Environmental Microbiology (2009) **11**(9), 2412–2424 doi:10.1111/j.1462-2920.2009.01969.x

Metagenomic approach studying the taxonomic and functional diversity of the bacterial community in a mesotrophic lake (Lac du Bourget – France)

Didier Debroas,[1]* Jean-François Humbert,[2,3] François Enault,[1] Gisèle Bronner,[1] Michael Faubladier[1] and Emmanuel Cornillot[1†]

[1]*Université Blaise Pascal – Laboratoire «Microorganismes: génome et environnement» – UMR CNRS 6023 63177 Aubiere cedex, France.*

[2]*INRA, UMR CARRTEL, BP 511, 74203 Thonon Cedex, France.*

[3]*Institut Pasteur-URA CNRS 2172, Unité des Cyanobactéries, 28 rue du Dr Roux, 75724 Paris Cedex 15, France.*

Summary

The main goals of this work were to identify the metabolic pathways of the bacterial community in a lacustrine ecosystem and to establish links between taxonomic composition and the relative abundances of these metabolic pathways. For this purpose, we analysed a 16S rRNA gene library obtained by gene amplification together with a sequence library of both insert ends on *c.* 7700 fosmids. Whatever the library used, *Actinobacteria* was the most abundant bacterial group, followed by *Proteobacteria* and *Bacteroidetes*. Specific aquatic clades such as acI and acIV (*Actinobacteria*) or LD12 and GOBB-C201 (*Alphaproteobacteria*) were found in both libraries. From comparative analysis of metagenomic libraries, the metagenome of this lake was characterized by overrepresentation of genes involved in the degradation of xenobiotics mainly associated with *Alphaproteobacteria*. *Actinobacteria* were mainly related to metabolic pathways involved in nucleotide metabolism, cofactors, vitamins, energy, replication and repair. *Betaproteobacteria* appeared to be characterized by the presence of numerous genes implicated in environmental information processing (membrane transport and signal transduction) whereas glycan and carbohydrate metabolism pathways were overrepresented in *Bacteroidetes*. These results prompted us to propose hypotheses on the ecological role of these bacterial classes in lacustrine ecosystems.

Received 10 November, 2009; accepted 8 May, 2009. *For correspondence. E-mail didier.debroas@univ-bpclermont.fr; Tel. (+33) 473 407 83771; Fax (+33) 473 407 87670. †Present address: Laboratoire de Biologie Cellulaire et Moléculaire UMR CNRS 5235 DIMNP/ERT 1038 'Vaccination antiparasitaire' Université Montpellier I, 34093 Montpellier.

Introduction

Freshwater accounts for only 2.5% of the total volume of water available on our planet, and much of it, being stored in form of ice, is not readily accessible. Despite being exposed to growing ecological and economical impacts and interests, freshwater ecosystems are paradoxically much less researched than marine ecosystems and less than 10% of limnology data are focused on viruses, bacteria, fungi and protists and their metabolic processes in freshwaters (Wetzel, 2002). Although there are numerous papers describing the composition and structure of freshwater microbial communities, most of them are based only on the study of rRNA genes and thus provide no information on the functional diversity of these communities. Thus, although the last 15 years has seen rRNA clones reveal numerous new lineages, we know very little on the specific biological properties of some of the most abundant and widespread organisms found in freshwater ecosystems and ecosystems in general.

Among the methods able to link function and diversity, sequencing clone libraries of environmental DNA appears as most promising (Handelsman, 2004). Random cloning and large-scale sequencing enable to simultaneously assess phylogenetic diversity, potential metabolic pathways and horizontal gene transfer in the environment without running amplification procedures like clone cultivation or PCR. Environmental genomics, like conventional genomics, is an excellent tool for opening access to information covering a wide range of aspects such as identification of antibacterial or antitumoural agents (e.g. Osburne *et al.*, 2000), or metabolism and physiological processes such as the identification of a novel type of rhodopsin (Beja *et al.*, 2000). Environmental genomics methods can also provide additional phylogenetic information to that provided by rRNA (Stein *et al.*, 1996) and detect divergent rRNAs that would be missed by PCR approaches (López-García *et al.*, 2004).

Metagenomics has been used to study prokaryotic community composition in a wide range of marine environments, including open oceans, coastal zones and estuaries at different latitudes, longitudes and depths (see DeLong, 2006). Metagenomics has also been applied for studying soils (e.g. Daniel, 2006; Riaz *et al*., 2008), the human gut microbiome (e.g. Frank and Pace, 2008) and sediments (e.g. Abulencia *et al*., 2006). A metagenomic study was recently performed on nine biomes (Dinsdale *et al*., 2008) using a 454 sequencing strategy in order to compare their functional diversity. These authors were able to clearly differentiate microbial and viral metagenomes. Nevertheless, few studies have been applied to freshwater ecosystems (Cottrell *et al*., 2005 in rivers; Pope and Patel, 2008 in freshwater cyanobacterial bloom; Dinsdale *et al*., 2008 in fish ponds).

Thus, in order to gain further insight into the species and functional diversity of a lacustrine microbial community, we developed a metagenomic approach studying the prokaryotic community collected in the largest natural lake in France, the Lac du Bourget. A clone library, named METAPROC, was constructed and the 3′ and 5′ fosmid ends of all these clones were sequenced. In parallel, we evaluated the taxonomic composition of the bacterial community by 16S rRNA gene amplification and sequencing in order to compare the composition and structure of this community with that obtained from our metagenomic data. The main metabolic pathways were identified in the bacterial community and compared with metagenomic results obtained from photic zones of estuary, coastal and open-ocean environments in order to highlight the specificities of lacustrine bacterial communities.

Results

General features of the fosmid library

The METAPROC library was constructed from a planktonic fraction smaller than 1.2 μm sampled from the photic zone of Bourget Lake. Amplification by universal 18S primers (Lefranc *et al*., 2005) and microscopic observations showed that there were no detectable picoeukaryotes in this size fraction. Our library contained 7746 fosmid clones that were subjected to bi-directional end-sequencing, yielding 12 Mbp of DNA sequences from the approximately 271 Mbp total archive. This represents raw sequences of approximately six prokaryotic genome equivalents, based on a 2 Mbp average genome size (Button and Robertson, 2001). Among the 7746 fosmid clones, 6709 were used for the phylogenetic affiliation because both fosmid-end sequences gave the same affiliation with a bit-score > 100. This fosmid library shows that the main microorganisms present in this sample were bacteria (6634 hits), whereas viruses, eukaryotes and archaea represented about 1% of clones.

Bacterial community composition

The composition and structure of the bacterial community from Bourget Lake were estimated both from analysis of the 282 sequences obtained after PCR amplification and cloning of a 550 bp 16S rRNA fragment, and from the taxonomic assignment of fosmid insert terminal sequences in our metagenomic library. More precisely, our metagenomic library returned hits for 20 16S rRNA genes with a sequence length > 300 bp, while screening BLASTx data against the non-redundant protein database (nr) pinpointed 190 sequences with phylogenetic markers among a total of 10712 sequences (bit-score > 100 and *e*-value < 10^{-15}).

Whatever the database used, *Actinobacteria* was the dominant bacterial group, followed by *Alpha*- and *Betaproteobacteria* and *Bacteroidetes* (Fig. 1). *Actinobacteria*, *Betaproteobacteria*, *Alphaproteobacteria* and *Bacteroidetes* represented, respectively, 43.5%, 22.9%, 14% and 7% of all fosmid library clones, and 47.7%, 23.6%, 19.5% and 5.6% if only phylogenetic markers are taken into account in the protein database. In this case, only three sequences are clustered in other phyla of the bacterial domain (Fig. 1), and the results concurred with those obtained from 16S rRNA gene analysis. There were no significant differences (Chi-square test) in terms of relative proportion of *Actinobacteria*, *Proteobacteria* and *Bacteroidetes* when comparing data resulting from the taxonomic analysis on the fosmid library and on the 16S rRNA gene library. However, there were significant differences (Chi-square test, $P < 0.05$) when *Alpha*- and *Betaproteobacteria* were considered separately.

The distribution among typical freshwater ecosystem clades was determined using 16S rRNA gene sequences (Fig. 2A and B). The composition of the bacterial community resulting from the phylogenetic affiliation of the 16S sequences detected by BLASTn (bit-score > 100) in the fosmid library was similar to that resulting from the analysis of the 16S rRNA gene library. In this library, 48.5% (33) of all operational taxonomic units (OTUs) belonged to *Actinobacteria* (Fig. 2A). The OTUs of the clades acI and acIV represented 72.9% (14 sequences) and 14.8% (9 sequences), respectively, of all *Actinobacteria* sequences. Similarly, the *Proteobacteria* were dominated by a handful of clades: clades LD12 (close to SAR11), GOBB-C201 (94.6% of clones belonging to *Alphaproteobacteria*) and Polynucleobacter (19% of clones belonging to *Betaproteobacteria*). Sequences belonging to *Bacteroidetes* were mainly retrieved in the 16S rRNA gene library (Fig. 2B).

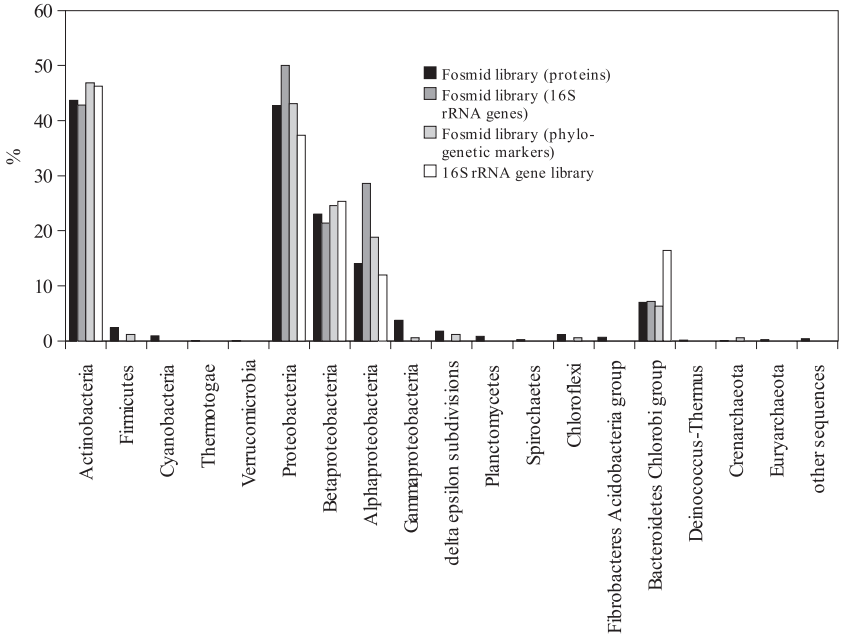

Fig. 1. Taxonomic affiliation of clones in METAPROC and 16S rRNA libraries expressed as the percentage of clones in each library (i.e. 7746 and 282 clones) respectively. This affiliation in METAPROC was established based on BLAST against the nr database, 16S rRNA genes and different proteic phylogenetic markers.

Gene content and metabolic potential

Fosmid-end sequences were stratified into functional classes according to the Kyoto Encyclopaedia of Genes and Genomes (KEGG). The BLASTx against the KEGG database gave 11 499 significant results according to our criteria (bit-score > 100 and *e*-value < 10^{-15}). According to this classification, 75.9% of CDS carried by fosmid-ends were related to metabolic function, 13.3% to housekeeping genes, 7% to environmental information processing and 3.2% to cellular processes.

The most abundant metabolic pathways were associated with membrane transport. The most important of these was ATP-binding cassette (ABC) transporters, which accounted for 5.4% (724 hits) of genes associated with a KEGG identifier (Fig. 3). The following best-represented pathways were associated with housekeeping genes, with DNA polymerase and aminoacyl-tRNA biosynthesis returning 670 and 495 hits respectively. The following 15 categories were related to metabolism pathways. Among them, the best-represented pathways were purine metabolism, oxidative phosphorylation and glycine, serine and threonine metabolism, representing 336, 334 and 325 hits respectively.

We compared the functional diversity estimated from our metagenomic library with the results obtained using the same approach in the central North Pacific Ocean (DeLong *et al*., 2006) and in coastal environment, estuary, coastal sea and open ocean (Rusch *et al*., 2007). We chose these metagenomic libraries because similarly to our study, they targeted aquatic microbial communities from the euphotic zone but in contrasted salinity and productivity conditions.

The first two axis generated by the correspondence analysis (COA) (Fig. 4) maximize the correspondence between the variations in the relative abundance of each metabolic pathway and the aquatic metagenomes selected. The percentage of total variance extracted by these axes was 90.5% (75.4% + 15.1%) and there was a significant relationship between these two variables (Chi-square test, $P < 0.001$). In particular, this analysis showed that the metagenome from the bacterial community of the Lac du Bourget was distinguished from other bacterial metagenomes by the relative importance of genes involved in xenobiotics biodegradation and metabolism pathways on axis 1 and glycan biosynthesis and metabolism, biosynthesis of polyketides and non-ribosomal peptides and signal transduction, on axis 2. On the other hand, genes involved in nucleotide metabolism and energy metabolism were strongly associated with the other metagenomes than with ours. A similar pattern was obtained by further analysis (Fig. 5) in which we estimated ratios between proportions of genes on the basis of the metabolic pathways defined in our project. Only the metabolic pathways where the abundance of genes found in METAPROC were significantly different from the average of the abundances in the other metagenomes were selected (Chi-square test, $P < 0.05$). To highlight the differences, Fig. 5 only shows the metabolic pathways where all the contributions of a given KEGG category divided by the same category in METAPROC were > 1 or < 1. For example, pyruvate metabolism was selected because cross-comparison between our results and those of all five selected metagenomes gave a > 1 result, whereas, for example, nucleotide sugar metabolism was not presented because this ratio was > 1 for the

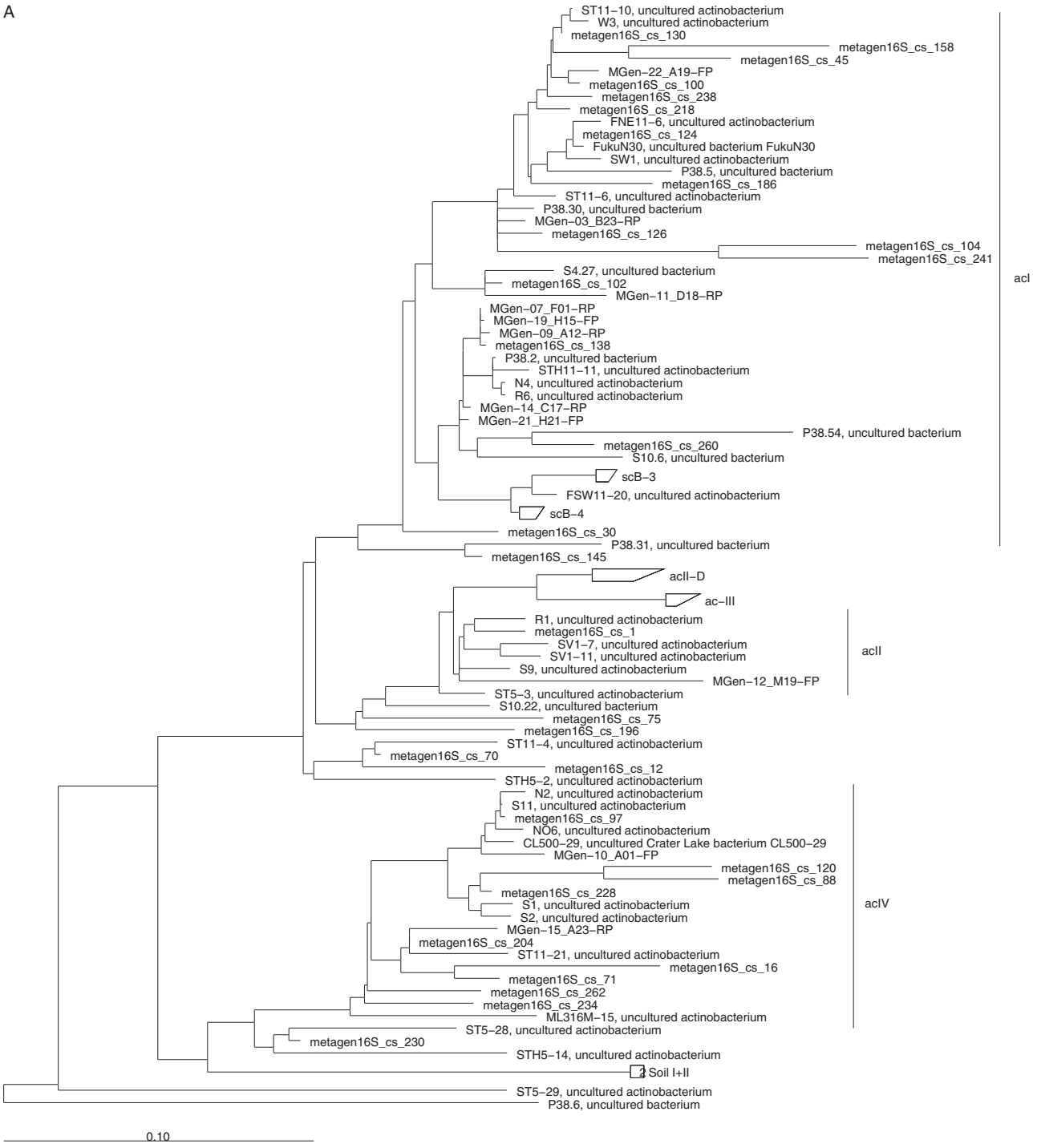

Fig. 2. Phylogenetic tree of bacterial small-subunit rRNA genes from the METAPROC library (Mgen) and the 16S rRNA library (metagen16S_cs).
A. Phylogenetic tree covering the diversity of *Actinobacteria*.
B. Phylogenetic tree covering the diversity of *Proteobacteria* and *Bacteroidetes*.

central North Pacific Ocean and Chesapeake Bay, but < 1 for the other three marine ecosystems.

This analysis confirmed that the metagenome of the bacterial community of the Lac du Bourget was characterized, for example, by an overrepresentation of genes involved in xenobiotics degradation and glycan metabolism and by an underrepresentation of genes involved in certain amino acid metabolisms. In marine ecosystems,

Fig. 2. *cont.*

energy metabolism is related to nitrogen metabolism while carbon fixation is related to reductive carboxylate cycle (CO_2 fixation) in phototrophic bacteria. These two metabolisms can be related to porphyrin and chlorophyll metabolism, which are overrepresented in estuary and marine systems compared with METAPROC.

Metabolic function related to the main phyla

In order to compare the metabolic potential of the four main lacustrine phyla (*Actinobacteria*, *Betaproteobacteria*, *Alphaproteobacteria* and *Bacteroidetes*), we performed a COA on the main KEGG categories associated with these different bacterial classes (Fig. 6). There was a significant relationship between these four phyla and the variations in the relative abundance of the main KEGG categories in each of them (Chi-square test, $P < 0.001$). The most important KEGG pathways associated with bacterial groups were those with the greatest scores on each axis. Thus, it appeared that genomes of freshwater *Alphaproteobacteria* (mainly SAR11-LD12 and GOBB-

Fig. 3. Distribution of METAPROC fosmid-end sequences among KEGG pathways. The percentages were calculated by dividing the abundance of genes in a KEGG category by the sum of genes identified.

C201) were enriched in genes belonging to xenobiotic degradation, biosynthesis of polyketides and non-ribosomal peptides (ansamycin biosynthesis), and the biosynthesis of secondary metabolite pathways and lipid metabolism. More precisely, this enrichment concerned genes implicated in benzoate degradation via CoA ligation, gamma-hexachlorocyclohexane degradation, naphthalene and anthracene degradation and caprolactam degradation (> 20 hits for each). Biodegradation of monocyclic (limonene) and bicyclic (pinene) terpenes were the main metabolic pathways classified as biosynthesis of secondary metabolites. In the Lac du Bourget, 11.1% of the *Alphaproteobacteria*-annotated genes belonged to the xenobiotic degradation metabolic pathway, whereas in the 188 prokaryotic genomes selection, this pathway represents only 6.2% of bacterial genes and 5.1% of *Alphaproteobacteria* genes like *Rickettsias* or *Rhizobacteria*. Similarly, this overrepresentation of genes involved in xenobiotic degradation was also found in *Betaproteobacteria* (7.6%) and *Bacteroidetes* (7.2%) from the Lac du Bourget compared with data present in main bacterial genome databases (4.6%).

At the opposite on the axis 1, *Actinobacteria* were mainly associated with nucleotide metabolism, and by decreasing order with replication and repair, metabolism of cofactors and vitamin and energy metabolism (Fig. 6). However, the proportions of these pathways mirrored those in bacterial genomes already sequenced. Among these KEGG pathways, aminoacyl-tRNA biosynthesis (260 hits), DNA repair (347 hits), purine metabolism (179 hits), ubiquinone biosynthesis (100 hits) and oxidative phosphorylation (199 hits) were the best-represented potential metabolic pathways.

Betaproteobacteria and *Bacteroidetes* were located between *Actinobacteria* and *Alphaproteobacteria* on axis 1 of the COA, meaning that these two bacterial groups were less discriminated by the metabolic pathways described above than the other two (Fig. 6). *Betaproteobacteria* appeared to be characterized by the presence of numerous genes involved in the environmental information processing category (membrane transport and signal transduction). In this group, the main pathways were ABC transporters (202 hits) and two-component systems (47 hits). Finally, genes implicated in the cell

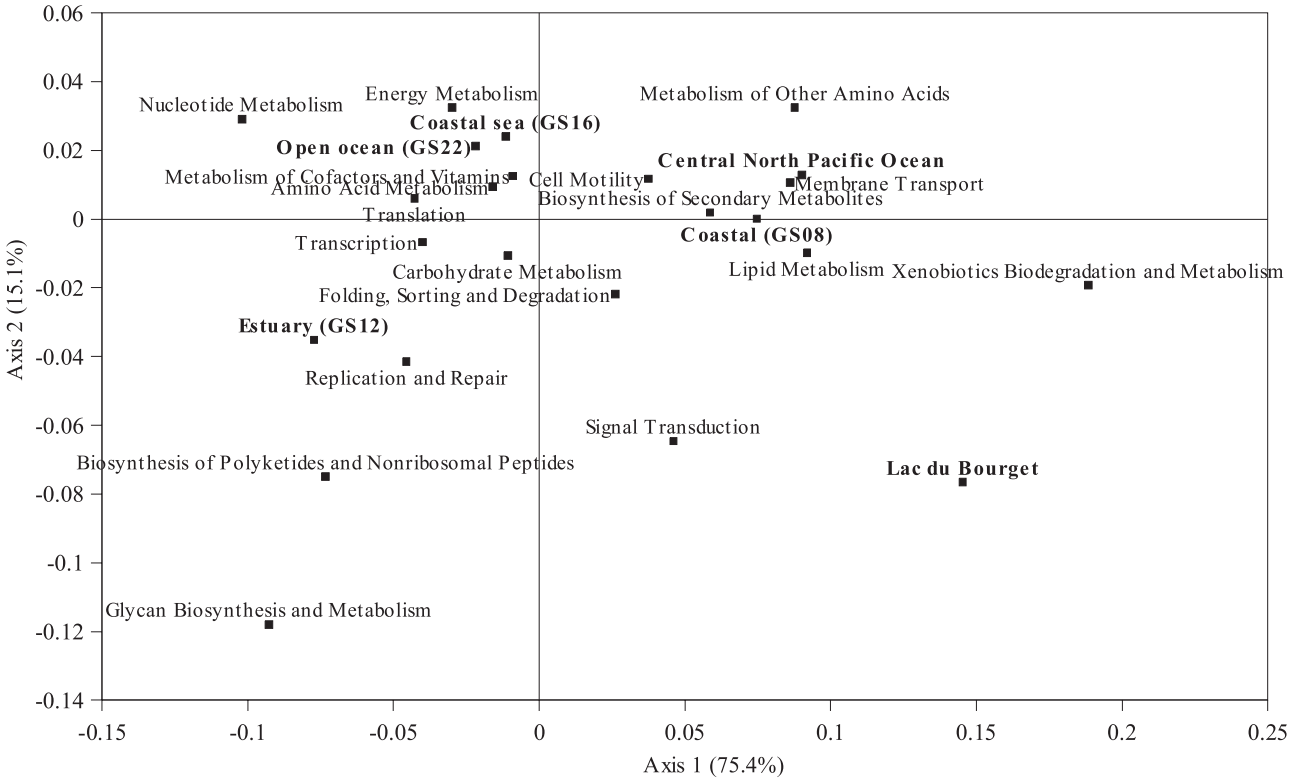

Fig. 4. COA obtained on major KEGG categories in five metagenomic libraries: METAPROC, the central North Pacific Ocean (DeLong *et al.*, 2006), coastal environments, estuaries, coastal sea and open ocean (Rusch *et al.*, 2007).

motility pathway were also able to distinguish this phylum from others.

Glycan biosynthesis (5.3%) and carbohydrate metabolism (20.1%) pathways were overrepresented in *Bacteroidetes* compared with other bacterial groups. Genes involved in metabolism of carbohydrates represented 16.3%, 14.1% and 17.2% of genes found in *Actinobacteria*, *Alphaproteobacteria* and *Betaproteobacteria* respectively. In particular, the amino sugars metabolism pathway was the most represented, which is interesting considering that amino sugars are a major component of bacteria and algae cells.

Discussion

Structure and composition of the bacterial community in Bourget Lake

The taxonomic binning of microbial protein homologues show that the main microbes detected in this metagenome belonged to *Bacteria* domain. The METAPROC metagenome shows that only 0.3% of fosmid-end sequences were affiliated to viruses, which is in agreement with other metagenomes constructed in the same way. For example, among the 964 094 open reading frames from the Sargasso sea metagenome, only 0.3% had significant similarity to phage genes stored in databases (Edwards and Rohwer, 2005). Concerning the global structure of the bacterial community, both 16S rRNA and fosmid libraries revealed the same dominance of *Actinobacteria* phylum, followed by *Alphaproteobacteria* and *Betaproteobacteria* and to a lesser extent *Bacteroidetes*. Depending on the method, significant differences were highlighted when *Alphaproteobacteria* and *Betaproteobacteria* were considered separately, while some groups escaped PCR detection. These differences could be explained by biases in PCR amplification or by non-accurate taxonomic classifications of our fosmid sequences. Accurate taxonomic classification of DNA fragments with high specificity can require up to 100 kb of training sequences as well as fragments > 1 kb (McHardy *et al.*, 2007), whereas fosmid-end sequences were generally < 1 kb (here, as well as in DeLong, 2006 and Martín-Cuadrado *et al.*, 2007). We sought to improve the phylogenetic assignment by comparing the best bit-scores obtained on both end-sequences from the same insert, and only clones having congruent taxonomic affiliations were retained for further analyses. This method was designed to eliminate biases of taxonomic annotation, such as those introduced by horizontal gene transfer.

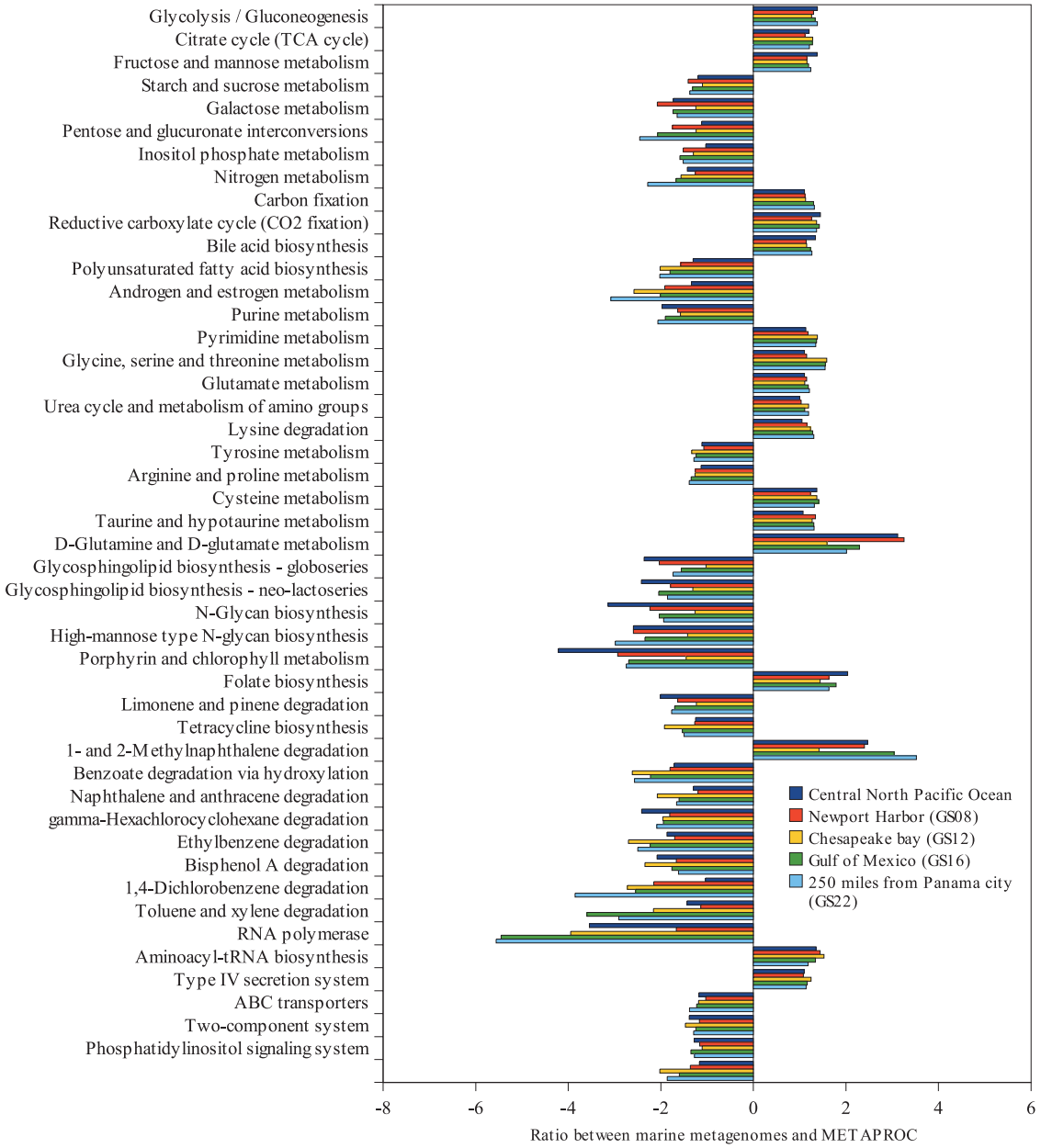

Fig. 5. Comparisons of major KEGG categories found in metagenomic libraries relative to METAPROC. For each category, a ratio was computed between the proportion of the pathway in one metagenome and METAPROC. This ratio was higher than 1 when the pathway studied was overrepresented in other aquatic ecosystems, and lower than 1 if this pathway was predominant in the METAPROC library.

The disagreement between 16S rRNA library and BLASTx records against nr could also be due to an overrepresentation of groups of bacteria for which the complete genome has been sequenced (e.g. *Gammaproteobacteria*), thereby generating biases in the gene annotation and, consequently, in the estimated relative proportions of the records for different bacterial groups. However, if we only consider the major proteic phylogenetic markers (see *Experimental procedures*) and rRNA gene sequences for taxonomic affiliation in the fosmid library, the two affiliation sets show strong congruence. Despite these potential biases and other well-known biases inherent to the PCR approach, we found a good agreement between both types of library to a high phylogenetic level, as well as between these results and established knowledge on the typical bacterial community composition patterns of lakes (Zwart *et al.*, 2002; Hahn, 2006). This means that our metagenomic library appears to be representative of the microbial diversity in the trophic zone of the Lac du Bourget.

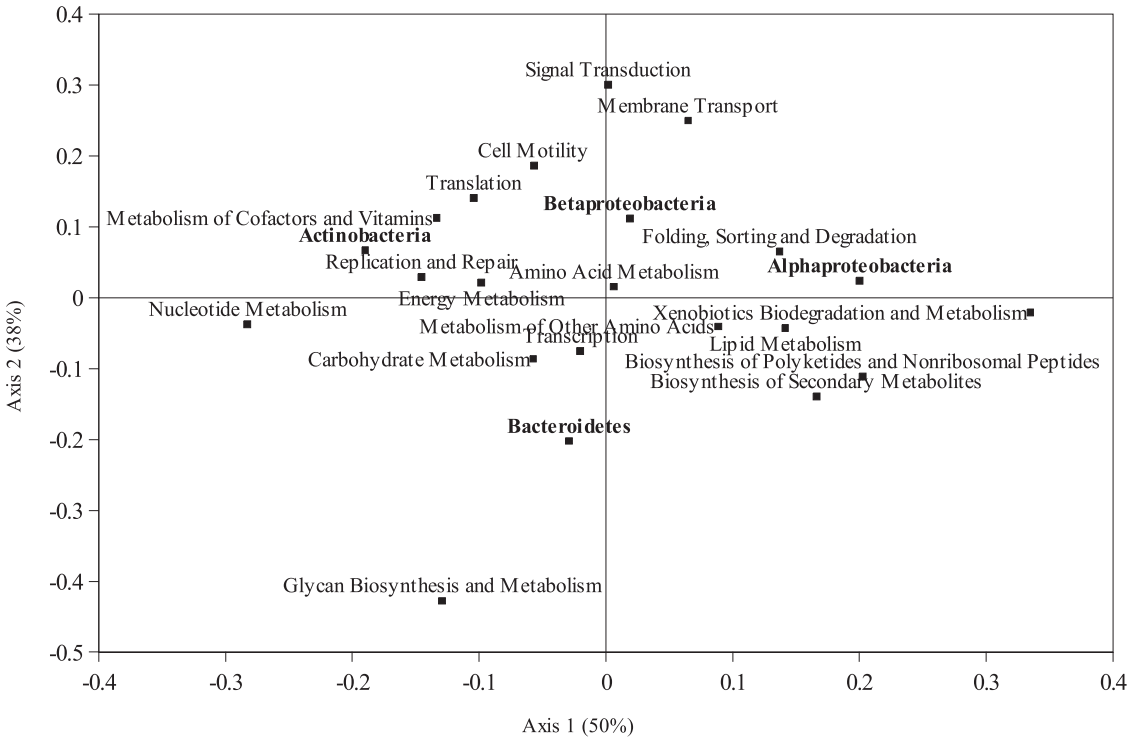

Fig. 6. Results of the COA obtained on major KEGG categories in the main bacterial groups identified in METAPROC: *Actinobacteria*, *Alphaproteobacteria*, *Betaproteobacteria* and *Bacteroidetes*.

However, Cottrell and colleagues (2005) also showed that PCR libraries were not essentially different from metagenomic libraries at a high phylogenetic level, although some differences can appear at a finer level. At a more fine taxonomic level, most of the METAPROC sequences appeared to belong to the *Actinobacteria* acI cluster defined by Warnecke and colleagues (2004). Most of the sequences from both libraries were distributed in the three subclusters acI-A, acI-B and acIV, in agreement with other studies on freshwater *Actinobacteria*. Moreover, in both libraries (16S rRNA and fosmid), a significant part of the sequence was distributed among *Proteobacteria* clusters: SAR11-LD12 and GOBB3-C201 in the *Alphaproteobacteria* class and polynucleobacter in the *Betaproteobacteria* class. *Bacteria* affiliated with the LD12 cluster (closely related to the ubiquitous marine SAR11 cluster) appear to be widely distributed in lakes of different types and from different regions around the world (Glöckner *et al*., 2000; Zwart *et al*., 2002; 2003). Thus, the METAPROC-based taxonomic composition of fosmid-end sequences was in agreement with 16S rRNA libraries showing that the most common clusters in freshwater ecosystems are present in the trophic zone of Bourget Lake. Most of these ubiquitous clusters clearly did not contain cultivated representatives as previously underlined by Hahn (2006).

Functional diversity of the bacterial community in the Lac du Bourget

Comparative genomics can yield comprehensive conclusions by comparing ecosystems or specific prokaryotic clusters. Water is an environment where substrate dilution is an important parameter. Thus, bacteria need sensor systems in order to optimize their use of all the resources available in their environment. ABC transporters couple ATP hydrolysis to the uptake and efflux of molecules (e.g. anions, small sugars, amino acids and even proteins) across the cell membrane and may play an essential role in competition between microbes and thus contribute to the selection of certain taxa. The METAPROC metagenome, and specifically the *Betaproteobacteria*, appeared to be enriched in genes involved in environmental information processing, such as these ABC transporters and two-component systems.

Among the other major bacterial classes in the lacustrine ecosystem, *Actinobacteria* were often in opposition to the other major taxonomic groups in terms of metabolic functions. The main metabolic pathways characterizing this bacterial class were clearly related to the production of bacterial compounds and the production of energy by the electron transport chain. These features may confer *Actinobacteria* a competitive advantage

materialized by more efficient growth in lacustrine ecosystems. Simek and colleagues (2005) showed that *Actinobacteria* growth in a grazer-free treatment (bacterial growth) was close to *Betaproteobacteria* growth. However, we cannot address the possible specific biogeochemical role of these organisms by sequence annotation alone. Another hypothesis is that *Actinobacteria* in freshwater ecosystems could very efficiently use the low-molecular-weight products (< 700 Da), which dominate compounds derived from algal excretion and which are an important carbon source for heterotrophic microorganisms (Maurin *et al*., 1997; Richardot *et al*., 2001). This interpretation is in agreement with Allgaier and colleagues (2007) who used analysis of covariance of bacterioplankton composition and environmental variables to show that phytoplankton-derived DOM was a determinant factor in *Actinobacteria* community dynamics. Another feature that can explain the success of this clade in freshwater is the overrepresentation of metabolic pathways involved in replication and repair compared with the other bacterial groups. These genetic mechanisms could explain the resistance to UV solar radiation observed by Warnecke and colleagues (2005) who demonstrated a correlation between this radiation and the per cent abundance of the acI clade.

The potential metabolic function described above could contribute to the dominance of *Betaproteobacteria* and *Actinobacteria* in lakes (Warnecke *et al*., 2004; Boucher *et al*., 2006). However, the abundance of organisms in lakes results from the balance between production of new biomass and loss by mortality. Several phenotypic features of aquatic bacteria have been interpreted as adaptations to escape protistan grazing pressure, and have been cited to explain the dominance of certain bacterial taxa (Pernthaler, 2005). Protection from protistan grazing could, moreover, be mediated by properties intrinsic to the bacterial cell wall. Gram-positive bacteria are consumed by protists at significantly lower rates than Gram-negative strains (Iriberri *et al*., 1994). Thus, *Actinobacteria* may benefit from a relatively fast growth rate and limited vulnerability to protistan grazing when cohabiting with other ecologically important bacterial groups in environments characterized by strong grazing pressure (Jezbera *et al*., 2005). In the same way, the large abundance of genes involved in cell motility in *Betaproteobacteria* could also decrease their vulnerability to predation by protists (Pernthaler, 2005) and thus contribute to their dominant occurrence in lake microbial communities.

The comparisons of metagenomes show the overrepresentation of METAPROC in some metabolic pathways related to xenobiotic degradation, especially benzene, toluene, ethylbenzene and xylene according to KEGG classification. The pollutant content of Bourget Lake is not known yet. However, monitoring programs throughout North America and Europe have demonstrated the widespread presence of pesticides in various freshwater bodies (Chèvre *et al*., 2006). The pollutants identified include priority pollutants and other well-known environmental pollutants such as polycyclic aromatic hydrocarbons and polychlorinated dibenzo-p-dioxins, but also other compounds that were either not previously considered environmental pollutants or that were not regulated against, such as substituted phenols, natural or synthetic estrogens and androgens (Brack *et al*., 2007). However, these genes could also be involved in plant organic matter degradation, composed in part by aromatic compounds (i.e. lignin), which could have originated from catchments. Moreover, exposure to sunlight may promote the maturation of humic substances. Assuming a certain degree of similarity, the metabolic pathway classified under xenobiotic degradation could also be associated with genes involved in the degradation of natural compounds such as fulvic acid. Only a fraction of total microbial diversity (i.e. the culturable fraction with metabolic potential) is known to be capable of metabolizing xenobiotics (Paul *et al*., 2005). Our results show that lacustrine degradation of aromatic compounds could be preferentially mediated by *Alphaproteobacteria* belonging to LD12 or GOBB-C201 clades. Although the main species reported as being involved in these processes were Pseudomonas or Burkholderia (Alfreider and Vogt, 2007), stable-isotope probing has also shown the importance of *Alphaproteobacteria* in xenobiotic degradation (Galvão *et al*., 2005), Thus, the *Proteobacteria* subclass can present new, potentially valuable metabolic pathways involved in the degradation of aromatic compounds, including xenobiotics, in freshwater ecosystems.

The *Bacteroidetes*, which generally constitute only a minor percentage of bacterial community composition in lake ecosystems (Boucher *et al*., 2006), could also degrade polymeric substrates of a labile and rather recalcitrant nature (Kirchman, 2002). This seems in agreement with our study that shows an overrepresentation in *Bacteroidetes* of genes involved in metabolism of carbohydrates like amino sugars. Our metagenomic analysis suggests also that *Bacteroidetes* were associated with the metabolism of cell wall component of both bacteria and microalgae. This finding seems to be in agreement with several reports evidencing a strong relationship between the high abundance of *Bacteroidetes* and the elimination of cyanobacterial blooms (van Hannen *et al*., 1999; Rashidan and Bird, 2001).

Conclusion

Our metagenomic approach studying the bacterial community of French Lac du Bourget evidenced the dominance of *Actinobacteria*, and to a lesser extent *Pro-*

teobacteria. Different functional capacities have been associated to the relative abundance of the main phyla. Furthermore, these functional capacities were able to differentiate the bacterial community of our freshwater ecosystem from the bacterial communities of marine environments. However, this study also suggested that metagenomic approaches still cannot explain species assemblage in complex communities. This is partly due to the fact that the species and function-based joint annotation of numerous sequences cannot be achieved due to the small number of genomes and sequences (apart from 16S rRNA gene) of environmental bacteria available in gene databases. Consequently, future research will need to be directed towards accumulating new sequences for the dominant species via the complete sequencing of fosmid inserts or by cultivating these species in order to resolve their complete genome.

Experimental procedures

Sample collection

Forty-five litres of freshwater was collected from the euphotic zone of the Lac du Bourget at a depth of 2 m, on 14 June 2006. The sample was sequentially filtered through a 30 μm pore-size filter and the filtrate was passed through a 1.2 μm carbon filter. Finally, this filtrate was concentrated on an ultrafiltration system (Amicon – Millipore – France) and the concentrate was filtered on 0.22 μm pore-size polycarbonate filters and stored at −20°C until further DNA extraction.

DNA extraction, library constructions and sequencing

Genomic DNA extraction was conducted according to Ausubel and colleagues (1987) as modified by Boucher and colleagues (2006). A fosmid gene library and a 16S rRNA gene library were constructed from the same DNA obtained from the 0.2–1.2 μm plankton fraction. Sequencing reactions were performed by GATC (http://www.gatc-biotech.com). The fosmid library was built using the pCC1FOS vector (Epicentre Biotechnologies – Tebu-bio, France) following the manufacturer's instructions. A total of 7746 fosmid clones were obtained with an average insert size of 30–40 kbp. All fosmid-end clones were sequenced, yielding 15 278 sequences with an average length of 778 bp (GenBank Accession: FI831481–FI846758). The 16S rRNA gene library was obtained after PCR amplifications of 30–60 ng of extracted DNA. Bacterial 16S rRNA gene fragments were amplified using primer combination 358 F (5′-CCT ACG GGA GGC AGC AG-3′) and 907R (5′-CCG TCA ATT CMT TTG AGT TT-3′), as described in Dorigo and colleagues (2006). Amplification products were cloned into the pGEMT vector following the manufacturer's instructions. Three hundred sequenced clones yielded 282 16S rRNA gene sequences (GenBank Accession: FJ447600–FJ447881).

Fosmid-end sequences analysis

For taxonomic binning, the fosmid-end sequences were queried against the NCBI nr database using BLASTx (*e*-value < 10^{-15}) and against the RDP (*e*-value < 10^{-3}) database using BLASTn. Top BLAST high scoring pairs (HSPs) were tabulated according to the NCBI taxonomic identifier. Furthermore, the taxonomic affiliation of BLASTx-generated results was validated if both the fosmid-end sequences gave the same result at bacterial class level. Taxonomic affiliation was also processed using the phylogenetic markers defined by Huson and colleagues (2007) and von Mering and colleagues (2007). Only bit-scores superior to 100 were kept to process this analysis (Huson *et al*., 2007).

To identify potential metabolic pathways in the METAPROC metagenome, sequences were compared with the KEGG database using BLASTx (*e*-value < 10^{-5}) by keeping the HSP with a bit-scores > 100. These data were cross-analysed with taxonomic composition to compare the main metabolic function associated with the main metabolic phyla found in this ecosystem.

The KEGG categories found in main bacterial groups present in this ecosystem were compared with the same categories in a set of complete genomes of prokaryotic organisms (188) selected in order to limit the bias due to organisms having more than one strain sequenced.

Comparative analysis of aquatic metagenomic libraries

To highlight the particularities of lacustrine ecosystems, these putative metabolic functions were compared against metagenomic studies from contrasting aquatic systems: the euphotic zone of the central North Pacific Ocean (10, 70 and 130 m) (DeLong *et al*., 2006), Newport harbor, Chesapeake Bay, Gulf of Mexico, and from 250 miles from Panama city (Rusch *et al*., 2007). These data were processed using the same bioinformatics procedures as for the fosmid-end analysis of the METAPROC metagenome. The sequences of these metagenomes were classified in KEGG categories by selecting the HSP from a BLASTx against KEGG database.

Phylogenetic analyses

The 16S rRNA sequences from the both libraries were aligned with complete sequences of an ARB database using the in-built automatic alignment tool (http://www.arb-home.de) (Ludwig *et al*., 2004). The resulting alignment was checked and corrected manually. Sequences were inserted into an optimized tree according to maximum parsimony criteria without allowing changes to existing tree topology. Putative chimeras were detected by the Bellerophon program (Huber *et al*., 2004). A distance matrix was generated and processed using DOTUR software (Schloss and Handelsman, 2005) to determine the different OTUs from the 282 16S rRNA gene sequences. Thus study set 98% as cut-off value.

Statistical methods

The COAs were performed on contingency tables in order to compare the main KEGG categories associated with: (i) the main bacterial groups (*Actinobacteria*, *Betaproteobacteria*, *Alphaproteobacteria* and *Bacteroidetes*) and (ii) the aquatic metagenomes. COA is an ordination method allowing the arrangement of two categorical variables (i.e. metabolic

pathways and bacterial groups or metagenomes) along gradient. The first two axis generated by this method maximize the correspondence between the metabolic pathways and bacterial groups or aquatic metagenomes. The relations between columns and rows of the matrices used were tested by a Chi-square test. This test was also used to compare the gene abundances in each KEGG category between METAPROC and the others aquatic metagenomes chosen. Statistical analysis was performed with R software using the ADE package for COA analysis (http://cran.r-project.org).

References

Abulencia, C.B., Wyborski, D.L., Garcia, J.A., Podar, M., Chen, W., Chang, S.H., *et al.* (2006) Environmental whole-genome amplification to access microbial populations in contaminated sediments. *Appl Environ Microbiol* **72:** 3291–3301.

Alfreider, A., and Vogt, C. (2007) Bacterial diversity and aerobic biodegradation potential in a BTEX-contaminated aquifer. *Water Air Soil Pollut* **183:** 415–426.

Allgaier, M., Brückner, S., Jaspers, E., and Grossart, H.P. (2007) Intra- and inter-lake variability of free-living and particle-associated *Actinobacteria* communities. *Environ Microbiol* **9:** 2728–2741.

Ausubel, F.M., Brent, R., Kingston, R.E., Moore, D.D., Smith, J.A., Sideman, J.D., and Struhl, K. (1987) *Current Protocols in Molecular Biology, Section 24*. New York, NY, USA: John Wiley and Sons.

Beja, O., Aravind, L., Koonin, E.V., Suzuki, M.T., Hadd, A., Nguyen, L.P., *et al.* (2000) Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. *Science* **289:** 1902–1906.

Boucher, D., Jardillier, L., and Debroas, D. (2006) Succession of bacterial community composition over two consecutive years in two aquatic systems: a natural lake and a lake-reservoir. *FEMS Microbiol Ecol* **55:** 79–97.

Brack, W., Klamer, H.J., López de Alda, M., and Barceló, D. (2007) Effect-directed analysis of key toxicants in European river basins a review. *Environ Sci Pollut Res Int* **14:** 30–38.

Button, D.K., and Robertson, B.R. (2001) Determination of DNA content of aquatic bacteria by flow cytometry. *Appl Environ Microbiol* **67:** 1636–1645.

Chèvre, N., Loepfe, C., Singer, H., Stamm, C., Fenner, K., and Escher, B.I. (2006) Including mixtures in the determination of water quality criteria for herbicides in surface water. *Environ Sci Technol* **40:** 426–435.

Cottrell, M.T., and Waidner, L.A., Yu, L., and Kirchman, D.L. (2005) Bacterial diversity of metagenomic and PCR libraries from the Delaware river. *Environ Microbiol* **7:** 1883–1895.

Daniel, R. (2006) The metagenomics of soil. *Nat Rev Microbiol* **3:** 470–478.

DeLong, E.F. (2006) Microbial community genomics in the ocean. *Nat Rev Microbiol* **3:** 459–469.

DeLong, E.F., Preston, C.M., Mincer, T., Rich, V., Hallam, S.J., Frigaard, N.U., *et al.* (2006) Community genomics among stratified microbial assemblages in the ocean′ interior. *Science* **311:** 496–503.

Dinsdale, E.A., Edwards, R.A., Hall, D., Angly, F., Breitbart, M., Brulc, J.M., *et al.* (2008) Functional metagenomic profiling of nine biomes. *Nature* **452:** 629–632.

Dorigo, U., Fontvieille, D., and Humbert, J.F. (2006) Spatial variability in the dynamic and the composition of the bacterioplankton community of the Lac du Bourget (France). *FEMS Microbiol Ecol* **58:** 109–119.

Edwards, R.A., and Rohwer, F. (2005) Viral metagenomics. *Nat Rev Microbiol* **3:** 504–510.

Frank, D.N., and Pace, N.R. (2008) Gastrointestinal microbiology enters the metagenomics era. *Curr Opin Gastroenterol* **24:** 4–10.

Galvão, T.C., Mohn, W.W., and de Lorenzo, V. (2005) Exploring the microbial biodegradation and biotransformation gene pool. *Trends Biotechnol* **23:** 497–506.

Glöckner, F.O., Zaichikov, E., Belkova, N., Denissova, L., Pernthaler, J., Pernthaler, A., and Amann, R. (2000) Comparative 16S rRNA analysis of lake bacterioplankton reveals globally distributed phylogenetic clusters including an abundant group of actinobacteria. *Appl Environ Microbiol* **66:** 5053–5065.

Hahn, M.W. (2006) The microbial diversity of inland waters. *Curr Opin Biotechnol* **17:** 256–261.

Handelsman, J. (2004) Metagenomics: application of genomics to uncultured microorganisms. *Microbiol Mol Biol Rev* **68:** 669–685.

van Hannen, E.J., Zwart, G., van Agterveld, M.P., Gons, H.J., Ebert, J., and Laanbroek, H.J. (1999) Changes in bacterial and eukaryotic community structure after mass lysis of filamentous cyanobacteria associated with viruses. *Appl Environ Microbiol* **65:** 795–801.

Huber, T., Faulkner, G., and Hugenholtz, P. (2004) Bellerophon: a program to detect chimeric sequences in multiple sequence alignments. *Bioinformatics* **20:** 2317–2319.

Huson, D.H., Auch, A.F., Qi, J., and Schuster, S.C. (2007) MEGAN analysis of metagenomic data. *Genome Res* **17:** 377–386.

Iriberri, J., Azua, I., Labirua-Iturburu, A., Artolozaga, I., and Barcina, I. (1994) Differential elimination of enteric bacteria by protists in a freshwater system. *J Appl Bacteriol* **77:** 476–483.

Jezbera, J., Hornák, K., and Simek, K. (2005) Food selection by bacterivorous protists: insight from the analysis of the food vacuole content by means of fluorescence in situ hybridization. *FEMS Microbiol Ecol* **52:** 351–363.

Kirchman, D.L. (2002) The ecology of Cytophaga-flavobacteria in aquatic environments. *FEMS Microbiol Ecol* **39:** 91–100.

Lefranc, M., Thénot, A., Lepère, C., and Debroas, D. (2005) Genetic diversity of small eukaryotes in lakes differing by their trophic status. *Appl Environ Microbiol* **71:** 5935–5942.

López-García, P., Brochier, C., Moreira, D., and Rodríguez-Valera, F. (2004) Comparative analysis of a genome fragment of an uncultivated mesopelagic crenarchaeote reveals multiple horizontal gene transfers. *Environ Microbiol* **6:** 19–34.

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.

McHardy, A.C., Martín, H.G., Tsirigos, A., Hugenholtz, P., and Rigoutsos, I. (2007) Accurate phylogenetic classifica-

tion of variable-length DNA fragments. *Nat Methods* **4:** 63–72.

Martín-Cuadrado, A.B., López-García, P., Alba, J.C., Moreira, D., Monticelli, L., Strittmatter, A., *et al.* (2007) Metagenomics of the deep Mediterranean, a warm bathypelagic habitat. *PLoS ONE* **2:** e914.

Maurin, N., Amblard, C., and Bourdier, G. (1997) Phytoplanktonic excretion and bacterial reassimilation in an oligomesotrophic lake: molecular weight fractionation. *J Plankton Res* **19:** 1045–1068.

von Mering, C., Hugenholtz, P., Raes, J., Tringe, S.G., Doerks, T., Jensen, L.J., *et al.* (2007) Quantitative phylogenetic assessment of microbial communities in diverse environments. *Science* **315:** 1126–1130.

Osburne, M.S., Grossman, T.H., August, P.R., and MacNeil, I.A. (2000) Tapping into microbial diversity for natural products drug discovery. *ASM News* **66:** 411–417.

Paul, D., Pandey, G., Pandey, J., and Jain, R.K. (2005) Accessing microbial diversity for bioremediation and environmental restoration. *Trends Biotechnol* **23:** 135–142.

Pernthaler, J. (2005) Predation on prokaryotes in the water column and its ecological implications. *Nat Rev Microbiol* **3:** 537–546.

Pope, P.B., and Patel, B.K. (2008) Metagenomic analysis of a freshwater toxic cyanobacteria bloom. *FEMS Microbiol Ecol* **64:** 9–27.

Rashidan, K.K., and Bird, D.F. (2001) Role of predatory bacteria in the termination of a cyanobacterial bloom. *Microb Ecol* **41:** 97–105.

Riaz, K., Elmerich, C., Moreira, D., Raffoux, A., Dessaux, Y., and Faure, D. (2008) A metagenomic analysis of soil bacteria extends the diversity of quorum-quenching lactonases. *Environ Microbiol* **10:** 560–570.

Richardot, M., Debroas, D., Thouvenot, A., Sargos, D., Berthon, J.L., and Dévaux, J. (2001) Influence of cladoceran grazing activity on dissolved organic matter, enzymatic hydrolysis and bacterial growth. *J Plankton Res* **23:** 1249–1261.

Rusch, D.B., Halpern, A.L., Sutton, G., Heidelberg, K.B., Williamson, S., Yooseph, S., *et al.* (2007) The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol* **5:** e77.

Schloss, P.D., and Handelsman, J. (2005) Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. *Appl Environ Microbiol* **71:** 1501–1506.

Simek, K., Hornák, K., Jezbera, J., Masín, M., Nedoma, J., Gasol, J.M., and Schauer, M. (2005) Influence of top-down and bottom-up manipulations on the R-BT065 subcluster of beta-proteobacteria, an abundant group in bacterioplankton of a freshwater reservoir. *Appl Environ Microbiol* **71:** 2381–2390.

Stein, J.L., Marsh, T.L., Wu, K.Y., Shizuya, H., and DeLong, E.F. (1996) Characterization of uncultivated prokaryotes: isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon. *J Bacteriol* **178:** 591–599.

Warnecke, F., Amann, R., and Pernthaler, J. (2004) Actinobacterial 16S rRNA genes from freshwater habitats cluster in four distinct lineages. *Environ Microbiol* **6:** 242–253.

Warnecke, F., Sommaruga, R., Seka, R., Hofer, J., and Pernthaler, J. (2005) Abundances, identity, and growth state of *Actinobacteria* in mountain lakes of different UV transparency. *Appl Environ Microbiol* **71:** 5551–5559.

Wetzel, R.G. (2002) Freshwater ecology: changes, requirements, and future demands. *Limnology* **1:** 3–9.

Zwart, G., Crump, B.C., Kamst-van Agterveld, M.P., Hagen, F., and Han, S. (2002) Typical freshwater bacteria: an analysis of available 16S rRNA gene sequences from plankton of lakes and rivers. *Aquat Microb Ecol* **28:** 141–155.

Zwart, G., van Hannen, E.J., Kamst-van Agterveld, M.P., Van der Gucht, K., Lindström, E.S., Van Wichelen, J., *et al.* (2003) Rapid screening for freshwater bacterial groups by using reverse line blot hybridization. *Appl Environ Microbiol* **69:** 5875–5883.