The ISME Journal (2007) 1, 419–435

www.nature.com/ismej

ORIGINAL ARTICLE

Fosmids of novel marine *Planctomycetes* from the Namibian and Oregon coast upwelling systems and their cross-comparison with planctomycete genomes

Dagmar Woebken[1], Hanno Teeling[2], Patricia Wecker[2,3], Alexandra Dumitriu[3], Ivaylo Kostadinov[2,3], Edward F DeLong[4], Rudolf Amann[1] and Frank O Glöckner[2,3]
[1]*Department of Molecular Ecology, Max Planck Institute for Marine Microbiology, Bremen, Germany;* [2]*Microbial Genomics Group, Max Planck Institute for Marine Microbiology, Bremen, Germany;* [3]*School of Engineering and Sciences, Jacobs University Bremen gGmbH, Bremen, Germany and* [4]*Division of Biological Engineering & Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA*

Planctomycetes **are widely distributed in marine environments, where they supposedly play a role in carbon recycling. To deepen our understanding about the ecology of this sparsely studied phylum six planctomycete fosmids from two marine upwelling systems were investigated and compared with all available planctomycete genomic sequences including the as yet unpublished near-complete genomes of *Blastopirellula marina* DSM 3645^T and *Planctomyces maris* DSM 8797^T. High numbers of sulfatase genes (41–109) were found on all marine planctomycete genomes and on two fosmids (2). Furthermore, C1 metabolism genes otherwise only known from methanogenic *Archaea* and methylotrophic *Proteobacteria* were found on two fosmids and all planctomycete genomes, except for '*Candidatus* Kuenenia stuttgartiensis'. Codon usage analysis indicated high expression levels for some of these genes. In addition, novel large families of planctomycete-specific paralogs with as yet unknown functions were identified, which are notably absent from the genome of '*Candidatus* Kuenenia stuttgartiensis'. The high numbers of sulfatases in marine planctomycetes characterizes them as specialists for the initial breakdown of sulfatated heteropolysaccharides and indicate their importance for recycling carbon from these compounds. The almost ubiquitous presence of C1 metabolism genes among *Planctomycetes* together with codon usage analysis and information from the genomes suggest a general importance of these genes for *Planctomycetes* other than formaldehyde detoxification. The notable absence of these genes in *Candidatus* K. stuttgartiensis plus the surprising lack of almost any planctomycete-specific gene within this organism reveals an unexpected distinctiveness of anammox bacteria from all other *Planctomycetes*.**
The ISME Journal (2007) **1,** 419–435; doi:10.1038/ismej.2007.63; published online 9 August 2007
Subject Category: integrated genomics and post-genomics approaches in microbial ecology
Keywords: *Blastopirellula marina*; C1 metabolism genes; *Planctomyces maris*; Planctomycete; marine upwelling; sulfatase genes

Introduction

Upwelling in marine coastal regions is a common phenomenon, which can, for example, be caused by alongshore winds that drive surface water away from the shore and bring nutrient-rich deep-sea water to the sun-lit surface. This results in an enhanced primary production, and the mineralization of the high amounts of biomass in marine upwelling systems results in depletion of oxygen in deeper waters forming the so-called oxygen minimum zone (OMZ). Massive losses of fixed nitrogen occur in the OMZ, mainly caused by planctomycetes living via anammox—the comproportionation of nitrite and ammonia to dinitrogen gas (Kuypers

Correspondence: H Teeling, Microbial Genomics Group, Max Planck Institute for Marine Microbiology, Celsiusstrasse 1, Bremen 28359, Germany.
E-mail: hteeling@mpi-bremen.de
Received 7 May 2007; revised 26 June 2007; accepted 27 June 2007; published online 9 August 2007

et al., 2005). Planctomycetes are also part of the microbial communities that are attached to macroscopic detrital aggregates (DeLong *et al*., 1993; Crump *et al*., 1999), where they are likely involved in the breakdown of complex heteropolysaccharides (Glöckner *et al*., 2003). Such marine snow particles play a major role in highly productive marine ecosystems like upwelling regions.

These observations notwithstanding, the overall ecological functions of the *Planctomycetes* are not well studied and hence not well understood. This might be attributed to the fact that in terms of abundance, *Planctomycetes* do not belong to the major players in marine surface waters. Typically, they amount only for a few percent of the total microbial biomass in coastal waters, and even much less in the marine pelagial (Rusch *et al*., 2007), while abundances are much higher in marine sediments (Rusch *et al*., 2003; Inagaki *et al*., 2006; Musat *et al*., 2006). However, it was recently shown that during diatom blooms in Oregon coastal waters *Pirellula*-related planctomycetes can reach cell numbers as high as $4 \times 10^7\,l^{-1}$. Since these planctomycetes have often been observed to be directly associated with algae, a direct interaction and carbon flow between algae and planctomycetes has been presumed (Morris *et al*., 2006). Moreover, abundances do not necessarily equal importance. Key metabolisms can be performed by minor groups, as we have learned from the long-neglected anammox planctomycetes, which have been shown to play such an important role in the global nitrogen cycle. Previous studies have indicated that other groups of marine planctomycetes are specialized on the initial breakdown of highly complex carbohydrates, thereby making them accessible to other organisms, and thus may play an important role in global carbon cycle (Glöckner *et al*., 2003).

In general, the *Planctomycetes* constitute an independent phylum within the domain *Bacteria* (Woese, 1987), consisting of only one single family (*Planctomycetaceae*) with six accepted (*Rhodopirellula*, *Blastopirellula*, *Pirellula*, *Planctomyces*, *Isosphaera*, *Gemmata*) (Schlesner *et al*., 2004) and four candidate genera (Kuenenia, Brocadia, Scalindua, Anammoxoglobus). The phylogeny of the *Planctomycetes* has been under debate for some time with conflicting views of them as being either rapidly evolving (Woese, 1987; Fuerst, 1995), deep (Stackebrandt *et al*., 1984) or even deepest branching within the bacterial domain (Brochier and Philippe, 2002) or being remotely related to the *Chlamydiae* (Weisburg *et al*., 1986; Liesack *et al*., 1992; Teeling *et al*., 2004). Recent studies suggest that the *Planctomycetes* are part of the so-called PVC superphylum, a monophyletic clade that besides the *Planctomycetes* is formed by the phyla *Chlamydiae*, *Verrucomicrobia, Lentisphaerae*, and the candidate phyla 'Poribacteria' and OP3 (the last two of which have no cultured representatives so far) (Fieseler *et al*., 2004; Wagner and Horn, 2006).

All *Planctomycetes* known to date are characterized by a unique set of characteristic morphological features. Their cells are organized in a polar manner within some non-prosthecate appendages (stalks) or polar holdfast structures. In addition, planctomycete cells have a generative pole from which daughter cells are produced in a yeast-like budding process (Fuerst, 1995). A life cycle has been described for some planctomycetes that resembles that of *Caulobacter crescentus* with flagellated, non-reproductive swarmer cells and non-motile but reproductive adult cells (Tekniepe *et al*., 1981; Franzmann and Skerman, 1984; Fuerst, 1995; Glöckner *et al*., 2003). The cell walls of *Planctomycetes* lack the common bacterial cell wall constituent peptidoglycan, but instead appear in some species to consist of proline and cystein-rich proteins that are stabilized by disulfide cross-links (König *et al*., 1984; Liesack *et al*., 1986). Likewise, S-layer-like arrayed proteins have been reported (Jetten *et al*., 2002). The cytoplasmic membranes of planctomycetes contain characteristic pore-like crateriform structures that are either evenly distributed over the whole surface (genus *Planctomyces*) or confined around the reproductive cell pole (genus *Pirellula*) (Liesack *et al*., 1986). In addition, the DNA of planctomycetes is highly compacted and hence often visible on electron micrographs as so-called nucleoids.

The most distinctive feature of the *Planctomycetes,* however, is their internal compartmentalization (Lindsay *et al*., 1997, 2001; Fuerst, 2005). Representatives of the genera *Rhodopirellula*, *Blastopirellula*, *Pirellula*, *Planctomyces* and *Isosphera* have been shown to contain a single, large, membrane-bounded compartment, termed the pirellulosome in *Pirellula* species. Additional structures have been reported for *Gemmata obscuriglobus* UQM 2246^T and the anammox planctomycete '*Candidatus* Brocadia anammoxidans' (Lindsay *et al*., 2001). *G. obscuriglobus* UQM 2246^T has an additional double-membrane-bounded compartment enclosing its nucleoid and *Candidatus* Brocadia anammoxidans a special compartment, termed the anammoxosome. The anammoxosome's membrane contains unique ladderane lipids (Sinninghe Damsté *et al*., 2002), and has been proposed to separate hazardous hydrazine from the cytoplasm in the course of the anammox process. Acccording to present knowledge, this process is carried out exclusively by a group of bacteria that branches deeply within the *Planctomycetes*. Since these anammox planctomycetes are of industrial importance for wastewater treatment (Jetten *et al*., 1997) they are intensely studied.

Apart from that, our knowledge on planctomycetal physiologies is quite limited. Most of the strains so far isolated in pure culture have been from aquatic and aerobic habitats (Bauld and Staley, 1976; Franzmann and Skerman, 1984; Giovannoni *et al*., 1987b; Schlesner, 1994) and they are all obligate or facultative aerobic chemoheterotrophs that use

carbohydrates as their major source of carbon. However, molecular methods like fluorescence *in situ* hybridization (FISH) and 16S ribosomal RNA (rRNA) gene sequencing as well as targeted cultivation efforts have revealed a much broader distribution of the *Planctomycetes* in the environment. They were detected in the water column and sediments of fresh water lakes (Neef *et al.*, 1998; Miskin *et al.*, 1999; Wang *et al.*, 2002; Kalyuzhnaya *et al.*, 2004, 2005a), hot springs (Giovannoni *et al.*, 1987a), in the water column (DeLong *et al.*, 1993; Vergin *et al.*, 1998; Gade *et al.*, 2004) as well as in shallow and deep sea sediments of marine systems (Llobet-Brossa *et al.*, 1998; Rusch *et al.*, 2003; Inagaki *et al.*, 2006; Musat *et al.*, 2006) and in oxic and anoxic soils (Wang *et al.*, 2002). Furthermore, planctomycetes have been detected in marine sponges (Fuerst *et al.*, 1998, 1999; Pimentel-Elardo *et al.*, 2003) and in the hepatopancreas of the crustacean *Panaeus monodon* (Fuerst *et al.*, 1991, 1997), in freshwater (Crump *et al.*, 1999), marine detritus particles (DeLong *et al.*, 1993; Fuerst, 1995; Crump *et al.*, 1999) and attached to diatoms (Morris *et al.*, 2006).

At present, only a few *Planctomycetes* have been investigated by whole genome sequencing. So far, the genome of the marine planctomycete *Rhodopirellula baltica* SH 1^T (formerly *Pirellula* sp. strain 1) is the only one that is completely closed (Glöckner *et al.*, 2003). Besides, four largely completed planctomycete genomes are available. The anammox planctomycete '*Candidatus* Kuenenia stuttgartiensis' was investigated by a metagenomic approach from a wastewater treatment plant enrichment culture (Strous *et al.*, 2006). In addition, in collaboration with us, the Gordon and Betty Moore foundation has funded draft sequencing of the two marine planctomycetes *Blastopirellula marina* DSM 3645^T (Schlesner *et al.*, 2004) and *Planctomyces maris* DSM 8797^T (Bauld and Staley, 1976) and Ward and co-workers at The Institute for Genomic Research (TIGR) have generated an early draft of the genome of the freshwater isolate of *Gemmata obscuriglobus* UQM 2246^T (Franzmann and Skerman, 1984). A further draft of a *Gemmata* sp. genome, strain Wa-1, has been produced by Integrated Genomics (Chicago, IL, USA), that has recently been investigated with respect to a possible role of the ancestor of the *Planctomycetes* in the evoluton of the *Eukarya* (Staley *et al.*, 2005).

Despite these genome sequencing projects, there is a need for further direct genomic and metagenomic studies of planctomycetes, since we know the group is diverse and its members are often difficult to culture. Hence, the true extent of the diversity of this group may only be accessible via metagenomic studies. To broaden our understanding about marine *Planctomycetes*, we present here comparative sequence analysis of six novel *Planctomycete* fosmids from two different upwelling systems and their comparison with all of the available *Planctomycete* genomic sequences (see Supplementary Figure 1 for sampling sites). Four of the fosmids were retrieved from the OMZ of the Benguela upwelling system at the Namibian coast, and two originate from 200 m depth off the coast of Oregon (Stein *et al.*, 1996; Vergin *et al.*, 1998). At the same time, this is the first study that includes results from our annotation of the almost complete genomes of *Blastopirellula marina* DSM 3645^T and *Planctomyces maris* DSM 8797^T. With these new draft genomes, we now have genomic sequences from four of the six accepted planctomycete genera. In this study we investigated planctomycete-specific genes as well as metabolic genes like sulfatases and those involved in the C1 metabolism.

Materials and methods

Metagenome sampling

In this study, two metagenome libraries were investigated: the first was constructed from picoplankton samples taken during a cruise in the eastern North Pacific at 200 m depth off the Oregon coast (44.012 N, 124.955 W) in August 1992 as described in Stein *et al.* (1996). The second was constructed from Namibian shelf water sampled during an R/V *Meteor* cruise in March/April 2003 at station M202 (22.64°S and 14.30°E) (Supplementary Figure 1). This sample was taken from 52 m depth, where 2 μM nitrite, 4.6 μM nitrate, and oxygen and ammonium concentration below the detection limit were measured (Kuypers *et al.*, 2005). About 500 l seawater were brought onto pre-combusted (at 450°C) fiber glass filters (GFF; nominal pore size, 0.7 μm) and stored at −80°C until further processing.

Fosmid library construction

The metagenome library from the North Pacific was constructed as described in Stein *et al.* (1996). The library from the Namibian upwelling system was constructed as follows: high molecular weight DNA was extracted according to the protocol from Zhou *et al.* (1996). Cell saver tips (that is, tips with an extra large opening) were taken for DNA extraction to avoid shearing the DNA. To avoid DNA loss due to DNA binding to the filter glass material, cells were washed off the filters with the extraction buffer without proteinase K and SDS before the extraction procedure.

After DNA extraction, the RNA was digested with RNAse A in 0.5 × TE (final RNAse A concentration: 100 μg/ml) for 1 h at room temperature. Subsequently, the RNAse was inactivated at 60°C for 10 min, an equal amount of chloroform/isoamyl alcohol mixture was added and the supernatant was precipitated with 1/10 volume of NaOAc and 0.6 volume of isopropanol (incubation for 1 h at room temperature and centrifugation). The DNA was stored in 0.5 × TE and the ends of the DNA were

filled blunt-ended according to the protocol of the Copy Control Fosmid Library Production Kit (EPICENTRE Biotechnologies, Madison, WI, USA).

We selected DNA of 30–45 kb for subsequent ligation into the vector via application of pulsed-field gel electrophoresis (PFGE) (1% LMT agarose; in 0.5 × TBE; program of PFGE: run time 12 h at 14°C; angle of 120°; 6 V/cm; initial switch time 1 s, final switch time 10 s; afterwards for 1.3 h initial and final switch time 2 min). After cutting the band of the desired size (30–45 kb), we equilibrated the DNA three times in 1 × TE (each 30 min) and digested the agarose with β-agarase (1000 U/ml; 1 μl/100 mg agarose gel). The DNA was concentrated and washed with 1 × TE using a Microcon tube (Millipore, Billerica, MA, USA) and subsequently eluted in PCR water.

The DNA was ligated into CopyControl pCC1FOS vectors (EPICENTRE Biotechnologies, Madison) at 4°C for 2 days. The vectors were packaged into the MaxPlax lambda phage (EPICENTRE Biotechnologies). *Escherichia coli* cells of the strain EPI300-T1R (EPICENTRE Biotechnologies) were infected with the phages upon reaching an OD of 0.8–1.0 and plated onto LB-chloramphenicol plates (12.5 μg/ml). About 10 000 grown colonies were transferred into LB medium containing chloramphenicol (12.5 μg/ml), $MgSO_4$ (10 mM), and glycerol (8%) and stored at −80°C. Additionally, clones were pooled according to the row-column-plate scheme (Asakawa *et al*., 1997) and incubated overnight at 37°C. The DNA was subsequently extracted using the 96-well Montage plasmid preparation kit (Millipore) and stored at −80°C.

Fosmid library screening

The Oregon coast metagenome library was screened for fosmids containing planctomycete 16S rRNA genes as described in Vergin *et al*. (1998). Four planctomycete fosmids were detected in 3552 screened clones and from those four, fosmids 5H12 and 6N14 were provided by courtesy of Edward DeLong.

The fosmid library from the Namibian upwelling system was screened by PCR with the planctomycete-targeting primer Pla46F (Neef *et al*., 1998) and the universal primer 1392R (Pace *et al*., 1986). In view of the pooling scheme, we could trace back the positive PCR products to single clones of the fosmid library. The 16S rRNA genes were sequenced and their phylogeny was reconstructed using the ARB software (Ludwig *et al*., 2004) in conjunction with the SILVA database (www.arb-silva.de). In total, 10 000 clones were screened and 12 clones with planctomycete DNA were obtained.

Insert size determination

The respective clones were induced for high copy number according to the instructions of the Copy-Control Fosmid Library Production Kit (EPICENTRE Biotechnologies). Fosmids were then isolated following the QIAprep Spin Miniprep Protocol (Qiagen, Hilden, Germany). Aliquots of 25 μl of fosmid DNA were digested with *Not*I (10 000 U/μl), and the digestion was stopped by heating at 65°C for 10 min and quickly placed on ice. The sizes of the resulting fragments were checked by PFGE (0.5 × TBE, 1% pulsed field certified agarose (Bio-Rad, Hercules, CA, USA), initial switch time 5 s, final switch time 15 s, 6 V/cm, angle: 120°; run time: 17 h, 14°C).

Fosmid sequencing

Clones 5H12 and 6N14 from the Oregon coast planctomycete fosmids were selected and sequenced at Integrated Genomics (Jena, Germany). From the 12 planctomycete fosmids from the Namibian metagenome library, four fosmids (3FN, 6FN, 8FN and 13FN) were chosen for sequencing based on the phylogenetic position of the planctomycete 16S rRNA gene and the determined insert size. For that purpose, large amounts of fosmid DNA were extracted after induction of the fosmids with the Qiagen Large-Construct Kit (Qiagen). Shotgun sequencing of the fosmids was conducted by AGOWA (Berlin, Germany) resulting in a single contig each.

Gene prediction and annotation

Gene prediction was adapted for each sequence individually. For the rather short fosmid sequences, all open reading frames (ORFs) exceeding 90 nucleotides were taken into account for annotation. Overpredicted genes were sorted out during manual annotation.

Preliminary sequence data from *Gemmata obscuriglobus* UQM 2246^T were obtained from The Institute for Genomic Research (TIGR) through the website at http://www.tigr.org and GLIMMER v2 (Delcher *et al*., 1999) was used for the gene prediction of the *G. obscuriglobus* UQM 2246^T draft genome. For the published genomes of *R. baltica* SH 1^T and '*Candidatus* K. stuttgartiensis', the original gene prediction was retained. For the two planctomycete draft genomes sequenced by the Moore foundation, the supplied gene prediction was evaluated against an in-house gene prediction pipeline (MORFind, unpublished) that post-processes the outputs of CRITICA (Badger and Olsen, 1999), GLIMMER v2 (Delcher *et al*., 1999) and Zcurve (Guo *et al*., 2003). In the case of *P. maris* DSM 8797^T the original gene prediction was kept since it could not be further enhanced, but for *B. marina* DSM 3645^T, a more accurate gene prediction could be generated by additional cross-comparisons with the other planctomycete genomes.

Annotation was accomplished with the GenDB v2 system (Meyer *et al*., 2003), using various bioinfor-

matic tools for each predicted gene ranging from similarity searches against sequence databases (NCBI nr, NCBI nt, SwissProt) and protein family databases (Pfam, Prosite, InterPro, COG) to signal peptide- (SignalP v2.0; Nielson *et al.*, 1997) and transmembrane helix predictions (TMHMM v2.0 Krogh *et al.*, 2001). From these predictions, an automatic annotation was generated using the fuzzy logic-based autoannotation tool MicHanThi (Quast, 2006). High-quality annotations were generated by manual revision of each gene's annotation for the fosmids and the draft genome of *B. marina* DSM 3645^T.

Transcriptome analysis

R. baltica SH 1^T cells were grown as batch cultures with glucose as carbon and ammonium chloride as nitrogen source on a rotary shaker (100 r.p.m.) at 28°C in the dark (Rabus *et al.*, 2002). The medium used was a modified M13a medium (Schlesner, 1994) with the original complex substrates (yeast extract, peptone) replaced by 10 mM glucose and 1 mM ammonium chloride. Cultures were harvested by centrifugation (Beckman CoulterTM AvantiTM J-20XP, JA10 Rotor, 20 min, 6000 r.p.m., 4°C), re-suspended in 10 × TAE buffer and then re-centrifuged to cell pellets that were shock-frozen in liquid nitrogen and stored at −80°C.

The set of oligonucleotides corresponding to the whole genome of *R. baltica* SH 1^T (*Pirellula* AROS Version 1.0) was purchased from Operon (Cologne, Germany) and diluted to 20 μM concentration in Micro Spotting Solution Plus spotting buffer (Telechem, Sunnyvale, CA, USA). Spotting was done in three replicates onto GAPS II aminosilane slides (Corning, Schiphol-Rijk, Netherlands) using a SpotArray 24 spotting device (Perkin Elmer, Wellesley, MA, USA). Post-processing and blocking of the slides were done according to the manufacturer's instructions. For hybridization at least 2 μg of Alexa 546 dye-labeled and 2 μg of Alexa 647 dye-labeled total cDNA were used. Blocking, hybridization and washing were carried out in an automated hybridization station HS400 (Tecan, Crailsheim, Germany).

Slides were scanned at a resolution of 5 μm using a ScanArray Express Microarray scanner (Perkin Elmer). The image analysis software provided with this scanner was used for automatic spot detection and signal quantification. Raw data were automatically processed using the microarray data analysis software tool MADA (http://www.mpi-bremen.de/en/mada).

Sequence availability

Fosmid sequences are available under the following accession numbers from GenBank: EF591884 (5H12), EF591885 (6N14), EF591886 (3FN), EF591887 (6FN), EF591888 (8FN), EF591889 (13FN). The almost complete genomes of *B. marina* DSM 3645^T and *P. maris* DSM 8798^T have been sequenced in the framework of the Microbial Genome Sequencing Project of the Gordon and Betty Moore foundation (http://www.moore.org/microgenome/). Gene predictions and annotations of these sequences are available on request.

Comparative genomics

All comparative analyses were performed using custom-made JAVA-based frameworks that operate directly on GenDB MySQL databases and allow working with multiple of these databases at once (details about authorships and code are available on request).

Planctomycete-specific genes were identified by searching all genes of the investigated planctomycete sequences, with BLAST against an in-house database (genomesDB). This database was constructed from the proteome FASTA files of all fully sequenced bacterial and archaeal genomes plus some almost complete genome drafts (440 on April 2007), in which each protein was tagged by a unique numerical identifier in the header that encodes the species, the chromosome and the gene itself. Two approaches were taken to determine whether or not a gene was planctomycete-specific. In the first, two E-value thresholds were used as criteria to separate planctomycete and non-planctomycete hits: an upper threshold (E-6) was used, above which all BLAST hits were regarded as insignificant noise. Hence, in order for a gene to be planctomycete-specific, no non-planctomycete hits were allowed below this boundary. In addition, a lower threshold (E-15) was used, below which BLAST hits were regarded as targeting the same gene. Only those genes having planctomycete hits below this boundary were considered. Genes complying with both of these criteria were regarded as planctomycete-specific. In the second approach, only those genes were regarded as planctomycete-specific where all BLAST hits better than E-10 exclusively targeted *Planctomycetes*. In both cases, self-matching BLAST hits and hits to intragenomic homologs were filtered from the analysis.

Besides planctomycete-specific genes, we also searched for orphaned, genome-specific genes, which are genes without known homologs in any other genome. A gene was regarded as orphaned, if it had no BLAST hit below E-7 in the non-redundant NCBI nr or the genomesDB databases.

Codon usage analysis

Codon usage analysis was carried out with CIAJava (Carbone *et al.*, 2003) and codonw (Peden, 1999). CAIJava was used for each genome with the full set of predicted genes and with the standard 15 iterations. Codonw was used for each of the genomes with a high-quality training set of genes

for the initial correspondence analysis that was generated by filtering all genes smaller than 300 bp as well as genes coding for hypothetical proteins, phage proteins, transposases and integrases. Thereafter, the extracted codon usage characteristics were used to analyze the codon usage of all genes.

Results

Characterization and phylogenetic assignment of the obtained fosmid sequences

The size of the fosmid inserts ranged from 34.6 kb (fosmid 8FN) to 42.5 kb (fosmid 6N14) (Table 1) with a G+C-content from ~48% to ~60%, which reflects the G+C-content variation of the *Planctomycete* genomes (Table 2). The number of ORFs per fosmid insert ranged from 27 to 33.

Based on 16S rRNA analysis, phylogenetic assignments were made for the fosmid insert sequences obtained from the Namibian (3FN, 6FN, 8FN, 13FN) and Oregon (5H12, 6N14) coasts. A phylogenetic reconstruction based on tree calculation with neighbor joining, maximum likelihood and maximum parsimony algorithms (without and with 50% position variability filter) is shown in Figure 1.

Fosmid insert 13FN exhibited highest 16S rRNA sequence similarity to an uncultured planctomycete sequence from Lake Bonney, a permanently ice-covered lake in Antarctica (DQ015853, 94.9% identity). Its closest cultivated planctomycete was *Planctomyces maris* DSM 8797^T (87.5% identity). Likewise, the 16S rRNA of *P. maris* DSM 8797^T was also the closest cultured relative to fosmid insert 3FN (88.7% identity), which however, was most similar to a planctomycete sequence from a hydrocarbon-contaminated soil (DQ298013, 89.2%). Fosmid insert 6FN had a very high similarity to the 16S rRNA sequence from an uncultured marine snow-associated planctomycete (L10942, 99.0% identity). The sequence identity to the next cultured relatives, two *Blastopirellula marina* DSM 3645^T-related strains (AJ231179 and AJ231180), was considerably lower (87.5% identity). Clone 8FN exhibited highest sequence identity to a sequence originating from a deep-sea mud volcano in the eastern Mediterranean (AY592598, 92% identity). The next cultured relative was a *Blastopirellula*-related planctomycete isolated from the postlarvae of the giant tiger prawn (*Penaeus monodon*) (X86391, 89.9% identity).

Fosmid 6N14 had highest sequence identity with an uncultured planctomycete from heavy metal-contaminated surface sediments in the North Sea (98.2%, DQ351810) and is related to *Rhodopirellula baltica* SH 1^T (95.7% identity). Clone 5H12 grouped, together with 8FN and 6FN, within a cluster of

Table 1 Characterization of fosmids: insert size, G+C-content, number of ORFs and sampling location

	3FN	*6FN*	*8FN*	*13FN*	*5H12*	*6N14*
Fosmid insert size (kb)	37.0	40.7	34.6	36.7	41.0	42.5
G+C-content (%)	59.7	47.8	50.8	55.0	51.0	52.4
Number of ORFs after manual annotation	27	28	33	31	31	32
Sampling location		Namibian upwelling system (Kuypers *et al.*, 2005)			Oregon upwelling system (Stein *et al.*, 1996)	

Table 2 Characterization of the planctomycete genomes: genome size, G+C-content, number of sulfatases, number of orphan and number of *Planctomycete* group-specific genes (GSG) and their domains of unknown functions (Pfam Accession numbers are indicated in parentheses)

	R. baltica	*P. maris*	*B. marina*	*G. obscuriglobus*	*K. stuttgartiensis*
Genome size	7145 kb	7775 kb	6655 kb	9106 kb	4218 kb
G+C content (%)	55.4	51.2	57.2	67.1	40.8
No. of predicted genes	7325	6531	5600	(13 024)[a]	4710
No. of orphan genes[b]	2897 (39.5%)	1559 (24.5%)	1088 (19.4%)	(5989)[a] (46.0%)[a]	1364 (29.0%)
No. of GSG w/o KS[c]	125	200	249	154	—
No. of GSG with KS[c]	3	3	3	3	2
No. of sulfatases[d]	109 (15.3/Mb)	83 (10.7/Mb)	41 (6.2/Mb)	12 (1.3/Mb)	3 (0.7/Mb)
No. of Pfam profile DUF1501[d] (PF07394)	41	94	31	61	0
No. of Pfam profile PSCyt2/DUF1549[d] (PF07583)	41	68	32	43	0
No. of Pfam profile PSD1/DUF1553[d] (PF07587)	41	68	32	43	0
No. of Pfam profile PSCyt1 (CytC)[d] (PF07635)	53	54	32	27	0
No. of Pfam profile SBP_bac_10/DUF1559[d] (PF07596)	74	134	197	125	0
No. of Pfam profile DUF1551[d] (PF07585)	6	2	1	7	0

[a]Overprediction by Glimmer 2.
[b]E value E-7 or better.
[c]Upper boundary: E-15, lower boundary: E-6.
[d]E value E-5 or better.

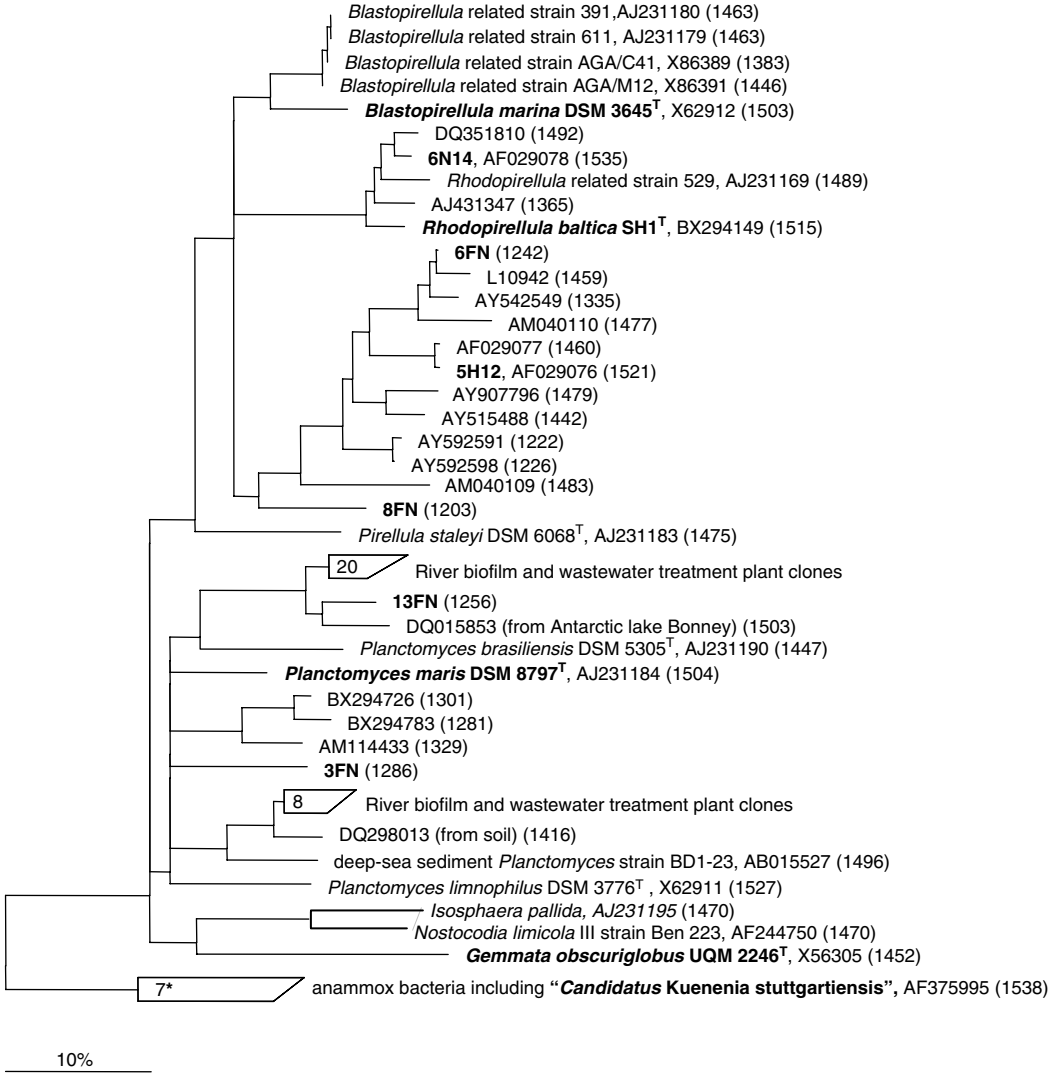

Figure 1 Phylogenetic tree based on 16S rRNA sequences showing the phylogenetic affiliation of fosmids and genomes used in this study. The consensus tree was constructed after tree calculation with neighbor joining, maximum parsimony and maximum likelihood algorithms with and without 50% position variability filters. Lengths of the sequences are indicated in parentheses. Anammox bacteria (group indicated by an asterisk comprising of '*Cand.* K. stuttgartiensis', '*Cand.* B. anammoxidans', '*Cand.* B. fulgida', '*Cand* A. propionicus', '*Cand.* J. asiatica', '*Cand.* S. sorokinii' and '*Cand.* S. brodae') were used as outgroup. The bar represents 10% estimated sequence divergence.

uncultured planctomycetes and showed highest sequence identity to a fosmid from the same metagenome library (uncultured *Pirellula* clone 6O13, AF029077, 99.7%, identity, (Vergin *et al.*, 1998)). The closest cultured representatives were a group of isolates retrieved from the postlarvae of *Penaeus monodon* (for example, X86389, 88.9% identity).

Genes involved in C1 metabolism

Previous studies have shown that the genomes of *R. baltica* SH 1^T and *G. obscuriglobus* UQM 2246^T both contain genes involved in tetrahydromethanopterin-(H_4MPT)-linked C1-compound conversions (Supplementary Figure 2) and the biosynthesis of the associated essential cofactor methanopterin (Glöckner *et al.*, 2003; Bauer *et al.*, 2004; Chistoserdova *et al.*, 2004).

In addition to the already described C1 metabolism genes in *R. baltica* SH 1^T and *G. obscuriglobus* UQM 2246^T, we could identify homologs to all known C1 genes in the draft genome of *B. marina* DSM 3645^T and, with the exception of *ORF21*, also in the *P. maris* DSM 9787^T draft. In contrast, none of these genes was found in the draft genome of *Candidatus* K. stuttgartiensis (Table 3 and Figure 2).

R. baltica SH 1^T and *G. obscuriglobus* UQM 2246^T have two copies of the formaldehyde-activating enzyme gene (*fae1* and *fae2*—31% and 28% amino acid (aa) identity, respectively). We also found two *fae* copies in *P. maris* DSM 8797^T (38% aa identity) but only one in *B. marina* DSM 3645^T. Within *R. baltica* SH 1^T, *fae1* is characterized by a rigidly

Table 3 Presence and absence of archaea-like and bacteria-like H_4MPT-dependent genes in the studied *Planctomycete* genomes and fosmids

Gene	*R.b.*[a]	*G.o.*[a]	*B.m.*[a]	*P.m.*[a]	*K.s.*[a]	*3FN*	*6FN*	*8FN*	*13FN*	*5H12*	*6N14*
ORF5	+	+	+	+	−	−	−	−	−	−	−
ORF7	+	+	+	+	−	−	−	−	−	−	−
ORF9	+	+	+	+	−	−	−	−	−	−	−
ORF17	+	+	+	+	−	−	−	−	−	−	−
ORF19	+	+	+	+	−	−	−	+	−	−	−
ORF20	+	+	+	+	−	−	−	+	−	−	−
ORF21	+	+	+	−	−	−	−	−	−	−	−
ORF22	+	+	+	+	−	−	−	−	−	−	−
ORFY	+	+	+	+	−	−	−	−	−	−	−
fae 1	+	+	+	+	−	−	−	−	−	−	−
fae 2	+	+	−	+	−	−	−	−	−	−	−
ftr[b]	+	+	+	+	−	−	−	−	−	−	−
Mch	+	+	+	+	−	−	−	−	−	−	−
mptG	+	+	+	+	−	−	−	+	−	−	−
mtdC	+	+	+	+	−	−	−	−	−	−	−
fmdABC[b]	+/−/+	+/+/+	+/+/+	+/+/+	−	−	−/−/+	−	−	−	−
ORF1	+	+	+	+	−	−	−	−	−	−	−
pabAB	+	−	+	+	−	−	−	+	−	−	−
ptS	+	+	+	+	−	−	−	−	−	−	−

[a]R.b.: *Rhodopirellula baltica* SH 1^T; G.o.: *Gemmata obscuriglobus* UQM 2246^T; B.m.: *Blastopirellula marina* DSM 3645^T and P.m.: *Planctomyces maris* 8797^T.
[b]Ftr together with the three subunits of Fmd forms the formyltransferase/hydrolase complex (fhc).

optimized codon usage indicating a high level of expression (Bauer *et al*., 2004). This is also the case within *B. marina* DSM 3645^T, whose *fae1* gene has the third highest codon adaptation index (CAI) of all of its genes, but not for the *P. maris* DSM 8797^T *fae* gene. In all four genomes, *fae1* is located upstream of a gene coding for the recently discovered methylene tetrahydromethanopterin gene *mtdC* (Vorholt *et al*., 2005). In *R. baltica* SH 1^T, *B. marina* DSM 3645^T and *P. maris* DSM 8797^T, *fae1* and *mtdC* are located adjacent to *ORFY*, a gene with as yet unknown function. Likewise, *mch* (encoding methenyl H_4MPT cyclohydrolase) is located directly upstream of *ORF5* (most likely involved in the biosynthesis of H_4MPT) in all four *Planctomycete* genomes. Clustering of formyl-H_4MPT-dehydrogenase subunits A and C (*fm(w)dA/C*) with formylmethanofuran:H_4MPT-formyltransferase (*ftr*) could be observed in *R. baltica* SH 1^T, while in *G. obscuriglobus* UQM 2246^T the *fm(w)dC* and *ftr* genes are co-located. Within *B. marina* DSM 3645^T and *P. maris* DSM 8797^T, the *ftr* gene was not located adjacent to any of the *fm(w)d* genes. In all four *Planctomycete* genomes, *ORF19* was found to be linked to *mptG*. *ORF19* is involved in H_4MPT biosynthesis (Chistoserdova *et al*., 2005), while *mptG* codes for β-RFAP synthase, an enzyme that catalyzes the first reaction distinguishing the methanopterin biosynthesis pathway from that of folate biosynthesis (Scott and Rasche, 2002; Chistoserdova *et al*., 2004).

Clustering of *ORF9* with *ORF21* was observed in three genomes: *R. baltica* SH 1^T, *B. marina* DSM 3645^T and *G. obscuriglobus* UQM 2246^T. These two genes are also involved in H_4MPT biosynthesis (Chistoserdova *et al*., 2005). Further co-localizations cover *ORF9* and *ORF1* in *B. marina* DSM 3645^T and *P. maris* DSM 8797^T, and *ORF7* and *pts* in *B. marina* DSM 3645^T, *P. maris* DSM 8797^T *and G. obscuriglobus* UQM 2246^T.

Genes involved in the H_4MPT-linked C1 metabolism were also found on two of the four fosmids from the Namibian OMZ (Table 3 and Figure 2). Fosmid 6FN harbors an instance of *fmdC* plus the gene *mobA* that encodes the molybdopterin-guanine dinucleotide biosynthesis protein A. Fosmid 8FN contains a larger C1 module comprising *ORF20-pabB-pabA-ORF19-mptG*. This complete module could also be identified in *B. marina* DSM 3645^T, while *P. maris* DSM 8797^T contained these genes in two separate clusters of *ORF20-pabB-pabA* and *ORF19-mptG*. *R. baltica* SH 1^T has the cluster *pabA-ORF19-mptG*, while ORF20 is isolated and pabB is co-localized with the C1 genes *ORF5* and *mch*. In *G. obsucriglobus* UQM 2246^T finally, only the genes *ORF19-mptG* from this cluster could be found. *ORF19* and *ORF20* are also co-localized in the methylotroph *M. extorquens* AM1 and have been shown to be essential for the biosynthesis of H_4MPT (Chistoserdova *et al*., 2005). *MptG* encodes a GHMP family kinase, and *pabA* and *pabB* the p-aminobenzoate synthase glutamine amidotransferase component II and the p-aminobenzoate synthase component I, respectively (Kalyuzhnaya *et al*., 2005b).

Expression profiling on glucose-grown *R. baltica* SH 1^T cells revealed that even in the absence of C1 compounds seven of the respective genes are expressed, namely *ORF21, fae, fm(w)dA, fm(w)dC, ORF7, mtdC* and *ORF1*.

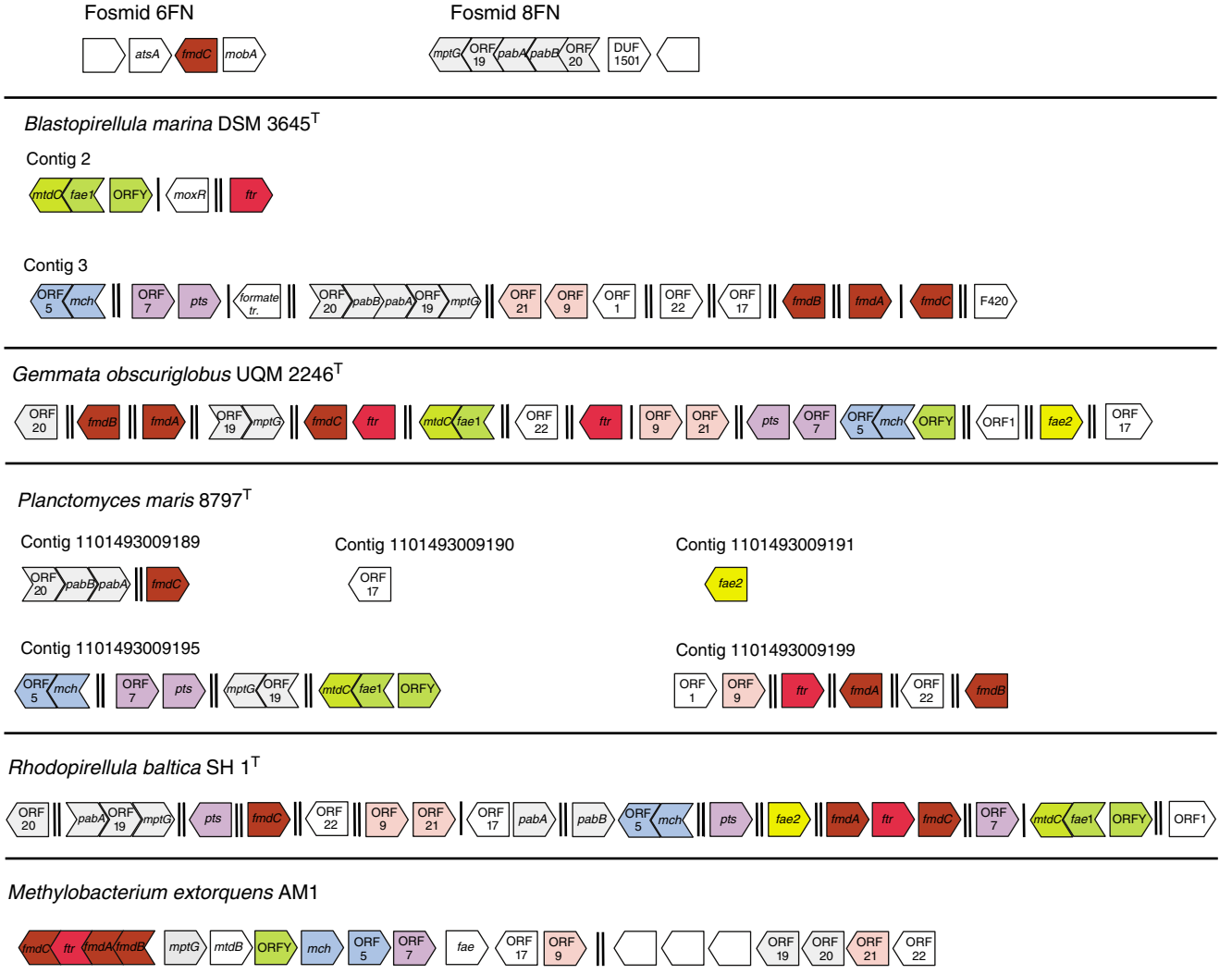

Figure 2 Genomic arrangement of genes involved in H_4MPT-dependent C1-transfer. Comparison of fosmids 6FN and 8FN from the Namibian upwelling system, genomes of the planctomycetes *Blastopirellula marina* DSM 3645^T, *Gemmata obscuriglobus* UQM 2246^T, *Planctomyces maris* 8797^T, *Rhodopirellula baltica* SH 1^T and the methylotrophic Alphaproteobacterium *Methylobacterium extorquens* AM1. Inserted arrows state that ORFs are organized in operons. A single line separating arrows indicate that ORFs are separated by at most 50 ORFs; double lines represent a separation by >50 ORFs.

Planctomycete-specific genes

Having multiple *Planctomycete* genomes at hand opens the possibility for the identification of planctomycete-specific genes. Among these, the ones that are present in all five planctomycetes investigated are of particular interest, since they define a *Planctomycete* and hence must include the genes coding for their unique cellular characteristics and planctomycete-specific metabolic traits.

To our surprise, only 2–3 genes were found when searching for such genes. A closer investigation revealed that '*Candidatus* K. stuttgartiensis' was almost devoid of genes that occur in all other planctomycetes. If left out, the remaining four planctomycetes had between 125 and 249 genes that were specific for them (Table 2). A considerable proportion of these planctomycete-specific genes were made up by large paralogous gene families that were defined by the occurrence of one or more specific domains of as yet unknown function (DUF), in particular the DUF1501, PSD1 (DUF1553) and DUF1559 as well as the planctomycete-specific cytochrome *c*-like domains PSCyt1 and PSCyt2 (DUF1549). These domains were previously identified as organism-specific domains within *R. baltica* SH 1^T (Studholme *et al.*, 2004). Frequently, the planctomycete-specific genes appeared in tandems. The most prominent of these tandems consisted of a gene carrying DUF1501 and a gene containing DUF1549 plus DUF1553 and sometimes an additional planctomycete-specific cytochrome *c* domain (Table 4). A remarkable high proportion of these genes had signal peptide predictions, especially the DUF1549/DUF1553 genes.

When the five planctomycete genomic sequences were searched for planctomycete-specific genes

Table 4 Statistical analysis of DUFs[a] in tandem structures on the planctomycete genomes

	R. baltica	*P. maris*	*B. marina*	*G. obscuriglobus*	*K. stuttgartiensis*
Number of tandem occurrences (search started with DUF1501)	39	ND	30	22	0
Percentage of genes with DUF1501 that appear in tandem	95.1%	ND	96.8%	36.1%	0
Number of genes containing DUF1549 and DUF1553 from tandems having signal peptides	21	ND	21	12	0
Percentage of secreted genes with DUF1549 and DUF1553 from tandems	53.8%	ND	70.0%	54.5%	0
Number of genes with DUF1549 and DUF1553 from tandems which also contain CytC domain	28	ND	18	11	0
Percentage of genes with DUF1549, DUF1553 and CytC domain present in tandems	71.8%	ND	60.0%	50.0%	0

DUF1553 = PSD1; DUF1549 = PSCyt2.
[a]DUF: domain of unknown function.

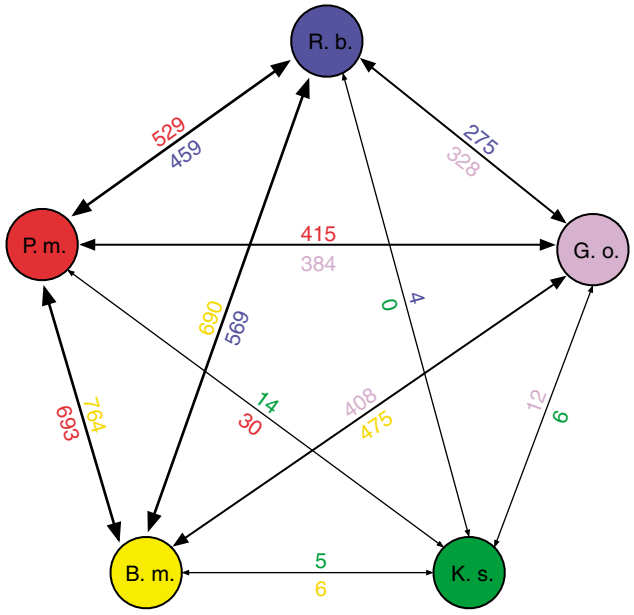

Figure 3 Genes for which BLAST hits below E-10 exclusively targeted *Planctomycetes*. Colors indicate the respective reference genome for the BLAST searches. Corresponding colors represent the number of hits in the respective target genome that is also expressed by thickness of the connection lines.

occurring in at least one other *Planctomycete* species, vastly diverging numbers were obtained (Figure 3). The genomes of *R. baltica* SH 1^T, *P. maris* DSM 8797^T and *B. marina* DSM 3645^T were found to share 429–764 genes, while *G. obscuriglobus* UQM 2246^T was more distant to these three with 275–475 shared genes. In *Candidatus* K. stuttgartiensis finally, merely 0–30 planctomycete-specific genes were found that were shared with at least one of the other four planctomycetes.

All five investigated genomes contained high proportions of hypothetical proteins of which large fractions were found to be orphaned, that is present only in one particular planctomycete with no known homolog in any other species (Table 2).

On all six sequenced fosmids, 26 out of 183 genes with reliable BLAST hits exclusively to *Planctomycetes* were found (Table 5). It is noteworthy that from these 26 planctomycete-specific genes, only a single one exhibited high similarity to a gene from *Candidatus* K. stuttgartiensis. As with the genome sequences, many of the planctomycete-specific genes on the fosmids contained the Pfam domains DUF1501, DUF1550 and DUF1559, and co-occurrences of the domains DUF1549 and DUF1533. Expression analysis of the closest homologs (BLAST) to these planctomycete-specific genes in glucose-grown *R. baltica* SH 1^T revealed two expressed genes, one homolog to 3FN_15 carrying a weak hit to a prenyltransferase repeat domain and one homolog to 8FN_23 that harbors a DUF1559 domain (Table 5).

Sulfatases

The number of genes encoding sulfatases varied considerably among the different *Planctomcetes* investigated (Table 2). While the marine planctomycetes *R. baltica* SH 1^T, *P. maris* DSM 8797^T and *B. marina* DSM 3645^T encode high numbers of sulfatases, only few sulfatases could be found in the non-marine planctomycetes *G. obscuriglobus* UQM 2246^T and '*Candidatus* K. stuttgartiensis'.

R. baltica SH 1^T codes for no less than 109 sulfatases (15.3 per Mb), which to date is the highest number found in any bacterial genome. In the *P. maris* DSM 8797^T draft genome we found 87 sulfatases (10.7/Mb) and in the one of *B. marina* DSM 3645^T 41 (6.16/Mb), whereas in the *G. obscuriglobus* UQM 2246^T draft sequence only 12 sulfatases were found (1.3/Mb) and in *Candidatus* K. stuttgartiensis only three (0.7/Mb). Additionally, one sulfatase could be found on fosmid 6FN (orf 6FN_6) from the Namibian upwelling system and one on fosmid 5H12 (orf 5H12_13) from the Oregon coast (Supplementary Tables 3 and 6). Using the whole-genome microarray, expression of the sulfatases with the best BLAST hits was checked in

Table 5 Overview of planctomycete-specific genes and their domains of unknown function (DUFs) detected on the investigated fosmids

Fosmid	*Gene*	*Function*	*Observations*
3FN	3FN_13		
	3FN_15	Protein containing terpenoid cylases/protein prenyltransferase alpha-alpha toroid domain	
	3FN_23	Protein containing DUF1549, DUF1553	
	3FN_25	Protein containing DUF1501, secreted	
	3FN_26	Protein containing DUF1549, DUF1553, secreted	
6FN	6FN_5		
	6FN_20	Protein containing DUF1501, secreted	
	6FN_21	Protein containing planctomycete cytochrome c domain, DUF1549 and DUF1553, secreted	
	6FN_26	Protein containing DUF1550, membrane	Hit also for *Candidatus* K. stuttgartiensis
8FN	8FN_8	Protein containing DUF1501, secreted	
	8FN_9		
	8FN_23	Protein containing DUF1559	
	8FN_33	Protein kinase superfamily protein, membrane	
13FN	13FN_31		
6N14	6N14_18		
	6N14_21		
	6N14_25		
5H12	5H12_7		
	5H12_8		
	5H12_9	Protein containing DUF1501	
	5H12_10	Protein containing DUF1553	
	5H12_11		
	5H12_17		
	5H12_25		
	5H12_26		
	5H12_29		

glucose-grown *R. baltica* SH 1^T. While the homolog to the sulfatase on fosmid 6FN did not show detectable levels of expression, the homolog on fosmid 5H12 was clearly expressed.

Discussion

In the present study, six *Planctomycete* fosmid insert sequences from two different marine upwelling systems were compared with all available complete or almost complete *Planctomycete* genomes. Taking the two new draft genomes of *B. marina* DSM 3645^T and *P. maris* DSM 8797^T into account, we now have genomic information from all described *Planctomycete* genera at hand except of the *Isosphaera* and *Pirellula* lineage.

R. baltica SH 1^T and *B. marina* DSM 3645^T were both isolated from the Kiel Fjord in the Baltic Sea (Schlesner, 1994) and are aerobic heterotrophs. They are marine representatives of the *Planctomycete* phylum, since they do not grow in freshwater media (Schlesner *et al.*, 2004). *P. maris* DSM 8797^T was isolated from shallow waters at Puget Sound, Washington, USA (Bauld and Staley, 1976, 1980) and is a heterotrophic, aerobic, marine planctomycete as well. *G. obscuriglobus* UQM 2246^T is a freshwater isolate from the Maroon Dam in Queensland, Australia (Franzmann and Skerman, 1984) and *Candidatus* K. stuttgartiensis was enriched from a wastewater treatment plant in Stuttgart, Germany (Schmid *et al.*, 2000).

Environmental function of Planctomycetes

The planctomycetes from marine habitats contained considerably higher sulfatase copy numbers in their genomes than the planctomycetes from freshwater habitats (Table 2). Likewise, one of the four fosmid sequences from the Namibian upwelling system and one of the two fosmids from the Oregon upwelling each contained one sulfatase. Moreover, recently a planctomycete fosmid was published from the North Pacific Subtropical Gyre (fosmid HF10_49E08, GenBank EF089402, 39.2 kb) that besides a proteorhodopsin contained no less than six sulfatases (McCarren and Delong, 2007). All these fosmids together amount to ~272 kb of sequence information with an average of 22 sulfatases per Mb. This exceeds even the sulfatase density observed within *R. baltica* SH 1^T, the marine planctomycete with the highest number of sulfatase encoding genes

known so far (Table 2). The presence of high sulfatase gene numbers might be attributed to a particular marine lifestyle. The most likely candidate substrates for these sulfatases are sulfated heteropolysaccharides, which are produced in large quantities in marine environments, for example by fish (chondritin in cartilage), red algae (agars and carrageenans) and brown algae (sulfated fucans). These compounds are of a great chemical complexity, and hence require a versatile repertoire of specific sulfatases for successful biodegradation, which explains the high sulfatase numbers within marine *Planctomycete* genomes.

We proposed earlier that sulfated polysaccharides are entrapped in marine snow aggregates, which are known to be colonized by planctomycetes. Sulfatase-rich planctomycetes are then supposed to be able to degrade these polysaccharides and subsequently use their carbon skeletons as an energy source (Glöckner *et al.*, 2003). Sulfatase activity has meanwhile been proven for *R. baltica* SH 1^T, and not only for linear but also for sterically demanding sulfated compounds (Wallner *et al.*, 2005). In addition, it has been demonstrated that *R. baltica* SH 1^T, while apparently not capable of degrading agar, does degrade carrageenan (Gurvan Michel, personal communication). Likewise, expression profiling showed that some of the sulfatases in *R. baltica* SH 1^T are expressed. This fits perfectly to the aforementioned proposed lifestyle for these organisms. A more in-depth analysis of the sulfatases in *B. marina* DSM 3645^T revealed that 39 of its sulfatases carried the essential canonical motif [CS]-x-[PA]-x-R (Dierks *et al.*, 1999; Berteau *et al.*, 2006), which in 37 cases could be extended to [CS]-x-[PA]-x-R-x(4)-[ST]-G. Thirty-four of these proteins had a proline at the third motif position, whereas four had a proline to alanine mutation. It is unclear whether this has any impact on the functioning or specificity of the respective enzymes, but the presence of the canonical motif indicates that these sulfatases are active. Considering the sheer amount of sulfated heteropolysaccharides that are produced in marine environments and the ubiquity of lateral gene transfer (LGT), it is highly unlikely that only representatives of the marine *Planctomycetes* have adapted to exploit this resource. Hence, it can be expected that high sulfatase gene numbers will be discovered in marine representatives from other lineages as well and likely play an important role in highly productive marine upwelling systems.

It was one of the most surprising findings in previous studies of the genomes of *R. baltica* SH 1^T and the draft genome of *G. obscuriglobus* UQM 2246^T that both code for proteins involved in tetrahydromethanopterin-(H_4MPT)-linked C1-compound conversions (Supplementary Figure 2) and the biosynthesis of the associated essential cofactor methanopterin (Glöckner *et al.*, 2003; Bauer *et al.*, 2004; Chistoserdova *et al.*, 2004). In subsequent studies, it could be demonstrated that such C1 genes are present in further uncultivated planctomycetes (Kalyuzhnaya *et al.*, 2005a, c). Before, these genes were believed to occur only in methanogenic and sulfate-reducing *Archaea*, and in methylotrophic *Alpha-, Beta- and Gammaproteobacteria*, where they play the central role in either the reductive or oxidative gain of energy from C1-substrates and in the detoxification of the hazardous metabolic intermediate formaldehyde (Chistoserdova *et al.*, 1998; Vorholt *et al.*, 1999; Marx *et al.*, 2003). Many of the H_4MPT-linked C1 genes have been identified by functional studies on the methylotroph *Methylobacterium extorquens* AM1 (Chistoserdova *et al.*, 2005) with the notable exception of *ORF1*, which is absent in *M. extorquens* AM1 but present in various other methylotrophs (see Supplementary Table 1 for an overview on the functions).

The parallel presence of H_4MPT-dependent genes in representatives of the *Archaea*, *Proteobacteria* and *Planctomycetes* has led to numerous theories about the origin and evolution of these genes (Bauer *et al.*, 2004; Chistoserdova *et al.*, 2004). One scenario assumes two LGT events by which the genes were first passed on from *Archaea* to *Proteobacteria* and from there to the *Planctomycetes*. Another scenario presumes that the last universal common ancestor was already equipped with C1-transfer genes, but that these genes were only preserved within few lineages during the microbial evolutionary radiation. This scenario is supported by a screening of the Sargasso Sea data set showing that H_4MPT-dependent genes occur in additional unknown bacterial lineages and thus are more wide-spread (Kalyuzhnaya *et al.*, 2005d). However, based on the present data it is not possible to decide between these alternate scenarios, especially since the phylogenetic position of the entire PVC superphylum is unclear. The easiest explanation would be to question the monophyly of the eubacteria, which has been proposed by Cavalier-Smith, (2006), but this discussion is clearly beyond the scope of this study.

Investigation of the studied planctomycete genomes and fosmids for genes involved in the conversion of C1 carbon compounds revealed that four of the genomes and two of the fosmids (8FN and 6FN) contained such genes (Table 3 and Figure 2). The notable exception in the genomes is the deep-branching anammox bacterium *Candidatus* K. stuttgartiensis, which lacks those genes completely. Earlier studies discussed that the clustering of C1-transfer genes in *Planctomycetes* is much looser than in the genomes of methylotrophic proteobacteria where the genes are arranged in a few main clusters (Bauer *et al.*, 2004; Kalyuzhnaya *et al.*, 2005b). Our study with four planctomycete genomes containing these genes supports this finding. The genes were scattered widely over the genomes and at most five genes were arranged in one cluster (Figure 2). This cluster of *Orf20-pabA-pabB-Orf19-mptG* was found on fosmid 8FN and in the genome

of *B. marina* DSM 3645^T. Most likely, this cluster reflects the original organization of these genes in the *Planctomycetes*, and in the course of evolution it was split into smaller modules, as for example in *P. maris* DSM 8797^T, where these genes are organized in the two distinct clusters *Orf20-pabA-pabB* and *Orf19-mptG*. The study also supports the previously postulated clustering patterns of C1 genes in *Planctomycetes*, in particular clustering of *mptG* with *orf19* and of *mtdC* with *fae1* (Bauer *et al.*, 2004; Kalyuzhnaya *et al.*, 2005b). However, the previous observation that the *fmdA*, *ftr* and *fmdC* are always co-located in all known bacterial genomes containing these *Archaea*-like genes (Bauer *et al.*, 2004; Kalyuzhnaya *et al.*, 2005b) could not be confirmed by the investigation of more *Planctomycete* genomes.

Up to now it has not been possible to answer the question whether these genes are functional in *Planctomycetes* and what role they play. So far, *R. baltica* SH 1^T could not be shown to grow on C1 substrates (Bauer *et al.*, 2004). On the other hand, codon usage analysis indicated high expression levels for *fae1* in *R. baltica* SH 1^T (Bauer *et al.*, 2004) and in *B. marina* DSM 3645^T. Furthermore, proteome analysis proved that *fae1*, *mtdC* and *mch* are expressed in *R. baltica* SH 1^T (Bauer *et al.*, 2004), and our microarray experiments with glucose-grown *R. baltica* SH 1^T cells have shown that even in the absence of an external C1 carbon source almost a third of the C1 genes in *R. baltica* SH 1^T are expressed. Another indication that the genes involved in H_4MPT-dependent C1-compound conversions are active in the planctomycetes is their arrangement in conserved modules (Figure 2). In summary, this suggests that the respective pathway is of great physiological and environmental importance to these planctomycetes. It has been proposed that the sole role of these genes within the *Planctomycetes* is the detoxification of formaldehyde (Chistoserdova *et al.*, 2004). However, to possess this pathway with its many steps solely as a means for formaldehyde detoxification does not seem to be necessary, since *R. baltica* SH 1^T, *B. marina* DSM 3645^T, *P. maris* DSM 8797^T and *G. obscuriglobus* UQM 2246^T contain glutathione-dependent formaldehyde dehydrogenases and *R. baltica* SH 1^T also contains a glutathione-independent formaldehyde dehydrogenase that can decompose formaldehyde in a much simpler way (Goenrich *et al.*, 2002). Therefore, it seems likely that these C1 metabolism genes in *Planctomycetes* play an important role in an as yet unknown context. The presence of C1 metabolism genes in the fosmids from the Nambian upwelling region indicates that this pathway is of importance in the suboxic waters of this highly productive region. The presence of non-anammox plancomycetes in this region further implies that these planctomycetes can cope with low oxygen concentrations. This is also supported by the fact that all of the investigated *Planctomycete* genomes with the exception of *Candidatus* K. stuttgartiensis harbor typical fermentation genes such as acetate/butyrate kinase or phosphoketolase. Nonetheless, these planctomycetes are considered as obligate aerobes since so far they have not been able to be cultivated under anoxic conditions.

All in all, the sulfatase-rich planctomycetes seem to be well adapted to the nutrient-rich conditions in marine upwelling systems and are also capable of thriving under oxygen-limiting conditions, which can occur seasonally (Oregon upwelling system) or permanently (OMZ of the Namibian upwelling system).

Features that set Candidatus *K. stuttgartiensis and the other* Planctomycetes *apart*

One striking result of our comparative genomics approach is the distinctness of '*Candidatus* K. stuttgartiensis'. It is almost completely devoid of genes that are shared with all other four investigated almost fully sequenced planctomycetes. In particular, the large paralogous gene families with otherwise planctomycete-specific domains that are even present on the small fosmids in this study are notably absent from '*Candidatus* K. stuttgartiensis'. Even when the E-value threshold is lowered to E-10, there are only 11 of such genes. Likewise, when searched for planctomycete-specific genes that are not present in all but shared by at least two planctomycetes, there are only very few of these genes in '*Candidatus* K. stuttgartiensis' (Figure 3). The same applies to the C1 carbon metabolism genes, which are absent from *Candidatus* K. stuttgartiensis but present in the four other almost fully sequenced planctomycetes, and also in the fosmids. As a consequence, the anammox bacteria have likely diverged at a very early stage from the last common planctomycete ancestor. The most parsimonious assumption is that this must have been before the radiation of planctomycete genera (*Pirellula*, *Blastopirellula*, *Rhodopirellula*, *Planctomyces*, *Gemmata* and *Isosphaera*), when the planctomycete-specific domains had not yet evolved and—if acquired by LGT—before the C1 metabolism genes were transferred. This is in agreement with phylogenetic studies based on 16S rRNA sequences and concatenated protein sequences, according to which anammox bacteria are deep branching within the *Planctomycetes* (Strous *et al.*, 1999, 2006). However, the extent of the distinctness of '*Candidatus* K. stuttgartiensis' is surprising. A distinctiveness of '*Candidatus* K. stuttgartiensis' within the PVC superphylum was already indicated by Wagner and Horn (2006), who found a class of specific nucleotide transporters to be shared among the *Chlamydiae* and *R. baltica* SH 1^T but notably absent from '*Candidatus* K. stuttgartiensis'. Since the respective transporters are thought to be typical for endosymbionts and endoparasites, this could have wide implications for the evolution of the *Planctomycetes* and the anammox bacteria that are beyond

the scope of this study. If the distinctness of the anammox bacteria is confirmed in the future by other anammox bacteria sequencing projects, one might consider whether a placement of the anammox bacteria within a separate phylum in the PVC superphylum is more appropriate.

Functions of the planctomycete-specific genes
So far one can only speculate on the functions of the planctomycete-specific genes reported here. Systematic investigations of their genetic neighborhoods did not lead to any conclusion concerning function. However, the sheer size of the paralogous gene families, in particular those of the Pfam domains DUF1501, PSCyt2, PSD1, PSCyt1 and SBP_bac_10, indicate that these genes are of great importance. It is noteworthy that these planctomycete-specific genes are often co-localized (Table 4). For example, DUF1559 genes are often followed downstream by a second planctomycete-specific gene of about 150 bp that carries a signal peptide. Considering the high proportion of genes with signal peptides among the paralogous gene families, they might play a role in the context of planctomycete compartmentalization. Likewise, planctomycete-specific genes are frequently found at the beginning of operon-like gene arrangements, indicating a regulatory function. Interestingly, an unusually low number of transcriptional regulators has been reported for *R. baltica* SH 1^T (Lombardot *et al.*, 2005), and hence it is not unlikely that some of the planctomycete-specific paralogs act in the context of transcriptional control. Of course, further functional studies are required to validate these hypotheses.

Conclusions

This comprehensive comparative study of six planctomycete fosmids from marine upwelling regions and all available planctomycete genomic sequences allowed striking new insights in the fascinating phylum *Planctomycetes*. We were able to confirm the hypothesis that sulfatases are of particular importance for the lifestyle of many marine *Planctomycetes* and therefore are likely to be of general importance for the recycling of carbon from complex sulfated heteropolysaccharides in marine habitats. Furthermore, C1 metabolism genes were found on some of the fosmids and in all planctomycete genomes except for the anammox bacterium '*Candidatus* Kuenenia stuttgartiensis'. This suggests a general relevance of these genes for *Planctomycetes*. In addition, very large families of planctomycete-specific paralogs were found that might serve as screening targets for *Planctomycetes* in flow-up studies. The notable lack of these genes within the only anammox planctomycete in this study reveals an as yet unknown distinctiveness of these organisms from all other *Planctomycete* genera.

Acknowledgements

The Gordon and Betty Moore Foundation (www.moore.org) has funded the almost complete sequencing of *B. marina* DSM 3645^T and *P. maris* DSM 8797^T within the framework of their Marine Microbiology Initiative. Furthermore, EFD is funded by a grant from the Gordon and Betty Moore foundation. Analysis of *Rhodopirellula baltica* SH 1T is supported by the European Union through the Network of Excellence Marine Genomics Europe. Sequencing of *G obscuriglobus* UQM 2246T is a collaboration between the group of Naomi Ward at The Institute for Genomic Research (www.tigr.org) and the group of John A. Fuerst at the University of Queensland, Australia and has been funded by the DOE (DoE grant DEFC0295ER61962). This study was funded by the Max Planck Society.

References

Asakawa S, Abe I, Kudoh Y, Kishi N, Wang Y, Kubota R *et al.* (1997). Human BAC library: construction and rapid screening. *Gene* **191**: 69–79.

Badger JH, Olsen GJ. (1999). CRITICA: coding region identification tool invoking comparative analysis. *Mol Biol Evol* **16**: 512–524.

Bauer M, Lombardot T, Teeling H, Ward NL, Amann RI, Glöckner FO. (2004). Archaea-like genes for C1-transfer enzymes in *Planctomycetes*: phylogenetic implications of their unexpected presence in this phylum. *J Mol Evol* **59**: 571–586.

Bauld J, Staley JT. (1976). *Planctomyces maris* sp. nov.: a marine isolate of the *Planctomycetes-Blastocaulis* group of budding bacteria. *J Gen Microbiol* **97**: 45–55.

Bauld J, Staley JT. (1980). *Planctomyces maris* sp. nov., nom. rev. *Int J Syst Bact* **30**: 657.

Berteau O, Guillot A, Benjdia A, Rabot S. (2006). A new type of bacterial sulfatases reveals a novel maturation pathway in prokaryotes. *J Biol Chem* **281**: 22464–22470.

Brochier C, Philippe H. (2002). Phylogeny: a non-hyperthermophilic ancestor for bacteria. *Nature* **417**: 244.

Carbone A, Zinovyev A, Kepes F. (2003). Codon adaptation index as a measure of dominating codon bias. *Bioinformatics* **19**: 2005–2015.

Cavalier-Smith T. (2006). Rooting the tree of life by transition analyses. *Biol Direct* **1**: 19.

Chistoserdova L, Jenkins C, Kalyuzhnaya MG, Marx CJ, Lapidus A, Vorholt JA *et al.* (2004). The enigmatic planctomycetes may hold a key to the origins of methanogenesis and methylotrophy. *Mol Biol Evol* **21**: 1234–1241.

Chistoserdova L, Rasche ME, Lidstrom ME. (2005). Novel dephosphotetrahydromethanopterin biosynthesis genes discovered via mutagenesis in *Methylobacterium extorquens* AM1. *J Bacteriol* **187**: 2508–2512.

Chistoserdova L, Vorholt JA, Thauer RK, Lidstrom ME. (1998). C1 transfer enzymes and coenzymes linking methylotrophic bacteria and methanogenic Archaea. *Science* **281**: 99–102.

Crump BC, Armbrust EV, Baross JA. (1999). Phylogenetic analysis of particle-attached and free-living bacterial communities in the Columbia river, its estuary, and the adjacent coastal ocean. *Appl Environ Microbiol S* **65**: 3192–3204.

Delcher AL, Harmon D, Kasif S, White O, Salzberg SL. (1999). Improved microbial gene identification with GLIMMER. *Nucleic Acids Res* **27**: 4636–4641.

DeLong EF, Franks DG, Alldredge AL. (1993). Phylogenetic diversity of aggregate-attached vs free-living marine bacterial assemblages. *Limnol Oceanogr* **38**: 924–934.

Dierks T, Lecca MR, Schlotterhose P, Schmidt B, von Figura K. (1999). Sequence determinants directing conversion of cysteine to formylglycine in eukaryotic sulfatases. *EMBO J* **18**: 2084–2091.

Fieseler L, Horn M, Wagner M, Hentschel U. (2004). Discovery of the novel candidate phylum *'Poribacteria'* in marine sponges. *Appl Environ Microbiol* **70**: 3724–3732.

Franzmann PD, Skerman VB. (1984). *Gemmata obscuriglobus*, a new genus and species of the budding bacteria. *Antonie Van Leeuwenhoek* **50**: 261–268.

Fuerst JA, Gwilliam HG, Lindsay M, Lichanska A, Belcher C, Vickers JE *et al.* (1997). Isolation and molecular identification of planctomycete bacteria from postlarvae of the giant tiger prawn, *Penaeus monodon*. *Appl Environ Microbiol* **63**: 254–262.

Fuerst JA, Sambhi SK, Paynter JL, Hawkins JA, Atherton JG. (1991). Isolation of a bacterium resembling *Pirellula* species from primary tissue culture of the giant tiger prawn (*Penaeus monodon*). *Appl Environ Microbiol* **57**: 3127–3134.

Fuerst JA, Webb RI, Garson MJ, Hardy L, Reiswig HM. (1998). Membrane-bounded nucleoids in microbial symbionts of marine sponges. *FEMS Microbiol Lett* **166**: 29–34.

Fuerst JA, Webb RI, Garson MJ, Hardy L, Reiswig HM. (1999). Membrane-bounded nuclear bodies in a diverse range of symbionts of great Barrier Reef sponges. *Mem Queensl Mus* **44**: 193–203.

Fuerst JA. (1995). The planctomycetes: emerging models for microbial ecology, evolution and cell biology. *Microbiology* **141**: 1493–1506.

Fuerst JA. (2005). Intracellular compartmentation in planctomycetes. *Annu Rev Microbiol* **59**: 299–328.

Gade D, Schlesner H, Glöckner FO, Amann R, Pfeiffer S, Thomm M. (2004). Identification of planctomycetes with order-, genus-, and strain-specific 16S rRNA-targeted probes. *Microb Ecol* **47**: 243–251.

Giovannoni SJ, Godchaux Wr, Schabtach E, Castenholz RW. (1987b). Cell wall and lipid composition of *Isosphaera pallida*, a budding eubacterium from hot springs. *J Bacteriol* **169**: 2702–2707.

Giovannoni SJ, Schabtach E, Castenholz RW. (1987a). *Isosphaera pallida*, gen. and comb. nov., a gliding, budding eubacterium from hot springs. *Arch Microbiol* **147**: 276–284.

Glöckner FO, Kube M, Bauer M, Teeling H, Lombardot T, Ludwig W *et al.* (2003). Complete genome sequence of the marine planctomycete *Pirellula* sp. strain 1. *Proc Natl Acad Sci USA* **100**: 8298–8303.

Goenrich M, Bartoschek S, Hagemeier CH, Griesinger C, Vorholt JA. (2002). A glutathione-dependent formaldehyde-activating enzyme (Gfa) from *Paracoccus denitrificans* detected and purified via two-dimensional proton exchange NMR spectroscopy. *J Biol Chem* **277**: 3069–3072.

Guo FB, Ou HY, Zhang CT. (2003). ZCURVE: a new system for recognizing protein-coding genes in bacterial and archaeal genomes. *Nucleic Acids Res* **31**: 1780–1789.

Inagaki F, Nunoura T, Nakagawa S, Teske A, Lever M, Lauer A *et al.* (2006). Biogeographical distribution and diversity of microbes in methane hydrate-bearing deep marine sediments on the Pacific ocean margin. *Proc Natl Acad Sci USA* **103**: 2815–2820.

Jetten MSM, Horn SJ, Van Loosdrecht MCM. (1997). Towards a more sustainable wastewater treatment system. *Water SciTechnol* **35**: 171–180.

Jetten MSM, Schmid M, Schmidt I, Wubben M, van Dongen U, Abma W *et al.* (2002). Improved nitrogen removal by application of new nitrogen-cycle bacteria. *Rev Environ Sci Biotechnol* **1**: 51–63.

Kalyuzhnaya MG, Bowerman S, Nercessian O, Lidstrom ME, Chistoserdova L. (2005c). Highly divergent genes for methanopterin-linked C1 transfer reactions in Lake Washington, assessed via metagenomic analysis and mRNA detection. *Appl Environ Microbiol* **71**: 8846–8854.

Kalyuzhnaya MG, Korotkova N, Crowther G, Marx CJ, Lidstrom ME, Chistoserdova L. (2005b). Analysis of gene islands involved in methanopterin-linked C1 transfer reactions reveals new functions and provides evolutionary insights. *J Bacteriol* **187**: 4607–4614.

Kalyuzhnaya MG, Lidstrom ME, Chistoserdova L. (2004). Utility of environmental primers targeting ancient enzymes: methylotroph detection in Lake Washington. *Microb Ecol* **48**: 463–472.

Kalyuzhnaya MG, Nercessian O, Lapidus A, Chistoserdova L. (2005d). Fishing for biodiversity: novel methanopterin-linked C transfer genes deduced from the Sargasso Sea metagenome. *Environ Microbiol* **7**: 1909–1916.

Kalyuzhnaya MG, Nercessian O, Lidstrom ME, Chistoserdova L. (2005a). Development and application of polymerase chain reaction primers based on *fhcD* for environmental detection of methanopterin-linked C1-metabolism in bacteria. *Environ Microbiol* **7**: 1269–1274.

König E, Schlesner H, Hirsch P. (1984). Cell wall studies on budding bacteria of the *Planctomyces/Pasteuria* group and on a *Prosthecomicrobium* sp. *Arch Microbiol* **138**: 200–205.

Krogh A, Larsson B, von Heijne G, Sonnhammer EL. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol S* **305**: 567–580.

Kuypers MM, Lavik G, Woebken D, Schmid M, Fuchs BM, Amann R *et al.* (2005). Massive nitrogen loss from the Benguela upwelling system through anaerobic ammonium oxidation. *Proc Natl Acad Sci USA* **102**: 6478–6483.

Liesack W, König H, Schlesner H, Hirsch P. (1986). Chemical composition of the peptidoglycan-free cell envelopes of budding bacteria of the *Pirellula/Planctomyces* group. *Arch Microbiol* **145**: 361–366.

Liesack W, Söller R, Steward T, Haas H, Giovannoni S, Stackebrandt E. (1992). The influence of tachytelically (rapidly) evolving sequences on the topology of phylogenetic trees—intrafamily relationships and the phylogenetic position of *Planctomycetaceae* as revealed by comparative analysis of 16S ribosomal RNA sequences. *Syst Appl Microbiol* **15**: 357–362.

Lindsay MR, Webb R, Fuerst JA. (1997). Pirellulosomes: a new type of membrane-bounded cell compartment in planctomycete bacteria of the genus *Pirellula*. *Microbiology* **143**: 739–748.

Lindsay MR, Webb RI, Strous M, Jetten MSM, Butler MK, Forde RJ *et al.* (2001). Cell compartmentalisation in planctomycetes: novel types of structural organisation for the bacterial cell. *Arch Microbiol* **175**: 413–429.

Llobet-Brossa E, Rossello-Mora R, Amann R. (1998). Microbial community composition of Wadden Sea sediments as revealed by fluorescence *in situ* hybridization. *Appl Environ Microbiol* **64**: 2691–2696.

Lombardot T, Bauer M, Teeling H, Amann R, Glöckner FO. (2005). The transcriptional regulator pool of the marine bacterium *Rhodopirellula baltica* SH 1T as revealed by whole genome comparisons. *FEMS Microbiol Lett* **242**: 137–145.

Ludwig W, Strunk O, Westram R, Richter L, Meier H, Yadhukumar *et al.* (2004). ARB: a software environment for sequence data. *Nucleic Acids Res* **32**: 1363–1371.

Marx CJ, Chistoserdova L, Lidstrom ME. (2003). Formaldehyde-detoxifying role of the tetrahydromethanopterin-linked pathway in *Methylobacterium extorquens* AM1. *J Bacteriol* **185**: 7160–7168.

McCarren J, Delong EF. (2007). Proteorhodopsin photosystem gene clusters exhibit co-evolutionary trends and shared ancestry among diverse marine microbial phyla. *Environ Microbiol S* **9**: 846–858.

Meyer F, Goesmann A, McHardy AC, Bartels D, Bekel T, Clausen J *et al.* (2003). GenDB—an open source genome annotation system for prokaryote genomes. *Nucleic Acids Res* **31**: 2187–2195.

Miskin IP, Farrimond P, Head IM. (1999). Identification of novel bacterial lineages as active members of microbial populations in a freshwater sediment using a rapid RNA extraction procedure and RT-PCR. *Microbiology* **145**: 1977–1987.

Morris RM, Longnecker K, Giovannoni SJ. (2006). *Pirellula* and OM43 are among the dominant lineages identified in an Oregon coast diatom bloom. *Environ Microbiol* **8**: 1361–1370.

Musat N, Werner U, Knittel K, Kolb S, Dodenhof T, van Beusekom JE *et al.* (2006). Microbial community structure of sandy intertidal sediments in the North Sea, Sylt-Romo Basin, Wadden Sea. *Syst Appl Microbiol* **29**: 333–348.

Neef A, Amann R, Schlesner H, Schleifer KH. (1998). Monitoring a widespread bacterial group: *in situ* detection of planctomycetes with 16S rRNA-targeted probes. *Microbiology* **144**: 3257–3266.

Nielson H, Engelbrecht J, Brunak S, von Heijne G. (1997). Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng* **10**: 1–6.

Pace NR, Olsen GJ, Woese CR. (1986). Ribosomal RNA phylogeny and the primary lines of evolutionary descent. *Cell S* **45**: 325–326.

Peden JF. (1999). *Analysis of Codon Usage*. PhD thesis, Department of Genetics, University of Nottingham.

Pimentel-Elardo S, Wehrl M, Friederich AB, Jensen PR, Henschel U. (2003). Isolation of planctomycetes from *Aplysina* sponges. *Aquatic Microbial Ecol* **33**: 239–245.

Quast C. (2006). MicHanThi - design and implementation of a system for the prediction of gene functions in genome annotation projects. Master thesis (available on request).

Rabus R, Gade D, Helbig R, Bauer M, Glöckner FO, Kube M *et al.* (2002). Analysis of N-acetylglucosamine metabolism in the marine bacterium *Pirellula* sp. strain 1 by a proteomic approach. *Proteomics* **2**: 649–655.

Rusch A, Huettel M, Reimers CE, Taghon GL, Fuller CM. (2003). Activity and distribution of bacterial populations in middle Atlantic bight shelf sands. *FEMS Microbiol Ecol* **44**: 89–100.

Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S *et al.* (2007). The sorcerer II global ocean sampling expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biol S* **5**: e77.

Schlesner H, Rensmann C, Tindall BJ, Gade D, Rabus R, Pfeiffer S *et al.* (2004). Taxonomic heterogeneity within the *Planctomycetales* as derived by DNA-DNA hybridization, description of *Rhodopirellula baltica* gen. nov., sp. nov., transfer of Pirellula marina to the genus Blastopirellula gen. nov. as Blastopirellula marina comb. nov. and emended description of the genus Pirellula. *Int J Syst Evol Microbiol* **54**: 1567–1580.

Schlesner H. (1994). The development of media suitable for the microorganisms morphologically resembling *Planctomyces* spp., *Pirellula* spp., and other *Planctomycetales* from various aquatic habitats using dilute media. *Syst Appl Microbiol* **17**: 135–145.

Schmid M, Twachtmann U, Klein M, Strous M, Juretschko S, Jetten MSM *et al.* (2000). Molecular evidence for genus level diversity of bacteria capable of catalyzing anaerobic ammonium oxidation. *Syst Appl Microbiol* **23**: 93–106.

Scott JW, Rasche ME. (2002). Purification, overproduction, and partial characterization of beta-RFAP synthase, a key enzyme in the methanopterin biosynthesis pathway. *J Bacteriol S* **184**: 4442–4448.

Sinninghe Damsté JS, Strous M, Rijpstra WI, Hopmans EC, Geenevasen JA, van Duin AC *et al.* (2002). Linearly concatenated cyclobutane lipids form a dense bacterial membrane. *Nature S* **419**: 708–712.

Stackebrandt E, Ludwig W, Schubert W, Klink F, Schlesner H, Roggentin T *et al.* (1984). Molecular genetic evidence for early evolutionary origin of budding peptidoglycan-less eubacteria. *Nature* **307**: 735–737.

Staley JT, Bouzek H, Jenkins C. (2005). Eukaryotic signature proteins of *Prosthecobacter dejongeii* and *Gemmata* sp. Wa-1 as revealed by in silico analysis. *FEMS Microbiol Lett* **243**: 9–14.

Stein JL, Marsh TL, Wu KY, Shizuya H, DeLong EF. (1996). Characterization of uncultivated prokaryotes: isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon. *J Bacteriol S* **178**: 591–599.

Strous M, Fuerst JA, Kramer EH, Logemann S, Muyzer G, van de Pas-Schoonen KT *et al.* (1999). Missing lithotroph identified as new planctomycete. *Nature* **400**: 446–449.

Strous M, Pelletier E, Mangenot S, Rattei T, Lehner A, Taylor MW *et al.* (2006). Deciphering the evolution and metabolism of an anammox bacterium from a community genome. *Nature* **440**: 790–794.

Studholme DJ, Fuerst JA, Bateman A. (2004). Novel protein domains and motifs in the marine planctomycete *Rhodopirellula baltica*. *FEMS Microbiol Lett* **236**: 333–340.

Teeling H, Lombardot T, Bauer M, Ludwig W, Glöckner FO. (2004). Evaluation of the phylogenetic position of the planctomycete 'Rhodopirellula baltica' SH 1 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.

Tekniepe BL, Schmidt JM, Starr MP. (1981). Life cycle of a budding and appendaged bacterium belonging to morphotype IV ofthe *Blastocaulis-Planctomyces* group. *Curr Microbiol* **5**: 1–6.

Vergin KL, Urbach E, Stein JL, DeLong EF, Lanoil BD, Giovannoni SJ. (1998). Screening of a fosmid library of marine environmental genomic DNA fragments reveals four clones related to members of the order *Planctomycetales*. *Appl Environ Microbiol* **64**: 3075–3078.

Vorholt JA, Chistoserdova L, Stolyar SM, Thauer RK, Lidstrom ME. (1999). Distribution of tetrahydromethanopterin-dependent enzymes in methylotrophic bacteria and phylogeny of methenyl tetrahydromethanopterin cyclohydrolases. *J Bacteriol* **181**: 5750–5757.

Vorholt JA, Kalyuzhnaya MG, Hagemeier CH, Lidstrom ME, Chistoserdova L. (2005). MtdC, a novel class of methylene tetrahydromethanopterin dehydrogenases. *J Bacteriol* **187**: 6069–6074.

Wagner M, Horn M. (2006). The *Plantomycetes*, *Verrucomicrobia*, *Chlamydiae* and sister phyla comprise a superphylum with biotechnological and medical relevance. *Curr Opin Biotechnol* **17**: 241–249.

Wallner SR, Bauer M, Würdemann C, Wecker P, Glöckner FO, Faber K. (2005). Highly enantioselective sec-alkyl sulfatase activity of the marine planctomycete *Rhodopirellula baltica* shows retention of configuration. *Angew Chem Int Ed Engl* **44**: 6381–6384.

Wang J, Jenkins C, Webb RI, Fuerst JA. (2002). Isolation of *Gemmata*-like and *Isosphaera*-like planctomycete bacteria from soil and freshwater. *Appl Environ Microbiol* **68**: 417–422.

Weisburg WG, Hatch TP, Woese CR. (1986). Eubacterial origin of *Chlamydiae*. *J Bacteriol* **167**: 570–574.

Woese CR. (1987). Bacterial evolution. *Microbiol Rev S* **51**: 221–271.

Zhou J, Bruns MA, Tiedje JM. (1996). DNA recovery from soils of diverse composition. *Appl Environ Microbiol S* **62**: 316–322.

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