To be submitted: *Environmental Microbiology*
Scope: genomics, functional genomics, environmental genomics/metagenomics, bioinformatic analyses and comparative genomics

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*]

Affiliations:
[1] Department of Marine Biology and Oceanography, Institute of Marine Science (ICM), CSIC, Passeig Marítim de la Barceloneta, 37-49, ES-08003, Barcelona, Spain.
[2] European Molecular Biology Laboratory, Meyerhofstr. 1, 69117 Heidelberg, Germany
[3] Federal University of Rio Grande do Norte, Department of Oceanography and Limnology, Pós-graduação em Ecologia, 59014-002 Natal, 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] CNRS, Université Pierre et Marie Curie, UMR 7144, Station Biologique de Roscoff, Roscoff, FR-29682, France.
[7] Research group of Bioinformatics and (eco-)systems biology, Department of Structural Biology, VIB, Pleinlaan 2, 1050 Brussels, Belgium
[8] Research group of Bioinformatics and (eco-)systems biology, Microbiology Unit (MICR), Department of Applied Biological Sciences (DBIT), Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium
[9] CEA, IG, Genoscope, 2 rue Gaston Crémieux 91057 Evry, France
[10] CNRS-UMR 8030, 2 rue Gaston Crémieux 91057 Evry, France
[11] Université d'Evry, boulevard François Mitterrand, 91025 Evry, France.

* To whom correspondence should be addressed: E-mail sacinas@icm.csic.es, Tel (+34) 93 230 8565; Fax (+34) 93 230 95 55.

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/1462-2920.12250

Running Title: Using $_{mi}$Tags as alternative approach to explore diversity and structure of microbial communities.

Summary:

Sequencing of 16S rDNA PCR-amplicons is the most common approach to investigate environmental prokaryotic diversity, despite the known biases introduced during PCR. Here we show that 16S rDNA fragments derived from Illumina-sequenced environmental metagenomes ($_{mi}$Tags) 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 gigabases). For comparative analyses, a subset of samples was also selected for Roche-454 sequencing using both shotgun ($_{m454}$Tags; 13 metagenomes, ca. 2.4 Gb) and 16S rDNA amplicon ($_{454}$Tags; ca. 0.075 Gb) approaches. Our results indicate that by overcoming PCR biases related to amplification and primer mismatch, $_{,mi}$Tags may provide more realistic estimates of community richness and evenness than amplicon $_{454}$Tags. In addition, $_{mi}$Tags can capture expected beta diversity patterns. Using $_{mi}$Tags is now economically feasible due to 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.

Keywords: $_{mi}$Tags, Illumina metagenomes, 16S rDNA, $_{454}$Tags, PCR biases, microbial diversity, microbial community structure; environmental prokaryotic communities

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 waiting for technological progress to be investigated. The advent of High-Throughput Sequencing (HTS) technologies (e.g. 454 & 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)_ENREF_5_ENREF_12_ENREF_7. So far, only a handful of studies have used Illumina-sequenced PCR amplicons ($_{i}$Tags) to explore natural microbial assemblages (Caporaso et al., 2011; Caporaso et al., 2012; Werner et al., 2012; Bokulich et al., 2013)_ENREF_6. 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). Since both technologies became popular in microbial ecology relatively recently, a careful evaluation of their performances and biases is still ongoing (Huse et al., 2007; Quince et al., 2009; Claesson et al., 2010; Huse et al., 2010; Minoche et al., 2011; Nakamura et al., 2011; Quince et al., 2011). A limited number of HTS cross-platform studies have indicated different biases associated to 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)_ENREF_18, thus validating these approaches. Still, regardless of the sequencing technology, the biases associated to 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)_ENREF_20_ENREF_21.

An alternative approach to circumvent PCR is to identify rDNA fragments from metagenomic data (hereafter $_{m}$Tags). Until recently this approach was unrealistic, since 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 4,100 turned out to be usable 16S rDNA reads (0.05%; see CAMERA, http://camera.calit2.net/). In the second release of GOS, a 1.4% was detected with a total of 142,783 16S rDNA fragments from 80 GOS metagenomes (Yilmaz et al., 2011). Other metagenomic studies based on 454 FLX Titanium sequencing identified hundred to thousand of rDNA fragments (hereafter $_{m454}$Tags) after sequencing several millions 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. Yet, the High-Throughput of Illumina HiSeq platforms circumvent 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 amount 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 to 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 (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 2x150bp and 2x250bp 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 depends on what region is being used (Claesson et al., 2009; Engelbrektson et al., 2010; Mizrahi-Man et al., 2013)_ENREF_21_ENREF_30, 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 reconstructing 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 profiling 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 vs. non-PCR-based (Tags vs. $_{m}$Tags) are very limited. Still, a recent study using synthetic microbial communities tested the capacity of PCR vs. non PCR-based sequencing at 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 to quantitatively uncover 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 which are in many cases 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 with 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 & 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, 1) avoids PCR biases and 2) a large amount of functional data is simultaneously

produced when miTags are generated. Thus, miTags are a powerful alternative to the commonly used amplicon-based Tags for community analyses. Using miTags is now feasible thanks to the dramatic decrease in sequencing costs.

Results

Thousands of 16S $_{mi}$Tags covering all 16S rDNA gene regions can be extracted from metagenomes and taxonomically classified to RDP. The 29 Illumina metagenomes from the three analyzed 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 Fig. S1 On average, 2.08 x10^4 16S $_{mi}$Tags > 100 bases were extracted per sample (metagenome), although 7.9x10^4 16S $_{mi}$Tags were retrieved from typically free-living bacterial size fraction (0.2-1.6 μm) (Table S2). Altogether, these $_{mi}$Tags covered all 16S rDNA hypervariable regions (V1 to V9) with a decrease in coverage at the 16S extremes (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 RDP database (Cole et al., 2009) when using the naïve Bayesian classifier (Wang et al., 2007), albeit $_{mi}$Tags had shorter sequence length (Fig. S3).

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 OTUs (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 region V1-V3 (Table S2), which was later used in comparative cross-platform analyzes. This number was expected when considering a more or less uniform read coverage of the 16S rDNA (about 1,300 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 highly used for marine $_{454}$Tags 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; Table S3). The number of $_{454}$Tags that could be assigned to OTUs was slightly smaller (about 86%; 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 region V1 (287-1204) and V3 (310-1443) was not different to what was obtained by assignation to reference OTUs (524-1070) (ANOVA; P-value >0.58) (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 (Table S5, Fig. 1A). When using a subsampling of 2,000 reads/sample, the increase is between 31.1 to 43.2% of OTUs per sample (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 (ACE) (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 ($_{454}$Tags-trimmed), both $_{mi}$Tags and $_{454}$Tags-trimmed recovered similar numbers of OTUs, ranging between 994-1178 for $_{mi}$Tags and 586-1824 for $_{454}$Tags-trimmed (Fig. 1B). Values were even closer when subsampling at 2,000 reads per sample ($_{mi}$Tags: 428-508 OTUs and $_{454}$Tags-trimmed: 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 μ m indicated a larger richness in the 0.8-5 μ m size fraction (Fig. S4). Interestingly, it was in the size fractions > 5 μ m wherein the number of mapped $_{mi}$Tags to reference OTUs dropped to 58% (Table S2) suggesting prokaryote novelty probably associated to larger particles.

We compared the capability of $_{mi}$Tags and $_{454}$Tags to detect prokaryote taxonomic diversity using both single reads as well as OTUs. At higher-rank taxonomic levels, $_{mi}$Tags uniquely recovered several phyla (e.g. *Fibrobacteres* and *Tenericutes*) and classes (*Halobacteria*, *Chloroflexi*) (Table S7) in RDP classifications (Cole et al., 2009). At lower-rank levels, we found 748 genera that were exclusively detected by $_{mi}$Tags (Fig. S5A; Table S7), whereas only nine genera were exclusively detected by $_{454}$Tags (Table S8). Similar results were obtained in OTU-based analyses; when using both the TARA-V1-V3 dataset with and without subsampling (see Fig.1S). 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 (Fig. S5B; left panel). For the subsampled dataset, normalization corrected artifacts that produced some of the differences between techniques, but still 446 OTUs were exclusively obtained by $_{mi}$Tags and 274 OTUs by $_{454}$Tags (Figure 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 $_{mi}$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 labeled with numbers in Fig. 2 and Table S7). Furthermore, $_{mi}$Tags retrieved Archaea, which were expectedly absent in $_{454}$Tags due to the use of bacterial primers.

The primer bias effect, as a potential explanation for the differences in OTU detection between both techniques, was furthered investigated in two fronts by (i)

analyzing the *in silico* coverage of the primer-pair set used for generating 16S rDNA amplicons 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 test the theoretical accuracy of the primer pair (27Fmod/533R). This pair covered 78.9% of the references and was well distributed across main phyla, where ranged between 60-100% coverage (Fig. S6) A few phyla were poorly represented in terms of coverage probably due to low number of sequences available in datasets (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). We found a strong and significant dependence between OTUs detected only by $_{mi}$Tags with $_{454}$Tags and the presence of mismatches (χ^2=53.04, df=1, p<0.0001) (Table S9). Conversely, when we selecting only the OTUs detected with $_{454}$Tags, the OTU detection with $_{mi}$Tags and the presence of mismatches appeared as independent factors (χ^2=1.45, df=1, p=0.2284) (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 to any phyla in particular although differences exist in the coverage within main phylum.

Further comparative analyses focused on the evenness patterns retrieved by $_{mi}$Tags and $_{454}$Tags (Fig. S7). First, similar rank-abundance curves were observed when samples were subsampled (Fig.S7, right panel); however, some differences emerged when using data non-subsampled. Interestingly, $_{mi}$Tags tended to recover a higher number of very low abundant taxa (<0.1%) from the rare biosphere (Pedrós-Alió, 2012) (Fig. S7, left panel). Despite the overall similarity in rank-abundance, different platforms (454 vs. Illumina) and approaches (Tags [amplicon-derived] vs. $_{mi}$Tags)

indicated, in several cases, different abundances for the same OTUs (Fig. S7, left panel; Fig 3, panels A and B). When OTU abundances derived from $_{mi}$Tag, $_{454}$Tag and $_{m454}$Tag were compared, a better agreement was found between approaches not involving PCR ($_{m454}$Tag vs. $_{mi}$Tag) resulting in a higher correlation and a fit closer to the 1:1 line (Fig.3; 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 overestimated with $_{454}$Tag compared to $_{m}$Tag (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: CARD-FISH counts (Fig. 4) and flow cytometry (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. 5; CARD-FISH vs. $_{mi}$Tags: Pearson r=0.866; p <0.001 and CARD-FISH vs. $_{454}$Tags: Pearson r=0.948; p<0.001). Similarly, a positive correlation was observed between cyanobacteria abundance (*Prochlorococcus* and *Synechococcus*) measured by flow cytometry and $_{mi}$Tags-derived abundance (*Prochlorococcus*: Pearson's r=0.782, p <0.001; *Synechococcus*: Pearson's r=0.603, p <0.001; Fig. S8).

Comparative community structure using $_{mi}$Tags and $_{454}$Tags. UPGMA clustering analysis based on Bray Curtis distances was performed for the four analyzed datasets (TARA-ALL, TARA-TRIMMED, TARA-V1-V3, TARA-V1-V3-TRIMMED; see methods and Fig. S1) after subsampling them to 2,000 reads per sample (Fig. S9, panels A-D). In three out of the four datasets, the $_{454}$Tag samples clustered together instead of

with their corresponding $_{mi}$Tag samples (Fig. S9, panels A-C). Only in the dataset considering trimmed $_{454}$Tags and the V1-V3 region (TARA-V1-V3-TRIMMED), one sample analyzed with $_{454}$Tags clustered with the same sample analyzed with $_{mi}$Tags (Fig. S9, panel D). Furthermore, in this latter dataset, samples from the prokaryote size fraction (0.2-1.6 μ m) analyzed with $_{454}$Tags and $_{mi}$Tags clustered together forming a tight group (Fig. S9, panel D). The absence of clustering of the same samples analyzed 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: *r* (pearson) =0.75, p=0.002) between the same set of samples analyzed 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: *r* (pearson) =0.44, p=0.023). This discrepancy could be associated to PCR biases affecting the relative abundance of taxa measured by $_{454}$Tags.

315 **Discussion**

In our metagenomic samples, miTags accounted for about 0.01-0.1% of the total reads, which is within the expected range. This 0.1% 16S rDNA recovery rate reported 320 here and in previous studies (Rusch *et al.*, 2007) seems to be independent from the sequencing technology (Sanger shotgun, Roche-454 and Illumina) providing a good plausibility check for metagenome sequencing projects. Due to the high throughput of Illumina platforms, the number of miTags recovered per sample (79,000 miTags on average for bacterial size fraction) can be considered more than sufficient for capturing 325 community composition patterns (Caporaso et al., 2011). As expected, the yield of miTags for the typical bacteria size-fraction was higher (about 0.09%) than for size fractions > 5 μ m (0.01%). Most miTags (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 330 samples. In fact, in another work, we have extracted all miTags for 72 globally distributed samples of 35 TARA-Oceans stations that represented surface, deep chlorophyll maximum (DCM), oxygen minimum zone (OMZ) and mesopelagic water samples, which showed similar miTags mapping percentages as for the three previous marine stations (Salazar et al., unpublished). Similarly, using RDP, most miTags (99%) 335 could be confidently classified and in all cases, as it was expected, classification confidence decreased with lowering taxonomic levels (Claesson et al., 2010).

In this work, we assigned miTags to the reference OTUs derived from clustering the SILVA 108 reference database at 97 % of similarity. This approach may have at least two drawbacks: (i) if a sample contains OTUs that are not present in the reference 340 database, then they will not be accounted. Nevertheless, we found that most (>94 %)

16S miTags from marine samples were assigned to reference OTUs, indicating that SILVA 108 is appropriate for typical marine surface studies. The second possible drawback (ii) is that miTags are shorter than 454Tags, and they contain less information for taxonomic assignment; this may be further complicated if a specific miTag cover a conserved 16S rDNA region. Thus, miTags may produce some diversity inflation, as different segments of the same 16S rDNA 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 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 to lineages not recovered with 454Tags (Fig. 2) due to primer mismatches (Table S9). A future 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, which 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 vs. 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 is omitted, which could be associated to potential PCR biases on amplicon-derived approaches.

$_{mi}$Tags as an alternative for probing microbial diversity. The generation of $_{mi}$Tags does not require long PCR steps, 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)_ENREF_37. 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 $_{mi}$Tags recovered more taxa at different taxonomic levels and OTUs than $_{454}$Tags. The recovery of more OTUs using $_{mi}$Tags could be related, to 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, classes as well as other lower rank taxonomic levels indicates that $_{mi}$Tags recover OTUs that are probably missed during the PCR step before $_{454}$Tag 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 $_{mi}$Tags (Fig. 2). Furthermore, the lack of detection of several OTUs with $_{454}$Tags was statistically proved to be related to primer mismatches, while there was no primer bias when testing the $_{mi}$Tags approach (Table S9).

Not only did $_{mi}$Tags and $_{454}$Tags 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 $_{m454}$Tags and $_{454}$Tags. Some OTUs were abundant among $_{454}$Tags and rare with e.g. $_{m454}$Tags or $_{mi}$Tags 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 $_{mi}$Tags captured more members of the rare biosphere than $_{454}$Tags Using a different dataset from deep ocean marine microbial communities, we performed a comparison between $_{mi}$Tags and $_{i}$Tags retrieving a similar picture as for $_{mi}$Tags vs. $_{454}$Tags (Salazar et al., unpublished).

Finally, we have analyzed the sequencing platform effect by comparing $_{mi}$Tags and $_{m454}$Tags and the approach effect (amplicon PCR $_{454}$Tags vs $_{m}$Tags). Despite the observed deviations from a linear relationship, the non-PCR scenarios provided the most compatible results, thus supporting the use of metagenomic Tags ($_{m}$Tags) for community profiling (Fig. 3, panel C). Lastly, quantitative techniques different from rDNA sequencing (i.e. FISH & 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, pointed out that metagenomics (both 454 and Illumina) outperformed amplicon 16S Tags sequencing to quantitatively reconstruct 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 summarized the characteristics of the analyzed approaches and platforms in Table 1. Depending on research goals different possibilities emerge. The longer sequences provided by 454-Roche platforms (up to 800-1000 bp) still are highly valuable to facilitate accurate assemblies 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 having a very large amount of samples. Illumina metagenomes can be done with as a little as 100 ng of DNA, and it is important to remark 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 x 250 bp paired-end reads) improving OTU assignation and taxonomic classifications. Thus, the $_{mi}$Tags approach will become more powerful and accessible in cost terms with the advance of High Throughput Sequencing technologies.

Experimental procedures

Detailed section of the experimental procedures can be found in the online version of this article under Supplementary Information

Building the $_{mi}$Tags, $_{m454}$Tags, and $_{454}$Tags datasets. From the 29 analyzed metagenomes, a total of 5.03 x 10^9 and 1.79 x 10^9 raw and merged paired-end metagenomic reads respectively were produced for Illumina (> 100 bp, GAIIx & HiSeq2000; Table S2). This represents about 700 giga bases (Gb) of metagenomic sequence data. From these libraries, 6.05 x10^5 16S $_{mi}$Tags > 100 bp were extracted (Table S2). Using 454 GS FLX Titanium platform, a total of 8.1 x 10^6 reads from 13 metagenomes were produced (about 2.4 Gb) and 3.30 x10^3 $_{m454}$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, Supplementary Table S2). Similar values were obtained using $_{m454}$Tags (mean 0.11 %; Table S3). Due to the higher sequencing depth allowed by the Illumina platform (about 15 Gb per metagenome in our samples), we were able to extract between 5-9 x 10^4 16S reads (> 100 bp) ($_{mi}$Tags) per metagenome from the prokaryote size-fraction (Supplementary Table S2). A much smaller number of $_{m454}$Tags was recovered due to the more limited throughput of the 454 GS FLX Titanium platform (Supplementary 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), totaling 2.63 x 10^5 reads. After a stringent quality filtering, this dataset was reduced to 1.53 x 10^5 $_{454}$Tags (Supplementary Table S4). Using $_{454}$Tags, we obtained between 2.88 – 7.00 x 10^4 reads (> 100 bp) per sample (Supplementary 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 (ENA) 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-Ti pyrosequencing technology ($_{m454}$Tags): ERA155563, ERA155562; (ii) Shotgun Sequencing of Tara Oceans DNA samples corresponding to size fractions for plankton larger size fractions (0.8-5, 5-20, 20-180 and 180-2000 μm) performed by Illumina technology ($_{mi}$Tags): ERA242028 and 454-Ti 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-Ti pyrosequencing technology: ERA242032.

Analyzed datasets (OTU tables). A total of four main OTU tables were constructed: the 1) TARA-ALL OTU table (*OT*), contained all $_{mi}$Tags, $_{m454}$Tags and $_{454}$Tags, while the 2) TARA-TRIMMED *OT* contained the same data as in 1) but here the $_{454}$Tags were trimmed to 100–150 bp. The *OT* 3) TARA-V1-V3 included only Tags that fell within the V1-V3 region, and the *OT* 4) TARA-V1-V3-TRIMMED, comprised $_{mi}$Tags within the V1-V3 region and trimmed $_{454}$Tags (100–150 bp). Finally, all four *OT* were subsampled (in QIIME) to 2,000 reads per sample, to correct for potential biases introduced by unequal sequencing effort. 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 and the following institutions and sponsor 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 and crew. TARA Oceans would not exist without the continuous support of the participating institutes (see Karsenti et al., 2011). This is contribution no. XXX of the Tara Oceans Expedition 2009-2012. We thank Dr. Josep M. Gasol for critical reading and Dr. Pedrós-Alió for his helpful comments. SGA was supported by a Ramon y Cajal contract from 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 (AGAUR) and MicroOcean PANGENOMICS (CGL2011-26848/BOS) to SGA by the Spanish Ministry of Science and Innovation (MICINN). RL 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 GS and FMC were supported by Ph.D. 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. Additionally funding was provided by the "Agence Nationale de la Recherche", ANR grants Prometheus ANR-09-GENM-031, Poseidon ANR-09-BLAN-0348 and Tara-Girus ANR-09-PCS-GENM-218. S.S. and P.B. were supported by EMBL core

500 funding and GL and JR are supported by the Fund for Scientific Research Flanders (FWO). Supplementary information is available at EMIs 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. *Applied and environmental microbiology* **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 biology* **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 U S A* **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. *Proceedings of the National Academy of Sciences of the United States of America* **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., 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 research* **38**: e200.

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.

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**: D141-145.

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. *Microbiology Open* **In press**.

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 research* **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 biology* **10**: R32.
Hillis, D.M., and Dixon, M.T. (1991) Ribosomal DNA - Molecular evolution and phylogenetic inference. *Quarterly Review of Biology* **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. *The ISME journal* **3**: 1365-1373.
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. *Environmental Microbiology* **12**: 1889-1898.
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 biology* **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.
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 biology* **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 research* **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. *Journal of microbiological 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 biology* **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 research* **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 in Microbiology* **14**: 257-263.
Pedrós-Alió, C. (2012) The rare bacterial biosphere. *Annual Review of Marine Science* **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. *Aquatic Microbial Ecology* **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**: D590-596.
Quince, C., Lanzen, A., Davenport, R.J., and Turnbaugh, P.J. (2011) Removing noise from pyrosequenced amplicons. *Bmc Bioinformatics* **12**: 38.
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.
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. In *Pacific Symposium on Biocomputing*, pp. 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 U S A* **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 U S A* **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.

Table 1. General comparison of the different platforms and approaches

Platform / Approach[1]	Template	Coverage	16S rDNA specificity	16S rDNA recovery[2]	PCR bias[3]	16S rDNA overlap[4]	Taxonomic definition[5]	OTU Clustering[6]	€/Mb[7]
$_{mi}$Tags	Metagenomic fragments	16S rDNA + functional metagenomic	Spanning all 16S rDNA	High / Medium	Absent	Low	Variable	Map to reference OTUs/ V region selection for de-novo	0.1 / 100* (HiSeq)
$_{m454}$Tags	Metagenomic fragments	16S rDNA + functional metagenomic	Spanning all 16S rDNA	Very Low	Absent	Low	Variable	Map to reference OTUs/ V region selection for de-novo	12 / 12000* (Titanium)
$_{454}$Tags	Amplicons	16S rDNA only	Specific 16S rDNA area	High / Very High	Present	High	High	De-novo & Map to Reference OTUs	12 (Titanium)
$_{i}$Tags	Amplicons	16S rDNA only	Specific 16S rDNA area	Very High	Present	High	High / Medium	De-novo & Map to Reference OTUs	0.7 (MiSeq)

[1] 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)

[2] Number of recovered 16S rDNA reads from the used template. Estimations depend on the throughput of the platform

[3] PCR bias refers mostly to known primer biases and chimera formation

[4] Overlapping of the recovered 16S rDNA fragments. 16S recovered from metagenomes show a limited overlapping that preclude typical clustering techniques

[5] Taxonomic information associated to 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)

[6] OTU clustering methods that can be used with the different approaches

[7] Approximate costs per million of base pairs. Based on Glenn ((2011)). *Costs to generate $_{mi}$Tags/$_{m454}$Tags disregarding all the remaining data that is not 16S; for example, about 1 Gb of metagenomic data needs to be sequenced to obtain 1 Mb of metagenomic Tags, and the cost to generate 1 Gb is reported.

Titles and legends to figures

Fig. 1. Rarefactions. Rarefaction analyses using two datasets. In panel A) only the dataset including all $_{mi}$Tags and the $_{454}$Tags was considered. In panel B) the dataset considered included $_{mi}$Tags falling into the V1-V3 region and trimmed $_{454}$Tags. Thus, panel A) represents the actual gathered data and panel B) the data most comparable between platforms and approaches. 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 due to the different characteristics of the datasets. Also note that the vertical axes have different lengths. The horizontal arrow indicates the maximum vertical value of B) in A).

Fig. 2. Phylogenetic Tree. Phylogeny of the OTUs recovered with $_{mi}$Tags (V1-V3) and $_{454}$Tags where all samples were subsampled to 2,000 reads per sample (TARA-V1-V3 OT with subsampling). $_{mi}$Tags are indicated in green and $_{454}$Tags 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 $_{454}$Tags and/or $_{mi}$Tags. A zoom of two selected areas of the tree is presented in boxes A & B. Box A exemplifies that relative abundance estimated by $_{mi}$Tags and $_{454}$Tags can be either very similar or different for evolutionary related OTUs. Box B exemplifies that several evolutionary related OTUs (probably groups) might be recovered by $_{mi}$Tags and not by $_{454}$Tags (and vice versa). Examples similar to the ones presented in Boxes A & B were observed throughout the entire phylogeny. Unique clusters of OTU from different phylogenetic taxa retrieved only by $_{mi}$Tags and absent by $_{454}$Tags are represented by

numbers from 1 to 5. Main taxonomic groups are indicated by the tree leave's color and corresponded to the legend at the bottom of the figure.

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

Fig. 4. Comparison between $_{mi}$Tags, $_{454}$Tags and CARD-FISH. Quantitative comparison of relative abundances of $_{mi}$Tags (empty circles) with CARD-FISH counts or $_{454}$Tags (full triangles) vs. CARD-FISH. Relative abundances (%) of four different prokaryote groups (Bacteroidetes, Gammaproteobacteria, Roseobacter and SAR11) estimated with CARD-FISH are compared to $_{mi}$Tags and $_{454}$Tags estimates. A linear model was adjusted and 95% confidence intervals were computed for the slope.

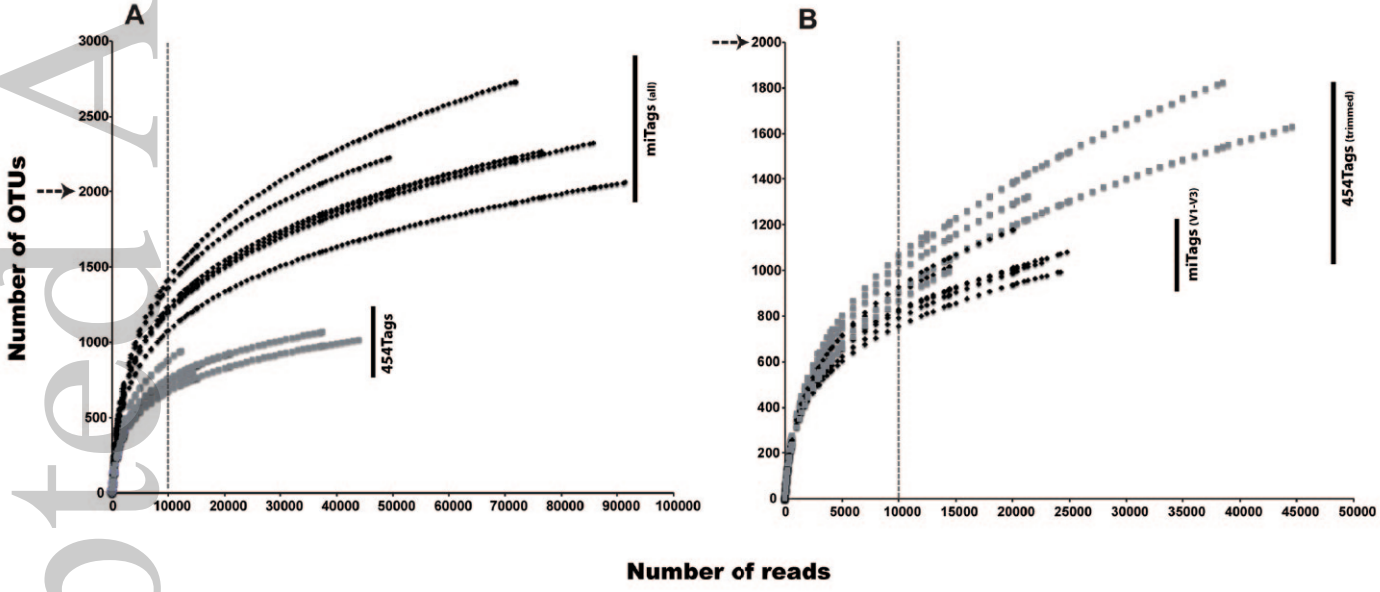

Figure1

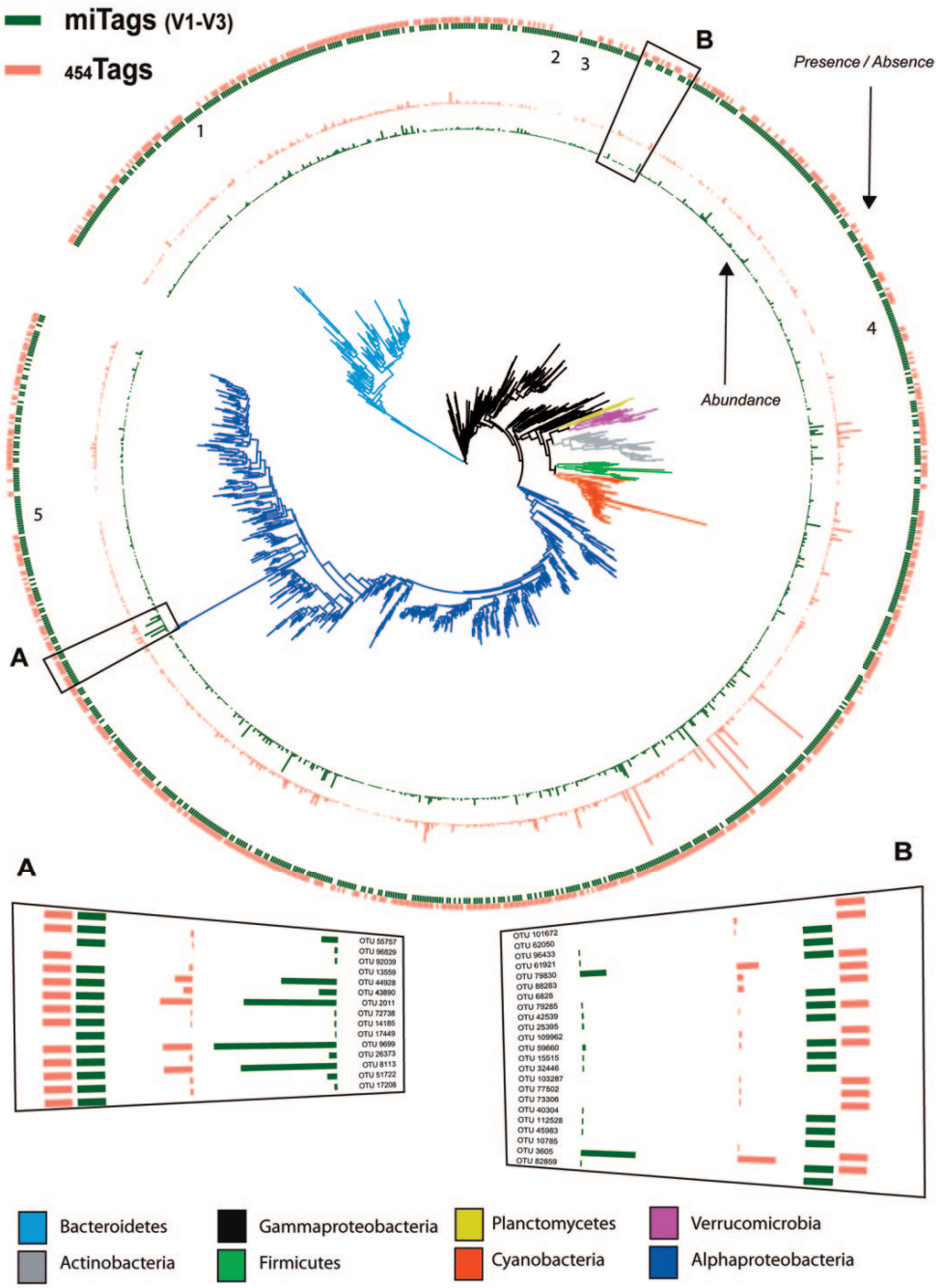

Figure2

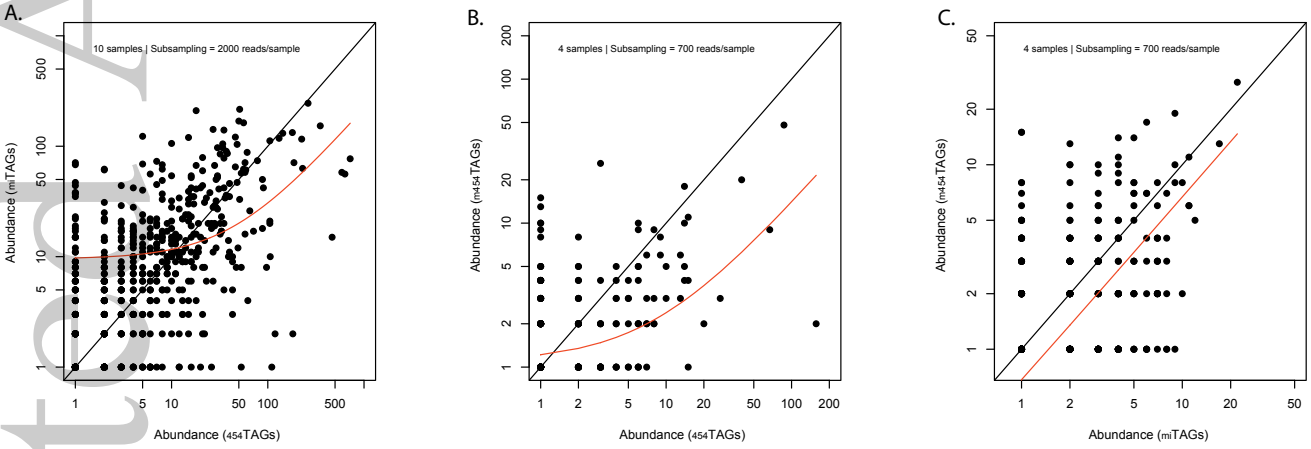

Figure3

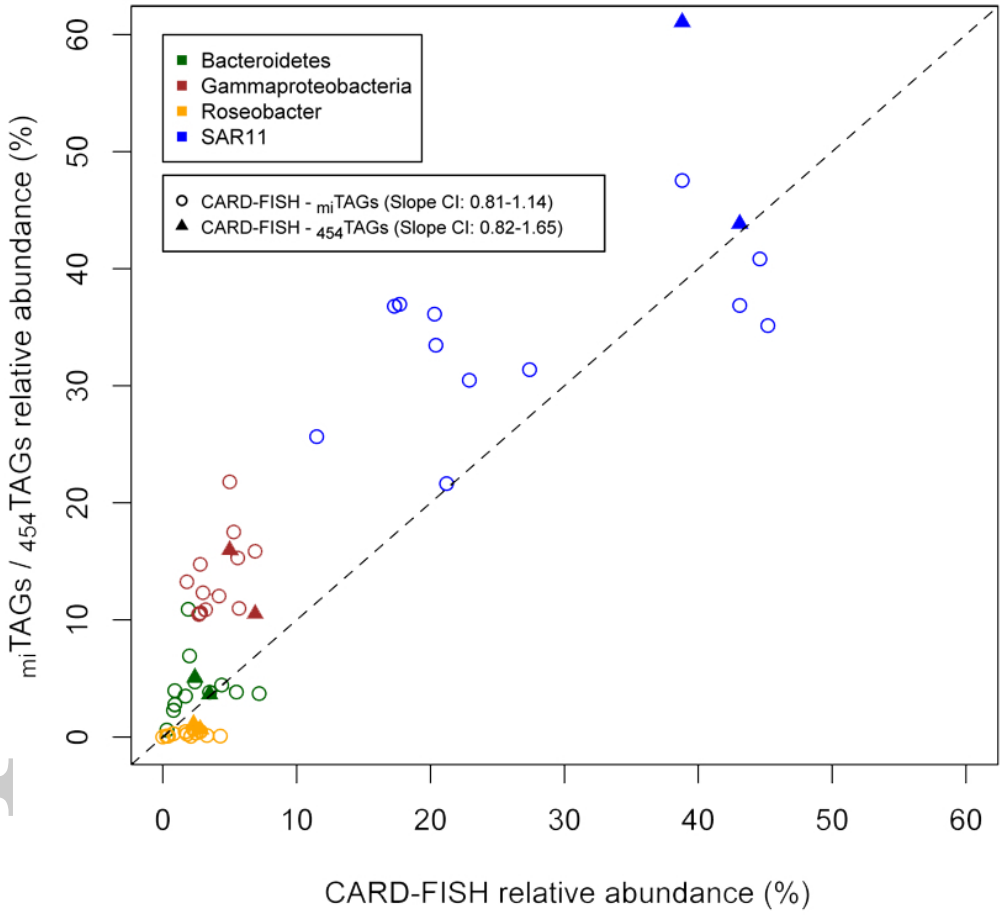

Figure4