Title: Reconstructing the evolutionary origins of extreme halophilic Archaeal lineages

Authors: Yutian Feng[1], Uri Neri[2], Sean Gosselin[1], Artemis S. Louyakis[1], R. Thane Papke[1], Uri Gophna[2], J. Peter Gogarten[1,3]

[1]Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT, 06268, USA

[2]School of Molecular Cell Biology and Biotechnology, Tel Aviv University, P.O. Box 39040, Tel Aviv 6997801, Israel

[3]Institute for Systems Genomics, University of Connecticut, Storrs, CT, 06268, USA

Correspondence: gogarten@uconn.edu

Running Head: Evolutionary origins of halophilic Archaea

Keywords: Nanohaloarchaea, Methanonatronarchaeia, gene concordance, metagenomic-assembled genome (MAG), single amplified genome (SAG)

Abstract

Interest and controversy surrounding the evolutionary origins of extremely halophilic Archaea has increased in recent years, due to the discovery and characterization of the Nanohaloarchaea and the Methanonatronarchaeia. Initial attempts in explaining the evolutionary placement of the two new lineages in relation to the classical Halobacteria (also referred to as Haloarchaea) resulted in hypotheses that imply the new groups share a common ancestor with the Haloarchaea. However, more recent analyses have led to a shift: the Nanohaloarchaea have been largely accepted as being a member of the DPANN superphylum, outside of the euryarchaeota; while the Methanonatronarchaeia have been placed near the base of the Methanotecta (composed of the class II methanogens, the halobacteriales, and archaeoglobales). These opposing hypotheses have far-reaching implications on the concepts of convergent evolution (unrelated groups evolve similar strategies for survival), genome reduction, and gene transfer. In this work, we attempt to resolve these conflicts with phylogenetic and phylogenomic data. We provide a robust taxonomic sampling of Archaeal genomes that spans the crenarchaeota, euryarchaeota, and the DPANN superphylum. In addition, we sampled and assembled 7 new representatives of the Nanohaloarchaea, from distinct geographic locations. Phylogenies derived from these data imply the highly conserved ATP synthase catalytic/non-catalytic subunits of Nanohaloarchaea share a sisterhood relationship with the Haloarchaea. This relationship, with strong support, was also observed for several other gene families. In addition, we present and evaluate data that argue for and against the monophyly of the DPANN superphylum. We employed phylogenetic reconstruction, constrained topology tests, and gene concordance factors to explore the support for and against the monophyly of the Haloarchaea, Nanohaloarchaea, and Methanonatronarchaeia.

Introduction

The status of Haloarchaea as the main inhabitant of hypersaline environments has been thrown into question in recent years with the discovery of the nanosized Nanohaloarchaea and the methanogenic Methanonatronarchaeia. Dissecting the evolutionary relationships between these new lineages and the Haloarchaea may inform on the origins of halophily and the role of genome streamlining. To thrive in extreme hypersaline environments (>150 g/L^{-1}), Haloarchaea employ a "salt-in" strategy through the import of potassium ions, in which the intracellular salt concentration equalizes with the external environmental condition (Oren, 2008). This acts to balance the cellular osmotic pressure but also has caused significant changes in amino acid usage, leading to an overabundance of acidic residues, aspartate and glutamate (D/E) in all Haloarchaea. The evolutionary origins of the Nanohaloarchaea have remained uncertain since their discovery (Ghai *et al*., 2011; Narasingarao *et al*., 2012). The composition of their proteome indicates that Nanohaloarchaea also use the "salt-in" strategy similar to Haloarchaea (Narasingarao *et al*., 2012). It was originally suggested that the Nanohaloarchaea are euryarchaeota that form a clade with the Haloarchaea, based on phylogenies of the 16S rRNA gene and ribosomal proteins (Narasingarao *et al*., 2012; Petitjean *et al*., 2014). Additional data obtained from individual cells via cell sorting followed by genome amplification and 16S rRNA sequencing analysis confirmed the original observations of the Nanohaloarchaea as a sister taxon to the Haloarchaea (Zhaxybayeva *et al*., 2013). More recently, based on analyses of concatenated conserved protein sequences, the Nanohaloarchaea were placed in a group together with similarly nanosized organisms, the Diapherotrites, Parvarachaeota, Aenigmarchaeota, and Nanoarchaeota, forming the DPANN superphylum (Andrade *et al*., 2015; Castelle *et al*., 2015; Rinke *et al*., 2013). Past analyses of this superphylum (Brochier-Armanet, Forterre, & Gribaldo,

2011; Petitjean *et al*., 2014; Raymann *et al*., 2014; Williams *et al*., 2015) suggested that the DPANN grouping may not be due to shared ancestry but may reflect an artifact due to long branches and/or small genomes. However, more recent analyses supported a monophyletic DPANN clade (Williams *et al*., 2017). Aouad *et al.* performed a multi-locus analysis using various models, which did not include DPAN sequences, and placed the Nanohaloarchaea with the methanocellales and the Haloarchaea with the methanomicrobiales (Aouad *et al*., 2018); *i.e.*, the Nanohaloarchaea were recovered as a member of the euryarchaeota, but not as a sister-group to the Haloarchaea. We note that a similar controversy surrounds the phylogenetic position of the Nanoarchaeota. *Nanoarchaeum equitans* was first considered a representative of a new deep branching archaeal phylum (Huber *et al.*, 2002), *i.e.*, an archaeon not a member of the euryarchaeotes or crenarchaeotes. However, later analyses of ribosomal proteins, phylogenetically informative HGTs, and signature genes led to the conclusion that *N. equitans* may represent a fast-evolving euryarchaeote instead of an early branching novel phylum (Brochier *et al.*, 2005; Dutilh *et al.*, 2008; Urbonavičius *et al.*, 2008).

Recently, another group of extreme halophiles, the Methanonatronarchaeia (also spelled as Methanonatronarcheia), were discovered and predicted to also use the "salt-in" strategy (Sorokin *et al*., 2017). Initial multi-locus phylogenetic analyses placed these methanogenic halophiles in a monophyletic clade with the Haloarchaea, suggesting they are an evolutionary intermediate between methanogens and modern halophiles. However, Aouad *et al*., (2019) contested this placement: a multi-locus dataset placed the Methanonatronarchaeia basal to a superclass named Methanotecta, a group that includes the Archaeoglobales, class II methanogens and Haloarchaea (Adam *et al*., 2017). Several conclusions can be drawn from these differing results with regard to adaptation to a halophilic lifestyle, most note-worthy of which is the convergent evolution of

the "salt-in" strategy among these three lineages. However, if Nanohaloarchaea, Haloarchaea, and Methanonatronarcheia form a monophyletic group, as seen with some analyses of 16S rRNA and ribosomal proteins, the hypothesis of common ancestral origins can more easily account for the evolutionary development of the salt-in strategy.

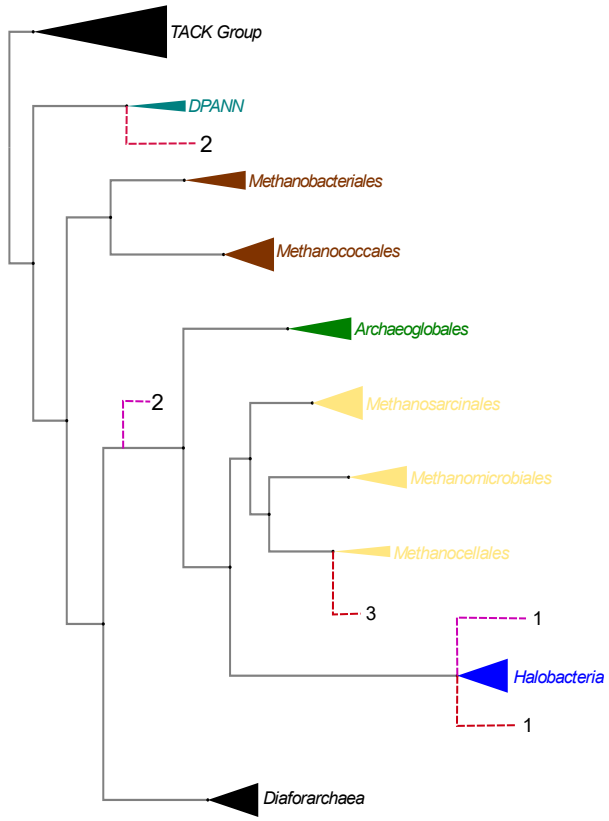

Figure 1. **Summary of proposed placements of halophilic lineages mapped on an Archaeal reference tree.** Individual taxa have been collapsed into higher taxonomic groups. The red (R) indicators represent the different placements proposed for the *Nanohaloarchaea*, while the purple (P) indicators are used for the *Methanonatronarchaeia.* Sources for each placement: R1 (Narasingarao *et al*., 2012), R2 (Andrade *et al*., 2015), R3 (Aouad *et al.*, 2018); P1 (Sorokin *et al*., 2017), and P2 (Aouad *et al.*, 2019).

The evolutionary relationships of the three halophilic lineages remain unresolved; Figure 1 summarizes the current controversies. This lack of resolution can be, at least in part, due to biases that are known to complicate phylogenetics. The genomes of the Methanonatronarchaeia and Nanohaloarchaea are comparatively small with average genome sizes of <2.1Mb and ~1.1 Mb, and most genome entries in public databases are incomplete. The Haloarchaea are known to

be highly recombinogenic (Mohan *et al*., 2014; Naor *et al*., 2012) and are likely physically associated with at least some of the Nanohaloarchaea (Andrade *et al*., 2015; Cono *et al*., 2019; Hamm *et al*., 2019).

Phylogenies based on many genes, like the ones discussed above, face many problems: 1) Genes have different evolutionary histories (e.g. duplication and transfer) and forcing the histories of all the genes on a single tree does not reflect the complex evolutionary history of the genomes (Lapierre *et al*., 2014). In particular, genes acquired from outside the group under consideration may create a strong signal for placing the recipient of the transferred gene at the base of the group. 2) Genes experience differing levels of purifying selection, which can lead to long branch attraction (LBA) artifacts (Felsenstein, 1978). 3) Heterotachy, or varying substitution rates among sites, in specific lineages can cause problems even if the individual genes evolved along the same history as the host species (Philippe *et al.*, 2005). 4) Substitution bias may create convergent signals in unrelated groups.

The work reported here was guided by the hypothesis that the phylogenetic reconstruction of a single, slowly evolving gene might be more robust against artifacts of phylogenetic reconstructions compared to analyses that are based on large sets of genes that may represent different evolutionary histories, include missing data, and contain genes with high substitution rates. We reconstruct single gene alongside multi-locus phylogenies to correct for these sources of bias and to critically assess the evolutionary relationships of the Haloarchaea, Nanohaloarchaea, and Methanonatronarchaeia. The ATP synthase catalytic and non-catalytic subunits, AtpA and AtpB, represent extremely slow evolving genes (J. P. Gogarten, 1994) conserved throughout Archaea and is likely the slowest evolving gene in cellular organisms. The evolution of these subunits is slow enough to ameliorate rate signal bias and minimize

compositional heterogeneity that otherwise plague reconstructions that includes DPANN and Haloarchaeal sequences. These ATP synthase subunits have been used successfully as a phylogenetic marker for large scale reconstructions (Gogarten & Taiz, 1992). To provide a more robust sampling of the Nanohaloarchaea, we include seven newly sequenced and assembled nanohaloarchaeal genomes together with existing genomes mined from the NCBI database. Robust sampling of the taxon, like the one we offer here, has the potential to improve the recovery of evolutionary relationships without adding more sites (genes) (Graybeal, 1998).

In maximum likelihood and Bayesian phylogenies, we find that the Nanohaloarchaea group robustly with the Haloarchaea in the single gene phylogenies, while the Methanonatronarchaeia was placed as a deeper branching euryarchaeal lineage, most likely at the base of the Methanotecta superclass. Clearly, phylogenies based on single gene or operons may reflect the transfer of the analyzed gene(s). However, we also demonstrate, through the Approximately Unbiased test (Shimodaira, 2002) that a constrained monophyletic grouping of the Nanohaloarchaea, Haloarchaea, and Methanonatronarchaeia in analyses based on the concatenation of the genome core is not a significantly worse explanation than the other proposed topologies in Fig. 1, and therefore may represent the true tree. We provide support for a sister-group relationship between Haloarchaea (Haloarchaea) and Nanohaloarchaea. Our analyses are marginally compatible with the hypothesis that all halophilic archaea form a monophyletic group.

Results

Increased genomic representation of the Nanohaloarchaea

We obtained five new Nanohaloarchaea single amplified genomes (SAGs) from solar salterns in Spain and two metagenome assembled genomes (MAGs) from Israel. The summary statistics and accompanying information of these genomes can be found in Table S2. These nine genomes greatly expand the number of Nanohaloarchaea assemblies available for analyses (18 at time of writing). Total average nucleotide identity (tANI) was used to delineate taxonomy amongst the newly described Nanohaloarchaea. Figure S5 is a distance-based tree calculated from corrected tANI distance (see Figure S1 for distance matrix) between the previously described and newly described Nanohaloarchaea. Using conservative cutoffs, it appears SAGs SCGC AAA188-M06 and M04 may belong to the genus *Ca.* Nanosalina. SAG M21 seems to be a member of *Ca*. Nanosalinarium, while the remaining new genomes (SAGs and MAGs) do not belong to any previously described genera.

The genome described as *Nanohaloarchaea archaeon* PL-Br10-U2g5 (Vavourakis *et al*., 2016) was likely miss-identified as a Nanohaloarchaeon. We find that this strain unequivocally groups within *Halorubrum* species in ribosomal (protein and rRNA), whole genome distance (estimated by ANI), and single gene phylogenies (Figs. 2, 3).

Phylogenetic placement of halophilic lineages

To shed light on the evolutionary origins of the Methanonatronarchaeia and the Nanohaloarchaea, we have produced three sets of trees from distinct markers that contain differing phylogenetic signals. All three tree sets contain >100 taxa, representing Archaea that span the Euryarchaeota, TACK group, and the candidate DPAN(N) superphylum. The phylogenies are depicted as rooted with the TACK Group. However, the root of the Archaeal

tree remains an open question with the emergence of Eukaryotes from the Archaea likely having rendered them paraphyletic (Fournier & Poole, 2018; Gribaldo *et al*., 2010; Williams *et al*., 2013). However, the placement of the Archaeal root does not impact the conclusions drawn from our phylogenies, presuming the Archaeal root is placed outside of the euryarchaeal crown group.

ATP synthase catalytic and non-catalytic subunit phylogenies

The ATP synthase catalytic and non-catalytic subunits are slow evolving, essential genes. Single protein phylogenies of these subunits may ameliorate LBA and deletion-transfer-loss (DTL) conflicts; both of which have plagued large scale, multi-locus attempts at reconstructing the Archaeal phylogeny. Maximum likelihood phylogenies (see File S3 for list of models used and the treefiles produced) of the AtpA and AtpB proteins were created using site-homogeneous and site-heterogeneous substitution models. All tree reconstructions based on the original, unaltered multiple sequence alignments confidently placed the Nanohaloarchaea as a sister group to the Haloarchaea ($\geq$ 91 Bootstrap Value (BV)). A representative example tree constructed with the concatenated AtpA+B proteins is shown below (Fig 2). Nanohaloarchaea and Haloarchaea are grouped together and are positioned as sisters to the class II methanogens**.** Curiously, the Methanonatronarchaeia are placed as a deeper branching euryarchaeal lineage, suggesting either gene transfer or convergent evolution in regard to the extreme halophilic "salt-in" strategy. In only one analysis did all three lineages group together, with poor support inside the Methanotecta, (File S3, LG+C50 AtpA d4). The branch lengths of the AtpA and AtpB phylogenies are relatively short, ameliorating possible LBA artefacts. The placement of the remaining DPAN (DPANN sequences minus the Nanohaloarchaea) taxa appears erratic. However, it is worth noting the other groups considered as members of DPANN fail to form a

monophyletic clade in all of the ATPase based trees and several of the branches breaking the DPAN(N) group apart are supported by high BVs (File S3).

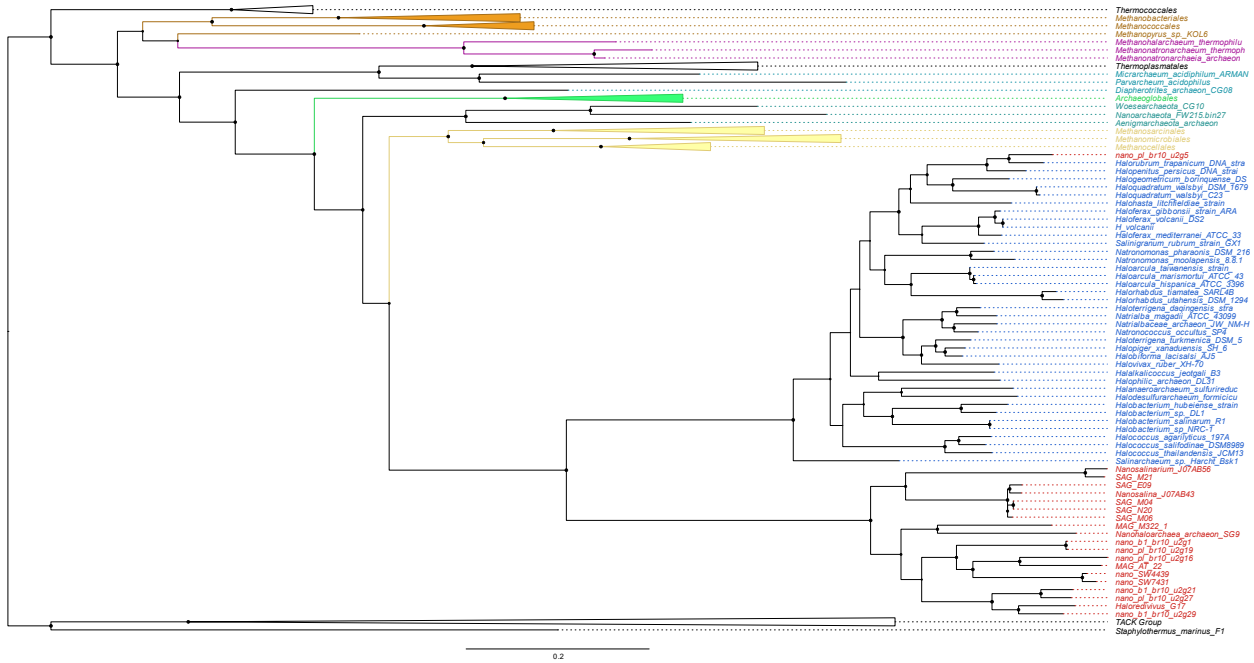

Figure 2. **Maximum likelihood phylogeny calculated from concatenated AtpA+B proteins.** The depicted tree contains most features of the other calculated ATP synthase phylogenies. Several taxa were collapsed into higher taxonomic ranks. Important taxa including the halophilic lineages and DPANN (teal) sequences have been colored; *Nanohaloarchaea* (red), *Haloarchaea* (blue), *Methanonatronarchaeia* (purple), class I methanogens (orange), and class II methanogens (yellow). Branch supports are non-parametric bootstraps, most relevant supports are displayed with numerical value, otherwise all supports greater than 80% BV are represented by circles at the nodes. The tree is drawn as rooted by the TACK Group, but should be considered as unrooted.

Compositional bias

Compositional bias in encoded amino acids can generate artifacts in large, domain-wide phylogenies (Aouad *et al.*, 2018; Aouad *et al.*, 2019). However, due to the slow rate of evolution in these ATPase subunits, compositional bias has been minimized. A chi-squared test of composition (File S1) for both protein alignments revealed only 10% and 6% of taxa fail the composition test in AtpA and AtpB sequences, respectively. None of the sequences that failed this composition test belong to a member of the halophilic lineages barring one sequence that

belonged to a *Nanohaloarchaeon* with an incompletely sequenced *atpA.* To minimize compositional bias, both alignments were recoded into 4 and 6 Dayhoff groups (Susko & Roger, 2007). These recoded alignments were used to create maximum likelihood and Bayesian phylogenies, which mostly recapitulated the groupings discussed earlier (Fig. S2, File S3). The only difference was that in several instances Methanonatronarchaeia moved either to the base of the Methanotecta, grouped with the Haloarchaea and Nanohaloarchaea, or with the TACK group.

A reason for bias in extreme halophilic lineages is an acidic proteome, *i.e.*, increased presence of aspartic and glutamic acid (D/E) in their protein sequence. This may lead to "compositional attraction", where those taxa that have an abundance of D/E sites are more likely to cluster together in a phylogeny. Sites that contained a conserved D/E residue among the Haloarchaea, Nanohaloarchaea, and the Methanonatronarchaeia were deleted from the AtpA and AtpB alignments. Maximum likelihood phylogenies of these new alignments were created (Fig. S2, File S3), and the topology discussed above was recovered, albeit with lower support due to the loss of phylogenetically informative sites.

Conflict between gene and genome trees

To explore the possible synergy of using multiple loci in reconstructing the history of these halophilic lineages, we constructed supermatrices of core genes and ribosomal proteins from a taxonomic sampling similar to the ATP synthase trees. We first created a core genome matrix composed of 146 loci (called the large core supermatrix); all of these genes are represented in every single nanohaloarchaeal genome considered complete. From this large core supermatrix we took a subset of 12 genes and formed a smaller core supermatrix; the 12 genes in the smaller core was found to be in the represented of all nanohaloarchaeal genomes, regardless of genome

completion. The relatively small length of this supermatrix ensured that these genes would be found in all sampled taxa and did not contribute conflicting phylogenetic signals due to DTL events. We also created a ribosomal supermatrix containing 44 concatenated ribosomal proteins. All supermatrices were analyzed with site-homogenous and heterogenous models, including models for each individual partition. The placement of the Haloarchaea and Methanonatronarchaeia calculated from the small core and ribosomal supermatrices (Fig 3a, S3) resemble many previously calculated large-scale phylogenies (Andrade *et al.*, 2015; Sorokin *et al.*, 2017). In contrast to the single protein trees, these multi-locus trees place the Nanohaloarchaea with several (not all) members of the DPANNs (Aenigmarchaeota and Woesearchaeota) in a monophyletic clade basal to the Methanotecta, with poor support values (Fig. 3a). In our small core phylogeny, the Methanonatronarchaeia are positioned as a sister group to the Haloarchaea, similar to Sorokin *et al.*, 2017, while they are basal to the class II methanogens in the ribosomal protein phylogeny (Fig. S3b, File S3). A phylogeny of the 16S rRNA gene, also recovers the Nanohaloarchaea as a sister-group to the Haloarchaea (File S3). Phylogenies reconstructed from the large core supermatrix offer conflicting information, Fig S4 and Table S3. On one hand all three halophilic lineages group together with poor supports using site homogenous models, however using partitioned analysis the Nanohaloarchaea form a monophlyetic clade within the DPAN(N) superphylum (Fig S4b).

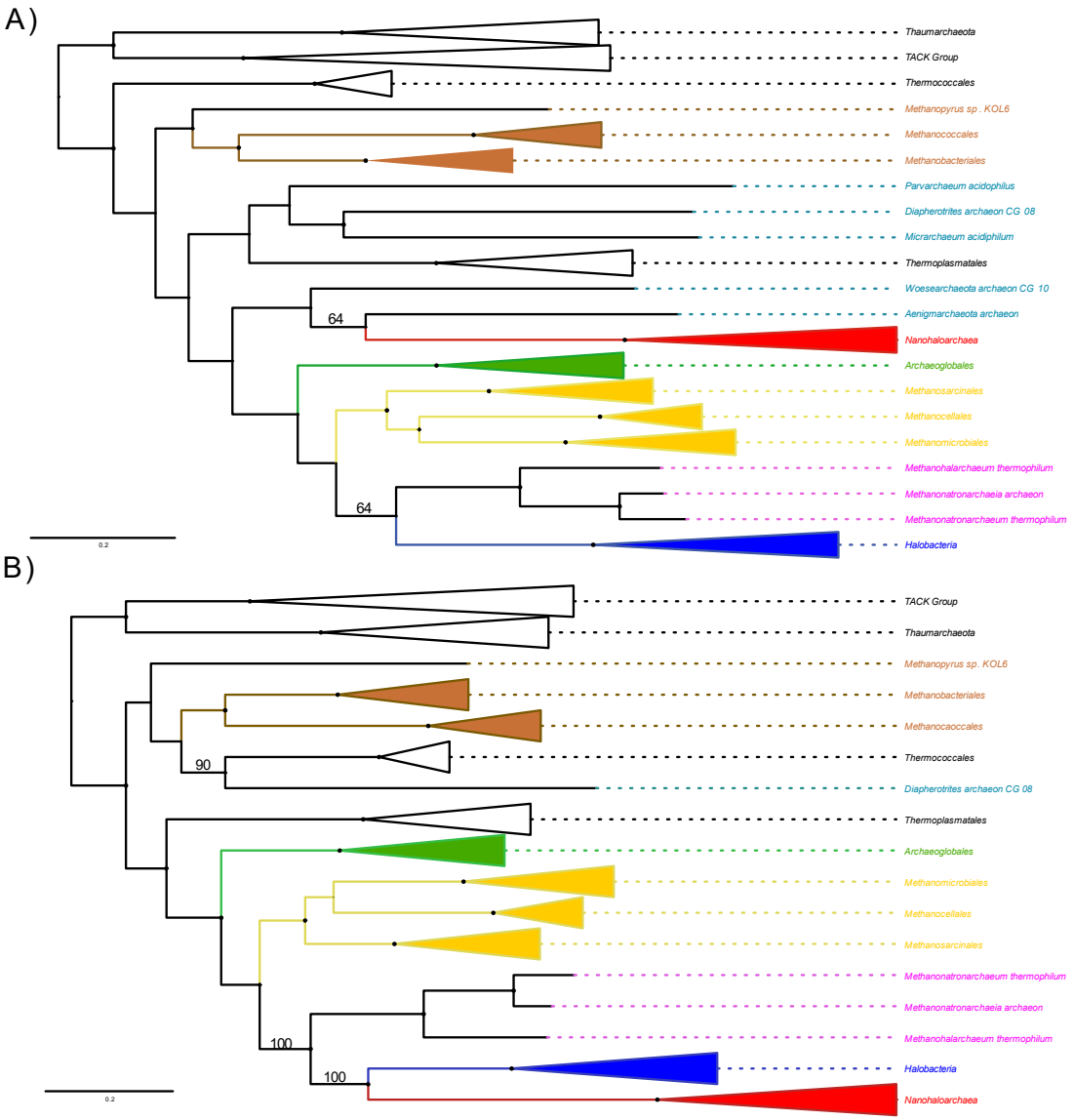

Figure 3**. Maximum likelihood phylogenies of Archaeal small core genome supermatrices**. A) Calculated from a supermatrix that contained the DPAN sequences (teal). Branch supports are non-parametric bootstraps, most relevant supports are displayed with numerical value, otherwise all supports greater than 80% BV are represented by circles at the nodes. The *Nanohaloarchaea* (red) group with several DPAN sequences, and are deep branching euryarchaeota. In B) we removed all but one of the DPAN sequences from the supermatrix and recalculated the tree. The *Nanohaloarchaea* group with the other halophilic lineages in this case, and the overall branch supports (BV) are higher. It is worth noting the *Nanohaloarchaea* have the longest branch in both phylogenies.

In an attempt to minimize the impact of long branch attraction, the DPAN sequences were removed from the smaller core genome supermatrix. Curiously, the phylogenetic tree calculated from this new supermatrix (File S3) place the Nanohaloarchaea,

Methanonatronarcheia, and Haloarchaea together as a monophyletic clade, with high support. To make sure compositional attraction was not a driving force of this topology, we removed conserved D/E sites from the supermatrix, which lead to an identical topology with lower supports (Table S3). It is clear that the inclusion of the DPAN sequences impact the tree building process and possibly leads to artifacts due to rate signal attraction. Adding these sequences in a large-scale multi-locus alignment further confounds the clustering. We recognize that excluding DPAN sequences altogether from the alignment precludes the possibility of Nanohaloarchaea being attracted to the group; however, we note that the Nanohaloarchaea did not group at the base of the euryarchaeota. Two locations appear to attract the nanohaloarchaeal sequences: the other halophiles and the small, divergent DPAN genomes. In Figure 3b, we have recalculated the phylogeny from the alignment in 3a, with all but one DPAN sequence (from a Diapherotrites archaeon) removed, and recover highly supported monophyly of the three halophilic lineages.

To further investigate the possible monophyly of the three lineages, we created 100 non-parametric bootstrap trees from the original small core genome supermatrix and the recoded ATPase (AtpA + d4) dataset with the DPAN(N)s included. We recovered 5 and 22 bootstrap trees from the core genome and the ATPase dataset that placed Nanohaloarchaea, Methanonatronarcheia, and Haloarchaea as a monophyletic group (e.g., contained a monophyletic grouping of all three halophilic lineages). The explanatory power of these constrained monophyletic trees was tested against the proposed alternative clusterings (Fig. 1) of the three lineages using the approximately unbiased test (AU-test, Shimodaira, 2002). The topology test revealed 2 of 5 and 10 of 22 trees from the core genome and ATPase dataset had p-values > 0.1 when evaluated against the best tree calculated from the respective original

alignments (File S2). The AU-test reveals that the explanatory power of the trees with a monophyletic grouping of halophilic archaea, is not significantly worse than the most likely phylogenies and in fact, the true tree of halophiles may be (or similar to) one of these trees.

We also employed gene concordance factor analysis (gCF (Ané *et al*., 2007; Gadagkar *et al*., 2005)) to our large core supermatrix to dissect which topology each individual merged gene-partition supported (summarized in Table 1). Using a reference tree that was constrained to group the Nanohaloarchaea with the Haloarchaea, 9 partitions (encompassing 30 genes and over 11K amino acid (AA) residues, Table 1) were found to be concordant with the internode that supports these groups' monophyly. Contained in these partitions, which were classified by similar rate categories, were highly conserved proteins such as the ATP synthase operon, ribosomal proteins, and elongation factors. In contrast, we also use an alternative reference tree that was constrained to group the Nanohaloarchaea within the DPANN superphylum, and only found marginal support for this grouping (4 partitions, 7 genes, 2.4K AA residues, Table 1). These DPANN supporting partitions contained genes for the RNA polymerase, FtsZ, and some ribosomal proteins as well. We also applied the gCF method to locate the best supported placement of the Methanonatronarchaeia, and found the highest concordance was at the base of Methanotecta super class. The full list of concordant partitions, as well as reference trees can be found in File S4; in addition, the full list of concordant genes (*i.e.*, from a similar analysis using unmerged partitions) can be found in File S5.

Table 1. Summary of bipartition support in gCF analyses.

Bipartition Explored	Number of Supporting Genes (AA sites)	Example Genes
Nanohaloarchaea + Haloarchaea	30 Genes (11,393 AA residues)	atpB, eif2a, rps8, ychF
Nanohaloarchaea + Aenigmarchaeota (DPANN)	7 Genes (2,389 AA residues)	rpoB, rps5, ftsZ, sucD
Methanonatronarchaeia + Methanotecta	24 Genes (11,646 AA residues)	rpoA, rpoB, ftsZ, eif2a

Discussion

Sisterhood of Nanohaloarchaea and Haloarchaea

Analysis of the catalytic and non-catalytic subunits of the archaeal ATP synthase group the enzyme from Nanohaloarchaea as a sister-group to the Halobacterial (Haloarchaeal) subunits (Fig. 2; Wang *et al.*, 2019). This strongly supported grouping is also recovered when the data are recoded to reduce compositional bias, when alignment columns containing acidic residues in both the Nanohaloarchaea and the Haloarchaea are deleted, and when the CAT-GTR model (a model that is less sensitive to compositional effects and long branch attraction artifacts) is used in phylogenetic reconstruction. None of these analyses recovered the DPANN clan. Given the consistent support for the Nanohaloarchaea-Haloarchaea clade in the AtpA and AtpB phylogenies, it is unlikely that this finding is due to compositional bias or long branch attraction. Two conflicting hypotheses can reconcile our findings with those of previous analyses based on concatenation of several genes or on gene tree/species tree reconciliations: (1) the ATP synthase

was acquired by the ancestor of the Nanohaloarchaea from a relative of the Haloarchaea or (2) the previous multi-locus analyses do not reflect evolutionary history, but are artifacts due to high substitution rates, gene transfer, and small genomes; and the Nanohaloarchaea and Haloarchaea share a common ancestor. The recent study by Wang *et al*. (2019) includes a phylogeny derived from the entire ATPase operon in Archaea, that also recovered the sisterhood between the Nanohaloarchaea and Haloarchaea. Wang *et al*. consider horizontal transfer of the operon as explanation for this grouping, and also observe an identical operon structure in both groups, which supports the monophyly of nanohaloarchaeal and haloarchaeal ATPases. Wang *et al.* (2019) recognized the clear conflict between a DPANN supergroup and the phylogeny of the ATPases, and they reconciled this conflict by assuming horizontal gene transfer of the ATPsynthase/ATPase. However, analyses presented here and by Rayman *et al.*, 2014 and Aouad *et al.*, 2018-9 reveal that the Nanohaloarchaea may not be a member of the DPANN group, weakening the argument for gene transfer.

Furthermore, we provide another layer of considerations with analyses of the genes that make up the nanohaloarchaeal core genome and the 16S rRNA gene, which garners support for the hypothesis that the Nanohaloarchaea and Haloarchaea are sister-groups. When genomes from DPAN members were included, the Nanohaloarchaea were recovered as part of the DPANN group. However, a phylogenetic reconstruction that was constrained to group Nanohaloarchaea with Haloarchaea resulted in a maximum likelihood phylogeny that the AU-test (Shimodaira, 2002) evaluated as not significantly worse than the best tree for this dataset. Furthermore, in the absence of the other DPANN genomes or with only one DPAN sequence, the Nanohaloarchaea formed a clade with the Haloarchaea (Fig. 3b), even after removing potential biases.

The radically different placements of the Nanohaloarchaea (Fig. 1, red indicators) can be at least partially attributed to the taxonomic sampling of the DPANN superphylum in the alignment supermatrix. In instances where the Nanohaloarchaea were recovered inside the euryarchaeota (Narasingarao *et al.*, 2012, Zhaxybayeva *et al.*, 2013; Aouad *et al.*, 2018; Aouad *et al.*, 2019), DPAN sequences were not included in the tree. However, including a robust sampling of DPAN sequences in the alignment (Andrade *et al.*, 2015; Sorokin *et al.*, 2017; Wang *et al.*, 2019, Figure 3) generally attracts the Nanohaloarchaea into the superphylum. It is obvious that one cannot recover the evolutionary relationship between the Nanohaloarchaea and the DPAN superphylum, without including DPAN sequences in the alignment. However in our study, in the absence of DPAN sequences, the nanohaloarchaeal sequences were not recovered at the base of the euaryarchaeota, *i.e.*, the place where the DPAN sequences were recovered (File S3, - DPAN, +1 DPAN supermatrices).

The gCF analysis revealed a larger list of conserved genes supporting the Nanohaloarchaea-Haloarchaea sister group relationship, versus the inclusion of the Nanohaloarchaea in the DPANN superphylum (see Table 1). Previous analyses have indicated high bootstrap support (Sorokin *et al*., 2017; Wang *et al*., 2019) for including the Nanohaloarchaea within the DPANN. This support may reflect the strong but artifactual signal in fast evolving genes, and phylogenetic signals created through gene transfers. In contrast, our gCF analysis dissected the concatenation based on individual gene trees, revealing opposing phylogenetic signals present in the original concatenated dataset. It is important to supplement the sampling variance measure for the singular branch (*i.e.*, bootstrap), with a measure of variance in the overall dataset with metrics like the concordance factors. The concordance factors reveal variance (conflict) within the multi-locus alignment datasets. These findings suggest that

using data for slowly evolving genes from more organisms, such as the ATP synthase, has a better chance of resolving deep phylogenetic relations than the reconstruction of a phylogenetic tree from the concatenation of many genes into a single phylogenomic analysis; an obvious caveat is that the better resolved single gene phylogeny represents only a single gene or operon, and that its phylogeny is embedded in the net-like, reticulated genome phylogeny.

Monophyly of extreme halophilic archaea

The Methanonatronarchaeia did not reveal a well-supported association with any particular Archaeal group in any of these phylogenies, except for the case where we removed DPAN sequences. In the ATP synthase-based phylogenies, the homologs from three members of this group were recovered as a deeper branching euryarchaeal lineage without well supported affinity to any other euryarchaeal group. The sequences from the Methanonatronarchaeia were, however, separated by at least one well supported bipartition from the other halophilic archaea grouping with non-halophilic methanogens (Fig. 2).

A concatenation of the genes from the nanohaloarchaeal core also did not reliably place the Methanonatronarchaeia. Removing the other members of DPAN from the same dataset results in a topology compatible with the hypothesis that the extreme halophilic archaea form a monophyletic clade and that the salt-in strategy evolved only once; however, the support for this grouping is marginal. In our assessment, the support values for monophyly of the extreme halophilic archaea in the analyses of the core genomes are too low to convince of monophyly for this clade. The most highly supported placement of the Methanonatronarchaeia, according to the gCF analysis, is at the base of the Methanotecta super-class, as proposed by Aouad *et al.*, 2019.

Aouad and colleagues provided evidence for three independent adaptation to high salt environments in Halobacteria, Nanohaloarchaea, and Methanonatronarchaeia (Aouad *et al.*, 2018; Aouad *et al.*, 2019). While we consider convergent evolution events rare, independent adaptations to hypersalinity through the salt-in strategy, revealed through a shift in the distribution of the theoretical isoelectric points of encoded proteins (Oren, 2008), have been observed in *Salinibacter* (Bacteroidetes) and *Salinicoccus* (Firmicutes) (see Fig S6), with minimal reliance on HGT from haloarchaea (Mongodin *et al*., 2005). The Methanonatronarchaeia have been deduced to employ the salt-in strategy, using intracellular potassium ion concentrations (Sorokin *et al.*, 2017), the same tactic used by the Nanohaloarchaea and Halobacteria. However, a proteomic analysis of the theoretical isoelectric point (pI) distributions reveals a less biased distribution of pIs in these methanogens compared to other proteomes of organisms that use the salt-in strategy (Halobacteria, Nanohaloarchaea, *etc*.) (Fig S6). The Methanonatronarchaeia may be an example of independent adaptation to hypersalinity; however, the concentration of intracellular potassium did not yet have a significant impact on the distribution of the theoretical isoelectric points of encoded proteins. This distribution of theoretical isoelectric point in Methanonatronarchaeia resembles that found in marine archaea (Fig S6).

Conclusion

Our analysis of ATPase subunits and ribosomal RNA supports the grouping of Haloarchaea and Nanohaloarchaea into a monophyletic lineage. Strong statistical support for this grouping either reflects a gene transfer event, or shared ancestry of the two groups. It remains an open question, if the analyses favoring separate origins of Haloarchaea and Nanohaloarchaea are impacted by lack of resolution and artifacts of phylogenetic reconstruction, or if the recovery of a strongly supported monophyly of these two groups in the analysis of a slowly evolving protein is due to gene transfer. Our analysis of a concatenated nanohaloarchaeal core, ribosomal proteins, and ribosomal rRNA is compatible with a monophyletic grouping of Haloarchaea and Nanohaloarchaea, weakening the argument in favor of gene transfer. A larger set of conserved genes supports this grouping rather than the inclusion of the Nanohaloarchaea in the DPANN superphylum. While we cannot exclude the possibility of massive gene transfer of conserved, slowly evolving genes (16S rRNA and ATP synthase) from the Haloarchaea to the Nanohaloarchaea, the monophyly of Nanohaloarchaea and Haloarchaea is viable alternative. In either case, our study documents the evolutionary relationships between the Haloarchaea and the Nanohaloarchaea, either through shared ancestry or through gene transfer.

Methods

Sample collection, DNA extraction, and sequencing of new genomes

Two hypersaline environments in Israel were sampled for metagenomic sequences: the Dead Sea and hypersaline pools at the Mediterranean coast in Atlit. Briefly, water samples from the Dead Sea (31°30'07.2"N 35°28'37.2"E) were extracted using Niskin bottles in late July 2018. To create the enriched media, the Dead Sea water (DSW) was diluted with autoclaved double distilled water (DDW) (final ratio ⅕ [DDW/DSW]), amended with 0.1% glycerol, 1 uM KH_2PO_4, 1 g/L peptone (Bacto, New South Wales, Australia), 1 g/L casamino acids (Difco, Detroit, MI USA). The media was incubated at 30 °C for 42 days.

The Atlit environmental samples were collected from high salt tide-pools on the coast of Israel (32°42′37.3″N 34°56′32.0″E) in mid-October 2018. Harvesting of the microbial communities was performed by serial passage through filters (0.45um, 0.22um, 0.1um) (Merck KGaA, Darmstadt, Germany). Environmental samples (Atlit) were first prefiltered using filter paper No. 1 (11um pore size) (Munktell & Filtrak, Bärenstein, Germany). The filters were then kept in -80 °C until DNA extraction. DNA was extracted from the filters using DNeasy PowerLyzer PowerSoil kit (QIAGEN, Hilden, Germany) following the manufacturer's protocol. For Dead Sea and Atlit samples, DNA purified from the 0.22um filters was used for library preparation (NuGen Celero enzymatic with UDI indexing). The libraries were ran on Illumina NovaSeq with SP flow cell, generating paired end reads (2x150bp).

Single amplified genomes (SAGs) were generated using fluorescence-activated cell sorting and multiple displacement amplification, as previously described (Zhaxybayeva *et al.*, 2013), from hypersaline salterns located in Santa Pola (Spain). Low coverage shotgun sequencing of SAGs was performed using Nextera library preparation and NextSeq 500 sequencers (Stepanauskas et al. 2017), resulting in an average of 377k, 2x150 bp reads per SAG. Although this number of reads is sub-optimal for high-quality genome reconstruction (Stepanauskas et al. 2017), thy were sufficient to perform the specific analyses of this study. SAG generation and raw sequence generation were performed at the Bigelow Laboratory for Ocean Sciences Single Cell Genomics Center (scgc.bigelow.org).

Sequence quality control

Raw reads obtained from single cell sequencing were trimmed and quality assured using Sickle v1.33 (Joshi & Fass, 2011) and FastQC v0.115 (Andrews, 2010). SPAdes v3.10.1 (Bankevich *et al*., 2012) was used to complete initial assemblies of single cell genomes, using option -sc. Contigs from the initial assembly were polished and bridged using the post-assembly Unicycler v0.4.7 pipeline (Wick *et al*., 2017), using normal and bold settings. Conflicts between normal and bold assemblies were investigated and reconciled in Bandage v0.8.1 (Wick *et al*., 2015). The taxonomy and completeness of the polished assemblies were verified with CheckM v1.0.7 (Parks *et al*., 2015), on default settings using a custom lineage marker developed specifically for Nanohaloarchaea (available on request). For the metagenome assembled genomes (MAGs), raw reads were trimmed using Trimmomatic-0.36 (Bolger *et al*., 2014) and quality assured using FastQC v0.10.1. SPAdes v3.11.0 was used to assemble the MAGs, using option -meta. Assembly Graphs were manually investigated using Bandage v0.8.1. The assembled genomes were

annotated with Archaeal mode Prokka v1.13.3 (Seemann, 2014). Sequences annotated as the ATP synthase alpha and beta subunits were retrieved from these genomes manually. These nine newly assembled genomes in addition to nine high quality assemblies on NCBI were compiled in a library to identify well-represented, conflict free core genes of the *Nanohaloarchaea.* Get_Homologues v03012018 (Contreras-Moreira & Vinuesa, 2013) with the COGtraingles v2.1 (Kristensen *et al*., 2010) and orthoMCL v1.4 (Li *et al*., 2003) algorithms (-t 0/1 option) were used to identify these "bona-fide" core genes.

Whole genome distance analysis

Average nucleotide identity (ANI) was calculated using a slight modification of the JSpecies method (Richter & Rosselló-Móra, 2009). Genomes were divided into 1,020 nt fragments and used as the query for pairwise BLAST searches. A 70% identity and 70% coverage cutoff was implemented in a manner akin to the global ANI (gANI) filtering method (Varghese *et al*., 2015). The filtered BLASTN (Camacho *et al*., 2009) searches were also used to calculate a modified gANI and alignment fraction (AF), which were used to construct a phylogenetic tree as per the tANI method (Gosselin *et al.*, 2019 *in prep*). The entire method and standalone script can be found at: **https://github.com/SeanGosselin/tANI_Matrix.git.**

Assembly of datasets

116 high quality genomes spanning the Archaea domain were collected through NCBI's ftp site (see script 1, and Table S1 for list), and were supplemented with the nine newly assembled Nanohaloarchaea genomes. AtpA and AtpB protein sequences were found in these genomes and gathered with BLASTP v2.7.1, using default parameters. Similarly, protein sequences of 146

Nanohaloarchaea core proteins and forty-four ribosomal proteins were found and gathered from these genomes using TBLASTN using default parameters. Sequences hits from each protein were categorized into their own respective files and aligned with Muscle v3.8.1551(Edgar, 2004) using default parameters. Each alignment file of the core and ribosomal protein dataset was concatenated using wrapper_supermatrix.py (Supp. Script 2), to generate supermatrices and the associated nexus partition files. A subset of 12 genes, present in >95% Nanohaloarchaea genomes, were gathered from the larger core (146 genes) to investigate potential DTL conflicts present in the Nanohaloarchaea genomes.

To reduce the influence of heterogeneous composition of sequences throughout the Archaea domain, we first recoded the AtpA and AtpB alignment sequences into 4 and 6 Dayhoff groups based on functional classes of amino acids, using PhyloBayes v4.1 (Lartillot *et al*., 2009). We also manually curated alternative alignments which had removed alignment columns if they contained an Aspartate or Glutamate (D/E) residue that was conserved in the Nanohaloarchaea, Haloarchaea, and Methanonatronarchaeia, to minimize compositional attraction. The same process was repeated for the smaller core genome supermatrix.

Phylogenetic Estimation

IQTREE v1.6.9 (Nguyen *et al*., 2015) was used to calculate maximum likelihood phylogenies for all alignments and supermatrices. The best site homogeneous models were used for the estimation as determined by the Bayesian Information Criterion using ModelFinder (Kalyaanamoorthy *et al*., 2017). The AtpA and AtpB alignments were also analyzed by the LG+C60 (Le *et al*., 2008) mixture model. The multi-locus sequence alignments and their

respective partition schemes were also used to calculate maximum likelihood phylogenies, using appropriate models for each partition and the merging of similar partitions with MFP+MERGE (Chernomor, Von Haeseler, & Minh, 2016) of IQTREE. Bayesian inference of Dayhoff recoded alignments were conducted within PhyloBayes v4.1(Lartillot *et al*., 2009; Quang, Gascuel, & Lartillot, 2008) using the CAT+GTR +G4 model in two independent chains for each alignment. These chains ran until convergence (maxdiff < 0.25), >400,000 trees sampled, with a burn-in of the first 10% of the trees, to calculate a majority rule consensus tree. All trees in this paper were drawn and editorialized with Figtree v1.4.3 (Rambaut, 2016). The approximately unbiased test was also carried out in IQTREE, with 10,000 RELL replicates for each sample tree (100 total). gCF analyses of the large core supermatrix was carried out in the IQTREEv1.7.17 beta.

Data Availability

The newly assembled nanohaloarchaeal genomes and accompanying information have been deposited into GenBank under BioProject PRJNA587522 for the SAGs, and PRJNA523480 for the MAGs. Alignments can be made available on request.

Acknowledgments and Funding

We thank the staff of the Bigelow Laboratory for Ocean Sciences' Single Cell Genomics Center for the generation of single cell genomic data. We specifically thank Ramunas Stepanauskas for helpful critical discussions and overseeing the single cell sequencing, which was funded by DEB-1441717 to Ramunas Stepanauskas. This work was supported through grants from the Binational Science Foundation (BSF 2013061 to UG, JPG, and RTP); the National Science

Foundation (NSF/MCB 1716046 to JPG, RTP and UG) within the BSF-NSF joint research program; and NASA exobiology (NNX15AM09G, and 80NSSC18K1533 to RTP).

Author Contributions

The project was conceived by RTP, JPG, and UG. Sampling, sequencing, and genome reconstruction of Dead Sea samples were conducted by UN and UG. Single cell samples were collected by RTP, and assemblies performed by ASL and YF. Phylogenies were calculated by YF, whole genome distance by SG. All authors contributed to writing and editing the manuscript.

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. *ISME Journal*. https://doi.org/10.1038/ismej.2017.122

Andrade, K., Logemann, J., Heidelberg, K. B., Emerson, J. B., Comolli, L. R., Hug, L. A., … Banfield, J. F. (2015). Metagenomic and lipid analyses reveal a diel cycle in a hypersaline microbial ecosystem. *ISME Journal*. https://doi.org/10.1038/ismej.2015.66

Andrews, S. (2010). FastQC. *Babraham Bioinformatics*. https://doi.org/citeulike-article-id:11583827

Ané, C., Larget, B., Baum, D. A., Smith, S. D., & Rokas, A. (2007). Bayesian estimation of concordance among gene trees. *Molecular Biology and Evolution*. https://doi.org/10.1093/molbev/msl170

Aouad, M., Borrel, G., Brochier-Armanet, C., & Gribaldo, S. (2019). Evolutionary placement of

Methanonatronarchaeia. *Nature Microbiology*. https://doi.org/10.1038/s41564-019-0359-z

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*. https://doi.org/10.1016/j.ympev.2018.04.011

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., … Pevzner, P. A. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology : A Journal of Computational Molecular Cell Biology*. https://doi.org/10.1089/cmb.2012.0021

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btu170

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

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

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics*. https://doi.org/10.1186/1471-2105-10-421

Castelle, C. J., Wrighton, K. C., Thomas, B. C., Hug, L. A., Brown, C. T., Wilkins, M. J., … Banfield, J. F. (2015). Genomic expansion of domain archaea highlights roles for organisms from new phyla in anaerobic carbon cycling. *Current Biology*. https://doi.org/10.1016/j.cub.2015.01.014

Chernomor, O., Von Haeseler, A., & Minh, B. Q. (2016). Terrace Aware Data Structure for Phylogenomic Inference from Supermatrices. *Systematic Biology*. https://doi.org/10.1093/sysbio/syw037

Cono, V. La, Messina, E., Rohde, M., Arcadi, E., Ciordia, S., Crisafi, F., … Yakimov, M. M. (2019). Differential polysaccharide utilization is the basis for a nanohaloarchaeon : haloarchaeon symbiosis. *BioRxiv*. https://doi.org/10.1101/794461

Contreras-Moreira, B., & Vinuesa, P. (2013). GET_HOMOLOGUES, a Versatile Software Package for Scalable and Robust Microbial Pangenome Analysis. *Applied and Environmental Microbiology*. https://doi.org/10.1128/aem.02411-13

Dutilh, B. E., Snel, B., Ettema, T. J. G., & Huynen, M. A. (2008). Signature genes as a phylogenomic tool. *Molecular Biology and Evolution*. https://doi.org/10.1093/molbev/msn115

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkh340

Felsenstein, J. (1978). Cases in which Parsimony or Compatibility Methods will be Positively Misleading. *Systematic Biology*. https://doi.org/10.1093/sysbio/27.4.401

Fournier, G. P., & Poole, A. M. (2018). A briefly argued case that Asgard Archaea are part of the eukaryote tree. *Frontiers in Microbiology*. https://doi.org/10.3389/fmicb.2018.01896

Gadagkar, S. R., Rosenberg, M. S., & Kumar, S. (2005). Inferring species phylogenies from multiple genes: Concatenated sequence tree versus consensus gene tree. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*. https://doi.org/10.1002/jez.b.21026

Ghai, R., Pašić, L., Fernández, A. B., Martin-Cuadrado, A. B., Mizuno, C. M., McMahon, K. D.,

… Rodríguez-Valera, F. (2011). New abundant microbial groups in aquatic hypersaline environments. *Scientific Reports*. https://doi.org/10.1038/srep00135

Gogarten, J. P. (1994). Which is the most conserved group of proteins? Homology-orthology, paralogy, xenology, and the fusion of independent lineages. *J Mol Evol*, *39*(5), 541–543.

Gogarten, Johann Peter, & Taiz, L. (1992). Evolution of proton pumping ATPases: Rooting the tree of life. *Photosynthesis Research*. https://doi.org/10.1007/BF00039176

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

Gribaldo, S., Poole, A. M., Daubin, V., Forterre, P., & Brochier-Armanet, C. (2010). The origin of eukaryotes and their relationship with the Archaea: Are we at a phylogenomic impasse? *Nature Reviews Microbiology*. https://doi.org/10.1038/nrmicro2426

Hamm, J. N., Erdmann, S., Eloe-Fadrosh, E. A., Angeloni, A., Zhong, L., Brownlee, C., … Cavicchioli, R. (2019). Unexpected host dependency of Antarctic Nanohaloarchaeota. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.1905179116

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*. https://doi.org/10.1038/417063a

Joshi, N., & Fass, J. (2011). sickle - A windowed adaptive trimming tool for FASTQ files using quality. *(Version 1.33)*. https://doi.org/10.1088/0022-3727/13/9/001

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., Von Haeseler, A., & Jermiin, L. S. (2017). ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nature Methods*. https://doi.org/10.1038/nmeth.4285

Kristensen, D. M., Kannan, L., Coleman, M. K., Wolf, Y. I., Sorokin, A., Koonin, E. V., & Mushegian, A. (2010). A low-polynomial algorithm for assembling clusters of orthologous groups from intergenomic symmetric best matches. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btq229

Lapierre, P., Lasek-Nesselquist, E., & Gogarten, J. P. (2014). The impact of HGT on phylogenomic reconstruction methods. *Briefings in Bioinformatics*. https://doi.org/10.1093/bib/bbs050

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

Le, S. Q., Lartillot, N., & Gascuel, O. (2008). Phylogenetic mixture models for proteins. *Philosophical Transactions of the Royal Society B: Biological Sciences*. https://doi.org/10.1098/rstb.2008.0180

Li, L., Stoeckert, C. J., & Roos, D. S. (2003). OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Research*. https://doi.org/10.1101/gr.1224503

Mohan, N. R., Fullmer, M. S., Makkay, A. M., Wheeler, R., Ventosa, A., Naor, A., … Papke, R. T. (2014). Evidence from phylogenetic and genome fingerprinting analyses suggests rapidly changing variation in Halorubrum and Haloarcula populations. *Frontiers in Microbiology*. https://doi.org/10.3389/fmicb.2014.00143

Naor, A., Lapierre, P., Mevarech, M., Papke, R. T., & Gophna, U. (2012). Low species barriers in halophilic archaea and the formation of recombinant hybrids. *Current Biology*. https://doi.org/10.1016/j.cub.2012.05.056

Narasingarao, P., Podell, S., Ugalde, J. A., Brochier-Armanet, C., Emerson, J. B., Brocks, J. J.,

… Allen, E. E. (2012). De novo metagenomic assembly reveals abundant novel major lineage of Archaea in hypersaline microbial communities. *ISME Journal*. https://doi.org/10.1038/ismej.2011.78

Nguyen, L. T., Schmidt, H. A., Von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*. https://doi.org/10.1093/molbev/msu300

Oren, A. (2008). Microbial life at high salt concentrations: Phylogenetic and metabolic diversity. *Saline Systems*. https://doi.org/10.1186/1746-1448-4-2

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*. https://doi.org/10.1101/gr.186072.114

Petitjean, C., Deschamps, P., López-Garciá, P., & Moreira, D. (2014). Rooting the domain archaea by phylogenomic analysis supports the foundation of the new kingdom Proteoarchaeota. *Genome Biology and Evolution*. https://doi.org/10.1093/gbe/evu274

Philippe, H., Zhou, Y., Brinkmann, H., Rodrigue, N., & Delsuc, F. (2005). Heterotachy and long-branch attraction in phylogenetics. *BMC Evolutionary Biology*. https://doi.org/10.1186/1471-2148-5-50

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

Rambaut, A. (2016). FigTree v1.4.3.

Raymann, K., Forterre, P., Brochier-Armanet, C., & Gribaldo, S. (2014). Global phylogenomic analysis disentangles the complex evolutionary history of DNA replication in Archaea. *Genome Biology and Evolution*. https://doi.org/10.1093/gbe/evu004

Richter, M., & Rosselló-Móra, R. (2009). Shifting the genomic gold standard for the prokaryotic species definition. *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.0906412106

Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N. N., Anderson, I. J., Cheng, J. F., … Woyke, T. (2013). Insights into the phylogeny and coding potential of microbial dark matter. *Nature*. https://doi.org/10.1038/nature12352

Seemann, T. (2014). Prokka: Rapid prokaryotic genome annotation. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btu153

Shimodaira, H. (2002). An approximately unbiased test of phylogenetic tree selection. *Systematic Biology*. https://doi.org/10.1080/10635150290069913

Sorokin, Di. Y., Makarova, K. S., Abbas, B., Ferrer, M., Golyshin, P. N., Galinski, E. A., … Koonin, E. V. (2017). Discovery of extremely halophilic, methyl-reducing euryarchaea provides insights into the evolutionary origin of methanogenesis. *Nature Microbiology*. https://doi.org/10.1038/nmicrobiol.2017.81

Susko, E., & Roger, A. J. (2007). On reduced amino acid alphabets for phylogenetic inference. *Molecular Biology and Evolution*. https://doi.org/10.1093/molbev/msm144

Urbonavičius, J., Auxilien, S., Walbott, H., Trachana, K., Golinelli-Pimpaneau, B., Brochier-Armanet, C., & Grosjean, H. (2008). Acquisition of a bacterial RumA-type tRNA(uracil-54, C5)-methyltransferase by Archaea through an ancient horizontal gene transfer. *Molecular Microbiology*. https://doi.org/10.1111/j.1365-2958.2007.06047.x

Varghese, N. J., Mukherjee, S., Ivanova, N., Konstantinidis, K. T., Mavrommatis, K., Kyrpides, N. C., & Pati, A. (2015). Microbial species delineation using whole genome sequences. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkv657

Vavourakis, C. D., Ghai, R., Rodriguez-Valera, F., Sorokin, D. Y., Tringe, S. G., Hugenholtz, P., & Muyzer, G. (2016). Metagenomic insights into the uncultured diversity and physiology of microbes in four hypersaline soda lake brines. *Frontiers in Microbiology*. https://doi.org/10.3389/fmicb.2016.00211

Wang, B., Qin, W., Ren, Y., Zhou, X., Jung, M.-Y., Han, P., … Jia, Z. (2019). Expansion of Thaumarchaeota habitat range is correlated with horizontal transfer of ATPase operons. *The ISME Journal*. https://doi.org/10.1038/s41396-019-0493-x

Wick, R. R., Judd, L. M., Gorrie, C. L., & Holt, K. E. (2017). Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Computational Biology*. https://doi.org/10.1371/journal.pcbi.1005595

Wick, R. R., Schultz, M. B., Zobel, J., & Holt, K. E. (2015). Bandage: Interactive visualization of de novo genome assemblies. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btv383

Williams, T. A., & Embley, T. M. (2014). Archaeal "dark matter" and the origin of eukaryotes. *Genome Biology and Evolution*. https://doi.org/10.1093/gbe/evu031

Williams, T. A., Foster, P. G., Cox, C. J., & Embley, T. M. (2013). An archaeal origin of eukaryotes supports only two primary domains of life. *Nature*. https://doi.org/10.1038/nature12779

Williams, T. A., Heaps, S. E., Cherlin, S., Nye, T. M. W., Boys, R. J., & Embley, T. M. (2015). New substitution models for rooting phylogenetic trees. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *370*(1678), 20140336. https://doi.org/10.1098/rstb.2014.0336

Williams, T. A., Szöllősi, G. J., Spang, A., Foster, P. G., Heaps, S. E., Boussau, B., … Embley,

T. M. (2017). Integrative modeling of gene and genome evolution roots the archaeal tree of life. *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.1618463114

Zhaxybayeva, O., Stepanauskas, R., Mohan, N. R., & Papke, R. T. (2013). Cell sorting analysis of geographically separated hypersaline environments. *Extremophiles*. https://doi.org/10.1007/s00792-013-0514-z