OPEN ACCESS Freely available online

PLoS one

Simultaneous Assessment of Soil Microbial Community Structure and Function through Analysis of the Meta-Transcriptome

Tim Urich[1,2], Anders Lanzén[3], Ji Qi[4], Daniel H. Huson[5], Christa Schleper[1,2]*, Stephan C. Schuster[4]*

1 Centre of Geobiology, Department of Biology, University of Bergen, Bergen, Norway, **2** Department of Genetics in Ecology, Vienna Ecology Center, University of Vienna, Vienna, Austria, **3** Computational Biology Unit, Bergen Center for Computational Science, University of Bergen, Bergen, Norway, **4** Center for Comparative Genomics and Bioinformatics, Pennsylvania State University, State College, Pennsylvania, United States of America, **5** Center for Bioinformatics, Tübingen University, Tübingen, Germany

Abstract

Background: Soil ecosystems harbor the most complex prokaryotic and eukaryotic microbial communities on Earth. Experimental approaches studying these systems usually focus on either the soil community's taxonomic structure or its functional characteristics. Many methods target DNA as marker molecule and use PCR for amplification.

Methodology/Principal Findings: Here we apply an RNA-centered meta-transcriptomic approach to simultaneously obtain information on both structure and function of a soil community. Total community RNA is random reversely transcribed into cDNA without any PCR or cloning step. Direct pyrosequencing produces large numbers of cDNA rRNA-tags; these are taxonomically profiled in a binning approach using the MEGAN software and two specifically compiled rRNA reference databases containing small and large subunit rRNA sequences. The pyrosequencing also produces mRNA-tags; these provide a sequence-based transcriptome of the community. One soil dataset of 258,411 RNA-tags of ~98 bp length contained 193,219 rRNA-tags with valid taxonomic information, together with 21,133 mRNA-tags. Quantitative information about the relative abundance of organisms from all three domains of life and from different trophic levels was obtained in a single experiment. Less frequent taxa, such as soil Crenarchaeota, were well represented in the data set. These were identified by more than 2,000 rRNA-tags; furthermore, their activity *in situ* was revealed through the presence of mRNA-tags specific for enzymes involved in ammonia oxidation and CO_2 fixation.

Conclusions/Significance: This approach could be widely applied in microbial ecology by efficiently linking community structure and function in a single experiment while avoiding biases inherent in other methods.

Citation: Urich T, Lanzén A, Qi J, Huson DH, Schleper C, et al. (2008) Simultaneous Assessment of Soil Microbial Community Structure and Function through Analysis of the Meta-Transcriptome. PLoS ONE 3(6): e2527. doi:10.1371/journal.pone.0002527

Editor: Naomi Ward, University of Wyoming, United States of America

Received April 29, 2008; **Accepted** May 12, 2008; **Published** June 25, 2008

Funding: TU was financially supported by an EMBO long-term fellowship. Part of this project was financed through funding of the University of Bergen given to CS. SCS is funded in part by the Gordon and Betty Moore Foundation.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: Christa.Schleper@univie.ac.at (CS); scs@bx.psu.edu (SS)

Introduction

Soils cover almost all of the terrestrial area on Earth and have an indispensable ecological function in the global cycles of carbon, nitrogen and sulfur. Due to their physico-chemical complexity with many micro-niches, they teem with bio-diversity, both phylogenetically and functionally. A single gram of soil has been estimated to contain thousands to millions of different bacterial, archaeal and eukaryotic species [1,2] interwoven in extremely complex food webs. Communities of soil microbes carry out a multitude of small-scale processes that underlie many environmentally important functions. However, the explicit functional and ecological roles of individual taxa remain uncertain because most microbes withstand laboratory cultivation [3,4]. Therefore the most basic questions in microbial ecology "who is out there?" and "what are they doing?" are still often unanswered for many environments and for many microbial taxa. Ideally, especially the second question requires simultaneously information about the identity of taxa within a community and about functional processes performed. While soils seem to harbor the most complex microbial communities, these considerations apply to many other environments as well, like e.g. oceans and sediments.

With metagenomic technologies new dimensions in the characterization of complex microbial communities have been reached. Large scale shotgun sequencing approaches allow the discovery of many novel genes found in the environments independent of cultivation efforts [5–8]. Sequencing of large genomic inserts that contain phylogenetic "anchors" allows a direct link to the microbial taxon. However, in almost all of the metagenomic studies, a separate accompanying molecular typing method-usually based on PCR-amplified 16S rRNA genes-is needed to characterize the gene discoveries in the context of the microbial phylogenetic diversity of the sample [7–9].

These PCR-based typing methods-though very powerful, in particular when combined with the novel pyro-sequencing technology-[10,11] have some well-known drawbacks: (1) bias is

introduced by primers and/or exponential amplification; (2) simultaneous quantitative assessment of all three domains of life is impossible and (3) persistence of free DNA can bias the measurement of community responses to environmental changes.

Furthermore, DNA-based metagenomic and diversity studies do not allow us to draw conclusions on the expression state of the genes and therefore the functional role of genes or organisms in the investigated environment remains uncertain. In analogy to postgenomic studies of cultivated organisms, a logical next step in the metagenomic area therefore includes meta-transcriptomic technology.

First attempts to study the transcription of genes in environmental samples have been performed. They involved specific purification steps to selectively enrich for mRNA of bacteria or eukarya by depleting ribosomal RNA or enriching for polyA-tailed mRNA, respectively [12,13]. A more large scale (pyro-)sequencing approach was recently adapted for use with bacterial and archaeal mRNA from an environmental marine sample [14]. It involved an *in vitro* amplification step to maintain small sample size and fast preparation.

In order to overcome some of the limitations of the approaches mentioned above, we explore here the possibility to analyse the total RNA pool of a community, as it is naturally enriched not only in functionally but also taxonomically relevant molecules, i.e. mRNA and rRNA, respectively.

This offers the opportunity to link community structure and function in a single experiment, and to reach beyond the community's genomic potential as assessed in DNA-based methods, towards its *in situ* activity. Furthermore, the use of total RNA avoids extensive cleaning or amplification steps of mRNA molecules, enabling fast preparation even from difficult samples, such as soils, which are notoriously enriched for humic acids and other substances inhibiting molecular biological applications.

We have published initial results of the approach earlier in the context of characterizing the transcriptional activity of particular organisms from soil [15]. Here we present and validate the "double-RNA" approach for in-depth characterization of a soil microbial community by studying mRNA and rRNA molecules simultaneously from the same sample. Our approach ensures (i) fast preparation of the RNA in the light of the short half life of mRNA molecules, it avoids (ii) biases introduced by PCR or cloning procedures and (iii) obtains qualitative and quantitative information simultaneously of genes from organisms of all three domains of life, the archaea, bacteria and eukarya. Since highly parallel sequencing was used for this approach, which produced many, but short reads (100 bp), we have set up an appropriate bioinformatic analysis pipeline that allows to reliably extract in-depth functional and taxonomic information from this dataset.

Results and Discussion

Community profiling based on SSU and LSU rRNA-tags

The sandy soil ecosystem studied here stems from a conservation area and is comparably poor in nutrients and of neutral pH. It has been the subject of a variety of studies. Many molecular data are therefore available which facilitate result interpretation [15–20]. We generated cDNA from the transcriptome of this soil community by a non-targeted approach applying random-hexamer primed reverse transcription on total RNA. The cDNA was subjected directly to pyrosequencing without any prior PCR or cloning steps (see materials and methods for details). Several potentially biasing steps were therefore avoided. One experiment resulted in 258,411 RNA-derived sequence tags of ~98 bp. We set up a two-step analysis process to identify rRNA- and mRNA-derived tags ("ribo-tags" and "mRNA-tags", respectively) for efficiently and reliably obtaining taxonomic and functional information (Fig. 1). In the first step, all RNA-tags were compared against a small subunit and a large subunit rRNA reference database (SSUrdb and LSUrdb, respectively) that we compiled from publicly available sources (SILVA[21], RDP-II[22], Genbank nucleotide; see materials and methods for details). These two databases contain sequences from all three domains of life (SSUrdb: 137,160 sequences; LSUrdb: 6,247 sequences; see Supplement Table S1 in Supporting Information, SI). This approach enables the first simultaneous taxonomic analysis of communities based on the two most commonly used taxonomic marker molecules. The output file was analyzed with the program MEGAN[23]. In the second stage, all unassigned RNA-tags were compared against the Genbank non-redundant protein database to identify mRNA-tags.

The validity of our taxonomic analysis was tested with two simulated datasets composed of 43 small or large subunit rRNAs from bacterial (32), archaeal (5) and eukaryotic (6) representatives (see Supplement Methods and Results S1 in SI for details). Both databases, though very different in size and sequence composition, produced a similarly high taxonomic resolution power not only at the taxonomic level of domain and phylum, but also at the level of order for most of the species tested (Tables S3 and S4 in SI give details). These results also showed that neither reference database introduced a major artificial "community" shift; rather, both correctly reflected the test set's "community" structure. In addition, none of the ribo-tags were taxonomically wrongly assigned. Even with a ribo-tag length of 100 bp the power of resolution sufficed for our approach (this has also been demonstrated by other simulations, e.g. [24,25]). Additionally, we simulated a more natural situation by removing all reference sequences similar to test rRNAs from the databases, according to similarity distributions of ribo-tags from our soil dataset against the

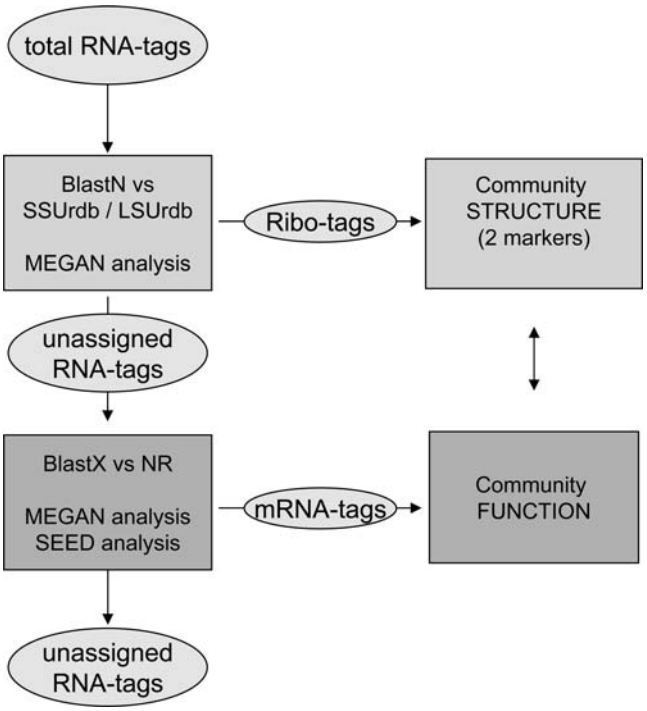

Figure 1. Overview about the double RNA analysis pipeline. Refer to text for details.
doi:10.1371/journal.pone.0002527.g001

reference databases (see Supplement Methods and Results S1 in SI for details). The identified thresholds reflected the median similarity of SSU and LSU ribo-tags (98% and 93% respectively; see Figures S1 and S2 in SI) to their best BLAST match in the reference databases. No significant decrease in taxonomic resolution was observed for the SSUrdb and no ribo-tag was incorrectly assigned (Table S5 and Figure S3 in SI). The much smaller LSUrdb assigned 5% of the ribo-tags incorrectly (Table S5 and Figure S4 in SI). When removing all reference sequences ≥86% similar to test sequences from the SSUrdb (to simulate a situation reflecting to the lowest decile similarity for SSU ribo-tags observed in the soil dataset), 5% of the ribo-tags were incorrectly assigned (Table S5 in SI). These results indicated a robust performance of the taxonomic binning approach. Encouragingly, the taxonomic resolution power will only improve further as (1) read lengths increase –250 bp are now possible, and 450 bp will soon be achievable; and (2) rRNA reference databases compile ever more sequences.

The soil dataset contained 193,219 RNA-tags which had significant BLAST hits against the rRNA reference databases (Table 1 and Fig 2). This dataset–which was generated without any PCR or cloning steps–was two to three orders of magnitude larger than traditional rDNA clone libraries, and was also far larger than recently published soil-derived pyro-sequencing datasets [26],

The SSUrdb and LSUrdb reported very similar relative community proportions on the domain level, with 10.3% and 13.3% of ribo-tags stemming from Eukarya, 87.2% and 83.8% from Bacteria and 1.5% and 1.4% from Archaea, respectively (Fig. 3a), which shows the reliability of our analytical approach. We further confirmed the experimental reproducibility by performing two independent cDNA syntheses from the same RNA pool (Fig. 3b, see materials and methods for details). In addition, abundances measured previously for some bacterial and archaeal groups in the same environment using various quantitative real-time PCR or metagenomic methodologies were in agreement with this study [15–19]. One percent of the SSU and LSU ribo-tags were consistently not sorted into one of the three domains, instead being classified as "cellular organisms". A closer inspection revealed that these represented short ribo-tags with comparably low taxonomic resolution power (not shown), rather than sequence tags of deeply rooting lineages or even of a "fourth domain of life". In principle, though, this non-targeted approach has the potential to identify lineages, which are currently not seen by primer-based PCR methodology, like e.g. the Nanoarchaeota [27].

Table 1. Statistics of soil community RNA-tag analysis.

	No. of RNA-tags	% of RNA-tags
Total RNA-tags	258,411	100
Total ribo-tags	193,219	74.8
SSU ribo-tags	99,061	38.3
LSU ribo-tags	94,165	36.4
mRNA-tags	21,133	8.2
unassigned RNA-tags	44,059	17.0

Note that the number of total ribo-tags differs from the sum of SSU and LSU ribo-tags. Seven ribo-tags contained regions with significant similarity against both databases (not shown). These putative chimeras were likely produced from SSU and LSU derived cDNA fragments during second-strand cDNA synthesis.

doi:10.1371/journal.pone.0002527.t001

Diversity and relative abundances of Bacteria and Archaea

With 165,246 ribo-tags (85.5% of total ribo-tags), bacteria were by far the most abundant domain in the sample; they were also extraordinarily diverse. Nineteen out of 24 validly described bacterial phyla were identified (Fig. 2 and Table S6 in SI) in addition to 20 candidate divisions. The latter are recently discovered, deep-branching, poorly characterized groups. The dominant bacterial phyla in our sample, such as Actinobacteria (47,513 ribo-tags) and Proteobacteria (38,477 ribo-tags) are typically found in high numbers in soil microbiotas. While both databases showed similar results for the five most abundant phyla (Fig. 4a), incongruencies arose, where LSUrdb sequences were underrepresented (e.g. Chloroflexi, Verrucomicrobia, Bacteroidetes), or completely missing (e.g. Nitrospirae, Gemmatimonadetes, various candidate divisions, see Table S6 in SI). Even rare phyla, such as Chlorobi and Dictyoglomi (ca. 0.02% of all bacterial tags) were reliably detected by both rRNA databases. It is also noteworthy that the analysis remains congruent at a more detailed taxonomic resolution (class and order, Figures 4b and 4c) for the taxonomic groups, which are comparably well covered in both databases (e.g. Proteobacteria, Firmicutes, Actinobacteria). Approximately 2% (1803) of the bacterial SSU ribo-tags were distributed over 20 candidate divisions (Table S6 in SI). The numbers of ribo-tags ranged from one (WS2 and KSB1 candidate phylum) to several hundred (SPAM, OP8, OP10, NKB19, VC2), or in other words, from indications of the presence of a taxon, to a high confidence including a wealth of sequence information.

In-depth taxonomic profiling enables the analysis of microbial communities from various functional perspectives. We demonstrate this for the process of nitrification, the conversion of ammonia to nitrate via nitrite. Both SSU-based and LSU-based analyses consistently identified the Crenarchaeal candidate division GroupI.1b as the predominant archaeal taxon in the soil sample (Table S7 and Figure S5 in SI); this is consistent with earlier studies of the same habitat [16] [15]. Members of this group were recently identified as important players in ammonia oxidation [15,20]. So it is perhaps not surprising that community-wide analysis of the different groups of bacteria and archaea harboring ammonia- and nitrite-oxidation capabilities (Fig. 4d) indicated that the groupI.1b Crenarchaeota were, at least from a numerical standpoint, the major ammonia-oxidizing group in our sample. The ribo-tag ratios between this group and ammonia-oxidizing bacteria were very similar to archaeal and bacterial amoA transcript ratios determined from the same cDNA preparation (12 vs. 16[15]). The subsequent nitrite-oxidation step appeared to be mainly performed by members of the Nitrospirae bacterial phylum. This result would have required seven independent quantitative real-time PCR assays to target the groups involved in nitrification when using traditional methods, whereas it represents only one of a variety of results obtained in this non-targeted approach. Additionally the respective functional groups are displayed in the context of the whole community.

As opposed to PCR-dependent approaches [10,11,26] that are confined to specific regions in the SSU rRNA molecule, our method allows the reassembly *in silico* of a full length "composite community" or consensus rRNA sequence for certain taxa. To illustrate this, the 1502 bp large SSU rRNA gene for groupI.1b of Crenarchaeota was assembled from 1105 ribo-tags. The resulting rRNA sequence differed by 3.1 % and 5.4 %, respectively, from the SSU rRNA sequences of the two fosmid clones, 29i4 and 54d9, isolated earlier from this habitat [17,20].

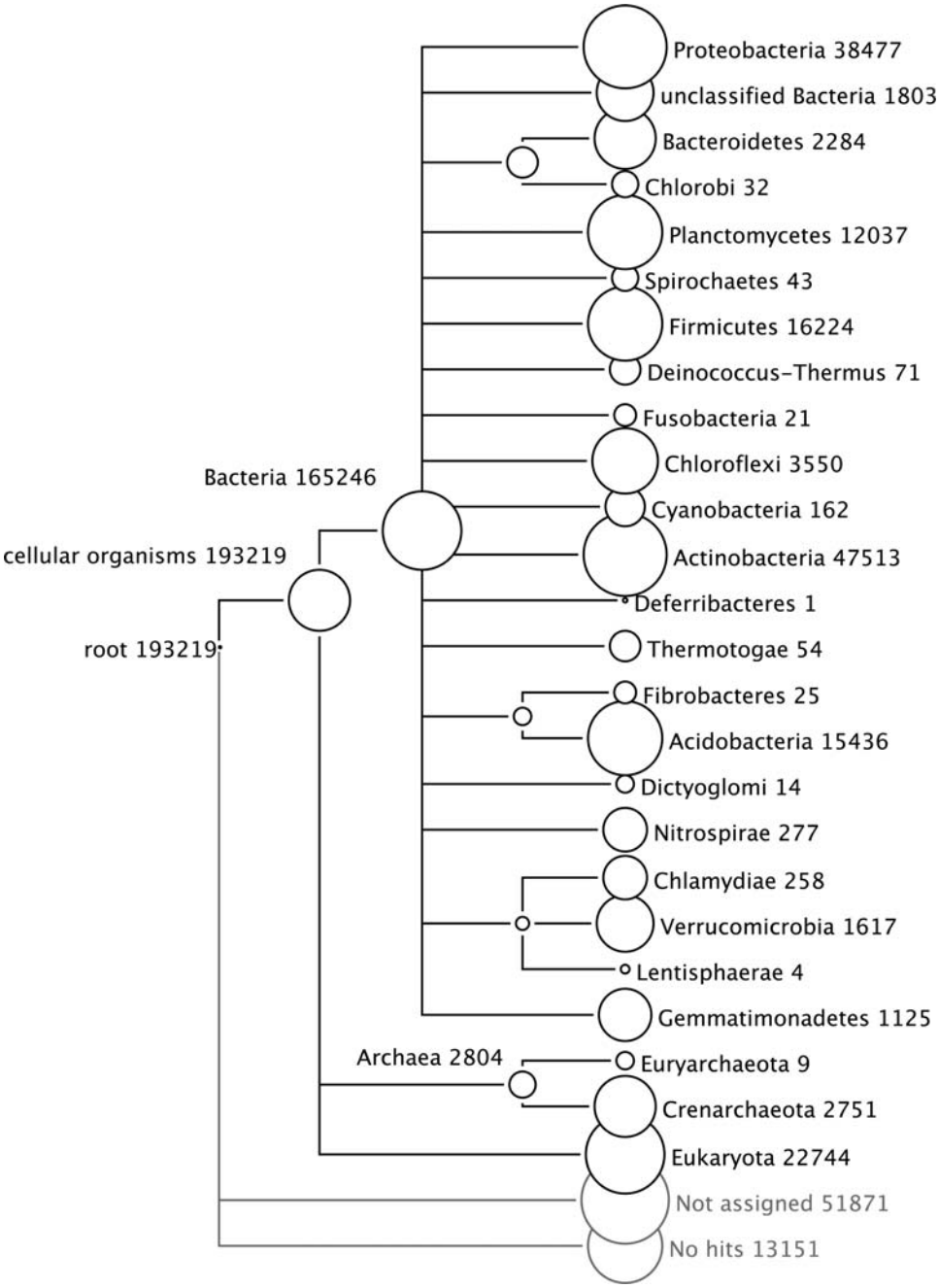

Figure 2. In-depth taxonomic community profiling based on SSU and LSU rRNA. MEGAN tree with the taxonomic affiliation of ribo-tags identified by BLASTN of all RNA-tags against our SSUrdb and LSUrdb according to the NCBI taxonomy. The numbers and sizes of the circles at the tree nodes refer to the ribo-tags affiliated with the respective taxon (absolute cutoff: BLASTN bit score 86, relative cutoff 10% of BLASTN top hit).
doi:10.1371/journal.pone.0002527.g002

Diversity and relative abundances of Eukaryotes

Although Eukaryotes are major players in soils and strongly influence the prokaryotic community structure, their diversity and abundance has received comparably little attention in molecular studies [28]. They account for ~11% of the ribo-tags (i.e. cellular biomass) in the soil community. Due to in-depth sampling, we obtained 22,740 eukaryotic ribo-tags, with more than 80% of the tags having a taxonomic resolution at least to the kingdom level. The numerically dominant kingdoms were the Fungi, known as the major destruents in soil, as well as plants (Viridiplantae) and Metazoa (Fig. 5a), with ~50%, ~20% and ~10% of the ribo-tags being consistently assigned by both reference databases. Within the Fungi, the phylum Ascomycota accounted for two thirds of the ribo-tags (Fig. 5b), followed by the phyla Glomeromycota and Basidiomycota.

As grazers, Protists play important roles in the soil food web by regulating microbial populations. Few LSU rRNA sequences are currently available for protists. Therefore, we estimated the Protist community composition using only the dataset of 1321 SSU ribo-tags (Fig. 5c). This is, to our knowledge, still the biggest molecular dataset of a protist community generated so far. It displayed the presence of slime molds (Mycetozoa) as the most abundant group, followed by

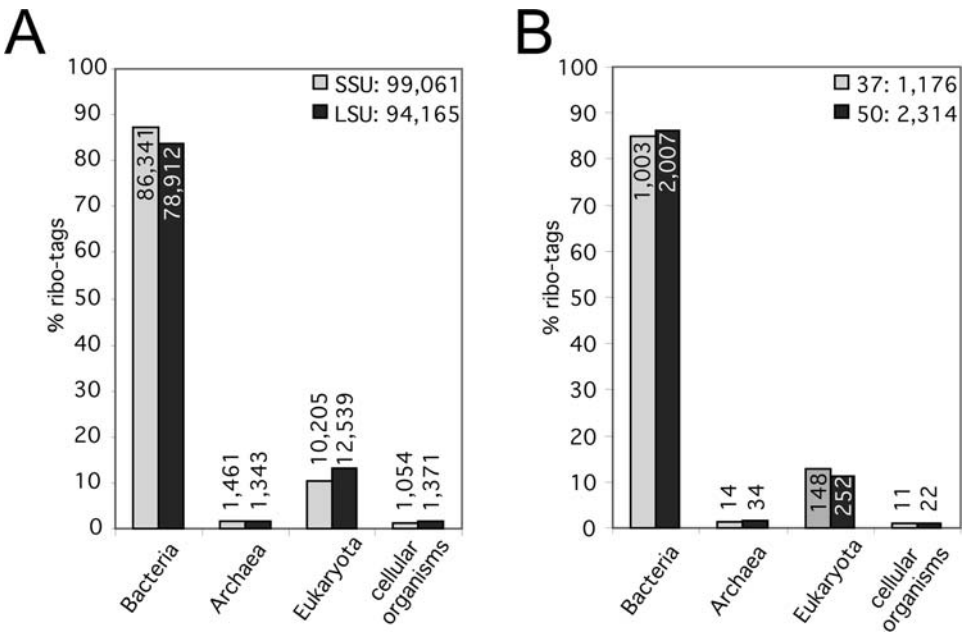

Figure 3. a, Relative ribo-tag distribution of the different cellular domains of life in the SSUrdb and LSUrdb. Absolute numbers are additionally given. Both reference databases report similar fractions, with bacteria-derived ribo-tags being most prominent. Note that approximately 1% of the ribo-tags are not affiliated to any domain by the reference databases. **b,** Comparison of relative ribo-tag distribution from two independent cDNA syntheses at 37°C (n = 4141 RNA-tags) and 50°C (n = 1985 RNA-tags) performed on the same RNA pool. Values represent the mean of SSU and LSU ribo-tags.
doi:10.1371/journal.pone.0002527.g003

Cercozoa, Plasmodiophorida and Alveolata. In addition, Lobosea, Acanthamoebidae, Heterolobosea, Euglenozoa and Stramenopiles were present with at least 40 ribo-tags in the sample.

The non-targeted and in-depth approach resulted in a dataset, which enabled a broad and holistic view onto a community and covered many of the trophic levels present. We are aware that the community profiles presented here are derived from the counting of ribosomal molecules, but not their genes, as is the case in DNA-based approaches. The ribosome content can vary considerably in the cells of an organism reflecting its physiological state; it also differs between taxa. Currently, few data are available; *E. coli* cells have been estimated to contain between 6,800 and 72,000 ribosomes, depending on the growth phase [29], unicellular eukaryotes likely have more ribosomes than microorganisms, and still higher is the ribosomal content of multi-cellular organisms like fungi, metazoans or plants. We consider the amount of ribo-tags as determined in this study to be a measure of a taxon's cellular biomass within a community [30,31].

Global functional analysis of putative mRNA-tags

The 65,192 RNA-tags that did not give a significant hit against the rRNA reference databases were aligned against the Genbank non-redundant protein database. Homologues to 21,133 RNA-tags were found, showing that a considerable amount of assignable mRNA had been reversely transcribed (Table 1). We subjected those presumable mRNA-tags (2.1 Mbp) to a global functional analysis using the SEED database[32] and compared the meta-transcriptomic data to metagenomic data from (1) the same soil habitat (4.3 Mbp[19]) and (2) a different farm soil (145 Mbp[7]). Overall, the DNA-based (metagenomic) functional repertoire in both soils was surprisingly similar, as judged from the relative distribution of the functional subsystems (Figure 6). This indicates that a generally similar "pool of functions" is present in both soil communities, which consequentially indicates that functional investigations of soil communities based on DNA might always give similar global patterns. In contrast to this, categories involved in RNA and protein metabolism (transcription, translation, protein folding and degradation) were significantly over-represented in the meta-transcriptome compared to both metagenomes (2.7 to 4.0-fold, see Figure 6 and Figure S6 in SI), as one would expect to see for active organisms. Further differences between the transcriptome and the metagenomes were related to carbohydrate metabolism: while transcripts of proteins involved in the aerobic degradation of mono-, di- and oligo-saccharides and amino-sugars seemed to be less frequent than suggested by the metagenomes (by 2-fold or more; Figure S7 in SI), transcripts for fermentation, degradation of sugar alcohols and CO_2-fixation were equally represented.

Metagenomic analyses frequently assign protein-encoding genes to taxonomic groups by comparing them against the content of sequenced genomes to derive a taxonomic community profile [5,6]. The simultaneously obtained rRNA and mRNA data provided us with the unique opportunity to validate these procedures. When we compared the ribo-tag profile with the community profile derived from taxonomic binning of the mRNA-tags using MEGAN, we observed a considerable shift for the five dominant bacterial phyla (Figure S8 in SI). These differences correlated strongly with the number of sequenced genomes of the different phyla (Fig. 7). This suggests that taxonomic binning solely based on protein encoding genes currently generates an artificial bias against groups with few sequenced genomes, and correspondingly over-represents phyla with many sequenced genomes. This problem, which is inherent to all metagenomic studies, will likely be overcome as more genome sequences of less represented phyla become available.

Probing the metabolism of an uncultured low-abundant group

1.7% of the mRNA-tags (ca. 36 kb) were identified as being of archaeal origin, similar to 1.5% of the ribo-tags (see Supplement

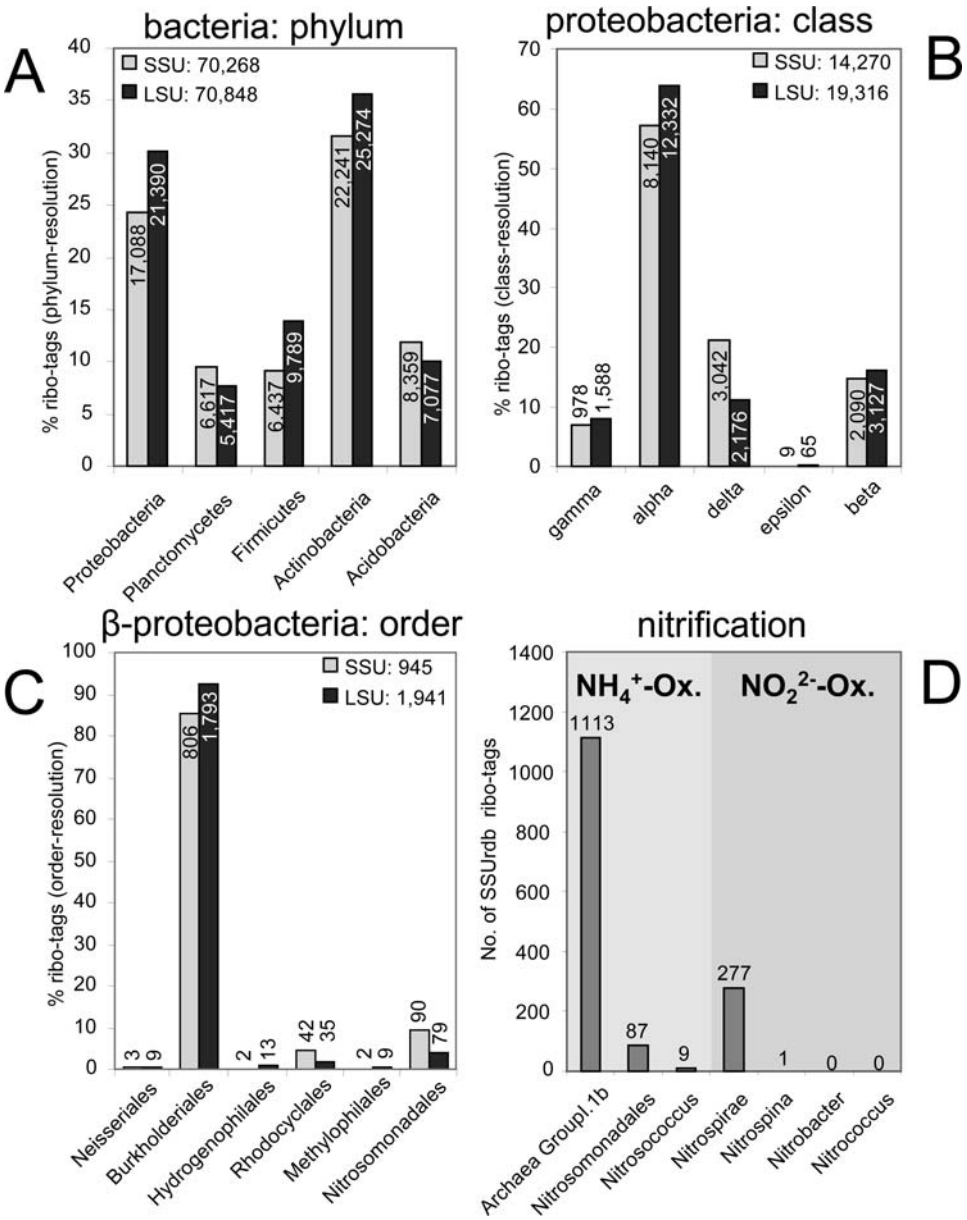

Figure 4. Relative distribution of bacterial SSU and LSU ribo-tags at different taxonomic resolutions and the abundance of prokaryotic groups involved in the nitrification process. a, fraction of phylum-resolution SSU and LSU ribo-tags affiliated to the five numerically dominant bacterial phyla (>5% of bacterial ribo-tags). Absolute numbers are additionally given. b, fraction of class-resolution SSU and LSU ribo-tags affiliated to the five proteobacterial classes. c, fraction of order-resolution SSU and LSU ribo-tags affiliated to the identified beta-proteobacterial orders. d, numbers of SSU ribo-tags from the seven archaeal and bacterial taxa involved in nitrification.
doi:10.1371/journal.pone.0002527.g004

Methods and Results S1 and Figure S5 in SI for details). These metatranscriptomic data allowed a first glimpse into the *in situ* activity of the yet uncultured soil Crenarchaeota from groupI.1b. Homologues to more than 80 mRNA-tags were functionally annotated in the databases. Besides transcripts of typical archaeal house-keeping gene products, those involved in ammonia oxdiation were predominant (Fig. 8); 13 mRNA-tags were derived from transcripts of the key metabolic enzyme ammonia monooxygenase (*amoA and amoC*). Furthermore, mRNA-tags of a putative copper-containing nitrite reductase (*nirk*) gene [20] indicated that this enzyme–as postulated for ammonia oxidizing bacteria–could be involved in the process of ammonia oxidation either under aerobic or anaerobic conditions[33]. These findings again hint for ammonia oxidation being the main energy metabolism in soil Crenarchaeota [34]. In addition, ten mRNA-tags could be related to the potential carbon metabolism. One mRNA-tag was derived from a homologue of methyl-malonyl-CoA mutase (MCM) and two from 4-hydroxybutyryl-CoA dehydratase (4-HBDH) homologues. These two gene products are, together with Acetyl-CoA/Propionyl-CoA carboxylase, diagnostic for a CO_2 fixation pathway recently characterised in hyperthermophilic Crenarchaeota and suggested for marine crenarchaeota [35]. This indicates that a similar pathway of CO_2 fixation might act in the soil crenarchaeota. Taken together, the metatranscriptomic data provide evidence for a chemolithoautotrophic lifestyle of this yet poorly characterised group.

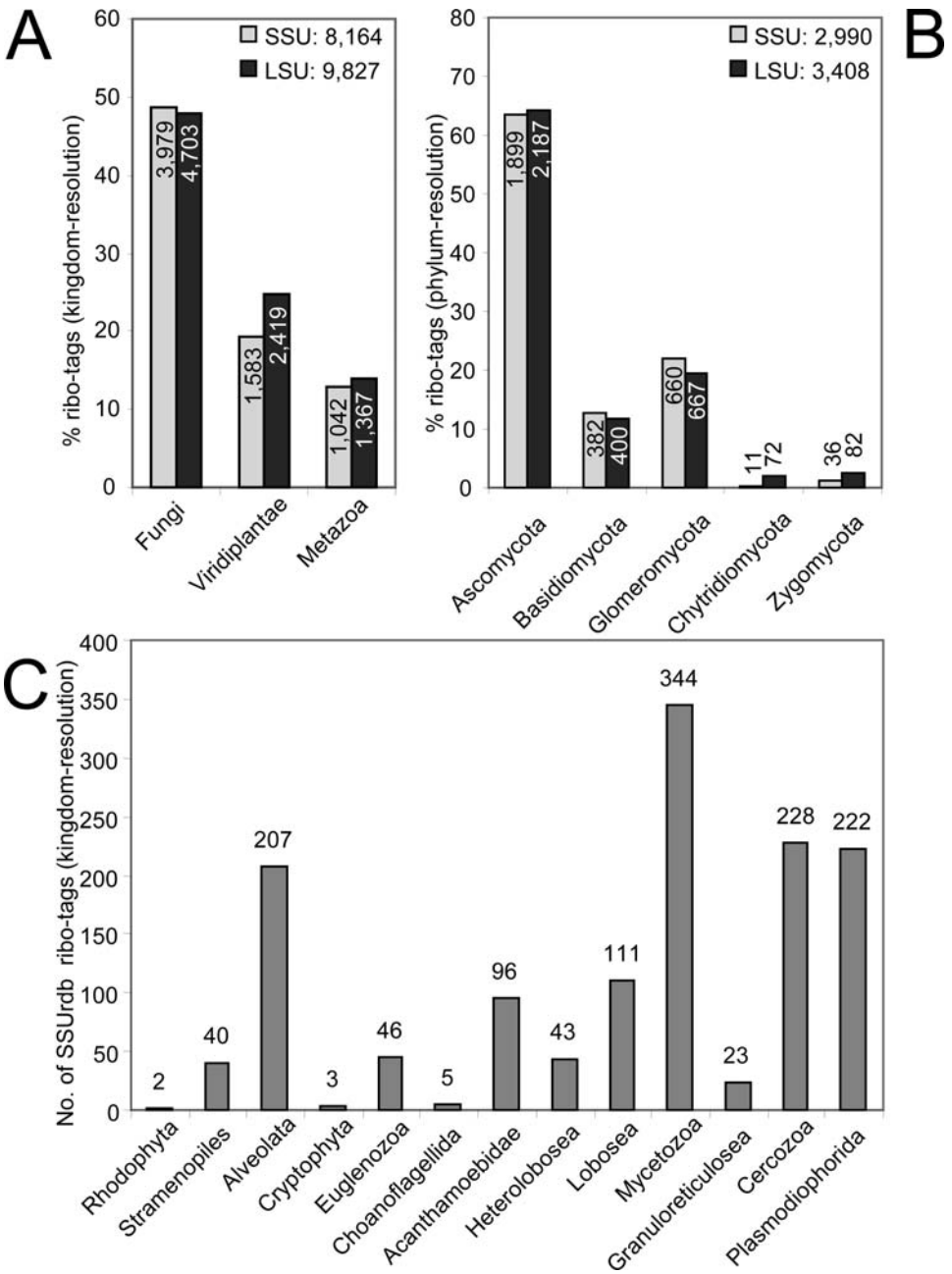

Figure 5. Relative and quantitative distribution of eukaryotic ribo-tags. a, fraction of kingdom-resolution SSU and LSU ribo-tags affiliated to the three numerically dominant eukaryotic kingdoms. **b,** fraction of phylum-resolution SSU and LSU ribo-tags affiliated to the five fungal phyla detected. The distribution was highly consistent in the SSUrdb and LSUrdb. **c,** numbers of SSU ribo-tags affiliated to the various protist kingdoms.
doi:10.1371/journal.pone.0002527.g005

With the non-targeted randomly primed reverse transcription approach, we can also assemble complete or nearly complete ''composite'' genes of specific lineages. A near full-length ''composite'' *amoC* transcript from the mRNA-tags of this analysis has been assembled (Figure S9 in SI). The deduced amino acid sequence covered 146 out of 189 positions (77%) of the archaeal sponge symbiont *Cenarchaeum symbiosum amoC* homolog [36] and had 88% identity to *C. symbiosum amoC*, similar to the 84% identity for *amoA* [15,20].

Conclusions

In conclusion, we have presented a rapid experimental and analytical approach that uses rRNA and mRNA to characterize microbial community structure and *in situ* function in-depth and simultaneously. This methodology will help (1) to identify microbial groups in complex communities; (2) to relate taxonomic groups to their ecological function (as demonstrated for the soil crenarchaeota); and (3) to efficiently monitor structural and functional community shifts caused by environmental changes. Furthermore, this approach enables for the first time the simultaneous quantitative assessment of the abundance of members of all three domains of life. The analytical power of this approach will continuously improve as sequence read lengths increase and as rRNA reference and genome databases continuously grow.

We believe that the analysis presented here could be used in parallel with mRNA enrichment procedures [13,14] to ensure an

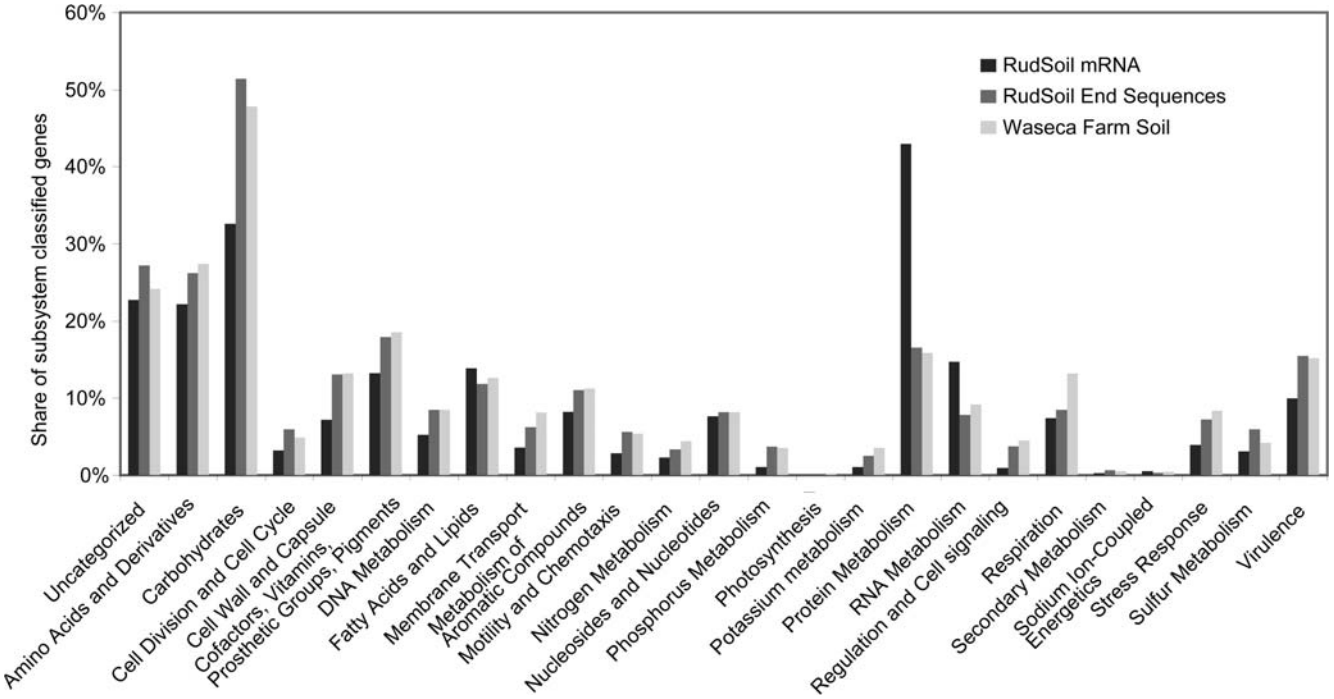

Figure 6. Global functional analysis of mRNA-tags, fosmid-derived end sequences from DNA of the same community [19] and shotgun-cloned DNA from a farm soil community [7]. All three datasets were subjected to automated analysis using the MG-RAST annotation procedure at the SEED (http://metagenomics.theseed.org). Percentages are expressed as the number of mRNA-tags assigned to a subsystem category, divided by the total number of mRNA-tags assigned to subsystems.
doi:10.1371/journal.pone.0002527.g006

efficient global analysis of the activity of naturally occurring assemblages, with one approach covering all trophic levels and domains as well as reasonable numbers of mRNA tags in one sequencing step (total RNA) and the other tool adding further in depth information from the enriched mRNA fraction.

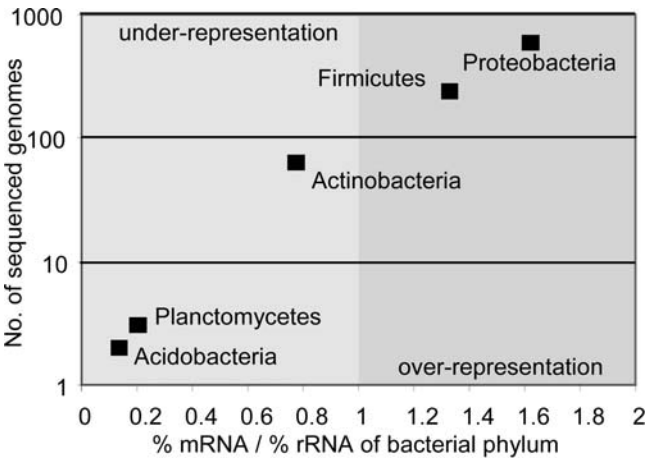

Figure 7. Logarithmic bi-variance plot of the number of publicly available genomes (as of September 2007) of the five numerically dominant bacterial phyla, as judged from the ribo-tags, versus the mRNA-tag based over- and under-representation, compared to the mean of SSU and LSU ribo-tag fraction. A ratio of one means that mRNA- and ribo-tags report the same fraction for the respective phylum.
doi:10.1371/journal.pone.0002527.g007

Materials and Methods

Site description and soil sample processing

The soil samples were obtained from a sandy lawn in the environmentally protected area "Am Rotböll" (Germany, Hessen, 49°55′34″N 8°37′21.6″E, [15]). The soil is nutrient-poor with a pH of 7.06, and with water extractible organic nitrogen (WEON) and carbon (WEOC) of 0.9 and 7.6 mg kg dry weight soil^{-1}, respectively. The top soil was sampled at a depth of 0–10 cm in January 2006 after a snow thawing period (soil temperature 5.5°C). Three replica within a distance of 1 m were withdrawn and kept in open plastic bags in the dark at 4°C for 40 hours. 100 g of the soil samples were sieved (2 mm mesh size) and equal amounts of the triplicates were subsequently mixed. Six grams of soil were then processed for RNA extraction.

cDNA synthesis

Nucleic acid extractions (RNA and DNA) were performed using a modification of the method of [37], and as briefly described in [15]. 0.5 ml of both CTAB buffer and phenol:chloroform:isoamylalcohol (25:24:1, pH 6.8) were added to a lysing matrix E tube (Q-Biogene) containing 0.5 g of soil. The cells were lysed in a FastPrep machine (Q-Biogene) at speed 5.5 for 30 seconds, followed by nucleic acid precipitation with PEG 8000. Total nucleic acids were subjected to Dnase treatment and the remaining total RNA was used as template for random hexamer-primed reverse transcription performed independently at 37°C and 50°C, respectively. The successful and complete hydrolysis of DNA was probed by archaeal and bacterial *amoA* specific quantitative real time PCR, which detected specific products in the cDNA samples after reverse transcription but not in parallel treated RNA samples, where no cDNA synthesis had taken place ([15], data not shown). The subsequently

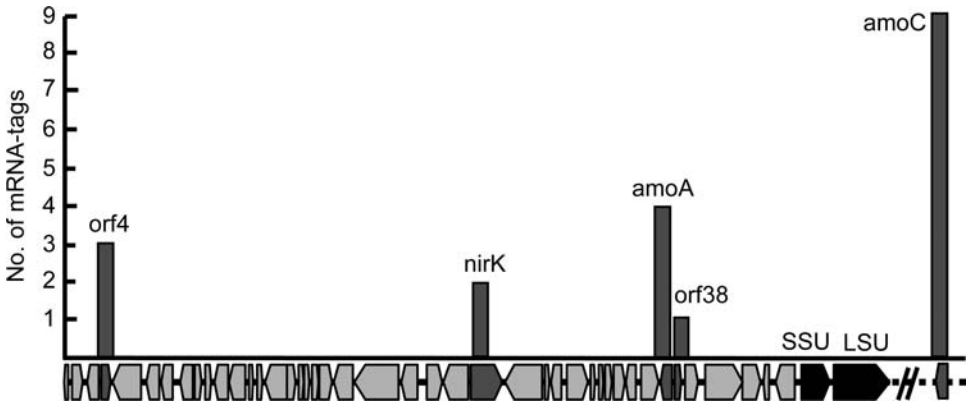

Figure 8. Abundance-dependent plot of presumably archaeal mRNA-tags onto the crenarchaeal fosmid clone 54d9 isolated from the same soil habitat (accession number: AJ627422). The x-axis represents the annotated open reading frames (orfs). Note that an *amoC* gene was not found on the fosmid and is therefore indicated as loosely affiliated.
doi:10.1371/journal.pone.0002527.g008

generated double-stranded (ds) cDNA ranged from approximately 100 to 1500 bp in length (not shown). Approximately one microgram of ds cDNA was generated.

Pyrosequencing

Sequencing was performed as described previously [15]. The 37°C and 50°C samples were kept separate during pre-sequencing processing. Independent test-sequencing was performed on a Roche GS20 sequencer (Roche Applied Sciences/454 Life Sciences, Branford, CT) for both samples, producing small subsets of RNA-tags (37°C = 4141 and 50°C: n = 1985). Both samples were subsequently pooled for one full run resulting in 314,041 sequences [15]. Re-analysis of the sequences resulted in 258,411 high-quality RNA-tags.

Reference Databases

Two rRNA reference sequence databases were constructed for community structure analysis using MEGAN [23]. The sequences therein were linked to taxa in the NCBI Taxonomy (as of June 2007), extended with 28 archaeal candidate divisions [38] (see Tables S1 and S2 in SI). The Small- and Large Subunit rRNA Reference Databases (hereafter SSUrdb and LSUrdb, respectively) were constructed by combining sequences from several public databases. The SSUrdb includes sequences from the RDP-II release 9.39 [22] (bacteria), the SILVA SSURef database release 89 [21] (eukaryotes and archaea), and sequences from the dataset described in [38] (archaea). The LSUrdb includes sequences from the Silva LSURef database release 90 [21]. Bacterial SSU rRNA sequences were retrieved from RDP-II [22]. More than 112,000 high quality sequences from isolates and uncultured strains were downloaded as FASTA files. The selection was made according to their taxonomic affiliation in the NCBI taxonomy. Sequences with low taxonomic resolution (e.g. "unclassified bacteria", "environmental samples") were mostly omitted until the "order" taxonomic level. Only in phyla with comparably few reference sequences and/or a relatively poorly refined taxonomy (e.g. the phyla Acidobacteria and Verrucomicrobia, and various candidate divisions), those sequences were included. All sequences were screened for vector contamination using the *cross_match* program from the Phred/Phrap package [39] against NCBI's UniVec database. All identified vector subsequences were removed from the database sequences. In some RDPII derived sequences, the entry contained more than the SSU rRNA gene, i.e. flanking regions and the LSU rRNA gene. In order to remove those, all sequences longer than 1,550 bp were aligned to the remainder of the sequences in the database (those within the expected length of an SSU rDNA gene), using BLASTN. Based on these alignments, any regions without significant similarity (E≤1e-6) were cropped, leaving only the SSU rRNA gene.

All eukaryotal and most archaeal SSU rRNA sequences were retrieved from the SILVA project [21]. The SSURef database version 89 was retrieved as an ARB file. 24,197 aligned eukaryotal sequences including 331 mitochondrial and 641 plastid sequences were analyzed and chosen for our database based on a correct taxonomic affiliation in the NCBI taxonomy and a comparatively high taxonomic resolution using the software package ARB [40]. Sequences were exported as FASTA files using a filter (including positions between 1000 and 43284 of the ARB alignment) for eliminating sequence information not belonging to the SSU rRNA gene.

The archaeal part of SSUrdb reference database contained 944 sequences from cultured strains, extracted from the SILVA SSURef database, again based on a correct and high resolution taxonomic affiliation in the NCBI taxonomy. Many archaeal lineages today consist exclusively of uncultured representatives. Those are not well resolved in the NCBI taxonomy, but deposited as "uncultured archaea" or "environmental samples", which prevents an informative taxonomic grouping of sequences belonging to those lineages. To overcome this, we have extended the NCBI taxonomy with 28 archaeal candidate divisions as described in [38] and exchanged the taxonomic affiliation of the effected sequences in the SSUrdb accordingly. Altogether 544 sequences are distributed over those groups (extracted from the ARB dataset used in [38]; see Table S2 in SI). The archaeal part therefore consists of 1,490 sequences and the whole SSUrdb of 135,160 sequences in total.

The large subunit reference database (LSUrdb) was generated from the SILVA database project. The LSURef database version 90 was retrieved as ARB file. The sequences therein were analyzed and chosen based on a correct taxonomic affiliation in the NCBI taxonomy reflected in the sequence alignment and a comparatively high taxonomic resolution using ARB. Sequences were exported as FASTA files using a filter (including positions between 66155 and 129011 of the ARB alignment) for eliminating sequence information not belonging to the LSU rRNA gene. Vector sequences were removed as described above. The LSUrdb consists of 6,247 sequences, of which 2,759 belong to Bacteria, 130 to Archaea (107 cultured and 23 uncultured) and 3,358 Eukaryotes (including 122 mitochondrial and 491 plastid sequences).

As MEGAN utilizes the FASTA header in the BLAST output to identify the corresponding taxon (see below), the headers of all files in the reference databases contain the Genbank, RefSeq or EMBL accession number as identifier, to enable taxon identification with MEGAN (see below).

Taxonomic assignment of RNA-tags using MEGAN

All RNA-tags in the sample were compared to both of the rRNA reference databases using the NCBI *blastall* implementation of BLASTN (default parameters, except setting the maximum number of hits to 100). Analysis of BLAST output files was performed using the MEGAN software version 1beta18 [23]. This software reads the results of a BLAST comparison as input and attempts to place each read on a node in the NCBI taxonomy. This is performed by the LCA algorithm that assigns each RNA-tag to the lowest common ancestor in the taxonomy from a subset of the best scoring matches in the BLAST result (absolute cutoff: BLAST bitscore 86, relative cutoff: 10% of the top hit). RNA-tags that have no BLAST matches are assigned to the special node "no hits" and those unassigned due to algorithmic reasons (e.g. below an applied threshold) are placed on the special node "unassigned". The result of the analysis is displayed as a tree representation of the NCBI taxonomy (as of June 2007). To enable identification of the species involved in the BLAST hits, we employed lookup-tables that map GI accessions to taxon IDs obtained from the NCBI website.

Database refinement, analysis pipeline adjustment and testing

In order to evaluate the performance and stability of our taxonomic binning method using MEGAN and the generated reference databases, we used LSU and SSU rRNA test sets containing simulated ribo-tags. Those consisted of 100 bp long sequence fragments and were obtained by randomly cropping out a 100 bp window from a given rRNA sequence. This procedure was repeated 200 times resulting in a statistical coverage of 13.3 for each position in SSU rRNA (for an assumed length of 1500 bp) and 6.7 for LSU rRNA sequences (for an assumed length of 3000 bp) within the test set. Datasets were iteratively compared by BLASTN against the reference database(s) and analyzed using MEGAN.

Simulated SSU ribo-tag datasets were aligned to the LSUrdb and vice versa, to estimate the potential of "cross-contamination", i.e. the assignment of SSU rRNA-derived ribo-tags as LSU rRNA ribo-tags and vice versa. No such assignments were observed above a BLASTN score of 71 bits. Manual inspections of the alignments from the simulated data lead us to choose 86 bits as the minimum BLASTN score for an RNA-tag to be considered as ribo-tag. These alignments are always longer than 40 bases and correspond to an e-value of 2e-16 for SSUrdb alignments. In addition, a relative threshold was introduced in the taxonomic binning procedure applying MEGANs LCA algorithm. Here, ten percent was chosen as cut-off, which means that the taxonomic information of all reference sequences on the BLAST hit list which are within 10% of the score of the BLAST top hit were included in the taxonomic binning of the respective ribo-tag. Increasing this cutoff to 20% or more, i.e. making the taxonomic assignment of a ribo-tag less strict, led to a very minor shift in taxonomic resolution (data not shown), showing that the chosen relative cut-off is already rather relaxed.

The rRNA reference databases were screened for taxonomically wrongly assigned reference sequences using simulated ribo-tag datasets. Since those consisted of ribo-tags of known origin, an assignment congruent with the taxonomic resolution power of the database for the respective taxon was expected. Where results deviated from these expectations, they were manually inspected. Especially the BLAST hit lists of ribo-tags with poor taxonomic resolution (mainly the taxonomic levels "cellular organism", "domain" and several prokaryotic "phyla") were analysed in detail. For example, a single acidobacterial reference sequence wrongly affiliated to another bacterial phylum would result in a drastic under-representation of acidobacteria, because the LCA algorithm would group the acidobacterial ribo-tags at the bacterial domain level. In the analysis of a natural community, this would introduce a strong artificial community shift, biasing against acidobacteria. Suspicious reference sequences were compared by BLAST against NCBI non-redundant nucleotide database and removed from the reference database. The test set was subsequently compared to the refined reference database, to verify an improved performance for the respective taxon and/or to identify additional wrongly affiliated reference sequences. Applying this iterative procedure, more than 600 bacterial sequences with wrong affiliations in the NCBI taxonomy were removed from the SSUrdb.

In order to extend the sensitivity and robustness analysis (see Supplement Methods and Results S1 in SI), we attempted to simulate the situation where ribo-tags would be as similar to the database as in median case, by removing entries from the database. Database filtering was carried out by aligning the full-length SSU and LSU rRNA sequences from seven selected test species to all entries in the respective reference database, using BLASTN with modified scoring parameters (mismatch penalty = 1, gap open cost = 2, gap extension cost = 1). In all cases where the full-length test sequence showed higher similarity than the threshold to a database sequence, this sequence was removed from the database. The procedure was carried out independently for each test species with the respective sample median similarities used as thresholds (98% for the SSUrdb and 93% for the LSUrdb), thus producing 7 filtered versions of the LSUrdb and SSUrdb. For the SSUrdb, the lowest decile similarity (86%) was also removed in a separate filtering. The simulated SSU and LSU ribo-tags from the test species were then aligned to respective filtered database versions using BLASTN and taxonomically assigned using MEGAN, as described above.

Assembly of a "composite community" rRNA sequence

Using MEGAN, all 1,113 putative SSU ribo-tags assigned to the archaeal GroupI.1b were extracted from the dataset. The assembly program CAP3 [41] was then used to attempt to assemble the data (default parameters) which resulted in a single contiguous sequence contig with a length of 1,502 bp assembled from 1105 of the extracted ribo-tags.

Functional analysis of putative mRNA-tags

All RNA-tags without a significant similarity against the rRNA reference databases (BLAST score threshold below 86 bits), were translated in all six reading frames and aligned to the NCBI non-redundant protein database (release of June 25, 2007) using BLASTX [42] and analysed with MEGAN. The resulting tags with at least one BLAST alignment producing a bitscore of 30 or above were assigned as putative mRNA transcripts. Manual inspection revealed that sequences with BLAST hits close to a bitscore of 30 produce apparently meaningful alignments. In addition, the coverage of the ribosomal reference databases, compared to the true distribution of all existing ribosomal sequences, is expected to be much better than the coverage of Genbank nr compared to all existing protein coding sequences. This assumption, together with the short lengths of the translated reads (approximately 30 amino acids) means that possibly true

mRNA transcripts would be missed, were we to use a higher bitscore limit.

Putative mRNA sequences were annotated using the MG-RAST (Meta Genome Rapid Annotation using Subsystem Technology; v1.2) server at the Argonne National Library (http://metagenomics.nmpdr.org), using subsystem-based annotation based on the SEED database [32]. Subsystems are groups of genes, or functional roles, acting together in a biological process, e.g. in a metabolic pathway. These are grouped into subsystem categories. The MG-RAST annotation pipeline assign some putative genes to more than one subsystem, i e predicts that these genes have multiple functions. In such cases, each assignment was counted in the statistic Total Subsystem Assignments in Table S8 in SI (so that the gene was counted more than once). As a consequence of this, the sum of the relative subsystem counts (''Share of subsystem classified genes'', Figures 6, Figures S6 and S7 in SI) add up to more than 100%. In addition to the putative mRNA-tags, subsystem annotation of two metagenome datasets was also carried out using the same methodology. The first of these contains genomic sequences from a sample collected previously from the same site[19] and the second from farm soil [7] (see for details Table S8 in SI). For each subsystem, the relative population was calculated, i.e. the putative genes assigned to a particular subsystem divided by the total number of genes assigned to subsystems. In addition to manual comparison of relative populations between the samples putative mRNA-tag dataset and genomic sequence from the same site, these datasets were compared statistically using the method described in[43] (Table S9 in SI). In order to rule out that the difference in abundance was not an artifact caused by the shorter length of the putative mRNA-tags, compared to the longer genomic reads obtained using Sanger sequencing, the genomic reads from [19] were fragmented to a number of shorter reads and re-analysed using the MG-RAST server. From each genomic read, seven fragments were randomly generated with a length identical to the average length of the mRNA-tags (98bp), such that the total sequence length was close to that of the un-fragmented sample. The distribution of subsystem categories for the short fragments did not show any deviances larger that 10% compared to the original dataset, for the subsystems or subsystem categories abundant in the mRNA sample (data not shown). This showed that the observed abundance is not an artifact caused by shorter read length of the mRNA-tags. A study where 454 transcripts were functionally categorized and compared to Sanger sequenced ESTs in the plant *Medicago*, also showed that 100bp length is sufficient for meaningful functional classification comparable to traditional ESTs [44].

The putative mRNA-tags were taxonomically binned by analysing the output file of the BLASTX comparison against the NCBI nr protein database with MEGAN. The same parameters as described above for the ribo-tag analysis were applied (except a minimum score of 30 bits). The resulting taxonomic community profile was manually compared to that from the corresponding ribo-tag analyses. In addition, higher score cutoffs (40 and 50 bit) were applied on the mRNA-tags dataset, which gave essentially the same community profile (not shown).

Nine mRNA-tags which were found to have similarity to the ammonia monooxygenase subunit C gene (*amoC*) of archaea, were translated to protein sequence using the reading frames corresponding to the BLASTX hits, and aligned to the *Cenarchaeum symbiosum* AmoC protein sequence using CLUSTAL X [45].

Data deposition

The SSUrdb and LSUrdb are made available for download as compressed fasta files from http://www.bioinfo.no/services/ community-profiling. The sequences reported in this paper have been submitted to GenBank, under accession number SRA001014.

Supporting Information

Methods and Results S1 This file contains supplementary methods and results
Found at: doi:10.1371/journal.pone.0002527.s001 (0.08 MB DOC)

Figure S1 Distribution of sequence difference between assigned SSU ribo-tags and their top scoring BLAST match in the SSUrdb. Similarity is defined as the number of nucleotide identities in the BLASTN alignment divided by the total length of the ribo-tag.
Found at: doi:10.1371/journal.pone.0002527.s002 (0.17 MB TIF)

Figure S2 Distribution of sequence difference between assigned LSU ribo-tags and their top scoring BLAST match in the LSUrdb. Similarity is defined as the number of nucleotide identities in the BLASTN alignment divided by the total length of the ribo-tag.
Found at: doi:10.1371/journal.pone.0002527.s003 (0.20 MB TIF)

Figure S3 3D Bar plot showing the number of correctly assigned simulated SSU ribo-tags at different taxonomical levels. 200 ribo-tags of length 100 bp were randomly simulated from seven test species and compared to a modified version of the SSUrdb, filtered in order to exclude all sequences more than 98% similar to the species test sequence (the median similarity of the SSU ribotags in the sample). Note that no order level in Crenarchaeote 54d9 is defined.
Found at: doi:10.1371/journal.pone.0002527.s004 (0.92 MB TIF)

Figure S4 3D Bar plot showing the number of correctly assigned simulated LSU ribo-tags at different taxonomical levels. 200 ribo-tags of length 100 bp were randomly simulated from seven test species and compared to a modified version of the LSUrdb, filtered in order to exclude all sequences more than 93% similar to the species test sequence (the median similarity of the SSU ribotags in the soil sample to reference sequences in the LSUrdb). Note that no order level in the Crenarchaeote 54d9 is defined.
Found at: doi:10.1371/journal.pone.0002527.s005 (0.74 MB TIF)

Figure S5 MEGAN comparison of archaeal rRNA based and mRNA based community profile. The taxonomic affiliation of an RNA-tag is based on the Blast hits within 10% of the top Blast Bit score.
Found at: doi:10.1371/journal.pone.0002527.s006 (0.94 MB TIF)

Figure S6 Functional analysis of DNA, protein and RNA metabolism subsystems in Rudsoil mRNA-tags, fosmid-derived end sequences from DNA of the same community (Treusch et al., 2004) and shotgun-cloned DNA from a farm soil community (Tringe et al., 2004). All three datasets were subjected to automated analysis using the MG-RAST annotation procedure at the SEED (http://metagenomics.theseed.org).
Found at: doi:10.1371/journal.pone.0002527.s007 (0.26 MB TIF)

Figure S7 Functional analysis of carbohydrate metabolism subsystems in Rudsoil mRNA-tags, fosmid-derived end sequences from DNA of the same community (Treusch, 2004) and shotgun-cloned DNA from a farm soil community (Tringe, 2004). All three datasets were subjected to automated analysis using the MG-RAST annotation procedure at the SEED (http://metagenomics.theseed.org).
Found at: doi:10.1371/journal.pone.0002527.s008 (0.20 MB TIF)

Figure S8 Fraction of phylum-resolution SSU and LSU ribo-tags (in %) affiliated to the five numerically dominant bacterial

phyla (>5% of bacterial ribo-tags) compared to mRNA-tag derived fraction for each phylum.
Found at: doi:10.1371/journal.pone.0002527.s009 (0.15 MB TIF)

Figure S9 The soil crenarchaeota "composite" AmoC protein. Alignment of 9 amoC mRNA-tags (translated into amino acid sequence) against the AmoC protein of Cenarchaeum symbiosum (DQ397580), a member of the marine groupI.1a Crenarchaeota. Mismatches between the sequences are shaded gray. The nine mRNA-tags form three fragments which cover 77% of the C. symbiosum AmoC with 88% sequence identity.
Found at: doi:10.1371/journal.pone.0002527.s010 (1.78 MB TIF)

Table S1 Schematic overview of the sequence content of SSUrdb and LSUrdb.
Found at: doi:10.1371/journal.pone.0002527.s011 (0.03 MB DOC)

Table S2 Archaeal candidate divisions implemented into the NCBI taxonomy.
Found at: doi:10.1371/journal.pone.0002527.s012 (0.05 MB DOC)

Table S3 Result of BLASTN of SSU rRNA test dataset derived from 43 species against SSUrdb.
Found at: doi:10.1371/journal.pone.0002527.s013 (0.12 MB DOC)

Table S4 Result of BLASTN of LSU rRNA test dataset derived from 43 species against LSUrdb.
Found at: doi:10.1371/journal.pone.0002527.s014 (0.12 MB DOC)

Table S5 Sensitivity analysis results for the SSUrdb and LSUrdb.
Found at: doi:10.1371/journal.pone.0002527.s015 (0.14 MB DOC)

Table S6 The bacterial community structure in soil.
Found at: doi:10.1371/journal.pone.0002527.s016 (0.11 MB DOC)

Table S7 The archaeal community structure in soil.
Found at: doi:10.1371/journal.pone.0002527.s017 (0.04 MB DOC)

Table S8 Statistics from functional analysis of putative mRNA-tags.
Found at: doi:10.1371/journal.pone.0002527.s018 (0.03 MB DOC)

Table S9 Functional subsystems with significantly higher abundance in putative mRNA-tags compared to a metagenomic sample.
Found at: doi:10.1371/journal.pone.0002527.s019 (0.03 MB DOC)

Acknowledgments

We thank the Bergen Center for Computational Science for access to the computer cluster. We acknowledge the invaluable work of the different ribosomal database projects by providing curated rRNA sequences to the public. We wish to thank the anonymous reviewers of a previous reviewing process for their very instructive comments. TU acknowledges the participants of the EuroDiversity workshop on Microbial diversity and Ecosystem functioning (Lunz, Austria) for valuable discussions.

Author Contributions

Conceived and designed the experiments: SS TU CS. Performed the experiments: TU. Analyzed the data: TU AL CS. Contributed reagents/materials/analysis tools: SS DH JQ TU AL. Wrote the paper: SS TU AL CS.

References

1. Torsvik V, Ovreas L, Thingstad TF (2002) Prokaryotic diversity–magnitude, dynamics, and controlling factors. Science 296: 1064–1066.
2. Gans J, Wolinsky M, Dunbar J (2005) Computational improvements reveal great bacterial diversity and high metal toxicity in soil. Science 309: 1387–1390.
3. Prosser JI (2007) Microorganisms Cycling Soil Nutrients and Their Diversity. In: van Elsas JD, Jansson JD, Trevors JT, eds. Modern Soil Microbiology, Second Edition. Boca Raton: CRC Press. pp 237–262.
4. Fierer N, Bradford MA, Jackson RB (2007) Toward an ecological classification of soil bacteria. Ecology 88: 1354–1364.
5. Martin-Cuadrado AB, Lopez-Garcia P, Alba JC, Moreira D, Monticelli L, et al. (2007) Metagenomics of the deep mediterranean, a warm bathypelagic habitat. PLoS ONE 2: e914.
6. Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, et al. (2007) The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. PLoS Biol 5: e77.
7. Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, et al. (2005) Comparative metagenomics of microbial communities. Science 308: 554–557.
8. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, et al. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. Science 304: 66–74.
9. Gill SR, Pop M, Deboy RT, Eckburg PB, Turnbaugh PJ, et al. (2006) Metagenomic analysis of the human distal gut microbiome. Science 312: 1355–1359.
10. Huber JA, Welch DB, Morrison HG, Huse SM, Neal PR, et al. (2007) Microbial population structures in the deep marine biosphere. Science 318: 97–100.
11. Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, et al. (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere". Proc Natl Acad Sci U S A 103: 12115–12120.
12. Bailly J, Fraissinet-Tachet L, Verner MC, Debaud JC, Lemaire M, et al. (2007) Soil eukaryotic functional diversity, a metatranscriptomic approach. ISME J 1: 632–642.
13. Poretsky RS, Bano N, Buchan A, LeCleir G, Kleikemper J, et al. (2005) Analysis of microbial gene transcripts in environmental samples. Appl Environ Microbiol 71: 4121–4126.
14. Frias-Lopez J, Shi Y, Tyson GW, Coleman ML, Schuster SC, et al. (2008) Microbial community gene expression in ocean surface waters. Proc Natl Acad Sci U S A 105: 3805–3810.
15. Leininger S, Urich T, Schloter M, Schwark L, Qi J, et al. (2006) Archaea predominate among ammonia-oxidizing prokaryotes in soils. Nature 442: 806–809.
16. Ochsenreiter T, Selezi D, Quaiser A, Bonch-Osmolovskaya L, Schleper C (2003) Diversity and abundance of Crenarchaeota in terrestrial habitats studied by 16S RNA surveys and real time PCR. Environ Microbiol 5: 787–797.
17. Quaiser A, Ochsenreiter T, Klenk HP, Kletzin A, Treusch AH, et al. (2002) First insight into the genome of an uncultivated crenarchaeote from soil. Environ Microbiol 4: 603–611.
18. Quaiser A, Ochsenreiter T, Lanz C, Schuster SC, Treusch AH, et al. (2003) Acidobacteria form a coherent but highly diverse group within the bacterial domain: evidence from environmental genomics. Mol Microbiol 50: 563–575.
19. Treusch AH, Kletzin A, Raddatz G, Ochsenreiter T, Quaiser A, et al. (2004) Characterization of large-insert DNA libraries from soil for environmental genomic studies of Archaea. Environ Microbiol 6: 970–980.
20. Treusch AH, Leininger S, Kletzin A, Schuster SC, Klenk HP, et al. (2005) Novel genes for nitrite reductase and Amo-related proteins indicate a role of uncultivated mesophilic crenarchaeota in nitrogen cycling. Environ Microbiol 7: 1985–1995.
21. Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, et al. (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Res.
22. Cole JR, Chai B, Farris RJ, Wang Q, Kulam-Syed-Mohideen AS, et al. (2007) The ribosomal database project (RDP-II): introducing myRDP space and quality controlled public data. Nucleic Acids Res 35: D169–172.

23. Huson DH, Auch AF, Qi J, Schuster SC (2007) MEGAN analysis of metagenomic data. Genome Res 17: 377–386.
24. Wang Q, Garrity GM, Tiedje JM, Cole JR (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol 73: 5261–5267.
25. Liu Z, Lozupone C, Hamady M, Bushman FD, Knight R (2007) Short pyrosequencing reads suffice for accurate microbial community analysis. Nucleic Acids Res.
26. Roesch LFW, Fulthorpe RR, Riva A, Casella G, Hadwin AKM, et al. (2007) Pyrosequencing enumerates and contrasts soil microbial diversity. The ISME Journal 1: 283–290.
27. Huber H, Hohn MJ, Rachel R, Fuchs T, Wimmer VC, et al. (2002) A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont. Nature 417: 63–67.
28. Fierer N, Breitbart M, Nulton J, Salamon P, Lozupone C, et al. (2007) Metagenomic and Small-Subunit rRNA Analyses Reveal the Genetic Diversity of Bacteria, Archaea, Fungi, and Viruses in Soil. Appl Environ Microbiol 73: 7059–7066.
29. Bremer H, Dennis PP (1996) Modulation of chemical composition and other parameters of the cell by growth rate. In: Neidhardt FC, Curtiss RI, Ingraham JL, Lin ECC, Low KB, et al., eds. Escherichia coli and Salmonella: cellular and molecular biology. Washington, D.C.: ASM Press. pp 1553–1569.
30. Schaefer M (1990) The soil fauna of a beech forest on limestone: trophic structure and energy budget. Oecologia 82: 128–136.
31. Weller R, Ward DM (1989) Selective Recovery of 16S rRNA Sequences from Natural Microbial Communities in the Form of cDNA. Appl Environ Microbiol 55: 1818–1822.
32. Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang HY, et al. (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res 33: 5691–5702.
33. Beaumont HJ, Lens SI, Reijnders WN, Westerhoff HV, van Spanning RJ (2004) Expression of nitrite reductase in Nitrosomonas europaea involves NsrR, a novel nitrite-sensitive transcription repressor. Mol Microbiol 54: 148–158.
34. Konneke M, Bernhard AE, de la Torre JR, Walker CB, Waterbury JB, et al. (2005) Isolation of an autotrophic ammonia-oxidizing marine archaeon. Nature 437: 543–546.
35. Berg IA, Kockelkorn D, Buckel W, Fuchs G (2007) A 3-hydroxypropionate/4-hydroxybutyrate autotrophic carbon dioxide assimilation pathway in Archaea. Science 318: 1782–1786.
36. Hallam SJ, Mincer TJ, Schleper C, Preston CM, Roberts K, et al. (2006) Pathways of carbon assimilation and ammonia oxidation suggested by environmental genomic analyses of marine Crenarchaeota. PLoS Biol 4: e95.
37. Griffiths RI, Whiteley AS, O'Donnell AG, Bailey MJ (2000) Rapid method for coextraction of DNA and RNA from natural environments for analysis of ribosomal DNA- and rRNA-based microbial community composition. Appl Environ Microbiol 66: 5488–5491.
38. Schleper C, Jurgens G, Jonuscheit M (2005) Genomic studies of uncultivated archaea. Nat Rev Microbiol 3: 479–488.
39. Ewing B, Hillier L, Wendl MC, Green P (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res 8: 175–185.
40. Ludwig W, Strunk O, Westram R, Richter L, Meier H, et al. (2004) ARB: a software environment for sequence data. Nucleic Acids Res 32: 1363–1371.
41. Huang X, Madan A (1999) CAP3: A DNA sequence assembly program. Genome Res 9: 868–877.
42. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389–3402.
43. Rodriguez-Brito B, Rohwer F, Edwards RA (2006) An application of statistics to comparative metagenomics. BMC Bioinformatics 7: 162.
44. Cheung F, Haas BJ, Goldberg SM, May GD, Xiao Y, et al. (2006) Sequencing Medicago truncatula expressed sequenced tags using 454 Life Sciences technology. BMC Genomics 7: 272.
45. Jeanmougin F, Thompson JD, Gouy M, Higgins DG, Gibson TJ (1998) Multiple sequence alignment with Clustal X. Trends Biochem Sci 23: 403–405.

Simultaneous assessment of soil microbial community structure and function through analysis of the meta-transcriptome

Tim Urich[1,2], Anders Lanzén[3], Ji Qi[4], Daniel H. Huson[5], Christa Schleper[1,2] and Stephan C. Schuster[4]

SUPPLEMENTARY INFORMATION

Sensitivity and robustness analysis of the taxonomic binning approach

In order to test the sensitivity and robustness of our taxonomic binning approach with MEGAN and the LSUrdb and SSUrdb, we generated an SSU and an LSU test set of simulated ribo-tags. They were prepared from the full-length small- or large subunit rRNA sequence from 43 selected test species, composed of 32 bacterial, five archaeal and six eukaryotic representatives (see Materials and Methods).

SSUrdb (see supplementary table ST3): On the domain level, 100% of the archaeal and eukaryal ribo-tags and 99.8% of bacteria-derived ribo-tags were correctly identified. Ten ribo-tags, which were only identified as "cellular organisms" were from cyanobacteria. Here, the close similarity to plastid sequences from eukaryotes resulted in a decrease in resolution. The resolution power of all archaeal and eukaryal ribo-tags and of 97% bacterial ribo-tags was higher than the domain level. More than 90% of bacterial and archaeal ribo-tags were assigned to the correct phylum, and in 22 out of 26 classes at least 80% of the ribo-tags were correctly grouped, highlighting also that the newly implemented archaeal taxonomy gives reliable results. Similarly, in 21 out of 26 orders, more than 70% of the simulated ribo-tags were correctly affiliated. In the 5 remaining orders, the lower relative representation was due to the usage of taxonomically less well defined sequences in the SSUrdb. The reason for this was a trade-off between providing sufficient reference sequence space and the taxonomic resolution of the reference sequences (e.g. for the acidobacteria). These results show that (1) the community composition can be robustly measured on the domain level, but also with a good resolution until the order level. (2) A comparably small artificial community shift is introduced by the database at the order level.

LSUrdb (see supplementary table ST4): The LSUrdb differs in size and sequence composition from the SSUrdb (see supplementary table ST1). Primarily, it should be regarded as a suitable internal control for the the SSUrdb in its present state. Founding the

community analysis on the two most commonly used marker molecules is a major advantage compared to purely SSU rRNA based studies. Our test dataset indicates that the LSUrdb had a similar taxonomic resolution as the SSUrdb; all eukaryotal and archaeal and 99.7% bacterial ribo-tags were correctly assigned, again without any incorrectly assigned ribo-tags. All microbial orders were represented with at least 70% of the respective sequences. The LSUrdb corroborates the results obtained with the SSUrdb, although with a restriction of considerably reduced sequence space (e.g. for many of the bacterial candidate divisions only SSU rRNA sequences are currently available). Overall, these tests on SSUrdb and LSUrdb did not result in any incorrect taxonomic assignment of any ribo-tag.

The tests described above used simulated ribo-tags with sequences identical to those in the reference databases. While this provides a good indication of the resolution and sensitivity of the taxonomic binning approach for sample ribo-tags essentially identical to the reference sequences (i.e. originating from organisms with known rRNA sequences), it is probable that a fraction of the ribo-tags in an environmental sample will differ significantly from reference sequences. To measure this difference, we analyzed the similarity of ribo-tags from the soil sample to their closest reference sequence. This was carried out separately for the LSU and SSU ribo-tags. Similarity was defined as the number of identical nucleotides in the BLASTN alignment divided by the length of the ribo-tag. The distributions of sequence difference are shown in supplementary figures SF6 and SF7. The median difference (50% of the ribo-tags) was 2% for SSUrdb and 7% for LSUrdb. Difference for the lowest decile (ten percent quantile) was 14% for SSUrdb and 31% for LSUrdb. Differences appeared to be approximately exponentially distributed.

In order to extend the sensitivity and robustness analysis, we simulated the situation where ribo-tags would be as similar to the database as in median cases for seven of the test species. This was done by removing reference sequences with ≥ 2% (SSUrdb) and ≥ 7% (LSUrdb) similarity to a test sequence from the database prior to the taxonomic assignment of simulated ribo-tags (see Online Methods). In addition, all reference sequences with ≥ 86% similarity to test sequences were removed from the SSUrdb, to simulate a situation similar to the lowest decile similarity, as determined above. Results are shown in supplementary table ST5 and in supplementary figures SF8 and SF9.

No significant decrease in taxonomic resolution followed by filtering the SSUrdb at the median similarity level in most cases, (Supplementary table ST5 A and B). At the class level, more than 71% of the simulated ribo-tags were correctly assigned for all seven test species and over 95% for four of them (see table ST5 and Figure SF8). At the domain level, all of the simulated ribo-tags were correctly assigned. Overall, no ribo-tag was incorrectly assigned,

again showing the robustness of our taxonomic binning approach with the SSUrdb. We calculated the probability for not observing no false assignments (out of 800) at the order level, using an exact binomial test with a true false discovery rate of 0.5% or more ($p=0.01813$). Thus, our simulation indicates that the false discovery rate is lower than this at median similarity. These predictions are obviously only valid given that our relatively small test set contains organisms representative of the database as a whole, in terms of "surrounding sequence space", or in other words the density of known sequences in the database compared to the sample.

When removing reference sequences ≥ 86% similar (at the lower decile level, see table ST5 C), assignments were significantly biased between test species at all taxonomical levels. Nonetheless, as many as 50% of the ribo-tags for all test species could be correctly assigned to phylum level or better. Only 5% of the total ribo-tags were incorrectly assigned and 18% remained unassigned (i.e. not classified as SSU rRNA). This indicates that our binning approach generates a comparably low number of false positives. However, this simulation approach is not sufficiently extensive to estimate the false discovery rate for sequences from organisms far from known sequence space. Interestingly, there seems to be no correlation in our test set between the number of unassigned ribo-tags from a test species to the number incorrect assignments.

In the simulated LSU ribo-tags, the resolution and sensitivity decreased significantly when filtering the reference database from sequences ≥ 93% similar (at the median level; see tables ST5). This is not surprising giving the smaller size of the database and the small number of reference sequences for some phyla. For five of the seven test species, however, more than 70% of the ribo-tags were still correctly assigned at least to the class level. In contrast, sequences from the two phyla Spirochaetes and Aquifacae were almost entirely filtered from the reference database with this cutoff, causing a drastic loss of taxonomic resolution and the "disappearance" of the entire phylum Aquifacae. Remarkably, only a small proportion of the *A. aeolicus* ribo-tags was incorrectly assingned to a different phylum, but the majority remained unclassified This shows that for most phyla, results based on LSU ribo-tags appear relatively unbiased, whereas some may introduce a significant bias. In total, five percent of the simulated LSU ribo-tags were incorrectly assigned and 24% remained unassigned. This is comparable to the SSUrdb perfomance at the lower decile level.
Given the current number of sequences in the LSUrdb, and ribo-tag read length of around 100bp, the LSUrdb is mainly regarded as an internal control for results obtained with the SSUrdb.

Out of the SSU ribo-tags in the soil sample, 955 (1%) had only one BLAST match above the applied bitscore cutoff. The corresponding number for LSU ribo-tags was 6,724 (7%). In these cases, assignment will only be based on the best match and thus expected to be less robust. This difference in robustness between the reference databases is comparable to false assignments at the median level in the simulations described.

In summary, the combined, independent usage of both reference databases on a dataset appears to provide a robust and relatively unbiased analysis of the taxonomic composition of a community, especially for SSU ribo-tags. The false discovery rate is expected to be lower than 0.5% for SSU ribo-tags and around 5% for LSU tags. As the ribo-tag length is limited by the pyrosequencing technology used (currently 250 bp are already feasible, with potential for up to 400bp in the near future), we simulated the performance of the taxonomic binning with 200bp long ribo-tags. Especially the SSUrdb performs considerably better with 200 bp at higher taxonomic resolution (data not shown), which will make the taxonomic community profile even more reliable.

Taxonomic binning of archaeal mRNA-tags

The metabolism of groupI.1b Crenarchaeota from soil is, due to the lack of cultured representatives largely unknown. We have taxonomically binned the putative mRNA-tags using MEGAN, applying the same relative threshold as with the ribo-tag analysis on a BLASTX comparison against the NCBI nr protein database (BLAST hits within 10% of the top-hit were included in the binning procedure). 237 mRNA-tags were identified as of archaeal origin. This accounts for 1.1% of all mRNA-tags, which is similar to the proportion of archaeal ribo-tags (1.5%). Remarkably, 2/3 of the mRNA-tags were affiliated with lineages within the Euryarchaeota and only 1/3 with Crenarchaeota, which is in strong contrast to the SSU and LSU ribo-tag derived community profile (supplementary figure SF3). The root cause for this discrepancy is likely the genetic information deposited in the NCBI nr database. There is very limited information about the genomic repertoire of groupI.1b Crenarchaeota, as no genome is sequenced, and only few genomic fragments have been deposited in the databases, obtained through metagenomic studies [1-3]. This probably leads to the taxonomic binning of mRNA-tags to other archaeal groups. Based on the ribo-tag analysis from the same experiment, where the crenarchaeal groupI.1b consistently accounts for more than 98% of the LSU and SSU ribo-tags (supplementary table ST6), we assumed that most of the archaeal mRNA-tags are indeed derived from this group and not from euryarchaeal or other crenarchaeal lineages. Including only the BLAST top hit into the taxonomic binning increased the number of putative archaeal mRNA-tags to 360. Although most of the identified homologues encoded hypothetical proteins, approximately 80 homologues were functionally annotated to a more or less extend.

1. Quaiser A, Ochsenreiter T, Klenk HP, Kletzin A, Treusch AH, et al. (2002) First insight into the genome of an uncultivated crenarchaeote from soil. Environ Microbiol 4: 603-611.
2. Treusch AH, Kletzin A, Raddatz G, Ochsenreiter T, Quaiser A, et al. (2004) Characterization of large-insert DNA libraries from soil for environmental genomic studies of Archaea. Environ Microbiol 6: 970-980.
3. Treusch AH, Leininger S, Kletzin A, Schuster SC, Klenk HP, et al. (2005) Novel genes for nitrite reductase and Amo-related proteins indicate a role of uncultivated mesophilic crenarchaeota in nitrogen cycling. Environ Microbiol 7: 1985-1995.