August 2015

Probing the rare biosphere of the North-West Mediterranean Sea

Running title: Probing the rare biosphere

Bibiana G. Crespo[1*], Philip J. Wallhead[2], Ramiro Logares[1] and Carlos Pedrós-Alió[1]

[1] Institut de Ciències del Mar, Consejo Superior de Investigaciones Científicas (ICM-CSIC).
Passeig Marítim de la Barceloneta, 37-49. 08003. Barcelona, Spain.

[2] Norwegian Institute for Water Research (NIVA), Thormøhlens gate 53D, N-5006 Bergen, Norway.

[*]Present address: Uni Research Environment, Center for Applied Biotechnology, Thormøhlens gate 49B, N-5006 Bergen, Norway.

Correspondence: Bibiana G. Crespo, Uni Research Environment, Center for Applied Biotechnology, Thormøhlens gate 49B, N-5006 Bergen, Norway.

Email: bibianagc@hotmail.com

Conflict of interest statement: The authors declare no competing financial interest.

Funding: Sampling was supported by the Spanish MICINN grants CTM2005-04795/MAR and CTM2008-03309/MAR. B. G. C. was supported by a Juan de la Cierva contract from the Spanish “Ministerio de Ciencia e Innovación”. Research was funded by the Spanish Plan Nacional de Investigación Científica y Técnica grants Marine Gems (CTM2010-20361) and Blue Genes (CTM2013-48292-C3-1-R).

Subject Category: Microbial population and community ecology.

Abstract: The relatively recent development of high-throughput sequencing (HTS) techniques has revealed a wealth of novel sequences found in very low abundance: the rare biosphere. We performed a deep (1 million sequences per sample) pyrosequencing analysis of two marine bacterial samples and isolated a culture collection from one of them. Species data were derived from the sequencing analysis (97% similarity criterion) and various parametric distributions were fitted to the observed counts. Using the best-fitting Sichel distribution we estimate a total richness of 1 568–1 669 (95% CI) and 5027–5196 for surface and deep water samples respectively, implying that 84–89% of the total richness was sequenced. We also predict that a quadrupling of the present sequencing effort should suffice to observe 90% the total richness in both samples. Comparing with isolate sequences we found that isolation retrieved mainly extremely rare taxa which were not obtained by HTS despite the high sequencing effort. Culturing therefore remains a useful tool for mapping marine bacterial diversity, in addition to its other uses for studying the ecology of the rare biosphere.

Introduction

The question of how many species of living beings there are on Earth has intrigued ecologists and evolutionary scientists for decades (May, 1988; Erwin, 1991). One of the most recent estimates suggested around 8.7 million species, but this excluded bacteria and archaea due to our ignorance of these microorganisms (Mora *et al.*, 2011). The International Census of Marine Microbes set out to map the diversity of microbes in the oceans with novel high-throughput sequencing (HTS) techniques (Amaral-Zettler *et al.*, 2010) but a global estimate was not attempted. Some estimates for marine bacterial species range from 10^4 to 10^6 based on different assumptions (Curtis *et al.*, 2002; Hagström *et al.*, 2002). Such a range of values, spanning several orders of magnitude, shows that we are far from an accurate estimate.

Traditionally, bacteria were isolated in pure culture and then characterized biochemically and genetically until a new species could be formally described. It was realized that the bacteria able to grow in culture media were a small fraction of the bacterial cells that could be directly counted on a filter, a discrepancy named the “great plate count anomaly” (Staley & Konopka, 1985). Different studies estimated that only about 1% of the cells in natural waters could be cultivated (Pace, 1997; Eilers *et al.*, 2000). Moreover, most of the cells in pure cultures were not the abundant ones in nature.

After the application of molecular cloning to natural systems (Giovannoni *et al.*, 1990; Pace, 1997) a wealth of new taxa were found and, this time, they were the abundant ones in the oceans (DeLong, 1997; Pace, 1997). The drawback was that a sequence of the 16S rDNA did not provide much information about the physiology of the organism. Further, the realization that bacteria obtained in culture were mostly different from bacterial sequences obtained in clone libraries produced what could be called the “great clone library anomaly”. Molecular methods could retrieve many sequences from the abundant organisms but missed the rare ones, and only occasionally a rare clone was found. Isolation, on the other hand, retrieved mostly rare bacteria. This anomaly was a consequence of the fact that natural assemblages are formed by a few taxa in very large concentrations and many taxa in very low concentrations. The pattern can be easily visualized by looking at a rank-abundance curve (Pedrós-Alió, 2006). Primers for clone libraries will hybridize with the most abundant sequences over and over again before they find a rare target. Thus, only a fraction of the community will be available to cloning and sequencing. The relatively recent development of HTS techniques and their application to natural microbial communities (Sogin *et al.*, 2006) now provides an opportunity to solve the “great clone library anomaly”.

The study of microbial communities with such technologies has revealed a wealth of novel sequences found in very low abundance – a rare biosphere (Sogin *et al.*, 2006) – of which various properties have been examined (Galand *et al.*, 2009; Jones & Lennon, 2010; Pedrós-Alió, 2012; Caporaso *et al.*, 2012; Lynch *et al.*, 2012; Gibbons *et al.*, 2013). Today, studies of microbial diversity are performed almost exclusively with such HTS techniques, yet culturing still seems indispensable (Donachie *et al.*, 2007; Shade *et al.*, 2012; Lekunberri *et al.*, 2014), especially if the aim is to explore the rare biosphere. Shade *et al*. (2012) compared the outputs of a shallow (~ 2 000 sequences per sample) pyrosequencing analysis of the bacteria collected from a soil sample and the isolates cultured from the same sample. They found that 61% of the cultured bacteria were not present in the pyrosequencing dataset, demonstrating that culturing provided a fruitful route to the rare biosphere that was complementary to sequencing.

In this study, we performed a deep (1 million sequences per sample) pyrosequencing analysis of two marine bacterial samples and isolated a culture collection from one of them. Comparing these data sets allowed us assess whether or not current HTS technologies are sufficient to sequence all the taxa that are observed in culture. By fitting a parametric statistical model to the sequencing count data (observed abundances) we were also able to make well-constrained estimates of total species richness and to predict the sequencing effort necessary to observe 90% of the total richness in both surface and bottom samples.

Material and methods

1. Study area and sampling

Samples were taken at Station D, an open sea station at 40º52'N and 02º47'E (Table 1, and Pedrós-Alió *et al.*, 1999) in the NW Mediterranean Sea, during cruise SUMMER between

13th and 22nd of September 2011, on board the RV "García del Cid". The surface sample was taken at 5 m on 15th September and the bottom sample was collected at 2 000 m depth on 17th September.

Sampling was done with Niskin bottles mounted on a rosette with a conductivity-temperature-depth (CTD) profiler. Water was prefiltered through a 200 µm mesh and processed on board. To collect microbial biomass, 5–15 L of sea-water were prefiltered through a 3 µm pore size Durapore filter (Millipore, Cork, Ireland) and free-living bacterial biomass was collected on a 0.22 µm pore size Sterivex filter (Durapore, Millipore). The filtration was done in succession using a peristaltic pump. The 0.22 µm pore size Sterivex unit was filled with 1.8 ml of lysis buffer (40 mM EDTA, 50 mM Tris-HCl, 0.75 M sucrose) and stored at –80ºC. DNA was extracted by a standard protocol using phenol/chloroform (details in Schauer *et al*., 2003). The sequencing was done with the same amount of DNA for every sample.

2. 454-pyrosequencing and noise removal

Purified DNA samples were submitted to the Research and Testing Laboratory (Lubbock, Texas, USA) and prokaryotic diversity was assessed by tag-pyrosequencing of the V1-V3 regions (~400 bp) with the Roche 454 Titanium platform using primers 28F/519R (details in *SI*). 713 076 and 970 346 tags were retrieved from the surface and the bottom samples, respectively (Table 1). These data have been deposited in EMBL with accession number PRJEB9061.

Sequence data were processed, including end-trimming, quality control and denoising, using QIIME (Caporaso *et al*., 2010). To identify potential chimera sequences, the dataset was subjected to the ChimeraSlayer implemented in Mothur (Schloss *et al*., 2011). The final number of tags was thus reduced to 500 262 and 574 960 for surface and bottom samples respectively (Table 1). The sequences were then clustered into Operational Taxonomic Units (OTUs) based on the relatedness of the sequences (97% similarity); the taxonomy assignment of consensus sequences was done using the SILVA v108 database (Quast *et al.*, 2013) (see details in *SI*).

3. Isolation of bacterial cultures

Isolates were obtained on board and incubated back in the laboratory where 326 bacterial colonies were selected, purified and stored (see details in *SI*). 200 µl of these cultures were placed in 96 well plates, diluted 1:4 and heated (95 ºC, 10 min) to cause cell lysis, so available DNA could be used as a template in Polimerase Chain Reactions (PCR). PCR of the Internal Transcribed Spacer (ITS) (see *SI*) was done to select as many different species as possible from the 326 isolates. ITS length is species specific and therefore allows to differentiate the isolates (Fisher & Triplett 1999; Scheinert *et al.* 1996). According to their different ITS patterns, 148 isolates were chosen out of 326, including some replicates, and their 16S rRNA genes were amplified (see *SI*). Nearly the full-length 16S rRNA gene (approx. 1 300 bp) was sequenced in GENOSCREEN (Lille Cedex, France). Taxonomical assignment was done by BLAST searches in the National Center for Biotechnology Information (NCBI) website. The 16S rRNA sequences have been deposited in EMBL with accession numbers LN845965 to LN846112.

4. Richness, sequencing effort estimates, and diversity of 454 pyrosequecing data

Observed species richness (S_{obs}) was computed as the total number of sequenced OTUs (97% similarity) in each DNA sample. Total species richness (S), defined as the total number of species represented in the water sample, was estimated by fitting a parametric distribution to the count data following the Bayesian Markov-Chain Monte Carlo (MCMC) method of Quince *et al*. (2008). We fitted four distributions: the Poisson log-normal, the Poisson log-Student, the Poisson inverse Gaussian, and the Poisson generalized inverse Gaussian (Sichel distribution). The best-approximating distribution for each sample was chosen using the Deviance Information Criterion (DIC; Spiegelhalter *et al*., 2002), which for our fits was almost identical to Akaike's Information Criterion (AICc; Burnham & Anderson, 2002; see Table S1). S was then estimated as the posterior mean value of the corresponding Bayesian parameter under the selected model, and 95% credible intervals (CIs, Bayesian equivalent of confidence intervals) were taken from the 2.5% and 97.5% quantiles of the posterior distribution. Note that by this method the total richness S is included in the likelihood function and estimated jointly with the two or three parameters describing the taxon abundance distribution, thus facilitating uncertainty calculations (Izsak, 2008; Connolly & Thibaut, 2012). Also, the Bayesian MCMC approach appears to mitigate the problem of trapping in local maxima which can severely compromise the calculation of maximum likelihood estimates (Connolly & Dornelas, 2011).

We also predicted the required sequencing effort (RSE) to observe 90% of the total water sample richness in a hypothetical repeat DNA sample. Higher percentages were not considered because due to uncertainties in the estimates they could not be meaningfully constrained. RSE was predicted by simulating an ensemble of 80 repeat sequences using the selected model and sampling from the posterior parameter distribution, then taking the ensemble mean RSE and (2.5%, 97.5%) percentiles as point predictions and 95% prediction intervals (PIs) respectively (see *SI* for details). Model selection uncertainty (Burnham & Anderson, 2002) was not accounted for in the PIs for RSE nor in the CIs for S; however, the only model with comparable DIC to the best-approximating model (to within 12 units of DIC or AICc) was merely a special case of the best-fitting model (Poisson inverse Gaussian vs. Sichel distribution, see Table S1) so the neglected uncertainty was likely small.

These simulations and others using the non-selected distributions were also used to test the performance of various simpler and faster methods to predict S and RSE, including several nonparametric methods (Chao, 1984; Chao & Lee, 1992; Chao *et al.*, 2000; Shen *et al.*, 2003; Chao *et al.*, 2009; Wang, 2011; Chao & Shen, 2012; Colwell *et al.*, 2012; Chao *et al.*, 2014; Chiu *et al.*, 2014) and a semiparametric method whereby multiple saturating functions are fitted to the collector's curve and the lowest-AICc function is used for prediction (cf. Flather 1996; Guilhaumon *et al*. 2008; Table S2). Unfortunately, none of these faster methods showed robust performance over all simulations (Table S3; O'Hara, 2005; Quince *et al*. 2008). Herein, we report only the Chao1 estimator for S (Chao, 1984) because it is widely quoted and thus useful for comparison with other studies.

To measure diversity we use the Shannon diversity ($H' = -\sum p_i \ln p_i$) and the Simpson diversity index ($D = 1 - \sum p_i^2$ where $p_{i=}$ N$_i$/N, the number of individuals of species *i* divided by the total number of individuals in the sequencing sample N). Evenness was computed with the Pielou index ($J' = H'/H_{\max}$ where H' is the Shannon diversity index and $H_{\max}$ is the maximal possible Shannon diversity index if all the species were equally abundant: $H_{\max} = -\sum S_{obs}^{-1} \ln S_{obs}^{-1} = \ln S_{obs}$). Diversity and evenness were calculated using the "vegan" package (Oksanen *et al.*, 2013) of R (R Core Team, 2013). Rank-abundance plots of the isolated cultures and the 454 pyrosequencing data were done using "BiodiversityR" (Kindt & Coe, 2005) and collector's curves with confidence intervals were computed using "iNEXT" (Chao *et al.*, 2014; Hsieh *et al.*, 2015).

5. Comparison of 454-pyrosequencing tags and isolates

Comparison between isolates and 454 tag-sequences was done by running BLASTn locally. The isolate sequences were searched for in the 454 tag-sequence datasets and vice versa, and only the reciprocal matches between these two searches were considered. The output was filtered using R (R Core Team, 2013), requiring 99% of identical nucleotide matches, ≥75% coverage of the isolate sequence, and a bit-score higher than 100. In all the cases the e-value was lower than 0.0001.

Since the primers used for Sanger sequencing of the isolates and those used for the pyrosequencing of the environmental DNA were different, the possibility existed of different biases that could prevent detection of the cultures in the 454 dataset. Multiple alignments of the sequences of the isolates and the sequences of the primers used in the pyrosequencing analysis were done using the software Geneious pro 3.5.4 (Kearse *et al.*, 2012), and allowed us to confirm that the 454 primers hybridized with the sequences of all the isolates.

Results

1. Pyrosequencing dataset

Observed richness (S_{obs}) was much higher in the bottom (4 460) than in the surface (1 400) sample (Table 1). In both samples only ~17% of the OTUs were singletons (an OTU represented by a single sequence) (Table 1). Evenness (J') and diversity (H' and D) were both higher in the bottom than in the surface sample (Table 1). Collector´s curves suggested that the bottom sample would be richer for a broad range of lesser, equal sampling efforts and that S_{obs} was approaching asymptotic values for both samples (Figure 1).

Among the four candidate parametric distributions fitted to the count data, the Sichel distribution was the best approximating model (lowest DIC and AICc) for both samples (Table S1). The goodness-of-fit of this distribution is illustrated in Supplementary Figure 1. The fitted frequencies at moderately low counts may suggest some room for improvement, but overall for the counts in the range 1–100 shown in Supplementary Figure 1 it appears that the model gives an adequate fit. Using the Sichel distribution, the total water sample richness was estimated as 1 568–1 669 (95% CI) and 5 027–5 196 for surface and deep samples respectively, suggesting that 84–89% and 86–89% of the total richness was observed by sequencing. By simulating from this distribution we predict that 0.6–4.3 (95% PI) and 1.0–3.2 times the present sequencing effort would suffice to observe 90% of the total richness in the surface and bottom water samples respectively (Table 1).

Rank-abundance curves (Figure 2) showed that the bacterial assemblages from both samples were characterized by few abundant and many rare OTUs. The most abundant OTU was more abundant in the surface than in the bottom sample, in agreement with the lower evenness found for the surface sample (Table 1). The abundance of the most abundant OTU in the bottom sample was close to the abundance of the second most abundant OTU in the surface sample.

2. Culture collection

Bacterial isolation from the sample collected at the surface retrieved 148 cultures belonging to 38 different species. The most frequent bacterium in the collection was

Erythrobacter citreus, isolated 37 times, while 17 species were isolated only once. A rank abundance plot of the 38 species is shown in Figure 3. The isolates belonged to the phyla Actinobacteria (4 isolates), Bacteroidetes (4 isolates) and Firmicutes (2 isolates) and to the Proteobacteria classes Alpha-proteobacteria (18 isolates) and Gamma-proteobacteria (10 isolates). The names of all the isolates are shown in Table 2 and Table 3.

3. Comparison of isolates and sequences

Only 14 (37%) of the 38 different isolated species were found in the 454 tag-sequence datasets: one Actinobacteria, two Bacteroidetes, two Firmicutes, four Alpha-proteobacteria and five Gamma-proteobacteria isolates (Figure 3, Table 2). Surprisingly, the number of cultures found in the 454 tag-sequence dataset was higher in the sample collected at 2 000 m (37%) than in the surface sample (24%), even though the latter was the sample used for isolation of the bacterial cultures (Figure 3, Table 2). Nine isolated species were found in the sequences from both samples (maroon in Figure 3), five were found in the bottom sample only (green in Figure 3) and 24 were not found in either sample (empty symbols in Figure 3). Practically all of the 454 tag-sequences that matched the sequences from the isolates belonged to rare OTUs (<1% of the total tags). Only the OTU matching the isolate *Alteromonas macleodii* str. ‘Balearic Sea AD45’ (Gamma-Proteobacteria) was somewhat abundant (1.3%) in the bottom sample (Table 2). Further, all the matching sequences made a larger percentage of the assemblage in the bottom sample than in the surface sample.

Discussion

1. Estimates of richness

In a previous study (Pommier *et al.*, 2010) we used pyrosequencing of the V6 region of the 16S rDNA gene to estimate richness of the bacterial assemblages in the NW Mediterranean Sea, at the same location and month as the present study but during a different year. Around 20 000 final tags were obtained per sample and we observed 632 and 2 065 OTUs in surface and deep samples respectively. It is well known that the number of new taxa retrieved increases with sample size and sampling effort (Preston, 1960; Magurran, 1988; Rosenzweig, 1995) and that a large part of the diversity may remain hidden due to sampling limitations (Gotelli & Colwell 2011), especially in microbial ecology (Øvreås & Curtis, 2011). In the present study, we took advantage of pyrosequencing capabilities to increase the sequencing depth (to around 500 000 final tags per sample) in an attempt to achieve realistic estimates of the whole bacterial diversity.

The resulting collector's curves appear to be approaching asymptotic values (Figure 1) and the reduced percentage of singletons (~17% vs. 40%–60% in Pommier *et al.*, 2010) suggests an improved coverage of the bacterial community. However, the order of magnitude of the Chao1 estimates of total richness are consistent with the earlier study (1 646 and 5 031 here vs. 1 289 and 4 156 in Pommier *et al.*, 2010), and our present Chao1 estimates agree with the 95% CIs from the best-approximating Sichel distribution (see Table 1). Also, the narrowness of the confidence intervals for expected richness in Figure 1, relative to the difference in surface vs. bottom values, suggests that the higher richness of the bottom sample could have been established with a much lower sequencing effort. The apparent availability of such basic results at lower effort is clearly good news for further routine and comparative studies.

In the surface sample, the most abundant OTU contributed a very large fraction of the total tags (36%), raising concerns that this may have caused less OTUs to be uncovered and forced the richness to appear lower. However, if this species is excluded from the analysis, the main effect on the collector's curves is to decrease the total number of tags for the surface sample by 36%, and the bottom sample is still clearly richer at this lower level of effort (Figure 1). We also reran the Sichel fit to the surface data with this OTU excluded and obtained a negligible change (1 species) in the estimated total richness (Table S1). The numbers of OTUs observed in both samples in this study are consistent with numbers estimated by other authors for the upper ocean (Rusch *et al.*, 2007; Quince *et al.*, 2008; Pommier *et al.*, 2010; Crespo *et al.*, 2013) and the deep ocean (Salazar *et al.*, 2015). Pommier *et al.* (2010) and Crespo *et al.* (2013) also found higher richness in the bottom than in the surface waters.

A study of a marine bacterial sample collected in the English Channel (Caporaso et al., 2012) is particularly relevant for our analysis. Station L4 was very deeply sequenced (10 million sequences) by Illumina and revealed ~100 000 OTUs, two orders of magnitude higher than our estimates of total richness for the Mediterranean samples. To explain this huge difference we see three possible reasons.

First, the English Channel may in reality have more species than the Mediterranean Sea. However, we consider it unlikely that the real difference is two orders of magnitude given that both environments correspond to relatively open seawater, albeit in different hydrographic and nutrient regimes.

Second, there could be a statistical issue that caused an underestimation of diversity based on the smaller number of tags in the present study. For example, when diversity of marine bacterial communities was estimated from conventional clone libraries (with a few hundred clones) the Chao1 total richness estimates were on the order of a few hundred OTUs, but when similar samples were analyzed by HTS (with tens of thousands of sequences per sample) the Chao1 estimators gave several thousands of OTUs. However, the Chao1 estimator is known to underestimate diversity in strongly heterogeneous and undersampled communities, of which microbial communities are a prime example (Quince *et al.*, 2008; Øvreås & Curtis, 2011; Chiu *et al.*, 2014). The Sichel distribution used for our estimates was selected by statistical criteria (Table S1) and gave an apparently good fit to the observed count frequencies (Supplementary Figure 1). Under this distribution, the total sample richness was well constrained to within 3–6% at 95% credibility (Table 1). Of course, some other distribution not considered here may fit the data better and predict a higher richness, but we have no reason to expect orders of magnitude revisions. Indeed, none of the other three candidate distributions produced upper CI limits of total richness more than 1 000–2 000 (40–70%) higher than the Sichel upper limits (Table S1).

A third possible reason is that the procedure chosen for the L4 OTU identification overestimated the number of OTUs. Caporaso et al. (2012) found that 45-48 % of their OTUs were singletons, which is a surprisingly large fraction for data sets consisting of over 10^7 tags. Increased sequencing depth is generally expected to reduce the fraction of singletons (Wall *et al.*, 2009; Penton *et al.*, 2013) and this appears to be the case for our Mediterranean samples (~17% here vs. 40%–60% in Pommier *et al.*, 2010). We believe that the current processing of pyrosequencing data is quite robust (Quince *et al.*, 2011) but processing of Illumina tags was still in its infancy when the L4 study was carried out, suggesting that diversity might have been overestimated due to misidentified OTUs, probably due to bias from short reads (Claesson *et al.*, 2010). This is actually what happened in the first application of pyrosequencing to marine bacterial diversity in Sogin *et al.* (2006). Later studies found ways to properly clean the sequences and estimates became lower (Huse *et al.*, 2010; Quince *et al.*, 2011).

2. Comparison of sequencing and isolation

The current power of massive parallel sequencing allows us to probe the rare biosphere (Caporaso et al., 2012; Pedrós-Alió, 2012; Gibbons *et al*., 2013), but culturing is an alternative avenue to explore it (Pedrós-Alió, 2006; Shade *et al*., 2012). Comparing both approaches we have found that isolation retrieves some of the rarest taxa since only 24–37% of the isolates were found in the 454-pyrosequencing data (Figure 2).

The observed and estimated total richnesses can be used to estimate the probability that a species *chosen at random* from the total list of species is retrieved by the present sequencing effort. This probability is $S_{obs}/S \approx 0.87$ for the surface sample, so if the 38 cultured species could be considered randomly chosen, we would expect to retrieve 33 of them by sequencing. Given this probability, the fact that we retrieved only 9 (24%) is highly significant ($P < 10^{-17}$ from binomial test; $P < 1/3001$ from simulation test, see *SI*). A similar argument applies to the bottom sample if we assume that all the cultured species (derived from the surface water sample) are also present in the bottom water sample. The cultured species are apparently less represented in the sequencing data sets than would be random selections from the lists of all species in the water samples.

Again we see three possible reasons for this discrepancy. First, there may have been a bias in the PCR and DNA amplification of the sequencing techniques (Berry *et al*., 2011; Pinto & Raskin, 2012). However, when tested *in silico*, the primers used for pyrosequencing covered the whole diversity captured by the primers used for Sanger sequencing of the isolates, and a bias affecting the diversity found using both methods seems unlikely.

Second, since the cultures were isolated from the whole water sample while the pyrosequencing data were obtained from the 0.2–3 μm fraction, some species attached to larger particles may have been excluded from the sequencing datasets. However, 18 of the 38 cultured species are expected to be free-living bacteria since they belong to the Alpha-proteobacteria class (DeLong *et al.*, 1993; Crespo *et al.*, 2013) and should therefore be present in the 0.2–3 µm fraction used for sequencing. If the comparison is restricted to this class we find that only 4 out of 18 isolates are retrieved in both surface and bottom sequences, which is still a highly significant deficit ($P < 10^{-9}$, binomial test; $P < 1/3001$, simulation test).

A third possible reason is that the special environment imposed by culturing may favour certain species that are generally less successful in natural oceanic conditions, and consequently too rare to retrieve with the present sequencing effort. The process of culturing might in this sense "select for the losers" in the natural environment. However, if this were a consistent effect, we would expect the few isolates that are retrieved by sequencing to have anomalously low tag counts, but this not in fact observed (Table 2). The surface counts, while low/rare in an absolute sense (<0.1% of total tags), are not low relative to a random sample from the observed or modelled count distributions ($P > 0.05$ from bootstrap and simulation tests on mean, median and maximum counts, see *SI*), and the bottom counts are if anything slightly higher than a random sample ($P < 0.05$ for mean, median and maximum counts). The culturing process might therefore have selected for a few moderately-rare species that grow better in deep water (Table 2), probably because the culturing was done in the dark, plus a larger number of extremely rare species that could not be retrieved with the present sequencing effort (Table 3).

Our results suggest that, with present HTS capacity, culturing remains an important complementary tool for mapping microbial diversity. Future improvements in sequencing depth will eventually uncover the isolated bacteria, though perhaps only slowly. However, even if the whole bacterial diversity were mapped by sequencing, culturing would remain essential for the study of marine bacterial communities, especially if the target is the rare biosphere (Donachie *et al*., 2007). Culturing provides complete genomes and allows the study of the physiology, metabolism and ecology of marine bacteria, yielding information that cannot be obtained by sequencing alone (Giovannoni & Stingl, 2007).

In conclusion, using deep sequencing (10^6 tags) we have been able to obtain robust estimates of the total richness of the bacterial assemblages in two samples from the surface and deep Mediterranean Sea. These estimates were on the order of 2 000 to 5 000 taxa, and current sequencing capacity appears to be in reach of observing 90% of the total diversity. Comparison with cultured isolates showed that many of the isolated species were from deep within the rare biosphere and not retrieved by sequencing, thus confirming that sequencing and culturing remain complementary strategies for probing the rare marine biosphere.

Acknowledgements

We thank the crews and scientists in cruises Modivus and SUMMER, both on the RV García del Cid, supported by the Spanish MICINN grants CTM2005-04795/MAR and CTM2008-03309/MAR respectively.

We thank F. M. Cornejo-Castillo for his advice on the method for differentiation of isolates. Help from E. Sa and V. Balagué with bacterial culturing and PCR work is greatly appreciated.

B. G. C. was supported by a Juan de la Cierva contract from the Spanish “Ministerio de Ciencia e Innovación”. Research was funded by the Spanish “Plan Nacional de Investigación Científica y Técnica” grants Marine Gems (CTM2010-20361) and Blue Genes (CTM2013-48292-C3-1-R).

Conflict of interest statement: The authors declare no competing financial interest.

Supplementary information (*SI*) is available at ISME Journal's website.

References

Alonso-Sáez L, Balagué V, Sà EL, Sánchez O, González JM, Pinhassi J, *et al.* (2007). Seasonality in bacterial diversity in north-west Mediterranean coastal waters: assessment through clone libraries, fingerprinting and FISH. *FEMS Microbiol Ecol* **60**:98–112.

Alonso-Sáez L, Sánchez O, Gasol JM, Balagué V, Pedrós-Alio C. (2008). Winter-to-summer changes in the composition and single-cell activity of near-surface Arctic prokaryotes. *Environ Microbiol* **10**:2444–2454.

Amaral-Zettler L, Artigas LF, Baross J, Bharathi LPA, Boetius A, Chandramohan D, *et al.* (2010). A Global Census of Marine Microbes — Census of Marine Life Maps and Visualization. In: McIntyre., A (ed). Life in the World's Oceans: Diversity, Distribution, and Abundance. Wiley-Blackwell, pp 223–345.

Berry D, Ben Mahfoudh K, Wagner M, Loy A. (2011). Barcoded primers used in multiplex amplicon pyrosequencing bias amplification. *Appl Environ Microbiol* **77**:7846–7849.

Burnham KP, Anderson DR. (2002). Model selection and multimodel inference: A practical information-theoretic approach.2nd edition. Springer-Verlag: New York.

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman F, Costello E, *et al.* (2010). QIIME allows analysis of high- throughput community sequencing data. *Nature* **7**:335–336.

Caporaso JG, Paszkiewicz K, Field D, Knight R, Gilbert JA. (2012). The Western English Channel contains a persistent microbial seed bank. *ISME J* **6**:1089–1093.

Chao A. (1984). Nonparametric estimation of the number of classes in a population. *Scand J Stat* **11**:265–270.

Chao A, Colwell RK, Lin C, Gotelli NJ. (2009). Sufficient Sampling for Asymptotic Minimum Species Richness Estimators. *Ecology* **90**:1125–1133.

Chao A, Gotelli N, Hsieh T, Sander E, Ma K, Colwell R, *et al.* (2014). Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species diversity studies. *Ecol Monogr* **84**:45–67.

Chao A, Hwang W, Chen Y, Kuo C. (2000). Estimating the number of shared species. *Statistica sinica* **10**:227–246.

Chao A, Lee S-M. (1992). Estimating the number of classes via sample coverage. *J Am Stat Assoc* **87**:210–217.

Chao A, Shen T. (2012). Program SPADE (Species Prediction And Diversity Estimation). Program and use's guide published at http://chao.stat.nthu.edu.tw.

Chiu C-H, Wang Y-T, Walther BA, Chao A. (2014). An improved nonparametric lower bound of species richness via a modified good-turing frequency formula. *Biometrics* **70**:671–682.

Claesson MJ, Wang Q, O'Sullivan O, Greene-Diniz R, Cole JR, Ross RP, *et al.* (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.

Colwell RK, Chao A, Gotelli NJ, Lin S-Y, Mao CX, Chazdon RL, *et al.* (2012). Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. *J Plant Ecol* **5**:3–21.

Connolly SR, Dornelas M. (2011). Fitting and empirical evaluation of models for species abundance distributions. In: Magurran, A & McGill, B (eds). Biological diversity: Frontiers in measurement and assessment. Oxford University Press: Oxford, UK, pp 123-141.

Connolly SR, Thibaut LM. (2012). A comparative analysis of alternative approaches to fitting species-abundance models. *J Plant Ecol* **5**: 32-45.

Crespo BG, Pommier T, Fernández-Gómez B, 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* **2**:541–552.

Curtis TP, Sloan WT, Scannell JW. (2002). Estimating prokaryotic diversity and its limits. *Proc Natl Acad Sci* **99**:10494–10499.

DeLong E. (1997). Marine microbial diversity: the tip of the iceberg. *Trends Biotechnol* **15**:203–207.

DeLong E, Franks D, Alldredge A. (1993). Phylogenetic diversity of aggregate-attached vs. free-living marine bacterial assemblages. *Limnol Oceanogr* **38**:924–934.

Donachie SP, Foster JS, Brown MV. (2007). Culture clash: challenging the dogma of microbial diversity. *ISME J* **1**:97–99.

Eilers H, Pernthaler J, Glöckner FO, Amann R. (2000). Culturability and *In situ* abundance of pelagic bacteria from the North Sea. *Appl Environ Microbiol* **66**:3044–3051.

Erwin T. (1991). How many species are there? Revisited. *Conserv Biol* **5**:1–4.

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

Flather CH. (1996). Fitting species-accumulation functions and assessing regional land use impacts on avian diversity. *J Biogeogr* **23**:155–168.

Galand PE, Casamayor EO, Kirchman DL, Lovejoy C. (2009). Ecology of the rare microbial biosphere of the Arctic Ocean. *Proc Natl Acad Sci* **106**:22427–22432.

Gibbons SM, Caporaso JG, Pirrung M, Field D, Knight R, Gilbert JA. (2013). Evidence for a persistent microbial seed bank throughout the global ocean. *Proc Natl Acad Sci* **110**:4651–4655.

Gilbert JA, Steele JA, Caporaso JG, Steinbrück L, Reeder J, Temperton B, *et al.* (2012). Defining seasonal marine microbial community dynamics. *ISME J* **6**:298–308.

Giovannoni S, Stingl U. (2007). The importance of culturing bacterioplankton in the " omics " age. *Nat Rev Microbiol* **2007**:820–826.

Giovannoni SJ, Britschgi TB, Moyer CL, Field KG. (1990). Genetic diversity in Sargasso Sea bacterioplankton. *Nature* **345**:60–63.

Gotelli NJ, Colwell RK. (2011). Estimating species richness. In: Magurran, A & McGill, B (eds). Biological diversity: Frontiers in measurement and assessment. Oxford University Press: Oxford, UK, pp 39-54.

Guilhaumon F, Gimenez O, Gaston KJ, Mouillot D. (2008). Taxonomic and regional uncertainty in species-area relationships and the identification of richness hotspots. *Proc Natl Acad Sci U S A* **105**:15458–15463.

Hagström Å, Pommier T, Rohwer F, Simu K, Svensson D, Zweifel U. (2002). Bio-informatics reveal surprisingly low species richness in marine bacterioplankton. *Appl Environ Microbiol* **67**:3628–3633.

Hsieh T, Ma K, Chao A. (2015). iNEXT: An R Package for interpolation and extrapolation of species diversity (Hill numbers). Submitted manuscript. http://chao.stat.nthu.edu.tw/blog/software-download/.

Huse SM, Welch DM, Morrison HG, Sogin ML. (2010). Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environ Microbiol* **12**:1889–1898.

Izsak R. (2008). Maximum likelihood fitting of the Poison lognormal distribution. *Environ Ecol Stat* **15**:143-156.

Jones SE, Lennon JT. (2010). Dormancy contributes to the maintenance of microbial diversity. *Proc Natl Acad Sci U S A* **107**:5881–5886.

Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, *et al.* (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**:1647–1649.

Kindt R, Coe R. (2005). Tree diversity analysis. A manual and software for common statistical methods for ecological and biodiversity studies. World Agroforestry Centre (ICRAF): Nairobi (Kenya).

Lekunberri I, Gasol JM, Acinas SG, Gómez-Consarnau L, Crespo BG, Casamayor EO, *et al.* (2014). The phylogenetic and ecological context of cultured and whole genome-sequenced planktonic bacteria from the coastal NW Mediterranean Sea. *Syst Appl Microbiol* **37**:216-228.

Lynch MDJ, Bartram AK, Neufeld JD. (2012). Targeted recovery of novel phylogenetic diversity from next-generation sequence data. *ISME J* **6**:2067–2077.

Magurran AE. (1988). Ecological diversity and its measurements. Princeton University Press: Princeton, New Jersey.

May RM. (1988). How many species are there on Earth? *Science* **241**:1441–1449.

Mora C, Tittensor DP, Adl S, Simpson AGB, Worm B. (2011). How many species are there on Earth and in the ocean? *PLoS Biol* **9**:e1001127.

O'Hara RB. (2005). Species richness estimators: how many species can dance on the head of a pin? *J Anim Ecol* **74**:375–386.

Oksanen J, Guillaume-Blanchet F, Kindt R, Legendre P, Minchin P, O'Hara R, *et al.* (2013). Vegan: Community Ecology Package.

Øvreås L, Curtis TP. (2011). Microbial diversity and ecology. In: Magurran, A & McGill, B (eds). Biological diversity: Frontiers in measurement and assessment. Oxford University Press: Oxford, UK, pp. 221–236.

Pace NR. (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. *Ann Rev Mar Sci* **4**:449–466.

Pedrós-Alió C, Calderón-Paz J-I, Guixa-Boixereu N, Estrada M, Gasol JM. (1999). Bacterioplankton and phytoplankton biomass and production during summer stratification in the northwestern Mediterranean Sea. *Deep Sea Res Part I Oceanogr Res Pap* **46**:985–1019.

Penton CR, St Louis D, Cole JR, Luo Y, Wu L, Schuur EAG, *et al.* (2013). Fungal diversity in permafrost and tallgrass prairie soils under experimental warming conditions. *Appl Environ Microbiol* **79**:7063–7072.

Pinto AJ, Raskin L. (2012). PCR biases distort bacterial and archaeal community structure in pyrosequencing datasets. *PLoS One* **7**:e43093.

Pommier T, Neal P, Gasol J, Coll M, Acinas S, Pedrós-Alió 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.

Preston FW. (1960). Time and Space and the Variation of Species. *Ecology* **41**:612–627.

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**:590–596.

Quince C, Curtis TP, Sloan WT. (2008). The rational exploration of microbial diversity. *ISME J* **2**:997–1006.

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

R Core Team. (2013). R: A language and environment for statistical computing.

Rosenzweig M. (1995). Species diversity in space and time. Cambridge University Press: Cambridge.

Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S, *et al.* (2007). The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol* **5**:e77.

Salazar G, Cornejo-Castillo, FM Benítez-Barrios V, Fraile-Nuez E, Álvarez-Salgado, XA Duarte C, Gasol J, Acinas S. (2015). Global diversity and biogeography of deep-sea pelagic prokaryotes. *ISME J*. (In press)

Schauer M, Balagué V, Pedrós-Alió C, Massana R. (2003). Seasonal changes in the taxonomic composition of bacterioplankton in a coastal oligotrophic system. *Aquat Microb Ecol* **31**:163–174.

Scheinert P, Krausse R, Ullmann U, Söller R, Krupp G. (1996). Molecular differentiation of bacteria by PCR amplification of the 16S–23S rRNA spacer. *J Microbiol Methods* **26**:103–117.

Schloss PD, Gevers D, Westcott SL. (2011). Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. *PLoS One* **6**:e27310.

Shade A, Hogan CS, Klimowicz AK, Linske M, McManus PS, Handelsman J. (2012). Culturing captures members of the soil rare biosphere. *Environ Microbiol* **14**:2247–2252.

Shen T-J, Chao A, Lin C-F. (2003). Predicting the number of new species in further taxonomic sampling. *Ecology* **84**:798–804.

Sogin ML, Morrison HG, Huber JA, Mark Welch D, Huse SM, Neal PR, *et al.* (2006). Microbial diversity in the deep sea and the underexplored “rare biosphere”. *Proc Natl Acad Sci U S A* **103**:12115–12120.

Spiegelhalter D, Best N, Carlin B, van der Linde A. (2002). Bayesian measures of model complexity and fit (with discussion). *J R Stat Sic Ser B* **64**:583-639.

Staley J, Konopka A. (1985). Measurement of in situ activities of nonphotosynthetic microroganisms in aquatic and terrestrial habitats. *Annu Rev Microbiol* **39**:321–383.

Wall PK, Leebens-Mack J, Chanderbali AS, Barakat A, Wolcott E, Liang H, *et al.* (2009). Comparison of next generation sequencing technologies for transcriptome characterization. *BMC Genomics* **10**:347.

Wang J. (2011). SPECIES : An R Package for Species Richness.

Figure legends

Figure 1. OTU collector's curves of the surface (orange line) and bottom (green line) samples. Black dashed lines indicate the 95% confidence intervals (95% CI).

Figure 2. Rank-abundance plots of surface (A) and bottom (B) samples. The red line is the rank-abundance plot calculated with the actual data. The dark blue line shows the estimates of the sequencing effort necessary to retrieve 90% of the total richness calculated by simulation from the best-approximating Sichel distribution (posterior mean estimate). The vertical black line separates the real data (left) from the estimates (right). The percentage of cultured isolates found in the 454-pyrosequencing datasets is indicated at the left side of the black vertical line. The percentage of cultured isolates not found in the 454-pyrosequencing datasets, and that would presumably be found by increasing the sequencing effort, is indicated at the right of the black vertical line. Insert pictures show some of the bacterial cultures grown from the surface sample. Font size and pictures are scaled according to the percentage of cultured isolates found or not found in the 454-pyrosequencing datasets.

Figure 3. Rank-abundance plot of the 38 isolated bacterial species. The maroon squares indicate the cultured isolates found in both the surface and bottom 454-pyrosequencing datasets, the green triangles indicate the cultures isolated found only in the bottom 454-pyrosequencing dataset, and the white circles indicate the cultures that were not found in any of the 454-pyrosequencing datasets. A list of the isolated bacterial species can be found in Table 2 and Table 3.

Table captions

Table 1. Summary of location and depth (m) of samples, total sequences before (Raw Tags) and after (Final Tags) cleaning, richness (S) computed as total Operational Taxonomic Units (OTUs) clustered at 97% identity, percentage of singletons. Diversity was estimated using the Shannon diversity index (H'), Simpson diversity (D) and Pielou's evenness (J). Total richness (S) was estimating using the Chao1 lower bound estimator (Chao, 1984) and the Sichel distribution, fitted to the count frequency data by the Bayesian method of Quince *et al*. (2008) and selected from four alternative candidate models using the Deviance Information Criterion. Using the Sichel distribution, point estimates and 95% credible intervals (CIs) for S were obtained from the mean and (2.5%, 97.5%) quantiles of the posterior distribution sampled 15000 times by Markov Chain Monte Carlo (after a burn-in period of 100 000 samples, see Quince *et al*., 2008). The Required Sequencing Effort (RSE) to sequence 90% of the total richness was predicting by hierarchical simulation (see *SI*) and is quoted in terms of the number of final tags and as a multiple of the present sequencing effort. Point estimates and 95% prediction intervals (PIs) for RSE were obtained from the mean and (2.5%, 97.5%) quantiles from an ensemble of 80 simulations using the Sichel distribution.

Table 2. Isolates' closest relative according to BLAST results, % of identity with the BLAST reference strain (identity BLAST), GenBank accession number of the BLAST reference strain, number of tags matching the isolates sequences in the surface and bottom samples (Tags in Surface, Tags in Bottom), percentage of the tags in the surface and bottom samples (% Surface, % Bottom) and number of isolates of each taxa sequenced. Actino (Actinobacteria), Bact (Bacteroidetes), Firm (Firmicutes), Alpha-P (Alpha-Proteobacetria) and Gamma-P (Gamma-Proteobacteria).

Table 3. Isolates not matching the tag sequences. Isolates' closest relative according to BLAST results, % of identity with the BLAST reference strain (identity BLAST), GenBank accession number of the BLAST reference strain and number of isolates of each taxa sequenced. Actino (Actinobacteria), Bact (Bacteroidetes), Firm (Firmicutes), Alpha-P (Alpha-Proteobacetria) and Gamma-P (Gamma-Proteobacteria).

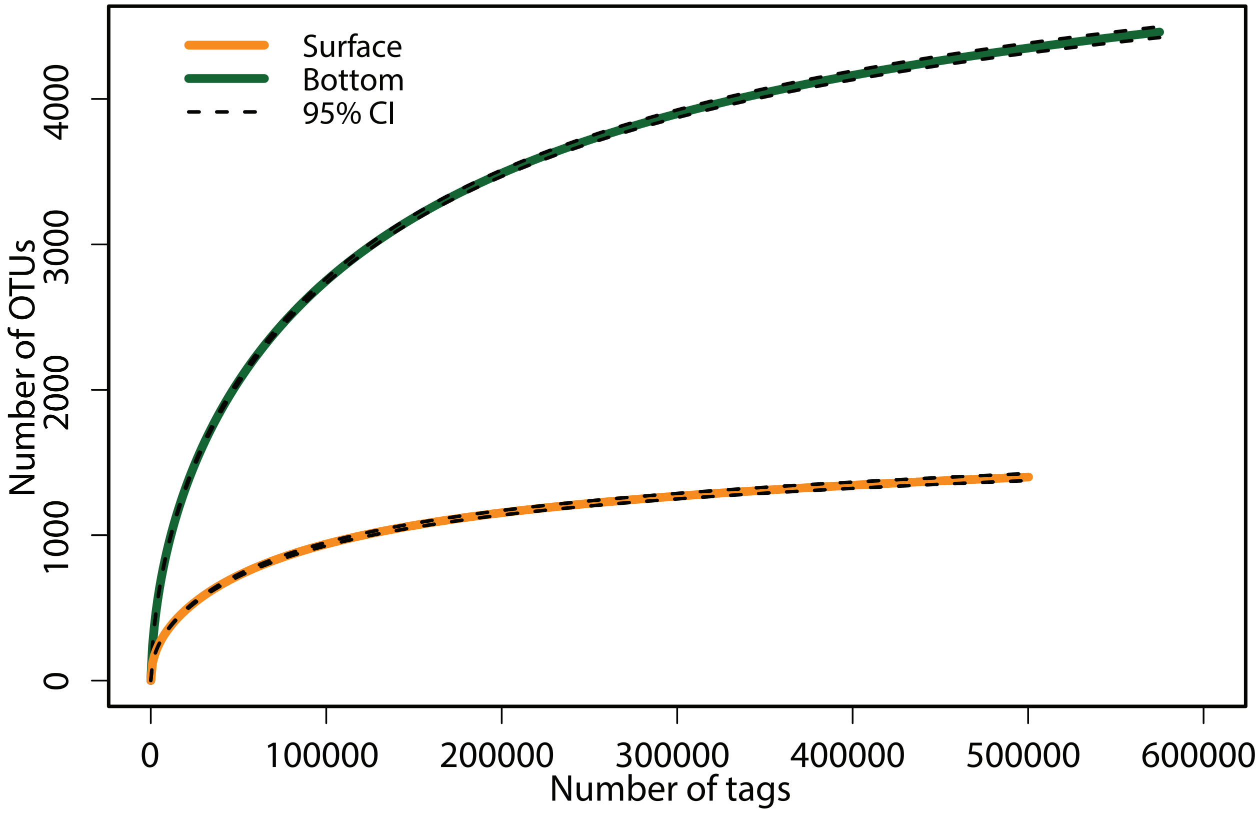

Figure 1
Crespo et al.

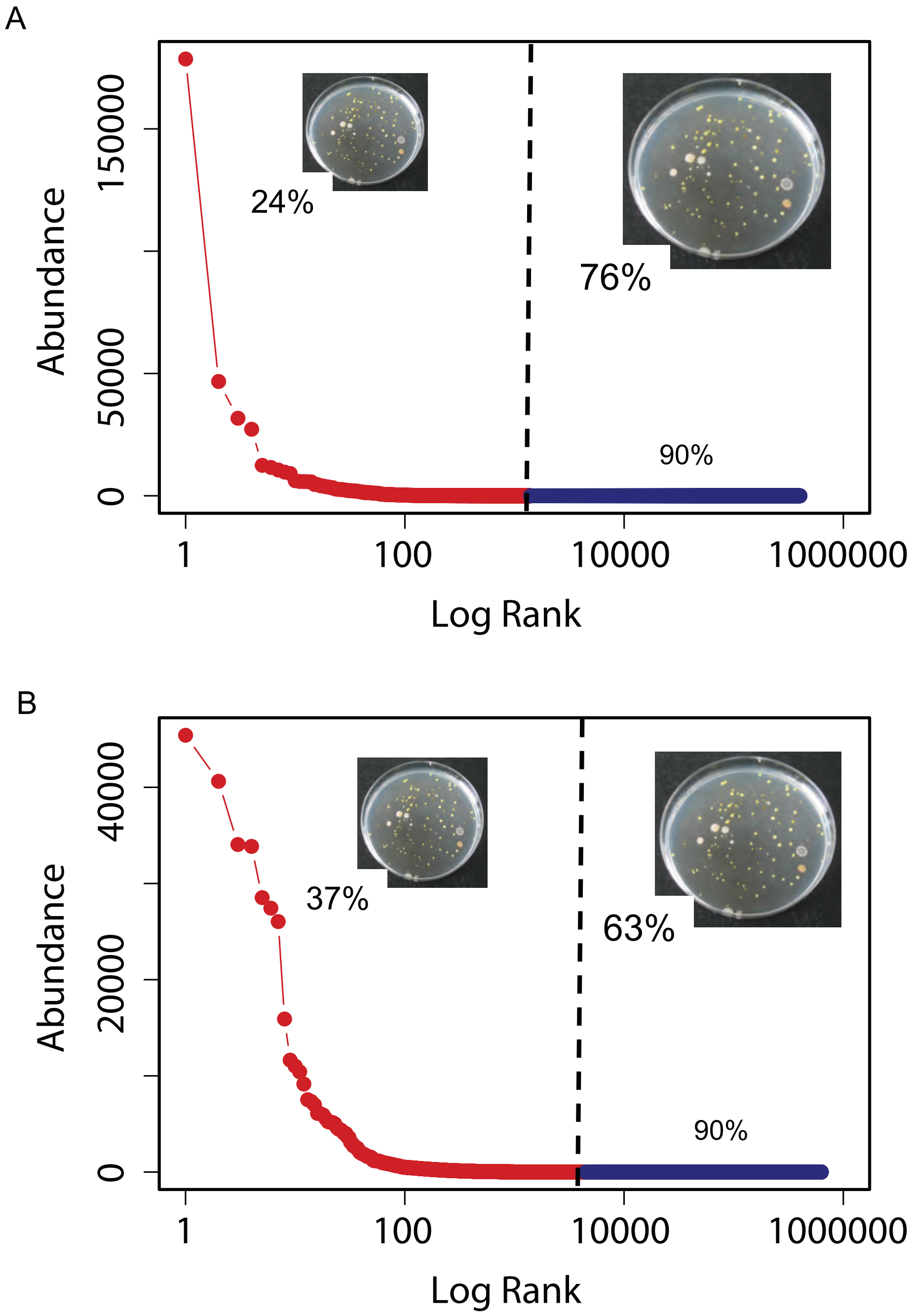

Figure 2
Crespo et al.

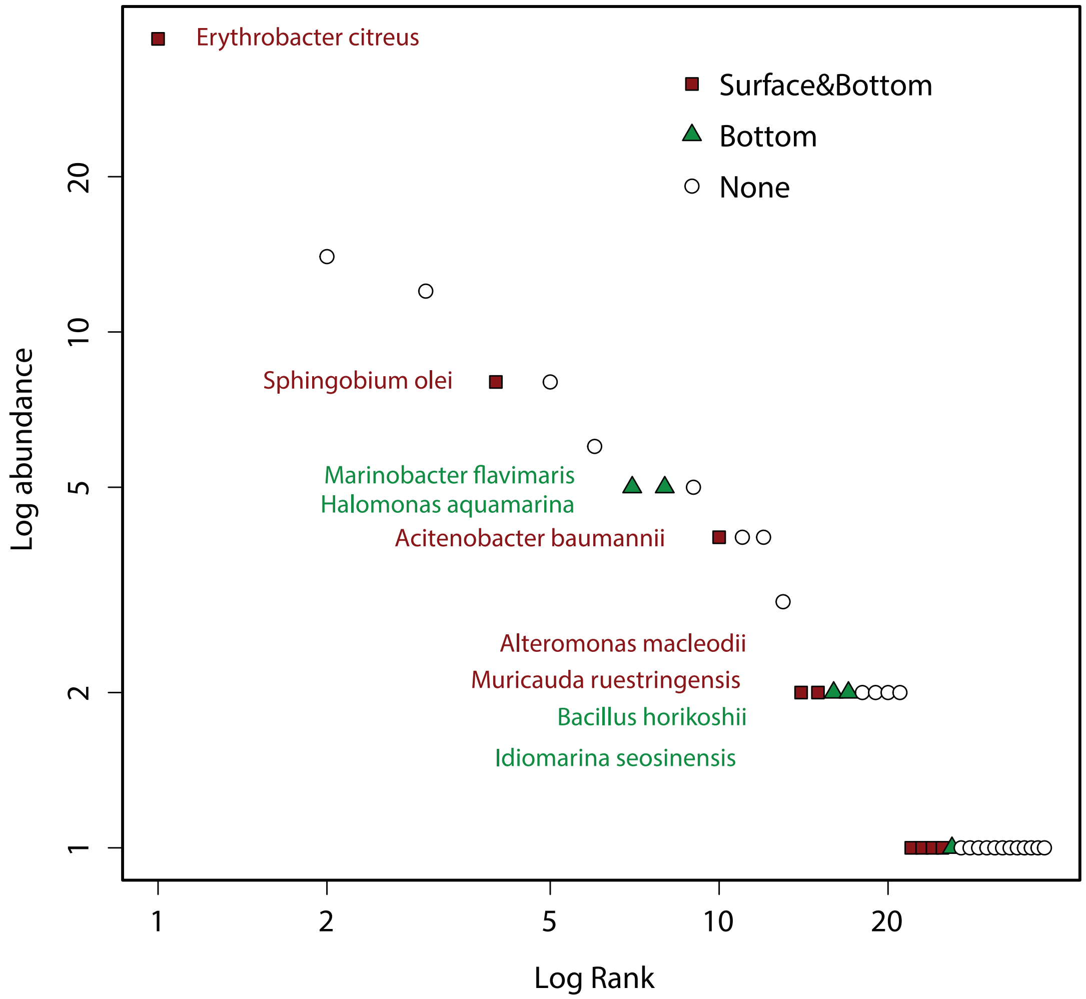

Figure 3
Crespo et al.

Table 1.

	Surface	Bottom
Lat, Long	40º52'N, 02º47'E	40º52'N, 02º47'E
Depth (m)	5	2 000
Raw Tags	713 076	970 346
Final Tags	500 262	574 960
OTUs 97 % identity (S_{obs})	1 400	4 460
Singletons (% OTUs)	17.86	17.2
Diversity estimates:		
H' (Shannon diversity index)	3.26	4.75
D (Simpson diversity)	0.45	0.66
J' (Pielou's eveness)	0.45	0.57
Total richness (S):		
Chao1 point estimate	1 646	5 031
Sichel point estimate	1 615	5 109
Sichel 95% CI	1 568–1 669	5 027–5 196
Required Sampling Effort (RSE) for 90% of total richness:		
Sichel point prediction (final tags)	$0.9x10^{6}$	$1.2x10^{6}$
Sichel 95% PI (final tags)	$(0.3\text{-}2.2)x10^{6}$	$(0.6\text{-}1.9)x10^{6}$
Sichel point prediction / present	1.8	2
Sichel 95% PI / present effort	0.6–4.3	1.0–3.2

Table 2.

Isolates’ closest relative	Identity BLAST	GenBank accession number
Uncultured Brevundimonas sp. (Alpha-P)	99.90%	JX047099
Alteromonas macleodii str. 'Balearic Sea AD45' (Gamma-P)	100%	CP003873
Sphingobium olei (Alpha- P)	100%	HQ398416
Erythrobacter citreus (Alpha- P)	100%	EU440970
Citromicrobium sp. (Alpha- P)	100%	HQ871848
Acinetobacter baumannii (Gamma- P)	100%	JX966437
Bizionia sp. (Bact)	100%	EU143366
Muricauda ruestringensis (Bact)	99%	JN791391
Microbacterium jejuense (Actino)	100%	AM778450
Marinobacter flavimaris (Gamma-P)	100%	AB617558
Bacillus sp. (Firm)	100%	AM950311
Bacillus horikoshii (Firm)	100%	JQ904719
Halomonas aquamarina (Gamma- P)	100%	AB681582
Idiomarina seosinensis (Gamma- P)	99.90%	EU440964

Tags in Surface	% Surface	Tags in Bottom	% Bottom	Number of isolates
76	$1.52\text{x}10^{-2}$	172	$2.99\text{x}10^{-2}$	1
40	$8.00\text{x}10^{-3}$	7526	1.31	2
34	$6.80\text{x}10^{-3}$	232	$4.04\text{x}10^{-2}$	8
31	$6.20\text{x}10^{-3}$	861	$1.50\text{x}10^{-1}$	37
22	$4.40\text{x}10^{-3}$	39	$6.78\text{x}10^{-3}$	1
16	$3.20\text{x}10^{-3}$	128	$2.23\text{x}10^{-2}$	4
13	$2.60\text{x}10^{-3}$	66	$1.15\text{x}10^{-2}$	1
4	$8.00\text{x}10^{-4}$	92	$1.60\text{x}10^{-2}$	2
1	$2.00\text{x}10^{-4}$	15	$2.61\text{x}10^{-3}$	1
0	0	174	$3.03\text{x}10^{-2}$	5
0	0	17	$2.96\text{x}10^{-3}$	1
0	0	8	$1.39\text{x}10^{-3}$	2
0	0	1	$1.74\text{x}10^{-4}$	5
0	0	1	$1.74\text{x}10^{-4}$	2

Table 3.

Isolates' closest relative	Identity BLAST	GenBank accession number	Number of isolates
Microbacterium aquimaris (Actino)	99.60%	HQ009858	14
Thalassospira sp. (Alpha-P)	100%	EU440837	12
Fulvimarina pelagi (Alpha-P)	96%	HQ622550	8
Alcanivorax sp. (Gamma-P)	99.70%	AB681671	6
Devosia subaequoris (Alpha-P)	100%	JQ844475	5
Alterierythrobacter sp. (Alpha-P)	100%	FM177586	4
Alteromonas macleodii (Gamma-P)	99.90%	CP003917	4
Erythrobacter sp. (Alpha-P)	100%	AB429073	3
Brevundimonas sp. (Alpha-P)	99.90%	HQ830182	2
Roseivirga spongicola (Bact)	99.80%	NR043531	2
Devosia hwasunensis (Alpha-P)	99%	HQ697727	2
Rhizobiales family (Alpha-P)	96%	HQ622550	2
Arthrobacter oxydans (Actino)	100%	EU086823	1
Emticicia sp. (Bact)	100%	JX426065	1
Halomonas sp. (Gamma-P)	100%	HE586874	1
Marinobacter hydrocarbonoclasticus (Gamma-P)	100%	JQ799097	1
Nitratireductor sp. (Alpha-P)	99.90%	AM981316	1
Nocardioides marinus (Actino)	99.90%	NR043787	1
Pseudomonas sp. (Gamma-P)	99.90%	JN244973	1
Sphingobium yanoikuyae (Alpha-P)	99.90%	DQ659593	1
Thalassospira permensis (Alpha-P)	99.90%	FJ860275	1
Alphaproteobacterium	99.80%	AY515421	1
Martelella mediterranea (Alpha-P)	99.80%	EU440955	1
Uncultured *Nitratireductor* sp. (Alpha-P)	99.70%	AM981316	1