The ISME Journal (2014) 8, 1353–1357

www.nature.com/ismej

COMMENTARY

'Geoarchaeote 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 JG Ettema

The ISME Journal (2014) **8,** 1353–1357; doi:10.1038/ismej.2014.6; published online 13 February 2014

Recent cultivation-independent studies have revealed a considerable diversity of uncultured and uncharacterized archaeal lineages. The genomic characterization of these lineages will be instrumental in order to shed light on their physiology and ecological significance and to provide an overall insight on the genomic diversity and evolution of the archaeal domain of life. In a recent issue of *The ISME Journal*, Kozubal *et al.* (2012) have used a metagenomic approach to obtain the composite genome of members of a 'novel archaeal group 1' (NAG1), isolated from a high-temperature acidic iron mat in Yellowstone National Park. Based on phylogenetic analysis of selected marker genes, Kozubal *et al.* (2012) concluded that NAG1 represented a novel phylum ('Geoarchaeota') in the Domain Archaea.

Although we want to stress that we value the overall findings presented by Kozubal *et al.* (2012) in terms of provided insights into metabolic and eco-physiological characteristics of the NAG1 lineage, we would like to contest the claim that NAG1 represents a novel archaeal phylum. As a main argument, we would like to point out that the phylogenetic approach used by Kozubal *et al.* (2012) is not suitable to place deeply rooting microbial lineages due to the nature of biological sequence data, which are, among others, affected by variations of evolutionary rates over time (heterotachy) and compositional bias. First, Kozubal *et al.* (2012) base their phylogenetic inference on 32 universal ribosomal proteins (r-proteins). It is a common practice to use the subset of universal r-proteins for phylogenetic analyses, most often successfully, even for deep phylogenies. Nonetheless, previous studies have shown that protein sequences of r-proteins are compositionally biased (Cox *et al.*, 2008; Fournier *et al.*, 2010) and perhaps not very well suited for addressing the existence of novel archaeal phyla, unless great care is taken to assess the possible effects of such bias (Brochier *et al.*, 2005). In addition, although very rarely, horizontal gene transfers of r-proteins are known to occur (Makarova *et al.*, 2001), and such events will complicate phylogenetic analyses of concatenated alignments. Second, by averaging over 20 trees based on different phylogenetic methods, Kozubal *et al.* (2012) put a disproportional weight on phylogenetic signals of distance-based approaches: 12 out of the 20 trees included in the analysis are based on minimum evolution and neighbor-joining methods, which are known to be sensitive to compositional bias and long branch attraction artifacts (Phillips *et al.*, 2004; Philippe *et al.*, 2005). Third, the substitution matrices used by Kozubal *et al.* (2012) (Dayhoff, JTT) are based on a small number of proteins with a relatively narrow taxonomic distribution, and their efficiency is surpassed by modern matrices (WAG, LG) (Le and Gascuel, 2008). Finally, the use of the number of (amino acid) differences (ND; basically a uniform distance matrix) and p-distance (ND normalized by the length of the alignment) as measure of distance between two proteins, is ignorant of the evolutionary processes at work and should hence not be used. The weakness of the methodology used is reflected by the fact that the overall branching patterns in Figures 3a and b (Supplementary Figure S3) in Kozubal *et al.* (2012) are conflicting with the generally accepted topology of the archaeal species tree, which generally recovers strong support for monophyly of the archaeal 'TACK' superphylum (Guy and Ettema, 2011; Williams *et al.*, 2012; Lasek-Nesselquist and Gogarten, 2013; Rinke *et al.*, 2013; Williams *et al.*, 2013).

To determine possible causes for these conflicting findings, we first inspected all 33 ribosomal protein trees inferred by maximum likelihood (ML) individually (see Supplementary Methods in the Supplementary Material for more details). This revealed that the different r-proteins harbor contrasting phylogenetic signals: r-proteins of NAG1 were grouping with Thaum-/Aigarchaeota (six trees), Thermoproteales (seven trees) or other sub-lineages of Crenarchaeota (eight trees), whereas only five r-proteins revealed a sister relationship to all Crenarchaeota (Supplementary Figure S1). In comparison, in a set of 25 conserved non-r-proteins from which potentially horizontally transferred genes have been removed by a discordance filtering

approach (Guy *et al.*, 2014), only 2 trees show NAG1 associated with Thaum-/Aigarchaeota, 4 with other crenarchaeal lineages, 3 as a separated phylum, whereas as many as 12 trees associate it with Thermoproteales (Supplementary Figure S2). Next, we re-analyzed the same concatenated set of 33 universally conserved r-proteins (5535 sites) with both ML and Bayesian methods. Although the ML analysis indicated that NAG1 is associated with the archaeal order Thermoproteales (albeit with insignificant bootstrap support (BS) = 23; Figure 1a and Supplementary Figure S3A), the Bayesian analysis placed NAG1 as a sister clade to Aig- and Thaumarchaeota with a posterior probability (PP) of 0.87 (Supplementary Figure S4). One possible explanation for the contradicting results of these phylogenetic analyses could be a biased amino-acid composition in r-protein sequences, which affect the placement of NAG1 in the archaeal species tree.

To assess and disentangle the effects of amino-acid compositional bias on the phylogeny inferred from the universal r-proteins, we divided all amino-acid sites of the concatenated alignment into two equal data sets using a χ^2 test (Viklund *et al.*, 2012; Guy *et al.*, 2014; see Supplementary Methods in the Supplementary Material for more details), with one data set comprised of the most and the other of the least compositionally biased sites. Whereas NAG1 branches at the root of Crenarchaeota in the set with most biased sites (BS = 94, Figure 1b, Supplementary Figure S3B), it is associated with Thermoproteales in the least biased data set (BS = 80; Figure 1c, Supplementary Figure S3C). In addition, we performed an ML phylogenetic analysis on the set of 25 non-r-proteins (9579 sites). This analysis, as well as an analysis of this data set in combination with 32 out of the 33 of the universal r-proteins (one universal r-protein was excluded by the discordance filter and was thus not included in the analysis; 15 051 sites), displays a strong association of NAG1 with Thermoproteales (BS = 100 and 92, respectively; Figures 1d and e; Supplementary Figures S3D and S3E). To investigate the impact of compositional biases on phylogenetic placement in more detail, the concatenated alignments of the 33 r-proteins and of the 25 non-r-proteins were further filtered, keeping increasing fractions (from 20% to 100%, with 10% increments) of the most or least biased sites. ML bootstraps inferred from these alignments reveal that the most biased sites of the

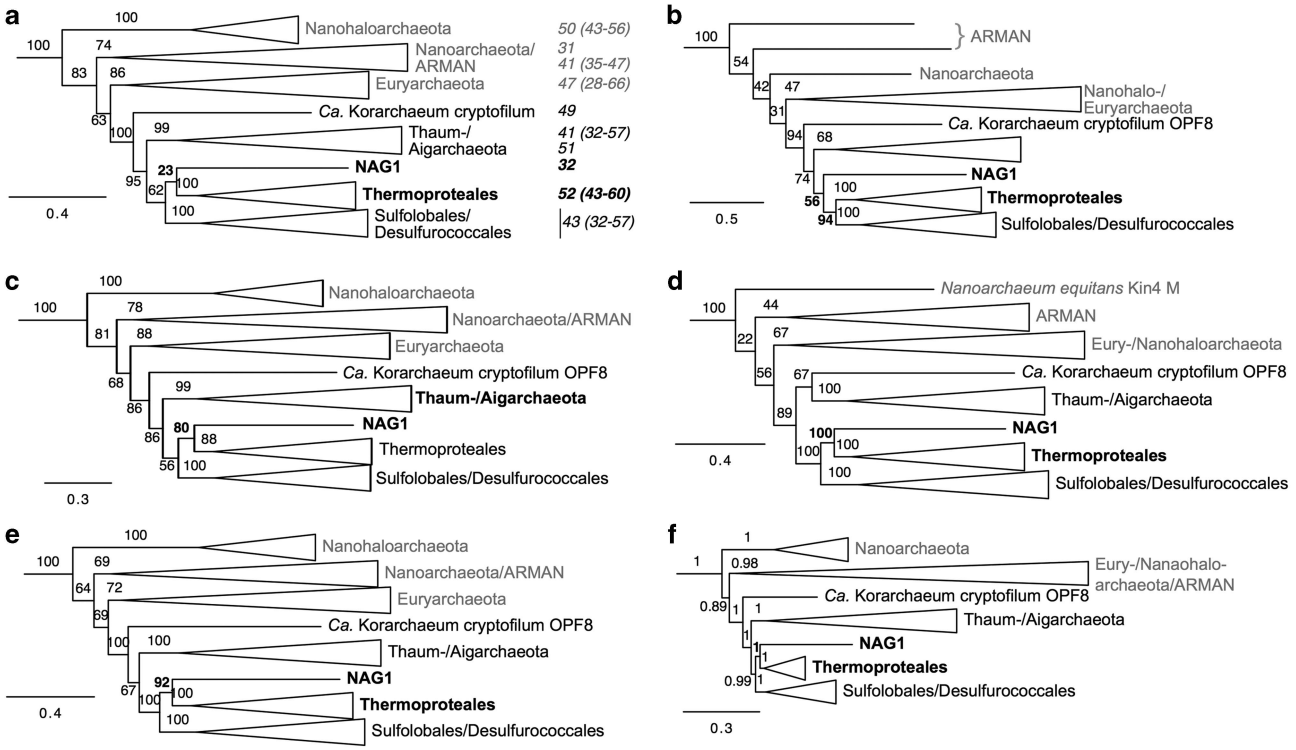

Figure 1 Phylogenetic placement of NAG1. In all the phylogenies, NAG1 and its sister clade are shown in bold font. Taxa not part of the TACK superphylum are shown in grey. Major taxa have been collapsed to improve readability. Numbers on branches represent bootstrap support (**a**–**e**) or posterior probability (**f**). Support for the branch linking NAG1 and its sister clade is shown in bold. Scale (number of substitutions per site) is shown next to each tree. Full trees showing all included organisms are displayed in Supplementary Figure S3. (**a**) ML phylogeny inferred from 33 universal r-proteins. Numbers in italics next to taxa names show the mean genomic GC content in the taxa and, whenever applicable, the range (between parentheses). (**b**, **c**) ML phylogeny inferred from the same proteins, keeping only the most and the least compositionally biased half of the sites, respectively. (**d**) ML phylogeny from a set of 25 conserved non-r-proteins, filtered for horizontal gene transfers. (**e**) ML phylogeny of the concatenation of 32 of the 33 r-proteins from panel (**a**) and the 25 non-r-proteins from panel (**d**). (**f**) Bayesian phylogeny inferred from the concatenated alignments of the sequences of the 16S and 23S rRNA sequences.

33 r-proteins firmly support NAG1 branching outside of Crenarchaeota (Supplementary Figure S5A), with about half of them associating it with Thaum-/Aigarchaeota, whereas the support for the placement of NAG1 as a sister clade to Thermoproteales is increasing when keeping only the least biased sites (Supplementary Figure S5B). The same analysis on the 25 non-r-protein set shows a uniformly strong support for NAG1 as a sister clade to Thermoproteales (Supplementary Figures S5C and D). This strong affiliation is corroborated by a Bayesian analyses of concatenated 16S and 23S ribosomal RNA (rRNA) genes, which places NAG1 as a deep sister lineage of the Thermoproteales with the strongest possible support (PP = 1; Figure 1f, Supplementary Figure S3F). It should be noted, however, that a ML analysis of the same data set places NAG1 as a sister clade to the Crenarchaeota with a fair support (BS = 76; Supplementary Figure S6). Yet, given that the CAT model implemented in the Bayesian analysis (see Supplementary Material for details) is specifically designed to reduce systematic errors caused by taxon sampling effects and compositional bias in sequence data (Lartillot *et al*., 2007), it is reasonable to assume that the placement of NAG1 obtained in the ML analysis of the ribosomal RNA data set is caused by such biases. Removing allegedly fast-evolving taxa (ARMAN, Nano- and Nanohaloarchaeota) confirmed the ambiguous phylogenetic signal in the r-proteins, for which the best ML tree shows moderate support (BS = 65) for NAG1 as a sister clade to Aig- and Thaumarchaeota (Supplementary Figure S7A) and a very strong support for NAG1 as a sister clade to Thermoproteales in the 25 non-r-proteins, 57 conserved proteins and concatenated 16S and 23S sets (BS = 100, BS = 100, PP = 0.99, respectively; Supplementary Figures S7B–D). Altogether, the re-assessment of the phylogenetic affiliation of NAG1 presented here indicates that NAG1 represents an archaeal lineage that is affiliated with the order Thermoproteales rather than a phylum-status archaeal lineage as was suggested by Kozubal *et al.* (2012). Our findings seem to be in line with those described in a recent study of single-cell amplified genomes (SAGs) belonging (among others) to Geoarchaeota, which also reported that this clade 'clusters within the Crenarchaeota' (Rinke *et al*., 2013). A possible explanation as to why NAG1

Species	Order and/or Phylum	L30e	L14e	L34e	S25e	S26e	S30e	L18ae	L35ae	L38e	L41e	L13e	L45a	L46a	L47a
Thermofilum pendens	*Thermoproteales* (Crenarchaeota)	◆	◆	◆	◆	◆	◆	◆	◆			◆			
Geoarchaeon NAG1	*Thermoproteales* (Crenarchaeota)	◆	◆	◆	◆	◆	◆	◆		◆		◆			
Thermoproteus neutrophilus	*Thermoproteales* (Crenarchaeota)	◆	◆	◆	◆	◆	◆	◆		◆		◆			
Pyrobaculum aerophilum	*Thermoproteales* (Crenarchaeota)	◆	◆	◆	◆	◆	◆	◆		◆		◆			
Pyrobaculum arsenaticum	*Thermoproteales* (Crenarchaeota)	◆	◆	◆	◆	◆	◆	◆		◆		◆			
Pyrobaculum calidifontis	*Thermoproteales* (Crenarchaeota)	◆	◆	◆	◆	◆	◆	◆		◆		◆			
Pyrobaculum islandicum	*Thermoproteales* (Crenarchaeota)	◆	◆	◆	◆	◆	◆	◆		◆		◆			
Caldivirga maquilingensis	*Thermoproteales* (Crenarchaeota)	◆	◆	◆	◆	◆	◆	◆		◆		◆			
Thermosphaera aggregans	*Desulfurococcales* (Crenarchaeota)	◆	◆	◆	◆	◆	◆	◆						◆	
Desulfurococcus kamchatkensis	*Desulfurococcales* (Crenarchaeota)	◆	◆	◆	◆	◆	◆	◆						◆	
Acidilobus saccharovorans	*Desulfurococcales* (Crenarchaeota)	◆	◆		◆	◆	◆	◆		◆				◆	
Aeropyrum pernix	*Desulfurococcales* (Crenarchaeota)	◆	◆	◆	◆	◆	◆	◆	◆	◆		◆		◆	
Hyperthermus butylicus	*Desulfurococcales* (Crenarchaeota)	◆	◆	◆	◆	◆	◆	◆	◆	◆		◆		◆	
Ignicoccus hospitalis	*Desulfurococcales* (Crenarchaeota)	◆	◆	◆	◆	◆	◆	◆	◆	◆		◆			
Staphylothermus hellenicus	*Desulfurococcales* (Crenarchaeota)	◆	◆	◆	◆	◆	◆	◆	◆	◆		◆		◆	
Metallosphaera sedula	*Sulfolobales* (Crenarchaeota)	◆	◆	◆	◆	◆	◆	◆					◆	◆	◆
Sulfolobus acidocaldarius	*Sulfolobales* (Crenarchaeota)	◆	◆	◆	◆	◆	◆	◆					◆	◆	◆
Sulfolobus islandicus YN1551	*Sulfolobales* (Crenarchaeota)	◆	◆	◆	◆	◆	◆	◆				◆	◆	◆	◆
Sulfolobus solfataricus	*Sulfolobales* (Crenarchaeota)	◆	◆	◆	◆	◆	◆	◆				◆	◆	◆	◆
Sulfolobus tokodaii	*Sulfolobales* (Crenarchaeota)	◆	◆	◆	◆	◆	◆	◆				◆	◆	◆	◆
Korarchaeum cryptofilum	Korarchaeota	◆	◆	◆	◆	◆	◆					◆			
Nitrosopumilus maritimus	Thaumarchaeota	◆			◆	◆	◆								
Caldiarchaeum subterraneum	Thaum/ (Aig)archaeota	◆	◆		◆	◆	◆								
Nanoarchaeum equitans	Nanoarchaeota	◆	◆	◆		◆		◆	◆		◆				
Thermococcus gammatolerans	*Thermococcales* (Euryarchaeota)	◆	◆	◆				◆	◆		◆				
Methanobrevibacter smithii	*Methanobacteriales* (Euryarchaeota)	◆	◆	◆				◆							
Methanococcus jannaschii	*Methanococcales* (Euryarchaeota)	◆	◆	◆				◆			◆				
Methanopyrus_kandleri	*Methanopyrales* (Euryarchaeota)	◆	◆	◆		◆		◆	◆						
Aciduliprofundum boonei	(Euryarchaeota)	◆						◆			◆				
Archaeoglobus profundus	*Archaeoglobales* (Euryarchaeota)	◆				◆		◆							
Methanoregula boonei	*Methanomicrobiales* (Euryarchaeota)	◆						◆							
Methanocella paludicola	*Methanocellales* (Euryarchaeota)	◆						◆							
Methanosarcina acetivorans	*Methanosarcinales* (Euryarchaeota)	◆						◆							
Thermoplasma acidophilum	*Thermoplasmatales* (Euryarchaeota)										◆				
Haloarcula marismortui	*Halobacteriales* (Euryarchaeota)							◆							

Figure 2 Overview of non-universal r-protein occurrence in different archaeal lineages. The figure is in part based on a study by Yutin *et al.* (2012) but has been updated to include novel members of the archaea, including NAG1. Black diamonds denote the presence of respective r-proteins.

appears as a sister clade to Crenarchaeota in Kozubal *et al.* (2012) is that many r-proteins in NAG1 and Aig- and Thaumarchaeota might share a similar compositional bias, whereas Thermoproteales and NAG1 have a large difference in genomic GC content (NAG1: 32%; Thermoproteales: 52% in average; see Figure 1a) although other factors (horizontal gene transfer, weak phylogenetic signal and so on) cannot be excluded as a cause of errors in phylogenetic inference. As a consequence, phylogenies might have placed NAG1 at an artifactual intermediate position, at the root of Crenarchaeota, as shown in Kozubal *et al.* (2012) and by our ML analyses of the full 33 r-protein data set as well as by the concatenated 16S and 23S rRNA sequence phylogeny.

An additional argument used by Kozubal *et al.* (2012) in support of the proposed phylum status of NAG1 entails the presence/absence pattern of a previously described set of informational processing and cell division genes (Spang *et al.*, 2010). We would like to argue that the observed phyletic distribution patterns do not support such conclusions. First, features that are currently regarded as missing from the NAG1 genome have to be interpreted with caution as the genome of this organism is incomplete. For example, the authors pointed out that NAG1 lacks genes for Topo IA, a topoisomerase generally present in Crenarchaeota and most other archaeal phyla. However, we identified Topo IA orthologs in several NAG1-related SAGs that were published recently (Rinke *et al.*, 2013). Second, a careful re-analysis of the described gene pattern of NAG1 fails to provide convincing support for the claim that this organism represents a new archaeal phylum. For instance, NAG1 encodes a set of r-proteins characterized by the presence of L13e, L18ae and L38e and the absence of L35ae, L41e and L45a-L47a (Figure 2). Surprisingly, the same set of r-proteins is encoded by all members of the Thermoproteales except for *Thermofilum pendens*, which might have lost L38e (Figure 2) rather than L35ae as suggested for other members of this group (Yutin *et al.*, 2012). The most parsimonious explanation for this observation is that, in line with the results of the phylogenetic analyses discussed above, Thermoproteales and NAG1 share a common ancestor to the exclusion of the other crenarchaeal lineages. Additional characteristics that are shared between Crenarchaeota and NAG1 include, among others, the presence of the transcription factor ELF1, two homologs of the crenarchaeal-type single-stranded binding protein involved in replication and *cdvABC* genes for cell division and the absence of FtsZ, ScpA and ScpB, SmcA and histones (also absent from NAG1-related SAGs). The presence of a single gene coding for sliding clamp protein PCNA in NAG1 is similar to *T. pendens*, but differs from later diverging Crenarchaeota that harbor two or three homologs (Supplementary Figure S8).

Indeed, only two proteins of this signature set differ between Crenarchaeota and Geoarchaeota: the NAG1 genome encodes a gene for the small subunit of polymerase D (the gene annotated as large subunit of PolD appears incorrectly annotated) and lacks an ortholog of RNA polymerase G (homolog of the eukaryotic rpb8 subunit) (Supplementary Figure S8). Yet, the phylogenetic distribution patterns of these few genes are not sufficient to warrant a phylum status of NAG1, as these can be easily explained by lineage-specific gene loss events. The apparent similarity in signature gene content between NAG1 and Crenarchaeota contrasts with that observed between Thaum- and Crenarchaeota, which differ by >10 proteins in this signature gene set and, in addition, harbor completely different sets of r-proteins and repair machineries (Spang *et al.*, 2010).

In summary, our re-assessment of the phylogenetic placement and presence/absence patterns of core genes strongly undermines the proposed phylum status of NAG1. Rather, our analyses suggest that NAG1 represents a deeply rooting sister clade of the archaeal order Thermoproteales. With powerful cultivation-independent approaches nearing maturity, the rate and magnitude at which novel microbial lineages can be explored at the genomic level will increase in the near future. The wealth of genomic data that will emerge from these efforts will allow us to gain an unprecedented insight into microbial diversity and evolution. Yet, the use of robust phylogenomic approaches will be a *sine qua non* to correctly infer the phylogenetic affiliations of these new microbial lineages.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgements

The work in Ettema laboratory is supported by the Swedish Research Council (Grant number 621-2009-4813), by the European Research Council (ERC) (Grant number 310039-PUZZLE_CELL), by a Marie Curie European Reintegration Grant (ERG) (Grant number 268259-RICKOCHET) and by grants of the Carl Tryggers Stiftelse (to AS) and Wenner-Gren Stiftelse (to JHS).

L Guy, A Spang, JH Saw and Thijs JG Ettema are at the Department of Cell and Molecular Biology, Science for Life Laboratory, Uppsala University, Uppsala, Sweden
E-mail: thijs.ettema@icm.uu.se

References

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.

Cox CJ, Foster PG, Hirt RP, Harris SR, Embley TM. (2008). The archaebacterial origin of eukaryotes. *Proc Natl Acad Sci USA* **105**: 20356–20361.
Fournier GP, Neumann JE, Gogarten JP. (2010). Inferring the ancient history of the translation machinery and genetic code via recapitulation of ribosomal subunit assembly orders. *PLoS One* **5**: e9437.
Guy L, Ettema TJG. (2011). The archaeal 'TACK' superphylum and the origin of eukaryotes. *Trends Microbiol* **19**: 580–587.
Guy L, Saw JH, Ettema TJG. (2014). The archaeal legacy of eukaryotes: a phylogenomic perspective. *Cold Spring Harb Persp Biol;* doi:10.1101/CSHPERSPECT.a016022.
Kozubal MA, Romine M, Rd Jennings, 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, 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.
Lasek-Nesselquist E, Gogarten JP. (2013). The effects of model choice and mitigating bias on the ribosomal tree of life. *Mol Phylogenet Evol* **69**: 17–38.
Le SQ, Gascuel O. (2008). An improved general amino acid replacement matrix. *Mol Biol Evol* **25**: 1307–1320.
Makarova KS, Ponomarev VA, Koonin EV. (2001). Two C or not two C: recurrent disruption of Zn-ribbons, gene duplication, lineage-specific gene loss, and horizontal gene transfer in evolution of bacterial ribosomal proteins. *Genome Biol* **2**: RESEARCH 0033.
Philippe H, Zhou Y, Brinkmann H, Rodrigue N, Delsuc F. (2005). Heterotachy and long-branch attraction in phylogenetics. *BMC Evol Biol* **5**: 50.
Phillips MJ, Delsuc F, Penny D. (2004). Genome-scale phylogeny and the detection of systematic biases. *Mol Biol Evol* **21**: 1455–1458.
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.
Viklund J, Ettema TJG, Andersson SGE. (2012). Independent genome reduction and phylogenetic reclassification of the oceanic SAR11 clade. *Mol Biol Evol* **29**: 599–615.
Williams TA, Foster PG, Nye TMW, Cox CJ, Embley TM. (2012). A congruent phylogenomic signal places eukaryotes within the Archaea. *Proc R Soc Lond B Biol Sci* **279**: 4870–4879.
Williams TA, Foster PG, Cox CJ, Embley TM. (2013). An archaeal origin of eukaryotes supports only two primary domains of life. *Nature* **504**: 231–236.
Yutin N, Puigbò P, Koonin EV, Wolf YI. (2012). Phylogenomics of prokaryotic ribosomal proteins. *PLoS One* **7**: e36972.

Supplementary Information accompanies this paper on The ISME Journal website (http://www.nature.com/ismej)