environmental microbiology

Environmental Microbiology (2014) **16**(9), 2672–2681 doi:10.1111/1462-2920.12308

Ecological coherence of diversity patterns derived from classical fingerprinting and Next Generation Sequencing techniques

Angélique Gobet,[1,2†] Antje Boetius[1,3] and Alban Ramette[1*]

[1]*HGF-MPG Group for Deep Sea Ecology and Technology, Max Planck Institute for Marine Microbiology, Bremen, Germany.*

[2]*Jacobs University Bremen GmbH, Bremen, Germany.*

[3]*Alfred Wegener Institute for Polar and Marine Research, Bremerhaven, Germany.*

Summary

Changes in richness and bacterial community structure obtained via 454 Massively Parallel Tag Sequencing (MPTS) and Automated Ribosomal Intergenic Analysis (ARISA) were systematically compared to determine whether and how the ecological knowledge obtained from both molecular techniques could be combined. We evaluated community changes over time and depth in marine coastal sands at different levels of taxonomic resolutions, sequence corrections and sequence abundances. Although richness over depth layers or sampling dates greatly varied [~ 30% and 70–80% new operational taxonomic units (OTU) between two samples with ARISA and MPTS respectively], overall patterns of community variations were similar with both approaches. Alpha-diversity estimated by ARISA-derived OTU was most similar to that obtained from MPTS-derived OTU defined at the order level. Similar patterns of OTU replacement were also found with MPTS at the family level and with 20–25% rare types removed. Using ARISA or MPTS datasets with lower resolution, such as those containing only resident OTU, yielded a similar set of significant contextual variables explaining bacterial community changes. Hence, ARISA as a rapid and low-cost fingerprinting technique represents a valid starting point for more in-depth exploration of community composition when complemented by the detailed taxonomic description offered by MPTS.

Received 11 July, 2013; accepted 13 October, 2013. *For correspondence. E-mail aramette@mpi-bremen.de; Tel. (+49) 421 2028 863; Fax (+49) 421 2028 690. †Present address: Marine Glycobiology group, UMR 7139 CNRS-UPMC, Station Biologique de Roscoff, Place Georges Teissier, CS 90074, 29680 Roscoff Cedex, France. E-mail angelique.gobet@sb-roscoff.fr.

Introduction

Long DNA sequences (e.g. 16S rRNA genes) derived from clone library-based approaches have originally been used to describe microbial diversity (Olsen *et al*., 1986; Zinger *et al*., 2012). These techniques are time consuming and rather expensive, especially if the aim is to process the many samples required for a robust statistical description of both structure and dynamics of microbial communities in their environmental context, including spatial and temporal variation (Zinger *et al*., 2011). Molecular fingerprinting techniques [e.g. terminal restriction fragment length polymorphism (T-RFLP, Avaniss-Aghajani *et al*., 1994), automated rRNA gene intergenic spacer analysis (ARISA, Fisher and Triplett, 1999)] represent tools of choice for the rapid, reproducible and cost-effective processing of many environmental samples (Fisher and Triplett, 1999), and have significantly contributed to advancing microbial community ecology (Zinger *et al*., 2012). ARISA targets the Intergenic Transcribed Spacer (ITS) between the 16S and the 23S rRNA gene regions (Fisher and Triplett, 1999), which is highly variable in nucleotide sequence and length [e.g. from 60 bp to 1529 bp (Gürtler and Stanisich, 1996)]. ARISA may produce hundreds of fluorescence intensity profiles (Cardinale *et al*., 2004) where each individual peak may correspond to one or several phylotypes (Crosby and Criddle, 2003; Yannarell and Triplett, 2005). Consequently, although suitable to study community changes over time, space or along environmental gradients, the technique does not allow determining neither the number of microbial types in a given sample (Bent and Forney, 2008) nor their taxonomy (Fisher and Triplett, 1999; Brown *et al*., 2005).

Although traditional sequence library-based techniques have already described a substantial fraction of microbial diversity, the major part of it has yet escaped our sampling efforts, and even large 16S rRNA clone libraries highly underestimate microbial diversity (Curtis and Sloan, 2005; Quince *et al*., 2008). For instance, Sanger sequencing of

coastal waters retrieved 516 unique operational taxonomic units (OTU), while the estimated richness reached 1633 OTU (Acinas *et al*., 2004). The advent of next generation sequencing (NGS) has revolutionized microbial ecology by giving a more comprehensive description of microbial diversity in any given sample. For instance, 454 massively parallel tag sequencing (MPTS) and Illumina may produce thousands to hundreds of thousands short variable sequences of the 16S rRNA gene per sample, which can be further taxonomically classified (Sogin *et al*., 2006; Degnan and Ochman, 2012). Using MPTS, about 4000 to 20 000 OTU were obtained in the pelagic realm (for 14 to 194 samples) and 2000 to 59 000 OTU per sample were obtained in the benthos [for 13 to 72 samples, (Zinger *et al*., 2011)].

Even though a deeper coverage of microbial community diversity is obtained, MPTS data output has to be analysed with care due to the presence of PCR and sequencing artifacts such as chimera and homopolymers, which may inflate microbial diversity estimates (Kunin *et al*., 2010). Consequently, several studies have provided various ways to trim and correct sequences (Quince *et al*., 2009; Kunin *et al*., 2010). Notably, different cyclic or seasonal microbial community patterns have been observed when using T-RFLP vs. 454 MPTS approaches (Gilbert *et al*., 2009), whereas similar bacterial community patterns could be observed along gradients in water depth when applying ARISA and 454 MPTS to polar deep-sea sediments (Bienhold *et al*., 2012). Fingerprinting techniques, which have a lower resolution than NGS, may exclude a significant proportion of rarer taxa, while NGS derived data may potentially provide an inflated number of spurious taxa. In both cases, the resulting community patterns and further ecological interpretations could be seriously impacted by those technical limitations.

Here, we systematically compared results obtained by high and low resolution molecular techniques (454 MPTS vs. ARISA) applied to bacterial communities from temperate coastal sediments (North Sea island Sylt, Germany). Those communities have previously been particularly well characterized using 16S rRNA-based libraries and fluorescence *in situ* hybridization (Musat *et al*., 2006). The application of ARISA further helped describe large depth-related and temporal patterns (Böer *et al*., 2009), and the application of 454 MPTS on the same DNA extracts provided a high-resolution description of the fluctuations of rare and resident OTU (Gobet *et al*., 2012). In this study, we compare patterns and their ecological interpretation by systematically taking into account varying levels of taxonomic classification, data correction and increasing removal of the rarest OTU in the datasets. Based on the different levels of resolution offered by the two approaches, we expect that OTU richness and OTU replacement between samples over depth and time obtained by MPTS are drastically different from that obtained by fingerprinting techniques, but that patterns of changes in richness and community structure are conserved.

Results and discussion

Local bacterial richness in temperate coastal sands as described by ARISA and NGS

The application of ARISA gave 306 different OTU [hereafter indicated as OTU_{ARISA}, which correspond to binned ARISA peaks (Böer *et al*., 2009)], with 100–202 unique OTU_{ARISA} per sample. The application of 454 MPTS generated 197 685 sequences in total, corresponding to 27 630 OTU_{unique} (two sequences are considered as belonging to two different OTU_{unique} when they differ by at least one bp), with a range of 1042–5577 OTU_{unique} identified per sample. Hence, as in previous studies, ARISA fingerprinting led to about a 10–55 times lower number of OTU per sample than the application of an NGS on the same extracted DNA (Roesch *et al*., 2009; Koopman *et al*., 2010; Bienhold *et al*., 2012).

In datasets produced by either ARISA or MPTS, the average number of OTU per sampling date did not change with time. However, OTU numbers increased with sediment depth (Fig. S1A and B), as observed in other coastal sediments by using T-RFLP (Urakawa *et al*., 1999). Numbers in the top 0–5 cm layer were clearly different from those of the deeper layers 5–10 cm and 10–15 cm with both molecular techniques (Student's *t*-tests, $P < 0.05$). Notably, patterns of community structure in the mid 5–10 cm layer differed between each technique (Fig. S1A): Samples from 0–5 cm were similar to those from 5–10 cm with ARISA, whereas samples from 5–10 cm were more similar to those from 10–15 cm with MPTS. Additionally, the total number of OTU, Chao and abundance-based coverage estimator (ACE) estimators from ARISA and different 454 MPTS datasets were significantly different due to the sensitivity of such indices to rare OTU (Fig. S2A–C).

The population detection limit of ARISA has been estimated to 10^3 cells per ml of sample as for any PCR-based technique (Ramette, 2009), which is valid for MPTS of an amplicon pool that also relies on PCR amplification. Some of the OTU_{ARISA} may not be distinguished as they probably represent several bacterial types with similar ITS length (Crosby and Criddle, 2003). Furthermore, rare bacterial types may not be represented (Bent and Forney, 2008), because the technique relies on chromatography, the data output is generally processed by only considering specific peak characteristics (e.g. peak intensity above 50 fluorescence units and sizes between 100–1000 bp length) and peak calling imprecisions are resolved by binning the data in a limited range of possible OTU [see details in (Brown

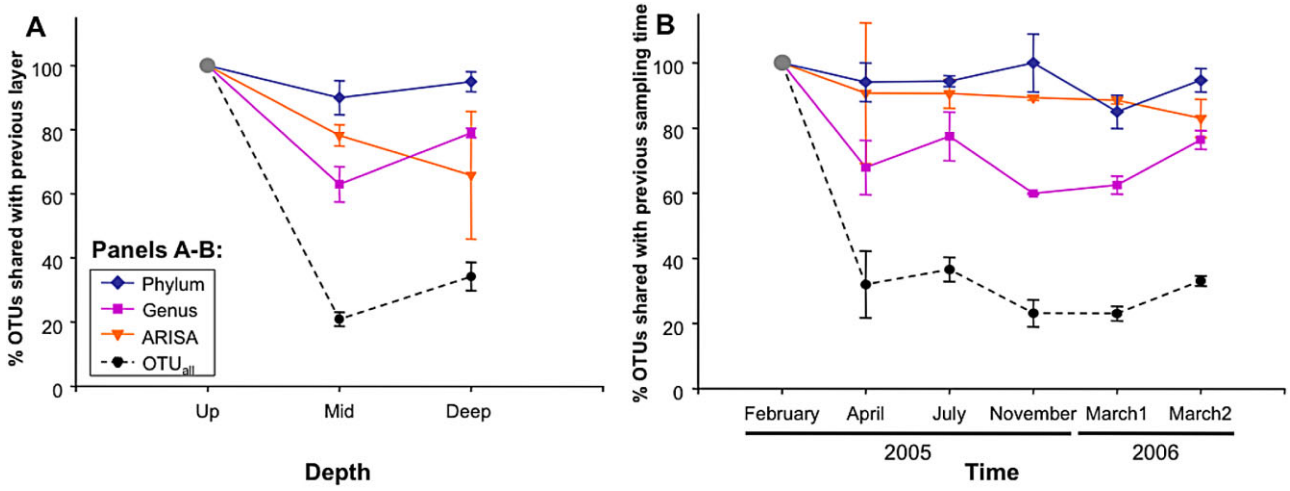

Fig. 1. Turnover of the bacterial community between consecutive (A) depth layers or (B) sampling dates. The percentage of shared OTU was calculated between two successive sampling depth layers (or sampling dates). The community turnover was compared between datasets at different taxonomic resolution levels [i.e. phylum, genus and OTU_{all}; here, we chose to use the OTU_{all} community turnover from a previous study (Gobet *et al.*, 2012) for a direct comparison between ARISA and MPTS data] and the ARISA dataset. OTU_{all} represents the original dataset with all OTU, used here as a reference to study the effects of the taxonomic classification of OTU on the interpretation of the dynamics of the bacterial community. Standard deviation bars are calculated over 4–6 sampling dates (A) and over three depth layers (B), except for July and November 2005 where two depth layers were considered. The top layer and the first sampling date (February 2005) are indicated by a grey point as 100% of shared OTU with themselves.

et al., 2005; Böer *et al.*, 2009)]. In contrast, the MPTS approach data offers a deeper description of the pool of PCR amplicons obtained from microbial communities, including many rare and less abundant OTU. The short V6 sequences targeted in our study should furthermore better resolve microbial diversity than larger sequences (Huber *et al.*, 2009). However, alpha-diversity indices (Shannon's index and Simpson's evenness) calculated from ARISA were highly correlated with those obtained from MPTS data at most levels of resolution, including taxonomic levels from class to genus, PyroNoise-corrected data and datasets with different proportions of rare OTU removed. The MPTS level of resolution indicating highest correlation with alpha-diversity from ARISA was the order level (Fig. S2).

Comparison of bacterial community turnover over depth and time

The community turnovers predicted by ARISA and MPTS were statistically compared to determine whether the two molecular techniques lead to similar ecological conclusions. While OTU_{all} (i.e. dataset considering all sequences before clustering) showed only 19–34% shared OTU_{unique} between two depth layers or any two sampling dates, the similarity of the bacterial community detected by ARISA was much higher, with 66–78% and 70–91% shared OTU_{ARISA} between two depth layers and between sampling times respectively (Fig. 1 and Fig. S3). Microbial community turnovers over depth and time were thus highly different between ARISA and original 454 MPTS datasets including the rare biosphere.

Even when excluding pyrosequencing artifacts, the large proportion of singletons may lead to an overestimation of the observed dynamics of the bacterial community at the OTU_{all} level (Quinlan *et al.*, 2008). Despite applying the PyroNoise algorithm and different levels of clustering, only about 20–40% of shared OTU in the bacterial community were present in all depth layers or at all sampling times (Fig. S3). If we consider that 3% sequence difference threshold for defining OTU roughly corresponds to cut-off levels defining bacterial species level (Stackebrandt and Goebel, 1994), patterns observed with PyroNoise-corrected data presented a continuum of taxonomic resolutions up to the genus level (Figs S3 and S4). The taxonomic assignment of MPTS-based OTU allowed comparing the amount of shared OTU_{unique} at successive taxonomic levels with the turnover from the ARISA dataset. When performing the analyses from the genus to the phylum levels, 63% to 97% OTU_{unique} were shared over sediment depth, and 54% to 100% OTU_{unique} were shared over time respectively (Fig. 1 and Fig. S4). Interestingly, the turnover in the ARISA dataset was more similar with that of the family level (Figs S3 and S4).

To test whether ARISA mostly detects abundant members of the microbial community, its community turnover values were compared to successively truncated subsets of the 454 OTU_{all} with increasing proportions of the rarest OTU_{unique} removed (Fig. S5). Removal of rare OTU_{unique} led to a decrease in community turnover

because of the low proportion of abundant types in the community (Gobet *et al.*, 2012). Despite the biases implied by each molecular technique, the amount of shared OTU_{ARISA} over depth or time seemed to indicate similar microbial community turnover as obtained when removing 20–25% of low abundant OTU_{unique} in the OTU_{all} dataset (Fig. 1, Figs S3 and S5). Additionally, when comparing the turnover at successive taxonomic levels or percentages of rarest OTU_{unique} removed, some patterns were found to be similar (Figs S4 and S5): For instance, it seemed that the turnover after removing 15% rare OTU_{unique} corresponds to the genus level or that the removal of 30% rare OTU_{unique} would lead to a similar turnover as the class or phylum levels (Figs S4 and S5). This is explained by the loss of community resolution at broader taxonomic levels, where the patterns of many different types are lumped together, and by the fact that most rare OTU_{unique} were not identified at genus to phylum levels. This overall supports the idea of high consistency of the ecological information obtained from beta-diversity analyses at various taxonomic levels, as also observed in global benthic and pelagic marine realms (Zinger *et al.*, 2011).

We then performed a systematic comparison of community structure for each dataset, assessing changes in community dissimilarity between samples based on Bray–Curtis dissimilarity index. The resulting matrices were correlated with each other using Pearson's correlation coefficient, and the resulting correlation coefficients were tested for significance (Fig. 2). Overall, the comparison of the community structure showed little variation at different taxonomic levels, after correcting for pyrosequencing noise, or after truncating the datasets by removing successive proportions of the rare fraction (Fig. 2) – as observed in deep-sea sediments from the Arctic (Bienhold *et al.*, 2012). The ARISA dataset structure was most similar to datasets truncated from 25–50% of their rarest OTU_{unique} (Fig. 2). This was also supported by significant Bonferroni corrected Mantel tests between ARISA data and resident OTU_{unique} (i.e. present at all times) with a correlation coefficient of 0.43 (data not shown). Therefore, main community patterns stayed consistent regardless of the chosen level of data resolution.

When patterns of community variation were visualized by non-metric multidimensional scaling (NMDS), similar depth-related patterns of the microbial community were obtained for all datasets produced by ARISA or MPTS (Fig. 3). These observations were also confirmed by analysis of similarity (ANOSIM), testing for differences between sampling depth layers ($R > 0.3$, P value ≤ 0.01, Fig. 3). When comparing the obtained NMDS ordinations by Procrustes correlation, a similar picture emerged: Sample ordinations based on ARISA data were highly correlated with those obtained from MPTS data when considering resident OTU_{unique} (not shown), OTU_{all} dataset and PyroNoise-corrected dataset at 3% clustering, with R values reaching 0.79, 0.88 and 0.74 respectively (Fig. S6). Together with the results on community turnover, we can conclude that patterns derived from ARISA data are consistent with the patterns of the most dominant microbial types in the community.

Ecological modelling of changes in community structure

By investigating the relationships between shifts in bacterial community structure and concomitant changes in environmental parameters, new insights into bacterial community ecology in temperate coastal sands have previously been obtained (Musat *et al.*, 2006; Böer *et al.*, 2009; Gobet *et al.*, 2012). Because different molecular techniques were used, it is not certain that the interpretations of the resulting ecological models are directly comparable with each other: Indeed, each molecular technique may best describe microbial communities at a specific phylogenetic resolution level, and this entails that only the effects of ecological factors specifically acting at those levels may be detected. Therefore, variations in community structure in the ARISA and the 454 MPTS datasets were more finely disentangled by applying a multivariate variation partitioning approach (Legendre and Gallagher, 2001; Ramette and Tiedje, 2007), using contextual parameters that included cell abundance, biogeochemical gradients (i.e. pigments, nutrients), extracellular enzyme activity and their combined effects.

Interestingly, the same combinations of significant biogeochemical variables could explain datasets that had a similar degree of phylogenetic resolution. Indeed, almost the same environmental model as for the taxonomically assigned MPTS dataset explained the biological variation in the ARISA dataset (Table S1). A model containing salinity, pigments, the same nutrients and extra-cellular enzymes, as well as cell abundance, explained 51–75% of the biological variation from the genus to the phylum level in sandy sediments (Fig. 4, Table S1). A similar environmental model applied to a more complex dataset (i.e. chlorophyll *a*, extra-cellular phosphatase activity, cell abundance) explained 14–20% of biological variation in the $OTU_{annotated}$ (i.e. fully taxonomically assigned sequences), the raw OTU_{all} and after PyroNoise-correction and clustering at 0% and 3% sequence difference (Table S1, Fig. 4). Noticeably, some truncated datasets were explained by similar environmental models as for datasets defined at specific taxonomic levels (Fig. 4, Fig. S7, Table S1): The same combination of environmental parameters could explain variation at the genus level and for the OTU_{all} dataset without 30% rarest OTU_{unique} (Fig. 4, Figs S1–S7, Table S1). Patterns observed at the family to the phylum levels corresponded

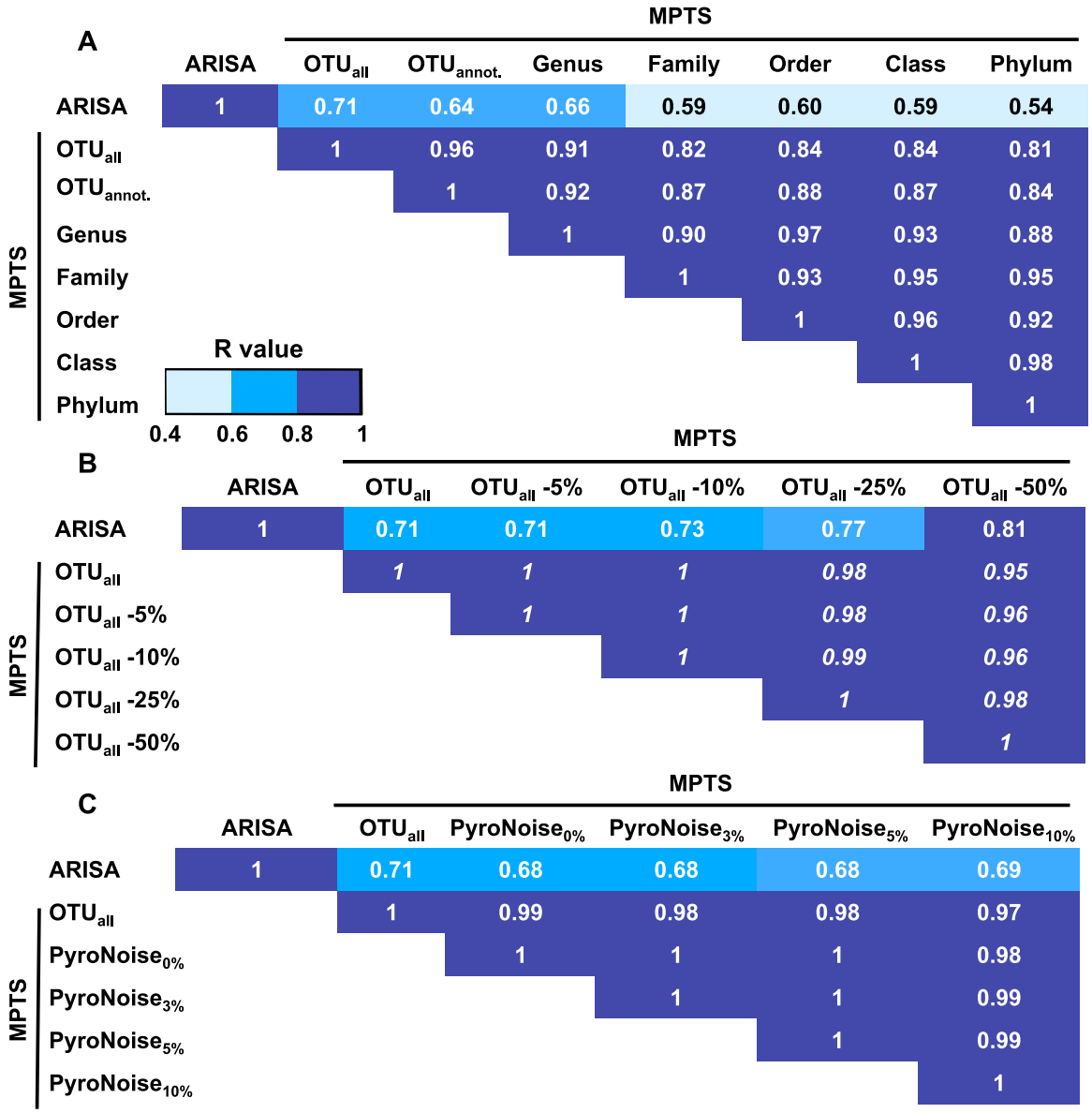

Fig. 2. Comparison of the structure of modified datasets. Pearson's correlation coefficient was used to compare the bacterial community structure present in the ARISA and the OTU_{all} datasets (A) at various levels of taxonomic annotation for the MPTS data, (B) after successive removal of rare OTU and (C) after successive clustering of PyroNoise-corrected data to define OTU. The correlation coefficient was calculated from the distance matrices resulting from the relative sequence abundances. Significances of the correlation were determined by Mantel tests with 1000 matrix permutations. All R values were significant after Bonferroni correction for multiple testing. In (B), values in italic indicate simple Pearson correlations of the truncated matrices, without a test of significance, as the truncated matrices are not statistically independent from each other (see Experimental procedures).

well to ecological patterns obtained with the OTU_{all} dataset without 35–50% rarest OTU_{unique} (Fig. 4, Figs S1–S7, Table S1).

Dynamic coastal sand bacterial communities showed consistent ecological patterns across subsets of the MPTS dataset investigated at different taxonomic resolutions (Gobet *et al.*, 2010; 2012). Here, comparing ARISA data and MPTS data, consistency was observed at intermediate taxonomic resolution levels: For instance, the amount of variation in ARISA data explained by the environmental parameters was similar to that obtained when considering $OTU_{Resident}$, the genus level, or truncated datasets without 30% rarest OTU ($R^2 = 51\%$, 55%, 51% and 55% respectively; Fig. 4, Table S1). Such findings may be explained by the presence of resident bacteria, which often were associated with large number of sequences, and thus dominated sub-datasets defined at different resolution levels (e.g. genus to phylum,

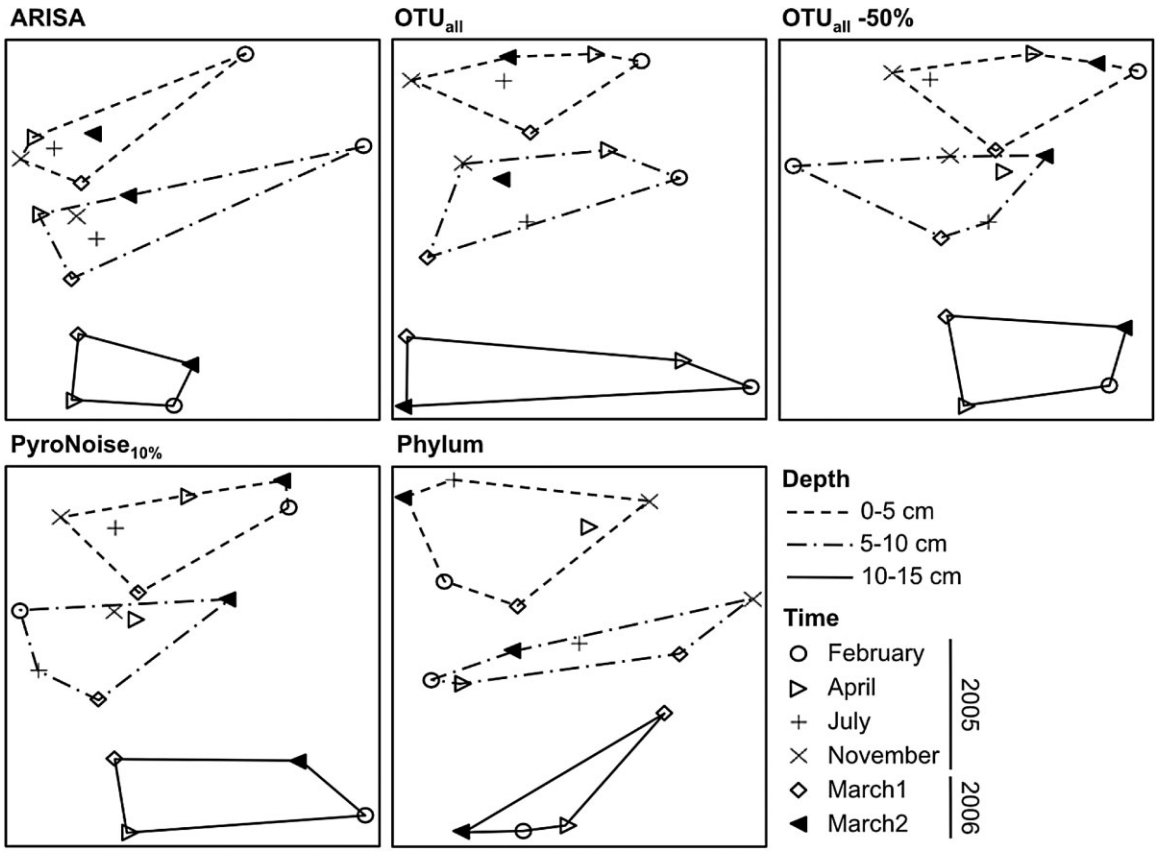

Fig. 3. Comparison of community patterns obtained from ARISA and 454 MPTS datasets. Those examples of NMDS ordination (based on Bray–Curtis distance matrices) are based on the relative abundance datasets from ARISA (stress = 0.062), the original OTU_{all} dataset (stress = 0.081), the OTU_{all} dataset with 50% of rarest OTU removed (stress = 0.095), the PyroNoise-corrected data at 10% sequence dissimilarities (stress = 0.083) and the phylum level for 454 MPTS data (stress = 0.047). Significant differences between sampling depth layer could be observed for all but the phylum level, using the ANOSIM followed by Bonferroni correction for multiple testing.

truncated datasets). This dominance across taxonomic ranks may lead to ecologically and functionally coherent patterns. Additionally, each molecular technique may target a specific range of that continuum of diversity, and NGS techniques that offer deeper insights into the rare biosphere could also lead to describing ecologically neutral variations or variations than cannot be assigned to the observed variation in the measured contextual parameters.

In conclusion, classical molecular fingerprinting approaches, as illustrated here by the high-throughput, low-cost method ARISA, are very well suited for a general overview of changes in abundant bacterial types. Our study shows that ARISA fingerprinting may also serve as a tool to track shifts in dominant and resident members of the community, and that the data output may be easier to process than larger datasets produced by MPTS. Large MPTS datasets provide, however, a deeper and direct insight into community composition, and may thus complement a shallower description of community shifts obtained across multiple samples. This combination of techniques can bridge gaps in the assessment of community patterns at various taxonomic and rarity levels. Importantly, we demonstrate that the ecological knowledge gained from classical microbial community studies, which were mostly based on low resolution community fingerprinting tools (e.g. T-RFLP, ARISA), is not obsolete, and can be further extended by the application of the new generation of high-throughput sequencing tools.

Experimental procedures

Sampling procedures

In February, April, July and November 2005, beginning and end of March 2006, sediment push cores were collected at low tide on the shallow subtidal sandy area of the island Sylt (55°00′47.7″N, 8°25′59.3″E, North Sea, Germany). Cores were sectioned every 5 cm down to 15 cm, and the sections were directly processed or stored at − 4°C or − 20°C until DNA extraction and environmental measurements were made. Additional environmental parameters from long-term records were included as well (Böer *et al.*, 2009).

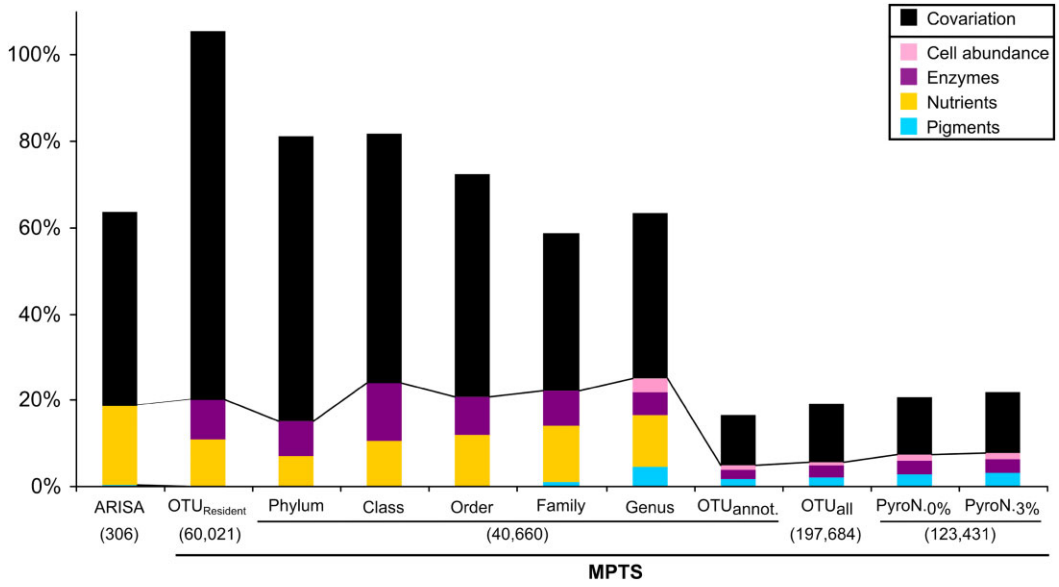

Fig. 4. Ecological interpretation of betadiversity patterns derived from ARISA and 454 MPTS. Environmental parameters accounted for include pigments (chlorophyll a and pheophytin), nutrients (silicate, phosphate, nitrite, nitrate, ammonium), extra-cellular enzyme activities (chitinase, α-glucosidase, β-glucosidase, lipase, aminopeptidase, phosphatase), cell abundance and their combined effects (as shown in black above the line). $OTU_{Resident}$ are OTU always present in the dataset, $OTU_{annot.}$ consists of sequences with a complete annotation from the phylum to the genus level (i.e. 20% of the total number of sequences in the original dataset), while OTU_{all} includes all sequences. $PyroN_{0\%}$ and $PyroN_{3\%}$ represent the PyroNoise-corrected OTU_{all} dataset, clustered at 0% and 3% of sequence identity respectively. The total number of sequences or OTU in each dataset is given in parentheses. White stars indicate pure factors that significantly explain the biological variation (P value ≤ 0.05) after 1000 Monte Carlo permutations. Variation partitioning for $OTU_{Resident}$, Phylum and OTU_{all} from Gobet *et al.* (2012) were represented here for comparison with variation partitioning from ARISA and other 454-derived datasets.

Community structure analysis

DNA from 16 sandy samples was extracted and purified as described earlier (Böer *et al.*, 2009). The same DNA templates were used to analyse the bacterial community structure samples by automated rRNA intergenic spacer analysis [(Fisher and Triplett, 1999), see (Böer *et al.*, 2009) for details] and 454 MPTS (for details see Gobet *et al.*, 2010; 2012). The analysis of ARISA profiles was done using the GeneMapper Software v 3.7 (Applied Biosystems, Carlsbad, CA, USA). Fragments above a threshold of 50 fluorescence units and between 100–1000 bp length were considered, and a binning strategy with a bin size of 2 bp was applied (for details see Böer *et al.*, 2009). MPTS data were obtained from the publicly available 'Visualization and Analysis of Microbial Populations Structure (VAMPS)' website (http://vamps.mbl.edu/; project AB_SAND_Bv6). Barcode, primer and low-quality sequences were removed, as reported earlier (Huse *et al.*, 2007). Taxonomic annotation of the sequences has been carried out with an automatic annotation pipeline (Sogin *et al.*, 2006), using several known databases (Entrez Genome, RDP, SILVA).

Data analyses

Datasets. In this study, analyses were performed by defining OTU (Operational Taxonomic Units) either as ITS phylotype (OTU_{ARISA}), or as unique 454 MPTS sequences (OTU_{unique}). For the 454 MPTS datasets, the following subsets were considered: (i) all unassigned sequences (OTU_{all}), (ii) the fully taxonomically assigned sequences (i.e. from phylum to genus levels and the corresponding $OTU_{annotated}$ level, each dataset representing 20% of the original OTU_{all} dataset), (iii) the PyroNoise-corrected data clustered at different percentages of sequence difference (0%, 3%, 5% and 10% sequence difference) (iv) the Multivariate Cutoff Level Analysis (MultiCoLA)-truncated datasets, consisting of the original OTU_{all} dataset without successive proportions of rarer OTU_{unique} [i.e. OTU_{unique} with number of sequences lower than a given cut-off are removed (Gobet *et al.*, 2010)].

Variation in bacterial diversity. OTU numbers from the ARISA, OTU_{all} and $PyroNoise_{3\%}$ datasets were compared by pairwise Student's *t*-tests (non-parametric tests yielded the same results; data not shown). Observed richness, diversity indices (Shannon index and Simpson's evenness) and richness estimators (Chao, ACE) were calculated after 1000 resampled sets of each community matrix [ARISA data, OTU_{all}, the fully assigned sequences, the PyroNoise-corrected and the truncated datasets; (Gobet *et al.*, 2010)]. Correlations between indices were calculated using the Pearson's product moment coefficient, and their significance was corrected for multiple testing by using the Bonferroni method. Finally, the proportion of shared OTU between either two sampling dates or two depth layers was calculated for all community matrices.

Pairwise distance matrices were calculated from the relative abundance data (ARISA and 454 MPTS datasets) using the Bray–Curtis dissimilarity index (Bray and Curtis, 1957). The dissimilarity matrices were then compared with

Pearson's product moment correlation coefficient and assessed for significance using the Mantel test (Pearson, 1901). The significance of Pearson's correlations between dissimilarity matrices from MultiCoLA-truncated datasets could not be assessed by Mantel tests because testing correlations is only valid when variables (here matrices) are independent from each other (Legendre and Legendre, 1998; Legendre *et al*., 2005).

Non-metric multidimensional scaling (NMDS; Gower, 1966) was applied to the distance matrices to explore the variation in community structure. The similarity between NMDS ordination results from the ARISA, and 454 MPTS datasets was then calculated by applying Procrustes rotation (Gower, 1966). The Procrustes approach quantifies the degree of agreement between two NMDS ordinations, producing R values ranging from 0 to 1 [a score closer to 1 indicates highest similarities between the NMDS results (Shepard, 1966)]. The microbial community composition from the three depth layers was compared and tested by ANOSIM (Clarke, 1993)].

Relationships between the structuring of the microbial community and the environment. In a previous study, multivariate regression approaches were applied to test the relationships between the variation of measured environmental parameters [salinity, pigments, nutrients, extracellular enzymatic activities and cell properties (Gobet *et al*., 2012)]. All explanatory variables (except salinity) were $\log_{10}$-transformed before describing the microbial community distribution in the Hellinger-transformed matrices (ARISA data, OTU_{all}, the fully assigned sequences, the PyroNoise-corrected data and the MultiCoLA-trimmed data). To reduce multicollinearity among environmental variables, environmental parameters were selected according to a stepwise selection procedure [based on 999 Monte Carlo permutation tests and Akaike Information Criterion (AIC)] before modelling the biological variation. Consequently, we obtained the best-fitting models that could significantly explain the variation in the Hellinger-transformed (Legendre and Legendre, 1998; Legendre and Gallagher, 2001) community tables. The effects of pure environmental variables (pigments, nutrients, extra-cellular enzymatic activities, cell abundance) selected previously and their covariation on microbial community structure were then tested by canonical variation partitioning (Legendre and Gallagher, 2001).

Statistical analyses were carried out using the R statistical environment [R version 3.0 (R Development Core Team, 2013)], including the *vegan* package (Oksanen *et al*., 2013) and custom R scripts (MultiCoLA; Gobet *et al*., 2010).

Acknowledgements

We acknowledge Lucie Zinger, Gunter Wegener and two anonymous reviewers for helpful advices and comments on the manuscript. This work was supported by the Marie Curie Early Stage Training fellowship in Marine Microbiology [MarMic EST contract MEST-CT-2004–007776 to A.G.] and by the International Max Planck Research School of Marine Microbiology. [A.G.], as well as by the Max Planck Society [A.R., A.B.], Helmholtz Association [A.B.] and Leibniz Program of the DFG [A.B]. The sequencing was financed by a Keck Foundation grant to the International Census of Marine Microbes led by Mitchell L. Sogin and Linda A. Amaral-Zettler. This is a contribution to the International Census of Marine Microbes (ICoMM).

References

Acinas, S.G., Klepac-Ceraj, V., Hunt, D.E., Pharino, C., Ceraj, I., Distel, D.L., and Polz, M.F. (2004) Fine-scale phylogenetic architecture of a complex bacterial community. *Nature* **430:** 551–554.

Avaniss-Aghajani, E., Jones, K., Chapman, D., and Brunk, C. (1994) A molecular technique for identification of bacteria using small-subunit ribosomal-RNA sequences. *Biotechniques* **17:** 144–149.

Bent, S.J., and Forney, L.J. (2008) The tragedy of the uncommon: understanding limitations in the analysis of microbial diversity. *ISME J* **2:** 689–695.

Bienhold, C., Boetius, A., and Ramette, A. (2012) The energy-diversity relationship of complex bacterial communities in Arctic deep-sea sediments. *ISME J* **6:** 724–732.

Böer, S.I., Hedtkamp, S.I.C., van Beusekom, J.E.E., Fuhrman, J.A., Boetius, A., and Ramette, A. (2009) Time- and sediment depth-related variations in bacterial diversity and community structure in subtidal sands. *ISME J* **3:** 780–791.

Bray, J.R., and Curtis, J.T. (1957) An ordination of the upland forest communities of Southern Wisconsin. *Ecol Monogr* **27:** 326–349.

Brown, M.V., Schwalbach, M.S., Hewson, I., and Fuhrman, J.A. (2005) Coupling 16S-ITS rDNA clone libraries and automated ribosomal intergenic spacer analysis to show marine microbial diversity: development and application to a time series. *Environ Microbiol* **7:** 1466–1479.

Cardinale, M., Brusetti, L., Quatrini, P., Borin, S., Puglia, A.M., Rizzi, A., *et al*. (2004) Comparison of different primer sets for use in automated ribosomal intergenic spacer analysis of complex bacterial communities. *Appl Environ Microbiol* **70:** 6147–6156.

Clarke, K.R. (1993) Non-parametric multivariate analyses of changes in community structure. *Aust J Ecol* **18:** 117–143.

Crosby, L.D., and Criddle, C.S. (2003) Understanding bias in microbial community analysis techniques due to rrn operon copy number heterogeneity. *Biotechniques* **34:** 790–794.

Curtis, T.P., and Sloan, W.T. (2005) Exploring microbial diversity – a vast below. *Science* **309:** 1331–1333.

Degnan, P.H., and Ochman, H. (2012) Illumina-based analysis of microbial community diversity. *ISME J* **6:** 183–194.

Fisher, M.M., and Triplett, E.W. (1999) Automated approach for ribosomal intergenic spacer analysis of microbial diversity and its application to freshwater bacterial communities. *Appl Environ Microbiol* **65:** 4630–4636.

Gilbert, J.A., Field, D., Swift, P., Newbold, L., Oliver, A., Smyth, T., *et al*. (2009) The seasonal structure of microbial communities in the Western English Channel. *Environ Microbiol* **11:** 3132–3139.

Gobet, A., Quince, C., and Ramette, A. (2010) Multivariate Cutoff Level Analysis (MultiCoLA) of large community data sets. *Nucleic Acids Res* **38:** e155.

Gobet, A., Böer, S.I., Huse, S.M., van Beusekom, J.E.E., Quince, C., Sogin, M.L., *et al*. (2012) Diversity and dynamics of rare and of resident bacterial populations in coastal sands. *ISME J* **6:** 542–553.

Gower, J.C. (1966) Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika* **53:** 325–338.

Gürtler, V., and Stanisich, V.A. (1996) New approaches to typing and identification of bacteria using the 16S-23S rDNA spacer region. *Microbiology* **142:** 3–16.

Huber, J.A., Morrison, H.G., Huse, S.M., Neal, P.R., Sogin, M.L., Welch, D.B.M., and Mark Welch, D.B. (2009) Effect of PCR amplicon size on assessments of clone library microbial diversity and community structure. *Environ Microbiol* **11:** 1292–1302.

Huse, S.M., Huber, J.A., Morrison, H.G., Sogin, M.L., and Mark Welch, D. (2007) Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol* **8:** R143.

Koopman, M.M., Fuselier, D.M., Hird, S., and Carstens, B.C. (2010) The carnivorous pale pitcher plant harbors diverse, distinct, and time-dependent bacterial communities. *Appl Environ Microbiol* **76:** 1851–1860.

Kunin, V., Engelbrektson, A., Ochman, H., and Hugenholtz, P. (2010) Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environ Microbiol* **12:** 118–123.

Legendre, L., and Legendre, P. (1998) *Numerical Ecology Edition, 2nd English (ed)*. Amsterdam, The Netherlands: Elsevier Science BV.

Legendre, P., and Gallagher, E.D. (2001) Ecologically meaningful transformations for ordination of species data. *Oecologia* **129:** 271–280.

Legendre, P., Borcard, D., and Peres-Neto, P.R. (2005) Analyzing beta diversity: partitioning the spatial variation of community composition data. *Ecol Monogr* **75:** 435–450.

Musat, N., Werner, U., Knittel, K., Kolb, S., Dodenhof, T., van Beusekom, J.E.E., *et al*. (2006) Microbial community structure of sandy intertidal sediments in the North Sea, Sylt-Romo Basin, Wadden Sea. *Syst Appl Microbiol* **29:** 333–348.

Oksanen, J., Guillaume Blanchet, F., Kindt, R., Legendre, P., O'Hara, R.B., Simpson, G.L., *et al*. (2013) *VEGAN: Community Ecology Package*. R package version 2.0–7.

Olsen, G.J., Lane, D.J., Giovannoni, S.J., Pace, N.R., and Stahl, D.A. (1986) Microbial ecology and evolution – a ribosomal-rna approach. *Annu Rev Microbiol* **40:** 337–365.

Pearson, K. (1901) Mathematical contributions to the theory of evolution – VII on the correlation of characters of not quantitatively measurable. *Philos Trans R Soc Lond A* **195:** 1–47.

Quince, C., Curtis, T.P., and Sloan, W.T. (2008) The rational exploration of microbial diversity. *ISME J* **2:** 997–1006.

Quince, C., Lanzén, A., Curtis, T.P., Davenport, R.J., Hall, N., Head, I.M., *et al*. (2009) Accurate determination of microbial diversity from 454 pyrosequencing data. *Nat Methods* **6:** 639–641.

Quinlan, A.R., Stewart, D.A., Stromberg, M.P., Marth, G.T., Strömberg, M.P., and Marth, T. (2008) Pyrobayes: an improved base caller for SNP discovery in pyrosequences. *Nat Methods* **5:** 179–181.

R Development Core Team. (2013) *R: a Language and Environment for Statistical Computing*.

Ramette, A. (2009) Quantitative community fingerprinting methods for estimating the abundance of operational taxonomic units in natural microbial communities. *Appl Environ Microbiol* **75:** 2495–2505.

Ramette, A., and Tiedje, J.M. (2007) Multiscale responses of microbial life to spatial distance and environmental heterogeneity in a patchy ecosystem. *Proc Natl Acad Sci USA* **104:** 2761–2766.

Roesch, L.F.W., Lorca, G.L., Casella, G., Giongo, A., Naranjo, A., Pionzio, A.M., *et al*. (2009) Culture-independent identification of gut bacteria correlated with the onset of diabetes in a rat model. *ISME J* **3:** 536–548.

Shepard, R.N. (1966) Metric structures in ordinal data. *J Math Psychol* **3:** 287–315.

Sogin, M.L., Morrison, H.G., Huber, J.A., Mark Welch, D., Huse, S.M., Neal, P.R., *et al*. (2006) Microbial diversity in the deep sea and the underexplored 'rare biosphere. *Proc Natl Acad Sci USA* **103:** 12115–12120.

Stackebrandt, E., and Goebel, B.M. (1994) Taxonomic note: a place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. *Int J Syst Bacteriol* **44:** 846–849.

Urakawa, H., Kita-Tsukamoto, K., and Ohwada, K. (1999) Microbial diversity in marine sediments from Sagami Bay and Tokyo Bay, Japan, as determined by 16S rRNA gene analysis. *Microbiology* **145:** 3305–3315.

Yannarell, A.C., and Triplett, E.W. (2005) Geographic and environmental sources of variation in lake bacterial community composition. *Appl Environ Microbiol* **71:** 227–239.

Zinger, L., Amaral-Zettler, L.A., Fuhrman, J.A., Horner-Devine, M.C., Huse, S.M., Welch, D.B.M., *et al*. (2011) Global patterns of bacterial beta-diversity in seafloor and seawater ecosystems. *PLoS ONE* **6:** e24570.

Zinger, L., Gobet, A., and Pommier, T. (2012) Two decades of describing the unseen majority of aquatic microbial diversity. *Mol Ecol* **21:** 1878–1896.

Supporting information

Additional Supporting Information may be found in the online version of this article at the publisher's web-site:

Fig. S1. Total OTU numbers along sediment depth or over time for ARISA, OTU_{all}, and $PyroNoise_{3\%}$ data sets.

Fig. S2. Alpha-diversity comparison of different data sets.

Fig. S3. Percentage of shared OTU of the bacterial community between sediment depth layers or sampling dates after correction and OTU clustering of the 454 MPTS data set and of the ARISA data set.

Fig. S4. Percentage of shared OTU of the bacterial community between sediment depth layers or sampling dates at successive taxonomic levels.

Fig. S5. Percentage of shared OTU of the bacterial community between sediment depth layers or sampling dates after applying MultiCoLA.
Fig. S6. Comparison of extracted variation from the different categories of data sets.
Fig. S7. Partitioning of the biological variation in the bacterial community structure.

Table S1. Contribution of environmental parameters to the variation in data sets at different levels of resolution (different techniques, taxonomic assignation, PyroNoise correction, removal of rarer OTU).