Extreme halophilic archaea derive from two distinct methanogen Class II lineages
Monique Aouad, Najwa Taïb, Anne Oudart, Michel Lecocq, Manolo Gouy, Céline Brochier-Armanet

To cite this version:

HAL Id: hal-01799910
https://hal.archives-ouvertes.fr/hal-01799910
Submitted on 25 May 2018
Extreme halophilic archaea derive from two distinct methanogen Class II lineages

Monique Aouad¹, Najwa Taib¹, Anne Oudart, Michel Lecocq, Manolo Gouy, Céline Brochier-Armanet

Université 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:
Stromosaarchaea
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 µm 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 constraints 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 RNA 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 sequences.
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, Paecarchaeota) 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 Methanomads, 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 artificial tree reconstructions where distant sequences with similar compositions are clustered together (Deliscu 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 Archeoglobales and 12 Difforarchaeae 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 (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, Difforarchaeae and Methanoceellales 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 Diahorarchae; 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)] and [(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) (Crisculo and Grisbaldo, 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 Diahorarchae, Archaeoglobales, ANME-I, methanogen Class II (i.e. Methanomicrobiales, Methanocellales and Methanosarcinales), and Halobacteria; Diahorarchae 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 archeal 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 archeal tree, as expected if they are member of the DPANN super-phyllum, they should branch on the stem of Diahorarchae.

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 Diahorarchae, and from Nanohaloarchae (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 under-represented 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 Diahorarchae 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 Diahorarchae. The Maximum Likelihood (ML) trees inferred with PHYML with the best-suited 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
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\textsubscript{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., 2015).

<table>
<thead>
<tr>
<th></th>
<th>Halobacteria</th>
<th>N + H</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 2 3 4 5 6 7 8 9</td>
<td>1 2 3 4 5 6 7 8 9</td>
</tr>
<tr>
<td><strong>FHalo</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LG + I + G4 + F 70</td>
<td></td>
<td>N + H</td>
</tr>
<tr>
<td>LG + PMSE + G4 70</td>
<td></td>
<td>N + H</td>
</tr>
<tr>
<td>LG + H + G4 70</td>
<td></td>
<td>N + H</td>
</tr>
<tr>
<td>LG + PMSE + G4 70*</td>
<td></td>
<td>N + H</td>
</tr>
<tr>
<td>LG + I + G4 95</td>
<td></td>
<td>N + H</td>
</tr>
<tr>
<td>LG + PMSE + G4 95</td>
<td></td>
<td>N + H</td>
</tr>
<tr>
<td>FNano</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LG + I + G4 70</td>
<td></td>
<td>N + H</td>
</tr>
<tr>
<td>LG + PMSE + G4 70</td>
<td></td>
<td>N + H</td>
</tr>
<tr>
<td>LG + PMSE + G4 70*</td>
<td></td>
<td>N + H</td>
</tr>
<tr>
<td>LG + H + G4 95</td>
<td></td>
<td>N + H</td>
</tr>
<tr>
<td>LG + PMSE + G4 95</td>
<td></td>
<td>N + H</td>
</tr>
<tr>
<td>FNanoHALO</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LG + I + G4 70</td>
<td></td>
<td>N + H</td>
</tr>
<tr>
<td>LG + PMSE + G4 70</td>
<td></td>
<td>N + H</td>
</tr>
<tr>
<td>LG + PMSE + G4 70*</td>
<td></td>
<td>N + H</td>
</tr>
<tr>
<td>LG + H + G4 95</td>
<td></td>
<td>N + H</td>
</tr>
<tr>
<td>LG + PMSE + G4 95</td>
<td></td>
<td>N + H</td>
</tr>
</tbody>
</table>

**Fig. 1.** Alternative branching positions for *Nanohaloarchaea* and *Halobacteria*. The tree represents the phylogenetic relationships among *Diaforarchaea*, *Archaeoglobales*, ANME-1, 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).

**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.
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$_{95}$ 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, irrespective 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 χ² 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 + PMSF + G4 model and 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, S3L–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).

**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$_{95}$ 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, irrespective 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 χ² 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 + PMSF + G4 model and 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, S3L–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).
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 artifacts such as the LBA (Delsuc et al., 2005; Felsenstein, 1978). 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 FNANOHALO95, FNANO95 and FHALO95 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 FNANOHALO95 and FHALO95 S-F supermatrices (Fig. S5A–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 FNANOHALO95 and FNANO95 supermatrices showed that the branching of Nanohaloarchaea on the stem of Difforarchaeae 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 Difforarchaeae 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-phyllum in Archaea (Rinke et al., 2013). The DPANN was challenged by several studies which suggested that the lineages composing this super-phyllum 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 Microarchaeota the closest relatives of Difforarchaeae (Petitjean et al., 2014), and Nanohaloarchaea the sister-group of Haloarchaea (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 Eurarchaeota (Williams et al., 2017). Furthermore, some DPANN proteomes harbor biased amino acid composition as a consequence of their extremophile 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
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 stenos, 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).
related to thermococcales? Genome Biol. 6, 842.


