ISME Journal, *Commentary*, Supplementary Material

'Geoarchaeon NAG1' is a deeply-rooting lineage of the archaeal order Thermoproteales rather than a new phylum

Lionel Guy, Anja Spang, Jimmy H. Saw and Thijs J. G. Ettema*

Supplementary Methods

A set of 57 conserved proteins was used as the basis of this study. The set is fully described in Guy *et al.* (2014). Briefly, archaeal clusters of orthologous genes (arCOGs) (Wolf et al 2012) were filtered, and only clusters present in 90% of a set of 58 representative archaea were retained. Homologs to these clusters were searched for in 10 representative bacteria, as well as in 8 novel archaeal genomes. Clusters that were not present in at least 3 bacterial genomes were discarded, leading to a set of 67 clusters. Paralogs and split proteins were curated manually. A discordance filter was then applied, following a method described elsewhere (Guy et al 2014, Viklund et al 2012), in an effort to remove horizontally transferred genes. Briefly, an individual gene tree is build for each cluster, and each tree is compared to all others, measuring the amount of shared, well-supported bipartitions. Trees are then ranked, with the ones having the least common bipartitions with other trees being the most discordant. Fifteen percent of the most phylogenetically discordant proteins were discarded, yielding to a set of 57 clusters. Of these, 32 are ribosomal protein genes. A 33rd cluster of ribosomal proteins (S14) was added to the set, as it was described to be universal previously (Lecompte et al 2002).

For each cluster, proteins were aligned with mafft-linsi v6.847b (Katoh and Toh 2008), and columns with 50% or more gaps were removed. Alignments were visually inspected for misaligned regions. Maximum likelihood phylogenies were inferred from individual alignments and concatenates with RAxML 7.9.5 (Stamatakis 2006) under CATLG model with 100 non-parametric bootstraps. Bayesian phylogenies were run with PhyloBayes

MPI 1.4f (Lartillot et al 2013) under CAT-Poisson model with a discrete gamma distribution of rates across sites. Four chains were run in parallel for ~12,000 generations, discarding the first 8000, and sampling every 50 trees. Convergence was assessed by plotting the variation of key statistics (log-likelihood, length of the tree, alpha parameter and number of modes) across generations, and verifying that the variation inside one chain was greater than across chains. In addition, the consensus trees for each chain were compared to ensure that they were similar and that the overall topology as shown in Figure 1 was identical in all four consensus trees. Trees were visualized with FigTree 1.4.0 (Andrew Rambaut, University of Edinburgh).

After alignment and concatenation, a χ^2-test (Viklund et al 2012) was applied to find the most and least compositionally sites. Alignments were built by adding increasing amounts (20 to 100%, by 10% increment) of the most and least biased sites. For each of these alignments, 100 non-parametric bootstraps were inferred as mentioned above. In addition, ML phylogenies were inferred for the most and least biased half of the alignment.

Small and large ribosomal RNA subunits from 90 archaea and 10 bacteria (used as outgroup) were individually aligned with mafft-linsi v6.847b (Katoh and Toh 2008) and columns consisting of 50% or more gaps were removed using trimAl (Capella-Gutiérrez et al 2009). The alignments were then concatenated and maximum likelihood and Bayesian phylogenies were inferred as above. In the Bayesian analysis, four chains were run for ~80,000 generations, discarding the first 40,000, and sampling every 100 tree. Convergence was assessed as above.

The occurrence of discussed arCOGs in NAG1 and in SAGs affiliating with Geoarchaeota (AAA471-B05, AAA471-B23, AAA471-C03, AAA471-L13, AAA471-L14, AAA471-O08)(Rinke et al 2013) was determined with PSI-BLAST (Altschul et al 1997) using aligned arCOGs as queries(Wolf et al 2012) (using an inclusion threshold of E-value < 0.001). The assignment of each specific protein to a particular arCOG cluster was subsequently checked by BLAST against the Refseq database (http://www.ncbi.nlm.nih.gov/refseq/) to avoid incorrect annotations due to the presence of truncated sequences only.

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W *et al* (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25:** 3389-3402.

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25:** 1972-1973.

Guy L, Saw JH, Ettema TJG (2014). The archaeal legacy of eukaryotes: a phylogenomic perspective. *Cold Spring Harbor Perspectives in Biology* **In press**.

Katoh K, Toh H (2008). Recent developments in the MAFFT multiple sequence alignment program. *Brief Bioinformatics* **9:** 286-298.

Kozubal MA, Romine M, Jennings Rd, Jay ZJ, Tringe SG, Rusch DB *et al* (2012). Geoarchaeota: a new candidate phylum in the Archaea from high-temperature acidic iron mats in Yellowstone National Park. *ISME J* **7:** 622–634.

Lartillot N, Rodrigue N, Stubbs D, Richer J (2013). PhyloBayes MPI. Phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. *Syst Biol* **62:** 611-615.

Lecompte O, Ripp R, Thierry J-C, Moras D, Poch O (2002). Comparative analysis of ribosomal proteins in complete genomes: an example of reductive evolution at the domain scale. *Nucleic Acids Res* **30:** 5382-5390.

Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng J-F *et al* (2013). Insights into the Phylogeny and Coding Potential of Microbial Dark Matter. *Nature* **499:** 431-437.

Spang A, Hatzenpichler R, Brochier-Armanet C, Rattei T, Tischler P, Spieck E *et al* (2010). Distinct gene set in two different lineages of ammonia-oxidizing archaea supports the phylum Thaumarchaeota. *Trends Microbiol* **18:** 331-340.

Stamatakis A (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22:** 2688-2690.

Viklund J, Ettema TJG, Andersson SGE (2012). Independent genome reduction and phylogenetic reclassification of the oceanic SAR11 clade. *Mol Biol Evol* **29:** 599-615.

Wolf YI, Makarova KS, Yutin N, Koonin EV (2012). Updated clusters of orthologous genes for Archaea: a complex ancestor of the Archaea and the byways of horizontal gene transfer. *Biol Direct* **7**.