environmental microbiology

Environmental Microbiology (2013) doi:10.1111/1462-2920.12250

Metagenomic 16S rDNA Illumina tags are a powerful alternative to amplicon sequencing to explore diversity and structure of microbial communities

Ramiro Logares,[1] Shinichi Sunagawa,[2] Guillem Salazar,[1] Francisco M. Cornejo-Castillo,[1] Isabel Ferrera,[1] Hugo Sarmento,[1,3] Pascal Hingamp,[4] Hiroyuki Ogata,[4,5] Colomban de Vargas,[6] Gipsi Lima-Mendez,[7,8] Jeroen Raes,[7,8] Julie Poulain,[9] Olivier Jaillon,[9,10,11] Patrick Wincker,[9,10,11] Stefanie Kandels-Lewis,[2] Eric Karsenti,[2] Peer Bork[2] and Silvia G. Acinas[1]*

[1]*Department of Marine Biology and Oceanography, Institute of Marine Science (ICM), Spanish National Research Council (CSIC), Passeig Marítim de la Barceloneta, 37–49, Barcelona ES-08003, Spain.*

[2]*European Molecular Biology Laboratory, Meyerhofstr. 1, Heidelberg 69117, Germany.*

[3]*Department of Oceanography and Limnology, Federal University of Rio Grande do Norte, Natal 59014-002, Brazil.*

[4]*Information Génomique et Structurale, Centre National de la Recherche Scientifique, Aix-Marseille Université, Institut de Microbiologie de la Méditerranée, 163 Avenue de Luminy, Marseille 13288, France.*

[5]*Education Academy of Computational Life Sciences, Tokyo Institute of Technology, 12-1, Ookayama, Meguro-ku, Tokyo 152-8552, Japan.*

[6]*Centre National de la Recherche Scientifique, Université Pierre et Marie Curie, Unité Mixte de Recherche 7144, Station Biologique de Roscoff, Roscoff FR-29682, France.*

[7]*Research Group of Bioinformatics and (Eco-)Systems biology, Department of Structural Biology, Vlaams Institut voor Biotechnologie, Pleinlaan 2, Brussels 1050, Belgium.*

[8]*Research Group of Bioinformatics and (Eco-)Systems biology, Microbiology Unit (MICR), Department of Applied Biological Sciences (DBIT), Vrije Universiteit Brussel, Pleinlaan 2, Brussels 1050, Belgium.*

[9]*Genoscope, Institut de Génomique, Commissariat á L'Energie Atomique, 2 rue Gaston Crémieux, Evry 91057, France.*

[10]*Centre National de la Recherche Scientifique– Unité Mixte de Recherche 8030, 2 rue Gaston Crémieux, Evry 91057, France.*

[11]*Université d'Evry, boulevard François Mitterrand, Evry 91025, France.*

Received 26 July, 2013; accepted 10 August, 2013. *For correspondence. E-mail sacinas@icm.csic.es; Tel. (+34) 93 230 8565; Fax (+34) 93 230 95 55.

Summary

Sequencing of 16S rDNA polymerase chain reaction (PCR) amplicons is the most common approach for investigating environmental prokaryotic diversity, despite the known biases introduced during PCR. Here we show that 16S rDNA fragments derived from Illumina-sequenced environmental metagenomes (*mitags*) are a powerful alternative to 16S rDNA amplicons for investigating the taxonomic diversity and structure of prokaryotic communities. As part of the *Tara* Oceans global expedition, marine plankton was sampled in three locations, resulting in 29 subsamples for which metagenomes were produced by shotgun Illumina sequencing (ca. 700 Gb). For comparative analyses, a subset of samples was also selected for Roche-454 sequencing using both shotgun (*m454tags*; 13 metagenomes, ca. 2.4 Gb) and 16S rDNA amplicon (*454tags*; ca. 0.075 Gb) approaches. Our results indicate that by overcoming PCR biases related to amplification and primer mismatch, *mitags* may provide more realistic estimates of community richness and evenness than amplicon *454tags*. In addition, *mitags* can capture expected beta diversity patterns. Using *mitags* is now economically feasible given the dramatic reduction in high-throughput sequencing costs, having the advantage of retrieving simultaneously both taxonomic (Bacteria, Archaea and Eukarya) and functional information from the same microbial community.

Introduction

Microbes have fundamental roles in the functioning of most ecosystems (Falkowski *et al.*, 2008), particularly in the vast ocean biome (DeLong, 2009). They also encompass a large taxonomic and metabolic diversity (Pace, 1997) that reflects their long history of evolutionary

diversification. Still, many important questions in microbial ecology remain unsolved and have been awaiting technological progress to be investigated. The advent of high-throughput sequencing (HTS) technologies (e.g. 454 and Illumina) (Logares *et al*., 2012) is enabling the exploration of microbial diversity at an unprecedented scale. One of the first applications of 454 pyrosequencing in microbial ecology was the sequencing of ribosomal DNA gene (rDNA) amplicons (hereafter $_{454}$tags) from environmental samples (Sogin *et al*., 2006). So far, only a handful of studies have used Illumina-sequenced polymerase chain reaction (PCR) amplicons ($_{i}$tags) to explore natural microbial assemblages (Caporaso *et al*., 2011; 2012; Werner *et al*., 2012; Bokulich *et al*., 2013). However, Illumina sequencers have a cost per base which can be 100 times lower than the 454 platform as well as a higher throughput (Glenn, 2011). As both technologies became popular in microbial ecology relatively recently, a careful evaluation of their performances and biases is still ongoing (Huse *et al*., 2007; 2010; Quince *et al*., 2009; 2011; Claesson *et al*., 2010; Minoche *et al*., 2011; Nakamura *et al*., 2011). A limited number of HTS cross-platform studies have indicated different biases associated with 454 and Illumina platforms (Harismendy *et al*., 2009). For example, comparisons between $_{454}$tags and $_{i}$tags derived from the same DNA samples showed different classification efficiencies (Claesson *et al*., 2010). In general terms, amplicon-based approaches using both $_{454}$tags and $_{i}$tags recovered previously observed global diversity patterns (Caporaso *et al*., 2011; Zinger *et al*., 2011), thus validating these approaches. Still, regardless of the sequencing technology, the biases associated with the PCR step in amplicon-based studies distort the estimations of richness and evenness in microbial communities (Acinas *et al*., 2005; Hong *et al*., 2009; Engelbrektson *et al*., 2010).

An alternative approach to circumvent PCR is to identify rDNA fragments from metagenomic data (hereafter $_{m}$tags). Until recently this approach was unrealistic, as the fraction of rDNA present in metagenomes was very low. For example, the Global Ocean Sampling (GOS) (Rusch *et al*., 2007) produced 7.7 million metagenomic reads, of which only 4100 turned out to be usable 16S rDNA reads (0.05%; see CAMERA, http://camera.calit2.net/). On the second release of GOS, the fraction of rDNA detected was 1.4%, with a total of 142 783 16S rDNA fragments from 80 metagenomes (Yilmaz *et al*., 2011). Other metagenomic studies based on 454 FLX Titanium sequencing identified hundreds to thousands of rDNA fragments (hereafter $_{m454}$tags) after sequencing several million reads (Bryant *et al*., 2012; Ghai *et al*., 2012). Thus, substantial sequencing is needed to recover enough rDNA reads from metagenomes for community taxonomic profiling. However, the high throughput of Illumina HiSeq platforms circumvents this limitation. For example, using the HiSeq2000 platform, we could expect about 10 000 16S rDNA fragments ($>$ 100 bp) out of 10 million metagenomic reads (assuming a 0.1% recovery) at a total cost of about €100; this number of reads would be enough to capture the structure of microbial communities (Caporaso *et al*., 2011). Although 16S rDNA fragments derived from Illumina-sequenced metagenomes have not been subjected to PCR, they have undergone amplification steps associated with the Illumina platform that may generate base-composition biases that, in many cases, are not randomly distributed (Aird *et al*., 2011; Nakamura *et al*., 2011). A number of protocols and base call algorithms have been developed to minimize such biases and improve the error rate of Illumina sequencing (Harismendy *et al*., 2009; Aird *et al*., 2011).

The short length of Illumina reads may represent a limitation, although 16S rDNA reads as short as 100 bp can be enough for an accurate taxonomic characterization of microbial communities (Liu *et al*., 2007). In addition, simulations have shown that 16S rDNA fragments $>$ 150 bp from multiple rDNA regions could be as accurate as the entire 16S rDNA sequence for taxonomic profiling of communities (Hao and Chen, 2012). Longer composite reads can be produced by merging paired-end reads from small insert-size libraries, a strategy that has been shown to produce results comparable to 454 FLX sequencing (Rodrigue *et al*., 2010). Read-length limitations are relaxing with the introduction of newer Illumina sequencers that produce longer reads (e.g. the HiSeq2500 and MiSeq produce 2×150 bp and 2×250 bp reads respectively, which after merging can generate reads up to, e.g., 290 and 490 bp).

Other limitations may be related to the intrinsic characteristics of the 16S rDNA. This gene has regions with different evolutionary rates (Hillis and Dixon, 1991). Diversity metrics and classification accuracy depend on what region is being used (Claesson *et al*., 2009; Engelbrektson *et al*., 2010; Mizrahi-Man *et al*., 2013), and 16S rDNA gene fragments extracted from metagenomes will more or less randomly cover different areas of the gene, thus providing a mixed taxonomic and evolutionary signal. Nevertheless, using different regions may allow reconstruction of the whole 16S rDNA sequence, which could improve diversity analyses (Miller *et al*., 2011), although this method may be affected by the generation of chimeric sequences between closely related taxa.

Altogether, considering the mentioned biases, it is not surprising that taxonomic profiles of microbial communities based on 16S rDNA derived from amplicons or metagenomes may disagree (Shah *et al*., 2011). In general, controlled quantitative studies comparing rDNA-based diversity using different sequencing platforms (e.g. Illumina vs 454) and PCR-based versus non-PCR-based tags are very limited. Still, a recent study using synthetic

microbial communities tested the capacity of PCR-based versus non PCR-based sequencing for recovering known diversity and indicated that the non-PCR-based approach performed better (Shakya *et al.*, 2013). Despite the obvious value of the latter approach for quantitatively uncovering biases and potential errors, synthetic communities are still a great simplification of natural microbial communities. The environmental DNA pool is highly complex, encompassing thousands of different genomes, many of which are unknown and normally present in very low abundances (Pedrós-Alió, 2006); therefore, kinetics and amplification PCR biases may behave differently than in controlled studies. Thus, studies based on natural samples are also needed to complement controlled laboratory experiments and in combination generate more realistic descriptions of microbial diversity.

Here we investigate whether 16S rDNA fragments derived from environmental metagenomes sequenced with Illumina (hereafter $_{mi}$tags) can capture diversity patterns of microbial communities. Our results are based on data from three marine stations that were part of the *Tara* Oceans global expedition (Karsenti *et al.*, 2011) and which were sequenced extensively using Illumina HiSeq2000 and GAIIx platforms. For comparative purposes, we generated metagenomes and 16S rDNA amplicon sequence data using the 454 GS FLX Titanium platform for a subset of these stations. We show that $_{mi}$tags can be used for taxonomic profiling of natural microbial communities as well as for richness, evenness and beta diversity estimations. Using $_{mi}$tags has at least two main advantages: (i) it avoids PCR biases; and (ii) a large amount of functional data are simultaneously produced when $_{mi}$tags are generated. Thus, $_{mi}$tags are a powerful alternative to the commonly used amplicon-based tags for community analyses. Using $_{mi}$tags is now feasible thanks to the dramatic decrease in sequencing costs.

Results

The 29 Illumina metagenomes from the three analysed marine stations consisted of about 700 Gb of sequence data covering five planktonic size fractions (0.2–1.6, 0.8–5, 5–20, 20–180 and 180–2000 μm). The approach used to extract and process $_{mi}$tags is displayed in Supporting Information (Fig. S1). On average, 2.08×10^4 16S $_{mi}$tags > 100 bases were extracted per sample (metagenome), although 7.9×10^4 16S $_{mi}$tags were retrieved from the typically free-living bacterial size fraction (0.2–1.6 μm) (Supporting Information Table S2). Altogether, these $_{mi}$tags covered all 16S rDNA hypervariable regions (V1 to V9), with a decrease in coverage at the 16S extremes (Supporting Information Fig. S2). A cross-platform analysis using $_{mi}$tags, $_{m454}$tags and $_{454}$tags indicated that the three methods showed similar degrees of taxonomic classification efficiency to the Ribosomal Database Project (RDP) (Cole *et al.*, 2009) when using the naïve Bayesian classifier (Wang *et al.*, 2007), albeit $_{mi}$tags had shorter sequence length (Supporting Information Fig. S3). These results show that thousands of 16S $_{mi}$tags covering all 16S DNA regions can be extracted from metagenomes and taxonomically classified to RDP.

Assignment of $_{mi}$tags, $_{m454}$tags and $_{454}$tags to reference OTUs

Most of the 16S $_{mi}$tags corresponded to the prokaryote size fraction (0.2–1.6 μm), and 94% of them were assigned to SILVA reference operational taxonomic units (OTUs) (Supporting Information Table S2). This indicates that the main fraction of bacterial taxa was represented in the SILVA reference database (Quast *et al.*, 2013). About 28% of the total number of $_{mi}$tags mapped to the V1–V3 region (Supporting Information Table S2), which was later used in comparative cross-platform analyses. This number was expected when considering a more or less uniform read coverage of the 16S rDNA (about 1300 bp) and the length of the V1–V3 region (about 500 bp). The V1–V3 was selected because it includes the V3 region, which is frequently used for marine $_{454}$tag rDNA amplicon studies and has a better resolution than the V6 region (Huse *et al.*, 2008). Similar results were obtained with $_{m454}$tags (about 92% of reads were assigned to SILVA reference OTUs, and of these about 20% were assigned to the V1–V3 segment; Supporting Information Table S3). The number of $_{454}$tags that could be assigned to OTUs was slightly smaller (about 86%; Supporting Information Table S4). The range of OTUs obtained per sample using *de novo* clustering (i.e. not based on a reference database) with the $_{454}$tags from regions V1 (287–1204) and V3 (310–1443) was not different from what was obtained by assignment to reference OTUs (524–1070) (ANOVA; $P > 0.58$) (Supporting Information Table S4).

Richness and evenness: a comparative analysis

When using all $_{mi}$tags from all 16S rDNA V regions, $_{mi}$tags recovered on average 61% more OTUs than $_{454}$tags (Supporting Information Table S5, Fig. 1A). When using a subsample of 2000 reads per sample, the increase is between 31.1 to 43.2% of OTUs per sample (Supporting Information Table S5). This increase translated to Chao-1 richness diversity estimator was 40.3% on average, and equivalent results were also observed using the abundance-based coverage estimator index (Supporting Information Table S6). Under the most comparable scenario, taking into account only $_{mi}$tags from the V1–V3 region and $_{454}$tags trimmed to the same length range as $_{mi}$tags, both $_{mi}$tags and trimmed $_{454}$tags recovered similar

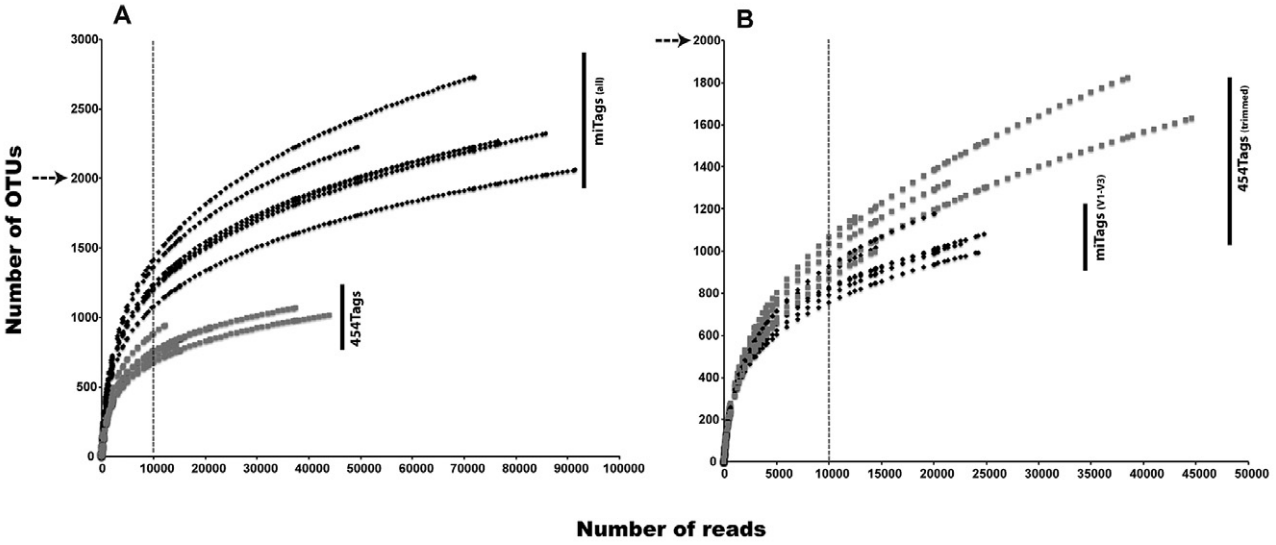

Fig. 1. Rarefaction analyses using two datasets. The dashed vertical line indicates a comparative sampling size for the datasets presented in A and B. Note that in A and B the sample size was different because of the different characteristics of the datasets. Also note that the vertical axes have different lengths.
A. Only the dataset including all *$_{mi}$tags* and *$_{454}$tags* was considered, representing the data actually gathered. The horizontal arrow indicates the maximum vertical value of B.
B. The dataset considered included *$_{mi}$tags* falling into the V1–V3 region and trimmed *$_{454}$tags*, representing the data most comparable between platforms and approaches.

numbers of OTUs, ranging between 994 and 1178 for $_{mi}$tags and between 586 and 1824 for trimmed $_{454}$tags (Fig. 1B).Values were even closer when subsampling at 2000 reads per sample ($_{mi}$tags 428–508 OTUs and trimmed $_{454}$tags 443–515 OTUs). Rarefaction analyses using all $_{mi}$tags (covering the entire 16S rDNA gene) from the size fractions 0.2–1.6 and 0.8–5 μm indicated a larger richness in the 0.8–5 μm size fraction (Supporting Information Fig. S4). Interestingly, it was in the size fractions > 5μm that the percentage of mapped $_{mi}$tags among reference OTUs dropped to 58% (Supporting Information Table S2), suggesting prokaryote novelty probably associated with larger particles.

We compared the capability of $_{mi}$tags and $_{454}$tags to detect prokaryote taxonomic diversity using both single reads and OTUs. At higher taxonomic levels, $_{mi}$tags uniquely recovered several phyla (e.g. *Fibrobacteres* and *Tenericutes*) and classes (*Halobacteria*, *Chloroflexi*) in RDP classifications (Cole *et al*., 2009) (Supporting Information Table S7). At lower levels, we found 748 genera that were exclusively detected by $_{mi}$tags (Supporting Information Fig. S5A; Supporting Information Table S7), whereas only nine genera were exclusively detected by $_{454}$tags (Supporting Information Table S8). Similar results were obtained in OTU-based analyses when using the TARA-V1–V3 dataset both with and without subsampling (see Supporting Information Fig. S1). Again, a higher number of unique OTUs were recovered by $_{mi}$tags than by $_{454}$tags. When using the complete dataset, we observed that 40.8% of the OTUs were recovered by both $_{mi}$tags and $_{454}$tags, while 43.7% and 15.5% were recovered exclusively by $_{mi}$tags and $_{454}$tags respectively (Supporting Information Fig. S5B, left panel). For the subsampled dataset, normalization corrected artefacts that produced some of the differences between techniques, but 446 OTUs were still exclusively obtained by $_{mi}$tags and 274 OTUs by $_{454}$tags (Supporting Information Fig. S5B, right panel).

We investigated the phylogenetic differences between the OTUs retrieved by $_{mi}$tags and $_{454}$tags from the same V1–V3 region (Fig. 2). Both methods, $_{mi}$tags and $_{454}$tags, presented a good agreement by recovering taxa from the same evolutionary groups (Fig. 2). Still, there were cases where $_{mi}$tags recovered small clusters that were not recovered by $_{454}$tags, as well as a few cases displaying the opposite pattern (Fig. 2). In general, unique OTUs from $_{mi}$tags were spread over all bacterial classes (see unique $_{mi}$tags clusters labelled with numbers in Fig. 2 and Supporting Information Table S7). Furthermore, $_{mi}$tags retrieved Archaea, which were expectedly absent in $_{454}$tags because of the use of bacterial primers.

The primer bias effect, as a potential explanation for the differences in OTU detection between the two techniques, was further investigated on two fronts by (i) analysing the *in silico* coverage of the primer pair set used for generating 16S rDNA amplicon tags and by (ii) statistical analyses comparing the number of OTUs detected by each approach to the presence of mismatches with the primer

pair used. First, we tested the theoretical accuracy of the primer pair (27Fmod/533R). This pair covered 78.9% of the references and was well distributed across main phyla, ranging between 60% and 100% coverage (Supporting Information Fig. S6). A few phyla were poorly represented in terms of coverage, probably because of the low number of sequences available in datasets (Supporting Information Fig. S6). Secondly, two χ^2-tests of independence were performed between these two datasets (OTUs detected by $_{454}$tags/$_{mi}$tags and primer detection with match/mismatch). Strong and significant dependence was found for OTUs detected only by $_{454}$tags and the presences of mismatches with respect the OTUs detected with mitags ($\chi^2 = 53.04$, df = 1, $P < 0.0001$) (Supporting Information Table S9). Conversely, when we selected only the OTUs detected with $_{454}$tags, OTU detection with $_{mi}$tags and the presence of mismatches appeared as independent factors ($\chi^2 = 1.45$, df = 1, $P = 0.2284$) (Supporting Information Table S9). This primer bias effect resulted in an underrepresentation of those OTUs having mismatches with the primer pair and an overrepresentation of those OTUs with a perfect match with the primer pair. However, this primer bias effect cannot be associated with any phyla in particular, although differences exist in the coverage within the main phylum.

Further comparative analyses focused on the evenness patterns retrieved by $_{mi}$tags and $_{454}$tags (Supporting Information Fig. S7). First, similar rank–abundance curves were observed when samples were subsampled (Supporting Information Fig. S7, right panel); however, some differences emerged when using data that were not subsampled. Interestingly, $_{mi}$tags tended to recover a higher number of very low-abundance taxa (< 0.1%) from the rare biosphere (Pedrós-Alió, 2012) (Supporting Information Fig. S7, left panel). Despite the overall similarity in rank–abundance curves, different platforms (454 vs Illumina) and approaches (amplicon-derived tags vs $_{mi}$tags) indicated, in several cases, different abundances for the same OTUs (Supporting Information Fig. S7, left panel; Fig. 3). When OTU abundances derived using $_{mi}$tags, $_{454}$tags and $_{m454}$tags were compared, a better agreement was found between approaches not involving PCR ($_{m454}$tag and $_{mi}$tag), resulting in a higher correlation and a fit closer to the 1:1 line (Fig. 3; Supporting Information Table S10). Interestingly, both comparisons involving PCR (i.e. involving $_{454}$tag) resulted in smaller slopes and positive intercepts, indicating that the abundance of rare OTUs was underestimated and that the abundance of abundant OTUs was overestimated with $_{454}$tags compared with $_{m}$tags (Supporting Information Table S10). Finally, to examine the performance of $_{mi}$tags for quantitative assessment of OTUs, we compared the relative abundance of several prokaryotic taxa obtained with $_{mi}$tags with those obtained by two well established quantitative approaches: catalysed reporter deposition fluorescence *in situ* hybridization (CARD-FISH) counts (Fig. 4) and flow cytometry (Supporting Information Fig. S8). First, we measured four bacterial groups, SAR11, *Gammaproteobacteria*, *Bacteroidetes* and *Roseobacter*, which exhibited distinct abundance in environmental samples. Our findings revealed a good agreement between CARD-FISH and $_{454}$tags/$_{mi}$tags (Fig. 4; CARD-FISH vs $_{mi}$tags: Pearson's $r = 0.866$, $P < 0.001$; CARD-FISH vs $_{454}$tags: Pearson's $r = 0.948$, $P < 0.001$). Similarly, a positive correlation was observed between cyanobacterial abundance (*Prochlorococcus* and *Synechococcus*) measured by flow cytometry and $_{mi}$tag-derived abundance (*Prochlorococcus*: Pearson's $r = 0.782$, $P < 0.001$; *Synechococcus*: Pearson's $r = 0.603$, $P < 0.001$; Supporting Information Fig. S8).

Comparative community structure using $_{mi}$tags and $_{454}$tags

UPGMA (unweighted pair-group method with arithmetic mean) clustering analysis based on Bray–Curtis distances was performed for the four analysed datasets (TARA-ALL, TARA-TRIMMED, TARA-V1–V3, TARA-V1–V3-TRIMMED; see methods and Supporting Information Fig. S1) after subsampling them to 2000 reads per sample (Supporting Information Fig. S9A–D). In three out of the four datasets, the $_{454}$tag samples clustered together instead of with their corresponding $_{mi}$tag samples (Supporting Information Fig. S9A–C). Only in the dataset considering trimmed $_{454}$tags and the V1–V3 region (TARA-V1–V3-TRIMMED) did one sample analysed with $_{454}$tags cluster with the same sample analysed with $_{mi}$tags (Supporting Information Fig. S9D). Furthermore, in this latter dataset, samples from the prokaryote size fraction (0.2–1.6μm) analysed with $_{454}$tags and $_{mi}$tags clustered together, forming a tight group (Supporting Information Fig. S9D). The absence of clustering of the same samples analysed with $_{mi}$tags and $_{454}$tags reflects the unequal estimation of richness and evenness by the different techniques and platforms. Nevertheless, we observed a relatively strong correlation using binary (i.e. presence–absence) Bray–Curtis dissimilarity values (Mantel test: Pearson's $r = 0.75$, $P = 0.002$) between the same set of samples analysed with $_{mi}$tags and $_{454}$tags (prokaryote fraction from dataset TARA-ALL subsampled). This means that samples that were more dissimilar in composition according to $_{mi}$tags were also more dissimilar according to $_{454}$tags and vice versa. However, a weaker correlation was observed for the same set of samples when using the regular Bray–Curtis index, which considers relative abundances (Mantel test: Pearson's $r = 0.44$, $P = 0.023$). This discrepancy could be associated with PCR biases affecting the relative abundance of taxa measured by $_{454}$tags.

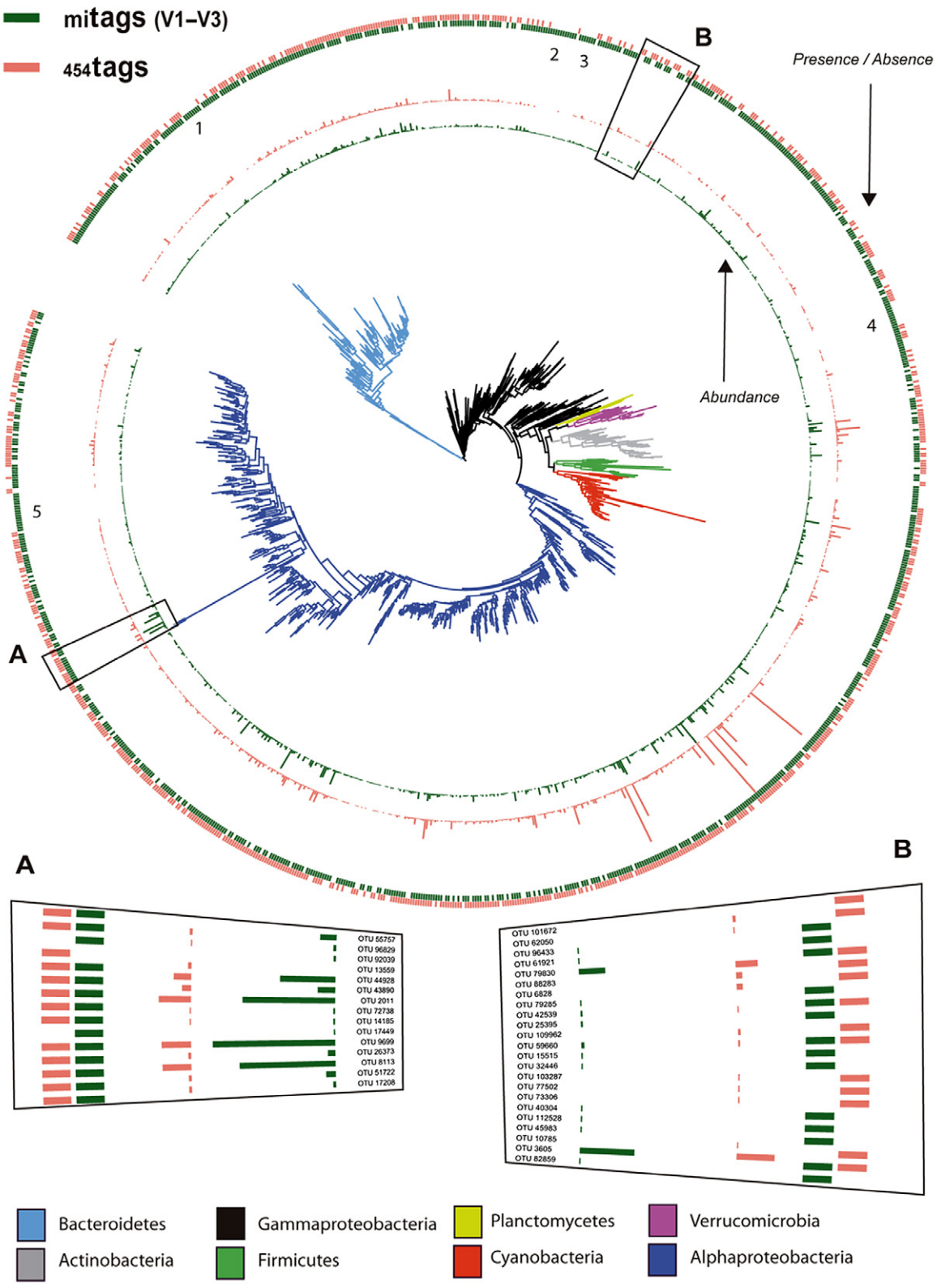

Fig. 2. Phylogeny of the OTUs recovered with $_{mi}$*tags* (V1–V3) and $_{454}$*tag*s where all samples were subsampled to 2000 reads per sample (TARA-V1–V3 OT with subsampling). $_{mi}$*tag*s are indicated in green and $_{454}$*tag*s in salmon color. The inner rings indicate OTU relative abundances (variable-length columns) and the outer rings (fixed-length columns) presence/absence of given OTUs in the $_{454}$*tag*s and/or $_{mi}$*tag*s. Zooms of two selected areas of the tree are presented in A and B. Examples similar to the ones presented in Boxes A and B were observed throughout the entire phylogeny. Unique clusters of OTUs from different phylogenetic taxa retrieved only by $_{mi}$*tag*s and not by $_{454}$*tag*s are represented by numbers from 1 to 5. Main taxonomic groups are indicated by the tree leaves' colors and correspond to the legend at the bottom of the figure.
A. Relative abundance estimated by $_{mi}$*tag*s and $_{454}$*tag*s can be either very similar or different for evolutionarily related OTUs.
B. Some evolutionarily related OTUs (probably groups) may be recovered by $_{mi}$*tag*s and not by $_{454}$*tag*s (and vice versa).

Discussion

In our metagenomic samples, $_{mi}$tags accounted for about 0.01–0.1% of the total reads, which is within the expected range. The 0.1% 16S rDNA recovery rate reported here and in previous studies (Rusch *et al*., 2007) seems to be independent of the sequencing technology (Sanger shotgun, Roche-454 or Illumina), providing a good plausibility check for metagenome-sequencing projects. Due to the high throughput of Illumina platforms, the number of $_{mi}$tags recovered per sample (79 000 $_{mi}$tags on average for bacterial size fraction) can be considered more than sufficient for capturing community composition patterns (Caporaso *et al*., 2011). As expected, the yield of $_{mi}$tags for the typical bacterial size fraction was higher (about 0.09%) than for size fractions > 5µm (0.01%). Most $_{mi}$tags (94%) could be mapped to reference OTUs present in the SILVA reference database. Although the latter results come from three Mediterranean stations, these findings can be extrapolated to other marine photic samples. In fact, in another work, we have extracted all $_{mi}$tags for 72 globally distributed samples of 35 *Tara* Oceans stations that represented surface, deep chlorophyll maximum, oxygen minimum zone and mesopelagic water samples, which showed similar $_{mi}$tag mapping percentages to the three previous marine stations (Salazar *et al*., unpublished). Similarly, using RDP, most $_{mi}$tags (99%) could be confidently classified, and in all cases, as expected, classification confidence decreased at lower taxonomic levels (Claesson *et al*., 2010).

In this work, we assigned $_{mi}$tags to the reference OTUs derived from clustering the SILVA 108 reference database at 97% similarity. This approach may have at least two drawbacks. First, if a sample contains OTUs that are not present in the reference database, then they will not be counted. Nevertheless, we found that most (> 94%) 16S $_{mi}$tags from marine samples were assigned to reference OTUs, indicating that SILVA 108 is appropriate for typical marine surface studies. The second possible drawback is that $_{mi}$tags are shorter than $_{454}$tags, and they contain less information for taxonomic assignment; this may be further complicated if a specific $_{mi}$tag covers a conserved 16S rDNA region. Thus, $_{mi}$tags may produce some diversity inflation, as different segments of the same 16S rDNA

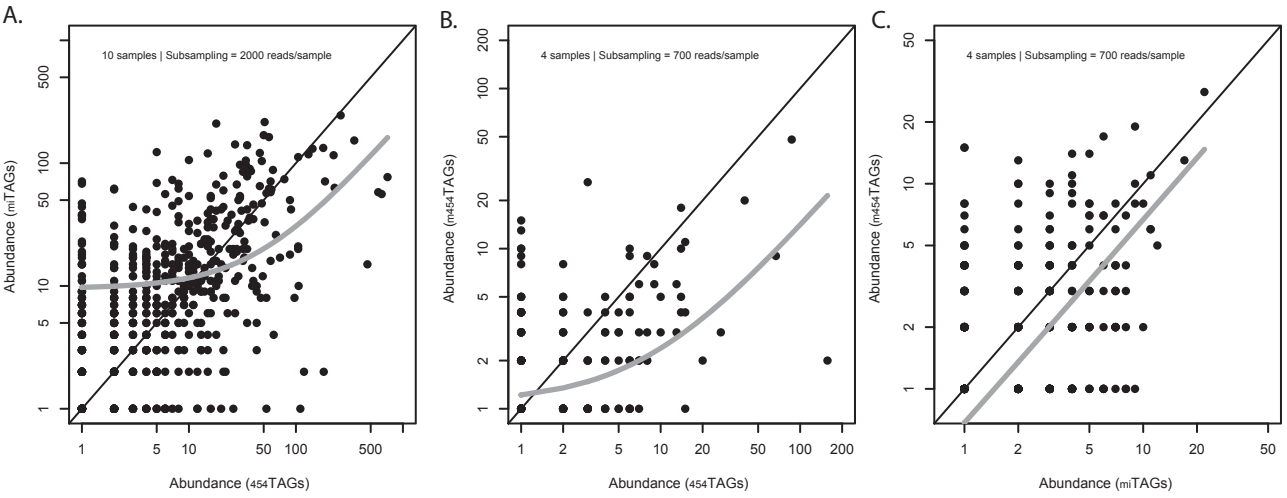

Fig. 3. Platform and PCR bias comparison. OTU abundances estimated with the three different techniques are compared for the pooled set of samples. All comparisons were done with subsampling and the greatest possible number of reads per sample. Samples with less than 500 reads were excluded from the comparison. The red line is the best fit to a linear model.
A. $_{454}$*T*ags versus $_{mi}$*tag*s, reflecting a potential joint cross-platform and PCR bias effect.
B) $_{454}$*T*ags versus $_{m454}$*tag*s, only reflecting a potential PCR bias effect within the same sequencing platform.
C) $_{m454}$*T*ags versus $_{mi}$*tag*s, only reflecting the cross-platform effect (no PCR involved).

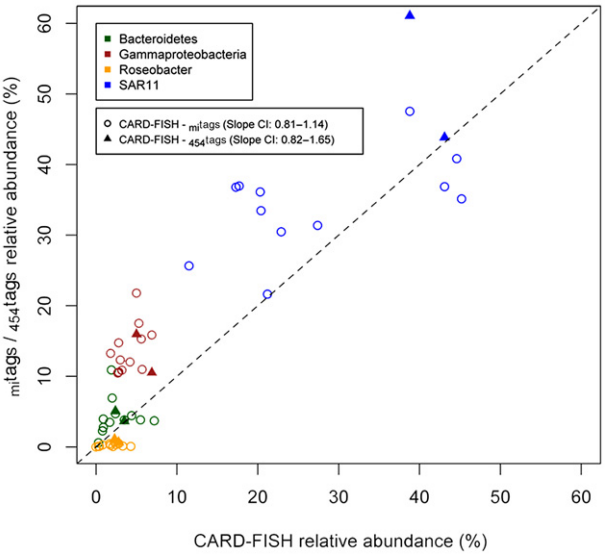

Fig. 4. Quantitative comparison of relative abundances of *mitag*s (empty circles) with CARD-FISH counts or *454tag*s (full triangles) versus CARD-FISH. Relative abundances (%) of four different prokaryote groups (*Bacteroidetes*, *Gammaproteobacteria*, *Roseobacter* and SAR11) estimated with CARD-FISH are compared to *mitag* and *454tag* estimates. A linear model was adjusted, and 95% confidence intervals were computed for the slope.

sequence (e.g. one conserved and another one variable) may be assigned to different OTUs. Nevertheless, the rarefaction analyses suggested that the potential inflation of diversity, if it exists, is not too large (Fig. 1). In addition, statistical analyses based on OTUs from hypervariable regions (V1–V3) detected by miTags and 454Tags indicated that the extra diversity recovered by miTags is at least partially associated with lineages not recovered with 454Tags (Fig. 2) due to primer mismatches (Supporting Information Table S9). A potential advantage of miTags is that specific 16S rDNA V regions could be selectively extracted to conduct *de novo* clustering with longer Illumina reads. This option is of particular importance when significant prokaryote novelty is expected that may not be represented in reference databases.

Using all miTags, the OTU numbers per sample (alpha richness) detected in different marine samples and size fractions were in the range of other marine studies (Pommier *et al.*, 2010; Crespo *et al.*, 2013; Sul *et al.*, 2013), supporting their use in microbial diversity analysis. Beta diversity analyses reflected the somewhat different community compositions indicated by miTags and 454Tags for the same samples of the prokaryote fraction, which formed different clusters (Fig. S9). Thus, it appears that the most reasonable approach is to avoid mixing data from different platforms (Illumina and 454 in this case) and approaches (PCR and non-PCR data). Our results indicated that both approaches (i.e. miTags and 454Tags) tend to provide a similar view of community differentiation if abundance data are omitted, which could be associated with potential PCR biases in amplicon-derived approaches.

miTags as an alternative for probing microbial diversity

The generation of miTags does not require prolonged PCR, a process well known to introduce biases. Generation of chimeric sequences and unequal amplification of targets during PCR may substantially distort microbial diversity estimations (Acinas *et al.*, 2005; Haas *et al.*, 2011). Furthermore, the primers used during PCR may not detect certain taxa (Hong *et al.*, 2009) and may have variable specificity to other taxa. Our analyses indicated that miTags recovered more taxa at different taxonomic levels and OTUs than 454Tags. The recovery of more OTUs using miTags could be related, to a certain extent, to errors during the OTU mapping step; limitations in the mapping algorithm could assign different fragments of the same 16S to different OTUs. However, the recovery of unique phyla and classes as well as other lower taxonomic levels indicates that miTags recover OTUs that are probably missed during the PCR step before 454Tag generation. These results were also supported by phylogenetic analyses, which showed that several clades (composed of more than a few reference OTUs) from different phylogenetic groups were only recovered by miTags (Fig. 2). Furthermore, the lack of detection of several OTUs with 454Tags was statistically proved to be related to primer mismatches, while there was no primer bias when testing the miTag approach (Supporting Information Table S9).

Not only did miTags and 454Tags differ in the number of recovered taxa, but also, and probably more markedly, in the registered relative abundances for the same OTUs. We have compared the effects of PCR using m454Tags and 454Tags. Some OTUs were abundant among 454Tags and rare with, for example, m454Tags or miTags, and vice versa. These differences are most likely related to PCR biases and agree with results indicating that PCR underestimates rare taxa and favors the detection of abundant ones (Gonzalez *et al.*, 2012). Probably for this reason, we observed that miTags captured more members of the rare biosphere than 454Tags. Using a different dataset from deep-ocean marine microbial communities, we performed a comparison between miTags and iTags, retrieving a picture similar to that for miTags vs 454Tags (Salazar *et al.*, unpublished).

Finally, we have analysed the sequencing platform effect by comparing miTags and m454Tags and the approach effect (amplicon PCR 454Tags vs mTags). Despite the observed deviations from a linear relationship, the non-PCR scenarios provided the most compatible results, thus

supporting the use of $_{m}$tags for community profiling (Fig. 3C). Lastly, quantitative techniques other than rDNA sequencing (i.e. FISH and flow cytometry) showed comparable results, suggesting that $_{mi}$tags exhibited an equally good quantitative performance, at least for the taxa compared (Fig. 4). Using data from controlled synthetic microbial communities where differences between them could be adequately quantified, it was found that metagenomics (both 454 and Illumina) outperformed amplicon 16S tag sequencing in quantitatively reconstructing community composition (Shakya *et al*., 2013).

In summary, $_{mi}$tags are a feasible alternative for diversity analysis and prokaryote community profiling that avoids PCR biases. We summarize the characteristics of the analysed approaches and platforms in Table 1. Depending on research goals, different possibilities emerge. The longer sequences provided by Roche-454 platforms (up to 800–1000 bp) are still highly valuable to facilitate accurate assembly for metagenomes or for designing new primers or probes for unknown microorganisms. Similarly, $_{i}$tags would be of interest for those studies focusing on diversity saturation or involving a very large number of samples. Illumina metagenomes can be done with as a little as 100 ng of DNA, and it is important to note that Illumina sequencers are rapidly increasing their throughput and sequence length. For example, $_{mi}$tags are already longer in newer platforms (e.g. Illumina MiSeq generates 2×250 bp paired-end reads), improving OTU assignment and taxonomic classifications. Thus, the $_{mi}$tag approach will become more powerful and accessible in cost terms with the advance of high-throughput sequencing technologies.

Experimental procedures

Detailed experimental procedures can be found in the online version of this article under Supporting Information.

Building the $_{mi}$tag, $_{m454}$tag, and $_{454}$tag datasets

From the 29 analysed metagenomes, a total of 5.03×10^9 and 1.79×10^9 raw and merged paired-end metagenomic reads respectively were produced for Illumina (> 100 bp, GAIIx and HiSeq2000; Supporting Information Table S2). This represents about 700 Gb of metagenomic sequence data. From these libraries, 6.05×10^5 16S $_{mi}$tags > 100 bp were extracted (Supporting Information Table S2). Using the 454 GS FLX Titanium platform, a total of 8.1×10^6 reads from 13 metagenomes were produced (about 2.4 Gb), and 3.30×10^3 $_{m454}$tags > 100 bp were extracted (Table S3). $_{mi}$Tags (> 100 bp) represented a small fraction of all merged paired-end reads (0.09% on average for the prokaryote size fraction; Supporting Information Table S2). Similar values were obtained using $_{m454}$tags (mean 0.11%; Supporting Information Table S3). Because of the greater sequencing depth allowed by the Illumina platform (about 15 Gb per metagenome in our samples), we were able to extract $5–9 \times 10^4$ 16S reads (> 100 bp) ($_{mi}$tags) per metagenome from the prokaryote size fraction (Supporting Information Table S2). A much smaller number of $_{m454}$tags was recovered with the 454 GS FLX Titanium platform because of the more limited throughput (Table S3).

Additionally, 16S $_{454}$tags (derived from amplicon sequencing of the V1–V3 region) were obtained from six samples from the prokaryote size fraction (0.2–1.6 μm), totalling 2.63×10^5 reads. After stringent quality filtering, this dataset was reduced to 1.53×10^5 $_{454}$tags (Supporting Information Table S4). Using $_{454}$tags, we obtained $2.88–7.00 \times 10^4$ reads (> 100 bp) per sample (Supporting Information Table S4). The sequence data of 16S $_{mi}$tags, $_{m454}$tags and $_{454}$tags used for this study were deposited in the European Nucleotide Archive as follows: (i) shotgun sequencing of *Tara* Oceans DNA samples corresponding to size fractions for prokaryotes (0.22–1.6 μm) done by Illumina technology ($_{mi}$tags) (ERA242033, ERA242034) and by 454 Titanium pyrosequencing technology ($_{m454}$tags) (ERA155563, ERA155562); (ii) shotgun sequencing of *Tara* Oceans DNA samples corresponding to size fractions for plankton and larger size fractions (0.8–5, 5–20, 20–180 and 180–2000 μm) performed by Illumina technology ($_{mi}$tags) (ERA242028) and 454 Titanium pyrosequencing technology ($_{m454}$tags) (ERA241291); and (iii) 16S rDNA gene sequencing ($_{454}$tags) of *Tara* Oceans DNA samples corresponding to size fractions for prokaryotes (0.22–1.6 μm) done by 454 Titanium pyrosequencing technology (ERA242032).

Analysed datasets (OTU tables)

A total of four main OTU tables (OTs) were constructed: (i) the TARA-ALL OT contained all $_{mi}$tags, $_{m454}$tags and $_{454}$tags; (ii) the TARA-TRIMMED OT contained the same data as in (i), but here the $_{454}$tags were trimmed to 100–150 bp; (iii) the TARA-V1–V3 OT included only tags that fell within the V1–V3 region; and (iv) the TARA-V1–V3-TRIMMED OT comprised $_{mi}$tags within the V1–V3 region and trimmed $_{454}$tags (100–150 bp). Finally, all four OTs were subsampled (in QIIME) to 2000 reads per sample to correct for potential biases introduced by unequal sequencing effort. Supporting Information Fig. S1 displays a simplified pipeline diagram of the datasets. From all OTU tables, we removed Archaea, chloroplasts and Eukarya. Singletons as well as OTUs present in only one sample were included, as the reference-based OTU assignment approach reduces the chances of generating false OTUs (i.e. $_{mi}$tags/$_{m454}$tags/$_{454}$tags are mapped to Sanger reference sequences, thus validating automatically the quality of the read).

Acknowledgements

We are keen to thank the commitment of the people, institutions and sponsors who made this singular expedition possible: CNRD, EMBL, Genoscope/CEA, UPMC, VIB, Stazione Zoologica Anton Dohrn, UNIMIB, ANR, FWO, BIO5, Biosphere 2, agnès b., the Veolia Environmental Foundation, Region Bretagne, World Courier, Cap L'Orient, the Foundation EDF Diversiterre, FRB, the Prince Albert II de Monaco Foundation, Etienne Bourgois, the Tara Foundations teams

Table 1. General comparison of the different platforms and approaches.

Approach[a]	Template	Coverage	16S rDNA specificity	16S rDNA recovery[b]	PCR bias[c]	16S rDNA overlap[d]	Taxonomic definition[e]	OTU clustering[f]	€/Mb[g]
$_{mi}$Tags	Metagenomic fragments	16S rDNA, functional metagenomic	Spanning all 16S rDNA	High/moderate	Absent	Low	Variable	Mapping to reference OTUs/V region selection for *de novo*	0.1/100[h] (HiSeq)
$_{m454}$Tags	Metagenomic fragments	16S rDNA, functional metagenomic	Spanning all 16S rDNA	Very low	Absent	Low	Variable	Mapping to reference OTUs/V region selection for *de novo*	12/12000[h] (Titanium)
$_{454}$Tags	Amplicons	16S rDNA only	Specific 16S rDNA area	High/very high	Present	High	High	*De novo* and mapping to reference OTUs	12 (Titanium)
$_{i}$Tags	Amplicons	16S rDNA only	Specific 16S rDNA area	Very high	Present	High	High/moderate	*De novo* and mapping to reference OTUs	0.7 (MiSeq)

a. The four basic approaches are indicated: $_{mi}$tags (metagenomic Illumina 16S tags), $_{m454}$tags (metagenomic 454 16S tags), $_{454}$tags (amplicon-based 454 16S tags) and $_{i}$tags (amplicon-based Illumina 16S tags).
b. Number of recovered 16S rDNA reads from the used template. Estimations depend on the throughput of the platform.
c. PCR bias refers mostly to known primer biases and chimera formation.
d. Overlapping of the recovered 16S rDNA fragments. 16S recovered from metagenomes show a limited overlapping that precludes typical clustering techniques.
e. Taxonomic information associated with the recovered fragments. Fragments extracted from metagenomes normally present different amounts of taxonomic information (e.g. reads could be assigned to one specific genus or several families depending on their variability).
f. OTU clustering methods that can be used with the different approaches.
g. Approximate costs per million base pairs. Based on Glenn (2011).
h. Costs to generate $_{mi}$tags/$_{m454}$tags disregarding all the remaining data that is not 16S; for example, about 1 Gb of metagenomic data need to be sequenced to obtain 1 Mb of metagenomic tags, and the cost to generate 1 Gb is reported.

and crew. TARA Oceans would not exist without the continuous support of the participating institutes (see Karsenti *et al*., 2011). This is contribution no. 008 of the *Tara* Oceans Expedition, 2009–2012. We thank Dr Josep M. Gasol for critical reading and Dr Pedrós-Alió for his helpful comments. S.G.A. was supported by a Ramon y Cajal contract from the Spanish Ministry of Science and Innovation and FP7-OCEAN-2011 'Micro B3'. This research was supported by grants BACTERIOMICS (CTM2010-12317-E), TANIT (CONES 2010-0036) from the Agència de Gestió d'Ajusts Universitaris i Reserca and MicroOcean PANGENOMICS (CGL2011-26848/BOS) to S.G.A. by the Spanish Ministry of Science and Innovation (MICINN). R.L. has been supported by a Marie Curie Intra European Fellowship (MASTDIEV; PIEF-GA-2009–235365, EU) and by the Spanish Ministry of Science and Innovation (Juan de la Cierva Fellowship, JCI-2010–06594), and G.S. and F.M.C. were supported by PhD JAE-Predoc (CSIC) and FPI (MICINN) fellowships respectively. High-throughput computing resources were provided by the Barcelona Supercomputing Center (http://www.bsc.es/) through the grants BCV-2010-3-0003 and 2011-2-0003/3-0005 to RL. Additional funding was provided by the Agence Nationale de la Recherche, ANR grants Prometheus ANR-09-GENM-031, Poseidon ANR-09-BLAN-0348, Oceanomics ANR-11-BTBR-0008 and Tara-GirusANR-09-PCS-GENM-218. S.S. and P.B. were supported by EMBL core funding, and GL and JR are supported by the Fund for Scientific Research Flanders (FWO). Supplementary information is available at EMI's website.

References

Acinas, S.G., Sarma-Rupavtarm, R., Klepac-Ceraj, V., and Polz, M.F. (2005) PCR-induced sequence artifacts and bias: insights from comparison of two 16S rRNA clone libraries constructed from the same sample. *Appl Environ Microbiol* **71:** 8966–8969.

Aird, D., Ross, M.G., Chen, W.S., Danielsson, M., Fennell, T., Russ, C., *et al*. (2011) Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome biol* **12:** R18.

Bokulich, N.A., Subramanian, S., Faith, J.J., Gevers, D., Gordon, J.I., Knight, R., *et al*. (2013) Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nat Methods* **10:** 57–59.

Brosius, J., Palmer, M.L., Kennedy, P.J., and Noller, H.F. (1978) Complete nucleotide sequence of a 16S ribosomal RNA gene from *Escherichia coli*. *Proc Natl Acad Sci USA* **75:** 4801–4805.

Bryant, J.A., Stewart, F.J., Eppley, J.M., and DeLong, E.F. (2012) Microbial community phylogenetic and trait diversity declines with depth in a marine oxygen minimum zone. *Ecology* **93:** 1659–1673.

Caporaso, J.G., Lauber, C.L., Walters, W.A., Berg-Lyons, D., Lozupone, C.A., Turnbaugh, P.J., *et al*. (2011) Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc Natl Acad Sci USA* **108** (Suppl 1): 4516–4522.

Caporaso, J.G., Lauber, C.L., Walters, W.A., Berg-Lyons, D., Huntley, J., Fierer, N., *et al*. (2012) Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J* **6:** 1621–1624.

Claesson, M.J., O'Sullivan, O., Wang, Q., Nikkila, J., Marchesi, J.R., Smidt, H., *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.

Claesson, M.J., Wang, Q., O'Sullivan, O., Greene-Diniz, R., Cole, J.R., Ross, R.P., and O'Toole, P.W. (2010) Comparison of two next-generation sequencing technologies for resolving highly complex microbiota composition using tandem variable 16S rRNA gene regions. *Nucleic Acids Res* **38:** e200.

Cole, J.R., Wang, Q., Cardenas, E., Fish, J., Chai, B., Farris, R.J., *et al*. (2009) The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res* **37:** D141–D145.

Crespo, B.G., Pommier, T., Fernández-Gómez, B., and Pedrós-Alió, C. (2013) Taxonomic composition of the particle-attached and free-living bacterial assemblages in the Northwest Mediterranean Sea analyzed by pyrosequencing of the 16S rRNA. *MicrobiologyOpen* **2:** 541–552.

DeLong, E.F. (2009) The microbial ocean from genomes to biomes. *Nature* **459:** 200–206.

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

Falkowski, P.G., Fenchel, T., and Delong, E.F. (2008) The microbial engines that drive Earth's biogeochemical cycles. *Science* **320:** 1034–1039.

Ghai, R., Hernandez, C.M., Picazo, A., Mizuno, C.M., Ininbergs, K., Diez, B., *et al*. (2012) Metagenomes of Mediterranean coastal lagoons. *Sci Rep* **2:** 490.

Glenn, T.C. (2011) Field guide to next-generation DNA sequencers. *Mol Ecol Resour* **11:** 759–769.

Gonzalez, J.M., Portillo, M.C., Belda-Ferre, P., and Mira, A. (2012) Amplification by PCR artificially reduces the proportion of the rare biosphere in microbial communities. *PLoS ONE* **7:** e29973.

Haas, B.J., Gevers, D., Earl, A.M., Feldgarden, M., Ward, D.V., Giannoukos, G., *et al*. (2011) Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Res* **21:** 494–504.

Hao, X., and Chen, T. (2012) OTU analysis using metagenomic shotgun sequencing data. *PLoS ONE* **7:** e49785.

Harismendy, O., Ng, P.C., Strausberg, R.L., Wang, X., Stockwell, T.B., Beeson, K.Y., *et al*. (2009) Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome biol* **10:** R32.

Hillis, D.M., and Dixon, M.T. (1991) Ribosomal DNA – molecular evolution and phylogenetic inference. *Q Rev Biol* **66:** 410–453.

Hong, S., Bunge, J., Leslin, C., Jeon, S., and Epstein, S.S. (2009) Polymerase chain reaction primers miss half of rRNA microbial diversity. *ISME J* **3:** 1365–1373.

Huse, S.M., Huber, J.A., Morrison, H.G., Sogin, M.L., and

Welch, D.M. (2007) Accuracy and quality of massively parallel DNA pyrosequencing. *Genome biol* **8:** R143.
Huse, S.M., Dethlefsen, L., Huber, J.A., Mark Welch, D., Relman, D.A., and Sogin, M.L. (2008) Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing. *PLoS Genet* **4:** e1000255.
Huse, S.M., Welch, D.M., Morrison, H.G., and Sogin, M.L. (2010) Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environ Microbiol* **12:** 1889–1898.
Karsenti, E., Acinas, S.G., Bork, P., Bowler, C., De Vargas, C., Raes, J., *et al*. (2011) A holistic approach to marine eco-systems biology. *PLoS biol* **9:** e1001177.
Liu, Z., Lozupone, C., Hamady, M., Bushman, F.D., and Knight, R. (2007) Short pyrosequencing reads suffice for accurate microbial community analysis. *Nucleic Acids Res* **35:** e120.
Logares, R., Haverkamp, T.H., Kumar, S., Lanzen, A., Nederbragt, A.J., Quince, C., and Kauserud, H. (2012) Environmental microbiology through the lens of high-throughput DNA sequencing: synopsis of current platforms and bioinformatics approaches. *J Microbiol Methods* **91:** 106–113.
Miller, C.S., Baker, B.J., Thomas, B.C., Singer, S.W., and Banfield, J.F. (2011) EMIRGE: reconstruction of full-length ribosomal genes from microbial community short read sequencing data. *Genome Biol* **12:** R44.
Minoche, A.E., Dohm, J.C., and Himmelbauer, H. (2011) Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems. *Genome Biol* **12:** R112.
Mizrahi-Man, O., Davenport, E.R., and Gilad, Y. (2013) Taxonomic classification of bacterial 16S rRNA genes using short sequencing reads: evaluation of effective study designs. *PLoS ONE* **8:** e53608.
Nakamura, K., Oshima, T., Morimoto, T., Ikeda, S., Yoshikawa, H., Shiwa, Y., *et al*. (2011) Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Res* **39:** e90.
Pace, N.R. (1997) A molecular view of microbial diversity and the biosphere. *Science* **276:** 734–740.
Pedrós-Alió, C. (2006) Marine microbial diversity: can it be determined? *Trends Microbiol* **14:** 257–263.
Pedrós-Alió, C. (2012) The rare bacterial biosphere. *Annu Rev Mar Sci* **4:** 449–466.
Pommier, T., Neal, P.R., Gasol, J., Coll, M., Acinas, S.G., and Pedros-Alio, C. (2010) Spatial patterns of bacterial richness and evenness in the NW Mediterranean Sea explored by pyrosequencing of the 16S rRNA. *Aquat Microb Ecol* **61:** 221–233.
Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., *et al*. (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* **41:** D590–D596.
Quince, C., Lanzen, 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.
Quince, C., Lanzen, A., Davenport, R.J., and Turnbaugh, P.J. (2011) Removing noise from pyrosequenced amplicons. *BMC Bioinformatics* **12:** 38.
Rodrigue, S., Materna, A.C., Timberlake, S.C., Blackburn, M.C., Malmstrom, R.R., Alm, E.J., and Chisholm, S.W. (2010) Unlocking short read sequencing for metagenomics. *PLoS ONE* **5:** e11840.
Rusch, D.B., Halpern, A.L., Sutton, G., Heidelberg, K.B., Williamson, S., Yooseph, S., *et al*. (2007) The Sorcerer II Global Ocean Sampling Expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol* **5:** e77.
Shah, N., Tang, H., Doak, T.G., and Ye, Y. (2011) Comparing bacterial communities inferred from 16S rRNA gene sequencing and shotgun metagenomics. *Paci Symp Biocomput***:** 165–176.
Shakya, M., Quince, C., Campbell, J.H., Yang, Z.K., Schadt, C.W., and Podar, M. (2013) Comparative metagenomic and rRNA microbial diversity characterization using archaeal and bacterial synthetic communities. *Environ Microbiol* **15:** 1882–1899.
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.
Sul, W.J., Oliver, T.A., Ducklow, H.W., Amaral-Zettler, L.A., and Sogin, M.L. (2013) Marine bacteria exhibit a bipolar distribution. *Proc Natl Acad Sci USA* **110:** 2342–2347.
Wang, Q., Garrity, G.M., Tiedje, J.M., and Cole, J.R. (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* **73:** 5261–5267.
Werner, J.J., Zhou, D., Caporaso, J.G., Knight, R., and Angenent, L.T. (2012) Comparison of Illumina paired-end and single-direction sequencing for microbial 16S rRNA gene amplicon surveys. *ISME J* **6:** 1273–1276.
Yilmaz, P., Kottmann, R., Pruesse, E., Quast, C., and Glockner, F.O. (2011) Analysis of 23S rRNA genes in metagenomes – a case study from the Global Ocean Sampling Expedition. *Syst Appl Microbiol* **34:** 462–469.
Zinger, L., Amaral-Zettler, L.A., Fuhrman, J.A., Horner-Devine, M.C., Huse, S.M., Welch, D.B., *et al*. (2011) Global patterns of bacterial beta-diversity in seafloor and seawater ecosystems. *PLoS ONE* **6:** e24570.

Supporting information

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

Fig. S1. Pipeline flowchart showing all steps performed for quality filtration, classification and OTU assignation of $_{mi}$tags.
Fig. S2. Coverage analysis of $_{mi}$tags against one reference 16S rDNA sequence from *Escherichia coli* (Brosius *et al*., 1978). The upper part of the figure shows the distribution of $_{mi}$tags (horizontal lines) along the 16S rRNA gene, taking into consideration the nucleotide coordinates and the percentage of identity (left scale) with regard to the reference gene as well as their density (grey scale; clear grey indicates low density). The lower part of the figure shows the coverage as the number of times that a nucleotide position in the reference gene is covered by a $_{mi}$tag (right scale). Horizontal red lines were plotted in accordance with the nucleotide coordinates of the hypervariable regions (V1 to V9).

Fig. S3. RDP classification. Mean classification confidence values across taxonomic levels calculated with RDP classifier for $_{mi}$tags, $_{m454}$tags and $_{454}$tags. The dashed lines indicate the standard deviation (±). Only confidence values > 0.5 were considered.

Fig. S4. Rarefaction analysis of 16S $_{mi}$tags for the size fractions 0.2–1.6 and 0.8–5 μm. All $_{mi}$tags were used, covering all possible V regions of the 16S rRNA gene.

Fig. S5. Venn diagrams. $_{mi}$Tags are indicated in salmon and $_{454}$tags in green.

A. Shared and unique taxonomic ranks (phylum, class, order, family and genus) recovered with $_{mi}$tags and $_{454}$tags using the entire dataset (i.e. $_{mi}$tags spanning the entire 16S rDNA and nontrimmed $_{454}$tags; no subsampling has been carried out). Classifications were done using the RDP classifier (see Methods).

B. Shared and unique OTUs recovered with $_{mi}$tags and $_{454}$tags; dataset restricted to the V1–V3 region, with (right side) and without (left side) subsampling.

Fig. S6. Phylum coverage of the pair set primers (27Fmod/533R) used for 16S rDNA amplicon tags ($_{454}$tags).

Fig. S7. Rank–abundance curves using different datasets from six samples retrieved from the bacterial fraction (see Table 1).

A. All $_{mi}$Tags are contrasted with $_{454}$tags considering full datasets as well as those with subsampling (2000 reads per sample) and trimmed $_{454}$tags (100–150 bp), i.e. using TARA-ALL and TARA-TRIMMED OTs with and without subsampling.

B. Here, results are presented for $_{mi}$tags falling into the V1–V3 region, while the rest of the comparisons are the same as in A, i.e. using TARA-V1–V3 and TARA-V1–V3-TRIMMED OTs with and without subsampling.

Fig. S8. Comparison between $_{mi}$tags, $_{454}$tags and flow cytometry (FC). A best linear fit was adjusted and Pearson's correlation coefficient was calculated for each plot.

A. Quantitative comparison of relative abundances of $_{mi}$tags and FC *Prochlorococcus* counts. Pearson's $r = 0.782$; $P < 0.001$.

B. $_{mi}$Tags vs FC *Synechococcus* counts. Pearson's $r = 0.603$; $P < 0.001$.

Fig. S9. Dendrogram based on UPGMA clustering of Bray–Curtis distances between samples analysed with $_{mi}$tags and $_{454}$tags. All the analysed samples belong to four size fractions (0.2–1.6, 0.8–5, 5–20 and 20–180 μm), and all were subsampled to 2000 reads per sample to correct for unequal sampling efforts (a number of samples, in particular all those analysed with $_{m454}$tags, did not reach that number, and for that reason they were excluded from this analysis). The four analysed datasets are shown in panels A–D. In all datasets, singletons and OTUs present in only one sample were included. Jackknife support (subsampling = 2000) is indicated with an asterisk (50–75% support) or double asterisk (> 75% support). In each panel, the position for the $_{454}$tags and $_{mi}$tags is indicated with a box. The size fraction from which each sample originated is indicated with colored dots.

A. $_{mi}$Tags spanning the entire 16S rDNA.

B. Same dataset as in A but including trimmed $_{454}$tags.

C. $_{mi}$Tags belonging to the V1–V3 rDNA region only.

D. $_{mi}$Tags belonging to the V1–V3 region plus trimmed $_{454}$tags.

The two branches in blue indicate the only two samples that clustered together despite having been analysed with different approaches and platforms.

Table S1. General description of the investigated *Tara* Oceans samples.

Table S2. List of the *Tara* Oceans metagenomes sequenced by Illumina used for this study.

Table S3. List of the *Tara* Oceans metagenomes sequenced by 454 Titanium pyrosequencing used for this study.

Table S4. List of the *Tara* Oceans samples sequenced by 454 Titanium pyrosequencing using 16S amplicons ($_{454}$tags). For comparative analysis, clustering analyses were done using two approaches: (i) using an OTU reference database (SILVA 108 release); and (ii) *de novo* clustering for two independent V regions (V1 and V3).

Table S5. OTU comparison of the six samples sequenced with both the 454 and Illumina sequencing platforms.

Table S6. Estimation of several diversity indexes for different *Tara* Oceans datasets.

Table S7. Unique phylums, classes, orders, families and genera detected by 16S $_{mi}$tags that were not detected by 16S $_{454}$tags.

Table S8. Unique genera detected by tags ($_{454}$tags) that were not detected by $_{mi}$tags.

Table S9. Contingency table of OTU detection versus presence of primer mismatches.

Table S10. Correlation and linear fit of OTU abundances estimated by $_{mi}$tag, $_{454}$tag and $_{m454}$tag.