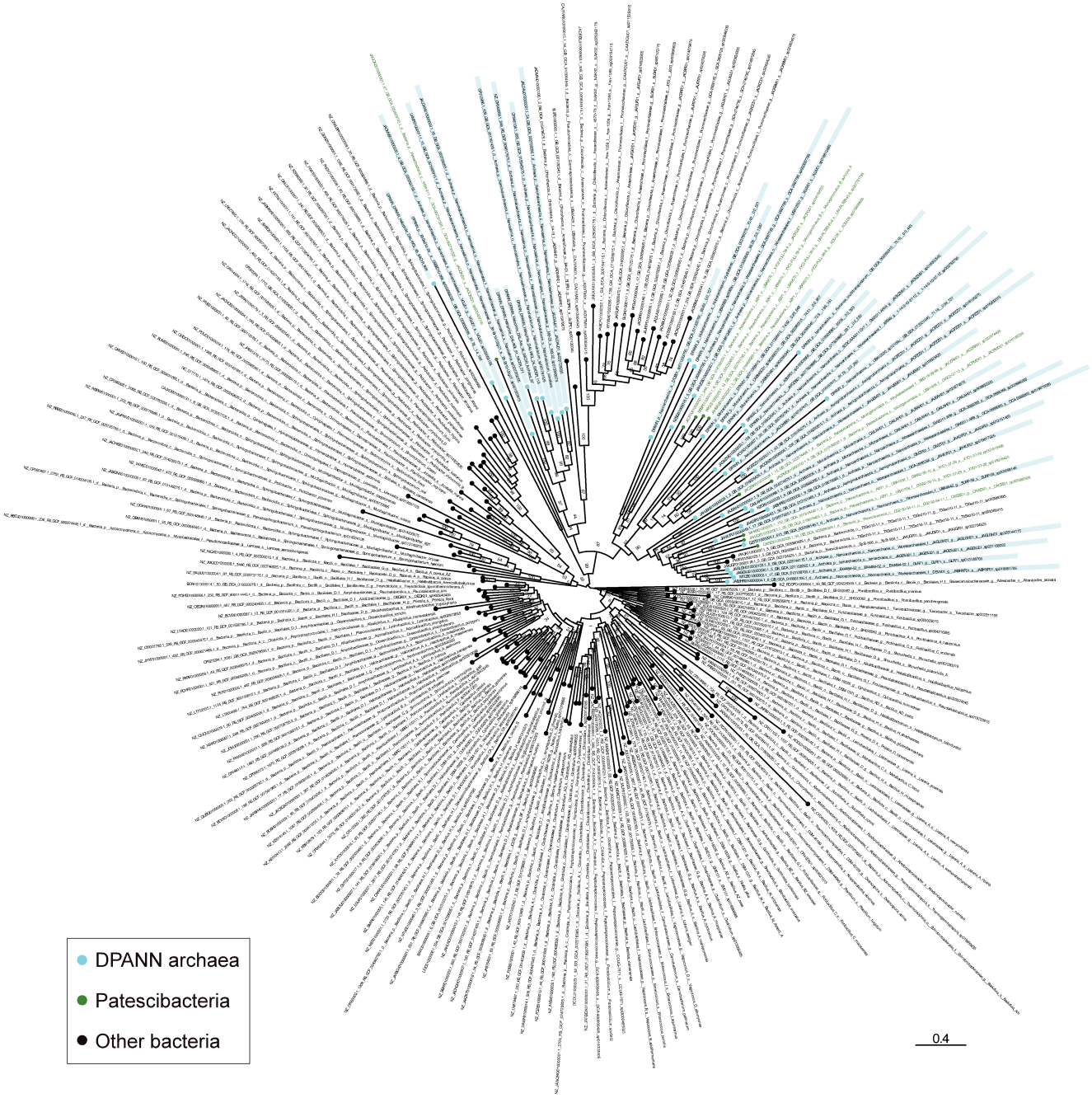
Supplementary Figure 32 | Presence and absence heatmap of the 14 OGs present in at least 7 DPANN phyla. The x-axis contains the OG identifier, and the y-axis contains the 11 DPANN phyla. The fuchsia squares indicate the presence, and the absence is cyan. The annotations of these 14 OGs can be found in Supplementary Data 3.



Supplementary Figure 33 | Maximum likelihood phylogeny of OG2458 tree. The ML tree was reconstructed using the LG+C20+F+Γ4 model. Ultrafast bootstrap support is shown above each branch via IQ-TREE implementation. The scale bar represents the estimated number of substitutions per site.



Supplementary Figure 34 | Maximum likelihood phylogeny of OG2142 tree. The ML tree was reconstructed using the LG+C20+F+Γ4 model. Ultrafast bootstrap support is shown above each branch via IQ-TREE implementation. The scale bar represents the estimated number of substitutions per site.



Supplementary Figure 35 | Maximum likelihood phylogeny of OG2651 tree. The ML tree was reconstructed using the LG+C20+F+Γ4 model. Ultrafast bootstrap support is shown above each branch via IQ-TREE implementation. The scale bar represents the estimated number of substitutions per site.