Expanded phylogeny of extremely halophilic archaea shows multiple independent adaptations to hypersaline environments

Brittany Baker, Ana Gutiérrez-Preciado, Álvaro Rodríguez del Río, Charley Mccarthy, Purificación López-García, Jaime Huerta-Cepas, Edward Susko, Andrew Roger, Laura Eme, David Moreira

Expanded phylogeny of extremely halophilic archaea shows multiple independent adaptations 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

Extremely halophilic archaea (Haloarchaea, Nanohaloarchaeota, Methanonatronarchaeia, and Halarchaeoplasmatales) thrive in saturating salt concentrations where they must maintain osmotic equilibrium with their environment. The evolutionary history of adaptations enabling salt tolerance remains poorly understood, in particular because the phylogeny of several lineages is conflicting. 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 uncultured lineages, Afararchaeaceae and Asbonarchaeaceae, which break the long branches at the base of Haloarchaea and Nanohaloarchaeota, respectively. We obtained 13 metagenome-assembled genomes (MAGs) of these archaea from metagenomes of hypersaline aquatic systems of the Danakil Depression (Ethiopia). Our phylogenomic analyses including these taxa 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 extremely halophilic lineages.

Introduction

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 previously undescribed 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

Two previously undescribed groups of 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, we identified 13 belonging to two lineages of extreme halophiles phylogenetically distant from already known groups, plus one additional MAG placed deep in the Haloarchaea (Fig. 1a,c, Supplementary Data 1).

The first group – a 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 and Supplementary Fig. 2). 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[22] indicates that Afararchaeaceae are likely heterotrophic aerobes that utilize branched-chain amino acids as a carbon source, similar to many known Haloarchaea[23] (Fig. 1b, Supplementary Data 3). They are probably mobile, possessing all genes for the archaeal flagellum (archaellum)[24] and a chemotaxis operon. Additionally, afararchaeal MAGs encode a single type-II sensory rhodopsin for phototaxis[25], but lack bacteriorhodopsin genes, suggesting that these archaea do not use light as an additional energy source like many Haloarchaea[26]. 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 Supplementary Fig. 3) 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[27]. 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%[28,29]. 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[28] (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[30–32] (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)[30]. 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 an independent family within the Nanohaloarchaeota.

Undescribed gene families in Afararchaeaceae and Asbonarchaeaceae

We identified gene families previously undescribed 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[33], Pfam[34], and EggNOG[35]), revealing a significant number of potentially unique 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[36], 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 genes encode proteins with an acidic pH isoelectric point, aligning with adaptation to hypersaline environments[37] (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 protein in

Afararchaeaceae next to a mechanosensitive ion channel (Fig. 1e), suggesting a potential role in osmotic regulation[38].

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[18,39]. 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[18,40]. To address these issues and accurately determine the positions of extreme halophilic archaea, we conducted a comprehensive phylogenomic analyses using a dataset of 136 non-ribosomal 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[41,42]. 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[43] 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[44].

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[10,45] or removed the fastest-evolving sites[8,10,45]. 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[44] 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 (Supplementary Fig. 13). 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 (Supplementary Fig. 13).

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. 5), 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. 14 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. 15). 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. 16 and 17).

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[9,46]. However, when we removed 15% of the highest D+E/I+K ratio sites from this dataset, Nanohaloarchaeota branched deeper (Extended Data Fig. 6), 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.

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 Fig. 7, and Supplementary Fig. 18). Haloarchaea, with some of the largest genome sizes among archaea[47], also experienced significant gene originations and duplications during their early evolution. This expansion involved genes encoding 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. 19-26, Extended Data Fig. 8a), and molecular chaperones like GrpE (Supplementary Fig. 27), which prevents protein aggregation during response to hyperosmotic stress[48]. Amino acid transporters, vital for species of these groups thriving on amino acids[23], also exhibited duplications (Extended Data Fig. 7). 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. 7). In Nanosalinaceae and Asbonarchaeaceae, gene transfer was dominant but less pronounced due to constraints in these small-sized archaea to maintain compact genomes[49]. 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[50,51]. 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. 7). Notably, we observed multiple copies of aerobic-type carbon monoxide dehydrogenase, found in other microorganisms adapted to the same nutrient-poor environments[52] (Supplementary Fig. 28).

Massive HGT from bacteria has likely played a significant role in the evolution of Haloarchaea, although its extent and timing are still debated[53–56]. 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[57], was acquired through HGT (Extended Data Fig. 8b). 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[57], which Methanonatronarchaeia acquired from bacteria (Supplementary Fig. 29).

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. 30-36).

HGT of organic molecule transporters is also observed, such as a transporter of Krebs cycle intermediates shared by Haloarchaea and Asbonarchaeaceae (Supplementary Fig. 37). We also identified genes of bacterial origin encoding various transporters subsequently transferred between different halophilic archaeal groups. These include genes encoding an AmiS/UreI urea transporter transferred between Haloarchaea and Nanohaloarchaea and a TauE/SafE sulfite exporter transferred between Haloarchaea and Methanonatronarchaeia (Supplementary Figs. 38-39), consistent with previous reports of inter-domain HGT followed by intra-domain HGT[58].

Discussion

Our study yields a robust archaeal phylogeny, including two halophilic lineages previously unknown, 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 those encoding 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 still undescribed halophilic archaea.

Taxonomic descriptions

Taxon names have been described under the SeqCode[59] 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 datasets[20-21]. 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 v2.0.3[60] (Supplementary Fig. 40). 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 divergent groups of extremely halophilic archaea based on their position compared to other known halophilic archaea. These 14 MAGs were taxonomically classified using GTDB-Tk[27] (version

2.3.0, r207; April 1st, 2022) and assigned to families within three GTDB orders: four MAGs were assigned to a family previously undescribed belonging to the order 'JAHENH01', which we have named Afararchaeaceae; nine MAGs were assigned to another family previously undescribed belonging to the order Nanosalinales, which we have named Asbonarchaeaceae; and one MAG belonged to a third family in the order Halobacteriales, 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 v1.34[61]. 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[62]. 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[63] and subjected to Pfam[34] and COG[64] functional annotations inside the Anvi'o v5 pipeline[65]. Genes were also annotated with KofamKOALA[66] and eggNOG-mapper v2.1.5[35]. 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.

Detection of undescribed protein families

We computed family clusters of the proteins predicted for the MAGs of the archaeal families Afararchaeaceae and Asbonarchaeaceae using Mmseqs2 v3.0[67] 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[35] (hits with an e-value <1e-3 were considered as significant) ii) the protein sequences encoded in the MAGs against PfamA domains using HMMER v3.3.2[68] (hits with an e-value <1e-5 were considered as significant), iii) the protein sequences encoded in the MAGs against PfamB domains using HMMER v3.3.2[68] (hits with an e-value < 1e-5 were considered as significant) and iv) the CDS sequences of the MAGs against RefSeq using Diamond BLASTx[69] ('sensitive' flag, hits with an e-value <1e-3 and query coverage >50% were considered as significant). We only considered as undescribed families those with no detectable homologs in these databases. To address the taxonomic breadth of these families, we used Diamond BLASTp v2.1.7[69] ('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)[70], the Global Microbial Gene Catalog (GMGC)[71], the Unified Human Gastrointestinal Genome collection (UHGG)[72], and the Ocean Microbiomics Database (OMD)[73]. 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 undescribed family set. We predicted signal peptides and transmembrane domains on the gene families using SignalP v6.0[74] and TMHMM v2.0[75]. 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 Afararchaeaceae and Asbonarchaeaceae). We reconstructed two phylogenomic datasets consisting of 48 ribosomal proteins (RP) and 136 non-ribosomal markers (NM) widely distributed in archaea (Supplementary Fig. 41). 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[63]. Next, sequences similar to the RP and NM proteins were identified using BLAST v2.10.0[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 v2.1.11[77] 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 phylogenetic 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[78] and trimmed with BMGE v1.12[79] (-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 v2.0.3[60] 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 IQ-TREE v2.0.3[60]. 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[80]. 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 v3.10.5 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 v3.4.2[81].

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 GFmix-DE/IK model by transforming the b parameter of the GFmix model[44] (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[44]. Branch length and alpha shape parameters for each tree tested were estimated using IQ-TREE v2.0.3[60] 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 (Supplementary Fig. 13).

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[82] accession number 1S72[83]) and the non-halophilic methanogen

Methanothermobacter thermautotrophicus (PDB[82] accession number 4ADX[84]). We located these positions on their respective structures using ChimeraX v1.7[85], 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[86] with Diamond BLAST v2.1.7 (--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 v7.450 --auto[78] with default settings and trimmed using trimAl v1.2[87] (-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 IQ-TREE v2.0.3[60]. 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 v0.4[88], 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[88]. ALE infers, for each gene family, duplications, losses, transfers, and originations events along a species tree[88]. 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 v6.8[89], and the ETE3 Toolkit v.3.1.2[90].

To detect possible genes of bacterial origin in halophilic archaea, we carried out BLAST v2.10.0[76] searches of the proteins considered by ALE as 'originations' in these archaea against the RefSeq[33] database. Proteins with similar sequences in bacteria were aligned using MAFFT v7.450 --auto[78] with default settings and trimmed using trimAl v1.2[87] (-automated1). Maximum likelihood trees were then reconstructed with IQ-TREE v2.0.3[60] 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). Additional data were obtained

from public databases, including GTDB (https://gtdb.ecogenomic.org/), Pfam (http://pfam.xfam.org/), COG (https://www.ncbi.nlm.nih.gov/research/cog), RefSeq (https://www.ncbi.nlm.nih.gov/refseq/), eggNOG (http://eggnog5.embl.de/#/app/home), the Genomic Catalog of Earth's Microbiomes (https://genome.jgi.doe.gov/portal/GEMs/GEMs.home.html), the Global Microbial Gene Catalog (https://gmgc.embl.de/), the Unified Human Gastrointestinal Genome collection (http://ftp.ebi.ac.uk/pub/databases/metagenomics/mgnify_genomes/), the Ocean Microbiomics Database (https://microbiomics.io/ocean/), and PDB (https://www.rcsb.org/).

Code availability

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

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.), Moore Foundation GBMF9739 (P.L.G.), and ANR DArchFolds ANR-22-CE02-0012-02 (D.M., P.L.G., and L.E.). 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 and A. Oren for his advice on taxonomic descriptions. 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 archaeal MAGs. A.R.R., B.B., and J.H.C. studied the protein families. C.G.P.MC., A.J.R., and E.S. conceived the binomial methods to identify significant shifts in amino acid composition, and E.S. implemented the changes of the GFmix model in the GFmix software. B.B., L.E., D.M., C.G.P.MC., 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.

Figure legends

Main figures

Fig. 1 | Phylogenetic position and metabolic potential of the families Afararchaeaceae and Asbonarchaeaceae. (a) Maximum likelihood phylogenetic tree of 35 euryarchaea, including four 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 Table 3. **(c)** Maximum likelihood phylogenetic tree of 24 DPANN archaea, including nine 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 Table 4. **(e)** Gene maps showing a previously undescribed 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 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").

Fig. 3 | Maximum likelihood phylogeny of archaea, including the 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.

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 Supplementary Fig. 18.

References

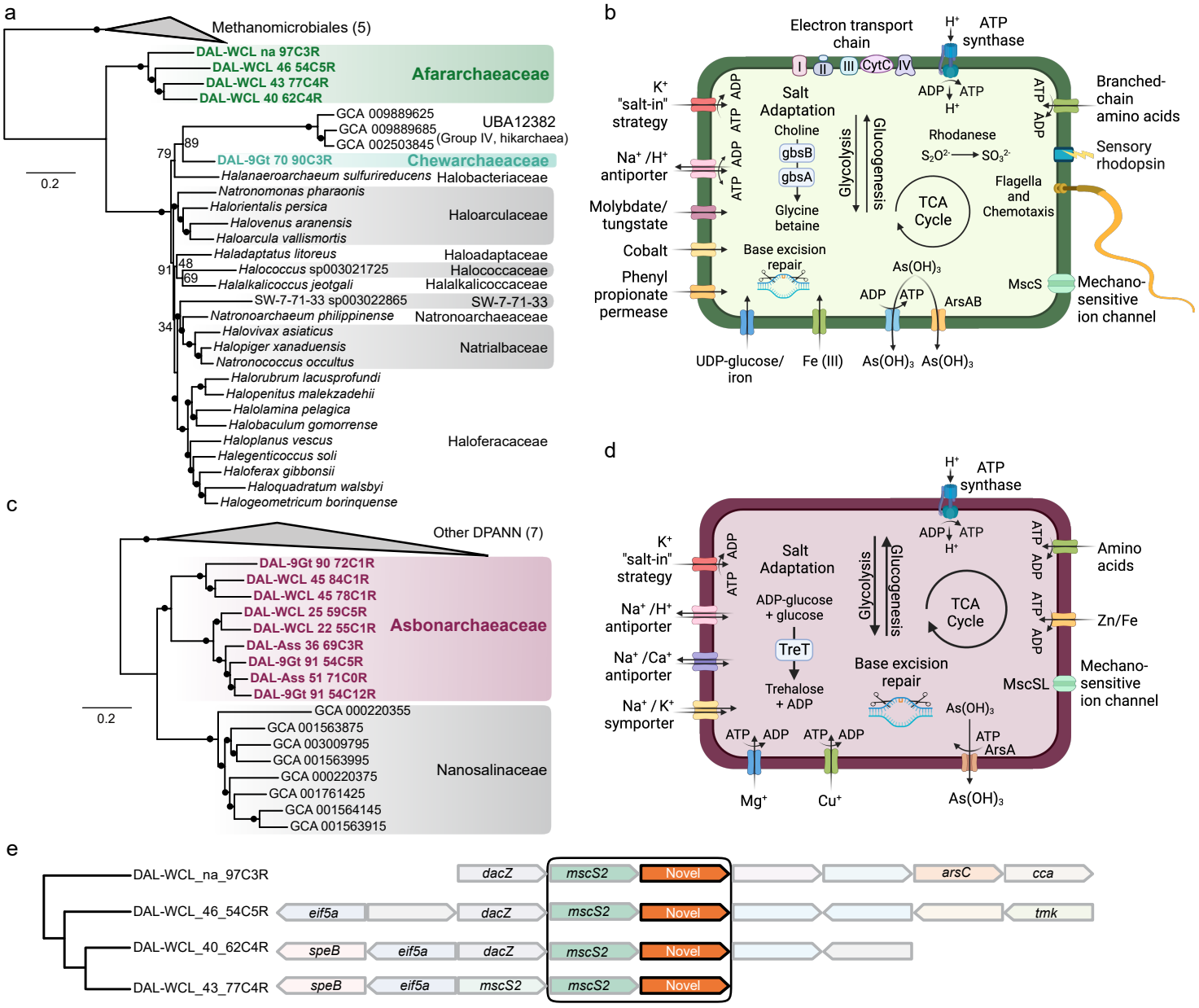
1. Oren, A. Diversity of halophilic microorganisms: Environments, phylogeny, physiology, and applications. *J. Ind. Microbiol. Biotechnol.* **28**, 56–63 (2002).
2. Oren, A. Molecular ecology of extremely halophilic Archaea and Bacteria. *FEMS Microbiol. Ecol.* **39**, 1–7 (2002).
3. Narasingarao, P. 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).

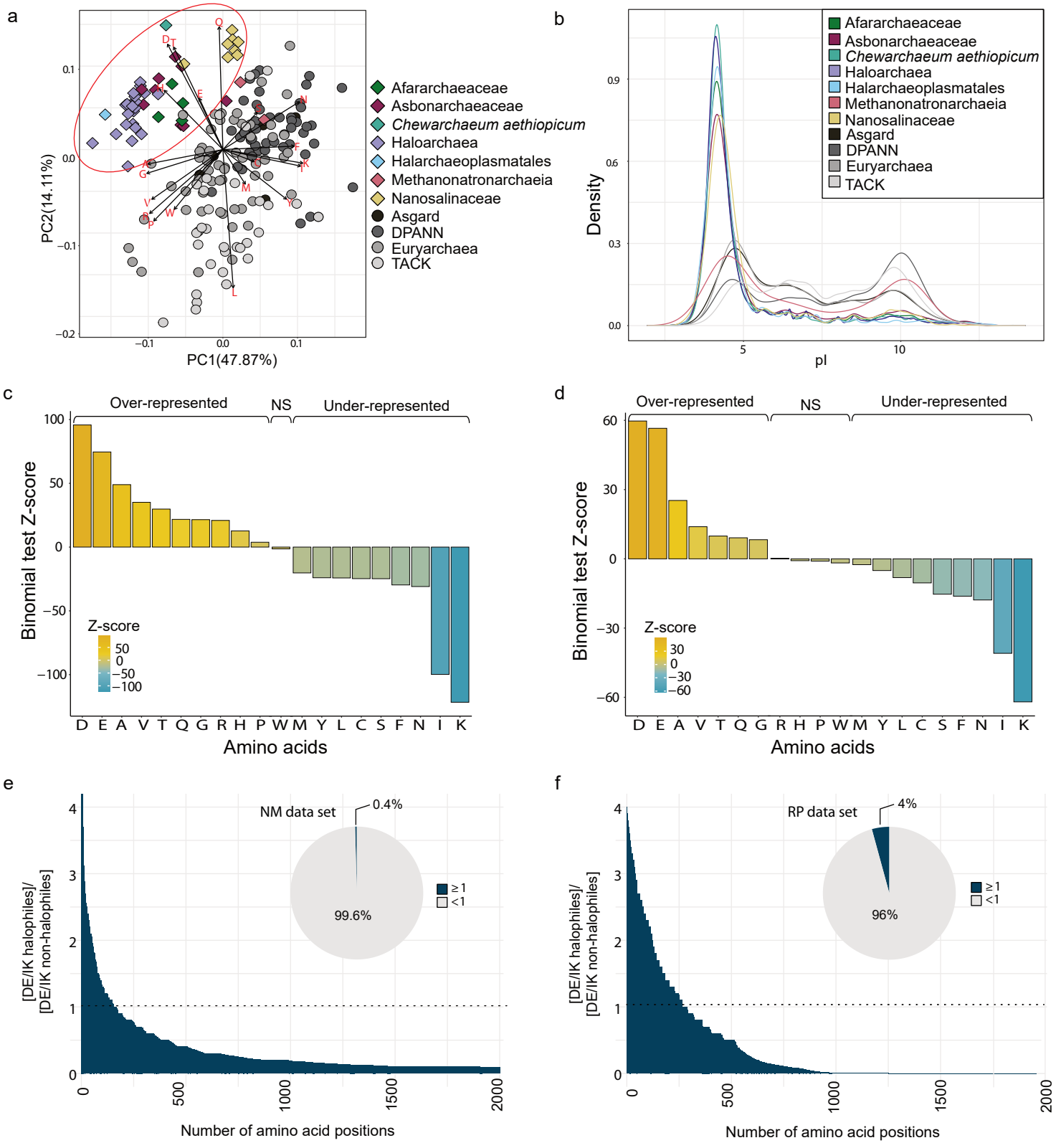
22. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **44**, D457–D462 (2016).
23. Falb, M. et al. Metabolism of halophilic archaea. *Extremophiles* **12**, 177–196 (2008).
24. Albers, S.-V. & Jarrell, K. F. The archaellum: how Archaea swim. *Front. Microbiol.* **6**, (2015).
25. Sasaki, J. & Spudich, J. L. Signal transfer in haloarchaeal sensory rhodopsin– transducer complexes. *Photochem. Photobiol.* **84**, 863–868 (2008).
26. 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).
27. 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).
28. 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).
29. 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).
30. 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).
31. Hamm, J. N. *et al.* Unexpected host dependency of Antarctic Nanohaloarchaeota. *Proc. Natl. Acad. Sci. USA* **116**, 14661–14670 (2019).
32. 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).
33. 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).
34. Mistry, J. et al. Pfam: The protein families database in 2021. *Nucleic Acids Res.* **49**, D412–D419 (2021).
35. 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).
36. 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).
37. Cabello-Yeves, P. J. & Rodriguez-Valera, F. Marine-freshwater prokaryotic transitions require extensive changes in the predicted proteome. *Microbiome* **7**, 117 (2019).
38. Rasmussen, T. How do mechanosensitive channels sense membrane tension? *Biochem. Soc. Trans.* **44**, 1019–1025 (2016).
39. 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).
40. Eme, L. et al. Inference and reconstruction of the heimdallarchaeial ancestry of eukaryotes. *Nature* **618**, 992–999 (2023).
41. Bergsten, J. A review of long-branch attraction. *Cladistics* **21**, 163–193 (2005).
42. Susko, E. & Roger, A. J. Long branch attraction biases in phylogenetics. *Syst. Biol.* **70**, 838–843 (2021).

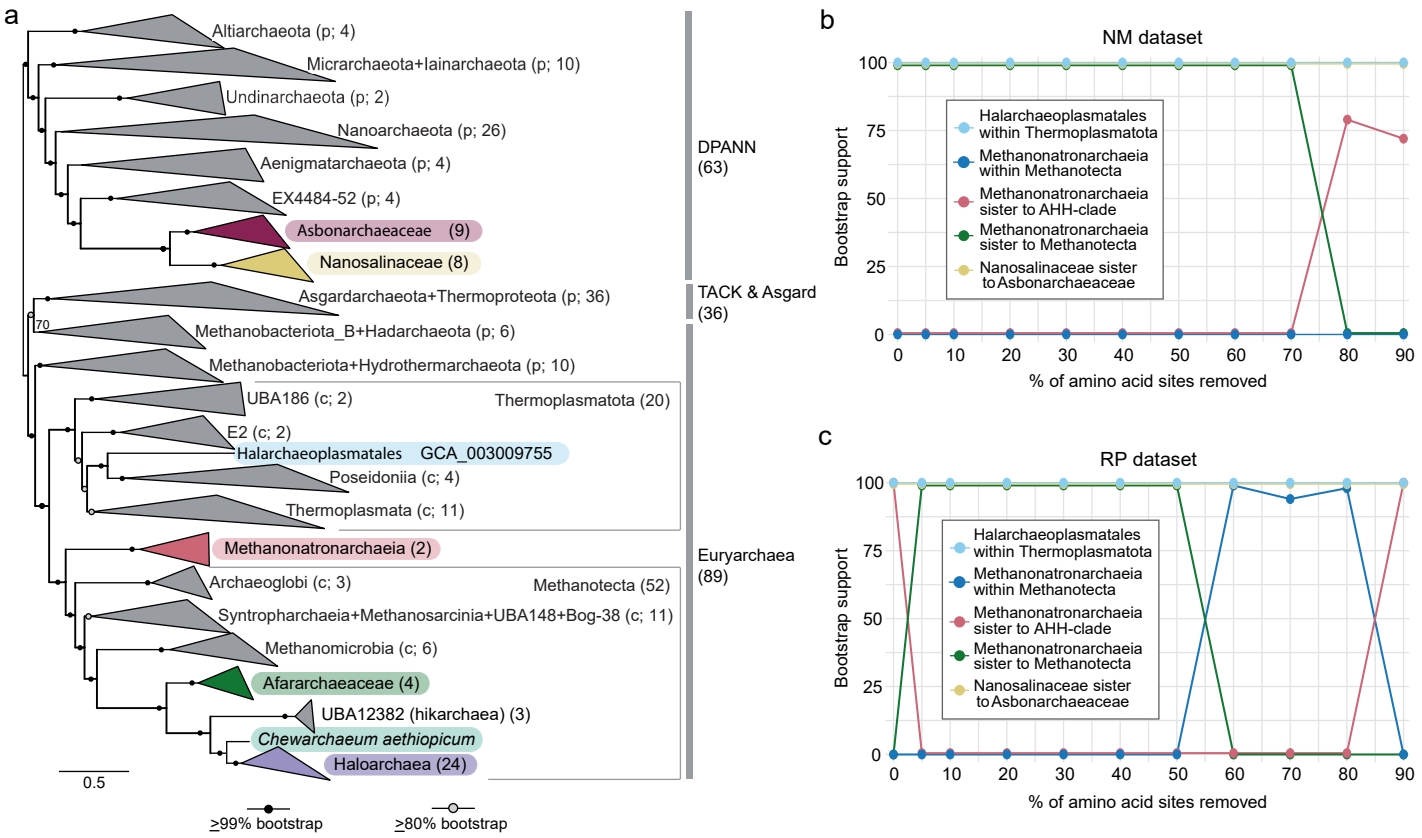
43. Shimodaira, H. An approximately unbiased test of phylogenetic tree selection. *Syst. Biol.* **51**, 492–508 (2002).
44. Muñoz-Gómez, S. A. et al. Site-and-branch-heterogeneous analyses of an expanded dataset favour mitochondria as sister to known Alphaproteobacteria. *Nat. Ecol. Evol.* **6**, 253–262 (2022).
45. Aouad, M. et al. Extreme halophilic archaea derive from two distinct methanogen Class II lineages. *Mol. Phylogenet. Evol.* **127**, 46–54 (2018).
46. 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).
47. Kellner, S. et al. Genome size evolution in the Archaea. *Emerg. Top. Life Sci.* ETLS20180021 (2018) doi:10.1042/ETLS20180021.
48. 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).
49. 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).
50. Giovannoni, S. J. *et al.* Genome streamlining in a cosmopolitan oceanic bacterium. *Science* **309**, 1242–1245 (2005).
51. 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).
52. 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).
53. 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).
54. 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).
55. 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).
56. Nelson-Sathi, S. et al. Origins of major archaeal clades correspond to gene acquisitions from bacteria. *Nature* **517**, 77–80 (2015).
57. 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).
58. 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).
59. 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., Sutcliffe, I.C., Venter, S.N. & Whitman, W.B. SeqCode: a nomenclatural code for prokaryotes described from sequence data. *Nat Microbiol*. 7, 1702-1708 (2022).
60. 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)

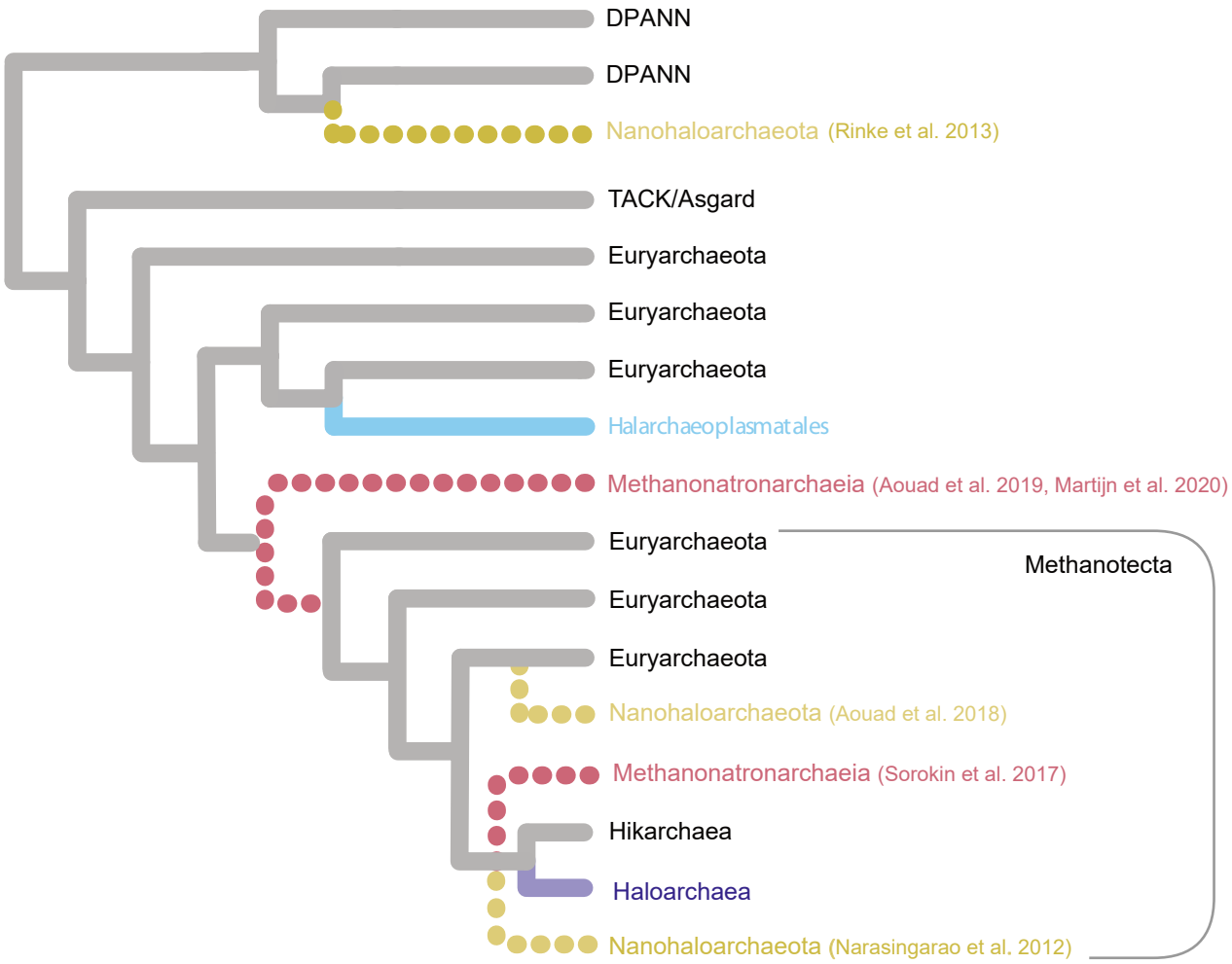
61. Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.* **9**, 1–8 (2018).
62. Rodriguez-R, L. M. & Konstantinidis, K. T. Bypassing cultivation to identify bacterial species: Culture-independent genomic approaches identify credibly distinct clusters, avoid cultivation bias, and provide true insights into microbial species. *Microbe Mag.* **9**, 111–118 (2014).
63. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).
64. Tatusov, R. L., Galperin, M. Y., Natale, D. A. & Koonin, E. V. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* **28**, 33–36 (2000).
65. Eren, A. M. et al. Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ* **3**, e1319 (2015).
66. Aramaki, T. et al. KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics* **36**, 2251–2252 (2020).
67. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).
68. Eddy, S. R. Accelerated Profile HMM Searches. *PLOS Comput. Biol.* **7**, e1002195 (2011).
69. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
70. Nayfach, S. et al. A genomic catalog of Earth's microbiomes. *Nat. Biotechnol.* **39**, 499–509 (2021).
71. Coelho, L. P. et al. Towards the biogeography of prokaryotic genes. *Nature* **601**, 252–256 (2022).
72. Almeida, A. et al. A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat. Biotechnol.* **39**, 105–114 (2021).
73. Paoli, L. et al. Uncharted biosynthetic potential of the ocean microbiome. Preprint at https://doi.org/10.1101/2021.03.24.436479 (2021).
74. Almagro Armenteros, J. J. et al. SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat. Biotechnol.* **37**, 420–423 (2019).
75. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. L. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. *J. Mol. Biol.* **305**, 567–580 (2001).
76. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
77. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLOS ONE* **5**, e9490 (2010).
78. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
79. 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).
80. 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).
81. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. (Springer, 2009).

82. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., Bourne & P. E. The Protein Data Bank. *Nucleic Acids Res*. **28**, 235-242 (2000).
83. 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).
84. 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).
85. Pettersen, E. F. et al. UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Sci. Publ. Protein Soc.* **30**, 70–82 (2021).
86. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
87. 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).
88. 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).
89. 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).
90. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: Reconstruction, analysis, and visualization of phylogenomic data. *Mol. Biol. Evol.* **33**, 1635–1638 (2016).


a
Methanomicrobiales (5)
DAL-WCL na 97C3R
DAL-WCL 46 54C5R
DAL-WCL 43 77C4R
DAL-WCL 40 62C4R
Afararchaeaceae
GCA 009889625
GCA 009889685
GCA 002503845
UBA12382
(Group IV, hikarchaea)
89
79
DAL-9Gt 70 90C3R
Chewarchaeaceae
Halanaeroarchaeum sulfurireducens
Halobacteriaceae
Natronomonas pharaonis
Halorientalis persica
Halovenus aranensis
Haloarcula vallismortis
Haloarculaceae
Haladaptatus litoreus
Haloadaptaceae
48
91
69
Halococcus sp003021725
Halococcaceae
Halalkalicoccus jeotgali
Halalkalicoccaceae
SW-7-71-33 sp003022865
SW-7-71-33
Natronoarchaeum philippinense
Natronoarchaeaceae
34
Halovivax asiaticus
Halopiger xanaduensis
Natronococcus occultus
Natrialbaceae
Halorubrum lacusprofundi
Halopenitus malekzadehii
Halolamina pelagica
Halobaculum gomorrense
Haloplanus vescus
Halegenticoccus soli
Haloferax gibbonsii
Haloquadratum walsbyi
Halogeometricum borinquense
Haloferacaceae
0.2
b
Electron transport chain
I
II
III
CytC
IV
H+
ATP synthase
K+ "salt-in" strategy
Na+ /H+ antiporter
Molybdate/ tungstate
Cobalt
Phenyl propionate permease
Salt Adaptation
Choline
gbsB
gbsA
Glycine betaine
Glycolysis
Gluconeogenesis
TCA Cycle
Rhodanese
S2O2- → SO32-
Flagella and Chemotaxis
Base excision repair
As(OH)3
ArsAB
MscS
Branched-chain amino acids
Sensory rhodopsin
Mechano-sensitive ion channel
UDP-glucose/ iron
Fe (III)
As(OH)3
As(OH)3
c
Other DPANN (7)
DAL-9Gt 90 72C1R
DAL-WCL 45 84C1R
DAL-WCL 45 78C1R
DAL-WCL 25 59C5R
DAL-WCL 22 55C1R
Asbonarchaeaceae
DAL-Ass 36 69C3R
DAL-9Gt 91 54C5R
DAL-Ass 51 71C0R
DAL-9Gt 91 54C12R
0.2
GCA 000220355
GCA 001563875
GCA 003009795
GCA 001563995
GCA 000220375
GCA 001761425
GCA 001564145
GCA 001563915
Nanosalinaceae
d
H+
ATP synthase
K+ "salt-in" strategy
Na+ /H+ antiporter
Na+ /Ca+ antiporter
Na+ / K+ symporter
Salt Adaptation
ADP-glucose + glucose
TreT
Trehalose + ADP
Glycolysis
Gluconeogenesis
TCA Cycle
Base excision repair
As(OH)3
ArsA
MscSL
Amino acids
Zn/Fe
Mechano-sensitive ion channel
Mg+
Cu+
As(OH)3
e
DAL-WCL_na_97C3R
DAL-WCL_46_54C5R
DAL-WCL_40_62C4R
DAL-WCL_43_77C4R
eif5a
speB
dacZ
mscS2
Novel
arsC
cca
tmk


a
PC2(14.11%)
PC1(47.87%)
Afararchaeaceae
Asbonarchaeaceae
Chewarchaeum aethiopicum
Haloarchaea
Halarchaeoplasmatales
Methanonatronarchaeia
Nanosalinaceae
Asgard
DPANN
Euryarchaea
TACK
b
Density
pI
c
Over-represented
NS
Under-represented
Binomial test Z-score
Z-score
Amino acids
d
e
NM data set
0.4%
99.6%
[DE/IK halophiles]/
[DE/IK non-halophiles]
Number of amino acid positions
f
RP data set
4%
96%


a
Altiarchaeota (p; 4)
Micrarchaeota+Iainarchaeota (p; 10)
Undinarchaeota (p; 2)
Nanoarchaeota (p; 26)
Aenigmatarchaeota (p; 4)
EX4484-52 (p; 4)
Asbonarchaeaceae (9)
Nanosalinaceae (8)
DPANN (63)
Asgardarchaeota+Thermoproteota (p; 36)
TACK & Asgard (36)
70
Methanobacteriota_B+Hadarchaeota (p; 6)
Methanobacteriota+Hydrothermarchaeota (p; 10)
UBA186 (c; 2)
Thermoplasmatota (20)
E2 (c; 2)
Halarchaeoplasmatales GCA_003009755
Poseidoniia (c; 4)
Thermoplasmata (c; 11)
Methanonatronarchaeia (2)
Archaeoglobi (c; 3)
Methanotecta (52)
Syntropharchaeia+Methanosarcinia+UBA148+Bog-38 (c; 11)
Methanomicrobia (c; 6)
Afararchaeaceae (4)
UBA12382 (hikarchaea) (3)
Chewarchaeum aethiopicum
Haloarchaea (24)
Euryarchaea (89)
0.5
≥99% bootstrap
≥80% bootstrap
b
NM dataset
Bootstrap support
% of amino acid sites removed
Halarchaeoplasmatales within Thermoplasmatota
Methanonatronarchaeia within Methanotecta
Methanonatronarchaeia sister to AHH-clade
Methanonatronarchaeia sister to Methanotecta
Nanosalinaceae sister to Asbonarchaeaceae
c
RP dataset
Bootstrap support
% of amino acid sites removed
Halarchaeoplasmatales within Thermoplasmatota
Methanonatronarchaeia within Methanotecta
Methanonatronarchaeia sister to AHH-clade
Methanonatronarchaeia sister to Methanotecta
Nanosalinaceae sister to Asbonarchaeaceae


DPANN
DPANN
Nanohaloarchaeota (Rinke et al. 2013)
TACK/Asgard
Euryarchaeota
Euryarchaeota
Euryarchaeota
Halarchaeoplasmatales
Methanonatronarchaeia (Aouad et al. 2019, Martijn et al. 2020)
Euryarchaeota
Methanotecta
Euryarchaeota
Euryarchaeota
Nanohaloarchaeota (Aouad et al. 2018)
Methanonatronarchaeia (Sorokin et al. 2017)
Hikarchaea
Haloarchaea
Nanohaloarchaeota (Narasingarao et al. 2012)