The ISME Journal (2007) 1, 632–642

www.nature.com/ismej

ORIGINAL ARTICLE

Soil eukaryotic functional diversity, a metatranscriptomic approach

Julie Bailly[1], Laurence Fraissinet-Tachet[1], Marie-Christine Verner[1], Jean-Claude Debaud[1], Marc Lemaire[2], Micheline Wésolowski-Louvel[2] and Roland Marmeisse[1]
[1]*Ecologie Microbienne, UMR CNRS, USC INRA, Université de Lyon, Université Lyon 1, Villeurbanne, France and* [2]*Microbiologie, Adaptation et Pathogénie, UMR CNRS/UCBL/INSA/BCS, Université de Lyon, Université Lyon I, Villeurbanne, France*

To appreciate the functional diversity of communities of soil eukaryotic micro-organisms we evaluated an experimental approach based on the construction and screening of a cDNA library using polyadenylated mRNA extracted from a forest soil. Such a library contains genes that are expressed by each of the different organisms forming the community and represents its metatranscriptome. The diversity of the organisms that contributed to this library was evaluated by sequencing a portion of the 18S rDNA gene amplified from either soil DNA or reverse-transcribed RNA. More than 70% of the sequences were from fungi and unicellular eukaryotes (protists) while the other most represented group was the metazoa. Calculation of richness estimators suggested that more than 180 species could be present in the soil samples studied. Sequencing of 119 cDNA identified genes with no homologues in databases (32%) and genes coding proteins involved in different biochemical and cellular processes. Surprisingly, the taxonomic distribution of the cDNA and of the 18S rDNA genes did not coincide, with a marked under-representation of the protists among the cDNA. Specific genes from such an environmental cDNA library could be isolated by expression in a heterologous microbial host, *Saccharomyces cerevisiae*. This is illustrated by the functional complementation of a histidine auxotrophic yeast mutant by two cDNA originating possibly from an ascomycete and a basidiomycete fungal species. Study of the metatranscriptome has the potential to uncover adaptations of whole microbial communities to local environmental conditions. It also gives access to an abundant source of genes of biotechnological interest.
The ISME Journal (2007) **1,** 632–642; doi:10.1038/ismej.2007.68; published onlline 20 September 2007
Subject Category: integrated genomics and post-genomics approaches in microbial ecology
Keywords: cDNA; eukaryotic micro-organisms; metatranscriptome; soil RNA

Introduction

Soil is a complex environment and a hotspot of microbial diversity with several thousands of different bacterial species in a single 1 g sample, a majority of them unknown and uncultivable on standard microbiological media (Rappe and Giovannoni, 2003). Soil also hosts numerous eukaryotic micro organisms that can represent a significant fraction of the microbial biomass in some ecosystems. Many prokaryotic species, and eukaryotic microbes cannot be easily isolated from complex environmental matrices and/or cannot be grown *in vitro*. To appreciate their true functional diversity and the activities they express *in situ* in the soil in response to different environmental constraints it is necessary to develop new experimental approaches adapted to these micro-organisms. One such approach developed in the recent years for prokaryotic micro-organisms is metagenomics.

The shotgun cloning of DNA extracted from complex environmental samples (soil, sediment, fresh or sea water) leads to the generation of metagenomic DNA libraries, which archive the genetic information present in the different genomes of the micro-organisms that colonize these environments (Rondon *et al.*, 2000). Such libraries have been generated for different purposes that encompass basic and applied research. The metagenome is considered as a treasure-trove for new enzymes (see, for example, Voget *et al.*, 2003; Yun *et al.*, 2004) and bioactive compounds whose biosynthetic pathways can be coded by full-length genes, operons or gene clusters present on single, long DNA inserts of the libraries (see, for example, Gillespie *et al.*, 2002;

Correspondence: R Marmeisse, Ecologie Microbienne, Université Lyon 1, Bâtiment Lwoff, 43 Boulevard du 11 Novembre 1918, 69622 Villeurbanne Cedex, France.
E-mail: roland.marmeisse@univ-lyon1.fr
Received 30 May 2007; revised 16 July 2007; accepted 16 July 2007; published onlline 20 September 2007

Courtois *et al.*, 2003; Schirmer *et al.*, 2005). Analysis of metagenomic libraries also offers the opportunity to get insights into genome organization and gene content of new bacterial species belonging to phyla with no cultivable representative (see, for example, Tyson *et al.*, 2004; Nesbø *et al.*, 2005; García Martín *et al.*, 2006; Strous *et al.*, 2006). Large-scale sequencing of metagenomic libraries from different environments and the comparison of their respective gene contents also revealed new or overlooked key physiological processes (see, for example, Venter *et al.*, 2004; Rusch *et al.*, 2007; Yooseph *et al.*, 2007) and illuminated adaptation of bacterial species and of entire communities to their respective environments (see for example, Tringe *et al.*, 2005; DeLong *et al.*, 2006).

None of these studies have included eukaryotic micro-organisms possibly because they were in a minority in the studied ecosystems or because they were physically excluded (by filtration or centrifugation on density gradients) from the biomass before DNA extraction. The study of the eukaryotic metagenome faces several specific problems. The genome size of a eukaryote can be several orders of magnitude higher than the genome size of a bacterium. Among free-living unicellular eukaryotes it can vary from, for example, 13.8 Mpb for the yeast *Schizosaccharomyces pombe* with an estimated number of protein coding genes of 4800 (Wood *et al.*, 2002) to 69 Mpb for the ciliate *Paramecium tetraurelia* (39 600 gene models; Aury *et al.*, 2006). As a consequence, it is unlikely that a workable metagenomic library based on genomic DNA can capture a significant fraction of the gene content of a eukaryotic microbial community. Furthermore, the frequent presence of introns and lack of conservation of motifs in promoter sequences prevent expression of genomic copies of eukaryotic protein-coding genes not only in a bacterial cell but also in most eukaryotic host. Finally, as there is no established protocols to easily separate eukaryotic cells from bacteria and from a complex environmental matrix such as soil, a DNA-based metagenomic library that include eukaryotic DNA would also necessarily include prokaryotic sequences.

For eukaryotes, the use of RNA extracted from environmental samples could circumvent these problems. Owing to their 3′ poly-A tails, eukaryotic mRNA can indeed be specifically isolated from a complex RNA mixture and converted into intronless cDNAs that can be cloned to generate environmental metatranscriptomic cDNA libraries, which are representative of the fraction of protein-coding genes expressed at the time of sampling. Grant *et al.* (2006) initiated this strategy using RNA extracted from hot springs water and activated sludge and Todaka *et al.* (2007) from a symbiotic protist community of termite gut. Both studies resulted in the identification of different eukaryotic protein-coding sequences.

The aim of the present study was to develop this approach for a Pine tree forest soil. The community of eukaryotic micro-organisms present in such an ecosystem is likely to be dominated by the extra-radical mycelia of ectomycorrhizal fungal symbionts that can contribute for up to one-third of the soil microbial biomass in temperate/boreal forests (Högberg and Högberg, 2002). To appreciate the taxonomic diversity of the micro-organisms that contributed to the extracted environmental RNA pool we amplified, cloned and sequenced from soil DNA and reverse-transcribed RNA a fragment of the eukaryotic 18S ribosomal gene. The cDNA library was constructed in *S. cerevisiae* expression plasmid to evaluate the possibility of screening environmental sequences by heterologous expression in this model eukaryotic micro-organism. This was successfully tested by complementing a histidine auxotrophic yeast mutant.

Materials and methods

Soil sampling

The sampling site was a monospecific *Pinus pinaster* forest planted on a stabilized coastal sand dune in SouthWest France (Truc Vert site, 44° 43′ *N*, 1° 15′ *W*; Gryta *et al.*, 1997; Guidot *et al.*, 2001). Understorey vegetation was composed of dispersed *Arbutus unedo* shrubs and seedlings of unidentified grasses on a bare floor devoid of litter. The substratum is a nutrient-poor, non-calcareous sand (*ca.* 98–99% sand, pH 5.5) with *ca.* 0.5% of organic matter. Sampling was performed on the 18 November 2004 (soil temperature, 11°C at 10 cm depth) by collecting 27, $10 \times 10 \times 20$ cm ($l \times L \times h$) blocks of soil underneath the fruit bodies of 25 different species of saprotrophic or symbiotic (ectomycorrhizal) basidiomycetes (Supplementary Table S1). Within 12 h the different samples were sieved (2 mm mesh size) to remove fine roots and large organic debris, and were then pooled to form a composite sample from which subsamples were taken and frozen at −70°C.

RNA extraction and purification

Frozen, 30 g soil samples were first ground for 3 min in a prechilled (−70°C) vibrating agate cup mill. To 1 g of milled soil were added glass beads and 750 μl of extraction buffer (50 mM NaCl, 10 mM Na_2EDTA, 1% SDS, 2 M guanidine isothiocyanate, 25 μl of β-mercaptoethanol, 50 mM Tris–HCl, pH 9.0). Cells were further broken by vortexing for 10 min at room temperature. After three successive extractions with a water-saturated phenol (pH 5.0) chloroform isoamyl alcohol (25:24:1, by vol.) mixture, nucleic acids were precipitated (30 min at –70°C) by adding to the aqueous solution 75 μl of 3 M Na-acetate (pH 5.2) and 2.1 ml of ethanol. After centrifugation, the nucleic acid pellet was resuspended in 40 μl of H_2O

and precipitated overnight at 4°C by adding 40 μl of 4 M LiCl. After centrifugation, the nucleic acid pellet was resuspended in 22 μl of H_2O. After saving an aliquot of genomic DNA, a 90 min DNA digestion at 37°C was performed by adding 14 U of DNase I and 4 μl of the appropriate buffer (MBI Fermentas). RNA was then precipitated by adding 40 μl of isopropanol, the pellet washed with 100 μl of 70% ethanol and resuspended in 50 μl of water. Low molecular weight contaminating molecules were eliminated by passing the samples through a Sephadex G50 spin column (GE Healthcare, Saclay, France). Polyadenylated eukaryotic mRNA was purified by affinity capture on paramagnetic beads coated with poly-dT oligonucleotides as described in the Dynabeads Oligo (dT) kit instruction manual (Dynal). Unbound rRNA was precipitated from the beads wash solutions using isopropanol and resuspended in water. mRNA was released from the beads using 20 μl of water.

RNA purity and concentration were estimated by spectrophotometry (NanoDrop ND-1000 Spectrophotometer). RNA quality was estimated by capillary electrophoresis by running samples on RNA 6000Nano lab on chips (Agilent).

Construction of the cDNA library

The cDNA library was constructed using the purified mRNA and the SMART cDNA construction kit following the manufacturer's recommendations (Clontech). Briefly, first strand cDNA synthesis was performed using a modified poly-dT oligonucleotide that binds the 3′-poly-A tail of the mRNA. The different cDNA were amplified by PCR (between 16 and 26 cycles) using a primer that binds the modified poly-dT oligonucleotide used for first-strand synthesis and a second primer that preferentially binds at the 3′-end of the single-stranded cDNAs. The resulting double-stranded cDNAs were bordered at their 5′-end by an *Sfi*1A and at their 3′-end by an *Sfi*1B restriction site. These sites were used for the directional cloning of the cDNA into the corresponding sites of the *Escherichia coli–Saccharomyces cerevisiae* pYESfi-URA3 shuttle plasmid. This plasmid was designed by inserting the *Sfi*1A and *Sfi*1B sites between the *Not*1 and *Bam*H1 cloning sites of pYES2-URA3 (Invitrogen, Cergy Pontoise, France). This plasmid possesses an ampicillin resistance gene for its maintenance in *E. coli* and a *URA3* gene for its maintenance in Ura3⁻ yeast strains. cDNAs were size-fractionated to remove molecules smaller than 400 bp and then cloned downstream of the strong glucose-repressible, galactose-inducible GAL1 promoter of pYESfi-URA3. Plasmids were introduced into electrocompetent *E. coli* DH10B′^{TM}T1^{R} cells (Invitrogen). The different bacterial colonies growing on a selective LB plus ampicillin solid medium were pooled and used for plasmid extraction (Qiafilter plasmid maxi kit, Qiagen, Courtaboeuf, France).

Transformation of S. cerevisiae

The haploid *S. cerevisiae* strain BY 4741 (*MAT*a *his3Δ1 leu2Δ0 lys2Δ0 met15Δ0 ura3Δ0*; from EUROSCARF) was used to screen the cDNA library. Yeast cultivation, transformation by the Li-acetate method and DNA extraction followed standard protocols (Rose *et al.*, 1990). For the selection of environmental *HIS3* genes, ura$^+$ yeast transformants were selected on a minimal medium containing 0.67% Yeast Nitrogen Base medium, 2% glucose and supplemented with auxotrophic requirements. Ura$^+$ transformants were then tested for their His phenotype by replica-plating on minimal medium containing 2% galactose, 2% glycerol, 2% ethanol, supplemented with auxotrophic requirements and missing histidine. The replicating plasmids present in the His$^+$ transformants were introduced by transformation in *E. coli* and were also reintroduced by transformation in BY 4741 to confirm that they really harbored *HIS3* complementing genes.

Amplification and cloning of the 18S rDNA

A *ca* 520 bp-long fragment located at the 5′-end of the eukaryotic 18S rDNA gene was amplified by PCR from both soil DNA and reverse-transcribed soil 18S rRNA using primers Euk1A (CTGGTTGATCCTGCCAG) and Euk516R (ACCAGACTTGCCCTCC) described by Diez *et al.* (2001). Soil-extracted rRNA (1 μg) was reverse-transcribed using primer Euk516R and 200 U of M-MuLV reverse transcriptase according to the manufacturer's instructions (MBI Fermentas). One tenth of the reverse-transcribed rRNA or *ca.* 100 ng of environmental DNA were used per PCR, which included in a final 25 μl volume, 200 nM of each primer, 1 mM $MgCl_2$, 200 μM of each dNTP, 0.25 mg ml^{-1} bovine serum albumin, 0.05% of W-1 detergent, 1 U of *Taq* DNA polymerase and the appropriate buffer (Invitrogen, Cergy Pontosie, France). After an initial denaturation of 1 min at 94°C, amplification was performed for 25 cycles comprising 45 s at 94°C, 1 min at 55°C and 1 min at 72°C. Amplification products of the expected size from seven different PCRs were pooled and isolated from an agarose gel (Nucleospin Extract II kit, Macherey-Nagel, Hoerdt, France), ligated in the plasmid pCR2.1 (TA cloning kit, Invitrogen) that was used to transform chemically competent DH5∝-T1 *E. coli* cells.

Sequencing and sequence analysis

Plasmids for sequencing were purified from *E. coli* cells using the NucleoSpin-plasmid kit (Macherey-Nagel). Sequencing was performed by Genoscreen (Lille, France) using universal primers T3 or T7 for the inserts cloned in pCR2.1 and primers T7 or YES12 (GCGTGAATGTAAGCGTGA) for the inserts cloned in pYESfi-URA3. Nucleotide sequences were deposited in the EMBL/GeneBank/DDJB databases under accession numbers AM409570 to AM409635

for the partial 18S rDNA sequences amplified from DNA; AM409518 to AM409569 for the partial 18S rDNA sequences amplified from reverse-transcribed RNA; AM409636 to AM409755 for the randomly selected cDNA clones and AM409756, AM409757 for the two *HIS3* genes.

Sequences were manually corrected and edited. BLAST (blastn or blastx) searches (Altshul *et al.*, 1997) were performed against various sequence databases at NCBI (http://www.ncbi.nlm.nih.gov/). 18S rDNA sequences were screened for putative chimeras using the Chimera Check program at the Ribosomal Database Project II website (http://rdp8.cme.msu.edu/). Potential chimeras were further analysed by blasting separately the two dissimilar segments of the sequences against GenBank. Functional and taxonomic annotation of protein coding sequences was performed on the basis of the Blast searches by looking not only at the 'best hit' but at the different 'best hits' that can correspond to sequences from different taxonomic groups and also by considering different criteria: expect value, percents of identity and similarity, length of the alignment. Protein coding sequences were also sorted in different functional categories as defined in the eukaryotic clusters of orthologous groups (KOG) database using Kognitor (http://www.ncbi.nlm.nih.gov/COG/grace/kognitor.html).

Multiple alignments were performed using CLUSTALW (Thompson *et al.*, 1994) and edited using SeaView (Galtier *et al.*, 1996). Phylogenetic trees were computed and drawn using the Phylo_win (Galtier *et al.*, 1996).

Data analysis

Rarefaction curves for the 18S rDNA sequences were computed using S. Holland's Analytical Rarefaction version 1.3 software (http://www.uga.edu/strata/software/). The abundance-based richness estimators S_{Chao1} and S_{ACE} were computed for different subsamples of different sizes drawn from the entire 18S rDNA data set as described by Kemp and Aller (2004) and as implemented at http://www.aslo.org/lomethods/free/2004/0114a.html.

Results

Soil nucleic acid extraction

The protocol optimized for this study allowed the simultaneous recovery of DNA and RNA from the studied sandy forest soil. Extractions were performed on 300 1-gram soil samples and a total of 107 μg of environmental RNA was recovered (yield of *ca.* 0.36 μg RNA/g of soil). Capillary electrophoresis of extracted total RNA separated three major peaks (upper curve, Figure 1a). The first two peaks, that eluted at 41.5 s and 43 s, comigrated with respectively a prokaryotic (*E. coli*) 16S rRNA and a eukaryotic (from the fungus *Hebeloma cylindrosporum*) 18S rRNA (data not shown). A third broader peak (44–48 s) corresponded to the pro- and eukaryotic rRNA large subunits. Interspecific size variations in the lengths of the rRNA large subunits could account for this lack of resolution. As reflected by the respective heights of the 16S and

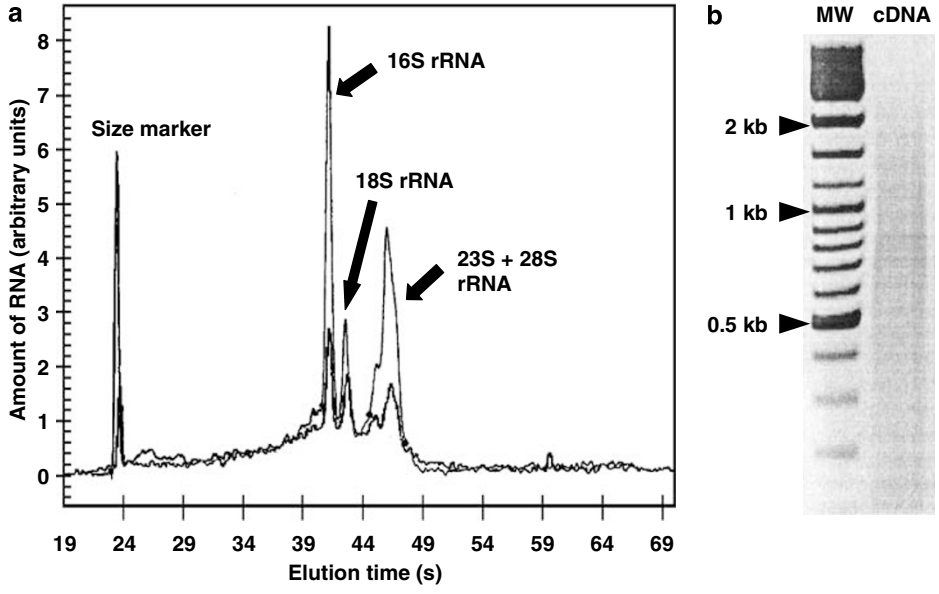

Figure 1 Soil RNA extraction and conversion into cDNAs. (**a**) Capillary electrophoresis profiles of total RNA (upper line) extracted from the Truc Vert forest soil and of RNA obtained after affinity capture on oligo-dT coated magnetic beads (lower, boldest line). The three major peaks comigrate with, from left to right, the prokaryotic 16S, the eukaryotic 18S and the pro- and eukaryotic large ribosomal RNA. Purification on oligo-dT beads lowers the proportion of rRNA relative to the baseline that represents the mRNA. (**b**) The double-stranded cDNAs obtained by PCR amplification of the reverse-transcribed mRNAs range in size from *ca.* 100 bp to more than 2 kb.

18S rRNA peaks, the RNA pool was predicted to contain *ca.* 20–25% of eukaryotic RNA.

The extract was enriched in eukaryotic polyadenylated mRNA by affinity capture on beads coated with poly-dT from which *ca.* 670 ng of mRNA were recovered (that is 0.6% of the total extracted RNA). Capillary electrophoresis of an aliquot sample (Figure 1a, lower curve in bold) showed a strong decrease in the height of the rRNA peaks but, the level of the baseline, which corresponds to the different mRNA molecules, was not affected. The purified extract was used for cDNA synthesis whose sizes ranged from 100 bp to more than 2 kb (Figure 1b).

Taxonomic diversity and richness of the soil eukaryotic community

An environmental cDNA library is likely to contain genes expressed by a variety of different eukaryotes living in soil. To appreciate this diversity, using eukaryote-specific PCR primers, we amplified and sequenced a *ca.* 520-bp-long 5′ fragment of the 18S rRNA gene that covers the variable domains V1, V2 and V3 (Borneman and Hartin, 2000). Sequences were obtained by using the soil DNA as PCR template (86 sequences) and by using reverse-transcribed soil rRNA as template (87 sequences). To minimize potential PCR biases we used a low number (25) of PCR cycles and we pooled seven separate PCRs before cloning. Despite these precautions, 32% of the sequences were identified as potential chimeras and were discarded. The remaining sequences were distributed in the different eukaryotic taxonomic groups on the basis of their 'best Blastn hits' against the nr database of GenBank and phylogenetic analyses.

Fungal sequences predominated among sequences derived from either environmental DNA or RNA and seventy of these sequences were from basidiomycete species (Figure 2). The other two main taxonomic groups were the protists and the metazoa and the ratio of protists vs metazoa was higher (2.6 vs 1.4) for sequences derived from reverse-transcribed RNA than from DNA (Figure 2). The two plant sequences (AM409540 and AM409634) were both related to conifer sequences. Twelve sequences (Figure 2, unknown) could not be clearly affiliated to a known taxonomic group; they could be artefacts or correspond to new or poorly studied eukaryotic phyla. Plant sequences were excluded from the subsequent analyses.

For both protists and metazoan, we identified sequences related to organisms commonly found in soils. In contrast to fungi and metazoa, protists form an unnatural taxonomic group that groups together mostly unicellular eukaryotes belonging to different kingdoms whose boundaries are still debated (see, for example, Simpson and Roger, 2004; Moreira *et al.*, 2007). Among protists, the cercozoa (Rhizaria) dominated and the euglenozoa (Excavata) were more

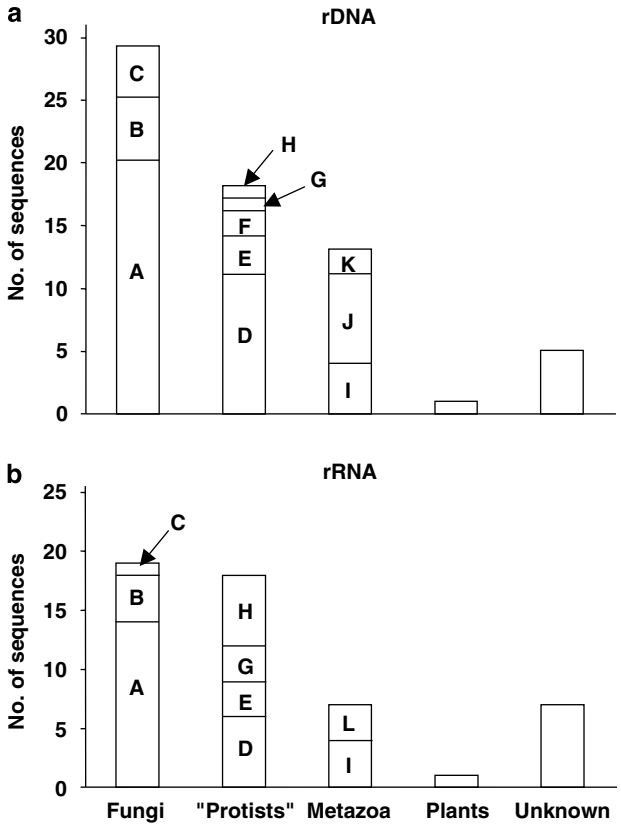

Figure 2 Taxonomic affiliation of the partial 18S rDNA sequences amplified from (**a**) soil DNA (66 sequences) and (**b**) reverse-transcribed soil RNA (52 sequences). A, Basidiomycetes; B, Ascomycetes; C, Zygo- and Chytidriomycetes; D, Cercozoa (Rhizaria); E, Amoebozoa; F, Apicomplexa (Alveolata); G, Ciliophora (Alveolata); H, Euglenozoa (Excavata); I, Arthropods and Tardigrads; J, Nematodes; K, Platyhelminthes; L, Annelids. Unknown sequences could not be clearly attributed to known eukaryotic taxonomic groups.

abundant among the sequences derived from environmental RNA (Figure 2). For the metazoa, the sequences were related to sequences of species that belong to taxonomic groups that contribute to the soil micro- and mesofauna. This was the case of acari and collembolans for the arthropods, of enchytrae for the annelids or of terricola worms for the platyhelminthes.

A phylogenetic analysis of the fungal sequences (Figure 3) shows that there is very little redundancy in our data set. If we used a cutoff value of 99% sequence identity to consider two sequences as belonging to the same phylotype, our 118 sequences identified 84 different phylotypes. Only one fungal phylotype, related to the homobasidiomycete *Cantharellus tubaeformis*, included more than two sequences and corresponded to the only phylotype that included sequences derived from both soil DNA and RNA (Figure 3). If we consider the distribution of sequences derived from either environmental DNA or RNA, they appear quite evenly distributed in the fungal tree (Figure 3). A slightly different

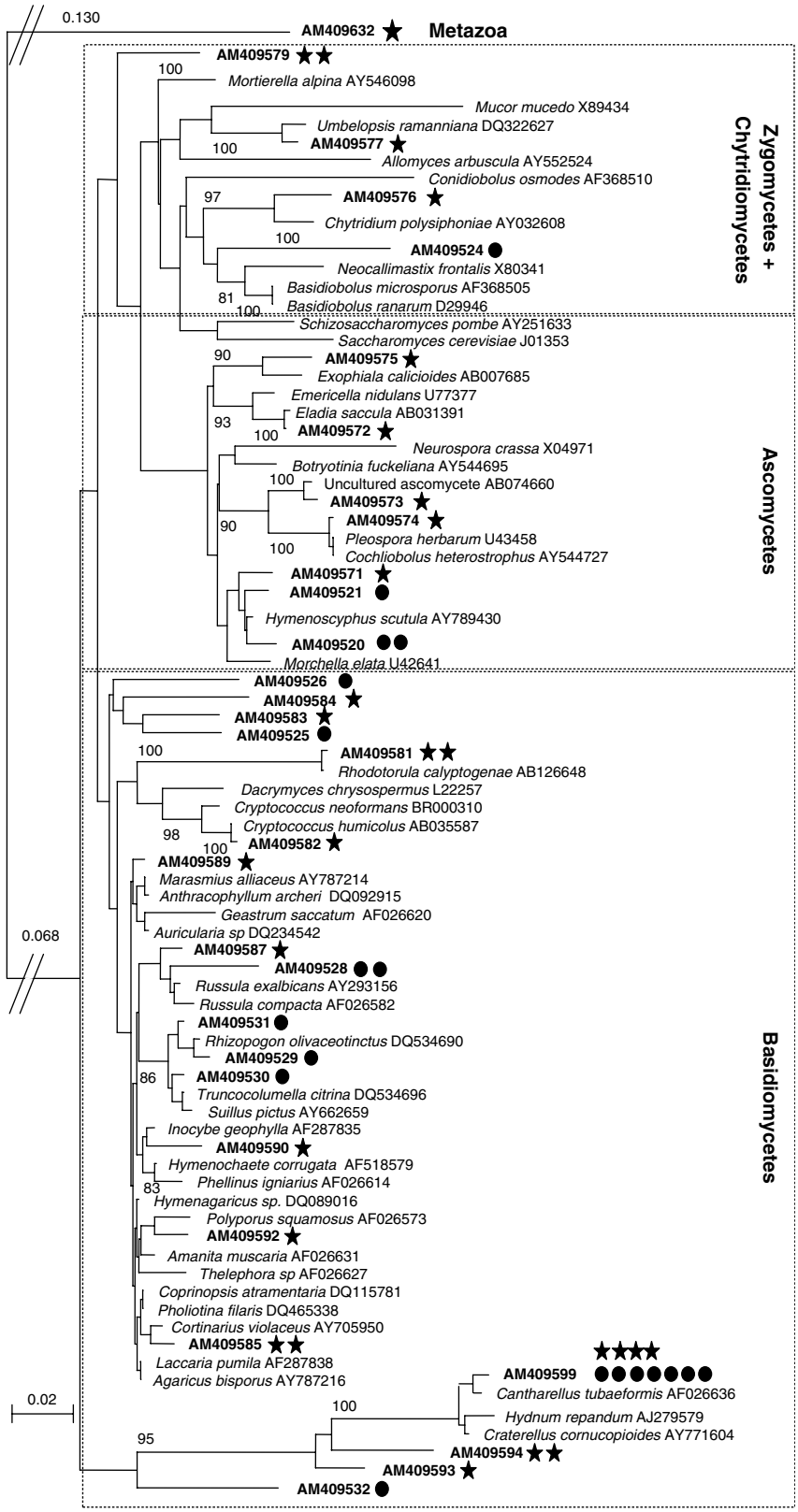

Figure 3 Neighbour-joining phylogenetic tree of the 30 different partial 18S rDNA fungal sequences retrieved from the *Pinus pinaster* forest soil studied. Stars indicate sequences amplified from soil DNA while black dots indicate those amplified from reverse-transcribed soil RNA. The numbers of stars and dots correspond to the number of times each sequence was identified. Bootstrap values (1000 replicates) above 80% are indicated; although most nodes are poorly supported, the representatives of the major fungal divisions group together with the notable exception of the fast-evolving cantharelloid sequences that branch outside. The tree was rooted using an environmental sequence related to Platyhelminth sequences (AM409632).

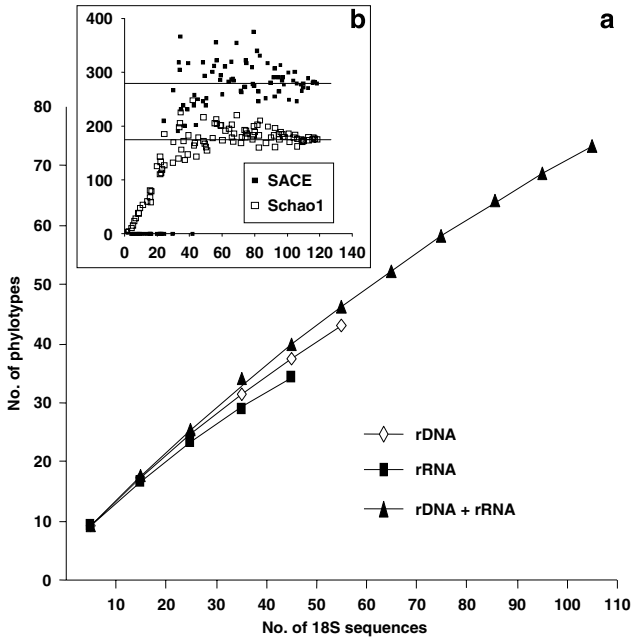

Figure 4 Evaluation of the 18S rDNA phylotype richness of the soil eukaryotic microbial community studied. (**a**) The phylotype accumulation curves drawn using either the sequences amplified from soil DNA, reverse transcribed RNA or both. (**b**) Prediction of the S_{ACE} and S_{Chao1} richness estimators for different subsamples of different sizes of the 18S library (ADNr and ARNr). For subsamples of more than *ca.* 60 sequences, values of the estimators tend to stabilize. Sequences from plants were excluded from these analyses.

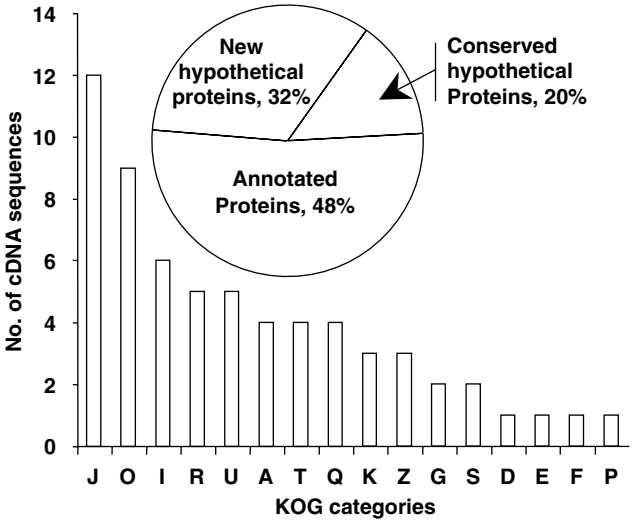

Figure 5 Functional classification of the putative proteins coded by the 119 sequenced cDNAs: 49% of the sequences fall into 15 different KOG functional categories. KOG categories are: A, RNA processing and modification; D, cell cycle control, cell division, chromosome partitioning; E, amino acid transport and metabolism, F, nucleotide transport and metabolism; G, carbohydrate transport and metabolism; I, lipid transport and metabolism; J, translation, ribosomal structure and biogenesis; K, transcription; O, post-translational modification, protein turnover, chaperones; P, inorganic ion transport and metabolism; Q, secondary metabolites biosynthesis, transport and catabolism; R, general function prediction only; S, function unknown; T, signal transduction mechanisms; U, intracellular trafficking, secretion and vesicular transport; Z, cytoskeleton. Sequences that are affiliated to two different KOG categories were counted in each of the categories.

situation was observed for protist and animal sequences (Figure 2). As already mentioned, euglenozoa were essentially identified among sequences derived from RNA (three phylotypes) and the arthropods were the only metazoa to be represented by sequences derived from both DNA and RNA (Figure 2). Additional sequences would be needed to test the significance of these observations.

The limited redundancy in the data set resulted in a sequence accumulation curve that clearly did not reach a plateau (Figure 4). Estimations of the abundance-based richness estimators S_{Chao1} and S_{ACE} were computed by resampling our data set according to Kemp and Aller (2004) (Figure 4, inset). For sample sizes above 60 sequences, S_{CHAO1} values levelled off to reach a stable asymptotic value of *ca.* 180 phylotypes while S_{ACE} estimates remained higher and did not converge towards a similar asymptotic value as already observed by Kemp and Aller (2004) for this estimator.

Construction and analysis of a cDNA library

PCR-amplified cDNAs larger than 400 bp were cloned directionally into the pYESfi yeast expression plasmid (see Materials and methods). The cDNAs were placed under the control of *GAL1* promoter. After transformation of *E. coli* cells, a library containing eight 10^6 independent plasmid clones with an insert was obtained.

One hundred and nineteen clones were randomly selected and their inserts sequenced from one or both ends. No ribosomal contamination was identified and all sequences seemed to be cloned in the proper orientation relative to the *S. cerevisiae GAL1* promoter of pYESfi. With the exception of two sequences that were represented each by two ESTs, all other sequences were single sequences. tBLASTX and BLASTN searches were conducted against respectively the nr and EST databases of GenBank (Figure 5). Thirty-two percent of the sequences did not give any positive hit (expected values above 1×10^{-8}) and were considered as coding new hypothetical proteins. Twenty percent were homologous to genes coding protein sequences of unknown functions (conserved hypothetical proteins). Among this category, we identified sequences that share between 80 and 99% identity at the nucleotide level to ESTs from organisms belonging to different taxonomic groups (fungi, plants and protists). The remaining 48% of the sequences corresponded to genes coding for proteins of known functions, some of which are listed in the functional KOG database (Figure 5 and Supplementary Table S1). Forty percent of these latter sequences could encode full-length functional proteins. Most sequences

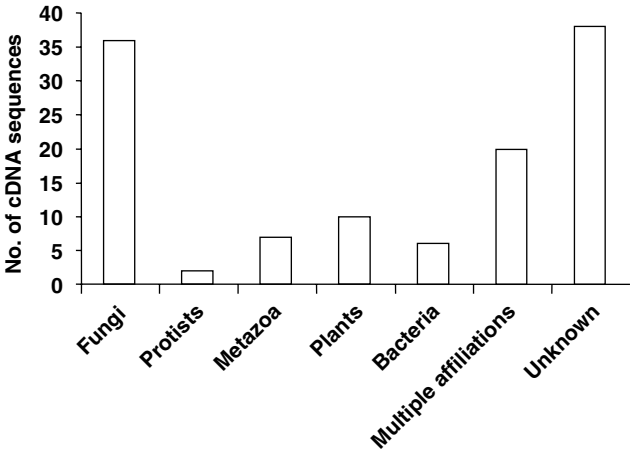

Figure 6 Taxonomic affiliation of the 119 sequenced cDNAs based on Blast analyses. Sequences that showed close similarities to sequences from different taxonomic groups (for example, from fungi and metazoa or from plants and bacteria…) were placed in the multiple affiliations category. Sequences that did not return any significant hit (usually E values greater than 10^{-8}) were classified as unknown.

corresponded to housekeeping genes involved in protein synthesis (KOG category J that includes ribosomal proteins) or post-translational modifications and protein turnover (KOG category O). However, we also identified genes that could be linked to biological soil processes such as the utilization of soil nutrients in the cases of fungal-related phosphate transporter (Accession no. AM409685) and glutamine synthetase (AM409660) or breakdown of phenolics/detoxication of xenobiotics in the case of a potentially full-length, fungal-related cytochrome *P*450 mono-oxygenase gene (AM409753).

Sequences were also tentatively attributed to taxonomic groups based on BLAST results (Figure 6 and Supplementary Table S2). Seventeen percent of the sequences presented comparable similarity values to different homologous sequences from organisms belonging to different taxonomic groups (for example, plants and bacteria for two sequences); these sequences were binned in a 'multiple affiliations' category (Figure 6). Six sequences could be of bacterial origin and the remaining 55 were attributed to a single eukaryotic taxonomic group (Figure 6). Interestingly, 5 the 10 plant sequences were between 89 and 99% identical at the nucleotide level to EST sequences from *Pinus pinaster* or *P. taeda*. When we compared taxon distribution for the ribosomal (Figure 2) and the cDNA sequences we observed (i) that the values are in good agreement for the fungi and (ii) a strong deficit in cDNA sequences attributed to the metazoa and the protists. This latter group was represented by only two cDNA sequences. One sequence (AM409746) encoded a putative full-length glutaredoxin homologous to *Dictyostelium* (Amoebozoa) but also bacterial sequences. The other (AM409697) was 90% identical at the nucleotide level to an EST and a genomic sequence of the filamentous oomycete (Heterokonta) plant pathogen *Phytophthora infestans*.

Functional complementation of a his3 yeast mutant
To validate this first environmental cDNA library we chose to complement a histidine auxotrophic phenotype of *S. cerevisiae* that has a mutation in the *HIS3* gene encoding an imidazole-glycerol phosphate dehydratase. Among *ca.* 300 000 independent yeast Ura^+ transformants we identified two transformants that grew on a medium supplemented with galactose without histidine. The complementing DNA of the replicating plasmids were sequenced. The deduced protein sequences coded by the inserts of the two plasmids aligned over their entire length to sequences of fungal imidazol-glycerol phosphate dehydratases. A phylogenetic analysis suggested that one of the genes could originate from a basidiomycete species and the other from an ascomycete one (Figure 7).

Discussion

This study demonstrates the feasibility of the metatranscriptomic approach, from soil-extracted RNA to the recovery of functional cDNAs expressed in a eukaryotic heterologous host cell (*S. cerevisiae*). In this respect, this approach developed for eukaryotic micro-organisms has the same potential as the traditional metagenomic approach for the cloning of single genes, coding enzymes of interest, from the environment (for example, Voget *et al.*, 2003; Yun *et al.*, 2004). In the context of natural product discovery from the environment the metatranscriptomic approach is however limited to the independent cloning of single genes and cannot be used to recover, at once, entire biosynthetic pathways despite the fact that genes coding for the synthesis of secondary metabolites often cluster in fungal genomes (Keller and Hohn, 1997).

Beyond its biotechnological applications, the metatranscriptome should reflect the pattern of gene expression of a microbial community in a complex environment and could thus be used to infer the physiological status of the corresponding community and to identify the environmental variables that have a major impact on gene expression *in situ*. It is therefore necessary to 'freeze' this pattern of gene expression soon after sampling. In our case, we choose to sieve the soil before storage at −70°C to exclude the macrofauna and the plant root biomass from the sample. As illustrated by the taxonomic distribution of the 18S rDNA sequences, this step was indeed effective in excluding plant biomass but this analysis also revealed that the soil metatranscriptome cannot be technically limited to eukaryotic micro-organisms *stricto sensu* and includes RNAs from the microfauna.

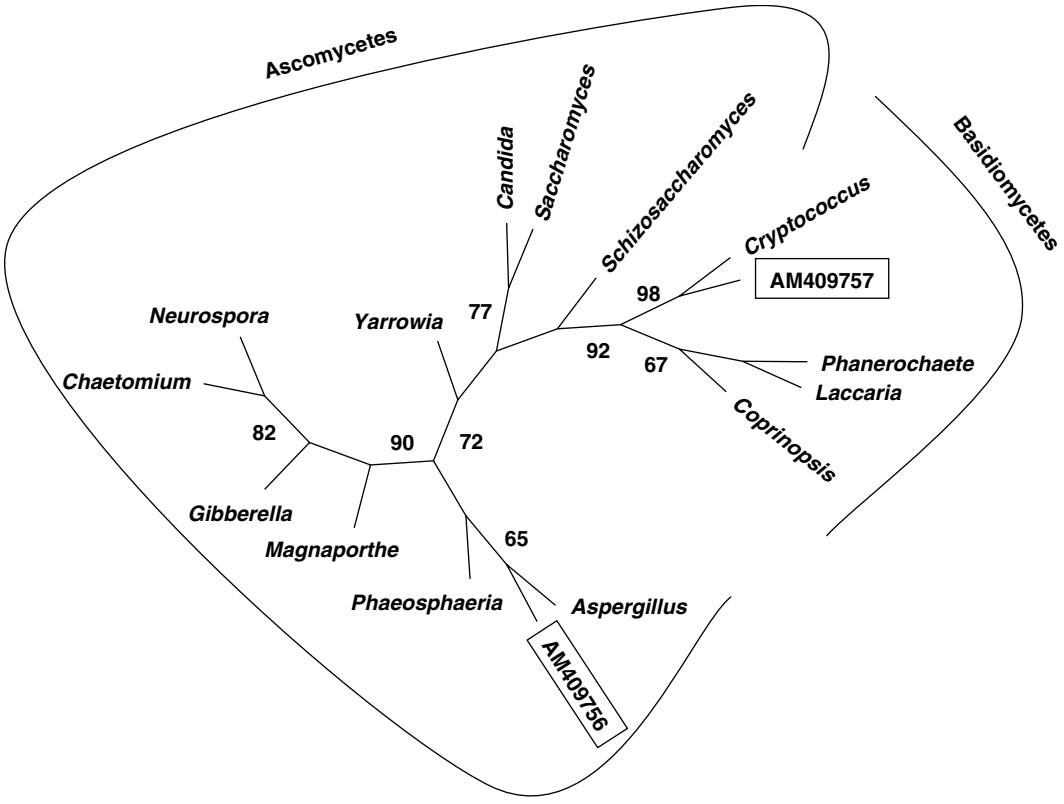

Figure 7 Unrooted most parsimonious phylogenetic tree of the two environmental imidazole-glycerol phosphate dehydratase amino acid (HIS3) sequences (AM409757 and AM409756), which complemented the yeast HIS3 auxotrophic mutation. Bootstrap values (1000 replicates) above 70% are given.

The survey of 18S sequences gives a global view of the biological diversity in the studied forest soil. The best-represented taxonomic group in terms of both number of sequences and number of phylotypes is the basidiomycete group known to account for a majority of ectomycorrhizal and saprobic fungal species in forest soils (see O'Brien *et al.*, 2005). Considering ectomycorrhizal species, their number commonly varies between *ca.* 50 >150 in a local, mature forest stand (see Horton and Bruns, 2001). Since the spatial distribution of many of these species is very patchy (Lilleskov *et al.*, 2004), the 27 soil cores that were collected to prepare the composite soil sample from which the nucleic acids were extracted, may not have captured the full diversity of the studied forest stand. The computed S_{Chao1} and S_{ACE} richness estimators therefore give us an indication of the total number of the species that may have contributed to the cDNA library but may underestimate the actual species richness of the forest stand.

An intriguing result concerns the large proportion of 18S sequences affiliated to the 'protists' whether the sequences were amplified from soil DNA (27% of the sequences) or from reverse-transcribed RNA (34%). Traditional studies, based for example on the estimation of cell volumes, tend to minimize soil protist biomass compared to the fungal one and ratios of protist over fungal biomass as low as 10^{-3} have been reported for forest soils (Ekelund *et al.*, 2001). The very few studies, including the present one, that have addressed soil eukaryotic diversity by amplification of ribosomal sequences using 'universal primers' give a far higher ratio of protist over fungal sequences that can even exceed one (O'Brien *et al.*, 2005, Supplementary Material of Tringe *et al.*, 2005). This raises a general problem in molecular microbial ecology of the existence of a correlation between the abundance of rDNA sequences and the relative biomass of the corresponding taxonomic groups. Abundance of sequences amplified by PCR reflects both the abundance of sequences added to the PCR and amplification biases. In eukaryotes, nuclear rDNA genes occur in unpredictable high copy number per haploid genome and the number of nuclei per unit of cytoplasm or of biomass varies also widely both between and within taxonomic groups. For sequences amplified from reverse-transcribed RNA it could nevertheless be assumed that some correlation should exist between sequence number and the volume of the corresponding 'biologically active' cytoplasm.

Data on the taxonomic distribution of ribosomal sequences contrast sharply with the similar set of data that concern cDNAs. With two possible exceptions, we failed to identify cDNA sequences that unambiguously originated from a 'protist'. With the exception of pathogenic species (*Cryptosporidium*,

Entamoeba, *Leishmania, Phytophthora*, *Plasmodium*, and so on) and of a few model species used in genetics or cell biology (*Paramecium*, *Dictyostelium* and so on), 'protists' are globally less studied at the molecular level compared to animals, plants and fungi. It is therefore tempting to hypothesize that a majority of the genes with no identified homologues originate from protists. However, this hypothesis is not fully supported by the fact that for several protist lineages (Ciliophora, Amoebozoa and so on) identified in the studied forest soil, the complete genome sequence of at least one species is now available (see http://www.genomesonline.org/gold.cgi). None of the 'unknown' environmental cDNA turned out to be homologous to sequences from fully sequenced 'protist' genomes. Similarly, none of the environmental cDNA encoding housekeeping eukaryotic proteins displayed strong similarities to 'protist' sequences. These discrepancies between 18S and cDNA taxonomic affiliations must be substantiated by additional data and illustrate one of the bioinformatics and practical problems that a large-scale analysis of the metatranscriptome may have to face.

Despite the limited number of sequences that were generated in this study, we identified sequences encoding enzymes that participate to major soil processes. Large scale sequencing of environmental cDNA sequences is likely to identify a significant number of such genes that could be used to infer the activities that are expressed *in situ*, in the soil, by the representatives of the different eukaryotic phyla.

We were also able to identify a few sequences that share a high level of identity at the nucleotide level (upto 99%) to sequences already deposited in databases. These sequences may belong to the same, or to a closely related, organism. This organism could therefore be specifically studied despite the complexity of the soil biota if it is naturally abundant in the studied ecosystem and if it has been used in a large-scale sequencing programme. In the present study, this is the case of *Pinus* trees despite the fact that most roots were eliminated before RNA extraction. The metatranscriptomic approach has therefore the potential to address ecophysiological questions, not only at the community level, but also at the level of a single, abundant species.

Acknowledgements

This work was supported by grants from the University Lyon 1 (BQR 2005), the French Ministry of Ecology and Sustainable Development (PNETOX programme) and the IFR 41. Sequencing of ribosomal genes was supported by an 'Etude et Appuis' grant of the Bureau des Ressources Génétiques. We thank Jacques Guinberteau and the other members of the INRA of Bordeaux for their help in field work and the determination of fungal species and the DTAMB of the University Lyon 1 for the use of the Agilent Bioanalyzer.

References

Altshul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, Miller W *et al.* (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl Acids Res* **25**: 3389–3402.

Aury J-M, Jaillon O, Duret L, Noel B, Jubin C, Porcer BM *et al.* (2006). Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. *Nature* **444**: 171–178.

Borneman J, Hartin RJ. (2000). PCR primers that amplify fungal rRNA genes from environmental samples. *Appl Environ Microbiol* **66**: 4356–4360.

Courtois S, Cappellano CM, Ball M, Francou FX, Normand P, Helynck G *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, Preston CM, Mincer T, Rich V, Hallam SJ, Frigaard NU *et al.* (2006). Community genomics among stratified microbial assemblages in the ocean's interior. *Science* **311**: 496–503.

Diez B, Pedros-Alio C, Massana R. (2001). Study of genetic diversity of eukaryotic picoplankton in different oceanic regions by small-subunit rRNA gene cloning and sequencing. *Appl Environ Microbiol* **67**: 2932–2941.

Ekelund F, Rønn R, Christensen S. (2001). Distribution with depth of protozoa, bacteria, and fungi in soil profiles from three Danish forest sites. *Soil Biol Biochem* **33**: 475–481.

Galtier N, Gouy M, Gautier C. (1996). SeaView and Phylo_win, two graphic tools for sequence alignment and molecular phylogeny. *Comput Applic Biosci* **12**: 543–548.

García Martín H, Ivanova N, Kunin V, Warnecke F, Barry KW, MacHardy AC *et al.* (2006). Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities. *Nat Biotechnol* **24**: 1263–1269.

Gillespie DE, Brady SF, Bettermann AD, Cianciotto NP, Liles MR, Rondon MR *et al.* (2002). Isolation of antibiotics turbomycin A and B from a metagenomic library of soil microbial DNA. *Appl Environ Microbiol* **68**: 4301–4306.

Grant S, Grant WD, Cowan DA, Jones BE, Ma Y, Ventosa A *et al.* (2006). Identification of eukaryotic open reading frames in metagenomic cDNA libraries made from environmental samples. *Appl Environ Microbiol* **72**: 135–143.

Gryta H, Debaud JC, Effosse A, Gay G, Marmeisse R. (1997). Fine-scale structure of populations of the ectomycorrhizal fungus *Hebeloma cylindrosporum* in coastal sand dune forest ecosystems. *Mol Ecol* **6**: 353–364.

Guidot A, Debaud JC, Marmeisse R. (2001). Correspondence between genet diversity and spatial distribution of above- and below-ground populations of the ectomycorrhizal fungus *Hebeloma cylindrosporum*. *Mol Ecol* **10**: 1121–1131.

Högberg MN, Högberg P. (2002). Extramatrical ectomycorrhizal mycelium contributes one third of microbial biomass and produces, together with associated roots, half of the dissolved organic carbon in a forest soil. *New Phytol* **154**: 791–795.

Horton TR, Bruns TD. (2001). The molecular revolution in ectomycorrhizal ecology: peeking into the black-box. *Mol Ecol* **10**: 1855–1871.

Keller NP, Hohn TM. (1997). Metabolic pathway gene clusters in filamentous fungi. *Fung Genet Biol* **21**: 17–29.

Kemp PF, Aller JY. (2004). Estimating prokaryotic diversity: when are 16S rDNA libraries large enough? *Limnol Oceanogr Methods* **2**: 114–125.

Lilleskov EA, Bruns TD, Horton TR, Taylor DL, Grogan P. (2004). Detection of forest stand-level spatial structure in ectomycorrhizal fungal communities. *FEMS Microbiol Ecol* **49**: 319–332.

Moreira D, von der Heyden S, Bass D, López-García P, Chao E, Cavalier-Smith T. (2007). Global eukaryote phylogeny: combined small- and large-subunit ribosomal DNA trees support monophyly of Rhizaria, Retaria and Excavata. *Mol Phylogenet Evol* **44**: 255–266.

Nesbø CL, Boucher Y, Dlutek M, Doolittle WF. (2005). Lateral gene transfer and phylogeneyic assignment of environmental fosmid clones. *Environ Microbiol* **7**: 2011–2026.

O'Brien HE, Parrent JL, Jackson JA, Montcalvo J-M, Vilgalys R. (2005). Fungal community analysis by large-scale sequencing of environmental samples. *Appl Environ Microbiol* **71**: 5544–5550.

Rappe MS, Giovannoni SJ. (2003). The uncultured microbial majority. *Ann Rev Microbiol* **57**: 369–394.

Rondon MR, August PR, Bettermann AD, Brady SF, Grossman TH, Liles MR *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.

Rose MD, Winston F, Hieter P. (1990). *Methods in Yeast Genetics: A Laboratory Course Manual*. Cold Spring Harbor Press: Cold Spring Harbor, NY.

Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S *et al.* (2007). The sorcerer II global sampling expedition: northwest atlantic through eastern tropical pacific. *PloS Biol* **5**: e77.

Schirmer A, Gadkari R, Reeves CD, Ibrahim F, DeLong EF, Hutchinson CR. (2005). Metagenomic analysis reveals diverse polyketide synthase gene clusters in microorganisms associated with the marine sponge *Discodermia dissoluta*. *Appl Environ Microbiol* **71**: 4840–4849.

Simpson AGB, Roger AJ. (2004). The real 'kingdoms' of eukaryotes. *Curr Biol* **44**: R693–R696.

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

Thompson JD, Higgins DG, Gibson TJ. (1994). CLUSTAL W: improving the sensivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucl Acids Res* **22**: 4673–4680.

Todaka N, Moriya S, Saita K, Hondo T, Kiuchi I, Takasu H *et al.* (2007). Environmental cDNA analysis of the genes involved in lignocellulose digestion in the symbiotic protist community of *Reticulitermes speratus*. *FEMS Microbiol Ecol* **59**: 592–599.

Tringe S, von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW *et al.* (2005). Comparative metagenomics of microbial communities. *Science* **308**: 554–557.

Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM *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, Halpern AL, Rusch D, Eisen JA *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.

Wood V, Gwilliam R, Rajandream M-A, Lyne M, Lyne R, Stewart A *et al.* (2002). The genome sequence of *Schizosaccharomyces pombe*. *Nature* **415**: 871–880.

Yooseph S, Sutton G, Rusch DB, Halpern AL, Williamson SJ, Remington K *et al.* (2007). The Sorcerer II global ocean sampling expedition: expanding the universe of protein families. *PLoS Biol* **5**: e16.

Yun J, Kang S, Park S, Yoon H, Kim MJ, Hew S *et al.* (2004). Characterization of a novel amylolytic enzyme encoded by a gene from a soil derived metagenomic library. *Appl Environ Microbiol* **70**: 7229–7235.

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