RESEARCH ARTICLE

Construction and preliminary analysis of a metagenomic library from a deep-sea sediment of east Pacific Nodule Province

Meixiang Xu, Fengping Wang, Jun Meng & Xiang Xiao

Key Laboratory of Marine Biogenetic Resources, Third Institute of Oceanography, State Oceanic Administration, Xiamen, China

Correspondence: Xiang Xiao, Key Laboratory of Marine Biogenetic Resources, Third Institute of Oceanography, State Oceanic Administration, Xiamen, 361005, China. Tel.: +86 592 2195236; fax: +86 592 2085376; e-mail: xxwfp@public.xm.fj.cn

Received 12 March 2007; revised 15 June 2007; accepted 21 July 2007.
First published online 11 September 2007.

DOI:10.1111/j.1574-6941.2007.00377.x

Editor: Patricia Sobecky

Keywords
metagenomic library; deep-sea sediment; Nodule Province; microbial diversity; *Magnetospirillum*.

Abstract

The Pacific Nodule Province is a unique ocean area containing an abundance of polymetallic nodules. To explore more genetic information and discover potentially industrial useful genes of the microbial community from this particular area, a cosmid library with an average insert of about 35 kb was constructed from the deep-sea sediment. The bacteria in the cosmid library were composed mainly of *Proteobacteria* including *Alphaproteobacteria*, *Gammaproteobacteria* and *Deltaproteobacteria*. The end sequences of some cosmid clones were determined and the complete insert sequences of two cosmid clones, 10D02 and 17H9, are presented. 10D02 has a length of 40.8 kb and contains 40 predicted encoding genes. It contains a partial 16S rRNA gene of *Alphaproteobacteria*. 17H9 is 36.8 kb and predicted to have 31 encoding genes and a 16S–23S–5S rRNA gene operon. Phylogenetic analysis of 16S and 23S rRNA gene sequence on the 17H9 both reveals that the inserted DNA from 17H9 came from a novel *Alphaproteobacteria* and is closely related to *Magnetospirillum* species. The predicted proteins of ORF 1–11 also have high identity to those of *Magnetospirillum* species, and the organization of these genes is highly conserved among known *Magnetospirillum* species. The data suggest that the retrieved DNA in 17H9 might be derived from a novel *Magnetospirillum* species.

Introduction

Microorganisms are believed to play important roles in metal cycling in many environments. Phylogenetically diverse metal-oxidizing or -reducing bacteria have been isolated and partially characterized from different environments including metal-rich environments such as manganese nodules and hydrothermal vents (Ghiorse, 1984; Gounot, 1994; Nealson & Saffarini, 1994; Nealson, 1997; Mitra *et al*., 1998; Stein *et al*., 2001). The microbial community associated with Lechuguilla and Spider Caves, environments rich in ferromanganese deposits formed during the weathering of Mn(II)-bearing carbonate host rock, was found to harbor a predominance of archaea (Northup *et al*., 2003). Interestingly, the clone libraries from the ferromanganese cave deposits and the enrichment cultures showed some similarity to the clone libraries and purified cultures from ferromanganese micronodules of Green Bay sediments (Stein *et al*., 2001; Northup *et al*., 2003). Marine *Crenarchaeota* were represented in the archaeal clones in libraries from both sites. Despite this information, little genetic information is available for these microorganisms, and nearly nothing is known about how the microorganisms in these environments cooperate, or compete, or how they interact with the environments.

Myriad environments on earth invite a deeper view of the network of genes, genomes, organisms and communities present in the natural world. Recent progress on genomic techniques has provided new opportunities to address challenging questions in ecology and evolution (DeLong, 2002, 2005; Doolittle, 2002; DeLong & Karl, 2005). The merging of cultivation-independent DNA sequencing with contemporary genomic approaches is providing a more comprehensive picture of the structure and function of indigenous microbial communities (DeLong, 2002, 2005; Doolittle, 2002; DeLong & Karl, 2005). Two approaches including random shotgun sequencing of environmental DNA and sequencing of large DNA fragments recovered in metagenomic libraries are now being utilized in environmental ecological, evolutionary, genomic and functional analyses (DeLong, 2002; Tyson *et al.*, 2004; Venter *et al.*, 2004). The metagenomic approach has begun to be used in genome analysis of uncharacterized microbial taxa (Rondon *et al.*, 2000; Liles *et al.*, 2003; Tyson *et al.*, 2004; Moreira *et al.*, 2006), expression of novel genes or pathways from uncultured environmental microorganisms (Henne *et al.*, 2000; Rondon *et al.*, 2000; Gillespie *et al.*, 2002; Voget *et al.*, 2003), elucidation of community-specific metabolism and comparison of gene contents in different communities (Knietsch *et al.*, 2003; Tyson *et al.*, 2004; DeLong *et al.*, 2006).

The deep-sea Nodule Province contains large amounts of polymetallic nodules mainly composed of manganese, iron, cobalt, copper and nickel (Piper & Williamson, 1977). It is suspected that microorganisms probably played a role in the formation of deep-sea polymetallic nodules. Bacterial 16S rRNA gene sequences analysis demonstrated that *Proteobacteria*, mainly *Alphaproteobacteria*, *Gammaproteobacteria* and *Deltaproteobacteria*, dominate in the sediment (Xu *et al.*, 2004, 2005). The genomic information of some of the deep-sea *Deltaproteobacteria* is beginning to be partly revealed through the metagenomic approach (Moreira *et al.*, 2006). Nothing is known about the deep-sea *Alphaproteobacteria*, which is also constantly detected in the deep-sea sediments (Xu *et al.*, 2004, 2005). To gain more comprehensive information on the microbial community and discover industrially useful genes from the Pacific Nodule Province, a cosmid library from a 5274 m deep-sea sediment of Pacific Nodule Province was constructed and analyzed. It represents a useful resource for genomic and ecological analyses of the microbial community in the Pacific Nodule Province.

Materials and methods

Deep-sea sediment sample collection

One deep-sea sediment core was collected by a multi-core sampler on September 2003, at the ES0303 station of east Pacific Nodule Province (8°21′11″N, 145°24′09″W), during the DY105-12/14 cruise of DaYang No.1. The depth of this site is 5274 m, the temperature is 1.5 °C and the salinity is 35‰. The sediment core was maintained at −20 °C for shipping and in the lab until usage.

DNA Isolation, fractionation and cosmid library construction

The DNA extraction method was a modification of an sodium dodecyl sulphate (SDS)-based DNA extraction method (Zhou *et al.*, 1996). Five grams of sediment around 5 cm below the surface was mixed with 13.5 mL DNA extraction buffer and incubated in a 60 °C water bath with an occasional gentle mixing for 2 h. Then, 1.5 mL 20% SDS was added and the samples were incubated at 60 °C for another 2 h. After centrifugation at 5000 ***g*** for 20 min, the DNA supernatant was precipitated with isopropanol and dissolved in TE buffer (pH 8.0). The pulsed field gel electrophoresis (PFGE) method (Gene navigator system, Amersham Pharmacia) was used to select DNA fragments of optimal size. The condition of PFGE is 250 V, 5–7 h, with a pulse time of 5 s. After DNA size estimation, a crude DNA sample was run on another PFGE on a 1% low-melting-point (LMP) agarose gel for optimal size DNA recovery. The gel was stained with SYBR® Gold (Molecular Probes, Inc., Eugene, Oregon), followed by illumination with a Dark Reader™ transilluminator (Clare Chemical Research, Denver, CO). The gel slice containing 35–45-kb DNA fragments was excised and digested with GElase (Epicentre, Madison, WI). Then, the DNA was concentrated, quantified and redissolved in TE buffer (pH8.0) for cosmid library construction. The cosmid library was constructed using the pWEB::TNC™ Cosmid Cloning Kit (Epicentre) following the user's manual. The cosmids were extracted according to the procedure of the QIAGEN Plasmid Mini Kit, and the sizes of the inserted fragments in the cosmid clones were determined by restriction enzyme digestion and agarose gel electrophoresis.

Analysis of microbial composition in the cosmid library

The DNA of pooled cosmid clones was extracted using the Plasmid DNA isolation kit (Plasmid Mini Kit I, Omega, Doraville, GA). Before PCR amplification, the cosmid DNA templates were digested with the plasmid-safe, ATP-dependent DNase (Epicentre) (Liles *et al.*, 2003) to remove the chromosome DNA contamination of the host strain (*Escherichia coli* EPI100). The bacterial universal primers EUB f933 and EUB r1387 (around 500 bp gene fragments) were used for 16S rRNA gene amplification. The PCR amplicons were cloned, and restriction fragment length polymorphism (RFLP) analysis was used to screen the unique 16S rRNA gene clones as described previously (Xu *et al.*, 2005). The representative unique clones were selected to sequence. The sequences were first checked for chimeric artifacts by the CHECK-CHIMERA program at the Ribosomal Database Project (http://rdp8.cme.msu.edu/html/) and were then blasted in the GeneBank for homologous sequences. The accession numbers of the 16S rRNA gene sequences are AM503078–AM503088, AM600911.

End sequencing of cosmid clones and analyses

Randomly selected cosmid clones were subjected to end sequencing (China Human Genome Centre at Shanghai). The average length of the sequences was around 350 bp. The primers P1 5′-TGCCACCTGACGTCTAAGA and P2 5′-CTGACTGCGTTAGCAA TTTAA, designed based on the sequence of pWEB::TNC cosmid vector, were used to amplify the integrated end sequences of the cosmid clones. The sequences obtained were blasted on NCBI and EMBL by BLASTN (http://www.ncbi.nlm.nih.gov/BLAST/), BLASTX (http://www.ncbi.nlm.nih.gov/BLAST/) and WU-BLAST (http: //www.ebi.ac.uk/blast2/). The information obtained was compared and summarized; only BLAST results with *E* value < 0.1 were used. Then, the end sequences were submitted to EMBL, and the accession numbers are AM160665–AM160788.

Sequencing of the entire genomic inserts of cosmid clones and genome analysis

The whole genomic sequences of selected cosmid clones were determined by shotgun sequencing. Briefly, the cosmid clones were isolated and fragmented by sonication. Then, the fragmented DNA was separated by gel electrophoresis. Random 2-kb fragments were recovered from gels, blunt end-repaired and cloned into pUC18 at the SmaI site. The plasmids were sequenced from both ends using an ABI3700 sequencer (Applied Biosystem Inc.). The sequences generated had around sevenfold coverage of the inserted DNA. The sequences were assembled using the program SEQUENCER. ORF analysis was performed using the GENEMARK program (http://opal.biology.gatech.edu/GeneMark/). The putative ORFs and their encoded protein sequences were blasted on NCBI and on the EMBL database.

Results

Cosmid library construction

Large fragment DNA was isolated from the sediment and a cosmid library was constructed using the extracted sediment DNA. The titer of the packaged cosmids of the deep-sea sediment metagenomic library was about 6×10^3 CFU mL^{-1}. Around 3500 cosmid clones were obtained without amplification, and most of these clones were picked into 96-well plates. The sizes of the inserted fragments of the clones in the library were checked by enzyme digestion and gel electrophoresis. It was calculated that the average size of the insert-DNA in the sediment library was about 35 kb. The metagenomic cosmid library contained at least 122-Mbp chromosomal DNA of deep-sea microorganisms.

Bacterial composition in the cosmid library

DNA were extracted from mixed pools of the clones in the library and used as templates for PCR amplification as described in Materials and methods. The amplified band was recovered and cloned; 37 clones were randomly chosen for RFLP analysis. Thirteen RFLP types were found, and the representative clones were sequenced. In total, the bacteria in the metagenomic library were found, including *Proteobacteria* (*Alphaproteobacteria*, *Gammaproteobacteria* and *Deltaproteobacteria*), *Actinobacteria* and Firmicutes (Table 1). *Proteobacteria* are dominant in the library, with *Deltaproteobacteria* comprising 48.6%, *Gammaproteobacteria*

Table 1. Summary of the bacterial community in the deep-sea sediment cosmid library

	Nearest phylogenetic neighbour and 16S rRNA gene accession number			
Clones	Culturable bacterium	Identity (%)	Uncultivated clones	Identity (%)
Proteobacteria				
Alphaproteobacteria				
Es8 (2)*	*Rhodovulum* sp. SMB1, DQ868668	94	Clone MBMPE43, AJ567557	99
Gammaproteobacteria				
Es66 (2)	*Nitrococcus mobilis*, L35510	88	Clone BD7-2, AB015578	99
ES76 (1)	*Pseudomonas* sp. NB1-h, AB013829	98	Clone Flyn1_17, DQ256717	99
Deltaproteobacteria				
Es2 (11)	*Geobacter argillaceus*, DQ145534	88	Clone OM27, U70713	96
Es6 (1)	*Desulfuromonadales* bacterium JN18_A94_J, DQ168651	89	Clone MBMPE45, AJ567559	99
Es35 (4)	*Geobacter rgillaceus*, DQ145534	89	B-BK96, AY622263	96
Es62 (1)	*Geobacter* sp. Ala-6, AF019929	92	Clone LC1-13, DQ289938	96
Es68 (1)	*Anaeromyxobacter dehalogenans*, AF382396	89	Clone s1uc25, DQ416291	96
Actinobacteria				
Es1 (7)	Bacterium Ellin5290, AY234641	88	Actinobacterium clone R171, AF333522	99
Firmicutes				
Es80 (1)	*Thermaerobacter* sp., AY094621	87	Clone KS77, AF328213	98
Es9 (2)	*Geobacillus subterraneus*, AY608956	87	Clone E52, AJ966599	99
Es40 (4)	*Thermaerobacter* sp. C4-1	88	Clone MSB-5D2, DQ811925	95

*The number in the brackets indicates the number of clones detected in the library.

8.1% and *Alphaproteobacteria* 5.4%, with *Actinobacteria* and Firmicutes each 18.9% (Table 1). Most of the bacteria retrieved in the library have < 90% identity to those of the cultivated strains, indicating that most of the strains in the library represent uncharacterized novel bacterial species. The 16S rRNA gene sequences in the library had a relatively high identity with 16S rRNA gene clones retrieved from various environments (Table 1) including the deep-sea sediment (Li *et al.*, 1999; Xu *et al.*, 2004), the shelf sediment (Hunter *et al.*, 2006), coastal water (Rappe *et al.*, 1997), surrogate minerals incubated in an acidic uranium-contaminated aquifer (Reardon *et al.*, 2004) and others.

End sequences of the cosmid clones

Cosmid clones were randomly chosen for end sequencing. The end sequences of 62 clones were obtained. The nucleotide and translated amino acid sequences were searched in the database for related information (supplementary Table S1). Of the 124 end sequences, 51 could be assigned to functional genes, 12 were genes of conserved hypothetical proteins, 10 were function unknown genes and 51 had no significant similarity to any known sequences. The information of the end sequences is summarized according to the clusters of orthologous groups (COGs, Tatusov *et al.*, 2003, Table 2). COGs analysis clearly showed that the sequences obtained in the deep-sea sediment cosmid library could be classified into several function groups: information storage and processing group (8.9%); cellular process and signaling group (7.3%); metabolism group (20.2%); and poorly characterized group, which was the majority and took up 63.7% (Table 2). Detailed information on the end sequences is given in supplementary Table S1; the nucleotide sequences showed from none to 88% identity to the sequences in the database. Most of the sequences were found to have the highest identity to genes from *Proteobacteria* and *Actinobacteria*.

Table 2. Classification of the end sequences in the deep-sea sediment cosmid library

Code and relative description	No. of sequences (%*)
Information storage and processing	11 (8.9%)
J: Translation, ribosomal structure and biogenesis	4 (3.2%)
K: Transcription	3 (2.4%)
L: Replication, recombination and repair	3 (2.4%)
B: Chromatin structure and dynamics	1 (0.8%)
Cellular process and signaling	9 (7.3%)
T: Signal transduction mechanisms	2 (1.6%)
M: Cell wall/membrane/envelope biogenesis	5 (4.0%)
O: Posttranslational modification, protein turnover, chaperones	2 (1.6%)
Metabolism	25 (20.2%)
C: Energy production and conversion	4 (3.2%)
G: Carbohydrate transport and metabolism	3 (2.4%)
E: Amino acid transport and metabolism	5 (4.0%)
F: Nucleotide transport and metabolism	1 (0.8%)
H: Coenzyme transport and metabolism	1 (0.8%)
I: Lipid transport and metabolism	4 (3.2%)
P: Inorganic ion transport and metabolism	5 (4.0%)
Q: Secondary metabolites biosynthesis, transport and catabolism	2 (1.6%)
Poorly characterized	79 (63.7%)
R: General function prediction only	7 (5.6%)
S: Function unknown	72 (58.1%)

*Represents the percentage of respective end sequences in the whole retrieved end sequences.

Genetic information of two cosmid clones containing 16S rRNA genes

The p1 end sequence of clone 10D02 had around 90% identity to the 16S rRNA gene partial sequence of *Oculina patagonica* endosymbiont Mucus bacterium 23 of *Alphaproteobacteria* (supplementary Table S1). As nothing is known about this type of bacteria, the whole sequence of 10D02 was determined to obtain more genetic information about the deep-sea bacterium from which the DNA is derived. The partial 16S rRNA gene on the terminus of 10D02 is 1005-bp long. It had 92% and 91% identities to 16S rRNA gene sequences of an unclassified bacterium from the mucus of coral *Oculina patagonica* and *Rhodovibrio* sp. 2Mb1, respectively. The insert sequence of 10D02 is 40 821 bp, has 67.09% G+C content and contains 40 predicted ORFs, among which 15% are hypothetical proteins, 20.5% are unknown proteins, 0.25% are proteins with unknown function and the remaining 62.5% are proteins with assigned functions (supplementary Table S2). Most of the ORFs encode proteins involved in cell metabolisms such as 3 hydroxy-3-methylglutaryl-CoA (HMG-CoA) lyase, which catalyzes the initial reactions of mevalonate catabolism, and carbamoyl-phosphate synthase, which initiates both the urea cycle and the biosynthesis of arginine and/or pyrimidines. Regulatory proteins and transport proteins were also identified. The gene organization on 10D02 is shown in Fig. 1a. Supplementary Table S2 provides detailed information on the putative ORFs and their encoded proteins. The predicted ORF functions include lyase, Biotin/lipoyl attachment, regulatory protein, phosphotransferase system (PTS), hydratase/isomerase, transferase, oxidoreductase, histidine kinase, dehydrogenase, thiolase, divalent cation transporter, ligase B, permease, acylphosphatases, carboxylase, cyclase, chaperone, AsmA, phosphatases, hypothetical protein and unknown.

Cosmid 17H9 was also selected for sequencing as it showed strong anti-bacterial activity (authors' unpublished data). The insert sequence of 17H9 is 36 847 bp, has a G+C

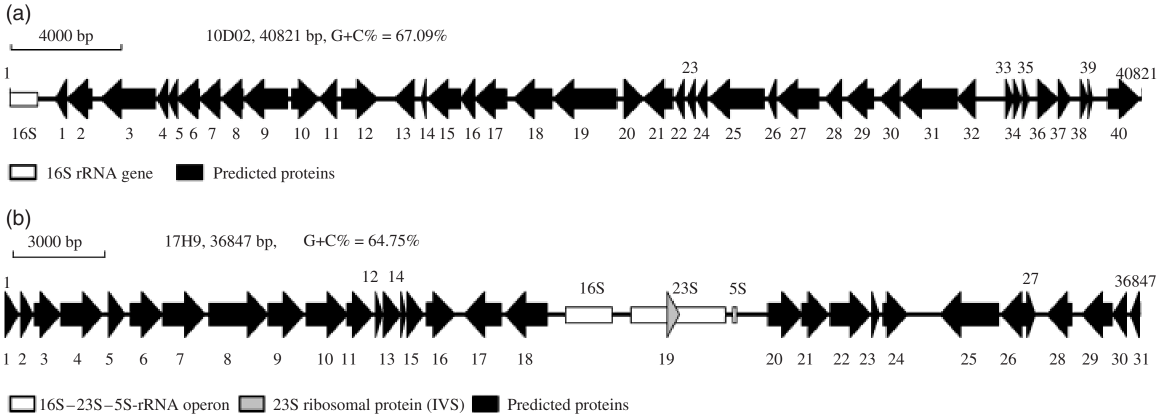

Fig. 1. Genomic map of cosmid clones 10D02 (a) and 17H9 (b) The complete DNA sequences of the clones were determined using the short-gun sequencing method as described in the 'Materials and methods'. The GENEMARK program was used to perform ORF analysis. Clone 10D02 is 40 821-bp long and contains 40 putative ORF, while clone 17H9 is 36 847-bp long and contains 31 putative ORFs. The putative ORFs in the clone 10D02 and 17H9 are represented by the numbers 1–40 and 1–31, respectively. 10D02 ORFs: 1, hypothetical protein; 2, lyase; 3, biotin/lipoyl attachment; 4, hypothetical protein; 5, regulatory protein; 6, unknown protein; 7, PTS system factor IIC; 8, hydratase/isomerase; 9, transferase; 10, hypothetical protein; 11, oxidoreductase; 12, kinase; 13, dehydrogenase; 14, hypothetical protein; 15, thiolase; 16, hypothetical protein; 17, dehydrogenase; 18, transporter; 19, dehydrogenase; 20, ligase B; 21, permease; 22, regulatory protein; 23, acylphosphatases; 24, unknown; 25, carboxylase; 26, hypothetical protein; 27, transferase; 28, regulatory protein; 29, cyclase; 30, chaperone; 3, AsmA; 32, phosphatases; 33, unknown; 34, unknown; 35, unknown; 36, unknown; 37, unknown; 38, unknown; 39, unknown; 40, structural protein; 17H9 ORFs: 1, transport proteins; 2, transport protein; 3, periplasmic protein; 4, periplasmic component, 5, outer membrane protein; 6, uncharacterized protein; 7, ATPase; 8, Zn protease; 9, synthase; 10, phosphomannomutase; 11, kinase; 12, unknown; 13, hypothetical protein; 14, synthetase; 15, hypothetical protein; 16, transmembrane protein; 17, dehydratase; 18, thiolase; 19, ribosomal protein; 20, synthase; 21, monophosphatase; 22, transferase; 23, synthase; 24, methyltransferase; 25, helicase; 26, desaturase; 27, regulator; 28, hypothetical protein; 29, oxidase; 30, unknown; 31, hypothetical protein. The detailed descriptions of the ORFs are presented in the supplementary Tables S2 and S3, respectively.

percentage of 64.75% and contains 31 predicted ORFs. Analysis results show that it contains a 16S–23S–5S rRNA gene operon. The 23S rRNA gene on 17H9 had a 423-bp intervening sequence (IVS) beginning at position 21 515 and ending at position 21 937 of the inserted genomic DNA (Fig. 1b). The 23S rRNA gene sequence without the IVS sequence showed the highest identity (90%) to those of *Magnetospirillum gryphiswaldense* and *Magnetospirillum magneticum* AMB-1. Phylogenetic trees showing the relationship of 17H9 with other reference species were constructed based on 16S rRNA and 23S rRNA gene sequences (Fig. 2a and b). As can be seen in Fig. 2a, the 16S rRNA gene sequence on 17H8 clusters with uncultured *Alphaproteobacterium* from Pacific deep-sea sediment (AJ966593 & AJ567563) and bacterioplankton from the Arctic Ocean (AF353236). The related 16S rRNA gene sequences in the *Alphaproteobacteria* subgroup could be divided into six clusters (Fig. 2a); Cluster I contains uncultured *Alphaproteobacterium* from coastal marine environment (AY033300), marine water in Monterey Bay (AY627380), forest soils (DQ451454), Green Bay ferromanganous micronodule (AF292999), heavy metal-contaminated environments (AJ581585) and Seafloor Basalts collected from East Pacific Rise and the Juan de Fuca Ridge (DQ070813). Cluster III consists of uncultured *Alphaproteobacterium* endosymbionts in the gutless marine worms. Cluster VI includes uncultured *Alphaproteobacterium* that arose from bioreactors treating wastewater (AF280850 and DQ146465) and alkaline, hypersaline lakes (DQ432379). The phylogenetic analysis indicated that the DNA fragment on 17H9 arose from a novel bacterium of *Alphaproteobacteria*, and closely related to *Magnetospirillum*.

The gene organization on 17H9 is shown in Figs 1 and 3. Upstream of the 16S rRNA gene, most of the genes (ORF 1–11) had high similarity to those from *Magnetospirillum* species. The organization of ORF 1–11 on 17H9 is highly conserved in *Magnetospirillum magnetotacticum* MS-1 and *Magenetospirillum* sp. AMB-1, whose genome sequences are available in the databank (Fig. 3). ORF 1–4 all encoded biopolymer transport proteins (supplementary Table S3) including ferric ion TonB transport protein. ORF 5–11 encode proteins probably involved in nucleic acid and protein biosynthesis (ORF 9), synthesis of vitamin B1 (ORF 11), protease (ORF 8), ATPase (ORF7), phosphomannomutase (ORF 10), Outer membrane protein and related peptidoglycan-associated lipo protein (ORF5) and uncharacterized conserved proteins (ORF6). Downstream from the 16S–23S–5S rRNA gene operon, four ORFs (ORF 20–23) encoded proteins with high similarities to those from archaea (supplementary Table S3). ORF23 encodes an 84 aa polypeptide, which is a truncated form of Myo-inositol-1-phosphate synthase (N-terminal part) encoded by ORF20. It may be a pseudogene (Moreira *et al.*, 2006). ORF29

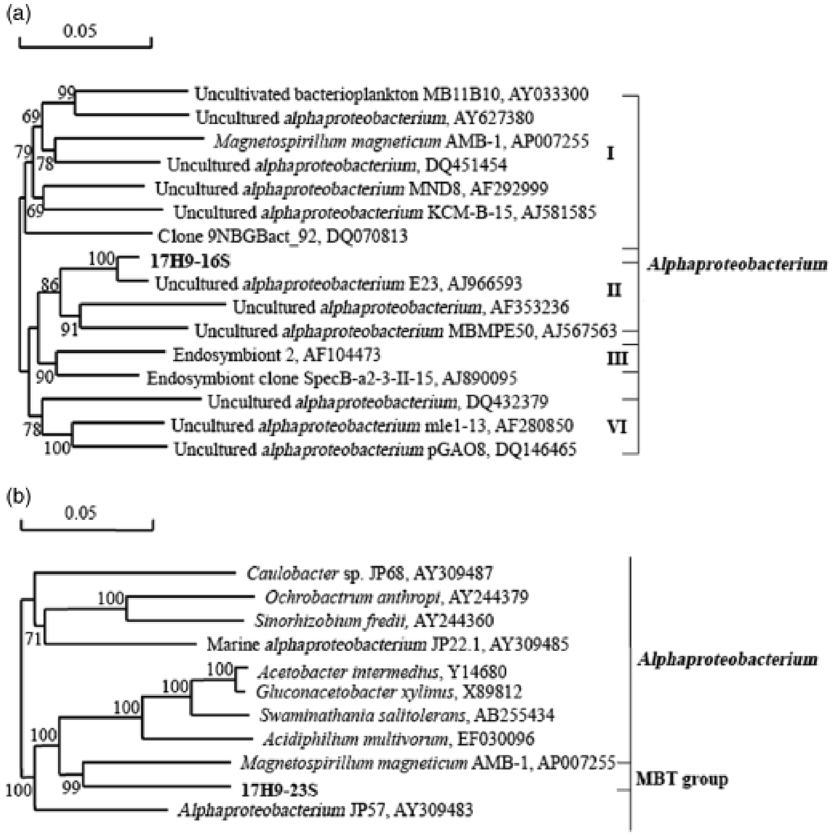
(a)
0.05
99
69
79
78
69
Uncultivated bacterioplankton MB11B10, AY033300
Uncultured alphaproteobacterium, AY627380
Magnetospirillum magneticum AMB-1, AP007255
Uncultured alphaproteobacterium, DQ451454
Uncultured alphaproteobacterium MND8, AF292999
Uncultured alphaproteobacterium KCM-B-15, AJ581585
Clone 9NBGBact_92, DQ070813
I
100
86
91
17H9-16S
Uncultured alphaproteobacterium E23, AJ966593
Uncultured alphaproteobacterium, AF353236
Uncultured alphaproteobacterium MBMPE50, AJ567563
II
90
Endosymbiont 2, AF104473
Endosymbiont clone SpecB-a2-3-II-15, AJ890095
III
78
100
Uncultured alphaproteobacterium, DQ432379
Uncultured alphaproteobacterium mle1-13, AF280850
Uncultured alphaproteobacterium pGAO8, DQ146465
VI
Alphaproteobacterium
(b)
0.05
100
71
100
100
100
100
99
100
Caulobacter sp. JP68, AY309487
Ochrobactrum anthropi, AY244379
Sinorhizobium fredii, AY244360
Marine alphaproteobacterium JP22.1, AY309485
Acetobacter intermedius, Y14680
Gluconacetobacter xylinus, X89812
Swaminathania salitolerans, AB255434
Acidiphilium multivorum, EF030096
Magnetospirillum magneticum AMB-1, AP007255
17H9-23S
Alphaproteobacterium JP57, AY309483
Alphaproteobacterium
MBT group

Fig. 2. Phylogenetic trees based on the 16S rRNA gene (a) and 23S rRNA gene (b) sequences. The phylogenetic relationship of the cosmid clone 17H9 with other reference strains was illustrated based on 16S rRNA gene and 23S rRNA genes analysis. The 16S rRNA gene and 23S rRNA gene in the clone 17H9 are highlighted by bold characters. The IVS in the 23S rRNA gene on 17H9 was removed for the phylogenetic analysis. The dendrogram was constructed using the neighbor-joining method by DNAMAN program. Only bootstrap values above 50% based on 1000 replicates are shown. The scalebar represents 0.05 substitution per nucleic acid site.

encoded a conserved protein that was similar to sulfite oxidase and related enzymes. The other predicted ORFs match methyltransferase, DNA helicase, Sterol desaturase, transcriptional regulator, hypothetical protein and unknown protein.

Discussion

The studied deep-sea sediment from east Pacific Nodule Province consists of light-brown siliceous clay. Large amounts of potato-like polymetallic nodules were dispersed in the sediments of east Pacific Nodule Province. It is still unknown how the deep-sea polymetallic nodules formed, but it is suspected that microorganisms may have some role in the process. The sediments surrounding the nodules may contain important information on the nodule formation. Using 16S rRNA gene sequence analysis, a great diversity of microorganisms has been detected, but only *Halomonas*, *Marinobacter*, *Psychrobacter* and *Pseudoalteromonas* species of *Gammaproteobacteria* have been retrieved by cultivation-dependent methods (Xu *et al.*, 2004, 2005). Therefore, in this study, a strategy combining metagenomic library preparation and sequence analysis was utilized to obtain genetic information on the microorganisms present in the deep-sea sediment of Pacific Nodule Province.

Work from several groups has revealed that the metagenomic library approach is very useful in gaining new perspectives on the microbial ecology of natural environments (Béjá *et al.*, 2000b; Rondon *et al.*, 2000; Gillespie *et al.*, 2002; Liles *et al.*, 2003; Quaiser *et al.*, 2003). Metagenomic libraries containing large DNA inserts in BAC, Fosmid, and cosmid libraries are suitable to isolate large gene clusters for bioactive compounds (Gillespie *et al.*, 2002; Courtois *et al.*, 2003; Liles *et al.*, 2003; Ginolhac *et al.*, 2004), or to partly elucidate the physiological properties of uncultivated microorganisms by partial genetic characterization (Béjá *et al.*, 2000a; Moreira *et al.*, 2006). In this study, a metagenomic cosmid library containing around 120 Mb DNA was constructed from a deep-sea sediment (5274 m in depth) of east Pacific Nodule Province. To the authors' knowledge, this is the first metagenomic library of a deep-sea sediment (water depth of more than 1000 m). The constructed cosmid library mainly consists of *Proteobacteria* including *Alphaproteobacteria*, *Gammaproteobacteria*, and *Deltaproteobacteria*, *Actinobacteria* and Firmicutes. The bacterial community in the cosmid library was not totally consistent with that of previous investigations by analysis of the deep-sea sediment directly (Xu *et al.*, 2004, 2005, and authors' unpublished data). Although it was also found that *Proteobacteria* were dominant in the bacterial community,

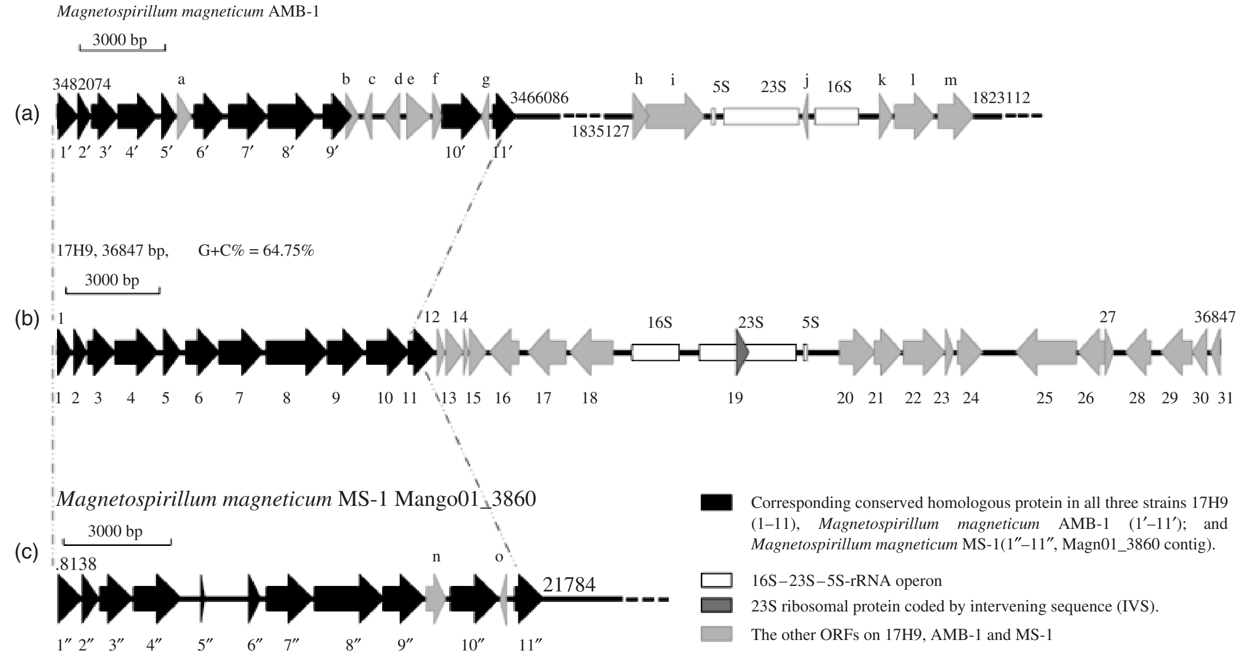
Magnetospirillum magneticum AMB-1
3000 bp
(a)
3482074
a
b c d e f g
3466086
1835127
h i 5S 23S j 16S k l m
1823112
1′ 2′ 3′ 4′ 5′ 6′ 7′ 8′ 9′ 10′ 11′
17H9, 36847 bp, G+C% = 64.75%
3000 bp
(b)
1
12 14
16S 23S 5S
27 36847
1 2 3 4 5 6 7 8 9 10 11 13 15 16 17 18 19 20 21 22 23 24 25 26 28 29 30 31
Magnetospirillum magneticum MS-1 Mango01_3860
3000 bp
(c)
.8138
n o
21784
1″ 2″ 3″ 4″ 5″ 6″ 7″ 8″ 9″ 10″ 11″
Corresponding conserved homologous protein in all three strains 17H9 (1–11), Magnetospirillum magneticum AMB-1 (1′–11′); and Magnetospirillum magneticum MS-1(1″–11″, Magn01_3860 contig).
16S–23S–5S-rRNA operon
23S ribosomal protein coded by intervening sequence (IVS).
The other ORFs on 17H9, AMB-1 and MS-1

Fig. 3. Gene organization of the conserved gene clusters and ORFs around the 5S–16S–23S rRNA gene operon in *Magnetospirillum magneticum* AMB-1 (a), cosmid clone 17H9 (b) and *Magnetospirillum magnetotacticum* MS-1 (c). The putative ORFs in the clone 17H9 are represented by 1–31. The description of ORF1–31 on 17H9 is given in the supplementary Table S3. The homologous ORFs in AMB-1 and MS-1 corresponding to the ORF1–11 in 17H9 are named as 1′–11′ (Amb3214–Amb3210, Amb3208–Amb3205, Amb3199, Amb3197), and 1″–11″ (Magn03010003–Magn03010011, Magn03010013, Magn03010015), respectively. Other ORFs in AMB-1 and MS-1 are represented by letters a to o (a, amb3209; b–f, amb3204–3200, g, amb3198, h–m, amb1688–amb1683; n, Magn03010012, o, Magn0301001), respectively. a–g represent gene insertions between ORF5′ and 6′, ORF9′ and 10′ and ORF 10′ and 11′ in AMB-1, while n and o represent gene insertions between ORF 9″ and 10″, and ORF10″ and 11″ in MS-1, as compared with 17H9.

Gammaproteobacteria, but not the *Deltaproteobacteria* was predominant (Xu *et al*., 2004, 2005). The inconsistency of the bacterial community composition as revealed by cosmid library and the direct 16S rRNA gene analysis has also been observed previously (Liles *et al*., 2003). The possible reason for this inconsistency could be due to the different cell lysis during the DNA preparation. This may also be part of the reason for this present case, although it is still not clear. Another reason for the inconsistency between the results obtained by metagenomic cloning and sequencing compared with direct sequencing of 16S rRNA genes is that very few sequences representative of the community were cloned using the metagenomic approach. The clone library was small, in the present case, only 37 clones were selected for RFLP analysis, and 13 clones of different RFLP types were sequenced. Therefore, one cannot expect to obtain a representation of all of the sequences present in the sample. More sequences should have been obtained by direct sequencing of 16S rRNA genes. However, this excludes analysis of functional genes that justifies use of the metagenomics approach. Most of the strains in the cosmid library represent uncharacterized novel bacterial species; the deep-sea metagenomic library provides an opportunity to look into the genome of the deep-sea microbial communities, infer the physiological properties of uncultivated microorganisms and to search for novel genes or gene clusters involved in specific biochemical pathways.

Alphaproteobacteria are ubiquitous in deep-sea environments; however, their physiology and roles in the biogeochemical cycling in the deep-sea environment remain elusive. In this study, two cosmid clones were found to contain Alphaproteobacterial 16S rRNA genes. The whole fragment sequences of these two clones were determined and analyzed. Important information was obtained from the genomic sequences, and this is the first time that the genomic information about the deep-sea *Alphaproteobacteria* has been partly revealed.

The 16S and 23S rRNA gene sequences on the 17H9 clone both revealed that the clone came from a novel *Alphaproteobacterium*, which is closely related to *Magnetospirillum* species (Fig. 2 and Table S3). In the 23S rRNA gene, an intervening sequence (IVS) containing an ORF was found. The ORF encoded a conserved hypothetical 23S ribosomal protein that had 62% identity and 81% similarity (supplementary Table S3) to those in the IVS identified from some species of *Coxiella* and *Leptospira* (Ralph & Mcclelland, 1993; Afseth *et al.*, 1995). Such IVSs have been found in some species of a broad range of bacteria (Ralph & Mcclelland, 1993; Afseth *et al.*, 1995); the IVSs may be spliced out posttranscriptionally, but why they persist in some bacteria and not others still remains a mystery. Upstream of the 16S–23S–5S rRNA gene operon, ORF 1-11 had high identity to those of the *Magnetospirillum* species and the gene organization was also highly conserved among the *Magnetospirillum* species (supplementary Table S3 and Fig. 3a–c). Among the conserved genes, four encode transport proteins including ferric ion uptake proteins essential for the magnetosome formation. The phylogenetic relationship, together with the highly conserved genes and their organization between 17H9 and those from *Magnetospirillum*, strongly suggested that the deep-sea bacterium may be a novel *Magnetospirillum* species. In AMB-1, there are gene insertion between ORF5′ and 6′, ORF9′ and 10′ and ORF 10′ and 11′, while in MS-1, gene insertions was observed between ORF 9″ and 10″, and ORF10″ and 11″, as compared with 17H9 (Fig. 3a–c). It is not clear whether insertion or deletion events occurred in the gene cluster evolution history. Magnetotactic bacteria (MTB) are a heterogeneous group of aquatic microorganisms that are abundant in the environment. MTB are known to be actively involved in Fe and S cycling. MTB are assumed to have a considerable impact in the biogeochemical cycling in nature sediment. It is not surprising that possible novel MTB bacteria exist in the deep-sea sediment of the Pacific Nodule Province. This environment contains a large number of metal-rich polymetallic nodules, and the MTB may play important roles in the metal cycling in this specific environment. Based on the genetic information obtained in this study, the authors are now designing media to enrich and isolate deep-sea MTB.

Interestingly, downstream from the 16S–23S–5S rRNA gene operon, a gene cluster containing four ORFs (ORF20-23) was identified to encode proteins that had high homology with those in thermophilic archaea. These ORFs encode putative myo-inositol-1-phosphate (I-1-P) synthase (EC 5.5.1.4), inositol monophosphatase (I-1-Pase) and glucose-1-phosphate thymidylyl transferase. ORF23 encodes a polypeptide that is the N-terminal part of the I-1-P synthase encoded by ORF20. ORF 23 may not be functional, probably a pseuodogene that has also been observed in metagenomic analysis of a novel Deltaproteobacterial group (Moreira *et al.*, 2006). I-1-P synthase and I-1-Pase are two crucial enzymes for the synthesis of Di-myo-inositol-1, 1′-phosphate (DIP) is found in hyperthermophilic archaea, a novel solute produced in response to heat stress (Shockley *et al.*, 2003). The existence of gene cluster of I-1-P synthase and I-1-Pase suggested that the deep-sea organism may be capable of synthesizing the inositol solute to aid survival in low-temperature, high-pressure deep-sea environments. This gene cluster may be derived from horizontal gene transfer; however, the possibility of heterogeneous cloning cannot be excluded.

Acknowledgements

The authors would like to thank the crews on Dayang N0.1 for help with collecting the samples. This work was financially supported by NSF fund (40625016, 40476054), 973 fund (2004CB719601).

Authors'contribution

M.X. and F.W. contributed equally to this study.

References

Afseth G, Mo YY & Mallavia LP (1995) Characterization of the 23S and 5S rRNA genes of *Coxiella burnetii* and identification of an intervening sequence within the 23S rRNA gene. *J Bacteriol* **177**: 2946–2949.

Béjá O, Aravind L, Koonin EV *et al.* (2000a) Bacterial rhodopsin: evidence for a newtype of phototrophy in the sea. *Science* **289**: 1902–1906.

Béjá O, Suzuki MT, Koonin EV *et al.* (2000b) Construction and analysis of bacterial artificial chromosome libraries from a marine microbial assemblage. *Environ Microbiol* **2**: 516–529.

Courtois S, Cappellano CM, Ball M *et al.* (2003) Recombinant environmental libraries provide access to microbial diversity for drug discovery from natural products. *Appl Environ Microbiol* **69**: 49–55.

DeLong EF (2002) Microbial population genomics and ecology. *Curr Opin Microbiol* **5**: 520–524.

DeLong EF (2005) Microbial community genomics in the ocean. *Nat Rev Microbiol* **3**: 459–469.

DeLong EF & Karl DM (2005) Genomic perspectives in microbial oceanography. *Nature* **437**: 336–342.

DeLong EF, Preston CM, Mincer T *et al.* (2006) Community genomics among stratified microbial assemblages in the ocean's interior. *Science* **311**: 496–503.

Doolittle RF (2002) Microbial genomes opened up. *Nature* **392**: 339–342.

Ghiorse WC (1984) Biology of iron- and manganese-depositing bacteria. *Annu Rev Microbiol* **38**: 515–550.

Gillespie DE, Brady SF, Bettermann AD, Cianciotto NP, Liles MR, Rondon MR, Clardy J, Goodman RM & Handelsman J (2002) Isolation of antibiotics turbomycin A and B from a metagenomic library of soil microbial DNA. *Appl Environ Microbiol* **68**: 4310–4306.

Ginolhac A, Jarrin C, Gillet B *et al.* (2004) Phylogenetic analysis of polyketide synthase I domains from soil metagenomic libraries allows selection of promising clones. *Appl Environ Microbiol* **70**: 5522–5527.

Gounot AM (1994) Microbial oxidation and reduction of manganese: consequences in groundwater and applications. *FEMS Microbiol Rev* **14**: 339–349.

Henne A, Schmitz RA, Bömeke M, Gottschalk G & Daniel R (2000) Screening of environmental DNA libraries for the presence of genes conferring lipolytic activity on *Escherichia coli*. *Appl Environ Microbiol* **66**: 3113–3116.

Hunter EM, Mills HJ & Kostka JE (2006) Microbial Community Diversity Associated with Carbon and Nitrogen Cycling in Permeable Shelf Sediments. *Appl Environ Microbiol* **72**: 5689–5701.

Knietsch A, Bowien S, Whited G, Gottschalk G & Daniel R (2003) Identification and characterization of coenzyme B_{12}-Dependent glycerol dehydratase- and diol dehydratase-encoding genes from metagenomic DNA libraries derived from enrichment cultures. *Appl Environ Microbiol* **69**: 3048–3060.

Li L, Kato C & Horikoshi K (1999) Bacterial diversity in deep-sea sediments from different depths. *Biodiv Conserv* **8**: 659–677.

Liles MR, Manske BF, Bintrim SB, Handelsman J & Goodman RM (2003) A census of rRNA genes and linked genomic sequences within a soil metagenomic library. *Appl Environ Microbiol* **69**: 2684–2691.

Mitra NG, Sachidanand B, Agarwal GD & Upadhyay A (1998) Microbial transformations of iron, manganese and copper in soil. *Acta Botanica Indica* **26**: 71–81.

Moreira D, Rodríguez-Valera F & López-García P (2006) Metagenomic analysis of mesopelagic Antarctic plankton reveals a novel deltaproteobacterial group. *Microbiology* **152**: 505–517.

Nealson KH (1997) Sediment bacteria: who's there, what are they doing, and what's new? *Annu Rev Earth Planet* **25**: 403–434.

Nealson KH & Saffarini D (1994) Iron and manganese in anaerobic respiration: environmental significance, physiology, and regulation. *Annu Rev Microbiol* **48**: 311–343.

Northup DE, Barns SM, Yu LE *et al.* (2003) Diverse microbial communities inhabiting ferromanganese deposits in Lechuguilla and Spider Caves. *Environ Microbiol* **5**: 1071–1086.

Piper DZ & Williamson M (1977) Composition of Pacific Ocean ferromanganese nodules. *Mar Geo* **23**: 285–303.

Quaiser A, Ochsenreiter T, Lanz C, Schuster SC, Treusch AH, Eck J & Schleper C (2003) *Acidobacteria* form a coherent but highly diverse group within the bacterial domain: evidence from environmental genomics. *Mol Microbiol* **50**: 563–575.

Ralph D & Mcclelland M (1993) Intervening sequence with conserved open reading frame in eubacterial 23S rRNA genes. *Proc Natl Acad Sci USA* **90**: 6864–6868.

Rappe MS, Kemp PF & Giovannoni SJ (1997) Phylogenetic diversity of marine coastal picoplankton 16S rRNA genes cloned from the continental shelf off Cape Hatteras, North Carolina. *Limnol Oceanogr* **42**: 811–826.

Reardon CL, Cummings DE, Petzke LM, Kinsall BL, Watson DB, Peyton BM & Geesey GG (2004) Composition and diversity of microbial communities recovered from surrogate minerals incubated in an acidic uranium-contaminated aquifer. *Appl Environ Microbiol* **70**: 6037–6046.

Rondon MR, August PR, Bettermann AD *et al.* (2000) Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. *Appl Environ Microbiol* **66**: 2541–2547.

Shockley KR, Ward DE, Chhabra SR, Conners SB, Montero CI & Kelly RM (2003) Heat Shock Response by the Hyperthermophilic Archaeon *Pyrococcus furiosus*. *Appl Environ Microbiol* **69**: 2365–2371.

Stein LY, La Duc MT, Grundl TJ & Nealson KH (2001) Bacterial and archaeal populations associated with freshwater ferromanganous micronodules and sediments. *Environ Microbiol* **3**: 10–18.

Tatusov RL, Fedorova ND, Jackson JD *et al.* (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**: 41.

Tyson GW, Chapman J, Hugenholtz P *et al.* (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**: 37–43.

Venter JC, Remington K, Heidelberg JF *et al.* (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**: 66–74.

Voget S, Leggewie C, Uesbeck A, Raasch C, Jaeger KE & Streit WR (2003) Prospecting for novel biocatalysts in a soil metagenome. *Appl Environ Microbiol* **69**: 6235–6242.

Xu M, Wang P, Wang F & Xiao X (2004) Microbial diversity in deep-sea sediments collected from different stations of Pacific. *Mar Biotechnol* **6**: S161–S167.

Xu M, Wang P, Wang F & Xiao X (2005) Microbial diversity at a deep-sea station of the Pacific Nodule Province. *Biodiv Conserv* **14**: 3363–3380.

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

Supplementary material

The following supplementary material is available for this article:

Table S1. Description of the end-sequences of cosmid clones in ES0303 cosmid library (E value > 0.1, Length > 200 bp).

Table S2. Putative ORFs on cosmid clone 10D02 and their information (E value > 0.005).

Table S3. Putative ORFs on cosmid clone 17H9 and their information (E value > 0.005).

This material is available as part of the online article from: http://www.blackwell-synergy.com/doi/abs/10.1111/j.1574-6941.2007.00377.x (This link will take you to the article abstract).

Please note: Blackwell Publishing are not responsible for the content or functionality of any supplementary materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.