APPLIED AND ENVIRONMENTAL MICROBIOLOGY, Oct. 2006, p. 6841–6844 Vol. 72, No. 10
0099-2240/06/$08.00+0 doi:10.1128/AEM.00429-06

Whole-Genome Reciprocal BLAST Analysis Reveals that *Planctomycetes* Do Not Share an Unusually Large Number of Genes with *Eukarya* and *Archaea*†

Clara A. Fuchsman and Gabrielle Rocap*

School of Oceanography, University of Washington, Seattle, Washington 98195

Received 21 February 2006/Accepted 28 July 2006

The genome sequences of *Rhodopirellula baltica*, formerly *Pirellula* sp. strain 1, *Blastopirellula marina*, *Gemmata obscuriglobus*, and *Kuenenia stuttgartiensis* were used in a series of pairwise reciprocal best-hit analyses to evaluate the contested evolutionary position of *Planctomycetes*. Contrary to previous reports which suggested that *R. baltica* had a high percentage of genes with closest matches to *Archaea* and *Eukarya*, we show here that these *Planctomycetes* do not share an unusually large number of genes with the *Archaea* or *Eukarya*, compared with other *Bacteria*. Thus, best-hit analyses may assign phylogenetic affinities incorrectly if close relatives are absent from the sequence database.

Microbiologists have debated the position of *Planctomycetes* in the tree of life. The canonical view of bacterial evolution identifies the *Planctomycetes* as a derived lineage with a high evolution rate and places hyperthermophiles at the root of the tree (16, 25). However, other reports have placed *Planctomycetes* at the root (3, 20). The genome sequence of the first member of the *Planctomycetes* to be sequenced, the marine organism *Rhodopirellula baltica*, formerly *Pirellula* sp. strain 1 (7), further complicated the debate. Of the deduced open reading frames from *R. baltica* that had a significant BLAST hit compared with the GenBank nonredundant database, 9% had closest BLAST hits to *Archaea* and 8% to *Eukarya* (7). Both of these percentages are much larger than has been observed when similar analyses have been applied to other bacterial genomes (7), apparently forming another line of evidence for the basal position of the *Planctomycetes*.

Planctomycetes are morphologically different from other bacteria and resemble eukaryotes in several ways. In particular, *Planctomycetes* do not have typical bacterial cell membranes. They lack peptidoglycan, and the fatty acids that constitute their phospholipids are mainly palmitic, palmitoleic, and oleic acids, which are typical of microeukaryotes, not *Bacteria* (9). Members of the *Planctomycetes* that utilize the anammox reaction have ether lipids (17, 18), once thought to be diagnostic of *Archaea* (11) but also found in some thermophiles and sulfate-reducing bacteria (10, 15). The planctomycete *Gemmata obscuriglobus* is capable of synthesizing sterols (16), a trait originally assumed to be eukaryotic but now also found in the bacterial lineages *Methylococcales* (16) and *Myxobacteriales* (2). Each of these sterol-utilizing bacterial lineages is characterized by cell compartmentalization (16). All *Planctomycetes* have at least a paryphoplasm: a membrane-bound, ribosome-free region (13). In addition to the paryphoplasm, *G. obscuriglobus* has a double membrane surrounding a nucleoid (6), and *Planctomycetes* that utilize the anammox reaction have an internal compartment called the anammoxosome where this reaction takes place (24).

Though morphological arguments suggest that *Planctomycetes* are similar to eukaryotes and therefore may be an ancient lineage, molecular analyses are in conflict with regard to the evolution of *Planctomycetes*. Phylogenetic analyses of the 16S rRNA gene have disagreed on the placement of the *Planctomycetes* in the bacterial phylogenetic tree (3, 5, 12, 14, 20, 25). A recent analysis resulted in significantly different 16S rRNA gene trees depending on the species of *Planctomycetes* included, consistent with the hypothesis that this group has experienced a high rate of evolution (12). Trees made from the most slowly evolving nucleotide bases of 16S rRNA have *Planctomycetes* at the root of the *Bacteria* (3). However, the limited number of positions in these analyses has led to the suggestion that these trees are not robust (5). Phylogenetic analysis using amino acid sequences of the elongation factor Tu also could not reliably resolve the division's position in the tree (8). Most recently, a comparison of 347 eukaryotic signature proteins to the unpublished genome of the planctomycete *Gemmata* sp. strain Wa-1 found a low number of high-scoring matches. Compared with matches to a proteobacterium, *Gemmata* did not appear significantly more closely related to *Eukarya* (21).

Genome sequences of two *Planctomycetes* (the marine organism *R. baltica* [7] and the uncultured anammox bacterium *Kuenenia stuttgartiensis* [22]) are now complete, and two more (*Blastopirellula marina* and *Gemmata obscuriglobus*) are in progress and publicly available, providing a molecular data set to address this question. Interestingly, both completed genomes have genes for peptidoglycan synthesis and other vestiges of a gram-negative cell wall (7, 22), which could suggest that the lineage *Planctomycetes* had once possessed and then lost a peptidoglycan cell wall. Phylogenetic trees of 39 concatenated ribosomal proteins (23), 49 concatenated protein sequences (22), and conserved proteins, such as ATP synthase and heat shock proteins 60 and 70 (7), indicate that *Plancto-*

* Corresponding author. Mailing address: School of Oceanography, Box 357940, University of Washington, Seattle, WA 98195. Phone: (206) 685-9994. Fax: (206) 685-6651. E-mail: rocap@ocean.washington.edu.

† Supplemental material for this article may be found at http://aem.asm.org/.

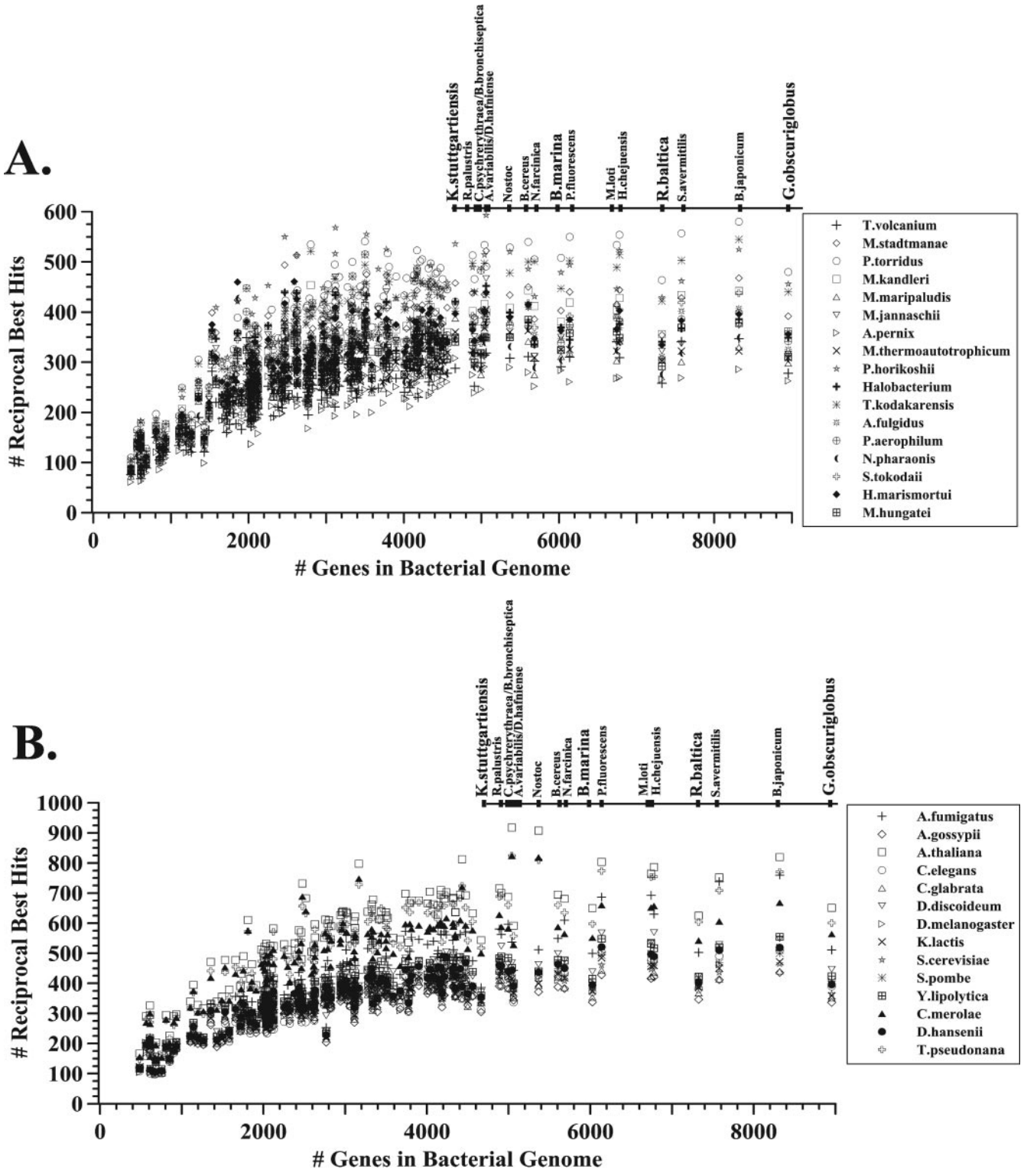

FIG. 1. Comparison between the number of best reciprocal BLAST hits ($E < e^{-10}$) between bacterial and archaeal (A) or bacterial and eukaryotic (B) genomes and the number of genes in the bacterial genome. Bacteria whose genomes have more than 4,000 genes are listed along the top axes. Each bacterial genome was separately compared to each archaeal or eukaryotic genome. Results are shown for the bacteria *Kuenenia stuttgartiensis*, *Rhodopseudomonas palustris*, *Colwellia psychrerythraea*, *Bordetella bronchiseptica*, *Anabaena variabilis*, *Desulfitobacterium hafniense*, *Nostoc* spp., *Bacillus cereus*, *Nocardia farcinica*, *Blastopirellula marina*, *Pseudomonas fluorescens*, *Mesorhizobium loti*, *Hahella chejuensis*, *Rhodopirellula baltica*, *Streptomyces avermitilis*, *Bradyrhizobium japonicum*, and *Gemmata obscuriglobus*. Information for bacteria with less than 4,000 genes can be found in Tables S2 and S3 in the supplemental material. Archaea used were *Thermoplasma volcanium*, *Methanosphaera stadtmanae*, *Picrophilus torridus*, *Methanopyrus kandleri*, *Methanococcus maripaludis*, *Methanococcus jannaschii*, *Aeropyrum pernix*, *Methanobacterium thermoautotrophicum*, *Pyrococcus horikoshii*, *Halobacterium* spp., *Thermococcus kodakarensis*, *Archaeoglobus fulgidus*, *Pyrobaculum aerophilum*, *Natronomonas pharaonis*, *Sulfolobus tokodaii*, *Haloarcula marismortui*, and *Methanospirillum hungatei*. Eukaryotes used were *Aspergillus fumigatus*, *Ashbya gossypii*, *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Candida glabrata*, *Dictyostelium discoideum*, *Drosophila melanogaster*, *Kluyveromyces lactis*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Yarrowia lipolytica*, *Cyanidioschyzon merolae*, *Debaryomyces hansenii*, and *Thalassiosira pseudonana*. Data for *Photorhabdus luminescens* and for *Methanosarcina acetivorans* were omitted for the sake of viewing clarity.

mycetes are not deeply branching. However, the higher-than-usual percentage of genes with BLAST hits to *Archaea* and *Eukarya* in *R. baltica* (7) might seem to support a basal position for the *Planctomycetes*. No trend was found for an organism of origin or a distinct functional category for these genes (7), reducing the likelihood that the result is due to a few instances of massive horizontal gene transfer.

R. baltica, the first planctomycete to be sequenced, is only distantly related to other divisions of *Bacteria* and hence to the majority of bacterial sequences in GenBank at the time of its sequencing. Thus, it is possible that the relatively large percentage of genes with best BLAST hits to *Archaea* and *Eukarya* was a result of the lack of close relatives available in the database. Other researchers have used comparisons of completed bacterial genomes to archaeal genomes to determine orthologs between genome pairs and to infer phylogeny (19). Although the number of orthologs was highly dependent on genome size (19), because this method compares one bacterial genome to one archaeal genome at a time, rather than examining the single best hit from a multiple-genome database, results are less susceptible to the taxonomic biases present in sequence databases. Here we employ a similar technique to compare bacterial genomes to archaeal and eukaryotic genomes in order to find out if the genomes of the planctomycetes *R. baltica*, *B. marina*, *G. obscuriglobus*, and *K. stuttgartiensis* contain an unusual number of eukaryotic and archaeal genes when biases in genome databases and genome sizes are taken into account.

The complete predicted protein sequences from 166 genomes (18 archaeal, 134 bacterial, and 14 eukaryotic) were downloaded from public databases in April 2006 (see Table S1 in the supplemental material). The genomes include a broad phylogenetic sampling, including both *Crenarchaeotes* and *Euryarchaeotes* and single and multicellular eukaryotes. In addition, preliminary sequence data for *G. obscuriglobus* were obtained from The Institute for Genomic Research through the website at http://www.tigr.org, and open reading frames were predicted using Glimmer, version 3.0 (4). Proteins in each bacterial genome were individually compared to proteins in each archaeal genome by reciprocal BLASTP (1). The best hit for each gene was extracted using a Perl script. The best hits for each bacterial-archaeal and archaeal-bacterial pair were compared in order to determine the number of reciprocal best hits for each pairwise comparison. The number of reciprocal best hits was counted by using an expectation value (E) of $<e^{-10}$ as the stringency threshold for determining a valid best hit. A subset of the genomes were also counted using a stringency threshold of $E < e^{-2}$. The same process was used to identify eukaryotic and bacterial protein pairs. Finally, archaeal genomes were compared to one another in order to confirm the underlying assumption that this method reflects organismal phylogeny. Indeed, *Archaea* had larger numbers of reciprocal best hits to other *Archaea* than did *Bacteria* of a similar genome size.

Regardless of the E value employed as a cutoff, the number of reciprocal BLAST hits between *Bacteria* and *Archaea* or *Eukarya* increased linearly with the number of genes in the bacterial genome until the number of genes in the bacterial genome reached approximately 4,000, after which the number of reciprocal best hits leveled off (Fig. 1) (see also Tables S2 and S3 in the supplemental material). Multiple regressions indicated that the number of reciprocal best hits depended on the numbers of both bacterial and archaeal genes ($P < 0.001$) but not on the number of eukaryotic genes. In the multiple regression of archaeal reciprocal best hit data, the number of genes in the bacterial genome was the most important variable used to explain the data (β coefficient, >0.7). The lack of importance of the number of eukaryotic genes for the number of reciprocal best hits could be due to the fact that the number of genes in all the eukaryotic genomes is relatively large compared to the number of genes in a bacterial or archaeal genome.

Importantly, these data show that *R. baltica*, *B. marina*, and *G. obscuriglobus* do not have more BLAST hits to *Archaea* and *Eukarya* than one would expect from their genome sizes (Fig. 1). In fact, all four *Planctomycetes* have fewer hits to *Eukarya* than do any four genomes of similar size examined ($P < 0.005$ by the Wilcoxon signed-rank test). These results are consistent with another genome analysis that found a low number of BLAST hits between the *Gemmata* Wa-1 genome and a list of eukaryotic signature proteins (21). *R. baltica* and *G. obscuriglobus* also have significantly fewer hits to *Archaea* than any four genomes of similar size ($P < 0.005$ by the Wilcoxon signed-rank test). *K. stuttgartiensis*, however, has more hits to *Archaea* than do three out of four bacteria with similar genome sizes ($P < 0.01$ by the Wilcoxon signed-rank test). Genomic analysis has revealed that *K. stuttgartiensis* is more deeply branching than *R. baltica* and *G. obscuriglobus* (22). Examination of protein pairs between *Bacteria* found that *R. baltica* and *G. obscuriglobus* have fewer reciprocal best hits to other *Bacteria* than all but 3 of the 32 nonplanctomycete bacterial genomes with more than 4,000 genes ($P < 0.005$ by the Wilcoxon signed-rank test), and *B. marina* has fewer reciprocal best hits than all but 5 ($P < 0.005$ by the Wilcoxon signed-rank test). The difference in the number of reciprocal best hits compared to other *Bacteria* was not as striking for *K. stuttgartiensis*. This difference may reflect the more recently derived lineages of *R. baltica*, *B. marina*, and *G. obscuriglobus*.

Our data suggest that the initial genome analyses found *R. baltica* to have an unusually large number of genes with best hits to *Archaea* and *Eukarya* (7) due to the dearth of sequences from closely related taxa in the database rather than because of its phylogenetic position. Indeed, when we repeated this analysis by comparing all 7,325 potential proteins in *R. baltica* to the NCBI nonredundant database in May 2006, 4,243 proteins had a significant hit (BLASTP expectation value, $<10^{-3}$), compared to 3,380 in 2003. Furthermore, now that the genomes of *K. stuttgartiensis* and *B. marina* are available in GenBank, the number of apparent BLAST hits of *R. baltica* to archaeal and eukaryotic sequences has decreased to 0.52% and 0.87% of the total number of genes (0.9% and 1.5% of the genes with significant BLAST hits). These results underscore the danger of assigning phylogenetic affinities from top BLAST hits, especially if close relatives are absent from the sequence database.

We thank James T. Staley and M. Claire Horner-Devine for valuable discussions and William J. Brazelton for technical assistance.

C.A.F. was funded by NSF MCB 0132101 to J. Murray and J. Staley. G.R. was funded by NSF OCE-0220826. Sequencing of the *G. obscuri-*

globus genome at TIGR was accomplished with support from the Department of Energy.

REFERENCES

1. **Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman.** 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25:**3389–3402.
2. **Bode, H. B., B. Zeggel, B. Silakowski, C. C. Wenzel, H. Reichenbach, and R. Muller.** 2003. Steroid biosynthesis in prokaryotes: identification of myxobacterial steroids and cloning of the first bacterial 2,3(*S*)-oxidosqualene cyclase from the myxobacterium *Stigmatella aurantiaca*. Mol. Microbiol. **47:**471–481.
3. **Brochier, C., and H. Philippe.** 2002. A non-hyperthermophilic ancestor for *Bacteria*. Nature **417:**244.
4. **Delcher, A. L., D. Harmon, S. Kasif, O. White, and S. L. Salzberg.** 1999. Improved microbial gene identification with GLIMMER. Nucleic Acids Res. **26:**544–548.
5. **Di Giulio, M.** 2003. The ancestor of the bacteria domain was a hyperthermophile. J. Theor. Biol. **224:**277–283.
6. **Fuerst, J. A., and R. I. Webb.** 1991. Membrane-bounded nucleoid in the eubacterium *Gemmata obscuriglobus*. Proc. Natl. Acad. Sci. USA **88:**8184–8188.
7. **Glöckner, F. O., M. Kube, M. Bauer, H. Teeling, T. Lombardot, W. Ludwig, D. Gade, A. Beck, K. Borzym, K. Heitmann, R. Rabus, H. Schlesner, R. Amann, and R. Reinhardt.** 2003. Complete genome sequence of the marine planctomycete *Pirellula* sp. strain 1. Proc. Natl. Acad. Sci. USA **100:**8298–8303.
8. **Jenkins, C., and J. A. Fuerst.** 2001. Phylogenetic analysis of evolutionary relationships of the planctomycete division of the domain *Bacteria* based on amino acid sequences of elongation factor Tu. J. Mol. Evol. **52:**405–418.
9. **Kerger, B. D., C. A. Mancuso, P. D. Nichols, D. C. White, T. Langworthy, M. Sittig, H. Schlesner, and P. Hirsch.** 1988. The budding bacteria, *Pirellula* and *Planctomycetes*, with atypical 16S rRNA and absence of peptidoglycan, show eubacterial phospholipids and uniquely high proportions of long chain beta-hydroxy fatty acids in the lipopolysaccharide lipid A. Arch. Microbiol. **149:**255–260.
10. **Langworthy, T. A., G. Holzer, J. G. Zeikus, and T. G. Tornabene.** 1983. Iso-and anteiso-branched glycerol diethers of the thermophilic anaerobe *Thermodesulfobacterium commune*. Syst. Appl. Microbiol. **4:**1–17.
11. **Langworthy, T. A., and J. L. Pond.** 1986. Archaebacterial ether lipids and chemotaxonomy. Syst. Appl. Microbiol. **7:**253–257.
12. **Liesack, W., R. Soller, T. Stewart, H. Haas, S. Giovannoni, and E. Stackebrandt.** 1992. The influence of tachytelically (rapidly) evolving sequences on the topology of phylogenetic trees—intrafamily relationships and the phylogenetic position of the *Planctomycetaceae* as revealed by comparative analysis of 16S ribosomal RNA sequences. Syst. Appl. Microbiol. **13:**357–362.
13. **Lindsay, M. R., R. I. Webb, M. Strous, M. S. Jetten, M. K. Butler, R. J. Forde, and J. A. Fuerst.** 2001. Cell compartmentalisation in planctomycetes: novel types of structural organisation for the bacterial cell. Arch. Microbiol. **175:**413–429.
14. **Pace, N. R.** 1997. A molecular view of microbial diversity and the biosphere. Science **276:**734–740.
15. **Pancost, R. D., I. Bouloubassi, V. Aloisi, J. S. Sinninghe Damsté, and the MEDINAUT Shipboard Scientific Party.** 2001. Three series of non-isoprenoidal dialkyl glycerol diethers in cold-seep cabonate crusts. Org. Geochem. **32:**695–707.
16. **Pearson, A., M. Budin, and J. J. Brocks.** 2003. Phylogenetic and biochemical evidence for sterol synthesis in the bacterium *Gemmata obscuriglobus*. Proc. Natl. Acad. Sci. USA **100:**15352–15357.
17. **Sinninghe Damsté, J. S., M. Strous, W. I. C. Rijpstra, E. C. Hopmans, J. A. J. Geenevasen, A. C. T. van Duin, L. A. van Niftrik, and M. S. M. Jetten.** 2002. Linearly concatenated cyclobutane lipids form a dense bacterial membrane. Nature **419:**708–712.
18. **Sinninghe Damsté, J. S., W. I. C. Rijpstra, J. A. J. Geenevasen, M. Strous, and M. S. M. Jetten.** 2005. Structural identification of ladderane and other membrane lipids of planctomycetes capable of anaerobic ammonium oxidation (anammox). FEBS J. **272:**4270–4283.
19. **Snel, B., P. Bork, and M. A. Huynen.** 1999. Genome phylogeny based on gene content. Nat. Genet. **21:**108–110.
20. **Stackebrandt, E., W. Ludwig, W. Schubert, F. Klink, H. Schlesner, T. Roggentin, and P. Hirsch.** 1984. Molecular genetic evidence for early evolutionary origin of budding peptidoglycan-less eubacteria. Nature **307:**735–737.
21. **Staley, J. T., H. Bouzek, and C. Jenkins.** 2004. Eukaryotic signature proteins of *Prosthecobacter dejongeii* and *Gemmata* sp. Wa-1 as revealed by in silico analysis. FEMS Microbiol. Lett. **243:**9–14.
22. **Strous, M., E. Pelletier, S. Mangenot, T. Rattei, A. Lehner, M. W. Taylor, M. Horn, H. Daims, D. Bartol-Mavel, P. Wincker, V. Barbe, N. Fonknechten, D. Vallenet, B. Segurens, C. Schenowitz-Truong, C. Medigue, A. Collingro, B. Snel, B. E. Dutilh, H. J. M. Op den Camp, C. van der Drift, I. Cirpus, K. T. van de Pas-Schoonen, H. R. Harhangi, L. van Niftrik, M. Schmid, J. Keltjens, J. van de Vossenberg, B. Kartal, H. Meier, D. Frishman, M. A. Huynen, H. Mewes, J. Weissenbach, M. S. M. Jetten, M. Wagner, and D. Le Paslier.** 2006. Deciphering the evolution and metabolism of an anammox bacterium from a community genome. Nature **440:**790–794.
23. **Teeling, H., T. Lombardot, M. Bauer, W. Ludwig, and F. O. Glöckner.** 2004. Evaluation of the phylogenetic position of the planctomycete 'Rhodopirellula baltica' SH1 by means of concatenated ribosomal protein sequences, DNA-directed RNA polymerase subunit sequences and whole genome trees. Int. J. Syst. Evol. Microbiol. **54:**791–801.
24. **van Niftrik, L. A., J. A. Fuerst, J. S. Sinninghe Damsté, J. G. Kuenen, M. S. M. Jetten, and M. Strous.** 2004. The anammoxosome: an intracytoplasmic compartment in anammox bacteria. FEMS Microbiol. Lett. **233:**7–13.
25. **Woese, C. R.** 1987. Bacterial evolution. Microbiol. Rev. **51:**221–271.