RESEARCH ARTICLE

Exploring the composition and diversity of microbial communities at the Jan Mayen hydrothermal vent field using RNA and DNA

Anders Lanzén[1,2], Steffen L. Jørgensen[1], Mia M. Bengtsson[1], Inge Jonassen[2,3], Lise Øvreås[1] & Tim Urich[1,4]

[1]Department of Biology and Centre for Geobiology, University of Bergen, Bergen, Norway; [2]Computational Biology Unit, Uni Computing, Uni Research AS, Bergen, Norway; [3]Department of Informatics, University of Bergen, Bergen, Norway; and [4]Department of Genetics in Ecology, Vienna Ecology Center, University of Vienna, Vienna, Austria

Correspondence: Anders Lanzén, Department of Biology and Centre for Geobiology, University of Bergen, PO Box 7803, Bergen 5020, Norway.
Tel.: +47 5558 4032; fax: +47 5558 4450;
e-mail: anders.lanzen@bio.uib.no

Received 23 December 2010; revised 16 May 2011; accepted 20 May 2011.
Final version published online 4 July 2011.

DOI:10.1111/j.1574-6941.2011.01138.x

Editor: Gary King

Keywords
metagenomics; community composition; diversity; bias; SSU rRNA; microbial mat.

Abstract

DNA sequencing technology has proven very valuable for analysing the microbiota of poorly accessible ecosystems such as hydrothermal vents. Using a combination of amplicon and shotgun sequencing of small-subunit rRNA and its gene, we examined the composition and diversity of microbial communities from the recently discovered Jan Mayen vent field, located on Mohn's Ridge in the Norwegian-Greenland Sea. The communities were dominated by the epsilonproteobacterial genera *Sulfurimonas* and *Sulfurovum*. These are mesophiles involved in sulphur metabolism and typically found in vent fluid mixing zones. Composition and diversity predictions differed systematically between extracted DNA and RNA samples as well as between amplicon and shotgun sequencing. These differences were more substantial than those between two biological replicates. Amplicon vs. shotgun sequencing differences could be explained to a large extent by bias introduced during PCR, caused by preferential primer–template annealing, while DNA vs. RNA differences were thought to be caused by differences between the activity levels of taxa. Further, predicted diversity from RNA samples was consistently lower than that from DNA. In summary, this study illustrates how different methods can provide complementary ecological insights.

Introduction

Deep-sea hydrothermal vents offer a wide range of habitats for microorganisms, with rich varieties of environmental conditions and steep physical and chemical gradients where hot vent fluids mix with cold oxygenated seawater (Reysenbach & Shock, 2002). In these environments, several different ecological niches for microorganisms exist, varying from high-temperature chimneys (black and white smokers), warm water vents close to chimneys and diffuse seepages from fissures underlying sediments more distant from the chimneys. From all these vents systems, emission of various reduced gases and solutes can serve as the primary sources of energy for specialized microbial populations, such as those involved in the sulphur cycle. Observations of dense, white suspensions of cells suggest a substantial microbial community in these sediments. A number of studies have been carried out using molecular and genomic techniques to analyse these complex microbial communities (reviewed in, e.g. Reysenbach & Shock, 2002; Nakagawa & Takai, 2008) and contributed towards a better understanding of their ecology.

Recent developments in high-throughput sequencing technology has facilitated the collection of enormous amounts of sequence data and helped to reveal an even greater microbial diversity than previously recognized from many environments including hydrothermal vents (Sogin *et al.*, 2006; Huber *et al.*, 2007, 2010; Brazelton *et al.*, 2010). However, even new technologies such as pyrosequencing are limited in their ability to reveal the complexity of microbial communities. Direct observation of the activities of microorganisms is difficult due to the small scale. Therefore, using sequencing of extracted nucleic acid to explore a complex microbial community can be likened to trying to understand the world by observing shadows on the wall of a cave, like the prisoners in Plato's Allegory of the Cave (Plato & Cornford, 1941). Sequence data often constitute a biased and incomplete representation of the microbiota in an

FEMS MICROBIOLOGY ECOLOGY

ecosystem and are challenging to interpret. Therefore, careful choice and understanding of sequencing strategies and methods is essential.

Pyrosequencing of PCR amplicons (or 'tags') from the small-subunit (SSU) rRNA gene is an increasingly common and cost-efficient technique for studying community composition and diversity (Tringe & Hugenholtz, 2008). This method can yield orders of magnitude more sequence data than Sanger sequencing of clone libraries for the same cost (Engelbrektson *et al.*, 2010). However, like the traditional clone library approach, the SSU region targeted (Liu *et al.*, 2008; Claesson *et al.*, 2009), primer mismatch (Bru *et al.*, 2008; Isenbarger *et al.*, 2008) and PCR conditions (Suzuki & Giovannoni, 1996; Sipos *et al.*, 2007) may bias the results. Further, there are no universal primers for targeting all three domains of life.

Sequencing of SSU rRNA amplicons from reverse-transcribed RNA (cDNA) is better suited for estimating the current *in situ* activity of a community, because cellular rRNA concentration is generally well correlated with growth rate and activity (Poulsen *et al.*, 1993; Bremer & Dennis, 1996). However, reverse transcription also introduces bias and may have lower reproducibility than PCR (Ståhlberg *et al.*, 2004). Amplicons from the SSU rRNA gene can instead provide broader insights into the presence of organisms within the community, because dormant and inactive cells are also targeted. A drawback is the inclusion of DNA from dead cells (Luna *et al.*, 2002). A promising holistic method that can provide both functional information and community composition data from all three domains of life simultaneously is the 'Double RNA Approach' (Urich *et al.*, 2008), where random hexamer primed reverse transcription of RNA is used and the resulting cDNA shotgun is sequenced. As opposed to many other metatranscriptomic studies, the rRNA molecules are retained rather than enriching for mRNA sequences. Thus, this method enables taxonomical profiling at the expense of functional coverage. Importantly, it also avoids the potential PCR and primer bias introduced with cDNA amplicon generation.

In 2005, the world's northernmost active vent fields at that time were discovered, located on the ultraslow spreading Mohn's Ridge, north of the island Jan Mayen in the Norwegian-Greenland Sea. The fields lack the typical macrofauna associated with vent fields to the south along the Mid-Atlantic Ridge (Pedersen *et al.*, 2010; Schander *et al.*, 2010). Large sediment areas were covered with white microbial mats, sustained by diffuse venting of hydrothermal fluids. The aim of this study was to investigate the community composition and diversity of these microbial mats, using a combination of different sequencing methods. In addition, we investigated how the choice of the community profiling method can influence predicted taxonomical composition and diversity estimates. From two samples from similar microbial mats, we extracted both DNA and RNA, and then prepared and pyrosequenced PCR amplicons from both extractions. In addition, shotgun pyrosequencing was carried out both from DNA and cDNA. Using these complementary techniques and marker molecules, the study was designed to obtain a holistic picture of the communities and any differences between them. In addition, this allowed us to analyse whether any systematic bias was introduced during PCR or otherwise, between amplicon sequencing and shotgun sequencing.

Materials and methods

Study site, samples and experimental design

Two sediment samples from white microbial mats were retrieved from the Trollveggen vent field (71°17.88, −5°46.34) at a depth of 564 m using a remotely operated vehicle, during the H2DEEP cruise, 25 July 2008. Two different sites from apparently similar habitats were sampled. One site was sampled using a slurping device [sample 1 (S1)], where the sample materials were collected in a closed container and the second sediment sample was collected using a metal box with a connected shovel [sample 2 (S2)]. The sediment sample was visibly stratified with a white top layer and underlying grey layers. Several kilograms of this sample were transported through the water column and on board the ship for immediate processing. The top layer including the upper 3–5 mm of the grey area was withdrawn using sterile spatulas. The slurping device was used in such a manner that roughly the same layers were collected. Again, the white layers and the surrounding grey material were sampled using spatulas. Samples were directly transferred into bead-beating tubes (MP Biomedicals), flash-frozen in liquid nitrogen and stored at −80 °C. The time from sampling until arrival onboard was approximately 60 min and the processing was another 30 min.

From the two biological replicates retrieved (S1 and S2), three different nucleic acid extractions were carried out and thereafter treated using different experimental schemes according to Fig. 1.

Nucleic acid extraction and cDNA synthesis

Nucleic acids (DNA and RNA) were simultaneously extracted from each of the two samples using a phenol: chloroform protocol and mechanical shearing (Urich *et al.*, 2008). To account for differences in extraction efficiencies between easily and difficultly lysed cells, bead beating was performed at two different speeds in the Fast-Prep instrument (4 and 5 ms^{-1}). The triplicate extractions at both conditions were later on merged into one nucleic acid pool per sample. After DNase treatment and RNA purification using MegaClear columns (Ambion), the success of the DNase treatment was verified by PCR with 16S rRNA gene primers specific for bacteria and archaea. Subsequently, total

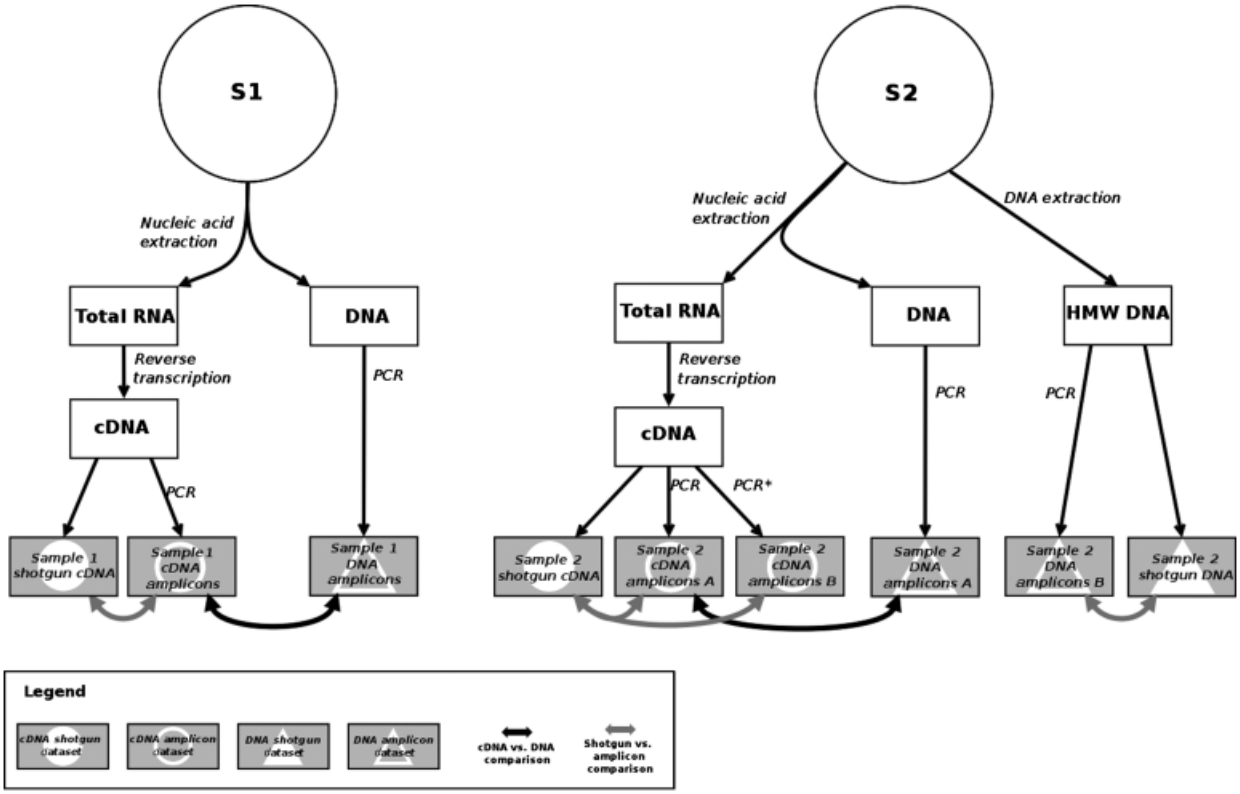

Fig. 1. Overview of the experimental design and the resulting datasets. The two samples are represented as circles. Total community DNA and RNA was extracted from both samples and a separate DNA extraction was made from Sample 2 to prepare a high-molecular-weight-DNA library ('HMW DNA'). RNA isolates were reverse transcribed into cDNA. Shotgun and amplicon sequencing of these produced nine datasets of sequence reads. Dataset types and comparisons made are indicated according to the legend. The amplification using a unique reverse primer (1492R) is indicated with a star.

RNA was subjected to double-stranded cDNA synthesis as described previously in Urich *et al.* (2008).

Preparation of high-molecular-weight DNA (HMW-DNA)

Extraction of HMW-DNA was carried out from S2, using the protocol described by Zhou *et al.* (1996). Five grams (wet weight) of sediment was mixed with 13.5 mL DNA extraction buffer with added Proteinase K and purified using phenol: chloroform. The extracted DNA was analysed on a 0.7% agarose gel and then on Nanodrop to measure its quality and concentration (absorption spectra, A_{260}/A_{280} and A_{260}/A_{230}).

Shotgun sequencing

Shotgun sequencing of the HMW-DNA and the two cDNA preparations was performed on a GS FLX pyrosequencer (Margulies *et al.*, 2005) using Titanium chemistry at the Norwegian High-throughput Sequencing Centre located at the University of Oslo, Norway. All sequencing reads shorter than 150 bp were removed as well as those containing degenerate bases ('Ns') or an average quality score below 25.

Sequencing of 16S amplicons

The V5-V6 region of 16S rRNA gene was targeted and amplified for all cDNA and DNA libraries prepared using the 'universal' forward primer Uni787F (ATTAGATACCC NGGTAG; Roesch *et al.*, 2007) and the 'universal' reverse primer 1391R (ACGGGCGGTGWGTRC; modified from Lane *et al.*, 1985). DNA and cDNA from each sample (10 ng) was PCR amplified in triplicates using the above-mentioned primers under the following thermal conditions: 95 °C/15 min, then 20–30 cycles of 94 °C for 45 s, 53 °C for 45 s and 72 °C for 1 min, followed by 72 °C for 7 min before cooling at 4 °C. A new round of PCR was performed to link the A and B pyrosequencing adaptors (454 Life Sciences/ Roche) and the unique barcodes to the amplicons as described by Hamady *et al.* (2008). The resulting amplicons were sequenced using multiplex GS FLX pyrosequencing. No protocol using Titanium chemistry was available for amplicon sequencing at this time, but became so later during the study. To validate that this generated similar results, a separate PCR and Titanium pyrosequencing was performed following the same PCR protocol, except using

1492R (GNTACCTTGTTACGACTT; Roesch *et al.*, 2007) as a reverse primer.

Filtering and removal of noise and chimeras from 16S amplicon sequences

Filtering and noise removal of 16S amplicon sequences was carried out using AMPLICONNOISE (Quince *et al.*, 2011). This method corrects the noise and errors introduced during pyrosequencing and PCR. Thus, a new set of denoised and chimera-filtered sequences were generated, each with a set of associated original sequence reads. Barcode and primer sequences were removed before further analysis such as taxonomic classification and clustering.

Preparation of a SSU rRNA reference database

A 16S rRNA gene reference sequence database was prepared from the Silva SSURef release 100 (Pruesse *et al.*, 2007; http://www.arb-silva.de/documentation/background/release-100). Using the ARB software package, we reviewed the SSURef alignment and removed all sequences with a pintail score below 75, an alignment quality score (*align_qual_slv*) below 75 or a length below 1200 bp. For most bacterial taxons, the Silva Taxonomy was used. In addition, we manually reviewed and edited taxonomy assignments for the *Epsilonproteobacteria*, *Acidobacteria*, *Chloroflexi* and the *Archaea*. Eukaryotic sequences including mitochondrial and plastid sequences were analysed and chosen for the database based on taxonomic affiliation in the NCBI taxonomy and a comparatively high taxonomic resolution. The database consisted of 341 683 sequences at the time of analysis.

All sequences were exported in FASTA format using a positional filter (including positions between 1000 and 43 284 of the ARB alignment) along with a tab-separated text file describing the accession number, Silva taxonomy placement and NCBI Taxonomy ID for each retained SSURef sequence. Based on the NDS file, configuration files for MEGAN (Huson *et al.*, 2007) were prepared (including a taxonomy tree in Newick format). The modified SSURef database is available for download from http://www.bioinfo.no/services/community-profiling and work is underway to release the database along with an rRNA version of the MEGAN software (Huson *et al.*, 2007).

Taxonomic classification and grouping

All filtered, denoised and chimera-filtered sequences were aligned to the modified SSURef database using BLASTN (default parameters) as implemented in the NCBI standalone BLAST suite. All sequences with a bit-score above 150 to any database reference sequence were classified as SSU rRNA. The BLAST result was analysed using MEGAN version 3.7 (Huson *et al.*, 2007), and sequences were assigned to taxa in the modified Silva Taxonomy described above (default parameters except *Min Support* = 1 and *Min Score* = 150). MEGAN assignments were then exported and, for amplicon sequences, also weighted according to the original number of associated reads. The abundance counts of each taxon at different ranks were calculated using these weighed assignments. The relative abundance for eukaryotic taxa was defined as the number of reads assigned to the eukaryotic taxon divided by the total prokaryotic reads in the dataset.

Operational taxonomic units (OTU) clustering, richness estimates (Chao1) and rarefaction analysis

All SSU sequences derived from amplicon sequencing were clustered into OTUs using maximum linkage clustering, based on pairwise distances generated using the exact pairwise Needleman–Wunsch algorithm, as described in Quince *et al.* (2011). A 3% distance cutoff was used to define OTUs.

Chao1 estimates of minimum diversity (Chao, 1987) and Simpson's index of diversity (1-D) were calculated, using a custom Python script. Rarefaction curves were calculated using the program E RAREFACTION distributed with AMPLICONNOISE (Quince *et al.*, 2011).

Comparison of taxon abundances across datasets

The relative abundance of each taxon was compared across datasets using a custom Python script. For each taxon in each pairwise comparison, we calculated the ratio of proportions (RP), the odds ratio and the difference between proportions, as recommended in Parks & Beiko (2010). Where the relative abundance of a taxon is p_1 in the first and p_2 in the second dataset, RP is defined as p_1/p_2.

As recommended in Parks & Beiko (2010), Fisher's exact test was used to determine the significance of the difference in observed relative abundance, or more precisely, the probability of the null hypothesis that the observed number of reads assigned to the taxon was drawn from the same underlying distribution. Two-tailed tests were performed using the Python package FISHER v0.1.4 (Tang & Pedersen, 2010). *P*-values were corrected for multiple hypotheses testing using Bonferroni correction. All taxa with corrected *P*-values < 0.05 were considered observed at significantly different relative abundance between datasets. Eukaryotes were excluded in the comparisons.

Linear regression, hierarchical clustering of datasets and other statistical analyses were carried out using the R programming language. Nonmetric multidimensional scaling (NMDS) and calculation of dataset dissimilarity indices were carried out using the functions 'metaMDS' and 'vegdist' in the R package VEGAN (Oksanen *et al.*, 2010).

Mismatches and base at degenerate primer positions

A number of factors were thought to influence primer binding in our samples and thus bias PCR efficiency. Using the shotgun sequence data, we studied the following factors: share of reads with mismatches to the forward (MM_F) and reverse primer (MM_R); and the share of reads with a *G* or a *C* in the degenerate base position of the forward (GC_F)- and reverse primer (GC_R)-binding sites. These factors were determined for each taxon, individually in each of the three shotgun datasets.

Using a custom Python script, all variants of the degenerate primer sequence were determined for the three primers. In addition, all transformations of the variants whose reverse complement has exactly one or two mismatches to one of the correct primer sequence variants were calculated and annotated with mismatch positions (see file primer_mismatch.fasta). The resulting set of sequences was aligned to each taxon dataset in each shotgun dataset using BLASTN (only full-length identical matches were retained). Values of GC_F and GC_R were estimated as the share of matches to versions with a G or a C in the degenerate position, while values of MM_F and MM_R were estimated as the share of matches to mismatch-transformed sequences.

Data submission

Pyrosequencing flowgrams (SFF files) containing only rRNA reads were submitted to the NCBI Sequence Read Archive with the study accession number SRP004929.

Results

Community composition

During initial filtering, 22% of all reads were removed and in total SSU rRNA sequences representing 199 736 sequence reads were obtained. Table 1 provides an overview of the datasets. The vast majority of reads from the metagenomic dataset (shotgun DNA) as well as mRNA reads from the metatranscriptomic datasets (shotgun cDNA) do not represent SSU rRNA and were not analysed here, but instead in a separate functional study of the community (T. Urich, A. Lanzén, R. Stokke, R.B. Pedersen, I.H. Thorseth, C. Schleper, I.H. Steen & L. Øvreås, unpublished data). The number of SSU rRNA reads ranged from 565 to 85 893 in shotgun sequencing datasets and from 3980 to 8903 in amplicon datasets.

In total, 96% of all SSU rRNA reads could be assigned at the class rank (Supporting Information, Table S2). Relative taxon abundances and OTU richness at this rank are shown in Fig. 2 for all datasets. In further analysis, candidate phyla for which no taxonomical grouping existed at class level were also included. In addition, because the *Epsilonproteobacteria* dominated all datasets, this class was divided further into genera. *Sulfurimonas*, being the most abundant taxon, contribute between 23% and 82% of prokaryotic abundance in the different datasets, followed by *Sulfurovum*, contributing between 5% and 39%.

The classification of S1 shotgun cDNA reads was also carried out using the RDP CLASSIFIER v2.0 (Wang *et al.*, 2007) with a confidence threshold of 0.7. Although the taxonomies used for classification differed slightly, the two methods agreed for the classification of 93% of the reads at the genus level, where 69% of the reads could be classified with the RDP

Table 1. Overview of the datasets

Sequence dataset	Total reads	PCR cycles	Reads after filtering	SSU rRNA reads*	rRNA OTUs	SDI†	Sequencing technology	Average sequence length (bp)‡	Classified at genus level (%)§
S1 shotgun cDNA	190 051	0	156 370	78 111	NA	NA	GS FLX Titanium	393	71
S1 cDNA amplicons	5708	20	4 341	4338	252	0.70	GS FLX	231	84
S1 DNA amplicons	7812	20	6 263	6254	385	0.88	GS FLX	231	65
S2 shotgun cDNA	172 930	0	145 160	85 893	NA	NA	GS FLX Titanium	382	78
S2 cDNA amplicons A	8505	20	6 495	6493	213	0.45	GS FLX	231	92
S2 cDNA amplicons B	12 568	25	8 904	8903	314	0.54	GS FLX Titanium	307	89
S2 DNA amplicons A	6735	20	5 206	5199	261	0.84	GS FLX	231	54
S2 DNA amplicons B	5579	30	3 986	3980	292	0.88	GS FLX	231	69
S2 shotgun DNA	637 973	0	484 176	565	NA	NA	GS FLX Titanium	433	59
All datasets	1 047 861		820 901	199 736	982¶			388	75

*Reads with a BLASTN alignment bit-score > 150 to the SSURef database.
†Simpson's diversity index (1-D; Simpson, 1949) calculated using all rRNA OTUs.
‡Average sequence length after removal of barcode and primer sequences.
§Share of prokaryotic reads classified at domain that were also unambiguously classified at genus level.
¶Including amplicon reads only.

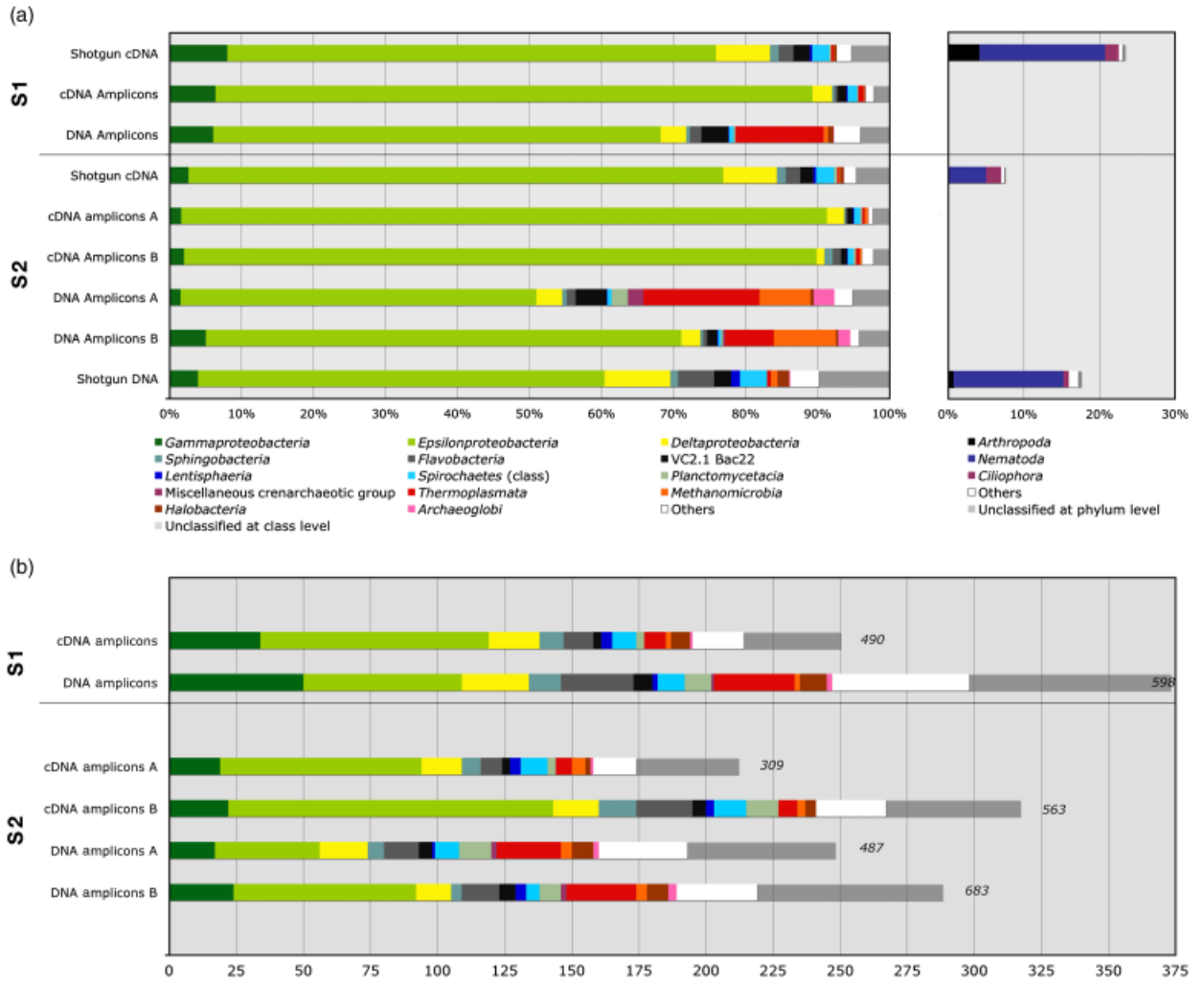

Fig. 2. Relative abundance (a) and OTU richness (b) for all datasets, given at the class rank for all prokaryotes and phylum rank for eukaryotes. Relative abundance is shown as the number of reads assigned to a taxa divided by the total number of prokaryotic reads. The Chao1 estimate of minimum OTU richness is written next to each bar in (b).

CLASSIFIER and 71% with the classification method presented in this study (see Fig. S1). As opposed to this MEGAN-based classification, the RDP CLASSIFIER cannot classify any of the 23% eukaryotic sequences in this dataset, because the default version of the program was not trained for classification of eukaryotic sequences.

The relative abundance and OTU richness of taxa when combining all amplicon datasets are plotted in Fig. S2. A clear log-linear relationship was observed between abundance and richness ($R^2 = 0.78$, $P < 2E-16$), similar to what has been observed in bacterial communities inhabiting the ocean surface (Kirchman *et al.*, 2010). The 'Miscellaneous Crenarchaeotic Group' and the *Archaeoglobi* are the two largest outliers from this relationship, showing an unexpectedly low diversity.

To identify systematic differences between the datasets, Bray–Curtis dissimilarities (Legendre & Legendre, 1998) were calculated using relative prokaryotic abundances (excluding eukaryotic and insufficiently abundant taxa; Table S3). In a hierarchical average linkage clustering of these distances, all RNA-derived datasets form a separate cluster (Fig. 3a). This indicates that the difference between DNA- and RNA-derived datasets was more substantial than other experimental factors. To visualize the patterns in predicted community composition, NMDS was also performed (Fig. 3b). The resulting two-dimensional map agrees with the clustering analysis.

Comparing composition and diversity between DNA and RNA

In order to determine differences in predicted community composition between 16S rRNA gene datasets from DNA and RNA, two pairs of amplicon datasets were compared.

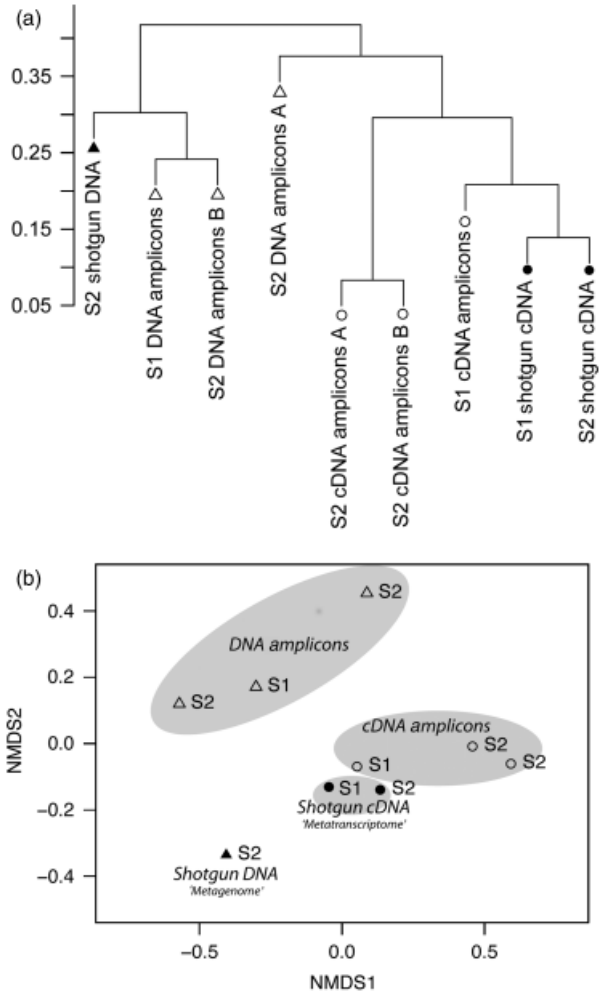

Fig. 3. Dendrogram resulting from hierarchical average linkage clustering (a) and NMDS map (b) of the datasets, based on Bray–Curtis dissimilarities between relative taxon abundances. Datasets obtained using sequencing of cDNA are plotted as circles and those from genomic DNA as triangles. Filled circles and triangles indicate datasets from shotgun sequencing of cDNA and DNA; empty ones show amplicon sequencing.

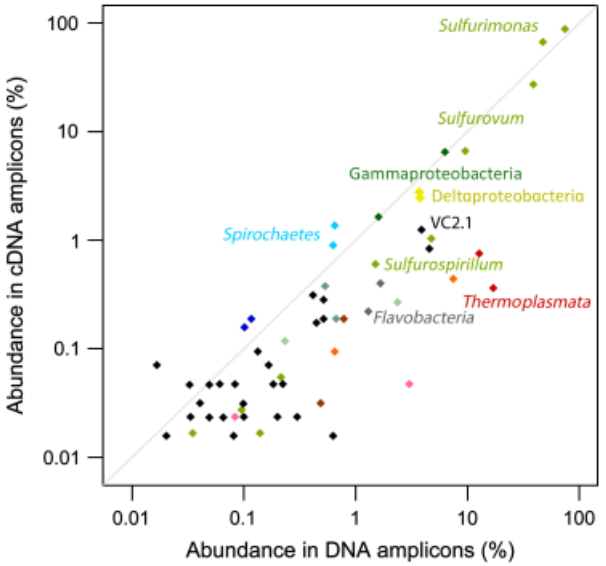

Fig. 4. Relative taxon abundance in amplicon datasets from DNA plotted vs. corresponding reverse-transcribed RNA (cDNA) datasets, in logarithmic scale.

Each pair thus contained two datasets: one from the DNA and one from the RNA of the same nucleic acid extraction (Fig. 1). The comparison revealed a similar abundance of the numerically dominant taxa, whereas the majority of rare taxa were more abundant in DNA than RNA (Fig. 4). The five most differentially observed taxa (lowest corrected *P*-values) are listed in Table 2 (all with $P < 0.05$ are listed in Table S4). Rarefaction curves of all amplicon datasets (Fig. S3) showed that the OTU richness was higher in all DNA-derived datasets compared with RNA from the same sample, when adjusting for sequencing depth. Simpson's diversity indices (1-D; Simpson, 1949) showed the same pattern (Table 1).

Comparing composition between amplicon and shotgun sequencing

To identify and quantify any taxon-specific PCR bias influencing the formation of amplicons, four comparisons were made between datasets originating from amplicon vs. shotgun sequencing. The comparisons were made so that both datasets always originated from the same pool of DNA or cDNA (see Fig. 1). The relative abundances for taxa between the two datasets in each pair are compared in Fig. 5 (and Table S5).

Taxa with less than five reads expected in the smallest dataset were excluded from further analysis, in order to compensate for difference in sequencing depth.

Whereas several taxa were similarly abundant, some taxa were consistently underrepresented in all amplicon sequencing datasets, indicating a negative amplification bias. The same is true for the five taxa with most significant differences (Table 3). When analysing their primer-binding sites, we identified a shared sequence feature between the taxa; in their degenerate primer-binding position, all are dominated by A or T. Only two taxa appear to be dominated by reads mismatching the forward primer, namely *Halobacteria* (> 90% of reads) and the candidate phylum 'BD-1-5' (> 95%), both significantly underrepresented in amplicon datasets.

In order to analyse the factors that could influence amplification efficiency during the PCR, a multidimensional linear regression was performed. We found that the RP between abundance in the amplicon and shotgun sequence datasets was significantly correlated with three parameters: (1) the shares of reads with mismatches to the forward primer (MM_F), (2) a G/C base in the degenerate position of the

Table 2. The five taxa with the most significant differences in relative abundance between sequencing of DNA and cDNA amplicons in samples 1 and 2

Comparison		DNA dataset		cDNA dataset		Change relative cDNA						
Taxon	Sample	Count	Share (%)	Count	Share (%)	+/−	RP*	OR†	DP (%)‡	$P^{\S}$	P_{corr}§	P_{all}¶
Flavobacteria	1	100	1.66	17	0.40	+	4.16	4.21	1.26	5E − 10	2E − 08	2E − 16
Flavobacteria	2	64	1.30	14	0.22	+	5.88	5.94	1.08	2E − 10	8E − 09	
Sulfurimonas	1	1982	47.24	2450	66.98	−	0.71	0.44	− 19.74	6E − 10	4E − 08	9E − 16
Sulfurimonas	2	2133	74.35	5294	88.32	−	0.84	0.38	− 13.98	6E − 10	2E − 08	
Sulfurovum	1	1626	38.75	997	27.26	+	1.42	1.69	11.50	7E − 10	5E − 08	4E − 12
Sulfurovum	2	274	9.55	398	6.64	+	1.44	1.48	2.91	2E − 06	9E − 05	
Thermoplasmata	1	764	12.72	32	0.75	+	16.88	19.19	11.96	4E − 10	2E − 08	3E − 16
Thermoplasmata	2	839	16.99	23	0.36	+	46.90	56.30	16.63	5E − 10	2E − 08	
VC2.1 Bac22	1	233	3.88	53	1.25	+	3.11	3.19	2.63	3E − 10	1E − 08	2E − 16
VC2.1 Bac22	2	225	4.56	53	0.83	+	5.46	5.67	3.72	3E − 10	1E − 08	

*Ratio of proportions (DNA dataset share/cDNA dataset share).

†Odds ratio.

‡Difference of proportions (DNA dataset share − cDNA dataset share).

§*P*-value for H_0 assuming equal abundance in both samples, tested using Fisher's exact test. P_{corr} gives the *P*-value after Bonferroni correction for multiple hypothesis testing.

¶Multiplication of all Bonferroni-corrected *P*-values for taxa throughout datasets yielding the total probability of H_0.

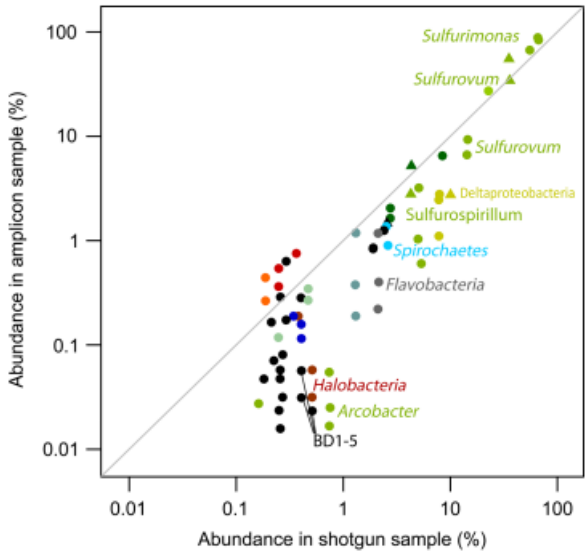

Fig. 5. Relative taxon abundance in amplicon datasets plotted vs. corresponding shotgun datasets, in logarithmic scale. Circles indicate the comparison of cDNA datasets, and triangles that of DNA.

forward primer (GC_F) and (3) a G/C in the degenerate position of the reverse primer (GC_R). Values of these parameters were estimated for each taxon using shotgun sequences (Table S5). MM_F had the largest influence on RP ($R^2 = 0.49$; $P = 2E - 9$; $RP = 0.39–1.2 \times MM_F + 0.91 \times GC_F + 0.26 \times GC_R$). About half of the variation could thus be explained by PCR bias caused by primer mismatch and preferential binding to templates matching the primer versions with G/C in the degenerate positions. Mismatches in the reverse primer did not add any significance as explanatory variables for PCR bias. The correlation between mismatching shares in the forward and reverse primers was weak and cannot explain this.

The number of PCR cycles differed between datasets (Table 1) and we expected this to influence the extent of the bias. In an attempt to use a simple model with an exponentially increasing bias, the $RP^{1/n}$ was instead used as dependent (where *n* is the number of cycles used), but it did not show a significant correlation to any explanatory variables. Instead, using a linear model where bias was additive with every cycle, as suggested by Polz & Cavanaugh (1998), increased the fit of regression ($R^2 = 0.51$; $P = 5E - 10$), suggesting that the bias was mainly template dependent, not influencing product–primer annealing to the same extent.

Discussion

Community composition and abundant taxa

All DNA- and RNA-based datasets were dominated by the epsilonproteobacterial genera *Sulfurimonas* and *Sulfurovum*, together contributing between 50% and 80% of relative abundance (Fig. 2). These genera are commonly found in vent fluid mixing zones of hydrothermal habitats. Typically restricted to areas with low oxygen concentrations, they are mesophilic chemoautotrophs capable of oxidizing sulphur compounds, coupled with the reduction of nitrate or oxygen and use of the rTCA cycle for CO_2 fixation (Campbell *et al.*, 2006; Sievert *et al.*, 2008; Yamamoto *et al.*, 2010). Most *Gammaproteobacteria* present in the samples appeared to be from the *Methylococcales* order, indicating that they oxidize methane rather than sulphur compounds. *Deltaproteo-*

Table 3. The five taxa with the most significant differences in relative abundance between sequencing of amplicon and shotgun libraries from the same DNA or cDNA pool

Comparison		Amplicon sequencing		Shotgun sequencing		Change relative shotgun abundance					
Taxon	DNA pool	Count	Share (%)	Count	Share (%)	RP*	OR†	DP (%)‡	P§	P_{corr}§	P_{all}¶
Arcobacter	S1 cDNA	2	0.05	338	0.74	0.074	0.073	−0.68	4E−09	2E−07	4E−23
Arcobacter	S2 cDNA[A]	1	0.02	468	0.74	0.023	0.022	−0.72	2E−10	1E−08	
Arcobacter	S2 cDNA[B]	2	0.03	468	0.75	0.033	0.033	−0.73	2E−10	2E−08	
Deltaproteobacteria	S1 cDNA	118	2.78	4759	7.92	0.35	0.33	−5.14	5E−10	1E−08	3E−32
Deltaproteobacteria	S2 cDNA[A]	156	2.46	5956	7.80	0.31	0.30	−5.35	6E−10	2E−08	
Deltaproteobacteria	S2 cDNA[B]	96	1.10	5956	7.82	0.14	0.13	−6.72	7E−10	4E−08	
Deltaproteobacteria	S2 HMW DNA	105	2.75	51	10.04	0.27	0.25	−7.29	7E−11	3E−09	
Flavobacteria	S1 cDNA	17	0.40	1289	2.14	0.19	0.18	−1.74	3E−10	9E−09	5E−33
Flavobacteria	S2 cDNA[A]	14	0.22	1609	2.11	0.10	0.10	−1.89	3E−10	9E−09	
Flavobacteria	S2 cDNA[B]	102	1.17	1609	2.11	0.56	0.55	−0.94	6E−10	3E−08	
Flavobacteria	S2 HMW DNA	23	0.60	28	5.51	0.11	0.10	−4.91	4E−11	2E−09	
Spirochaetes	S1 cDNA	58	1.37	1518	2.53	0.54	0.53	−1.16	4E−07	1E−05	6E−30
Spirochaetes	S2 cDNA[A]	57	0.90	1996	2.61	0.34	0.34	−1.72	4E−10	1E−08	
Spirochaetes	S2 cDNA[B]	71	0.82	1996	2.62	0.31	0.31	−1.81	4E−10	2E−08	
Spirochaetes	S2 HMW DNA	16	0.42	21	4.13	0.10	0.10	−3.72	4E−11	2E−09	
Sulfurospirillum	S1 cDNA	22	0.60	2449	5.36	0.11	0.11	−4.75	4E−10	2E−08	3E−23
Sulfurospirillum	S2 cDNA[A]	62	1.03	3156	5.00	0.21	0.20	−3.96	4E−10	2E−08	
Sulfurospirillum	S2 cDNA[B]	256	3.20	3156	5.08	0.63	0.62	−1.88	6E−10	5E−08	

*Ratio of proportions (PCR amplicon dataset share/shotgun dataset share).

†Odds ratio.

‡Difference of proportions (PCR amplicon dataset share − shotgun dataset share).

§P-value for H_0 that assumes equal abundance in both samples, tested using Fisher's exact test. P_{corr} gives the P-value after Bonferroni correction for multiple hypothesis testing.

¶Multiplication of all Bonferroni-corrected P-values for taxa throughout datasets yielding the total probability of H_0.

bacteria, mainly from the sulphate-reducing genus *Desulfobacterium*, accounted for up to 9% of abundance in the datasets. The co-occurrence of sulphur-oxidizing *Epsilonproteobacteria* and sulphate-reducing *Deltaproteobacteria* has been reported previously in hydrothermal environments and may suggest an internal sulphur cycle within the sediments (Teske *et al*., 2002; Perner *et al*., 2007). Other abundant taxa include *Bacteriodetes* and *Spirochaetes*, phyla consisting mainly of heterotrophs, also detected previously in hydrothermal environments (Kirchman, 2002; Imachi *et al*., 2008). *Sulfurimonas* and *Sulfurovum* have also been found to dominate several 16S rRNA gene clone libraries from hydrothermal vent systems and their associated microbial mats (e.g. Moyer *et al*., 1995; Kormas *et al*., 2006; Davis & Moyer, 2008). In the 'Marker 52' dataset taken from vent fluids above a microbial mat, collected at the Juan de Fuca Ridge (Sogin *et al*., 2006; Huber *et al*., 2007), these two groups accounted for about 60% of all sequences. The third and fourth most abundant taxa from Marker 52 were *Deferribacterales* and *Desulfobacterales*. The latter is present at a similar abundance in our samples, whereas the former is present at a much lower abundance. Therefore, we can assume that Marker 52 was sampled from a similar community. It also shares similar chemical characteristics such as a relatively low pH (5.1 vs. 4.5 in our sample) and a high H_2S concentration. Four out of 14 low-temperature vent samples from active seamounts in the Mariana Arch (FS447-449, FS473), studied by Huber *et al*. (2010), were also dominated by *Sulfurimonas* and *Sulfurovum*. These four samples also had a low pH (2.6–5.9) and many were from sites with 'abundant white microbial mats'.

The shotgun sequencing datasets (especially cDNA) allowed insights into the eukaryotic community of the microbial mats studied (Fig. 1). Between 7% and 19% of the total SSU rRNA reads were eukaryotic. A majority were identical to the 18S rRNA sequence of *Halomonhystera disjuncta*, a cosmopolitan, bacterivorous nematode previously found in similar environments, known for its high resistance to environmental stress (Vanreusel *et al*., 2010). In addition, a diversity of eukaryotic taxa were present, most notably ciliates and copepods.

Diversity and OTU richness

OTU richness of the datasets in this study ranged between 213 and 385 and clustering of all amplicon datasets resulted in 982 OTUs. As indicated by rarefaction analysis (Fig. S3) and Chao1 estimates of minimum diversity (Fig. 2b), our sampling was far from exhaustive. The deeper sequencing carried out by Huber *et al*. (2007) resulted in over 20 000 bacterial and archaeal OTUs in the two datasets from Marker 52 and Bag City. Higher sequencing depth is clearly

a part of the explanation for the much higher diversity estimate. However, sequencing noise is likely to contribute even more. The filtering and noise-removal method used in this study (Quince *et al.*, 2011) instead yields more accurate richness values that cannot be directly compared with studies where a corresponding method was not used (Quince *et al.*, 2009). Huber and colleagues also noted that about 1/3 of reads were more than 10% different from the closest reference sequence, whereas only about 3% of shotgun reads and 0.5% of amplicon reads in our datasets share this low similarity to the reference database used here. This could be due to better sequence coverage in our database, as well as another effect of noise reduction and improved filtering of low-quality reads.

Apart from the dramatic difference in richness estimates, diversity structures are very similar in our datasets and Marker 52. In both studies, the dominating *Epsilonproteobacteria* contribute about one third of the bacterial OTU richness while the archaea appear to be less diverse compared with their abundance. In the study where Huber *et al.* (2010) targeted only *Epsilonproteobacteria*, they used a more aggressive OTU clustering algorithm to better compensate for sequencing noise. Rarefied OTU richness at 3000 reads for the four datasets mentioned (FS447-449, FS473) is similar to our DNA-based datasets with around 100 OTUs.

Variance and systematic methodological differences

The experimental design allowed us to explore the differences between the perceived community structures when using DNA vs. RNA as marker molecules, as well as amplicon vs. shotgun sequencing (see Fig. 1). The two samples were used to compare the magnitude of putative methodological differences with that of sample-to-sample variation. Many statistically significant differences between the resulting community composition predictions were consistently observed in both samples, indicating that they were systematic. The clustering and NMDS analyses indicate that the choice of marker molecule and sequencing strategy (DNA vs. cDNA and amplicon vs. shotgun) influenced the predicted community composition, and that this influence was stronger than the variation between the two biological replicates (S1 and S2), even though different sampling methods were used to collect the two samples

Community differences between RNA and DNA

The predicted community composition differed significantly and systematically depending on whether DNA or RNA was extracted. Several taxa had a significantly lower relative abundance in RNA-derived datasets. The most noticeable difference in relative abundance was at the domain level, where Archaea constituted between 16% and 30% in DNA amplicon datasets and only around 1% in the cDNA datasets. In addition, amplicon sequencing from cDNA resulted in datasets with a consistently lower Simpson's diversity index than from DNA (Table 1).

The abundances of rRNA and that of its gene indicate different aspects of the microbial community and are known to differ from each other over shorter and longer time scales (e.g. Rodriguez-Blanco *et al.*, 2009; Jones & Lennon, 2010; McCarren *et al.*, 2010). While ribosome abundance generally reflects activity, rRNA gene abundance includes slow-growing, dormant and dead organisms. Inactive cells, including spores, are thought to act as a seed bank and play an important role in the maintenance of the high biodiversity observed in most ecosystems (Pedrós-Alió, 2006; Sogin *et al.*, 2006; Jones & Lennon, 2010). Thus, these differences are very likely to reflect actual biological differences between the RNA and the DNA pools. In addition, bias introduced during the reverse-transcription step may have contributed to the differences. Although less likely than for mRNA, partial degradation of rRNA during transport to the surface may also have biased this comparison.

Influence of PCR bias on predicted community composition

Comparisons between the relative taxa abundances obtained using amplicon and shotgun sequencing suggested that at least three factors significantly influenced the differences observed. The most influential factor was the presence of mismatches between the forward primer and the template. The majority of mismatches occurred near the 3′ end of the primer (between positions 13 and 17), which has been shown to be detrimental for primer annealing (Bru *et al.*, 2008; Wu *et al.*, 2009). The other two factors were the presence of a G/C- vs. an A/T-base in the degenerate base of the forward and reverse primer-binding sites. This agrees with previous studies (e.g. Polz & Cavanaugh, 1998) and the strength of this bias is thought to increase with annealing temperature (Sipos *et al.*, 2007). Therefore, we used the lowest possible temperature still leading to a specific product. The best option would certainly be to avoid primers with degenerate bases (except A/T or G/C), but this is impossible for conventional PCR primers to combine with targeting a broad spectrum of archaeal and bacterial groups. We also investigated the effect of differing sequencing chemistry (FLX vs. Titanium) and reverse primers (1391R vs. 1492R) on amplicon generation from the same template. The resulting Bray–Curtis distance between them was the smallest between any datasets (Table S3, Fig. 3a), showing that the applied PCR protocol yielded a reproducible, stable estimation of community composition,

even though biased by preferential primer–template annealing, as discussed above.

Concluding remarks

Applied to samples from the Jan Mayen hydrothermal vent field, we have demonstrated how different community sequencing strategies can provide complementary and, in some cases, contrasting views of taxonomical composition and diversity. The systematically differing results obtained illustrate the importance of considering the study design carefully. The methods compared here are also complementary and we have shown that combining them can provide additional insights into the ecology of microorganisms at the hydrothermal vents studied.

Acknowledgements

We would like to thank Rolf Birger Pedersen, Ingunn Thorseth, the crew of the research vessel G.O. Sars and all participants of the CGB cruise 2008. Jørn Einen and Håkon Dahle are acknowledged for help during sampling. Christopher Quince and Christa Schleper are acknowledged for valuable discussions and help with data analysis; Ave Tooming Klunderud, Lex Nederbragt and others at the Norwegian High-Throughput Sequencing Centre for sequencing; and the High Performance Computing unit of the University of Bergen (Parallab) for keeping the machines running, allowing our sequence analysis to run smoothly. This work was supported by the Norwegian Research Council (Project number 179560).

References

Brazelton WJ, Ludwig KA, Sogin ML *et al.* (2010) Archaea and bacteria with surprising microdiversity show shifts in dominance over 1000-year time scales in hydrothermal chimneys. *P Natl Acad Sci USA* **107**: 1612–1617.

Bremer H & Dennis PP (1996) Modulation of chemical composition and other parameters of the cell by growth rate. *Escherichia coli and Salmonella: Cellular and Molecular Biology* (Neidhardt FC, ed), pp. 1553–1569. ASM Press, Washington, DC.

Bru D, Martin-Laurent F & Philippot L (2008) Quantification of the detrimental effect of a single primer-template mismatch by real-time PCR using the 16S rRNA gene as an example. *Appl Environ Microb* **74**: 1660–1663.

Campbell BJ, Engel AS, Porter ML & Takai K (2006) The versatile epsilon-proteobacteria: key players in sulphidic habitats. *Nat Rev Microbiol* **4**: 458–468.

Chao A (1987) Estimating the population size for capture-recapture data with unequal catchability. *Biometrics* **43**: 783–791.

Claesson MJ, O'Sullivan O, Wang Q *et al.* (2009) Comparative analysis of pyrosequencing and a phylogenetic microarray for exploring microbial community structures in the human distal intestine. *PLoS One* **4**: e6669.

Davis RE & Moyer CL (2008) Extreme spatial and temporal variability of hydrothermal microbial mat communities along the Mariana Island Arc and southern Mariana back-arc system. *J Geophys Res* **113**: B08S15.

Engelbrektson A, Kunin V, Wrighton KC, Zvenigorodsky N, Chen F, Ochman H & Hugenholtz P (2010) Experimental factors affecting PCR-based estimates of microbial species richness and evenness. *ISME J* **4**: 642–647.

Hamady M, Walker JJ, Harris JK, Gold NJ & Knight R (2008) Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. *Nat Methods* **5**: 235–237.

Huber JA, Welch DBM, Morrison HG, Huse SM, Neal PR, Butterfield DA & Sogin ML (2007) Microbial population structures in the deep marine biosphere. *Science* **318**: 97–100.

Huber JA, Cantin HV, Huse SM, Mark Welch DB, Sogin ML & Butterfield DA (2010) Isolated communities of *Epsilonproteobacteria* in hydrothermal vent fluids of the Mariana Arc seamounts. *FEMS Microbiol Ecol* **73**: 538–549.

Huson DH, Auch AF, Qi J & Schuster SC (2007) MEGAN analysis of metagenomic data. *Genome Res* **17**: 377–386.

Imachi H, Sakai S, Hirayama H, Nakagawa S, Nunoura T, Takai K & Horikoshi K (2008) *Exilispira thermophila* gen. nov., sp. nov., an anaerobic, thermophilic spirochaete isolated from a deep-sea hydrothermal vent chimney. *Int J Syst Evol Micr* **58**: 2258–2265.

Isenbarger TA, Finney M, Rìos-Velázquez C, Handelsman J & Ruvkun G (2008) Miniprimer PCR, a new lens for viewing the microbial world. *Appl Environ Microb* **74**: 840–849.

Jones SE & Lennon JT (2010) Dormancy contributes to the maintenance of microbial diversity. *P Natl Acad Sci USA* **107**: 5881–5886.

Kirchman DL (2002) The ecology of *Cytophaga–Flavobacteria* in aquatic environments. *FEMS Microbiol Ecol* **39**: 91–100.

Kirchman DL, Cottrell MT & Lovejoy C (2010) The structure of bacterial communities in the western Arctic Ocean as revealed by pyrosequencing of 16S rRNA genes. *Environ Microbiol* **12**: 1132–1143.

Kormas KA, Tivey MK, Damm KV & Teske A (2006) Bacterial and archaeal phylotypes associated with distinct mineralogical layers of a white smoker spire from a deep-sea hydrothermal vent site (9°N, East Pacific Rise). *Environ Microbiol* **8**: 909–920.

Lane DJ, Pace B, Olsen GJ, Stahl DA, Sogin ML & Pace NR (1985) Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. *P Natl Acad Sci USA* **82**: 6955–6959.

Legendre P & Legendre L (1998) *Numerical Ecology*. Elsevier, Amsterdam.

Liu Z, DeSantis TZ, Andersen GL & Knight R (2008) Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers. *Nucleic Acids Res* **36**: e120.

Luna GM, Manini E & Danovaro R (2002) Large fraction of dead and inactive bacteria in coastal marine sediments: comparison

of protocols for determination and ecological significance. *Appl Environ Microb* **68**: 3509–3513.

Margulies M, Egholm M, Altman WE *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**: 376–380.

McCarren J, Becker JW, Repeta DJ *et al.* (2010) Microbial community transcriptomes reveal microbes and metabolic pathways associated with dissolved organic matter turnover in the sea. *P Natl Acad Sci USA* **107**: 16420–16427.

Moyer CL, Dobbs FC & Karl DM (1995) Phylogenetic diversity of the bacterial community from a microbial mat at an active, hydrothermal vent system, Loihi Seamount, Hawaii. *Appl Environ Microb* **61**: 1555–1562.

Nakagawa S & Takai K (2008) Deep-sea vent chemoautotrophs: diversity, biochemistry and ecological significance. *FEMS Microbiol Ecol* **65**: 1–14.

Oksanen JF, Blanchet FG, Kindt R *et al.* (2010) *VEGAN: community ecology package*. Available at http://CRAN.R-project.org/package=vegan

Parks DH & Beiko RG (2010) Identifying biologically relevant differences between metagenomic communities. *Bioinformatics* **26**: 715–721.

Pedersen RB, Rapp HT, Thorseth IH *et al.* (2010) Discovery of a black smoker vent field and vent fauna at the Arctic Mid-Ocean Ridge. *Nat Commun* **1**: 126.

Pedrós-Alió C (2006) Marine microbial diversity: can it be determined? *Trends Microbiol* **14**: 257–263.

Perner M, Seifert R, Weber S *et al.* (2007) Microbial CO(2) fixation and sulfur cycling associated with low-temperature emissions at the Lilliput hydrothermal field, southern Mid-Atlantic Ridge (9°S). *Environ Microbiol* **9**: 1186–1201.

Plato & Cornford FM (1941) *The Republic*. "Allegory of the cave" from Book VII. (Cornford FM, ed), pp. 514A–521B. Oxford University Press, London.

Polz MF & Cavanaugh CM (1998) Bias in template-to-product ratios in multitemplate PCR. *Appl Environ Microb* **64**: 3724–3730.

Poulsen LK, Ballard G & Stahl DA (1993) Use of rRNA fluorescence *in situ* hybridization for measuring the activity of single cells in young and established biofilms. *Appl Environ Microb* **59**: 1354–1360.

Pruesse E, Quast C, Knittle K, Fuchs BM, Ludwig W, Peplies J & Glöckner FO (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res* **35**: 7188–7196.

Quince C, Lanzén A, Curtis TP, Davenport RJ, Hall N, Head IM, Read LF & Sloan WT (2009) Accurate determination of microbial diversity from 454 pyrosequencing data. *Nat Methods* **6**: 639–641.

Quince C, Lanzén A, Davenport RJ & Turnbaugh PJ (2011) Removing noise from pyrosequenced amplicons. *BMC Bioinformatics* **12**: 38.

Reysenbach A-L & Shock E (2002) Merging genomes with geochemistry in hydrothermal ecosystems. *Science* **296**: 1077–1082.

Rodriguez-Blanco A, Ghiglione J-F, Catala P, Casamayor EO & Lebaron P (2009) Spatial comparison of total vs. active bacterial populations by coupling genetic fingerprinting and clone library analyses in the NW Mediterranean Sea. *FEMS Microbiol Ecol* **67**: 30–42.

Roesch LFW, Fulthorpe RR, Riva A *et al.* (2007) Pyrosequencing enumerates and contrasts soil microbial diversity. *ISME J* **1**: 283–290.

Schander C, Rapp HT, Kongsrud JA *et al.* (2010) The fauna of hydrothermal vents on the Mohn Ridge (North Atlantic). *Mar Biol Res* **6**: 155–171.

Sievert SM, Scott KM, Klotz MG *et al.* (2008) Genome of the epsilonproteobacterial chemolithoautotroph *Sulfurimonas denitrificans*. *Appl Environ Microb* **74**: 1145–1156.

Simpson EH (1949) Measurement of diversity. *Nature* **163**: 688.

Sipos R, Székely AJ, Palatinszky M, Révész S, Márialigeti K & Nikolausz M (2007) Effect of primer mismatch, annealing temperature and PCR cycle number on 16S rRNA gene-targetting bacterial community analysis. *FEMS Microbiol Ecol* **60**: 341–350.

Sogin ML, Morrison HG, Huber JA *et al.* (2006) Microbial diversity in the deep sea and the underexplored 'rare biosphere'. *P Natl Acad Sci USA* **103**: 12115–12120.

Ståhlberg A, Håkansson J, Xian X, *et al.* (2004) Properties of the reverse transcription reaction in mRNA quantification. *Clin Chem* **50**: 509–515.

Suzuki MT & Giovannoni SJ (1996) Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. *Appl Environ Microb* **62**: 625–630.

Tang H & Pedersen B (2010) *Fisher*. Available at http://github.com/brentp/fishers_exact_test

Teske A, Hinrichs K-U, Edgcomb V *et al.* (2002) Microbial diversity of hydrothermal sediments in the Guaymas Basin: evidence for anaerobic methanotrophic communities. *Appl Environ Microb* **68**: 1994–2007.

Tringe SG & Hugenholtz P (2008) A renaissance for the pioneering 16S rRNA gene. *Curr Opin Microbiol* **11**: 442–446.

Urich T, Lanzén A, Qi J, Huson DH, Schleper C & Schuster SC (2008) Simultaneous assessment of soil microbial community structure and function through analysis of the meta-transcriptome. *PLoS One* **3**: e2527.

Vanreusel A, Groote AD, Gollner S & Bright M (2010) Ecology and biogeography of free-living nematodes associated with chemosynthetic environments in the deep sea: a review. *PLoS One* **5**: e12449.

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 Microb* **73**: 5261–5267.

Wu J-H, Hong P-Y & Liu W-T (2009) Quantitative effects of position and type of single mismatch on single base primer extension. *J Microbiol Meth* **77**: 267–275.

Yamamoto M, Nakagawa S, Shimamura S, Takai K & Horikoshi K (2010) Molecular characterization of inorganic sulfur-compound

metabolism in the deep-sea epsilonproteobacterium *Sulfurovum* sp. NBC37-1. *Environ Microbiol* **12**: 1144–1153.
Zhou J, Bruns MA & Tiedje JM (1996) DNA recovery from soils of diverse composition. *Appl Environ Microb* **62**: 316–322.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Fig. S1. Comparison of the sensitivity and agreement at the different taxonomical ranks, given in share of reads classified and the percent overlap between the RDP CLASSIFIER and the method used in this study (MEGAN assignment using the modified Silva SSURef100 reference database).
Fig. S2. Relative abundance plotted vs. total OTU richness, from total assignments across all amplicon datasets.
Fig. S3. Rarefaction analysis of the number of OTUs found in each of the datasets.
Table S1. Number of assigned reads, for each dataset at all taxonomic ranks from domain to genus.
Table S2. Relative abundance for each dataset at all taxonomic ranks from domain to genus.
Table S3. Matrix of pairwise Bray–Curtis dissimilarities between the datasets, based on relative taxon abundances.
Table S4. Comparison of relative abundance between sequencing of cDNA and DNA-amplicons in Samples 1 and 2.
Table S5. Comparison of relative abundance between sequencing of amplicons and shotgun fragments.
Appendix S1. File primer mismatch (FASTA file). All variants of the three degenerate primer sequences plus all possible transformations of the variants whose reverse complement has exactly one or two mismatches to one of the correct primer sequence variants.