Extreme halophilic archaea derive from two distinct methanogen Class II lineages

Monique Aouad, Najwa Taïb, Anne Oudart, Michel Lecocq, Manolo Gouy, Céline Brochier-Armanet

Molecular Phylogenetics and Evolution 127 (2018) 46–54

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

Extreme halophilic archaea derive from two distinct methanogen Class II lineages

Monique Aouad[1], Najwa Taib[1], Anne Oudart, Michel Lecocq, Manolo Gouy, Céline Brochier-Armanet*

Univ Lyon, Université Lyon 1, CNRS, UMR5558, Laboratoire de Biométrie et Biologie Évolutive, 43 bd du 11 novembre 1918, F-69622 Villeurbanne, France

ARTICLE INFO

Keywords:
Stenosarchaea
Compositional bias
Long branch attraction
Slow-fast method
Rate signal
Substitutional saturation

ABSTRACT

Phylogenetic analyses of conserved core genes have disentangled most of the ancient relationships in *Archaea*. However, some groups remain debated, like the DPANN, a deep-branching super-phylum composed of nanosized archaea with reduced genomes. Among these, the *Nanohaloarchaea* require high-salt concentrations for growth. Their discovery in 2012 was significant because they represent, together with *Halobacteria* (a Class belonging to *Euryarchaeota*), the only two described lineages of extreme halophilic archaea. The phylogenetic position of *Nanohaloarchaea* is highly debated, being alternatively proposed as the sister-lineage of *Halobacteria* or a member of the DPANN super-phylum. Pinpointing the phylogenetic position of extreme halophilic archaea is important to improve our knowledge of the deep evolutionary history of *Archaea* and the molecular adaptive processes and evolutionary paths that allowed their emergence. Using comparative genomic approaches, we identified 258 markers carrying a reliable phylogenetic signal. By combining strategies limiting the impact of biases on phylogenetic inference, we showed that *Nanohaloarchaea* and *Halobacteria* represent two independent lines that derived from two distinct but related methanogen Class II lineages. This implies that adaptation to high salinity emerged twice independently in *Archaea* and indicates that emergence of *Nanohaloarchaea* within DPANN in previous studies is likely the consequence of a tree reconstruction artifact, challenging the existence of this super-phylum.

1. Introduction

Archaea gather many lineages with very diverse metabolic capacities and living in a broad range of ecosystems, including the human body (Eme and Doolittle, 2015). Recent advances in high-throughput sequencing technologies have revealed many new major uncultured environmental groups, most of them being known only through ribosomal RNA or genomic sequences (Castelle et al., 2015; Rinke et al., 2013; Schleper et al., 2005). This is for instance the case of *Nanohaloarchaea*, a group of extreme halophilic nanosized *archaea* ($< 0.8\,\mu m$ diameter), discovered recently in Lake Tyrell, Australia (Narasingarao et al., 2012), and detected in Spanish solar salterns (Ghai et al., 2011). Environmental surveys indicated that *Nanohaloarchaea* exist worldwide (Narasingarao et al., 2012). This discovery was significant because, at the time, extreme halophilic archaea were restricted to *Halobacteria*, a euryarchaeotal class of heterotrophs (Oren, 2014, 2008). Thus, *Nanohaloarchaea* represent the second lineage of extreme halophilic archaea described so far.

Similar to *Halobacteria*, *Nanohaloarchaea* could use the "salt-in" strategy, which involves the accumulation of molar concentrations of potassium and chloride within cells (Oren, 2008), to maintain their osmotic balance (Narasingarao et al., 2012). This generates strong constrains on intracellular proteins and cellular apparatus, and requires specific adaptations. In fact, proteins of extreme halophiles are depleted in large hydrophobic residues (Wright et al., 2002) and accumulate negative charges at their surface to maintain proper conformation and activity, and prevent aggregation (Ban et al., 2000; Britton et al., 2006). However, *Nanohaloarchaea* show differences with *Halobacteria*, preferring glutamic acid to aspartic acid, serine to threonine, and reduced frequencies of proline and histidine (Narasingarao et al., 2012). They have also smaller genomes (approximately 1.2 Mb), a single-copy rRNA operon, and globally a lower G+C genomic content (Narasingarao et al., 2012).

The phylogenetic position of extreme halophilic archaea is still unresolved. Phylogenetic analyses of the RNA component of the small subunit of the ribosome and large supermatrices of conserved core

* Corresponding author.
E-mail address: celine.brochier-armanet@univ-lyon1.fr (C. Brochier-Armanet).
[1] The two authors have equally contributed to the work.

https://doi.org/10.1016/j.ympev.2018.04.011
Received 17 October 2017; Received in revised form 12 March 2018; Accepted 9 April 2018
Available online 21 April 2018
1055-7903/

genes have revealed a close relationship between *Halobacteria* and methanogen Class II, a group encompassing *Methanomicrobiales, Methanosarcinales* and *Methanocellales*, and indicated that *Halobacteria* could derive from a methanogenic ancestor (Forterre et al., 2002). However, the identity of the closest relative of *Halobacteria* remains debated (Supplementary Table S1). In fact, recently published large scale phylogenomic analyses supported *Halobacteria* as the sister-lineage of the whole methanogen Class II lineage (Gao and Gupta, 2007; Nelson-Sathi et al., 2012; Wolf et al., 2012; Yutin et al., 2012), of *Methanomicrobiales* (Brochier-Armanet et al., 2011; Petitjean et al., 2015; Raymann et al., 2015; Williams et al., 2017), or of *Methanocellales* (Becker et al., 2014; Petitjean et al., 2014; Adam et al., 2017). Regarding *Nanohaloarchaea*, some studies suggested that they represent the sister-lineage of *Halobacteria* (Narasingarao et al., 2012; Petitjean et al., 2014), while other analyses, based on different sets of markers, different methods and/or different taxonomic samplings, suggested instead that *Nanohaloarchaea* belong to the recently proposed DPANN super-phylum (Castelle et al., 2015; Rinke et al., 2013; Williams et al., 2017; Williams and Embley, 2014) (Supplementary Table S1). This super-phylum, distinct from the two other main archaeal lineages (i.e. the *Euryarchaeota* and the *Thaumarchaeota-Aigarchaeota-Crenarchaeota-Korarchaeota* (TACK) super-phylum) encompasses diverse, fast evolving, and possibly nanosized archaea (e.g. *Diapherotrites, Parvarchaeota*, *Micrarchaeota, Aenigmarchaeota, Nanoarchaeota*, *Woesearchaeota*, *Pacearchaeota*) and was proposed to represent the first diverging lineage within *Archaea*. Other studies, such as the analysis performed by Raymann et al (2014) challenges the DPANN hypothesis, as in this study nanosized lineages do not form a monophyletic group due to the grouping of *Micrarchaeum acidiphilum* with *Thermococcales* and *Methanomada*, while other nanosized lineages form a monophyletic group nested within *Euryarchaeota* and do not represent a lineage separated from TACK and *Euryarchaeota* as postulated by the DPANN hypothesis.

Elucidating the precise position of *Halobacteria* and *Nanohaloarchaea* is particularly challenging because their proteomes harbor atypical amino acid compositions as a consequence of their extremophilic lifestyle. This can generate a compositional signal that may conflict with and dominate over the phylogenetic signal (Jeffroy et al., 2006), and lead to artifactual tree reconstructions where distant sequences with similar compositions are clustered together (Delsuc et al., 2005; Woese et al., 1991). Another source of bias could be linked to the fast evolutionary rate of nanohaloarchaeal and halobacterial proteomes highlighted by their very long branches in phylogenetic trees compared to other archaeal lineages (Narasingarao et al., 2012; Petitjean et al., 2014). The phylogenetic position of fast-evolving species and long branches is particularly difficult to determine because differences in evolutionary rates among lineages can generate a rate signal that may blur the phylogenetic signal (Jeffroy et al., 2006) and cause tree reconstruction artifacts such as the long branch attraction (LBA) (Felsenstein, 1978). This well-known tree reconstruction artifact tends to group fast-evolving sequences/long branches and slow-evolving sequences/short branches in different parts of phylogenetic trees when the rate signal dominates over the phylogenetic signal (Jeffroy et al., 2006). Accordingly, we may wonder to what extent the conflicting positions observed for *Nanohaloarchaea* and *Halobacteria* are the consequence of tree reconstruction artifacts and if it is possible to overcome them.

To address this issue, we performed an in-depth phylogenomic analysis designed to limit the impact of the non-phylogenetic signal on phylogenetic inferences. We showed that *Nanohaloarchaea* and *Halobacteria* group robustly with *Methanocellales* and *Methanomicrobiales*, respectively, meaning that they derive from two distinct but related methanogen Class II ancestors. This implies that adaptation to very high salinity occurred at least twice in *Archaea*, and that the phenotypical similarities of *Nanohaloarchaea* and *Halobacteria* likely result from convergent evolutionary processes, possibly accompanied by horizontal gene transfers. Finally, our results indicate also that the grouping of *Nanohaloarchaea* with DPANN lineages is likely the consequence of a tree reconstruction artifact, challenging the existence of this candidate super-phylum.

2. Materials and methods

2.1. Dataset assembly

155 complete (or nearly complete) proteomes of 102 *Halobacteria*, 3 *Methanocellales*, 12 *Methanomicrobiales*, 15 *Methanosarcinales*, 3 *Nanohaloarchaea,* 1 ANME-I, 7 *Archeaoglobales* and 12 *Diaforarchaea* were retrieved at the NCBI (www.ncbi.nlm.nih.gov) and gathered in a local database (Supplementary Table S2). Pairwise comparisons of the corresponding 539,902 protein sequences were performed with BLASTP version 2.2.26 (Altschul et al., 1997) (default parameters, excepted the filter of low complexity regions that was turned off) and used to assemble homologous protein families with SILIX version 1.2.9 (Miele et al., 2011). More precisely, protein sequences displaying more than 45% of identity and 80% of sequence coverage were gathered in the same family. The inferred protein families were refined using HIFIX version 1.0.5 (Miele et al., 2012), which performs a three-step high-quality sequence clustering guided by network topology and multiple alignment likelihood. This led to the assembly of 108,007 protein families. Among these, 106,688 presented a narrow taxonomic distribution (i.e. being present in less than 85 proteomes) or corresponded to multigenic protein families (i.e. containing more than 200 sequences) with many paralogues and complex evolutionary histories. The remaining 1319 families were accurately aligned with MAFFT version 7.215 using the L-INS-I option (Kazutaka Katoh, 2013). The resulting multiple alignments were trimmed using BMGE version 1.12 with the BLOSUM45 substitution matrix (Criscuolo and Gribaldo, 2010). Maximum likelihood (ML) phylogenies were inferred with PhyML version 3.1 (Guindon et al., 2010) with the Le and Gascuel (LG) evolutionary model (Le and Gascuel, 2008) and a gamma distribution with four site categories (G4) to model the heterogeneity of evolutionary rates across sites. NNI and SPR strategies were used to search for the optimal tree topology. The robustness of the inferred ML trees was estimated with the non-parametric bootstrap procedure implemented in PhyML (100 replicates of the original alignments). The visual inspection of the resulting trees revealed that 1004 protein families presented complex patterns of horizontal gene transfers, gene duplications and losses. In contrast, 315 recovered the monophyly of archaeal classes and orders. We used a semi-automatic procedure in order to control the delineation of these 315 protein families done by SILIX and HIFIX. More precisely, representative sequences of *Halobacteria*, *Diaforarchaea* and *Methanocellales* were selected in each family and used as seeds to query the local database with BLASTP (default parameters, excepted the filter of low complexity regions that was turned off and the max_target sequences, which was set to 500). Each BLASTP output was inspected in order to identify homologous proteins. These sequences were aligned and used to infer ML trees as described above. Careful examination of the resulting ML trees revealed that 57 of the 315 markers corresponded in fact to complexe multigenic families, for which the delineation performed by SILIX/HIFIX could be questionable, while 258 markers, corresponding mainly to single copy protein families, were accurately delineated (Supplementary Table S3). A few sequences showing evidence of punctual HGT among archaea orders or gene duplications were omitted from these 258 families at this step. Multiple alignments were built using MAFFT and trimmed with BMGE as described above.

2.2. Supermatrix construction

The trimmed alignments corresponding to the 258 protein families were combined to build various supermatrices. To test the impact of missing data on phylogenetic inference, different versions of these supermatrices were built by gathering protein families present in more

than 95% or 70% of the studied proteomes (supplementary Table S4). Supermatrices FMETHA gathered sequences from methanogen Class II (i.e. *Methanocellales, Methanomicrobiales, Methanosarcinales*), ANME-I, *Archaeoglobales* and *Diaforarchaea*; FHALO contained in addition sequences from *Halobacteria*; FNANO gathered sequences from all taxa (including the three nanohaloarchaea) excepted *Halobacteria*; while FNANOHALO gathered sequences from all taxa (Supplementary Table S4).

2.3. Phylogenetic analyses of the supermatrices

ML trees were inferred with PhyML using the best-suited evolutionary models identified with IQ-TREE (BIC criteria) (Nguyen et al., 2015) and with the PMSF + G4 model implemented in IQ-TREE (Nguyen et al., 2015). Bayesian inferences were performed with PhyloBayes version 4.1 (Lartillot et al., 2009) with the CAT + GTR + G4 model, which allows the amino acid replacement patterns at different sites of a protein alignment to be described by distinct substitution processes. Two chains were run in parallel for at least 10,000 cycles. The first 1,500 trees were discarded as "burnin" and one out of two of the remaining trees from each chain was sampled to test for convergence (maxdiff < 0.3) and to compute 50% majority rule consensus trees. All trees in this paper were drawn with iTOL version 3 (Letunic and Bork, 2016).

2.4. Reducing the impact of rate signal

The Slow-Fast (SF) method is an effective way to limit the influence of rate signal and thus to reduce long branch attraction artifact, by progressively removing the fastest evolving positions from large multiple alignments (Brinkmann and Philippe, 1999; Delsuc et al., 2005). At each step, phylogenetic trees are inferred. This allows testing the impact of the removed positions on important branches and can highlight potential tree reconstruction artifacts. The S-F approach was applied on the FHALO, FNANO and FNANOHALO supermatrices using SLOW-FASTER (Kostka et al., 2008). To avoid biases in the estimation of evolutionary rates at each position due to missing data, the S-F method was applied to the supermatrices built with markers present in at least 95% of the studied taxa (Supplementary Table S4). Furthermore, to avoid biases due to unbalanced taxonomic sampling among lineages, we kept only three to seven representative sequences for each archaeal class/order. As a consequence, the ANME-I lineage,represented by a single proteome, was not included in this analysis.

Amino acid recoding is another way to reduce the impact of rate signal in protein datasets (Delsuc et al., 2005). This consists in masking the amino acid substitutions that are most frequently observed in protein sequences. Two frequently used types of amino acid recoding schemes (dayhoff4 and dayhoff6) were applied. The four- and six-dayhoff's amino acid families corresponded to [(A,G,P,S,T) (D,E,N,Q) (H,K,R) (F,Y,W,I,L,M,V)] with cysteine treated as missing data (C = ?) and to [(A,G,P,S,T) (D,E,N,Q) (H,K,R) (F,Y,W) (I,L,M,V) (C)], respectively.

2.5. Reducing the impact of compositional signal

Multivariate analyses of proteome amino acid composition of FHALO, FNANO and FNANOHALO supermatrices were conducted in R version 3.3.0 (R Core Team, 2014), using the ade4 package (Dray and Dufour, 2007). The amino acid composition homogeneity of the supermatrices was tested using IQ-TREE (Nguyen et al., 2015). The sites responsible for amino acid composition heterogeneity were identified and removed with BMGE version 1.12 (option -s FAST) (Criscuolo and Gribaldo, 2010).

3. Results

Determining the phylogenetic position of *Nanohaloarchaea* and *Halobacteria* is challenging because of their fast evolutionary rates and the atypical amino composition of their proteomes. To tackle this issue, we designed a specific strategy maximizing the number of analyzed markers and limiting the impact of rate and compositional signals. While previous studies focused on the whole archaeal domain (Supplementary Table 1), our analysis focused on the euryarchaeotal part of the tree encompassing *Diaforarchaea*, *Archaeoglobales*, ANME-I, methanogen Class II (i.e. *Methanomicrobiales*, *Methanocellales* and *Methanosarcinales*), and *Halobacteria; Diaforarchaea* representing the first diverging lineage (Castelle et al., 2015; Petitjean et al., 2015, 2014; Adam et al., 2017). This allowed to work at a smaller evolutionary scale, to use more phylogenetic markers (and thus more amino acid positions) for phylogenetic analyses, and to consider large taxonomic samplings for these groups. Importantly this reduced also the risk of tree reconstruction artifacts and biases introduced by divergent archaeal lineages that are not directly linked to this issue. In this context, two different scenarios were expected for *Nanohaloarchaea*. If they are related to *Halobacteria*, any methanogen Class II lineage, ANME-I or even *Archaeoglobales,* they should branch on the stem of the corresponding lineage in the reconstructed trees. Alternatively, if they occupy a deeper position in the archaeal tree, as expected if they are member of the DPANN super-phylum, they should branch on the stem of *Diaforarchaea*.

3.1. Identification of a conserved core of genes

The comparison of 155 proteomes from *Halobacteria*, ANME-I, methanogen Class II, and their closest relatives: *Archaeoglobales* and *Diaforarchaea*, and from *Nanohaloarchaea* (Supplementary Table S2) led to the delineation of 108,007 families, among which 258 presented a broad taxonomic distribution and no or very few evidences of horizontal gene transfers (HGT) among these lineages and/or gene duplications (Supplementary Table S3). As expected, due to the small size of their proteomes, the three *Nanohaloarchaea* were less represented than other species, being present altogether in only 68 out of 258 protein families (Supplementary Fig. S1). A few proteomes deduced from incomplete and/or badly annotated genome sequences were also underrepresented in protein families (Supplementary Fig. S1A).

According to arCOG annotations (Makarova et al., 2015), half of the 258 markers (i.e. 128 markers) were involved in information storage and processing, while 81 were involved in metabolism, 12 in cellular processes and signaling, and 37 corresponded to conserved proteins of unknown function (Supplementary Fig. S1B). This functional diversity is important because the phylogenetic signal carried by markers that are functionally linked might reflect the history of the corresponding process, and not that of the organisms (Philippe, 2000).

3.2. Nanohaloarchaea are related neither to Halobacteria nor to methanogen Class II

As a first step, we reconstructed a reference phylogeny of the three orders of methanogen Class II, ANME-I and *Archaeoglobales*, rooted by *Diaforarchaea* in order to have a robust framework to study the phylogenetic position of *Nanohaloarchaea* and *Halobacteria*. Such a reference phylogeny may help to detect potential tree reconstruction artifacts resulting from the introduction of *Nanohaloarchaea* and *Halobacteria* in subsequent analyses. We combined the protein families present in more than 95% of *Methanosarcinales*, *Methanomicrobiales*, *Methanocellales*, ANME-I, *Archaeoglobales* and *Diaforarchaea*. The Maximum Likelihood (ML) trees inferred with PHYML with the best evolutionary model according to IQ-TREE and with the LG + PMSF + I + G4 model were very well resolved (most bootstrap values (BV) > 90%, Supplementary Figs. S2D and S3D) indicating that the

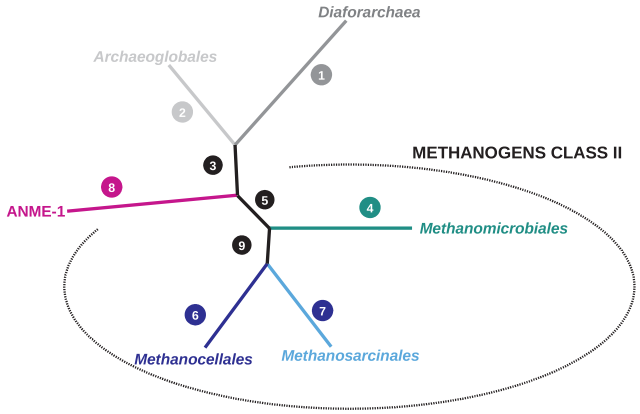

Fig. 1. Alternative branching positions for *Nanohaloarchaea* and *Halobacteria*. The tree represents the phylogenetic relationships among *Diaforarchaea*, *Archaeoglobales*, ANME-I, and methanogen Class II. Worth noticing, among these lineages, *Diaforarchaea* represent the most external one and can be used as outgroup. According to this tree, *Nanohaloarchaea* and *Halobacteria* can branch at nine different positions (circles 1-9).

corresponding supermatrix, and thus the protein markers used, contained a strong signal. Regarding relationships among methanogen Class II, *Methanocellales* were closely related to *Methanosarcinales* (BV = 100%), in agreement with previous works (Raymann et al., 2015; Sakai et al., 2008). This topology indicated that nine branching positions are possible for *Nanohaloarchaea* and *Halobacteria* (Figs. 1 and 2).

The protein markers present in the three *Nanohaloarchaea* and in more than 70% of the 155 studied proteomes (including *Halobacteria*) were combined to build the FNANOHALO$_{70}$ supermatrix (Supplementary Table S4). The corresponding ML tree inferred with the LG + I + F + G4 model, as suggested by IQ-TREE, was overall well resolved (Fig. 3 and Supplementary Fig. S2A). The relationships among methanogen Class II and especially, the sistership between *Methanocellales* and *Methanosarcinales*, were recovered, indicating that adding *Nanohaloarchaea* and *Halobacteria* did not distort the relationships among methanogen Class II. Regarding extreme halophiles, *Nanohaloarchaea* and *Halobacteria* did not group together. In fact, *Halobacteria* clustered with *Methanomicrobiales* (BV = 99%), corresponding to position number 4 on Figs. 1 and 2, in agreement with recent studies (Brochier-Armanet et al., 2011; Petitjean et al., 2015; Raymann et al.,

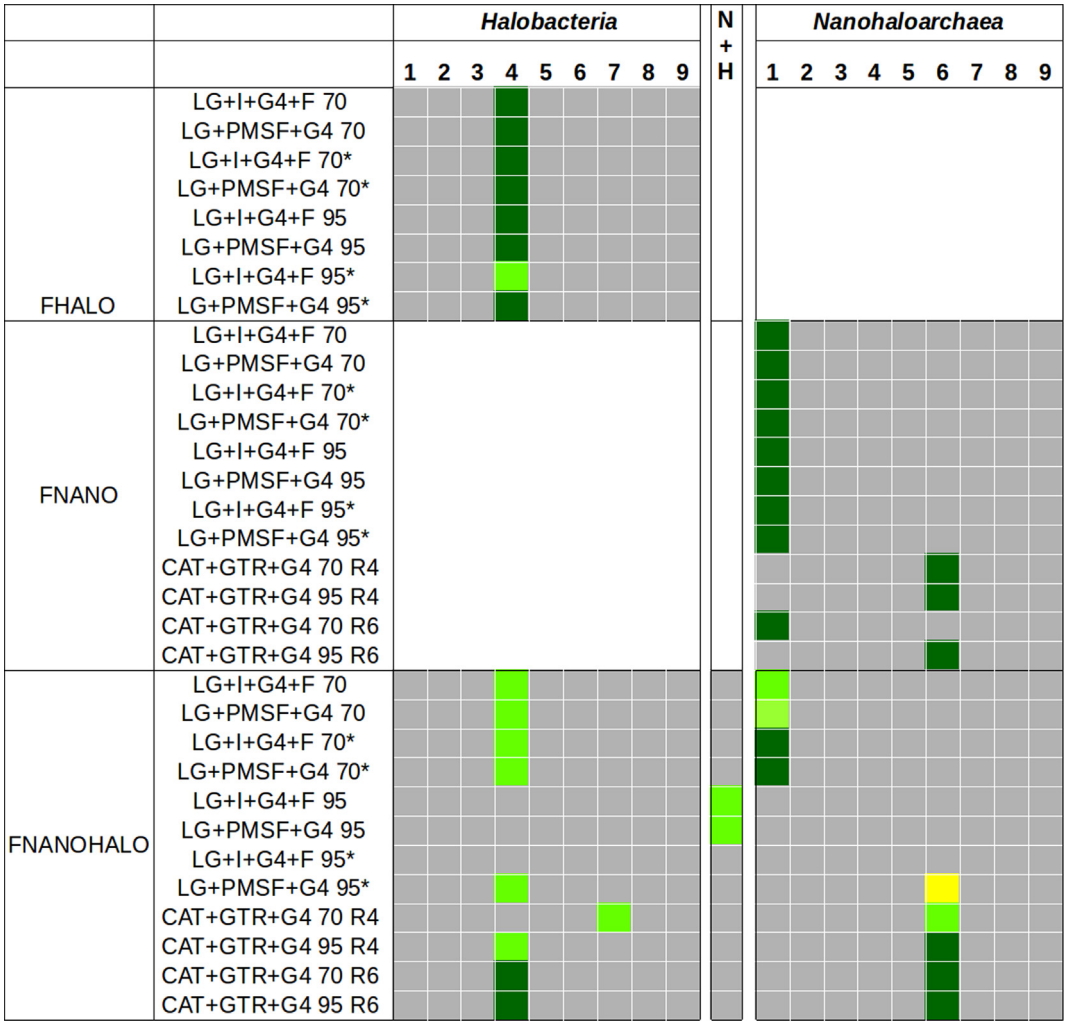

Fig. 2. Phylogenetic relationships of *Nanohaloarchaea* and *Halobacteria*. This figure summarizes the supports (BV and PP) observed for the branching of *Nanohaloarchaea* and *Halobacteria* in ML and BI phylogenetic trees inferred with the FHALO, FNANO and FNANOHALO supermatrices. Dark green square: maximal support (BV = 100%/PP = 1), light green square: strong support (100% > BV ≥ 85% and 1 > PP ≥ 0.95), yellow square: moderate support (85% > BV > 75%), grey square: no or weak support (75% ≥ BV and 0.90 > PP). (N + H) corresponds to the grouping of *Nanohaloarchaea* and *Halobacteria*, while the other positions correspond to the nine possible branching points of *Nanohaloarchaea* and *Halobacteria* in agreement with Fig. 1. Stars indicate that amino acid positions responsible of compositional biases have been removed with BMGE. "R4" and "R6" indicate that amino acids were recoded according to the dayhoff4 and dayhoff6 recoding schemes, respectively.

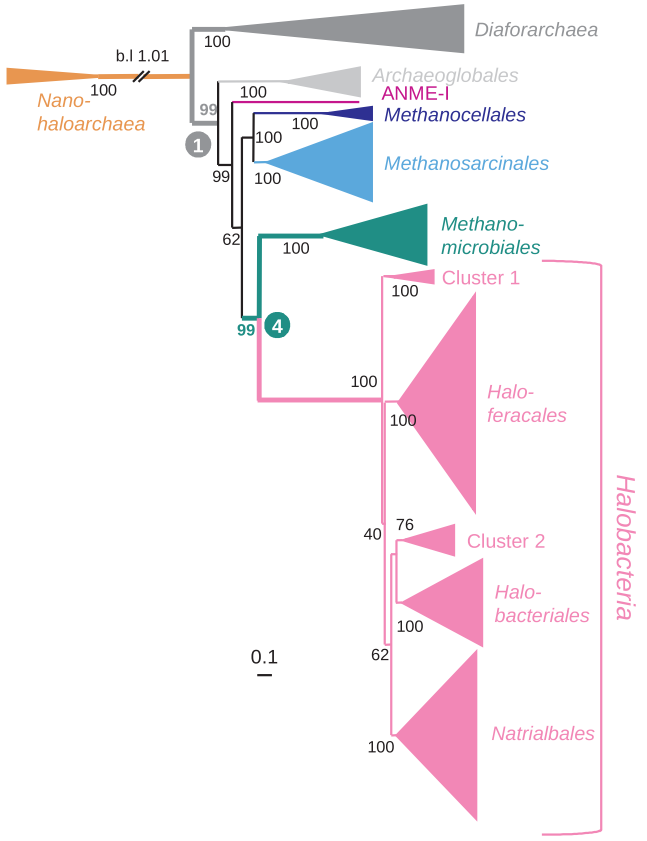

Fig. 3. ML phylogeny inferred with the FNANOHALO$_{70}$ supermatrix (18,309 amino acid positions, 155 sequences). The tree was inferred with PHYML 3.1 using the LG + F+I+Γ4 model as suggested by IQ-TREE. The scale bar corresponds to the average number of substitutions per site. Numbers at nodes correspond to bootstrap supports (100 replicates of the original dataset). For clarity, the branch leading to *Nanohaloarchaea* has been shortened, and the real length is indicated as b.l. Numbers in colored circles refer to Fig. 1.

2015), whereas *Nanohaloarchaea* branched on the stem of *Diaforarchaea* (BV = 99%), corresponding to position number 1 (Figs. 1 and 2). This suggested that *Nanohaloarchaea* are not related to any methanogen Class II lineage, ANME-I or *Archaeoglobales*. Their position is compatible with a deep-branching within *Archaea*, as postulated by the DPANN hypothesis (Castelle et al., 2015; Rinke et al., 2013; Williams et al., 2017; Williams and Embley, 2014). The grouping of *Halobacteria* with *Methanomicrobiales* and the branching of *Nanohaloarchaea* on the stem of *Diaforarchaea* were supported by ML analysis of the FHALO$_{70}$ and FNANO$_{70}$ supermatrices (Fig. 2 and Supplementary Fig. S2B–C), meaning that the phylogenetic position of *Nanohaloarchaea* observed in the FNANOHALO$_{70}$ tree was not impacted by the sequences from *Halobacteria* and reciprocally.

Most evolutionary models, as the LG model, postulate that all sites in a given multiple alignment evolve according to the same evolutionary process, meaning that a single substitutional matrix can be applied at each amino acid position. Yet, this assumption is frequently ruled out by biological data. Accordingly, site-heterogeneous models, such as the Bayesian CAT profile mixture model and its ML counterparts, the C10 to C60 models, have been developed (Lartillot and Philippe, 2004; Quang et al., 2008) and were shown to over-perform classical site-homogeneous models (Delsuc et al., 2005; Lartillot et al., 2007). However, these profile mixture models are time and memory consuming, and impose a very heavy burden in calculation time. In this study, using the CAT profile mixture model in a Bayesian (BI) framework failed to converge (not shown). Thus, we used the Posterior Mean Site Frequency (PMSF) model, a rapid and efficient approximation to CAT and C10 to C60 models. All trees inferred with the LG + PMSF + G4 model were consistent with those inferred with the LG model (Fig. 2 and Supplementary Figs. S3A–C), suggesting that the observed branching positions of *Halobacteria* and *Nanohaloarchaea* are not the consequence of a tree reconstruction artefact introduced by the use of a site-homogeneous model.

3.3. The phylogenetic position of Nanohaloarchaea is not biased by missing data or amino acid compositional signal

Recent studies have shown that large sparse supermatrices could be more sensitive to phylogenetic artifacts than smaller but less incomplete ones (Roure et al., 2013). The impact of missing data on the phylogenetic position of *Halobacteria* and *Nanohaloarchaea* was investigated by comparing the ML trees inferred with markers present in more than 70% and 95% of the studied proteomes (FNANOHALO$_{70}$, FHALO$_{70}$, and FNANO$_{70}$ versus FNANOHALO$_{95}$, FHALO$_{95}$ and FNANO$_{95}$) (Supplementary Table S4). ML trees inferred with the LG + I + F + G4 and the LG + PMSF + G4 models on supermatrices gathering markers present in more than 95% of the studied proteomes were consistent with those inferred with all markers (Fig. 2, Supplementary Figs. S2A–C, S2E-G, S3A–C and S3E–G). In particular, the FHALO$_{95}$ and FNANO$_{95}$ ML trees, as the FHALO$_{70}$, FNANO$_{70}$ an FNANOHALO$_{70}$ ML trees, supported the grouping of *Halobacteria* with *Methanomicrobiales* and the branching of *Nanohaloarchaea* on the stem of *Diaforarchaea* (Fig. 2). In contrast, the ML trees inferred with the FNANOHALO$_{95}$ supermatrix supported the grouping of *Nanohaloarchaea* and *Halobacteria* (BV = 85% and 98%, Fig. 2 and Supplementary Figs. S2E and S3E). However, this relationship was probably artefactual because it was not recovered when the amino acid positions responsible of amino acid composition heterogeneity were removed from the analysis (see below).

Sequence compositional heterogeneity is another important source of bias in molecular phylogeny (Jeffroy et al., 2006). This is due to the fact that most evolutionary models used in molecular phylogenetics (including the LG model) assume that the evolutionary process is at equilibrium from the root to the leaves. As a consequence, the overall sequence composition is not expected to change through time and thus studied sequences should harbor similar compositions. This assumption is rarely verified by biological sequences, a situation that leads often to the artefactual grouping of sequences with similar compositions, irrespectively of their true evolutionary relationships (Delsuc et al., 2005; Jeffroy et al., 2006; Woese et al., 1991; Ramulu et al. 2014). The amino acid composition homogeneity of sequences composing each supermatrix was tested using IQ-TREE (Nguyen et al., 2015). As expected, most of the sequences failed to pass the χ^2 test (Supplementary Table S4). Correspondence analyses of the amino acid composition of the supermatrices showed that the genomic G + C content of genomes explained most of the observed variation, while the impact of optimal growth temperature and optimal growth salinity was less pronounced (Supplementary Table S5). To test the impact of compositional biases on the inferred phylogenies, we removed from the supermatrices the positions displaying highest compositional biases and thus responsible for the observed amino acid composition heterogeneity among sequences. The removal of these positions did not impact the phylogenetic position of extreme halophilic archaea in ML trees inferred with the LG + I + F + G4 or the LG + PMSF + G4 models. The clustering of *Halobacteria* with *Methanomicrobiales* and the branching of *Nanohaloarchaea* on the stem of *Diaforarchaea* were recovered again with high supports (Fig. 2 and Supplementary Figs. S2H–J, S2L-M, S3H-J and S3L-M). Interestingly, the only exceptions concerned the FNANOHALO$_{95}$ ML tree which was not resolved when the LG + I + F + G4 was used and that supported the grouping of *Nanohaloarchaea* and *Methanocellales* albeit with a moderate support when the LG + PMSF + G4 model was used (BV = 79%, Fig. 2 and Supplementary Figs. S2K and S3K).

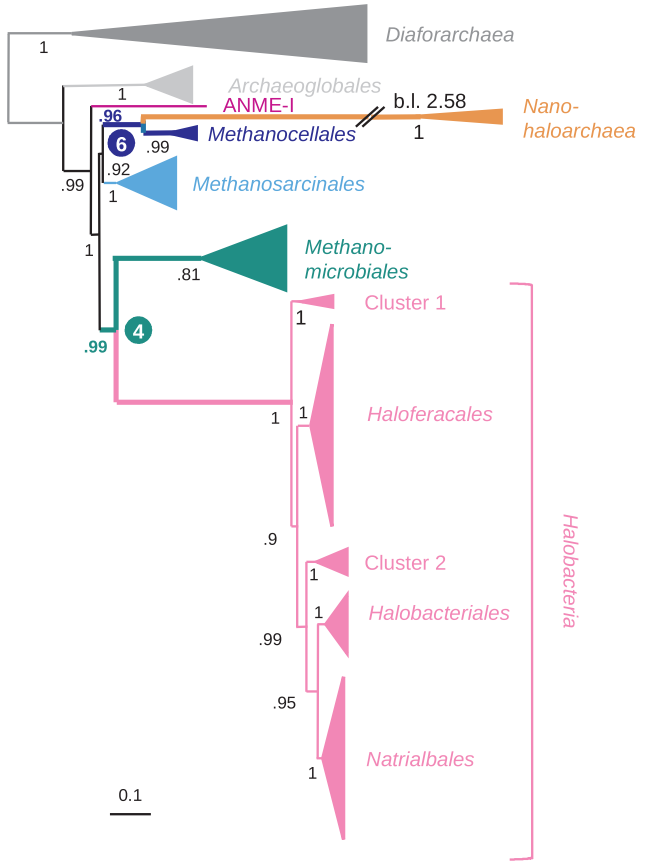

Fig. 4. BI phylogeny inferred with the FNANOHALO$_{70}$ supermatrix recoded according to the Dayhoff4 scheme (18,309 amino acid positions, 155 sequences). The tree was inferred with PHYLOBAYES using the CAT + GTR + G4 model. The scale bar corresponds to the average number of substitutions per site. Numbers at nodes correspond to posterior probabilities. For clarity, the branch leading to *Nanohaloarchaea* has been shortened, and the real length is indicated as b.l. Numbers in colored circles refer to Fig. 1.

3.4. The phylogenetic position of Nanohaloarchaea is the consequence of a LBA artifact

The evolutionary rate signal is a major cause of tree reconstruction artifacts such as the LBA (Delsuc et al., 2005; Felsenstein, 1978). This artifact is due to multiple substitutions occurring at the same site, a process which erases progressively the most ancient phylogenetic signal and results in the grouping of sequences according to their evolutionary rates in different parts of the inferred trees. To overcome this issue, dedicated methods have been developed.

Among them, the recoding of amino acids allows to hide substitutions among similar amino acids. This can reduce the impact of the rate signal because substitutions among similar amino acids occur more frequently than other substitutions. Thus, these are more prone to undergo multiple substitutions. To test the impact of the rate signal, we applied two different recoding schemes (dayhoff4 and dayhoff6) to the FNANOHALO$_{70}$, FNANOHALO$_{95}$, FNANO$_{70}$ and FNANO$_{95}$ supermatrices in a Bayesian framework. The inferred FNANOHALO$_{70}$ and FNANOHALO$_{95}$ BI trees strongly confirmed the sister-ship between *Halobacteria* and *Methanomicrobiales* (Figs. 2 and 4, and Supplementary Figs. S4A–E and S4G). Yet, surprisingly, *Nanohaloarchaea* branched with *Methanocellales* in seven out of the eight recoded FNANOHALO and FNANO supermatrices (Figs. 2 and 4, and Supplementary Fig. S4A–E and S4G), suggesting that the branching of *Nanohaloarchaea* on the stem of *Diaforarchaea* could result from a tree reconstruction artefact due to the rate signal. The only exception concerned BI tree inferred with the FNANO$_{70}$ supermatrix recoded in Dayhoff6 (Fig. 2 and Supplementary Fig. S4F) where *Nanohaloarchaea* branched on the stem of *Diaforarchaea*.

To test the hypothesis of a sister-ship between *Nanohaloarchaea* and *Methanocellales*, we used a second and independent approach aiming at limiting tree reconstruction artifacts resulting from the rate signal. This approach, called the Slow-Fast method (S-F), consists in the progressive removal of the fastest-evolving sites from large multiple alignments (Brinkmann and Philippe, 1999). The S-F method was shown to be very efficient to reduce tree reconstruction artifacts because the fastest evolving sites are the most susceptible to be impacted by multiple substitutions (Delsuc et al., 2005; Philippe, 2000). This approach allows monitoring the support associated to a given branch of a tree throughout the removal process and thus to determine if the corresponding relationship reflects the phylogenetic or the rate signal contained in the sequences (Delsuc et al., 2005). We applied the S-F method to the FNANOHALO$_{95}$, FNANO$_{95}$ and FHALO$_{95}$ supermatrices in ML and BI frameworks. The removal of the fastest-evolving sites did not impact the phylogenetic position of *Halobacteria*. In fact, the grouping of *Halobacteria* with *Methanomicrobiales* was strongly supported in ML and BI trees inferred with the FNANOHALO$_{95}$ and FHALO$_{95}$ S-F supermatrices (Fig. 5A–B and 5E-F and Supplementary Fig. S5). This suggested that this relationship was not the consequence of the rate signal. In sharp contrast, applying the S-F method to FNANOHALO$_{95}$ and FNANO$_{95}$ supermatrices showed that the branching of *Nanohaloarchaea* on the stem of *Diaforarchaea* was recovered in ML and BI trees inferred only when the fastest-evolving sites were included in the analysis, while a robust grouping with *Methanocellales* was observed when these sites were not considered (Fig. 5A–D and Supplementary Fig. S5).

Altogether, these results strongly suggested that the branching of *Nanohaloarchaea* on the stem of *Diaforarchaea* was the result of LBA artifact caused by the rate signal.

4. Discussion

The past few years have witnessed spectacular advances in genome sequencing methods. Applying these methods to environmental surveys has expanded the Tree of Life by disclosing a myriad of new major microbial lineages (Castelle et al., 2015; Hug et al., 2016; Rinke et al., 2013). Interestingly, many of them corresponded to very small organisms with reduced and often divergent genomes. This is for instance the case of the Candidate Phyla Radiation bacteria (Brown et al., 2015) or the DPANN super-phylum in *Archaea* (Rinke et al., 2013). The DPANN was challenged by several studies which suggested that the lineages composing this super-phylum are not related and have different origins. As an example, *Nanoarchaeota* were proposed to be the sister-lineage of *Thermococcales* (Brochier et al., 2005; Petitjean et al., 2014), *Parvarchaeota* and *Micrarchaeota* the closest relatives of *Diaforarchaea* (Petitjean et al., 2014), and *Nanohaloarchaea* the sister-group of *Halobacteria* (Narasingarao et al., 2012; Petitjean et al., 2014). The grouping of these lineages within the DPANN was then interpreted as the result of tree reconstruction artifacts.

Pinpointing the phylogenetic position of nanosized archaeal lineages is difficult because these organisms are fast-evolving, it could be unstable and very sensitive to LBA artifacts (Felsenstein, 1978; Quang et al., 2008). As an example, a recent study combining supermatrix and a supertree approaches showed that all DPANN lineages formed a monophyletic group when they were analyzed altogether, while when they were considered separately some of these lineages (including *Nanohaloarchaeota*) branched within *Euryarchaeota* (Williams et al. 2017). Furthermore, some DPANN proteomes harbor biased amino acid composition as a consequence of their extremophilic lifestyles. Thus, determining the phylogenetic position of nanosized archaeal lineages requires the design of specific protocols using cutting edge approaches to disentangle the various signals contained in their sequences. Here, we have investigated the phylogenetic position of

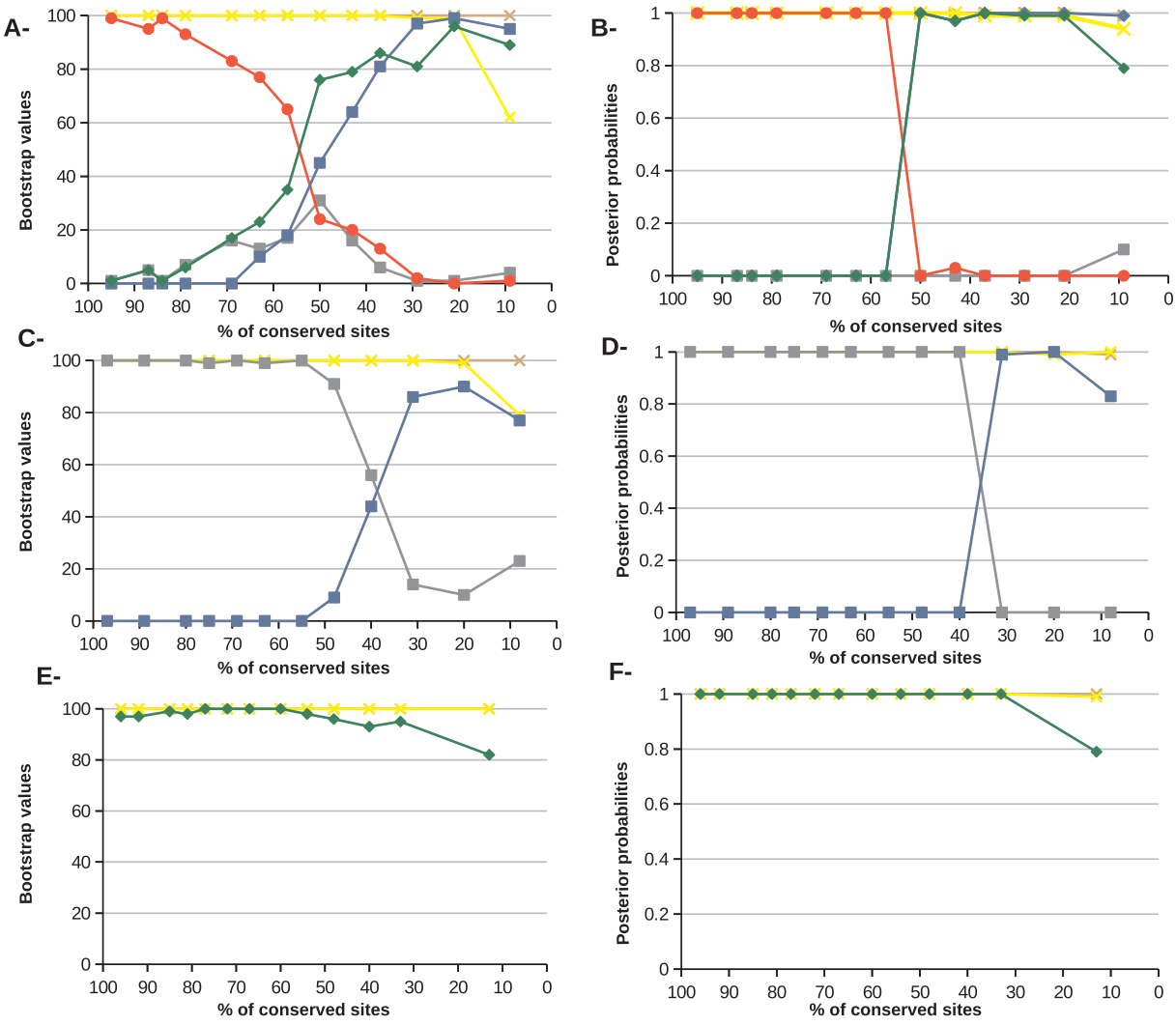

Fig. 5. Impact of removal of fast-evolving positions in the FNANOHALO$_{95}$, FNANO$_{95}$ and FHALO$_{95}$ supermatrices on the phylogenetic position of *Halobacteria* and *Nanohaloarchaea*. The removal of fastest-evolving sites proceeds from left to right on the x axes. The y axes correspond to BV (A, C, and E) or BI supports (B, D, and F). Green: support for the grouping of *Halobacteria* with *Methanomicrobiales*, red: *Nanohaloarchaea* with *Halobacteria*, blue: *Nanohaloarchaea* with *Methanocellales*, grey: *Nanohaloarchaea* on the stem of *Diaforarchaea*. Supports for the monophyly of *Methanosarcinales* (yellow) and *Methanomicrobiales* (brown) were used as positive control to visualize the amount of information contained in the supermatrices along the removal process of fast-evolving sites.

Nanohaloarchaea. By considering only this lineage of DPANN and focusing our analysis on the part of the euryarchaeotal tree that contains *Halobacteria* and methanogen Class II, we were able to assemble larger datasets of conserved markers and to use more intense taxonomic samplings compared to previous studies (supplementary Table S1). By using various methods allowing to decouple the different types of signal contained in protein sequences, we showed that the compositional signal did not significantly impact the phylogenetic positions of *Nanohaloarchaea* and *Halobacteria*, while the rate signal had a major impact on the phylogenetic position of *Nanohaloarchaea*. In fact, two independent methods allowing to reduce the impact of multiple substitutions on phylogenetic inferences, the removal of the fastest evolving sites (the S-F method) and the recoding of amino acids, provided consistent results supporting the grouping of *Halobacteria* with *Methanomicrobiales* and of *Nanohaloarchaea* with *Methanocellales*. The robust and recurrent grouping of *Halobacteria* and methanogen Class II in many studies, suggested that they could represent a new super-class that we propose to call *'Stenosarchaea'* (from the Greek stenós, meaning close/joint). Our data strongly suggested that *Nanohaloarchaea* are also part of the *Stenosarchaea*, and more precisely, that they represent the closest relatives of *Methanocellales*.

The robust grouping of *Nanohaloarchaea* with *Methanocellales* rules out alternative hypotheses for the branching of *Nanohaloarchaea*, and in particular a branching outside of *Stenosarchaea*, as expected if *Nanohaloarchaea* were part of the candidate DPANN super-phylum. As a consequence, this challenges the existence of this group as it is currently described and questions to what extent similar artifacts could also impact the position of the other DPANN lineages, which are all fast-evolving. Testing this hypothesis would require accurate and dedicated analyses, each focused on one single nanosized lineage. Finally, the branching of *Nanohaloarchaea* within *Stenosarchaea* shed a new light on the origin of atypical DNA primases found in nanosized archaeal lineages. In fact, previous studies have revealed the presence of atypical DNA primases, corresponding to the fusion of the catalytic domain of PriS and the Fe–S cluster-binding domain of PriL, in DPANN lineages, excepted in *Micrarchaeota* and *Diapherotrites* that encoded canonical versions of PriL and PriS (Adam et al. 2017; Raymann et al. 2014). The hypothesis of independent origins of these atypical DNA primases was discarded because these proteins are closely related at the sequence level (Raymann et al. 2014). This left open two possibilities: either independent HGT have led to the replacement of the canonical DNA primases by the fused DNA primase in some nanosized archaeal lineages or lineages harboring a fused DNA primase share a common ancestry, consistently with the DPANN hypothesis (Raymann et al.

2014). According to the second hypothesis, in *Micrarchaeota* and *Diapherotrites,* the fused DNA primase has been secondarily replaced by a canonical DNA Primase via HGT. Our study clearly positioned *Nanohaloarchaea* as a member of the *Stenosarchaea* and not as a member of the DPANN super-phylum, indicating that their atypical DNA primase has been likely acquired secondarily by HGT. This indicates also that the presence of a fused DNA primase can not be considered as a synapomorphy of the DPANN super-phylum and as an argument in favor of the DPANN hypothesis.

The grouping of *Nanohaloarchaea* with *Methanocellales* and that of *Halobacteria* with *Methanomicrobiales* within *Stenosarchaea* has major implications and opens new perspectives. First, it implies that adaptation to extreme high salt concentrations occurred at least twice independently during the evolution of *Archaea*. It also implies that both lineages derive from two distinct but related methanogen ancestors. This was consistent with the fact that most halophilic or halotolerant archaeal lineages that can survive at high salt concentrations belong to methanogen Class II (Oren, 2014). Thus, the phenotypic properties shared by *Nanohaloarchaea* and *Halobacteria* should be interpreted as the consequence of a convergent evolution that could have been facilitated by the possible existence of favorable genomic and phenotypic backgrounds in methanogen Class II lineages. In that context, it would be interesting to reevaluate the evolutionary history of these lineages, and the role played by HGT in the emergence of *Halobacteria* and *Nanohaloarchaea* from methanogenic ancestors.

Acknowledgements

This work was supported by Investissement d'Avenir (grant number ANR-10-BINF-01-01) and Agence Nationale de la Recherche (grant number ANR-16-CE02-0005) grants. M.A. held a doctoral fellowship from the Région Rhône-Alpes–ARC 1 Santé. M.L. held a doctoral fellowship from the French Ministère de l'Enseignement Supérieur et de la Recherche. N.T. held a fellowship from the Agence Nationale de la Recherche (grant number ANR-12-BSV7-0003). A.O. held a doctoral fellowship from the Investissement d'Avenir project (grant number ANR-10-BINF-01-01). We warmly acknowledge Simonetta Gribaldo and Laura Eme for stimulating discussions and helpful comments, and Thomas Bigot for technical help. We thank also the PRABI (Pôle Rhône-Alpes de Bioinformatique) for providing computing facilities.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.ympev.2018.04.011.

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 J. http://dx.doi.org/10.1038/ismej.2017.122.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.

Ban, N., Nissen, P., Hansen, J., Moore, P.B., Steitz, T.A., 2000. The complete atomic structure of the large ribosomal subunit at 2.4 A resolution. Science 80 (289), 905–920.

Becker, E.A., Seitzer, P.M., Tritt, A., Larsen, D., Krusor, M., Yao, A.I., Wu, D., Madern, D., Eisen, J.A., Darling, A.E., Facciotti, M.T., 2014. Phylogenetically driven sequencing of extremely halophilic archaea reveals strategies for static and dynamic osmo-response. PLoS Genet. 10, e1004784.

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

Britton, K.L., Baker, P.J., Fisher, M., Ruzheinikov, S., Gilmour, D.J., Bonete, M.-J., Ferrer, J., Pire, C., Esclapez, J., Rice, D.W., 2006. Analysis of protein solvent interactions in glucose dehydrogenase from the extreme halophile Haloferax mediterranei. Proc. Natl. Acad. Sci. USA 103, 4846–4851.

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

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

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

Castelle, C.J., Wrighton, K.C., Thomas, B.C., Hug, L.A., Brown, C.T., Wilkins, M.J., Frischkorn, K.R., Tringe, S.G., Singh, A., Markillie, L.M., Taylor, R.C., Williams, K.H., Banfield, J.F., 2015. Genomic expansion of domain archaea highlights roles for organisms from new phyla in anaerobic carbon cycling. Curr. Biol. 25, 690–701.

Criscuolo, A., Gribaldo, S., 2010. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. BMC Evol Biol. 10, 210. http://dx.doi.org/10.1186/1471-2148-10-210.

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

Dray, S., Dufour, A.B., 2007. The ade4 package: implementing the duality diagram for ecologists. J. Stat. Softw. 22, 1–20.

Eme, L., Doolittle, W.F., 2015. Archaea. Curr. Biol. 25, R851–R855.

Felsenstein, J., 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. 27, 401.

Forterre, P., Brochier, C., Philippe, H., 2002. Evolution of the Archaea. Theor. Popul. Biol. 61, 409–422.

Gao, B., Gupta, R.S., 2007. Phylogenomic analysis of proteins that are distinctive of Archaea and its main subgroups and the origin of methanogenesis. BMC Genomics 8, 86.

Ghai, R., Pašić, L., Fernández, A.B., Martin-Cuadrado, A.-B., Mizuno, C.M., McMahon, K.D., Papke, R.T., Stepanauskas, R., Rodriguez-Brito, B., Rohwer, F., Sánchez-Porro, C., Ventosa, A., Rodríguez-Valera, F., 2011. New abundant microbial groups in aquatic hypersaline environments. Sci. Rep. 1, 739–751.

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

Hug, L.A., Baker, B.J., Anantharaman, K., Brown, C.T., Probst, A.J., Castelle, C.J., Butterfield, C.N., Hernsdorf, A.W., Amano, Y., Ise, K., Suzuki, Y., Dudek, N., Relman, D.A., Finstad, K.M., Amundson, R., Thomas, B.C., Banfield, J.F., 2016. A new view of the tree of life. Nat. Microbiol. 1, 16048.

Jeffroy, O., Brinkmann, H., Delsuc, F., Philippe, H., 2006. Phylogenomics: the beginning of incongruence? Trends Genet. 22, 225–231.

Kazutaka Katoh, D.M.S., 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability.

Kostka, M., Uzlikova, M., Cepicka, I., Flegr, J., 2008. SlowFaster, a user-friendly program for slow-fast analysis and its application on phylogeny of Blastocystis. BMC Bioinform. 9, 341.

Lartillot, N., Brinkmann, H., Philippe, H., 2007. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. BMC Evol. Biol. 7 (Suppl 1), S4.

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

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

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

Letunic, I., Bork, P., 2016. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res.

Makarova, K.S., Wolf, Y.I., Koonin, E.V., 2015. Archaeal clusters of orthologous genes (arCOGs): an update and application for analysis of shared features between thermococcales, methanococcales, and methanobacteriales. Life (Basel, Switzerland) 5, 818–840.

Miele, V., Penel, S., Daubin, V., Picard, F., Kahn, D., Duret, L., 2012. High-quality sequence clustering guided by network topology and multiple alignment likelihood. Bioinformatics 28, 1078–1085.

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

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

Nelson-Sathi, S., Dagan, T., Landan, G., Janssen, A., Steel, M., McInerney, J.O., Deppenmeier, U., Martin, W.F., 2012. Acquisition of 1,000 eubacterial genes physiologically transformed a methanogen at the origin of Haloarchaea. Proc. Natl. Acad. Sci. USA 109, 20537–20542.

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. Mol. Biol. Evol. 32, 268–274.

Oren, A., 2014. Taxonomy of halophilic Archaea: current status and future challenges. Extremophiles 18, 825–834.

Oren, A., 2008. Microbial life at high salt concentrations: phylogenetic and metabolic diversity. Saline Syst. 4, 2.

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 Biol. Evol. 7, 191–204.

Petitjean, C., Deschamps, P., López-García, P., Moreira, D., Brochier-Armanet, C., 2015. Extending the conserved phylogenetic core of archaea disentangles the evolution of the third domain of life. Mol. Biol. Evol. 32, 1242–1254.

Philippe, H., 2000. Opinion: long branch attraction and protist phylogeny. Protist 151, 307–316.

Quang, L.S., Gascuel, O., Lartillot, N., 2008. Empirical profile mixture models for phylogenetic reconstruction. Bioinformatics 24, 2317–2323.
Ramulu, H.G., Groussin, M., Talla, E., Planel, R., Daubin, V., Brochier-Armanet, C., 2014. Ribosomal proteins: toward a next generation standard for prokaryotic systematics? Mol. Phylogenet. Evol. 75, 103–117.
R Core Team, 2014. R: A Language and Environment for Statistical Computing. R Found. Stat. Comput. Vienna, Austria 2014.
Raymann, K., Forterre, P., Brochier-Armanet, C., Gribaldo, S., 2014. Global phylogenomic analysis disentangles the complex evolutionary history of DNA replication in Archaea. Genome Biol. Evol. 6, 192–212. http://dx.doi.org/10.1093/gbe/evu004.
Raymann, K., Brochier-Armanet, C., Gribaldo, S., 2015. The two-domain tree of life is linked to a new root for the Archaea. Proc. Natl. Acad. Sci. USA 112, 6670–6675.
Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N.N., Anderson, I.J., Cheng, J.-F., Darling, A., Malfatti, S., Swan, B.K., Gies, E.A., Dodsworth, J.A., Hedlund, B.P., Tsiamis, G., Sievert, S.M., Liu, W.-T., Eisen, J.A., Hallam, S.J., Kyrpides, N.C., Stepanauskas, R., Rubin, E.M., Hugenholtz, P., Woyke, T., 2013. Insights into the phylogeny and coding potential of microbial dark matter. Nature 499, 431–437.
Roure, B., Baurain, D., Philippe, H., 2013. Impact of missing data on phylogenies inferred from empirical phylogenomic data sets. Mol. Biol. Evol. 30, 197–214.
Sakai, S., Imachi, H., Hanada, S., Ohashi, A., Harada, H., Kamagata, Y., 2008. Methanocella paludicola gen. nov., sp. nov., a methane-producing archaeon, the first isolate of the lineage “Rice Cluster I”, and proposal of the new archaeal order Methanocellales ord. nov. Int. J. Syst. Evol. Microbiol. 58, 929–936.
Schleper, C., Jurgens, G., Jonuscheit, M., 2005. Genomic studies of uncultivated archaea. Nat. Rev. Microbiol. 3, 479–488.
Williams, T.A., Embley, T.M., 2014. Archaeal "dark matter" and the origin of eukaryotes. Genome Biol. Evol. 6, 474–481.
Williams, T.A., Szöllősi, G.J., Spang, A., Foster, P.G., Heaps, S.E., Boussau, B., Ettema, T.J.G., Embley, T.M., 2017. Integrative modeling of gene and genome evolution roots the archaeal tree of life. Proc. Natl. Acad. Sci. USA 114, E4602–E4611.
Woese, C.R., Achenbach, L., Rouviere, P., Mandelco, L., 1991. Archaeal phylogeny: reexamination of the phylogenetic position of Archaeoglobus fulgidus in light of certain composition-induced artifacts. Syst. Appl. Microbiol. 14, 364–371.
Wolf, Y.I., Makarova, K.S., Yutin, N., Koonin, E.V., 2012. Updated clusters of orthologous genes for Archaea: a complex ancestor of the Archaea and the byways of horizontal gene transfer. Biol. Direct 7, 46.
Wright, D.B., Banks, D.D., Lohman, J.R., Hilsenbeck, J.L., Gloss, L.M., 2002. The effect of salts on the activity and stability of Escherichia coli and Haloferax volcanii dihydrofolate reductases. J. Mol. Biol. 323, 327–344.
Yutin, N., Puigbò, P., Koonin, E.V., Wolf, Y.I., 2012. Phylogenomics of prokaryotic ribosomal proteins. PLoS One 7, e36972.