Large-scale biogeography of marine pelagic Bacteria and Archaea

Guillem Salazar Guiral

Universitat Politècnica de Catalunya

Institut de Ciències del Mar (ICM-CSIC)

Large-scale biogeography of marine pelagic Bacteria and Archaea

Guillem Salazar Guiral

Directores:

Dr. Josep M. Gasol

Dra. Silvia G. Acinas

Dpto. Biologia Marina y Oceanografia

Instituto de Ciencias del Mar (ICM-CSIC)

Diciembre 2018

Tesis doctoral presentada para la obtención del título de Doctor por la Universidad Politécnica de Catalunya

Programa de doctorado de Ciencias del Mar

Al meu pare, a la meua mare,
A Maria.

A Pep, a Silvia, açò es tan meu com vostre!

Contents

Summary

The dark ocean contains about 70% of the ocean's microbial cells and 60% of its heterotrophic activity, which is mainly fueled by the flux of organic particles produced in the surface ocean and exported to the bathypelagic ocean (1,000 – 4,000 m depth). The bathypelagic ocean represents a nonhomogeneous environment and contains a variety of particles that are considered as the main supply of organic carbon to this environment. The microorganisms inhabiting this realm play a pivotal regulatory role in the biogeochemical cycles at a planetary scale. Accordingly, the study of these microorganisms is an essential step to decipher the ecological functioning of the deep ocean.

Chapters 1 to 3 in this Thesis are dedicated to the description of the prokaryotic community composition in the bathypelagic ocean at a global scale through the sequencing of ribosomal DNA and RNA fragments using data collected during the Malaspina 2010 expedition. Chapter 1 identifies the dominant prokaryotes in the deep ocean and reveals a high proportion (~50%) of previously undescribed prokaryotes. The water masses and the structure of the deep ocean's floor, organized into basins, are identified as the main drivers of their biogeography. Chapter 2 addresses the differences between free-living and particle-attached bathypelagic prokaryotic communities. This is shown to be a phylogenetically conserved trait, indicating that the bathypelagic particles and the water surrounding them constitute two distinct niches and that transitions from one to the other have been rare at an evolutionary timescale. Finally, in Chapter 3 we identify a linear relationship between the 16S RNA/DNA ratio and particle attachment preference, suggesting a global relationship between the prokaryote's preference for a particle-attached lifestyle and their growth rate.

While the deep ocean is a highly unexplored environment, a more complete knowledge exists for the epipelagic ocean (0 – 200 m depth). Steep gradients of light intensity and quality, temperature and nutrient availability characterize the oceans and impact on the distribution of species. However, different processes, such as the sinking of particles and the vertical movement of water masses, have been described as mechanisms capable of connecting the surface and deep layers of the ocean. These same processes could transport entire prokaryotic communities, a process theoretically proposed but never tested. In Chapter 4 we develop a tool (mtagger) for the extraction of short 16S ribosomal reads from metagenomes to describe the taxonomical composition of microbial communities. We propose and evaluate technical improvements compared to previous versions as a benchmark for its use in the last chapter. Chapter 5 is dedicated to the development of

a modeling tool (disperflux) for the analysis of prokaryotic communities' connectivity using data collected during the Tara Oceans expedition. We observe and describe a fast-decay relationship between community similarity and depth, which is consistently fitted by a power-law across the whole dataset, with the exception of 5 stations that are compatible with events of whole community export from the photic ocean to the mesopelagic.

In summary, this Thesis significantly contributes to the knowledge on the ecological functioning of marine prokaryotes by describing the structure of prokaryotic communities along the bathypelagic realm and the vertical gradient of the ocean and by the development of original methodological tools that may be applied to a variety of environments.

Resum

L'oceà profund conté el al voltant d'un 70% de las cèl·lules microbianes de l'oceà, les quals suposen el 60% de la activitat heterotròfica. Aquesta activitat biològica està mantinguda por un flux de partícules orgàniques produïdes a oceà superficial y exportades al batipelàgic (1,000 - 4,000 m de profunditat). L'ecosistema batipelàgic no és un ambient homogeni, sinó que conté una varietat de partícules considerades la font dominant de carboni orgànic. Els microorganismes d'aquest ambient tenen, doncs, un paper regulador central als cicles biogeoquímics planetaris. Conseqüentment, l'estudi d'aquests microorganismes suposa un pas essencial per desxifrar el funcionament ecològic de l'oceà profund.

Els Capítols 1 a 3 d'aquesta Tesi estan dedicats a la descripció a nivell global de la composició de les comunitats de procariotes a oceà batipelàgic mitjançant la seqüenciació de fragments de l'ADN y ARN ribosomal fent servir dades recollides a l'Expedició Malaspina 2010. Al Capítol 1 s'identifiquen els procariotes dominants a l'oceà profund y es revela l'existència d'una elevada proporció (~50%) de procariotes prèviament no descrits. Es reconeixen, a més, les masses d'aigua y l'estructura del sòl oceànic, organitzat en conques, como factors clau per la seva biogeografia. Al Capítol 2 s'estudien las diferencies entre les comunitats de procariotes de vida lliure y aquells adherits a partícules. Aquest tret es demostra estar conservat filogenèticament, indicant que les partícules del batipelàgic y l'aigua que las envolta constitueixen dos nínxols clarament diferenciats y que les transicions entre l'un y l'altre per part dels procariotes han sigut esdeveniments poc freqüents a escales evolutives. Finalment s'identifica al Capítol 3 una relació lineal entre el quocient de ARN/ADN ribosomal y la preferència per un mode de vida adherit a partícules, que suggereix una relació a nivell global entre l'adherència a partícules y la seva taxa de creixement.

Mentre l'oceà profund es un ambient amplament inexplorat, existeix un major coneixement de l'oceà superficial o epipelagic (0 - 200 m de profunditat). Els oceans es caracteritzen per gradients intensos en la quantitat i qualitat de la llum, la temperatura y la concentració de nutrients que n'influeixen en la distribució vertical de les especies. No obstant això, diferents processos, com la deposició de partícules o els moviments verticals de masses d'aigua, s'han descrit com mecanismes capaços de connectar les capes superficials y profundes de l'oceà. Aquests mateixos processos podrien teòricament exportar comunitats completes de microorganismes, un procés teòricament proposat però encara mai avaluat empíricament.

Al Capítol 4 es desenvolupa una eina informàtica (mtagger) per l'extracció de fragments del gen 16S ribosomal de metagenomes y la seva utilització en la descripció taxonòmica

de comunitats de procariotes. En aquest capítol es proposen y avaluen millores respecte a versions anteriorment utilitzades, com a pas previ al seu ús al darrer capítol. El Capítol 5 està dedicat al desenvolupament d'un model matemàtic (disperflux) per a la descripció de la connectivitat entre comunitats de procariotes fent servir dades recollides durant l'Expedició Tara Oceans. S'observa y descriu una disminució abrupta de la similitud de les comunitats de procariotes amb la profunditat. Aquesta relació s'ajusta a una equació potencial que resulta consistent al llarg de tot l'oceà, a excepció de 5 localitzacions, que es demostren compatibles amb esdeveniments d'exportació massiva de comunitats de la superfície a l'oceà profund.

En resum, aquesta tesis ha contribuït significativament al coneixement del funcionament ecològic dels procariotes marins mitjançant la descripció a nivell global d'aquestes comunitats a l'oceà profund y a través del gradient vertical y mitjançant el desenvolupament d'eines metodològiques novedoses aplicables a una gran varietat d'ambients.

Resumen

El océano profundo contiene el 70% de las células microbianas del océano las cuales suponen el 60% de la actividad heterotrófica. Dicha actividad biológica está mantenida por un flujo de partículas orgánicas producidas en el océano superficial y exportadas al océano batipelágico (1,000 - 4,000 m de profundidad). Éste no es, por tanto, un ambiente homogéneo, sino que contiene una variedad de partículas consideradas el aporte dominante de carbono orgánico. Los microorganismos de este ambiente tienen, por tanto, un papel regulatorio central en los ciclos biogeoquímicos planetarios. Consecuentemente, el estudio de estos microorganismos supone un paso esencial para descifrar el funcionamiento ecológico del océano profundo.

Los Capítulos 1 a 3 de esta Tesis están dedicados a la descripción a nivel global de la composición de las comunidades de procariotas en el océano batipelágico mediante la secuenciación de fragmentos del ADN y ARN ribosomal utilizando datos recolectados durante la Expedición Malaspina 2010. En el Capítulo 1 se identifican los procariotas dominantes en el océano profundo y se revela la existencia de una alta proporción (~50%) de procariotas previamente no descritos. Se reconocen además las masas de agua y la orografía del fondo oceánico, organizado en cuencas, como factores claves en su biogeografía. En el Capítulo 2 se estudian las diferencias entre las comunidades de procariotas de vida libre y aquellos adheridos a partículas. Este rasgo se demuestra estar conservado filogenéticamente, indicando que las partículas del batipelágico y el agua que las rodea constituyen dos nichos claramente diferenciados y que las transiciones entre uno y otro por parte de los procariotas han sido eventos poco frecuentes a escalas evolutivas. Finalmente se identifica en el Capítulo 3 una relación lineal entre el cociente de 16S ARN/ADN ribosomal y la preferencia a un modo de vida adherido a partículas, lo que sugiere una relación a nivel global entre la adherencia a partículas y su tasa de crecimiento.

Mientras el océano profundo es un ambiente ampliamente inexplorado, existe un mayor conocimiento del océano superficial o epipelágico (0 - 200 m de profundidad). Gradientes intensos en la cantidad y calidad de la luz, temperatura y concentración de nutrientes caracterizan a los océanos e influyen en la distribución vertical de las especies. Sin embargo, diferentes procesos, tales como la deposición de partículas o los movimientos verticales de masas de agua, se han descrito como mecanismos capaces de conectar las capas superficiales y profundas del océano. Estos mismos procesos podrían teóricamente exportar comunidades enteras de microorganismos, un proceso teóricamente propuesto pero no evaluado hasta la fecha.

En el Capítulo 4 se desarrolla una herramienta informática (mtagger) para la utilización de fragmentos del gen 16S ribosomal extraídos de metagenomas y su utilización para la descripción taxonómica de comunidades de procariotas. En este capítulo se proponen y evalúan mejoras respecto a versiones anteriormente utilizadas, como paso previo a su uso en el último capítulo. El Capítulo 5 está dedicado al desarrollo de un modelo matemático (disperflux) para la descripción de la conectividad vertical entre comunidades de procariotas usando datos recolectados durante la Expedición Tara Oceans. Se observa y describe una disminución abrupta de la similitud de las comunidades de procariotas con la profundidad. Esta relación se ajusta a una ecuación potencial que resulta consistente a lo largo de todo el océano, a excepción de 5 localizaciones, que se demuestran compatibles con eventos de exportación masiva de comunidades desde la superficie al océano profundo.

En resumen, esta tesis ha contribuido significativamente al conocimiento del funcionamiento ecológico de los procariotas marinos mediante la descripción a nivel global de estas comunidades en el océano profundo y en el gradiente vertical así como mediante el desarrollo de herramientas metodológicas novedosas aplicables a una gran variedad de ambientes.

Introduction

Marine prokaryotes and the ocean structure

The oceans cover 3.6 x 10^8 km^2 (71% of the Earth's surface) and contain 1.4 x 10^{21} liters of water, 97% of the total water on Earth. It is thus the largest ecosystem of the planet and it is also the place where life originally evolved. Microbes, referred here as single-celled organisms, have accounted for all known forms of life for more than 80% of Earth's history and the development of all forms of life has depended and depends on present and past microbes. Currently, the marine microbes may account for as much as a 90% of the total ocean's biomass (Fuhrman *et al.* 1989; Whitman *et al.* 1998), for more than the 95% of the respiration in the oceans (del Giorgio & Duarte 2002) and carry out 45% of the primary production of the planet (Field *et al.* 1998). Among microbes, the prokaryotes, that is the Bacteria and Archaea, carry out virtually every known biologically-mediated chemical reaction (Kirchman 2008) and are thus essential for the ecological functioning of the ocean being involved in the biogeochemical cycles that channel matter and energy at a planetary scale.

The ocean is an interconnected ecosystem that possesses, however, a strong structure both in the horizontal and vertical axes. A series of steep gradients exist in the vertical structure of the ocean, ultimately determined by the fact that the surface water is the part of the ocean exposed to the atmosphere and thus to the major input of heat and energy. The bulk of primary production in the ocean is consequently carried out in the photic layer by phototrophic prokaryotes and eukaryotes, a process that involves the fixation of inorganic carbon into biomass through photosynthesis. A considerable part of this carbon is directly transferred to higher trophic levels by the consumption of primary producers or indirectly by its transformation into dissolved organic matter (DOM) and its re-incorporation to higher trophic levels through the microbial loop (Azam *et al.* 1983; Fenchel 2008). The rest of the primary production, between a 1 and 40%, is exported out of the photic zone in the form of particles originated from dead organic mater and faecal material (Ducklow *et al.* 2001). This particulate organic matter (POM) is consumed by heterotrophic microbes, mainly Bacteria, while it sinks to the ocean's floor. This mechanism of carbon export, from the atmosphere to the ocean's interior, the *biological pump* (Fig. 1), is the main biologically mediated mechanism of carbon sequestration in the ocean and jointly with chemolithoautotrophy fuels life in the deep ocean.

The ocean is not uniform in its horizontal axes, either. The waters that compose the surface ocean move constantly due to the winds originated by the differential heating of air masses and the rotation of the Earth. Apart from this physical forcing, the abiotic environment that surface marine microbes cope with is influenced by latitudinal and seasonal changes in most of their relevant features. Consequently, differences in the composition

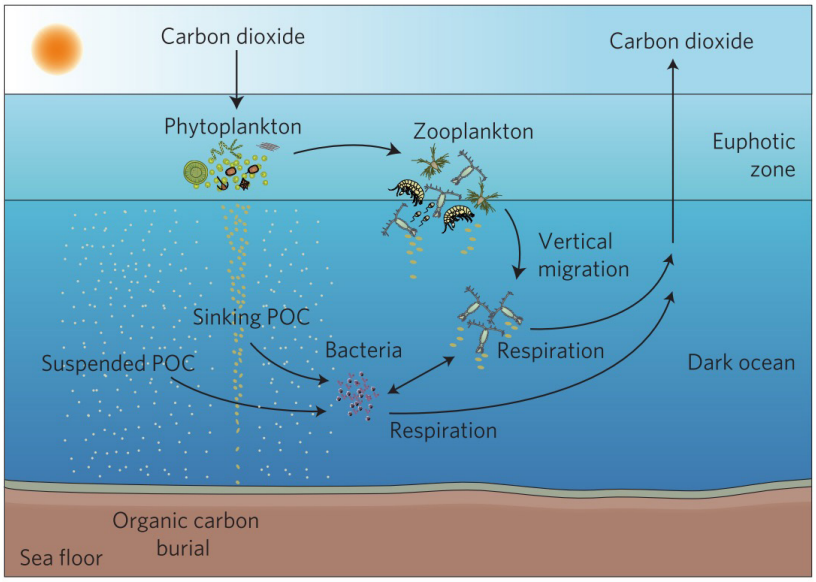


Figure 1: The biological pump. Phytoplankton in the euphotic zone use solar energy to fix carbon dioxide into organic carbon (OC), which is grazed on by herbivorous zooplankton, or consumed directly or indirectly by heterotrophic microbes. A fraction of primary production is exported out of the euphotic zone and remineralized in the oceanic water column. Only about 1% of the surface production reaches the seafloor as a large part gets respired/remineralized during sedimentation. In addition to the sedimenting particulate OC, another pool of OC is currently identified that can not be collected by sediment traps (the so-called "suspended POC"). (Figure from Herndl & Reinthaler, 2013).

of bacterial communities along space (Martiny *et al.* 2006; Zinger *et al.* 2011; Ghiglione *et al.* 2012; Amend *et al.* 2013; Sunagawa *et al.* 2015) and time (Brown *et al.* 2005; Fuhrman *et al.* 2006; Gilbert *et al.* 2009, 2012; Karl & Church 2014) have been reported in the past years. This effort for characterizing the biogeography of bacterial communities, although not exclusively, has been mainly centered in the surface ocean. Although most of the energy that fuels the deep ocean's life is dependent on processes occurring at shallower waters, the deep ocean is not subject to such strong seasonal fluctuations and is characterized by a higher homogeneity in its physico-chemical environment, at least in terms of salinity, temperature and nutrient concentration. In addition, the deep ocean is composed of different deep water-masses that hardly mix and maintain distinctive physical features and organic and inorganic nutrient concentrations. Additionally, below 200 m depth there are no effects of the wind-generated currents that act at the ocean surface. Instead, the differences of temperature and salinity between water masses create density differences

that fuel a relatively slow worldwide circulation of the deep oceanic water, the thermohaline circulation system or *global conveyor belt* (Rahmstorf 2003). The action of the water masses transporting entire microbial communities, jointly with their physico-chemical differences are likely to influence the distribution of Bacteria and Archaea in the deep ocean, as has been reported for vertically segregated water masses (Agogué *et al.* 2011). Additionally, the bathypelagic ocean contains specific geographical features, mainly submarine mountains, not that prevalent in the surface ocean, that compartmentalize it into different basins. This may influence water circulation, organism dispersal and determine the deep oceans' connectivity and thus, prokaryotic biogeography. However, the study of the biogeography of deep ocean's prokaryotic communities has not been possible until very recently.

Finally, apart from the purely geographical axes of variation in the ocean described above (horizontal and vertical) a third axis is thought to be relevant for the marine prokaryotes in potentially any location. Any volume of water in the ocean contains a mixture of organic matter that is composed of a continuum between purely dissolved (DOM) and particulate organic matter (POM). This constitutes the material basis of prokaryotic life and, consequently, differences between the prokaryotes that live attached to the particles (PA) and the ones living freely (FL) in the water have been long reported, mainly in surface waters (DeLong *et al.* 1993; Turley & Mackie 1994). However, the differences in composition of these two sets of communities are poorly known in the deep ocean.

In addition, high taxonomic ranks (such as Orders or Phyla) have been shown to exhibit some degree of ecological coherence in respect to the PA-FL axis (Eloe *et al.* 2011; Smith *et al.* 2013; Crespo *et al.* 2013). That is, entire Orders or Phyla have been described as either FL or PA. Despite this, and the fact that different metabolic capabilities are needed for the exploitation of the DOM and POM resource types, few efforts have been conducted to understand, from an evolutionary perspective, the phylogenetic organization and evolution of this key prokaryotic trait.

The general broad goal of this Thesis is to describe the taxonomical variation of prokaryotic communities along the three aforementioned axes of variation in the global ocean: the horizontal and vertical geographical axes and the FL-PA axis.

High-throughput sequencing and worldwide diversity surveys

The emergence of HTS techniques

The study of the taxonomical composition of marine prokaryotic communities has experienced a revolution since the emergence of DNA-based techniques (Fig. 2). This transformation of our understanding of marine microbial diversity was triggered by the

pioneering work of Carl Woese and George Fox that used the sequencing of the 16S subunit of the ribosomal gene (rDNA) to organize the life's diversity within a phylogenetic framework (Woese & Fox 1977). This started a completely new approach for the description of the prokaryotic diversity based on full or partial sequences of the 16S rDNA. The possibility of describing uncultured prokaryotes was the next key step through the development of clone libraries containing the 16S rDNA sequences of a modest numbers of individuals (a few tenths of sequences) from an environmental sample (Pace *et al.* 1986). Later on, the samples contained hundreds, and a maximum of a few thousand sequences (Acinas *et al.* 2004). In the following decade these molecular tools revealed the magnitude of diversity contained in marine microbial communities among Bacteria (Giovannoni *et al.* 1990) and Archaea (DeLong 1992; Fuhrman *et al.* 1992). Construction and sequencing of environmental clone libraries of marine communities were used in the following decades for the description of the magnitude of the marine prokaryotic diversity, its varia-

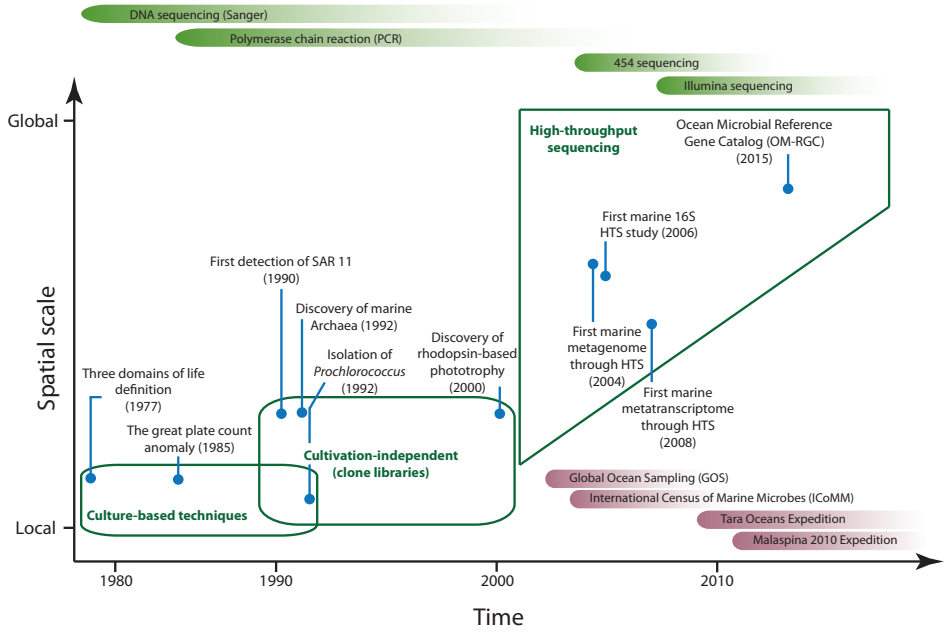


Figure 2: Timeline of the DNA-based marine microbial ecology history. The major events (blue dots), technological developments (green bands) and global sampling expeditions (red bands) are placed along time. Major events are approximately placed in a y-axis of spatial scale with arbitrary units from local studies to global studies. References matching the figure in chronological order (Woese & Fox 1977; Staley 1985; Giovannoni *et al*. 1990; DeLong 1992; Chisholm *et al*. 1992; Béjà *et al*. 2000; Venter et al. 2004; Sogin *et al*. 2006; Gilbert *et al*. 2008; Sunagawa *et al*. 2015)

tion through space and time (that is the biogeography), and its phylogenetic structure (e.g. Hagström *et al.* 2002; Acinas *et al.* 2004; Fuhrman *et al.* 2006; Pommier *et al.* 2007). Clone libraries have coexisted with a variety of fingerprinting molecular techniques, such as TRFLP (Liu *et al.* 1997), DGGE (Muyzer *et al.* 1993) or ARISA (Fisher & Triplett 1999). These techniques, without the need for sequencing and thus faster and more affordable, have been instrumental for addressing key ecological questions by allowing the comparison of a higher number of microbial communities with higher spatial and temporal resolution than clone libraries and thus making possible more sophisticated and extensive sampling designs.

A second era in the description of the diversity of marine microbial communities started with the developement of what nowadays are termed *high-throughput sequencing* (HTS) techniques. These are a series of technological advances that since 2007 have allowed the sequencing of a massive amount of short DNA sequences from a sample or combination of samples. While the traditional sequencing approach introduced in 1986 (i.e. "Sanger sequencing", Applied Biosystems) yielded on the order of hundreds of sequences per run, nowadays HTS techniques allow obtaining 5 to 7 orders of magnitude more sequences per run (Goodwin *et al.* 2016). HTS techniques opened the possibility of massively sequencing variable regions of marker genes (particularly the 16S rDNA) and thus characterizing the taxonomical diversity of microbial communities with a resolution that was previously impossible (Sogin *et al.* 2006). With these methods, it became also possible to directly sequence the whole DNA content of environmental samples (i.e. metagenomics) without the need for targeting (i.e. amplifying) a specific gene (Venter et al. 2004). This allowed obtaining information on the genomic content of the members of a microbial community beyond its taxonomical characterization. The use of HTS techniques applied to the 16S rDNA represents the "state of the art" technique for microbial ecology and has been applied in the last decade to virtually any marine environment. An explosion of studies using this approach has produced a detailed description of the microbial diversity of marine communities and has yielded a better understanding of the ecological organization of these communities.

DNA vs RNA: presence vs. activity

The targeting of the conserved gene coding for the small subunit of the prokaryotic ribosome, the 16S rDNA, is instrumental for the detection and quantification of the prokaryotes present in a community, as described above. It is possible, however, to directly target the ribosomal subunit (16S rRNA), instead of the gene coding for it, to inform on the activity of such prokaryotes. This is done through the extraction of the RNA from

the sample and the synthesis of the complementary DNA, from which the amplification through PCR and sequencing proceeds, from this point, similar to the rDNA-based analysis. Because growing cells require ribosomes for protein synthesis, rRNA can be used to characterize the active portion of communities, specifically the microbes with capacity for protein synthesis (Blazewicz *et al.* 2013). In the last decade the joint study of 16S rDNA and rRNA has been applied to characterize the active and inactive members of prokaryotic communities from aquatic environments, such as lakes (Jones & Lennon 2010; Denef *et al.* 2016) or the surface ocean (Campbell *et al.* 2009, 2011; Ghiglione *et al.* 2009; Campbell & Kirchman 2012; Hugoni *et al.* 2013; Hunt *et al.* 2013; Zhang *et al.* 2014). This exercise has not yet been conducted for the bathypelagic ocean. The ocean is fueled by the organic matter produced in the photic layer that reaches the deep ocean through the biological pump in form of sinking particles (Herndl and Reinthaler, 2013). Thus, the PA prokaryotes are likely to have access to a concentrated organic pool composed of polymeric materials (Minor *et al.*, 2003) in contrast to the FL prokaryotes in the deep ocean, that face an environment composed of diluted organic compounds (Arrieta *et al.*, 2015). This may drastically affect the activity of the FL and PA bathypelagic communities, affecting the deep ocean's biogeochemical carbon cycle. Here we not only describe the composition of the prokaryotic bathypelagic communities but also apply the combination of 16S rDNA and rRNA sequencing for elucidating which are the active and inactive members, and to test whether the PA communities contain more active members than the FL communities.

Global efforts at describing marine microbial diversity

Most of the steps involved in HTS-based studies targeting the 16S rDNA or rRNA lack standardized procedures and the use by different researchers different approaches can introduce a considerable degree of variation in the results. The RNA and DNA extraction step is known to impact the relative abundance of the sequences obtained in marine samples (Cruaud *et al.* 2014; Lekang *et al.* 2015). Even a greater impact on the final abundance estimates has been shown to be due to the dynamics of PCR amplification (Acinas *et al.* 2005; Haas *et al.* 2011) and to the PCR primer pair used (Hong *et al.* 2009; Engelbrektson *et al.* 2010). Moreover, several variable regions of the 16S gene have commonly been used for HTS-based studies (Klindworth *et al.* 2013), and thus, even if some of these biases are minimized, cross-study comparisons are impossible. This represents an unsolved problem yet, preventing comparisons of datasets that have been produced and analyzed with different methodologies, which is the case for virtually every study, particularly considering the velocity at which sequencing technologies and data analysis technics are evolving. As

a consequence, a worldwide scale study of marine microbial diversity has been hampered by the impossibility of merging the data from different studies. This limitation has been overcome by initiatives to survey the marine diversity at a worldwide scale with coherent sampling, sequencing and analysis protocols.

The pioneer expedition applying HTS techniques to marine microbial diversity was the Global Ocean Sampling expedition (GOS; www.jcvi.org/cms/research/projects/gos/overview). This included a pilot sampling project in 2003 in the Sargasso Sea and continued as a two-year expedition in 2004 - 2006. It was a proof-of-concept study of the metagenomic approach, that is, the sequencing of all the genomic material of all the organisms present within a sample, and demonstrated how this type of data could be used to extract functional gene information (Venter 2004). It also provided evidence for 1,300 distinct 16S rRNA sequences in seven surface seawater samples, and a total of 6 million gene families could be identified from an extended dataset in a transect between the central equator and the North Atlantic. However, the first worldwide survey to study marine microbial diversity through consistent standardized HTS was the International Census of Marine Microbes (ICoMM, 2005-2010) (Amaral-Zettler *et al.* 2010). This was an international effort aimed at collecting ~500 samples from a variety of marine environments (water column, thermal vents, sediments and others) around the world. The variable region V6 from the 16S rDNA was sequenced and analyzed using a common pipeline. This initiative dramatically changed the view of the marine microbial communities by describing the unprecedented extent of the microbial richness (i.e. number of different taxa). Key concepts in today's marine microbial ecology were termed as a result of the ICoMM, such as the rare biosphere (Sogin *et al.* 2006), and online databases were developed, such as the VAMPS (Visualization and Analysis of Microbial Population Structures; vamps.mbl.edu) and MICROBIS (Microbial Oceanic Biogeographic Information System; icomm.mbl.edu/microbis). The first worldwide biogeographical studies were possible thanks to this initiative, describing how microbial communities varied in their taxonomical composition along space (Galand *et al.* 2010; Pommier *et al.* 2010; Zinger *et al.* 2011) and time (Gilbert *et al.* 2009).

The two main worldwide surveys conducted to date after the launch of the ICoMM initiative have been the *Tara* Oceans Expedition (Karsenti *et al.* 2011) and the Malaspina 2010 Circumnavigation Expedition (Duarte 2015). The whole work of this Thesis has been conducted within these two initiatives. The *Tara* Oceans Expedition was launched in 2009 by an international consortium of more than 100 scientists from a variety of expertises (oceanographers, taxonomists, biologists, bioinformatics, microbial ecologists, etc.) and consisted on a 3-year expedition for the study of the global ocean ecosystem. In contrast

to the previous survey efforts, it emphasized a holistic approach to marine ecosystems by simultaneously sampling a wide range of organisms, from marine viruses to zooplankton covering up to six orders of magnitude in size (from 10^{-2} to 10^{5} µm) (Pesant *et al.* 2015). It covered most of the main oceans except the Arctic along 20 biogeographical provinces (Fig. 3) where samples for prokaryotic diversity were collected from surface waters down to the mesopelagic (i.e. up to 1,000 m depth). A variety of molecular approaches have been applied to the material collected through the cruise and as a result a considerable amount of studies have been conducted describing, among other aspects, the genetic diversity of the prokaryotic plankton (Sunagawa *et al.* 2015) as well as their biological interactions (Lima-Mendez *et al.* 2015). The scientific data generated, as well as the sampling and computational methods have also been made public (Pesant *et al.* 2015).

The data describing marine microbial diversity consists on a collection of metagenomes from the surface waters, the deep chlorophyll maximum (DCM) and the mesopelagic, a data set referred to as the Ocean Microbial Reference Gene Catalog (http://ocean-microbiome.embl.de/companion.html) that contains >40 million non-redundant sequences from viruses, prokaryotes, and picoeukaryotes from 139 samples. This constitutes the most complete dataset to date describing the genetic diversity of the marine microbes.

The Malaspina 2010 Circumnavigation Expedition was a 7-month sampling effort run by 35 Spanish research groups and 28 international partners to asses the state of the ocean in 2011 including the exploration, using advanced HTS tools, of the diversity of life in the ocean, with a particular emphasis in the dark ocean. It covered the main oceans but with a higher coverage of some regions in the southern latitudes than the Tara Oceans Expedi-

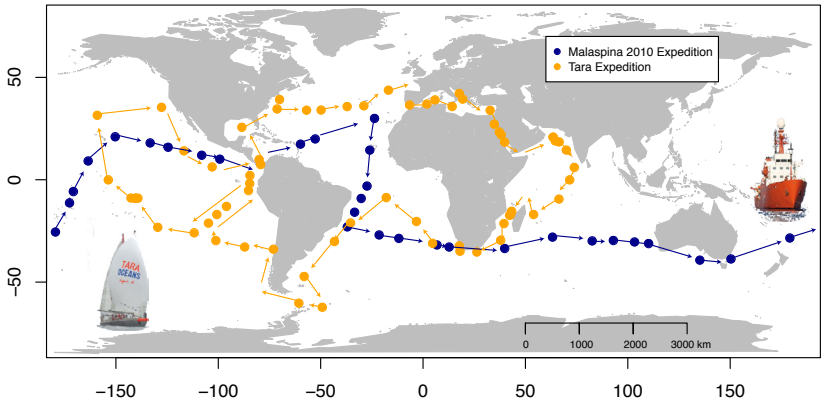


Figure 3: The *Tara* Oceans and Malaspina 2010 expeditions. The tracks (arrows) and the stations (points) from where samples were taken and are in this Thesis.

tion, such as the Indian Ocean (Fig. 3). A total of 146 sampling stations were profiled covering the water column from the surface to a maximum of 4,000 m depth with an especial effort in the bathypelagic realm (i.e. the layer contained between 1,000 and 4,000 m depth). More than 600 microbial biomass samples designated for DNA-based analyses were collected, a subset of which are the starting material of a considerable portion of this Thesis. A considerable number of studies have already been conducted based on the data collected through the Malaspina 2010 Expedition describing a variety of aspects of the ocean's ecology, such as an assessment of the magnitude of fish biomass in the mesopelagic (Irigoien *et al.* 2014), surface ocean plastic debris (Cozar *et al.* 2014), the turnover of the fluoresce organic matter (Catalá *et al.* 2015) or the factors limiting DOC consumption in the bathypelagic ocean (Arrieta *et al.* 2015). The global abundance and diversity of the heterotrophic protists in the bathypelagic have also recently been assessed (Pernice *et al.* 2015) and the study of prokaryotic diversity and its drivers in the bathypelagic ocean constitute three chapters of this Thesis.

Analysis of microbial diversity

Alpha-diversity and beta-diversity

The study of the diversity of ecological communities is the study of the number of species (richness) and how the abundances of these species are distributed within and among samples (evenness). This often is divided into the study of *alpha-diversity* and *beta-diversity*. This distinction was made by R.H. Whittaker in order to distinguish three spatial scales in the description of diversity (Whittaker 1960): a) *alpha-diversity*, i.e. the diversity of a defined assemblage or sample unit, b) *beta-diversity*, i.e. the change in diversity along transects or environmental gradients and c) *gamma-diversity*, i.e. the diversity of a complete landscape (generally considered as the total diversity of the area under study).

The study of alpha- and gamma-diversity with data derived from HTS is sensitive to several methodological features that make it problematic. On one hand, virtually all the studies of prokaryotic communities lack a sufficient sampling effort for capturing the whole richness (i.e. the total number of species) of a sample. When this is addressed through sequencing of the 16S rDNA, the sampling effort of a sample is the number of sequences (i.e. reads) obtained. This number is rarely sufficient to saturate the diversity within a sample, that is, to reach an amount of sequences above which no new species are detected. The sufficiency of sampling is usually explored with the use of individual-based accumulation or rarefaction curves, where the increase in the number of species uncovered is represented against the number of reads (i.e. the sampling effort). Complete sampling effort would result in a curve that reaches an asymptotic value, but the usu-

ally incomplete sampling effort results in curves far from reaching that plateau. Thus, the measurement of the richness is usually conducted through the use of statistical estimators (that attempt to correct the richness value trough the estimation of the unseen richness) or through the computation of the richness after standardizing all the samples to the same number of reads (for comparative purposes). However, the processing of HTS data may also affect richness and diversity estimates, especially the definition of the unit used as a proxy for species, usually referred as the operational taxonomic unit (OTU). These are defined through the use of a sequence similarity cutoff (generally 97%) and variations in this cutoff drastically alter the richness estimates. All these limitations make comparisons of richness and diversity values between different studies complicated. Despite that, estimates of alpha-diversity of different prokaryotic communities from different studies at different locations and depths and with different methodological approaches, seem to range between several hundreds to a few thousands OTUs (Pommier *et al.* 2007; Zinger *et al.* 2011; Ghiglione *et al.* 2012; Sunagawa *et al.* 2015). one of the aims of this Thesis is the description of the alpha-diversity of free-living and particle-attached communities in the global bathypelagic ocean alongside with the understanding of the mechanisms maintaining the richness and diversity of the prokaryotic communities along the water column from the photic layer to the mesopelagic. Gamma-diversity estimates including samples from the surface and the deep ocean are scarce yet range between 35,650 OTUs and 65,545 (Zinger *et al.* 2011; Sul *et al.* 2013; Sunagawa *et al.* 2015). No systematic effort has been conducted for the estimation of the total number of prokaryotic species inhabiting the bathypelagic ocean (i.e. its gamma-diversity), which will constitute another aim of the present Thesis.

The original definition of beta-diversity mentioned above was made with the final goal of quantifying and partitioning the diversity into three components, alpha, beta and gamma-diversity (Tuomisto 2010). However, there is no reason for restricting its original use to environmental gradients or transects (as originally defined) and not applying it to other spatial organizations or even to variations with time. This has gradually generated a variety of measures of beta-diversity and the use of this term for referring to any measure of the extent to which the diversity of two or more samples differs (Magurran 2004; Anderson *et al.* 2011). We have assumed this relaxed definition during this Thesis for referring to any estimate of the compositional difference between prokaryotic communities. Beta-diversity, understood as the measure of the difference in species composition between two communities, is the starting point for the study of biogeography at the community level. The study of the beta-diversity of prokaryotic communities along both

the horizontal and vertical geographical axes of the ocean, as well as the beta-diversity differences between PA and FL communities in the bathypelagic ocean constitute also central aims of this Thesis.

Biogeography: the classical biogeographical framework

Biogeography is the study of the distribution of organisms across space and time. It aims to reveal where organisms live, at which abundances, and why. The study of biogeography offers insights into the mechanisms that generate and maintain ecological diversity (Lomolino 2010). One of the object of biogeography, thus, is to demonstrate non-random patterns in the composition of communities, in other words, non-random beta-diversity patterns. The existence of microbial biogeographic patterns has been well established in the last decade (Martiny *et al.* 2006; Ramette & Tiedje 2007; Lindström & Langenheder 2012). However, the ultimate goal of microbial biogeography is to understand and reconstruct, if possible, the ecological processes underlying the present biogeographical patterns.

Four are the main processes capable of creating biogeography (Vellend 2010): *environmental selection* (also named species sorting), *ecological drift*, *dispersal* and *speciation*. Speciation acts by adding new species diversity through an evolutionary process and is generally excluded from most of the analyses, as is supposed to act at long time-scales, compared to the remaining three processes. Environmental selection is the effect of environmental factors over the survival capacity of organisms and thus creates biogeographical patterns by selecting different organisms in different locations based on the survival capacity of each species to each environment. Ecological drift is the result of changes in the relative abundance of the various species in a location due to chance demographic fluctuations. Pure ecological drift assumes that the different species are demographically identical and not subject to any kind of environmental selection, and thus, drift as the only mechanism acting to generate biogeography is unlikely. In adddition, the effect of ecological drift is inversely related to population sizes, and thus unlikely to be relevant for microbial populations except in special situations such as populations experiencing drastic fluctuations in population sizes. Finally, dispersal is defined as the movement and successful establishment of a species to a new location. High dispersal rates between locations increase the similarity between communities through the exchange of individuals. The most extreme case would be the arrival of enough individuals of a species in events of massive immigration to prevent competitive extinction of that species in one location, a process named *mass effects* (Lindström & Langenheder 2012). However, species dispersal limitation is generally the process considered as capable of generating biogeography. If

present, dispersal limitation leads to community differentiation, because of the inability of species to colonize different locations or because the dispersal of individuals is not sufficiently strong to erase the differences produced by drift or environmental selection.

The four ecological processes involved in shaping biogeography are usually grouped into "present environmental selection" or "historical processes" in a long-standing theoretical framework that in the absence of a consensus denomination, we will call the *classical biogeographical framework* (reviewed in Hanson *et al.* 2012). Present environmental selection corresponds to the influence of the current environment on the current distribution of microbial diversity and is detected when finding a significant correlation between community composition and the variables that define the environment. Historical processes correspond to the past action of environmental selection or ecological drift in combination with some degree of dispersal limitation: if dispersal is not completely efficient, drift or past environmental selection will leave a legacy on the current distribution of communities. Thus, a significant correlation between geographical distance and community composition, after controlling for the current environmental effect, would be indicative of the action of historical processes, i.e. implying some degree of dispersal limitation (Nekola *et al.* 1999; Martiny *et al.* 2006). Microbial communities inhabiting different environments (marine, inland waters, soils and others) have been studied within this framework in recent years in order to decipher the relative importance of the processes shaping microbial biogeography (reviewed in Hanson *et al.* 2012): the effect of environmental selection has generally a greater impact although some effect of dispersal limitation is detected in most of the studies. The bathypelagic ocean possesses specific features compared to shallower ocean layers. As mentioned above, the bathypelagic ocean is connected by a system of slow currents independent from the ones connecting the photic waters. It is also thought to be more homogeneous in its physico-chemical environment than the surface waters, to which is mainly connected through the sinking of organic particles, and it is also compartmented into basins. All these specific features may impact the relative importance of the processes governing its biogeography yet the biogeography of the prokaryotes inhabiting the deep ocean and the study of the processes generating it has not been properly assessed and constitutes an aim of this Thesis too. In contrast, the vertical structure of the upper ocean is characterized by steep gradients in most environmental factors (light, nutrients, temperature, etc.) and the presence of contrasting water masses in a short vertical distance. This has been described to have an impact, segregating the prokaryotic diversity into distinct communities according to water mass composition (Agogué *et al.* 2011) and the vertical axis has been identified as one of the most important axes structuring prokaryotic community composition at global scales (Zinger *et al.* 2011;

Sunagawa *et al.* 2015). However, the photic and aphotic oceans are not completely unconnected environments, nor the filtering by the environmental differences between these two layers is the only possible mechanism shaping their communities. A variety of purely physical processes as well as the sinking of organic particles may modulate the biological connectivity of these two realms. A final objective of this Thesis is to gain a more mechanistic understanding of the connectivity between the prokaryotic communities from the photic layers and those of the mesopelagic ocean through the development of a process-based model that we test with Tara Oceans data.

A process-based approach to biogeography

Process-based approaches in community ecology have increasingly been applied since the formulation of the Unified Neutral Theory of Biodiversity (UNTB) (Hubbell 2008). This theory establishes a neutral model for the assembly of ecological communities, which is understood as the result of the combined action of speciation, ecological drift and dispersal limitation. Thus, it constitutes a null hypothesis to niche theory (Pocheville 2015), as the role of environmental filtering (i.e. the existence of niche differences between species) is not necessary to explain several features of real communities. Such approach has been applied to reproduce the rank abundance of microbial communities (Sloan *et al.* 2006, 2007; Woodcock *et al.* 2007; Ofiteru *et al.* 2010) and their patterns of evolutionary distance (Jeraldo *et al.* 2012). Not intended to be a detailed representation of the complex processes shaping the composition of ecological communities, process-based approaches have the strength of linking patterns and processes. That is, they consist on the development of simplified mathematical models that explicitly assume the existence of a reduced number of ecological processes (speciation, drift and dispersal limitation in the case of the UNTB). The predictions of these models may be tested against the patterns observed in real communities. At the cost of constituting a simplified representation of the reality, their strength compared to other approaches, such as the *classical biogeographical framework,* is that they may be understood as null models (Gotelli & McGill 2006) that explicitly link the patterns observed in real communities to the potential ecological processes responsible for these patterns by modeling them.

One type of poorly studied processes that may create biogeographical patterns in ecological communities are the *mass effects*. As mentioned above, the vertical structure of the ocean is characterized by abrupt changes in physico-chemical variables that produce contrasting environments in a short vertical distance. This constitutes a feasible scenario where mass effects may be important for the biogeography of vertically segregated prokaryotic communities. We develop, as the final objective of this Thesis, a process-based

model that explicitly takes into account the effect of events of directional transport of prokaryotic communities on prokaryotic community structure, a special case of mass effect theoretically proposed for microbes but not yet developed (Rillig *et al.* 2015). We use this process-based approach for understanding at a global scale the connectivity between the photic ocean and the mesopelagic.

Aims of this thesis

The general aim of this thesis is to draw a general picture of the structure of the marine prokaryotic communities along the three main axis of variation evaluated as relevant in the introduction: a) the horizontal geographic axis in the deep ocean, b) the vertical geographical axis through the water column along the epipelagic and mesopelagic ocean and c) the axis defined by the dissolved/particulate nature of organic matter in the bathypelagic ocean that corresponds to the free-living and particle-attached communities.

The achievement of this goal is structured in five chapters. The first chapter (*Global diversity and biogeography of deep-sea pelagic prokaryotes*, ISME J. 2016) is designed as a first global description of the diversity of the bathypelagic prokaryotic communities and an analysis of the process responsible for their biogeography. In the second chapter (*Particle-association lifestyle is a phylogenetically conserved trait in bathypelagic prokaryotes*, MolEcol, 2015), the PA and FL lifestyles of bathypelagic prokaryotes are analyzed in order to gain insights into their distribution following phylogeny and the evolutionary processes that may have shaped them. The third chapter (*The activity of deep ocean prokaryotes is driven by their particle-association lifestyle*, unpublished) is dedicated to the identification of the active and inactive members of the bathypelagic communities and the relation between the activity and the particle-related lifestyle of bathypelagic prokaryotes. The fourth chapter (*mtagger: an R package for the characterization of microbial communities through rDNA metagenomic fragments*, unpublished) consists on a methodological development facilitating the use of metagenomic rDNA data as an alternative to amplicon sequencing. This is used to develop the fifth chapter (*Vertical microbial connectivity in the global ocean*, unpublished) where the vertical structure of epipelagic and mesopelagic communities are analyzed trough the lens of a process-based model in order to better understand the microbial connectivity of these two oceanic layers.

The outline of different topics studied can be explained under four general objectives and several specific ones, as follows:

Objective 1: *Description of the prokaryotic communities of the bathypelagic ocean at a global scale* (Chapter 1).

1.1 Evaluate the species richness of the bathypelagic Ocean compared to the photic layers.

1.2 Asses how the bathypelagic prokaryotic biogeography is associated to specific fea-

tures of this realm.

1.3 Given the differential ecological characteristics of particle-attached (PA) and free-living (FL)prokaryotes, testshow these two sets of prokaryotic communities differ in their biogeography.

Objective 2: *Phylogenetic analysis of the particle-attached and free-living lifestyles of bathypelagic prokaryotes* (Chapter 2).

2.1 Asses and measure the existence of two distinct prokaryote particle-association lifestyles (PA and FL) in the bathypelagic ocean.

2.2 Test wether these PA and FL lifestyles are phylogenetically conserved.

Objective 3: *Identification of the active and inactive members of bathypelagic communities and their relation to the particle-associated lifestyle* (Chapter 3).

3.1 Measure the proportion of inactive species (i.e. detected in the 16S rDNA pool but not in the 16S rRNA pool) within the bathypelagic communities.

3.2 Test wether the active/inactive state of each species is related to their particle-association lifestyle, expecting the PA prokaryotes to be more active that the FL ones.

Objective 4: *Understanding of the mechanisms maintaining the vertical structure of the epipelagic and mesopelagic prokaryotic communities* (Chapters 4 and 5).

4.1 Thest wether the use of 16S rDNA reads from metagenomic data introduces biases in the estimate of community diversity and if it may be corrected by a modified mapping strategy to reference databases (Chapter 4).

4.2 Asses the existence of universal patterns of community similarity in the vertical organization of prokaryotic communities from the surface to the mesopelagic ocean at a global scale (Chapter 5).

4.3 Design and evaluate measures of directional connectivity between communities trough a process-based model and test wether they might reveal patterns not detectable with classical community similarity indices (Chapter 5).

References

Acinas SG, Klepac-Ceraj V, Hunt DE *et al.* (2004) Fine-scale phylogenetic architecture of a complex bacterial community. *Nature*, **430**:551–4.

Acinas SG, Sarma-Rupavtarm R, Klepac-Ceraj V, Polz MF (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**:8966–9.

Agogué H, Lamy D, Neal PR, Sogin ML, Herndl GJ (2011) Water mass-specificity of bacterial communities in the North Atlantic revealed by massively parallel sequencing. *Molecular Ecology*, **20**:258–274.

Amaral-Zettler L, Artigas LF, Baross J *et al.* (2010) A Global Census of Marine Microbes. In: *Life in the World's Oceans* (ed McIntyre AD), pp. 221–245. Wiley-Blackwell, Oxford, UK.

Amend AS, Oliver T a., Amaral-Zettler L a. *et al.* (2013) Macroecological patterns of marine bacteria on a global scale (J Lambshead, Ed,). *Journal of Biogeography*, **40**:800–811.

Anderson MJ, Crist TO, Chase JM *et al.* (2011) Navigating the multiple meanings of β diversity: a roadmap for the practicing ecologist. *Ecology Letters*, **14**:19–28.

Arrieta JM, Mayol E, Hansman RL *et al.* (2015) Dilution limits dissolved organic carbon utilization in the deep ocean. *Science*, **348**:331–333.

Azam F, Fenchel T, Field JG *et al.* (1983) The ecological role of water-column microbes in the sea. *Marine Ecology Progress Series*, **10**:257–264.

Blazewicz SJ, Barnard RL, Daly R.A, Firestone MK (2013) Evaluating rRNA as an indicator of microbial activity in environmental communities: limitations and uses. *The ISME Journal*, **7**:2061–8.

Brown M V, Schwalbach MS, Hewson I, Fuhrman JA (2005) Coupling 16S-ITS rDNA clone libraries and automated ribosomal intergenic spacer analysis to show marine microbial diversity: development and application to a time series. *Environmental Microbiology*, **7**:1466–79.

Campbell BJ, Kirchman DL (2012) Bacterial diversity, community structure and potential growth rates along an estuarine salinity gradient. *The ISME Journal*, **7**, 210–220.

Campbell BJ, Yu L, Heidelberg JF, Kirchman DL (2011) Activity of abundant and rare bacteria in a coastal ocean. *Proceedings of the National Academy of Sciences of the United States of America*, **108**:12776–12781.

Campbell BJ, Yu L, Straza TRA, Kirchman DL (2009) Temporal changes in bacterial rRNA and rRNA genes in Delaware (USA) coastal waters. *Aquatic Microbial Ecology*, **57**:123–135.

Catalá TS, Reche I, Fuentes-Lema A *et al.* (2015) Turnover time of fluorescent dissolved organic matter in the dark global ocean. *Nature Communications*, **6**:5986.

Cózar A, Echevarría F, González-Gordillo JI *et al.* (2014) Plastic debris in the open ocean. *Proceedings of the National Academy of Sciences USA*, **111**:10239–10244.

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**:541–52.

Cruaud P, Vigneron A, Lucchetti-Miganeh C *et al.* (2014) Influence of DNA extraction method, 16S rRNA targeted hypervariable regions, and sample origin on microbial diversity detected by 454 pyrosequencing in marine chemosynthetic ecosystems. *Applied and Environmental Microbiology*, **80**:4626–4639.

DeLong EF (1992) Archaea in coastal marine environments. *Proceedings of the National Academy of Sciences USA*, **89**:5685–5689.

DeLong EF, Franks DG, Alldredge AL (1993) Phylogenetic diversity of aggregate-attached vs. free-living marine bacterial assemblages. *Limnology and Oceanography*, **38**:924–934.

Denef VJ, Fujimoto M, Berry MA, Schmidt ML (2016) Seasonal succession leads to habitat-dependent differentiation in ribosomal RNA:DNA ratios among freshwater lake bacteria. *Frontiers in Microbiology*, **7**.

Duarte CM (2015) Seafaring in the 21St Century: The Malaspina 2010 Circumnavigation Expedition. *Limnology and Oceanography Bulletin*, **24**:11–14.

Ducklow HW, Steinberg DK, Buesseler KO (2001) Upper ocean carbon export and the biological pump. *Oceanography*, **14**:50–58.

Eloe EA, Shulse CN, Fadrosh DW *et al.* (2011) Compositional differences in particle-associated and free-living microbial assemblages from an extreme deep-ocean environment. *Environmental Microbiology Reports*, **3**:449–58.

Engelbrektson A, Kunin V, Wrighton KC *et al.* (2010) Experimental factors affecting PCR-based estimates of microbial species richness and evenness. *The ISME Journal*, **4**:642–7.

Fenchel T (2008) The microbial loop - 25 years later. *Journal of Experimental Marine Biology and Ecology*, **366**:99–103.

Field CB, Behrenfeld MJ, Randerson JT, Falkowski P (1998) Primary production of the biosphere: integrating terrestrial and oceanic components. *Science*, **281**:237–240.

Fisher MM, Triplett EW (1999) Automated approach for ribosomal intergenic spacer analysis of microbial diversity and its application to freshwater bacterial communities.

Applied and Environmental Microbiology , **65**:4630–4636.

Fuhrman JA, Hewson I, Schwalbach MS *et al.* (2006) Annually reoccurring bacterial communities are predictable from ocean conditions. *Proceedings of the National Academy of Sciences USA*, **103**:13104–9.

Fuhrman JA, McCallum K, Davis AA (1992) Novel major archaebacterial group from marine plankton. *Nature*, **356**:148–9.

Fuhrman J, Sleeter T, Carlson C, Proctor L (1989) Dominance of bacterial biomass in the Sargasso Sea and its ecological implications. *Marine Ecology Progress Series*, **57**:207–217.

Galand PE, Potvin M, Casamayor EO, Lovejoy C (2010) Hydrography shapes bacterial biogeography of the deep Arctic Ocean. *The ISME Journal*, **4**:564–76.

Ghiglione J-F, Conan P, Pujo-Pay M (2009) Diversity of total and active free-living vs. particle-attached bacteria in the euphotic zone of the NW Mediterranean Sea. *FEMS microbiology Letters*, **299**:9–21.

Ghiglione J-F, Galand PE, Pommier T *et al.* (2012) Pole-to-pole biogeography of surface and deep marine bacterial communities. *Proceedings of the National Academy of Sciences USA*, **109**:17633–8.

Gilbert JA, Field D, Swift P *et al.* (2009) The seasonal structure of microbial communities in the Western English Channel. *Environmental Microbiology*, **11**:3132–3139.

Gilbert JA, Steele JA, Caporaso JG *et al.* (2012) Defining seasonal marine microbial community dynamics. *The ISME Journal*, **6**:298–308.

del Giorgio PA, Duarte CM (2002) Respiration in the open ocean. *Nature*, **420**, 379–84.

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

Goodwin S, McPherson JD, McCombie WR (2016) Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, **17**:333–351.

Gotelli NJ, McGill BJ (2006) Null versus neutral models: what's the difference? *Ecography*, **29**:793–800.

Haas BJ, Gevers D, Earl AM *et al.* (2011) Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Research*, **21**:494–504.

Hagström Å, Pommier T, Rohwer F *et al.* (2002) Use of 16S ribosomal DNA for delineation of marine bacterioplankton species. *Applied and Environmental Microbiology*, **68**:3628–3633.

Hanson CA, Fuhrman JA, Horner-Devine MC, Martiny JBH (2012) Beyond biogeographic patterns: processes shaping the microbial landscape. *Nature Reviews Microbiol-*

ogy, **10**:1–10.

Hong S, Bunge J, Leslin C, Jeon S, Epstein SS (2009) Polymerase chain reaction primers miss half of rRNA microbial diversity. *The ISME Journal*, **3**:1365–73.

Hubbell SP (2008) *The Unified Neutral Theory of Biodiversity and Biogeography*.

Hugoni M, Taib N, Debroas D *et al.* (2013) Structure of the rare archaeal biosphere and seasonal dynamics of active ecotypes in surface coastal waters. *Proceedings of the National Academy of Sciences of the United States of America*, **110**:6004–9.

Hunt DE, Lin Y, Church MJ *et al.* (2013) Relationship between abundance and specific activity of bacterioplankton in open ocean surface waters. *Applied and Environmental Microbiology*, **79**:177–184.

Irigoien X, Klevjer TA, Røstad A *et al.* (2014) Large mesopelagic fishes biomass and trophic efficiency in the open ocean. *Nature Communications*, **5**:3271.

Jeraldo P, Sipos M, Chia N *et al.* (2012) Quantification of the relative roles of niche and neutral processes in structuring gastrointestinal microbiomes. *Proceedings of the National Academy of Sciences USA*, **109**:9692–8.

Jones SE, Lennon JT (2010) Dormancy contributes to the maintenance of microbial diversity. *Proceedings of the National Academy of Sciences USA*, **107**:5881–6.

Karl DM, Church MJ (2014) Microbial oceanography and the Hawaii Ocean Time-series programme. *Nature reviews. Microbiology*, **12**.

Karsenti E, Acinas SG, Bork P *et al.* (2011) A holistic approach to marine eco-systems biology. *PLoS Biology*, **9**:e1001177.

Kirchman DL (2008) *Microbial Ecology of the Oceans: Second Edition*. John Wiley and Sons.

Klindworth A, Pruesse E, Schweer T *et al.* (2013) Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Research*, **41**:e1.

Lekang K, Thompson E, Troedsson C (2015) A comparison of DNA extraction methods for biodiversity studies of eukaryotes in marine sediments. *Aquatic Microbial Ecology*, **75**:15–25.

Lima-Mendez G, Faust K, Henry N *et al.* (2015) Determinants of community structure in the global plankton interactome. *Science*, **348**.

Lindström ES, Langenheder S (2012) Local and regional factors influencing bacterial community assembly. *Environmental Microbiology Reports*, **4**:1–9.

Liu WT, Marsh TL, Cheng H, Forney LJ (1997) Characterization of microbial diversity by determining terminal restriction fragment length polymorphisms of genes encoding 16S rRNA. *Applied and environmental microbiology*, **63**:4516–22.

Lomolino M (2010) *Biogeography*. Sinauer Associates, Sunderland, Mass.

Magurran AE (2004) *Measuring Biological Diversity*. Blackwell Science.

Martiny JBH, Bohannan BJM, Brown JH *et al.* (2006) Microbial biogeography: putting microorganisms on the map. *Nature reviews. Microbiology*, **4**:102–12.

Muyzer G, de Waal EC, Uitterlinden AG (1993) Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA. *Applied and environmental microbiology*, **59**:695–700.

Nekola JC, White PS, Carolina N, Hill C (1999) The distance decay of similarity in biogeography and ecology, *Journal of Biogeography*, **26:**867–878.

Ofiteru ID, Lunn M, Curtis TP *et al.* (2010) Combined niche and neutral effects in a microbial wastewater treatment community. *Proceedings of the National Academy USA*, **107**:15345–50.

Pace NR, Stahl DA, Lane DJ, Olsen GJ (1986) The Analysis of Natural Microbial Populations by Ribosomal RNA Sequences. In: *Advances in Microbial Ecology* , pp. 1–55.

Pernice MC, Forn I, Gomes A *et al.* (2015) Global abundance of planktonic heterotrophic protists in the deep ocean. *The ISME Journal*, **9**:782–792.

Pesant S, Not F, Picheral M *et al.* (2015) Open science resources for the discovery and analysis of Tara Oceans data. *Scientific Data*, **2**:150023.

Pocheville A (2015) The Ecological Niche: History and Recent Controversies. In: *Handbook of Evolutionary Thinking in the Sciences* (eds Heams T, Huneman P, Lecointre G, Silberstein M), pp. 547–586. Springer Netherlands, Dordrecht.

Pommier T, Canbäck B, Riemann L *et al.* (2007) Global patterns of diversity and community structure in marine bacterioplankton. *Molecular Ecology*, **16**:867–880.

Pommier T, Neal P, Gasol JM *et al.* (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.

Rahmstorf S (2003) Thermohaline circulation: The current climate. *Nature*, **421**:699–699.

Ramette A, Tiedje JM (2007) Biogeography: An emerging cornerstone for understanding prokaryotic diversity, ecology, and evolution. *Microbial Ecology*, **53**:197–207.

Rillig MC, Antonovics J, Caruso T *et al.* (2015) Interchange of entire communities: microbial community coalescence. *Trends in Ecology & Evolution*, **30:**470-476.

Sloan WT, Lunn M, Woodcock S *et al.* (2006) Quantifying the roles of immigration and chance in shaping prokaryote community structure. *Environmental microbiology*, **8**:732–40.

Sloan WT, Woodcock S, Lunn M, Head IM, Curtis TP (2007) Modeling taxa-abun-

dance distributions in microbial communities using environmental sequence data. *Microbial Ecology*, **53**:443–55.

Smith MW, Zeigler Allen L, Allen AE, Herfort L, Simon HM (2013) Contrasting genomic properties of free-living and particle-attached microbial assemblages within a coastal ecosystem. *Frontiers in Microbiology*, **4**:1–20.

Sogin ML, Morrison HG, Huber JA *et al.* (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proceedings of the National Academy of Sciences USA*, **103**:12115–20.

Sul WJ, Oliver T a, Ducklow HW, Amaral-Zettler L a, Sogin ML (2013) Marine bacteria exhibit a bipolar distribution. *Proceedings of the National Academy of Sciences USA*, **110:**2342-2347.

Sunagawa S, Coelho LP, Chaffron S *et al.* (2015) Structure and function of the global ocean microbiome. *Science*, **348**:1261359.

Tuomisto H (2010) A diversity of beta diversities: Straightening up a concept gone awry. Part 2. Quantifying beta diversity and related phenomena. *Ecography*, **33**:23–45.

Turley C, Mackie P (1994) Biogeochemical significance of attached and free-living bacteria and the flux of particles in the NE Atlantic Ocean. *Marine Ecology Progress Series*, **115**:191–203.

Vellend M (2010) Conceptual Synthesis in Community Ecology. *The Quarterly Review of Biology*, **85**:183–206.

Venter JC (2004) Environmental Genome Shotgun Sequencing of the Sargasso Sea. *Science*, **304**:66–74.

Whitman WB, Coleman DC, Wiebe WJ (1998) Prokaryotes: The unseen majority. *Proceedings of the National Academy of Sciences USA*, **95**:6578–6583.

Whittaker RH (1960) Vegetation of the Sisiyou Mountains, Oregon and California. *Ecological Monographs*, **30**:279–338.

Woese CR, Fox GE (1977) Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proceedings of the National Academy of Sciences USA*, **74**:5088–5090.

Woodcock S, van der Gast CJ, Bell T *et al.* (2007) Neutral assembly of bacterial communities. *FEMS microbiology ecology*, **62**:171–80.

Zhang Y, Zhao Z, Dai M, Jiao N, Herndl GJ (2014) Drivers shaping the diversity and biogeography of total and active bacterial communities in the South China Sea. *Molecular Ecology*, **23**:2260–2274.

Zinger L, Amaral-Zettler L a., Fuhrman J a. *et al.* (2011) Global patterns of bacterial beta-diversity in seafloor and seawater Ecosystems (JA Gilbert, Ed,). *PLoS One*, **6**:e24570.

Global diversity and biogeography of deep-sea pelagic prokaryotes

Chapter I

Abstract:

The deep-sea is the largest biome of the biosphere, and contains more than half of the whole ocean's microbes. Uncovering their general patterns of diversity and community structure at a global scale remains a great challenge since only fragmentary information of deep-sea microbial diversity exists based on regional-scale studies. Here we report the first globally comprehensive survey of the prokaryotic communities inhabiting the bathypelagic ocean using high-throughput sequencing (HTS) of the 16S rRNA gene. This work identifies the dominant prokaryotes in the pelagic deep ocean and reveals that 50% of the operational taxonomic units (OTUs) belong to previously unknown prokaryotic taxa, most of which are rare and appear in just a few samples. We show that whereas the local richness of communities is comparable to that observed in previous regional studies, the global pool of prokaryotic taxa detected is modest (~3,600 OTUs), as a high proportion of OTUs are shared among samples. The water masses appear to act as clear drivers of the geographical distribution of both particle-attached and free-living prokaryotes. In addition, we show that the deep oceanic basins in which the bathypelagic realm is divided contain different particle-attached (but not free-living) microbial communities. The combination of the aging of the water masses and a lack of complete dispersal are identified as the main drivers for this biogeographical pattern. All together, we identify the potential of the deep ocean as a reservoir of still unknown biological diversity with a higher degree of spatial complexity than hitherto considered.

Introduction:

The pelagic dark ocean (the water column > 200 m deep) contains 70% of the ocean's microbial cells and 60% of its heterotrophic activity, with a pivotal regulatory role in planetary biogeochemical cycles (Arístegui *et al.*, 2009). Yet, current knowledge of the pelagic microbial community structure of the dark ocean, the largest biome in the biosphere, is based on a pool of samples collected at specific locations (DeLong *et al.*, 2006; Martín-Cuadrado *et al.*, 2007; Brown *et al.*, 2009; Galand *et al.*, 2010; Agogué *et al.*, 2011; Eloe *et al.*, 2011; Quaiser *et al.*, 2011; Smedile *et al.*, 2012; Wang *et al.*, 2013; Wilkins *et al.*, 2013; Ganesh *et al.*, 2014) (Fig. S1) and thus are dwarf in comparison with the analyses of upper ocean microbial communities which have indeed been assessed at global scales (Yooseph *et al.*, 2007; Rusch *et al.*, 2007; Zinger *et al.*, 2011; Sunagawa *et al.*, 2015). Whereas the deep ocean is often considered to be a rather uniform environment, the connectivity of pelagic microbial communities may be reduced by the limited mixing between water masses (Agogué *et al.*, 2011; Hamdan *et al.*, 2013) or modulated by advection (Wilkins *et al.*, 2013) imposing limitations on the dispersion of marine microbes in this low-turbulence environment. In addition, the spatial structure of the bathypelagic ocean, organized in partially isolated basins created by the emergence of submarine mountains, has not been tested as a potential factor affecting the biogeography of pelagic microbial communities, as happens for specialized deep-sea fauna (Moalic *et al.*, 2012) and bacteria inhabiting deep-sea surface sediments (Schauer *et al.*, 2010), either by imposing limits to deep-ocean connectivity or by delineating different environments that select for distinct microbial communities. Therefore, the deep pelagic ocean may present a mosaic of biogeographical domains with distinct microbial assemblages, a hypothesis not yet fully tested.

We created a global collection of samples retrieved during the Malaspina 2010 circumnavigation expedition (cf. Irigoien *et al.*, 2014) and we have used high-throughput sequencing of the 16S rRNA genes jointly with ARISA profiles and metagenomic data of the prokaryotes present in bathypelagic waters of the main world's oceans to describe their diversity, community structure and biogeographical distribution and identify the cosmopolitan and/or abundant prokaryotes in the dark ocean at a global scale. Moreover, we aimed to test whether deep-sea pelagic prokaryotic communities are uniform or present biogeographical patterns delineated by water mass and/or deep-oceanic basins.

Material and Methods:

A total of 60 water samples were taken during the Malaspina 2010 expedition (http://scientific.expedicionmalaspina.es/) corresponding to 30 different sampling stations glob-

ally distributed across the world's oceans (Fig. 1). We focused on the samples at the depth of 4,000 m, although a few samples were taken at lower depths, all well within the bathypelagic realm.

Sample collection and processing

Two different size fractions were analyzed in each station representing the free-living (0.2-0.8 μm) and particle-attached (0.8-20 μm) prokaryotic communities (Crump *et al.*, 1999; Ghiglione *et al.*, 2009; Allen *et al.*, 2012). For each sample 120 L of seawater were sequentially filtered through a 200 μm and a 20 μm mesh to remove large plankton. Further filtering was done by pumping water serially through 142 mm polycarbonate membrane filters of 0.8 μm (Merk Millipore, Isopore polycarbonate) and 0.2 μm (Merck Millipore, Express Plus) pore size with a peristaltic pump (Masterflex, EW-77410-10). The filters were then flash-frozen in liquid N_2 and stored at -80ºC until DNA extraction. The time span from bottle closing to filter freezing was ~ 4 h. and except for the time needed to empty the rosette bottles, the water was kept at 4°C. DNA extractions were performed using the standard phenol-chloroform protocol (see SI), and prokaryotic diversity was assessed by amplicon sequencing of the V4 region of the 16S rDNA with the Illumina MiSeq platform (iTags) using paired-end reads (2 x 250 bp) and primers F515/R806 (*details in SI*) targeting both Archaea and Bacteria (Caporaso *et al.*, 2011). Sequence data processing included the paired-end reads assembly, end-trimming, sequence quality control and chimera checking process integrated in the JGI pipeline. OTUs were obtained by clustering the processed data at a 97% identity and the taxonomic annotation of consensus sequences was performed using the SILVA v111 database (*details in SI*). ARISA and metagenomic data analyses were also applied to the same samples as an independent validation of the iTag approach (*see SI*).

Statistical data analyses

Statistical analyses (see details below) included richness estimation and rarefaction curves. The analysis of differences in community composition among samples and their relation to potential drivers assessed by means of a combination of multivariate exploratory techniques based on Bray-Curtis similarities (Non-Metric Multidimensional Scaling, NMDS) and hypothesis testing methods (PERMANOVA). The novelty of the obtained 16S rDNA sequences was checked against the SILVA, NCBI and RDP public databases using BLAST. All the sequences used in this study are publicly available at the NCBI Sequence Read Archive (SRA, http://www.ncbi.nlm.nih.gov/Traces/sra) under accession ID SRP031469. All statistical analyses and data treatment were conducted with the R Statistical Software (R Core Team, 2014) using version 3.0.1 and the following packages:

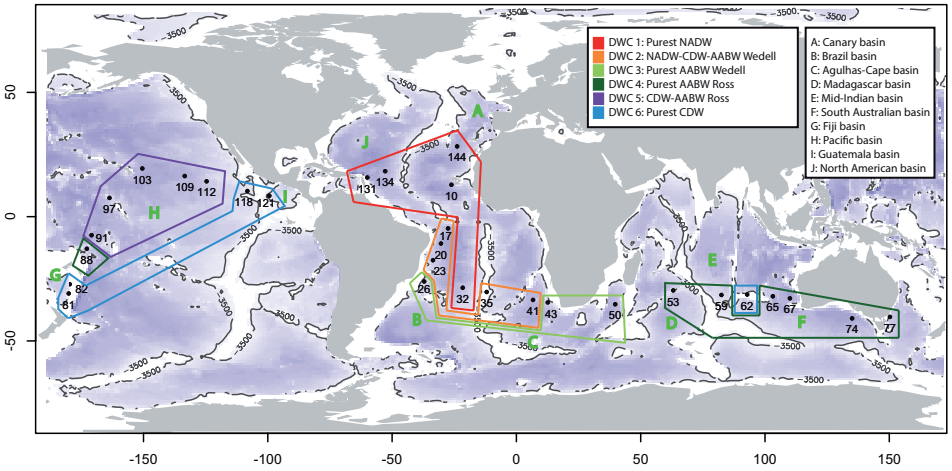


Figure 1: World map showing the location of the Malaspina sampling stations in the present study. The deep-water cluster derived from dominant water masses found at each station are color-coded, and the deep oceanic basins defined according to bathymetry below 3,500 m depth (see *Methods* for details) are indicated with letters.

BiodiversityR, ecodist, gdistance, marelac, and *vegan*. The iTags were used as the primary dataset for the whole study. ARISA and metagenomic data were used in specific analyses to compare with the iTag-derived data.

Novelty of the deep-ocean 16S rDNA sequences

In order to evaluate the novelty of the obtained 16S rDNA, the 3,507 representative OTU sequences were compared to RDP (Cole *et al.*, 2014), SILVA v111 (Pruesse *et al.*, 2007) and NCBI RefSeq (Pruitt *et al.*, 2012) public databases. The nucleotide subsets of the databases were downloaded (January 2013) and served as reference for a BLAST (Altschul *et al.*, 1990) comparison. A cutoff e-value of 1E-05 was used, a maximum of 5 target sequences were allowed for each query sequence and only the matches with coverage >90% were considered. When more than one match existed with an acceptable coverage and e-value, the one with the highest identity was chosen and identity values to the closest match for each sequence were collected.

Beta-diversity patterns of prokaryotic community composition

To infer the variation of the prokaryotic assemblages in space and along environmental gradients (i.e. beta-diversity), the Bray-Curtis dissimilarity index was used on community composition. The OTU-abundance table obtained from the sequence clustering was

sampled down to the lowest sampling effort (10,617 reads/sample), and dissimilarities between all pairs of samples were calculated using Bray-Curtis dissimilarity coefficient in order to obtain a beta-diversity matrix. The resulting dissimilarity matrix was used to perform an NMDS (Minchin, 1987) analysis using random starts. Permutational MANOVA (or PERMANOVA) using 1000 permutations was used to test for significant differences and to partition the beta-diversity matrix variance between groups of samples (Anderson, 2001; Anderson & Walsh, 2013).

The differential contribution ($D_{i,b}$) of a specific basin to the total abundance of a specific OTU was computed for the 30 most abundant OTU. This was calculated for each OTU and each basin following the formula:

$$D_{i,b} = (X_{i,b} - N_b) * 100$$

where $X_{i,b}$ is the contribution of the basin b to the total abundance of OTU i (i.e. the number of reads of OTU i in the samples belonging to the basin b divided by the total number of reads of OTU i) and N_b is proportion of samples in the dataset belonging to basin b. Thus an OTU with a percentage of reads coming from a specific basin higher/lower than would be expected under an even distribution across samples would have a positive/negative $D_{i,b}$ value.

Processes shaping prokaryotic biogeography

In order to infer the relative importance of the processes shaping the biogeographical patterns, the relative contribution of environmental drivers and geographical distance to the beta-diversity of deep ocean prokaryotic communities was assessed by means of a combined statistical strategy applied separately to both size-fractions. First, the best subset of environmental drivers was selected using the BIOENV approach (Clarke & Ainsworth, 1993). Secondly, permutation-based multiple regression on matrices (MRM) was used to partition the variance of the Bray-Curtis dissimilarity into (i) pure environmental variation, (ii) pure geographical variation, (iii) spatially structured environmental variation and (iv) and the unexplained variation (*see details in SI*).

Additionally, the scale of geographical variation was studied by means of Mantel correlograms (Oden & Sokal, 1986), that assesses the spatial correlation of multivariate data by computing a Mantel statistic (r) between the Bray-Curtis dissimilarity matrix and a matrix where pairs of sites belonging to the same geographic distance class receive value 0 and the other pairs, value 1. The process is repeated for each distance class and each r value can be tested for significance by permutation. Distance classes of 1,500 km were used.

Mantel correlograms were run for each size-fraction separately.

Dominant Phylum level analysis

In order to study the composition of prokaryotic communities at a broad taxonomic level a Phylum-abundance table was derived from the OTU-abundance table by adding up all the OTUs belonging to the same Phyla based on their SILVA taxonomy affiliation. For comparison with similar studies the Phylum Proteobacteria was divided into its Classes. OTUs that could not be assigned to any Phyla were included into an extra category (named as Others). Only the Phyla represented by more than a 0.5% of the reads in the whole dataset were considered. Differences in abundance for every Phylum between Oceans (categorized as North Atlantic, South Atlantic, Indian, South Pacific and North Pacific), deep oceanic basins and "deep-water clusters" – see SI– were statistically tested using ANOVA. P-values were then Bonferroni-corrected for the

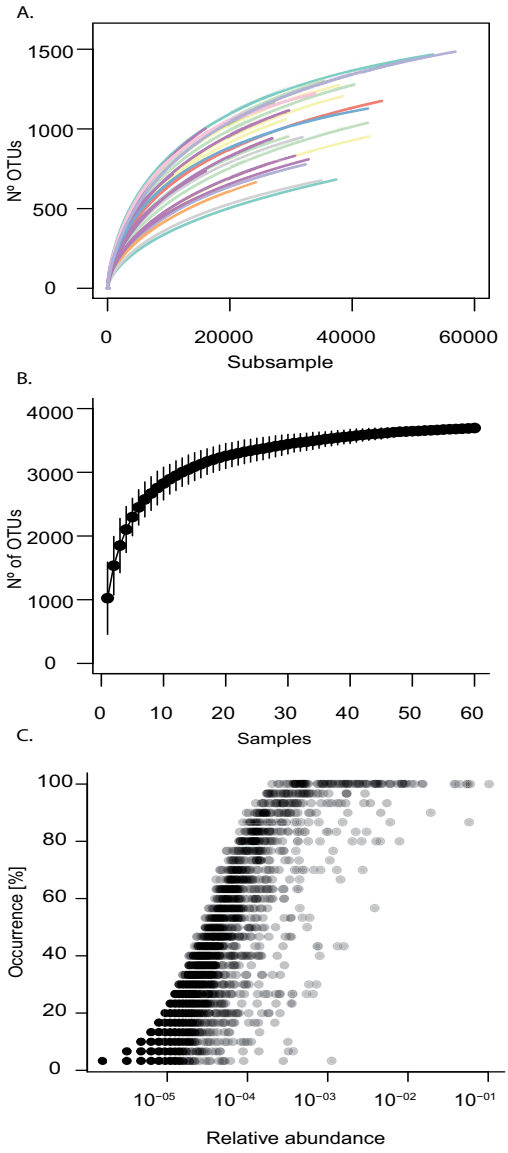


Figure 2 Rarefaction curves: (A) within samples, individual-based and (B) sample based. Global dataset relative abundance vs. occurrence (i.e. the percentage of samples in which an OTU occurs) for all the OTUs (C). The sample based rarefaction curve has been calculated for the entire dataset. The deep oceanic basins to which each station belongs are indicated with different colors in A. (legend in Fig. 5). No significant differences were detected for richness/diversity (neither OTU number, Chao1 nor Shannon index) between basins.

effect of multiple comparisons.

Results and Discussion:

We examined pelagic prokaryotic diversity from two size fractions considered free-living and particle-attached (see *Methods*), in 30 deep-ocean sites distributed in the North Atlantic (4 sites), South Atlantic (8), Indian (6) and South (5) and North Pacific Oceans (6), and an additional set of samples which was taken from the Southern Ocean in waters close to Australia (Fig. 1). We targeted 4,000 meters as the water depth of study taken as representative of the bathypelagic ocean, yet some samples were taken from shallower depths (always >2,000m). The sites were assigned to each of six different deep-water clusters according to their water mass composition (Fig. S2 and Table S1), as well as to "deep oceanic basins" based on the global ocean's bathymetry (Fig. 1; details in SI). Using Illumina sequencing of the V4 region of the 16S rRNA gene, we obtained a final dataset of 1,789,427 sequences (iTags) that could be constrained into a total of 3,695 operational taxonomic units (OTUs), which represents a minimum estimate of the richness of prokaryote taxonomic units in the deep-ocean (Table S2). The iTag data was compared to ARISA-derived and metagenomic data revealing a good consistency between techniques (see SI and Fig. S3 and S4). Hereafter, the downstream analyses were performed using iTags as the default dataset.

Richness of bathypelagic prokaryotic communities

Two kinds of rarefaction curves were computed in order to check whether prokaryotic richness was close or far from saturation, both at the local (individual site/sample) and at the global (all samples) scale. For each sample a rarefaction curve (or individual-sample-based rarefaction curves) was drawn by sequentially computing the number of OTUs for an increasing number of reads. Additionally, a sample-based rarefaction curve was drawn by randomly accumulating an increasing number of samples for the whole dataset. Rarefaction curves for individual samples showed that prokaryotic diversity at the OTU level (97% identity cutoff) was far from saturation locally with the sequencing effort used (Fig. 2a). In contrast, when considering the global set of samples, the sample-based rarefaction curves reached a considerably flat plateau at around 3,500 OTUs (Fig. 2b). The number of OTUs increased rapidly with the addition of the first 10 samples but once ~20 samples were considered, the addition of extra samples resulted in a small additional discovery of new OTUs. In fact, on average 42.0% of the OTUs present in one sample were shared with a second one taken at random from our data set (min=15.7%, max=76.2%), being these shared OTUs the ones with higher abundances (Fig. 2c). This indicates that the global

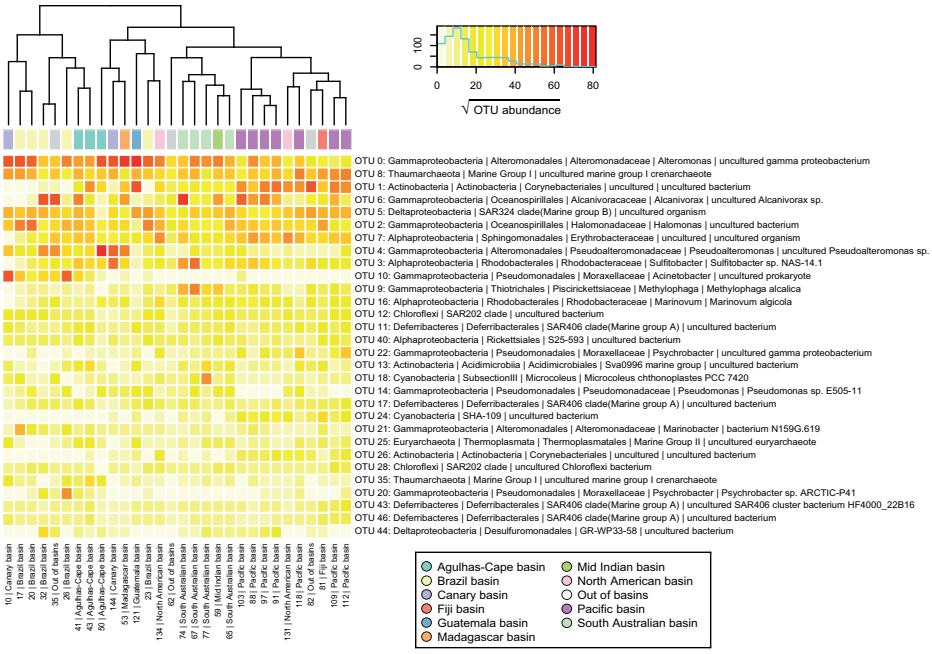


Figure 3: Heatmap representing the square root of abundances (number of reads) of the 30 most abundant OTUs (rows) along the 30 stations (columns). Subsampled abundances to the minimum sequencing depth (10,617 reads/sample) have been used for comparison and data from the two size-fractions within a station was summed after subsampling. The deep oceanic basins to which each station belongs are indicated at the top (see color legend). Taxonomical annotation for each OTU is based on the SILVA taxonomic assignment of each OTU representative sequence. OTUs are ordered top to bottom based on their global abundance in the whole dataset.

deep ocean contains a relatively modest number of prokaryotic phylotypes, likely in the order of a few thousands. The total number of OTUs identified here, 3,695, represents, however, a minimum estimate, as strict data cleaning criteria have been used in the data processing (specially, the removal of possible chimeric sequences and singletons) and as additional OTUs are likely to be present in areas not sampled in this study, such as the Arctic or Antarctica.

Each deep-sea prokaryotic community sampled here can be thus considered to be composed of i) a set of dominant species shared with the rest of the stations in varying proportions, which we estimate at about 42 % of the OTUs identified, and ii) a set of low abundant and relatively sample-specific (i.e. highly unshared) set of taxa comprising a "rare biosphere" (Sogin *et al.*, 2006; Pedrós-Alió, 2012) of the global pool of deep-sea prokaryotes.

In order to prevent artifacts during diversity/richness estimations due to uneven sampling efforts among samples, the dataset was randomly sampled down to the lowest sequencing effort (10,617 reads/sample). This resulted in 637,020 reads corresponding to a total of 3,543 OTUs. The number of OTUs in each community ranged from 248 to 896 (mean=659.1, sd=146.0), comparable to the mean local richness reported before in the Atlantic Ocean (mean=835, sd=421; Agogué *et al.*, 2011). Slightly higher values (mean=1,037.3, sd=173.6) had previously been reported in bathypelagic samples from polar and mid-latitudes (see samples below 1,000 m from Supplementary Table 1 in Ghiglione *et al.*, 2012). Thus, the global assessment of bathypelagic prokaryotic communities that we report delivers local richness values comparable to those observed in previous regional-scale bathypelagic surveys. However, our estimate of the total prokaryotic richness in the bathypelagic ocean estimated at around 3,600 OTUs, is consistent with previous estimates which found a total richness of 10,846 OTUs with half of them corresponding to singletons (Zinger *et al.*, 2011).. This represents a small fraction (~ 3% and 5.5%) of the total oceanic plankton bacterial richness found by recent surveys with comparable methodologies: A previous study combining 509 benthic and pelagic marine samples ranging from 0 to 5,400 m depth found a total richness of ~ 120,000 OTUs (Zinger *et al.*, 2011) while a total richness of ~ 65,500 OTUs was detected in a different study using data from 277 epipelagic samples (243 of which were also included in the previous one) from the Arctic, Atlantic, Pacific, and Southern Oceans (Sul *et al.*, 2013). This would suggest that only a small fraction of all oceanic microbes are found in the deep ocean.

Novelty of bathypelagic prokaryotic lineages

We assessed the degree of novelty of bathypelagic prokaryotic diversity by comparing the detected 16S rDNA sequences to those present in public databases. OTU representative sequences were compared to the RDP, SILVA and NCBI databases using 95% and 97% / 99% identity values as proxies for genus and "species" level, respectively. The three databases provided comparable identity distributions with two clear peaks, the first one at around 95% identity and a second peak near the 100% identity (Fig. S5). Interestingly, around one third of the OTUs had identity values lower than 95%, half of the OTUs lower than 97% and 2/3 of the OTUs had values lower than 99%. Although the 97% identity is widely used in microbial ecology studies as a broad proxy for "species" cut-off (Stackebrandt & Goebel, 1994; Hagström *et al.*, 2002; Cohan, 2002), it is well known that this value may integrate different species and overlook putative ecotypes within species with different ecological roles (Fox *et al.*, 1992; Acinas *et al.*, 2004; Stackebrandt, 2006). Therefore, it is safe to assume that we detected at least between 1,687 (at 97%) and 2,385

(at 99%) putative new prokaryotic OTUs as well as 986 OTUs belonging to putative new genera (at 95%) not present in the standard prokaryotic rDNA databases (corresponding to 45.7%, 64.6% and a 26.7% of the total OTUs, respectively). However, these novel lineages represented a minor fraction of the reads (4.5%, 9.1% and 2.2% respectively) and thus they are likely members of the bathypelagic "rare biosphere". This pattern had already been observed in a single sample from a hydrothermal vent (Sogin *et al.*, 2006) where novel sequences belonged to very low abundant OTUs. This result suggests that the prokaryotic assemblages in the bathypelagic ocean are composed of a combination of a set of relatively abundant and widely distributed species, already detected in previous environmental surveys, and a set of rare species with limited distributions where most of the genetic novelty accumulates.

Dominant prokaryotes in the bathypelagic ocean

We aimed at identifying the most abundant prokaryotes present in the deep ocean at a global scale. The relative abundance (i.e. proportion of reads) of every Phylum (except Proteobacteria, which were divided into Classes) was highly similar among samples (Fig. S6): Gammaproteobacteria, Alphaproteobacteria, Actinobacteria, Thaumarchaeota and Deltaproteobacteria dominated in all the stations in terms of relative abundance. Gammaproteobacteria was the most abundant group in all the sites, consistent with the previously described increase of their contribution to the total number of bacteria with depth (López-García *et al.*, 2001). Here, the proportion of members of Archaea ranged from 2.2% to 16.3% of the total reads in both fractions combined. This estimation is considerably lower than previous studies in which Archaea had been reported to be between 20-30% of the total of bacterioplankton in the deep ocean (DeLong, 1992, 2003; Massana *et al.*, 1997) or even higher (39%) (Karner *et al.*, 2001). Yet, Archaea in the free-living samples reached up to 25-30% of the total in specific locations (stations 10, 81, 112, 118 and 121 located in the North Atlantic and North Pacific). Our findings are not attributable to PCR biases or primer mismatches as relative abundances of both Euryarchaeota and Thaumarchaeota obtained from the metagenomic dataset were similar (Fig. S4). In all samples, Thaumarchaeota dominating over Euryarchaeota (1.7 – 14% vs. 0.29 – 3.7%) as described before for bathypelagic waters (Herndl *et al.*, 2005; Teira *et al.*, 2006). Only the Actinobacteria phylum had differences in relative abundance between oceans and between deep oceanic basins (as tested by ANOVA with bonferroni correction: $F = 5.8$, $P = 0.001$, $P_{corrected} = 0.016$ and $F = 4.9$, $P = 0.003$, $P_{corrected} = 0.045$ for Ocean and Basins, respectively) being more abundant in the North and South Pacific and in particular the Pacific and Guatemala Basins.

Despite the invariant composition of prokaryotic communities at a the Phylum level, the distribution of the 30 most abundant OTUs (Fig. 3, Table S3) included only a few cosmopolitan organisms that were relatively evenly distributed along the whole dataset: e.g. the first, second and fifth most abundant OTUs, representatives of the *Alteromonas* genus, the Marine Group I Thaumarchaeota (MGI) and the SAR324 clade, respectively. This is consistent with current knowledge on the ecology of these three groups: the existence of a deep *Alteromonas macleodii* ecotype (identical at 16S rDNA sequence level to our most abundant OTU) with specific adaptations to deep ocean conditions is well known (López-López *et al.*, 2005; Ivars-Martinez *et al.*, 2008) and the MGI archaeal group, jointly with Marine Group II (MGII) Euryarchaeota, are the most abundant Archaea in the ocean (Massana *et al.*, 2000). The SAR324 Deltaproteobacteria clade has also been described as a typical deep-sea group (López-García *et al.*, 2001; Agogué *et al.*, 2011). Most of the rest of the dominating OTUs exhibited uneven abundances throughout the world's deep oceans, with most of them restricted to a specific geographical region: e.g. *Alcanivorax* sp. and an uncultured Actinobacteria representatives were nearly absent from the Atlantic while a *Pseudoalteromonas* sp. representative was nearly absent from the Pacific but abundant in the rest of the sites. As a result of the heterogeneity in the distribution of the most abundant organisms, the samples tended to cluster with other geographically close samples (Fig. 3).

Differential OTU distribution through deep oceanic basins

In addition, we calculated the differential contribution (in %) of a specific basin to the abundance of a specific OTU for the 30 most abundant OTUs (Fig. 4). Within these, only a few did not have a clear differential contribution associated to a specific basin and thus, and as mentioned before (see Fig. 3), these OTUs were equally distributed among basins (i.e. were cosmopolitan): e.g. members of the SAR406 clade and a representative of MGI and SAR324 clade. For the rest of the 30 most abundant OTUs, some were consistently over-represented or under-represented in each deep-ocean basin indicating that at least a fraction of the community exhibited an uneven distribution across basins. Representatives of the *Acinetobacter* and *Pseudoalteromonas* genus and MGI Thaumarchaeota were over-represented in the Canary basin and under-represented in the Pacific basin. Despite their proximity, the Brazil basin was characterized by a different combination of over-represented OTUs: the same representative of the *Acinetobacter* genus was over-represented in this basin but in combination with two OTUs assigned to the *Marinobacter* and *Psychrobacter* genus. A different OTU also assigned to the *Psychrobacter* genus was under-represented in this basin but was especially abundant in the Pacific basin, together

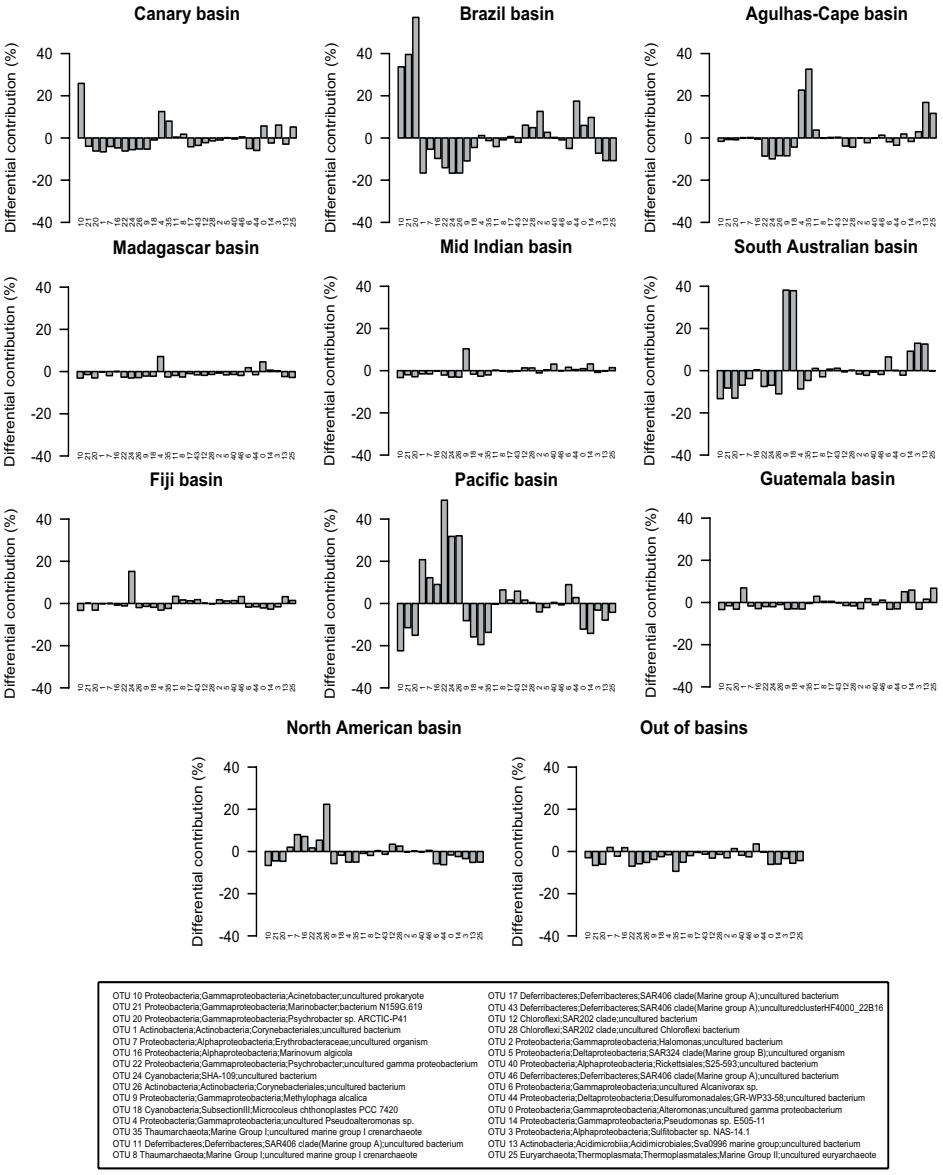


Figure 4 Differential contribution ($D_{i,b}$; in %) of each basin to the total abundance of each of the 30 most abundant OTUs (see SI for calculation details). Numbers below each bar represent each OTU, whose taxonomical affiliation is described in the legend, based on SILVA taxonomy. OTUs are the same as in Fig. 3 but ordered using a clustering based on $D_{i,b}$ values (details not shown) for a clearer visualization.

with two OTUs assigned to the Corynebacteriales order. The South Australian basin was characterized by the overrepresentation of two OTUs assigned to the *Methylophaga* and *Microcoleus* genus while the Agulhas-Cape basin had above-average contributions of the two OTUs assigned to *Pseudoalteromonas* genus and MGI that were abundant in the Canary basin but without the co-presence of *Acinetobacter*. Although the differential contribution was computed correcting for the different number of samples in each basin (*details in the SI*), the deviation from an even distribution was higher for the basins with a higher number of samples (Brazil, South Australian and Pacific basin). Future studies with a higher spatial detail and sampling size within each basin would allow to define these basins in terms of prokaryotic community composition and to describe, if they exist, indicator OTUs or clades for the distinct basins.

Beta-diversity patterns of bathypelagic prokaryotic communities

NMDS was applied in order to represent the Bray-Curtis dissimilarities (i.e. beta-diversity patterns) of the 60 samples (Fig. 5) based on the relative abundance of all the OTUs. The samples belonging to different size-fractions were clearly separated along the first axis. Detailed analysis of these differences is in the process (*Salazar et al.* Chapter 2). Particle-attached samples within a deep oceanic basin tended to have similar community composition, and thus clustered together in the NMDS. The seven samples located in the Pacific basin formed a tight cluster together with stations 81 (Fiji basin) and 88 (located also at the Pacific Ocean but at 2,150 m depth and thus out of the basins defined below 3,500 m). Samples belonging to the Brazil basin also clustered together and close to the samples from the Canary basin, both in the Atlantic Ocean. In contrast, the samples from stations 131 and 134 located also in the Atlantic Ocean and belonging to the same deep-water cluster but in a different basin, the North American basin, were more similar to the Pacific group. A third group of samples was composed by the stations situated in the Indian Ocean, in the South Australia basin, Madagascar basin and Mid Indian basin. This geographical ordering of the samples was not as evident for the free-living group of samples (Fig. 5 and Fig. 1 for sample location).

A deep-ocean study has recently emphasized the role of distinct deep-sea water masses as potential bio-oceanographical islands for prokaryotic communities (Agogué *et al.*, 2011). Additionally, physical transport processes such as advection, have been proved to act as ecological drivers of marine bacterial communities (Wilkins *et al.*, 2013) but the effect of the deep ocean's floor morphology over the composition of microbial communities was only explored in few locations such as the Walvis Ridge or the Challanger Deep (Schauer *et al.*, 2010; Nunoura *et al.*, 2015). Here we test, at a global scale, the possibility

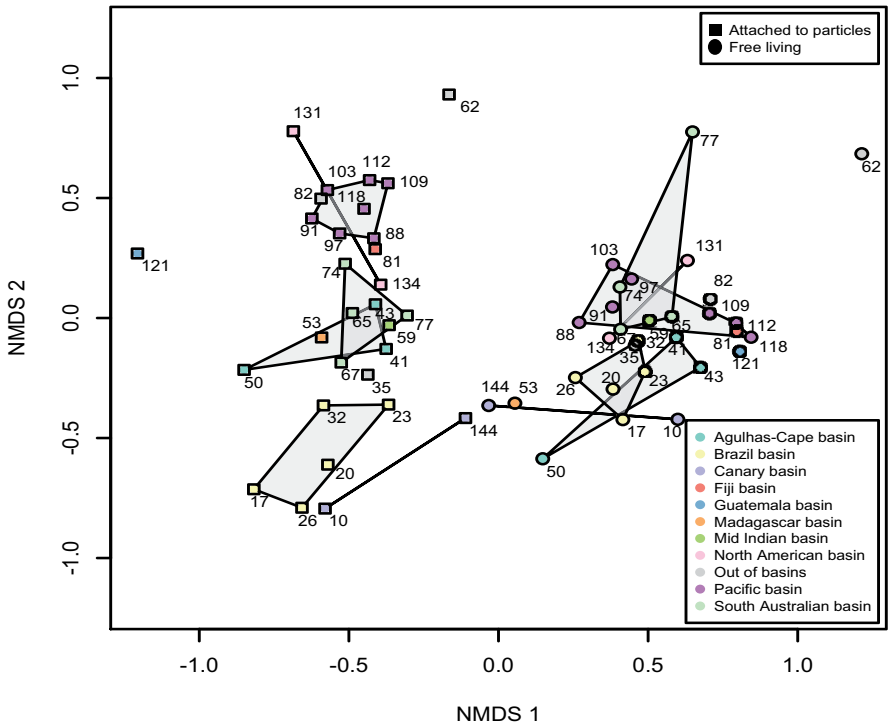


Figure 5: Non Metric Multidimensional Scaling (NMDS) analysis of beta-diversity (Bray-Curtis distances) for the 60 samples in the dataset based on iTags. Size-fraction is coded with point style (squares, attached and circles, free-living) and deep oceanic basins following color codes (see legends). Numbers close to each sample represent the station number (see Fig. 1).

that submarine mountains that divide the deep ocean into deep-oceanic basins may act as "ecological barriers" for prokaryotic communities: either by (i) imposing a reduced dispersion between basins or (ii) by compartmentalizing the bathypelagic ocean into contrasting environments that exert an ecological selection on the prokaryotes that inhabit it. Additionally, we tested the relative role of water masses and deep oceanic basins in shaping the biogeography of bathypelagic prokaryotes by examining the significance, tested using Permutational MANOVA, of the sample's clustering in accordance to the size-fraction, the deep-water cluster and deep oceanic basins they belonged to (Table S4). There were significant differences in the three factors (all $P<0.0001$) and they together explained almost 70% of the variance (size-fraction, $R^2=0.31$; deep-water cluster, $R^2=0.18$; deep oceanic basins, $R^2=0.18$). None of the interactions between factors was significant. As most

of the variance was explained by the size-fraction, we split the analyses by size-fraction to further test the deep-water cluster and basin as explanatory variables. Both factors were significant for the particle-attached fraction, and differences in oceanic basin origin explained 35% of the variance, even when taking the effect of deep-water clusters into account. For the free-living fraction there were no significant differences between oceanic basins once the deep-water clusters were considered. The date of sampling was included in the analyses to take into account seasonal differences as a possible confounding factor, as it has been shown that dark ocean prokaryotic communities can be as dynamic as those of the surface ocean (Winter *et al.*, 2009) where the seasonal patterns are extremely relevant (Brown *et al.*, 2005; Fuhrman *et al.*, 2006; Gilbert *et al.*, 2009; Giovannoni, 2012; Gilbert *et al.*, 2012). In all cases the date of sampling appeared to be a significant factor but its inclusion in the analyses did not modify the variance explained by the other factors (Table S5). As the particulate matter in which particle-attached communities develop may ultimately come from the surface ocean through sinking, the Longhust provinces were also tested as a potential factor structuring the betadiversity of free-living and particle attached communities. In none of the two cases the grouping of the stations in the corresponding Longhust provinces was significant (details not shown). Thus, in summary, particle-attached prokaryotic communities exhibited a significant basin-specific composition while this basin-specificity was not observed for free-living prokaryotes. Consequently, different processes need to be structuring the particle-attached and free-living beta-diversity and thus generating differential biogeographical patterns.

Ecological processes shaping the biogeography of deep-ocean prokaryotic communities

The bathypelagic free-living and particle-attached prokaryotic communities clustered according to the water masses while only particle-attached communities exhibited a significant basin-specificity. However, the biogeographical patterns observed can arise as the result of different ecological processes well established within a theoretical framework (Hanson *et al.*, 2012): a) the existence of environmental differences between basins or water masses that exert a differential selection of prokaryotes, i.e. "environmental selection" or b) a reduced dispersal of microbes between basins or water masses, i.e. "historical effects". We estimated the relative contribution of both processes by relating community composition to a set of environmental variables, and to the geographical distance between sampling locations, using multiple regression on matrices (MRM; *see SI for details and a further explanation of the theoretical framework*). The MRM analysis explained a total of 23.1% (particle-attached fraction) and 10.7% (free-living fraction) of the total vari-

ance in the Bray-Curtis dissimilarity matrix. For the free-living fraction, only the effect of the environmental variables on community composition was statistically significant (P=0.005; explaining 6.4% of the variance) and the effect of the geographical distance was not (P>0.1), indicating that dispersal limitation and historical processes are not relevant in shaping the biogeography of free-living prokaryotes. However, dispersal limitation seemed to play a minor, yet significant, role for the particle-attached communities, as the pure effect of the geographical distance between locations explained a small but significant fraction of the variance (P<0.005; explaining a 5% of the variance). In addition, most of the variance was explained by the effect of the environmental variables (P<0.005), i.e. the pure and the spatially structure environmental variation. These accounted for a 6.3% and 11.8% of the variance, respectively. Thus, both environmental selection and historical effects appear to shape the biogeography of particle-attached communities, although spurious distance effects may arise as a result of unmeasured environmental variables (Hanson *et al.*, 2012).

Table 1. Environmental drivers of free-living and particle-attached communities.

Environmental variable	**Particle-attached**		**Free-living**	
	Mantel r	P-value	Mantel r	P-value
Depth	0.145	0.11	0.261	0.06
Salinity	0.332	<0.01*	0.060	0.24
Potential Temperature (P.Temp)	0.156	0.01*	0.124	0.07
Apparent Oxygen Utilization (AOU)	0.426	<0.01*	0.077	0.17
Prokaryotic heterotrophic activity	-0.156	0.93	-0.170	0.94
Prokaryote Abundance	-0.156	0.90	-0.145	0.86
Percentage of HNA-content prokaryotes	-0.100	0.78	-0.077	0.67
Prok. Biomass Duplication Time (DT)	0.209	0.01*	0.100	0.15
Prokaryote Abundance (0-200m)	-0.145	0.95	-0.059	0.67
Best BIOENV Model	0.427 (AOU + DT)	<0.01*	0.280 (Depth + P.Temp)	0.02*

Abbreviation: HNA, high nucleic acid. Mantel correlation (using Pearson correlation) between the Bray–Curtis dissimilarity and the Euclidean distance for particle-attached and free-living prokaryotes of the environmental variables used in the BIOENV approach (see Supplementary Information). Single variables and the best BIOENV model for each fraction are tested. Significant P-values (0.05) are indicated with an asterisk.

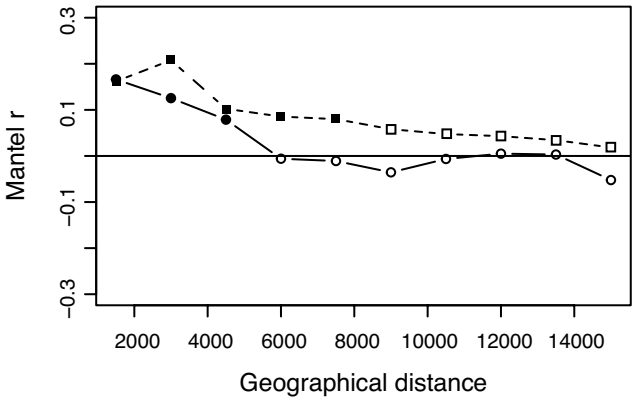


Figure 6: Mantel correlogram for particle-attached (squares) and free-living (circles) prokaryotic communities testing the autocorrelation on community composition by performing sequential Mantel tests between the Bray-Curtis dissimilarities and the grouping of samples using geographical distance classes set at 1,500 m. Filled points represent significant correlations after Bonferroni correction. Mantel correlograms were run up to a maximal distance of 15,000 km.

The best subset of environmental drivers for the free-living fraction was temperature and depth of the sampling location (Table 1), pointing to a pure environmental selection process exerted by these two drivers for these communities. Temperature has also been found to be the main environmental driver for upper ocean microbial communities explaining the spatial variation within the epipelagic ocean (Sunagawa *et al.*, 2015). In contrast, the best subset of environmental drivers for the particle-attached fraction were the Apparent Oxygen Utilization (AOU) and the prokaryotic biomass duplication time (DT), although most of the effect was solely due to the AOU (Table 1). The AOU (i.e. the difference between the saturation and measured dissolved oxygen) indicates the modification of oxygen concentration through the mixing of water masses and various biogeochemical processes and correlates with the aging of a water mass (Jenkins, 1982). Additionally, in this case the AOU is clearly reflecting the deep-water clusters each sample belongs to (Fig. S2b). Thus, the fact that the AOU of the samples where particle-attached communities are found is the best explanatory variable suggest that water mass mixing and aging play an important role in the assembly of particle-attached bathypelagic communities.

Finally, the scale of geographical variation for the two size-fractions was studied using Mantel correlograms. We tested how far in space the samples maintain a significant autocorrelation in community composition. For particle-attached communities, there was a significant spatial autocorrelation, which expanded until 7,500 km (Fig. 6a). These

relatively short distances, considering the global scale of the dataset (i.e. the ship covered ~ 45,700 km), are consistent with the basin-specificity of particle-attached prokaryotic community composition described above (Fig 5). In fact, the mean and maximal distance between all the samples belonging to the same deep-oceanic basin is 4,950 km and 9,800 km respectively. These distances are also coherent with the only study with a similar approach, which explored the effect of the Walvis Ridge on the bacterial communities in the deep-sea sediments at the Guinea, Angola and Cape basins, reporting an effect of the geographical distance on community composition detectable at distances >3,000 km (Schauer *et al.*, 2010). Although the effect of the geographical distance on community composition was not significant for the free-living communities once the effect of the environmental drivers is considered, there was a significant autocorrelation when tested at short distances, which expanded until 4,500 km (Fig. 6b). This significant autocorrelation found at short distances for the free-living communities does not correspond to the basin organization of the deep ocean, as tested before (Table S4 and S5), and may be due to the effect of potentially relevant environmental variables structuring the free-living bathypelagic communities at shorter scales which were not measured in this study.

Thus, in summary, it seems that although both, the free-living and the particle-attached prokaryotic communities, exhibited autocorrelation at short distances and differ between water masses, they appeared to be structured by contrasting processes and drivers. The free-living prokaryotic communities appears to respond to an environmental selection process exerted by temperature and depth variations, although a high proportion of the variance remains unexplained (89.3%), as in similar studies (Hanson *et al.*, 2012). In contrast, the particle-attached communities appear to respond to a more complex set of processes where the ageing and global circulation of the water masses and some degree of dispersal limitation create basin-specific communities not evident for the free-living fraction. This could be an indication that at least a fraction of the deep oceanic particles where the prokaryotes are associated, instead of coming from the surface ocean through sinking, correspond to presumably buoyant or slow-sinking particles that are produced autochthonously at depth, as it has recently been hypothesized (Herndl & Reinthaler, 2013). This hypothesis would explain why particle-attached prokaryotes reflect the deep-water mass circulation and why a signal of reduced dispersal between basins is found only for particle-attached, and not for free-living communities.

References

Acinas SG, Klepac-Ceraj V, Hunt DE, Pharino C, Ceraj I, Distel DL, *et al.* (2004). Fine-scale phylogenetic architecture of a complex bacterial community. *Nature,* **430**:551–4.

Agogué H, Lamy D, Neal PR, Sogin ML, Herndl GJ. (2011). Water mass-specificity of bacterial communities in the North Atlantic revealed by massively parallel sequencing. *Mol Ecol,* **20**:258–74.

Allen LZ, Allen EE, Badger JH, McCrow JP, Paulsen IT, Elbourne LD, *et al.* (2012). Influence of nutrients and currents on the genomic composition of microbes across an upwelling mosaic. *ISME Journal,* **6**:1403–1414.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. (1990). Basic local alignment search tool. *Journal of Molecular Biology,* **215**:403–10.

Anderson MJ. (2001). A new method for non-parametric multivariate analysis of variance. *Austral Ecology,* **26**:32–46.

Anderson MJ, Walsh DCI. (2013). PERMANOVA, ANOSIM, and the Mantel test in the face of heterogeneous dispersions: What null hypothesis are you testing? *Ecological Monographs,* **83**:557–574.

Arístegui J, Gasol JM, Duarte CM, Herndl GJ. (2009). Microbial oceanography of the dark ocean's pelagic realm. *Limnology adn Oceanography,* **54**:1501–1529.

Brown M V, Philip GK, Bunge JA, Smith MC, Bissett A, Lauro FM, *et al.* (2009). Microbial community structure in the North Pacific ocean. *ISME Journal,* **3**:1374–86.

Brown M V, Schwalbach MS, Hewson I, Fuhrman J a. (2005). Coupling 16S-ITS rDNA clone libraries and automated ribosomal intergenic spacer analysis to show marine microbial diversity: development and application to a time series. *Environmental Microbiology,* **7**:1466–79.

Caporaso JG, Lauber CL, Walters W a, Berg-Lyons D, Lozupone C a, Turnbaugh PJ, *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 USA,* **108**:4516–22.

Clarke K, Ainsworth M. (1993). A method of linking multivariate community structure to environmental variables. *Marine Ecology Progress Series,* **92**:205–219.

Cohan FM. (2002). What are bacterial species? *Annual Reviews of Microbiology,* **56**:457–87.

Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, *et al.* (2014). Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Research,* **42**:D633–42.

Crump BC, Armbrust E V, Baross JA. (1999). Phylogenetic analysis of particle-attached

and free-living bacterial communities in the Columbia river, its estuary, and the adjacent coastal ocean. *Applied and Environmental Microbiology,* **65**:3192–204.

DeLong EF. (1992). Archaea in coastal marine environments. *Proceedings of the National Academy of Sciences USA,* **89**:5685–5689.

DeLong EF. (2003). Oceans of archaea. *ASM News-American Soc Microbiol* **69**:503.

DeLong EF, Preston CM, Mincer T, Rich V, Hallam SJ, Frigaard N-U, *et al.* (2006). Community genomics among stratified microbial assemblages in the ocean's interior. *Science,* **311**:496–503.

Eloe EA, Shulse CN, Fadrosh DW, Williamson SJ, Allen EE, Bartlett DH, *et al.* (2011). Compositional differences in particle-associated and free-living microbial assemblages from an extreme deep-ocean environment. *Environmental Microbiology Reports,* **3**:449–58.

Fox GE, Wisotzkey JD, Jurtshuk P. (1992). How Close Is Close: 16S rRNA Sequence Identity May Not Be Sufficient To Guarantee Species Identity. *International Journal of Systematic Bacteriology,* **42**:166–170.

Fuhrman JA, Hewson I, Schwalbach MS, Steele J a, Brown M V, Naeem S. (2006). Annually reoccurring bacterial communities are predictable from ocean conditions. *Proceedings of the National Academy of Sciences USA,* **103**:13104–9.

Galand PE, Potvin M, Casamayor EO, Lovejoy C. (2010). Hydrography shapes bacterial biogeography of the deep Arctic Ocean. *ISME Journal,* **4**:564–76.

Ganesh S, Parris DJ, DeLong EF, Stewart FJ. (2014). Metagenomic analysis of size-fractionated picoplankton in a marine oxygen minimum zone. *ISME Journal,* **8**:187–211.

Ghiglione J-F, Conan P, Pujo-Pay M. (2009). Diversity of total and active free-living vs. particle-attached bacteria in the euphotic zone of the NW Mediterranean Sea. *FEMS Microbiol ogy Letters,* **299**:9–21.

Ghiglione JF, Galand PE, Pommier T, Pedrós-Alió C, Maas EW, Bakker K, *et al.* (2012). Pole-to-pole biogeography of surface and deep marine bacterial communities. *Proceedings of the National Academy of Sciences USA,* **109**:17633–8.

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

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

Giovannoni SJ. (2012). Seasonality in Ocean Microbial Communities. **671**.

Hagström Å, Pommier T, Rohwer F, Simu K, Stolte W, Svensson D, *et al.* (2002). Use of 16S ribosomal DNA for delineation of marine bacterioplankton species. *Applied and*

Environtal Microbiology, **68**:3628–3633.

Hamdan LJ, Coffin RB, Sikaroodi M, Greinert J, Treude T, Gillevet PM. (2013). Ocean currents shape the microbiome of Arctic marine sediments. *ISME Journal,* **7**:685–96.

Hanson CA, Fuhrman JA, Horner-Devine MC, Martiny JBH. (2012). Beyond biogeographic patterns: processes shaping the microbial landscape. *Nat Rev Microbiology,* **10**:1–10.

Herndl GJ, Reinthaler T. (2013). Microbial control of the dark end of the biological pump. *Nature Geoscies,* **6**:718–724.

Herndl GJ, Reinthaler T, Teira E, van Aken H, Veth C, Pernthaler A, *et al.* (2005). Contribution of Archaea to total prokaryotic production in the deep Atlantic Ocean. *Applied and Environmental Microbiology,* **71**:2303–9.

Irigoien X, Klevjer TA, Røstad A, Martinez U, Boyra G, Acuña JL, *et al.* (2014). Large mesopelagic fishes biomass and trophic efficiency in the open ocean. *Nat Commun* **5**:3271.

Ivars-Martinez E, Martin-Cuadrado A-B, D'Auria G, Mira A, Ferriera S, Johnson J, *et al.* (2008). Comparative genomics of two ecotypes of the marine planktonic copiotroph Alteromonas macleodii suggests alternative lifestyles associated with different kinds of particulate organic matter. *ISME Journal,* **2**:1194–212.

Jenkins WJ. (1982). Oxygen utilization rates in North Atlantic subtropical gyre and primary production in oligotrophic systems. *Nature,* **300**:246–248.

Karner MB, DeLong EF, Karl DM. (2001). Archaeal dominance in the mesopelagic zone of the Pacific Ocean. *Nature,* **409**:507–10.

López-García P, López-López A, Moreira D, Rodríguez-Valera F. (2001). Diversity of free-living prokaryotes from a deep-sea site at the Antarctic Polar Front. *FEMS Microbiology Ecology,* **36**:193–202.

López-López A, Bartual SG, Stal L, Onyshchenko O, Rodríguez-Valera F. (2005). Genetic analysis of housekeeping genes reveals a deep-sea ecotype of Alteromonas macleodii in the Mediterranean Sea. *Environmental Microbiology,* **7**:649–59.

Martín-Cuadrado A-B, López-García P, Alba J-C, Moreira D, Monticelli L, Strittmatter A, *et al.* (2007). Metagenomics of the deep Mediterranean, a warm bathypelagic habitat. *PLoS One,* **2**:e914.

Massana R, DeLong EF, Pedrós-Alió C. (2000). A Few Cosmopolitan Phylotypes Dominate Planktonic Archaeal Assemblages in Widely Different Oceanic Provinces. *Appl Environmental Microbiology,* **66**:1777–1787.

Massana R, Murray A, Preston C, DeLong E. (1997). Vertical distribution and phylogenetic characterization of marine planktonic Archaea in the Santa Barbara Channel. *Appl Environmental Microbiology,* **63**:50–56.

Minchin PR. (1987). An evaluation of the relative robustness of techniques for ecological ordination. *Vegetation,* **69**:89–107.

Moalic Y, Desbruyères D, Duarte CM, Rozenfeld AF, Bachraty C, Arnaud-Haond S. (2012). Biogeography revisited with network theory: retracing the history of hydrothermal vent communities. *Systematic Biology,* **61**:127–37.

Nunoura T, Takaki Y, Hirai M, Shimamura S, Makabe A, Koide O. (2015). Hadal biosphere: Insight into the microbial ecosystem in the deepest ocean on Earth. *Proceedings of the National Academy of Sciences USA,* **112:***1230-1236.*

Oden NL, Sokal RR. (1986). Directional Autocorrelation: An Extension of Spatial Correlograms to Two Dimensions. *Systematic Zoology,* **35**:608.

Pedrós-Alió C. (2012). The rare bacterial biosphere. *Ann Rev Mar Sci* **4**:449–466.

Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J, *et al.* (2007). SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Research,* **35**:7188–96.

Pruitt KD, Tatusova T, Brown GR, Maglott DR. (2012). NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Research,* **40**:130–5.

Quaiser A, Zivanovic Y, Moreira D, López-García P. (2011). Comparative metagenomics of bathypelagic plankton and bottom sediment from the Sea of Marmara. *ISME Journal,* **5**:285–304.

R Core Team. (2014). R: A language and environment for statistical computing.

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 Biology,* **5**:e77.

Schauer R, Bienhold C, Ramette A, Harder J. (2010). Bacterial diversity and biogeography in deep-sea surface sediments of the South Atlantic Ocean. *ISME Journal,* **4**:159–170.

Smedile F, Messina E, La Cono V, Tsoy O, Monticelli LS, Borghini M, *et al.* (2012). Metagenomic analysis of hadopelagic microbial assemblages thriving at the deepest part of Mediterranean Sea, Matapan-Vavilov Deep. *Environmental Microbiol*ogy, **1**:167-182

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'. *Proceedings of the National Academy of Sciences USA,* **103**:12115–20.

Stackebrandt E. (2006). Defining Taxonomic Ranks. In:*The Prokaryotes SE - 3*, Dworkin, M, Falkow, S, Rosenberg, E, Schleifer, K-H, & Stackebrandt, E (eds), Springer New York, pp. 29–57.

Stackebrandt E, Goebel BM. (1994). Taxonomic note: a place for DNA-DNA reassocia-

tion and 16S rRNA sequence analysis in the present species definition in bacteriology. *International Journal Systematic Bacteriology,* **44**:846–849.

Sul WJ, Oliver T a, Ducklow HW, Amaral-Zettler L a, Sogin ML. (2013). Marine bacteria exhibit a bipolar distribution. *Proceedings of the National Academy of Sciences USA,* **110**:2342-2347.

Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, *et al.* (2015). Structure and function of the global ocean microbiome. *Science,* **348**:1261359–1261359.

Teira E, Lebaron P, van Aken H, Herndl GJ. (2006). Distribution and activity of Bacteria and Archaea in the deep water masses of the North Atlantic. *Limnology and Oceanography,* **51**:2131–2144.

Wang Y, Cao H, Zhang G, Bougouffa S, Lee OO, Al-Suwailem A, *et al.* (2013). Autotrophic microbe metagenomes and metabolic pathways differentiate adjacent red sea brine pools. *Scientific Reports,* **3**:1748.

Wilkins D, van Sebille E, Rintoul SR, Lauro FM, Cavicchioli R. (2013). Advection shapes Southern Ocean microbial assemblages independent of distance and environment effects. *Nature Communications,* **4**:2457.

Winter C, Kerros M-E, Weinbauer MG. (2009). Seasonal changes of bacterial and archaeal communities in the dark ocean: Evidence from the Mediterranean Sea. *Limnology and Oceanography,* **54**:160–170.

Yooseph S, Sutton G, Rusch DB, Halpern AL, Williamson SJ, Remington K, *et al.* (2007). The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. *PLoS Biology,* **5**:e16.

Zinger L, Amaral-Zettler L a., Fuhrman J a., Horner-Devine MC, Huse SM, Welch DBM, *et al.* (2011). Global Patterns of Bacterial Beta-Diversity in Seafloor and Seawater Ecosystems Gilbert, JA (ed). *PLoS One* **6**:e24570.

Particle-association lifestyle is a phylogenetically conserved trait in bathypelagic prokaryotes

Chapter II

Abstract

The free-living (FL) and particle-attached (PA) marine microbial communities have repeatedly been proved to differ in their diversity and composition in the photic ocean and also recently in the bathypelagic ocean at a global scale. However, although high taxonomic ranks exhibit preferences for a PA or FL mode of life, it remains poorly understood whether two clear lifestyles do exist and how these are distributed across the prokaryotic phylogeny. We studied the FL (<0.8 μm) and PA (0.8 – 20 μm) prokaryotes at 30 stations distributed worldwide within the bathypelagic oceanic realm (2,150 – 4,000 m depth) using high throughput sequencing of the small subunit ribosomal RNA gene (16S rRNA). A high proportion of the bathypelagic prokaryotes were mostly found either attached to particles or freely in the surrounding water but rarely in both types of environments. In particular, this trait was deeply conserved through their phylogeny suggesting that the deep-ocean particles and the surrounding water constitute two highly distinct niches and that transitions from one to the other have been rare at an evolutionary time-scale. As a consequence, PA and FL communities had clear alpha- and beta-diversity differences that exceeded the global-scale geographical variation. Our study organizes the bathypelagic prokaryotic diversity into a reasonable number of ecologically coherent taxa regarding their association to particles, a first step for understanding which are the microbes responsible for the processing of the dissolved and particulate pools of organic matter that have a very different biogeochemical role in the deep ocean.

Introduction

The deep ocean contains 70% of the ocean's microbial cells and represents 60% of its heterotrophic activity (Arístegui *et al*. 2009). This activity is supported by a flux of particles produced in the upper ocean, which is considered to be the dominant input of organic carbon to the deep-sea (Ducklow *et al*. 2001; Arístegui *et al*. 2002) as dissolved organic carbon is largely consumed within the mesopelagic layer (Arístegui *et al*. 2002). Hence, the deep-sea pelagic ecosystem is not an homogenous habitat where microbes grow in suspension, but contains a variety of particles that represent important sources of organic matter fueling the dark ocean food web (Herndl & Reinthaler 2013). These include fast-sinking particles, sinking through the deep-ocean over a few weeks, as well as buoyant or slow-sinking organic particles, which remain suspended in the deep ocean over annual times scales (Herndl & Reinthaler 2013). Thus, the identification of the microorganisms inhabiting deep ocean's organic particles and the ones living freely in the water is a crucial first step for deciphering the ecological functioning of the deep ocean.

Differences between free-living (FL) and particle-attached (PA) microbial communities have been observed in relation to their local abundance and biomass, substrates incorporation or adaptation to different ecological features such as the degradation of organic matter compounds (Caron *et al*. 1982; Pedrós-Alió & Brock 1983; Fernández-Gómez *et al*. 2013). However, comparative analyses of FL and PA microbial communities have been restricted mostly to the photic ocean, where particles remain in suspension less than one month (Lande & Wood 1987) and/or to specific locations in the aphotic ocean and have used a diversity of approaches that made robust comparisons difficult among studies (Acinas *et al*. 1999; Ghiglione *et al*. 2007; Eloe *et al*. 2011; Smith *et al*. 2013; Crespo *et al*. 2013). Hence, there is a need for geographically extensive and coherent sampling efforts to examine the consistency of the differences between PA and FL communities across distant locations in the deep ocean.

The diversity and biogeography of bathypelagic prokaryotic communities has only been recently described at a global scale showing that PA and FL communities differ greatly in composition and appear to be structured by different ecological drivers (Salazar *et al*. 2016). Additionally, high taxonomic ranks (such as Orders or Phyla) have been shown to exhibit contrasting abundance patterns between FL and PA communities (Eloe *et al*. 2011; Smith *et al*. 2013; Crespo *et al*. 2013) suggesting that some degree of ecological coherence exists for high taxonomic ranks in relation to the degree of particle-association. Despite that, only the recent emergence of high-throughput sequencing techniques allows the exploration of the diversity of the FL and PA communities from a phylogenetic

point of view and thus understanding whether and how these two lifestyles are linked to the prokaryotic evolutionary history.

Here we examine the phylogenetic patterns of free-living and particle-associated prokaryote communities in the global deep ocean. We do so on the basis of a global survey of samples collected between 2,000 - 4,000 m depth along the track of the Malaspina 2010 Circumnavigation Expedition (Duarte 2015). Samples were divided into two size fractions: 0.2-0.8 (FL) and 0.8-20 μm (PA) in order to operationally separate free-living cells from those attached to small particles, and analyzed by high-throughput sequencing of the16S rRNA prokaryotic genes. We combined ecological and phylogenetic analyses in order to: i) test whether two clear lifestyles (FL and PA) exist in the bathypelagic ocean and whether these are consistent at a global scale, ii) quantify the proportion of prokaryotic taxa associated with each lifestyle, iii) test the phylogenetic conservation of these two lifestyles and iv) identify the abundant and cosmopolitan members of the FL and PA communities in the dark ocean.

Material and Methods

Sample collection

A total of 60 samples of deep, bathypelagic water were obtained during the Malaspina 2010 expedition corresponding to 30 different sampling stations globally distributed across the world's oceans (Fig. S1, Table S1). We focused our efforts in sampling at a depth of 4,000 m, although a few samples were taken at shallower depths within the bathypelagic zone where orographic constraints prevented deeper sampling. Two different size fractions were analyzed in each sample to characterize each of the free-living (FL, 0.2-0.8 μm) and particle-attached (PA, 0.8-20 μm) prokaryotes. For each sample 120 L of seawater were sequentially filtered through a 200 μm and a 20 μm mesh to remove large plankton. Further filtering was done by filtering water serially through 142 mm polycarbonate membrane filters of 0.8 μm (Merk Millipore, Isopore polycarbonate) and 0.2 μm (Merk Millipore, Express Plus) pore size with a peristaltic pump (Masterflex, EW-77410-10) obtaining a final set of 60 samples. The filters were then flash-frozen in liquid N_2 and stored at -80ºC until DNA extraction. For that purpose, the filters were cut in small pieces with sterile razor blades and half of each filter was used for DNA extractions, which were performed using the standard phenol-chloroform protocol with slight modifications (Logares *et al.* 2013). Details regarding the methodological approach have been presented before (Salazar *et al.* 2015).

Sample sequencing and processing

Prokaryotic diversity was assessed using amplicon sequencing of the V4 region of the

16S rRNA gene with the Illumina MiSeq platform using paired-end reads (2 X 250 bp) and primers targeting prokaryotes (i.e. both Bacteria and Archaea). All library construction and sequencing was carried out at the JGI (www.jgi.doe.gov) following a standard protocol (Caporaso *et al.* 2011). Briefly, the variable region V4 of the 16S rRNA gene was amplified using primers F515/R806. Primer 806R has been recently shown to underestimate the abundance of SAR11 and Thaumarchaeota (Apprill *et al.* 2015; Parada *et al.* 2015). This dataset, however, was shown to be in good agreement with data derived from metagenomes, and thus not dependent on primers. Although SAR11 abundances were underestimated, Thaumarchaeota abundances derived from 16S rRNA sequencing and metagenomes were highly consistent (Salazar *et al.* 2015). Sequence processing included the removal of contaminants, disrupted pair-end reads and PhiX spike-in shotgun library reads included as internal standards, trimming and assembling of remaining pair-end reads, removal of primer sequences, quality control using sliding window and clustering at 97% identity for the construction of Operational Taxonomic Units (OTUs). Singletons (i.e. OTUs occurring once in just one sample) and chimerical OTUs were removed. The remaining OTUs were taxonomically annotated using both the online RDP Naïve Bayesan Classifier (Wang *et al.* 2007) and the BLAST-based classifier within the QIIME pipeline (Caporaso *et al.* 2011) using the SILVA database (release 115) as reference. The best match to SILVA (minimum similarity of 70%) was used to annotate each OTU by this approach. A minimum confidence value of 90 was used as the criterion for the RDP-based annotation. An OTU abundance table was constructed containing the number of reads belonging to every OTU in each sample. Details on the sequence processing have also been described before (Salazar *et al.* 2016).

Phylogenetic reconstruction

Short sequences, such as the reads obtained through Illumina sequencing, may be problematic for phylogenetic reconstruction, especially for the satisfactorily resolving the evolutionary relations between broad taxonomic groups (Moret *et al.* 2002). However, new tools have been developed for this purpose that use reference phylogenies, usually constructed with longer sequences, and add the short reads through the use of new algorithms developed for this purpose (Matsen *et al.* 2010; Berger *et al.* 2011). This approach, which has been applied before for microbial eukaryotes (Dunthorn *et al.* 2014; Monier *et al.* 2014), bacteria (Brazelton *et al.* 2013; Larsson *et al.* 2014) and viruses (Mengual-Chuliá *et al.* 2012) was used for the present study: A phylogeny was inferred for all the representative OTU sequences with an average of 250 bp (from 219 to 278 bp) through its phylogenetic placement into a previously constructed phylogenetic tree with full-length 16S rRNA sequences. The closest sequence to each OTU in SILVA v.115 database was found

and collected using BLAST (Altschul *et al.* 1990) and used for the construction of an initial phylogeny (the alignment provided by SILVA v.115 release was used). The phylogeny was constructed using maximum likelihood inference with RAxML v. 8.0.19 (Stamatakis 2014) and the GTR evolutionary model with optimization of substitution rates and of sitespecific evolutionary rates (GTRCATI). The best tree was selected from a total of 100 trees constructed for the topology and 100 extra trees were generated for computing the bootstrap values. This initial phylogeny was used for the insertion of the representative OTU sequences using the evolutionary placement algorithm (Berger *et al.* 2011) as implemented in RAxML v. 8.0.19 (Stamatakis 2014). For that purpose we used the previosuly constructed tree and an alignment containing both set of sequences (the representative OTU sequences and the sequences used for the first tree). This alignment was constructed with MOTHUR (Schloss *et al.* 2009) by aligning the first set of sequences while using the second as a reference. The alignment was trimmed to the common 16S rRNA gene fragment covered by both sets. The same evolutionary model (GTRCATI) was used for the inclusion of the representative OTU sequences within the initial phylogeny. The final phylogeny was visually inspected and 8 OTUs were removed because they corresponded to very large branches and were closely related to mithocondria sequences (confirmed using BLAST against NCBI). An additional reduced phylogeny was constructed containing only the OTUs with more than 10 reads (see motivation in *Results*).

Statistical analyses

All data treatment and statistical analyses were conducted with the R Statistical Software (R Core Team 2015) using version 3.1.0 and *vegan* (Oksanen *et al.* 2015), *ape* (Paradis *et al.* 2004), *picante* (Kembel *et al.* 2010), *geiger* (Harmon *et al.* 2008), *MASS* (Venables & Ripley 2002) and *indicspecies* (De Cáceres & Legendre 2009) packages. All the analyses were performed using an OTU abundance table that was previously sampled down to the minimum number of reads (10,617 reads/sample) in order to avoid artifacts due to an uneven sequencing effort among samples.

Alpha and beta-diversity

We calculated prokaryotic richness/diversity metrics using two approaches: an OTU-based approach (i.e. considering the OTUs as unrelated biological entities) and a phylogenetic approach (i.e. considering the evolutionary relationships among OTUs using the complete computed phylogeny). The number of OTUs, the Chao extrapolative richness estimator (Colwell & Coddington 1994) and the Shannon entropy index (Shannon 1948) were computed as OTU-based metrics and the Faith's phylogenetic diversity (PD) (Faith

1992), the PD divided by the number of OTUs (PD/OTUs, hereafter) and the mean nearest taxon distance (MNTD) (Webb *et al.* 2002) were used as phylogenetic measures of diversity. Differences between FL and PA for richness/diversity measures were tested using Mann-Whitney test, as data normality was not assured.

The study of how prokaryotic assemblages vary along sites (i.e. beta-diversity), was also approached by using both OTU-based and phylogenetic beta-diversity distances. The Bray-Curtis dissimilarity index of community composition was used as the OTU-based beta-diversity distance, and betaMNTD (Webb *et al.* 2008) as the phylogenetic beta-diversity distance. Non-metric multidimensional scaling (NMDS) (Minchin 1987) analysis using random starts was used for visualization of beta-diversity and Permutational MANOVA (McArdle & Anderson 2001; Anderson 2001) using 1000 permutations was used to test for significant differences and to partition the beta-diversity matrix variance between FL and PA group of samples.

Particle-association niche index analyses

To numerically characterize each OTU in relation to its occurrence and relative abundance in the PA and FL sets of samples we defined a "particle-association niche index" (PAN index) for each OTU as a measure of the position of an OTU in a continuous niche space ranging from a completely free-living to a completely particle-attached lifestyle. We computed the PAN index by using an abundance-weighted mean: for a given OTU we recorded its abundance in every sample and recorded the size-fraction that every sample belonged to. FL samples were given a value of 0 and PA a value of 1. We then found the abundance-weighted mean of these values. Thus, an OTU occurring only in PA samples would have a PAN-index value of 1 and an OTU strictly occurring in FL samples would have a value of 0. An OTU equally distributed across FL and PA samples would have a PAN-index value of 0.5. This index allows positioning every OTU in a continuum describing its lifestyle preference. This approach has been previously used to define microbial niches in soil microbial ecology in relation to variables such as subsurface depth or soil mud (Stegen *et al.* 2012, 2013; Wang *et al.* 2013).

PAN-index values were compared to null model communities constructed with randomization methods in order to test whether bathypelagic prokaryotes exhibit an association to particles different from what is expected if populations had unlimited dispersal across samples (and thus, across PA and FL fractions) and were free of selection pressures. We constructed null communities by randomly permuting the counts across our abundance OTU table, maintaining row and column sums, (i.e., the total number of counts per sample and the global absolute abundance of each OTU), and thus controlling for

sampling design and global taxon abundance. We permuted the matrix 1,000 times and computed the PAN index for each OTU for these null matrices. These 1,000 null PAN-index values were compared to the real PAN-index values for each OTU. Randomizations were performed using the *quasiswap* algorithm (Miklós & Podani 2004) in *permatswap* function within the *vegan* R package.

PAN-index values were also used to explore the phylogenetic signal of the particle-association lifestyle (i.e. to ask whether closely related OTUs have similar associations to PA or FL habitats). To summarize major trends in this relationship, the between-OTU niche differences (the difference for each pair of OTUs in its PAN index) were placed in phylogenetic distance bins and the mean niche difference was computed for each bin (Stegen *et al.* 2012). A bin interval of 0.01 units was used (the maximum phylogenetic distance was 3.74 arbitrary units). This allowed the identification of the phylogenetic distance threshold beyond which niche differences no longer increased with phylogenetic distance. The phylogenetic signal was also statistically tested using Pagel's l index (Pagel 1999), which yields values close to 0 when there is phylogenetic independence of a trait and close to 1 when species' traits are distributed as expected under a Brownian model of trait evolution (Münkemüller *et al.* 2012). Pagel's l was estimated by maximum likelihood using the *fitContinuous* function within the *geiger* R package. To test whether the estimate was significantly different from 0 (i.e. whether there was a significant phylogenetic signal) it was compared to a model assuming a l equal to 0 (i.e. no phylogenetic signal) using a likelihood ratio test (LRT).

Analysis of specific lineages analyses in PA and FL microbial communities

The detection of specific lineages with a PA or FL lifestyle was performed using two different strategies: one at broad taxonomic levels (Phylum and finer levels) and a second approach at the OTU level. For the first strategy, OTUs belonging to the same Phylum were grouped using SILVA-derived taxonomy. The group PAN-index values were tested for significance using the null hypothesis that the mean PAN-index of all the OTUs belonging to a same Phylum equaled 0.5 (i.e. the expected PAN index for an OTU that is equally distributed across PA and FL samples). The significance was tested with one-sample Wilcoxon signed rank tests, as data normality was not assured. In order to assure a minimum sample size, only Phyla containing more than 40 OTUs were tested. P-values were adjusted for multiple comparisons using False Discovery Rate correction (Benjamini & Hochberg 1995). The same was done for the main lineages clearly annotated within each Phylum, which were selected according to the SILVA-based taxonomical annotation of OTUs. A second strategy consisted on detecting "indicator OTUs" for each of the two

lifestyles, i.e. cosmopolitan OTUs (widely distributed OTUs across sampling stations) but restricted to the PA or FL group of samples. This was addressed using the "indicator species" approach (Dufrene & Legendre 1997; De Cáceres & Legendre 2009). Indicator OTUs for the two size-fractions were identified using the IndVal index, which is a combined measure of "specificity" (A, the proportion of the total reads of an OTU that appear in a given size-fraction) and "fidelity" (B, the proportion of samples of a given size-fraction where an OTU occurs) (De Cáceres & Legendre 2009). The significance of the association was tested using permutation tests. Those indicator OTUs with a p-value<0.05 and both, a fidelity and specificity value ≥0.8, were considered valid. This assures that potential indicator OTUs are both widely distributed among stations and restricted to any of the two size-fractions.

Results

The samples of bathypelagic prokaryotes formed two non-overlapping clusters that exactly corresponded to the PA and FL samples in an NMDS ordination space built from OTU-based Bray-Curtis distances (Fig. 1a), accounting for a third of the variance in community composition across samples (Permutational MANOVA: F=26.295, R^2= 0.312, P-value=0.001). The mean Bray-Curtis distance between samples belonging to the same size-fraction was in both cases (i.e. in FL and PA samples) lower than the mean distance between the two size-fractions of the same station (Fig. S2; Mann-Whitney test: FL:

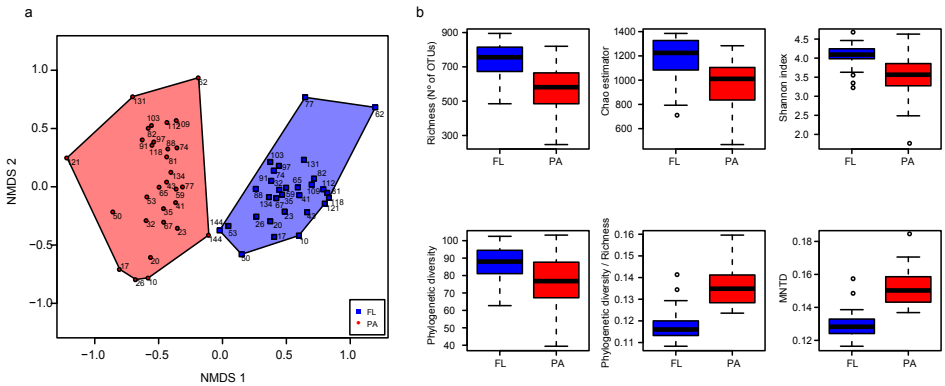


Figure 1: Beta-diversity and alpha-diversity of deep-sea prokaryotic communities. a) Beta-diversity visualized using Non-Metric Multidimensional Scaling (NMDS). Samples belonging to particle-attached (PA) and free-living (FL) are color-coded. The number close to each sample corresponds to the sampling station (see Fig. S1 and Table S1). b) Alpha-diversity measures using OTU-based (top panels) and phylogenetic (bottom panels) approaches.

U=11612, P-value<0.0001; PA: U=20252, P-value<0.0001).

Both, phylogenetic beta-diversity (i.e. betaMNTD) and OTU-based beta-diversity (i.e. Bray-Curtis distance) were highly correlated (Mantel test: r=0.77, P-value<0.001) across samples although, irrespectively of the absolute Bray-Curtis value, betaMNTD values within the FL group of samples tended to be lower than within PA (Fig. S3). All three metrics of OTU-based alpha-diversity used here differed significantly between FL and PA samples (Mann-Whitney test: nº of OTUs: U=745.5, P-value<0.0001; Chao richness estimator: U=714, P-value<0.0001; Shannon diversity index: U=749, P-value<0.0001) with the FL set of samples being richer and more diverse in OTUs than their PA counterpart (Fig. 1b). The three measures of phylogenetic diversity were also significantly different (Mann-Whitney test: PD: U=630, P-value=0.007; PD/OTUs: U=44, P-value<0.0001; MNTD: U=37, P-value<0.0001) with a higher phylogenetic diversity in the FL group of samples while the phylogenetic diversity per OTU and MNTD were higher in the PA fraction (Fig. 1b).

The placement of each OTU into a continuous niche space described by the PAN-index, ranging from a completely free-living to a completely particle-attached lifestyle, showed a bimodal distribution (Fig. 2), with most OTUs accumulating in the extreme values (close to 0 or 1) in

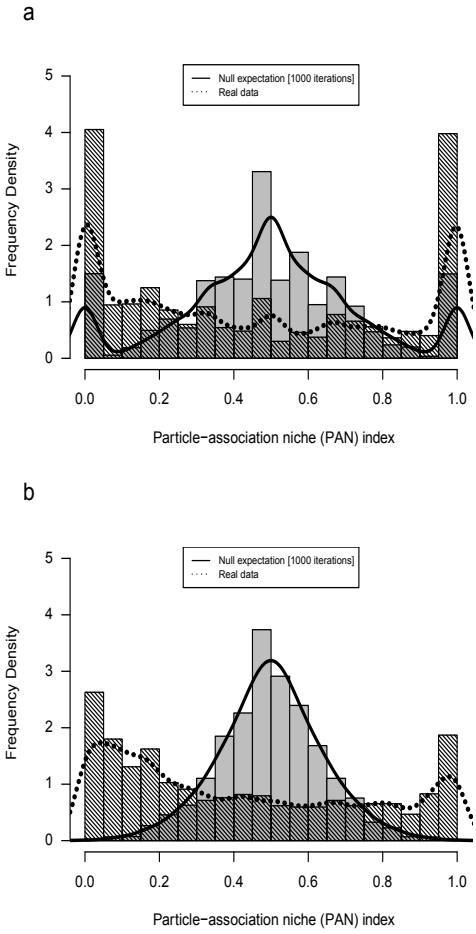


Figure 2: **Histogram of the distribution of real particle-associated niche index (PAN index) for each OTU compared to the null community model expectation (see Materials and Methods for details) ba**sed on 1,000 randomizations. a) Comparison done using the whole dataset and b) a reduced version using the OTUs with more than 10 reads. The lines correspond to kernel density estimates for each distribution: the null distribution (grey bars and the solid line) and the real distribution (hatched bars and dashed line).

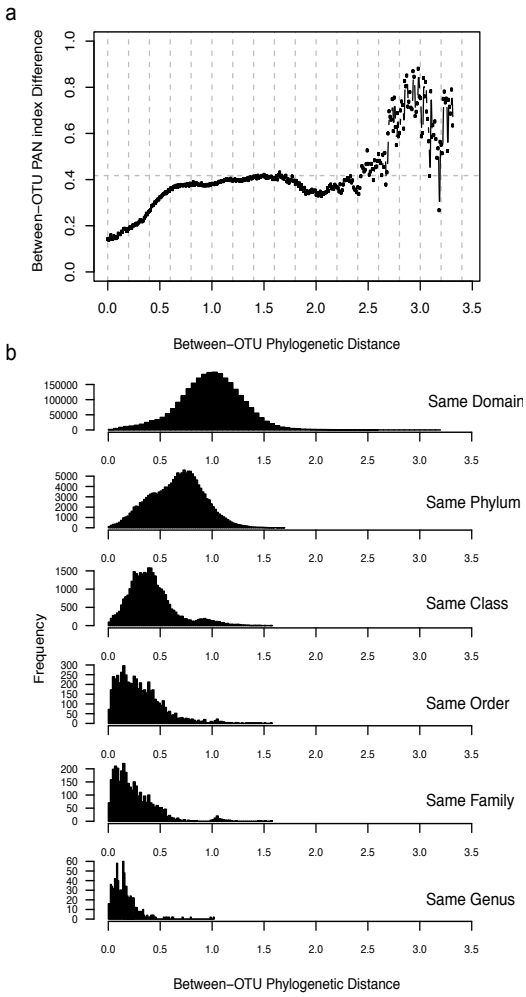


Figure 3: **Phylogenetic signal of particle-associated niche index (PAN index).** a) Mean PAN index differences between pairs of OTUs as a function of between-OTUs phylogenetic distance. Mean values are computed using 0.01 unit bins. The mean PAN index difference for all the OTUs is indicated with a horizontal line. b) Histograms of phylogenetic distances between OTUs belonging to the same Domain, Phylum, Class, Order, Family and Genus based on RDP fix-rank taxonomical annotation. Only taxonomic annotations with confidence values ≥90 were used. The phylogenetic distance scale in both panels is the same. Only the OTUs containing more than 10 reads were used for both panels.

contrast to the unimodal aggregation around 0.5, with two secondary peaks at the extreme values, expected for null, randomization-based communities (Fig. 2a). One third of the OTUs (1,163 out of 3,534) had significantly more extreme PAN index than their null expectation (i.e. the observed values were beyond the 2.5% percentile of the simulated ones to either side of the PAN range). These 1,163 OTUs represented 85.1% of the total reads. The removal of low-abundance OTUs (those having 10 or fewer reads) resulted in the disappearance of these peaks at extreme values in the null expectation but not in the real PAN-index values (Fig. 2b). For this reduced dataset, 60% of the OTUs (1,018 out of 1,712), representing an 82.1% of reads, had more extreme values than their null expectation.

Pairwise differences in PAN index between OTUs were closely and positively correlated with between-OTU phylogenetic distances across short phylogenetic distances (up to 0.6-0.8 arbitrary units, representing a 16-20% of the maximum phylogenetic distance across the entire tree), but there was no systematic relationship for OTU pairs at greater phylogenetic distances (Fig. 3a). Most of the OTUs belonging to the same Class or even Phylum (and lower taxonomic ranks) exhibited pairwise differences in the

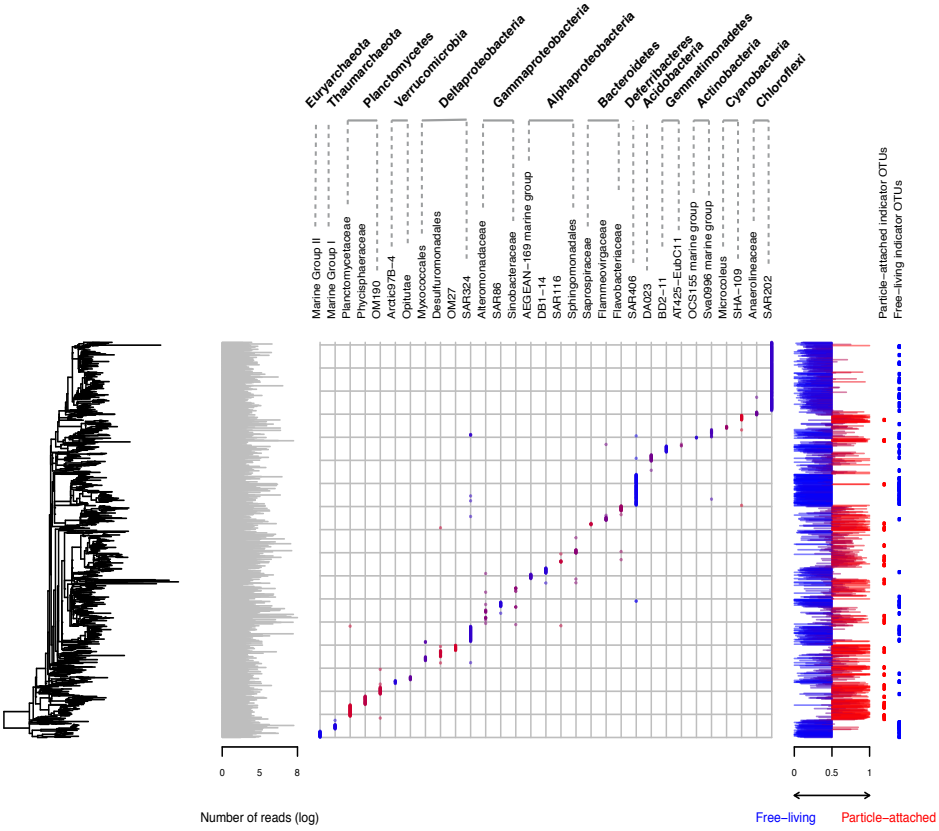


Figure 4: **Phylogenetic placement of PAN index**. Phylogenetic tree representing the evolutionary history of all the OTUs with more than 10 reads. Abundance (log of number of reads) and PAN-index values are represented as bars for each OTU. Mean PAN index is additionally color-coded by a blue-red gradient for each of the main lineages (selected based on the SILVA-based taxonomical annotation). Indicator OTUs for both lifestyles are indicated using blue (free-living) and red (particle-attached) dots.

PAN index equal or lower than this value (Fig. 3b). The analysis of the phylogenetic consistency of the PAN index showed that most OTUs exhibited extreme PAN-index values (i.e. close to 0 or 1) and that these values were consistently conserved for relatively broad clades across the phylogeny (Fig. 4). The phylogenetic signal test for the whole phylogeny resulted in a Pagel's λ estimation of 0.887, significantly different from the value of 0 corresponding to the null-hypothesis (LRT: likelihood ratio=1340.132, P-value≤0.0001).

Seven out of the 15 Phyla and 16 out of the 31 lower-level lineages deviated significantly from the mean PAN index of 0.5 expected in the absence of preference toward the FL or PA life styles (Fig. 5, Table S2 and S3). These differences were also evident

by the positioning of the OTUs belonging to each of the Phyla or finer lineages within the NMDS space (Fig. S4 and Fig. S5). Deferribacteres, Choloroflexi, Euryarchaeota and Thaumarchaeota and all the lower-level lineages tested within them (e.g. SAR406, SAR202, Marine Group I and II) had mean PAN indexes significantly < 0.5, indicating a clear preference for the free-living lifestyle. Other lineages (Arctic 97B-4, SAR324, SAR86, AEGEAN-169, DB1-14 and BD2-11) also had a significantly lower mean PAN index compared to the null hypothesis but the Phyla they belong to did not (Verrucomicrobia, Deltaproteobacteria, Gammaproteobacteria, Alphaproteobacteria, and Gemmatimonadetes). Planctomycetes, Bacteroidetes and Firmicutes had mean PAN index significantly greater than 0.5 supporting a preference for the particle-attached lifestyle. All the lineages tested within Planctomycetes (Planctomycetaceae, Phycisphaeraceae and OM90) had also significantly high mean PAN index, together with Desulfuromonadales and OM27 (Deltaproteobacteria) and the SHA-109 lineage (Cyanobacteria). Despite

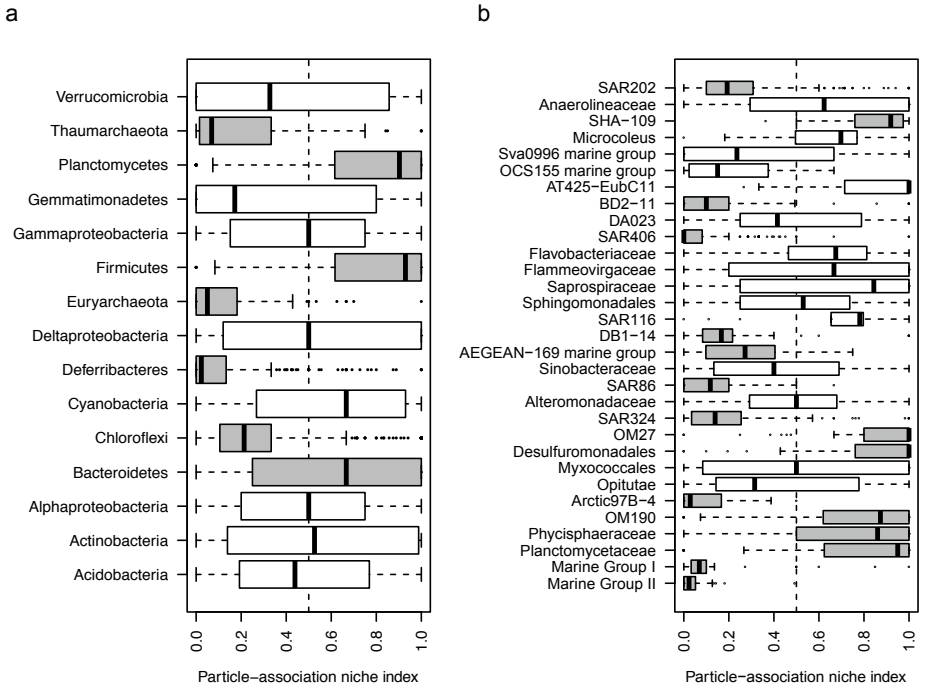


Figure 5: **Boxplots of the particle-association niche index (PAN index)** for a) the main Phyla and b) the main lineages within each Phylum. Only Phyla containing more than 40 OTUs are presented. Lineages within each Phylum were selected based on the SILVA-based taxonomical annotation. Boxplots with PAN-index values significantly different from 0.5 (P-value ≤ 0.05 using Wilcoxon signed rank test and after FDR correction, see *Material and Methods*) are in gray. The vertical broken line corresponds to a 0.5 PAN-index.

Bacteroidetes, as a Phylum, had a mean PAN index significantly greater than 0.5, none of the lineages tested within it yielded significant departures from 0.5 after FDR correction (although they had significant or almost significant P-values before correction). As OTUs within Firmicutes did not group into finer lineages that could be clearly annotated by SILVA, none were included in the analysis.

We detected 100 OTUs with value as indicator OTUs for the FL and 35 for the PA lifestyles (phylogenetically placed in Fig. 4; analysis results in Table S4 and S5). SAR324, SAR406, SAR202, Marine Group I and II dominated the indicator OTUs for the FL lifestyle. These same lineages comprised the 30 most abundant OTUs across the whole dataset (Fig. S6). PA indicator OTUs were mainly composed of several representatives of Planctomycetes, Alphaproteobacteria, Deltaproteobacteria and Gammaproteobacteria.

Discussion

Comparison of the free-living and particle attached microbial communities in the bathypelagic ocean

The results presented here demonstrate niche partitioning in deep-sea prokaryotes as reflected in clear differences in the composition of free-living and particle-associated bathypelagic bacteria. This confirms previous evidence from surface waters in various marine sites (Delong *et al*. 1993; Acinas *et al*. 1999; Hollibaugh & Wong 2000; Kellogg & Deming 2009) as well as indications for the limited set of deep-ocean communities examined in the past (Ghiglione *et al*. 2007; Eloe *et al*. 2011; Smith *et al*. 2013; Crespo *et al*. 2013). Free-living and particle-attached fractions had consistent beta-diversity differences at the scale of the global bathypelagic ocean examined here (Fig. 1a), with these differences accounting for a considerable proportion (~31%) of the variance in community composition (Salazar *et al*. 2016). Moreover the communities within any of the two size-fractions from any location worldwide were, on average, more similar than the FL and PA communities within the same location (Fig. S2). That these differences were consistent both with an OTU-based and a phylogenetic-based approach to characterize beta-diversity (Fig. S3) confirms sufficiently strong niche partitioning between the two size fractions (i.e. affecting a sufficiently large number of prokaryotic species) as to be detected at the community level. Indeed, community differences between these two lifestyles were stronger than geographical variation within each lifestyle at the global scale (Salazar *et al*. 2015).

The free-living and particle-associated bathypelagic communities exhibited also significant differences in alpha-diversity, with the FL communities being, on average, richer and more diverse at an OTU-based (Fig. 1b) and phylogenetic-based level than their PA

counterparts. Phylogenetic Diversity was defined as the sum of the lengths of all those branches in the phylogeny that are members of the target sample and, thus, depends on the numbers of OTUs, which is different for both FL and PA fractions. Indeed, a higher phylogenetic diversity, standardized to the number of OTUs, was observed within PA communities compared to FL. Most studies in surface waters (Acinas *et al*. 1999; Hollibaugh & Wong 2000; Ghiglione *et al*. 2007; Kellogg & Deming 2009) also reported higher gross OTU richness for free-living bacteria. Our result contrasts with a recent study in the Northwestern Mediterranean Sea based also on 16S rRNA sequencing in which PA assemblages were found to be richer in OTUs than the FL fraction, both in photic and aphotic samples (Crespo *et al*. 2013), yet the deep samples in the Mediterranean Sea maintain a relatively high temperature year-round (12 ºC) and deep mixing to the bottom of the bathypelagic is a frequent phenomenon in the area (MEDOC Group 1970). The samples we analyzed had average potential temperatures of 1.4 ºC and never mixed with surface waters. In addition, the FL communities in the present study were less diverse communities when taking into account the phylogenetic relatedness of the OTUs, indicating that, on average, the FL communities are composed of more closely related taxa than their PA counterparts, a pattern that, to our knowledge, had not been described in the past.

The distinction between PA and FL microbes has traditionally been made by size fractionation using a variety of filter's pore sizes ranging from 0.5 μm to 5 μm and no consensus exists on the optimal filter pore size. And it is not clear whether an optimal size can be applicable to different ecosystems. In this work, the PA prokaryotes were defined as those retained in a 0.8 μm filter, this cut-off has been repeatedly used in many other studies (Crump *et al*. 1999; Ghiglione *et al*. 2009; Allen *et al*. 2012). The existence of clear alpha- and beta-diversity differences between PA and FL communities indicates that the 0.8 μm delineation was effective. The PA fraction includes prokaryotes attached to particles and may also include endosymbiotic or parasitic prokaryotes within small protists (<20 μm) , as well as some elongated or aggregated microbes if they are present. Although some authors have observed large cells, individual cells larger than 0.8 μm do not seem to be abundant in the bathypelagic ocean. Our measured biovolumes for bacteria from these same samples (details not shown) corresponded to a mean diameter of 0.495 μm assuming spherical shape (min = 0.417 μm; max = 0.631 μm), thus lower than the 0.8 μm cutoff. However, the existence of some organisms presenting very elongated cells can not be discarded, and thus could be partially retained in the PA fraction.

Attachment to particles or living freely is a phylogenetically conserved trait

Our next goal was to define an index for the strength of the association of an OTU to particles, providing statistical evidence of the significance of the association of a given OTU to the attached or free-living lifestyle. For that purpose we defined the "particle-association niche index" (PAN index) (see *Material and Methods*) by applying an approach used in the past to delineate "microbial niches" from co-occurrence with specific sets of environmental variables (Stegen *et al*. 2012, 2013). Consistent with the observed differences in alpha and beta-diversity for the two lifestyles, the distribution of the particle-association niche index (PAN index) departed from the randomization-based null model expectation (Fig. 2a). This observation implies that the empirical PAN-index values are differently distributed from the distribution expected for communities assembled under unlimited dispersal between size fractions and under the absence of selective differences between the attached and free-living lifestyles. The PAN-index showed a preference for associations of OTUs to either the free-living or particle-attached fractions higher than expected by chance for a third of the OTUs, which represented 85.1% of the total reads. However, the PAN index is unreliable for OTUs represented by a few reads, for which the chance of finding them only in one size fraction, even if they had a random distribution across samples, is not negligible: being *i* the number of reads of an OTU across the whole dataset, the probability (P) of finding the i reads in samples belonging only to one size fraction (that is, half of the samples) under a random distribution is $P=0.5^i$; for i=5 reads, P=3.13% and for i=10 reads, P=0.09%. Thus, statistically-robust tests of association with either lifestyle are not possible based on the PAN index for rare OTUs . As a result, the OTUs with 10 or less reads across the whole dataset were excluded from this analysis (Fig. 2b). The majority of the deep-ocean prokaryotes (60%), thus, showed a preference to either be attached to particles or free-living with only about 40% of OTUs being randomly distributed among fractions. These results provide strong evidence for the existence of a dichotomous lifestyle for most bathypelagic prokaryotes regarding the association to particles in the deep-ocean. Indeed, that is the case for the majority of the 30 most abundant OTUs of the dataset, that, irrespectively of their distribution along stations, are abundant only in one of the two size-fractions (Fig. S6A-D). However, members of abundant OTUs such as the genera *Alteromonas*, *Alcanivorax* or *Pseudoalteromonas* were found to be evenly distributed between size-fractions (Fig. S6A-D).

Previous indications suggested that high bacterial taxonomic ranks have consistent lifestyles regarding their association to particles (Eloe *et al*. 2011; Smith *et al*. 2013; Crespo *et al*. 2013). Despite this evidence, the phylogenetic coherence of the particle-association lifestyle, i.e. the hypothesis that closely related prokaryotes have similar as-

sociation to particles, has never been formally tested. Our results confirmed that closely related prokaryotes exhibit a similar lifestyle in relation to particle attachment, whereas the coherence between their lifestyle decreased with increasing phylogenetic distance between OTUs. This positive linear relation between phylogenetic proximity and lifestyle held up to a 0.6-0.8 phylogenetic distance units (corresponding to a 16-20% of the maximum distance), a threshold that corresponds to the distance separating most of the OTUs belonging to the same Class or even Phylum (Fig. 3b). A secondary increase in the relationship between pairwise OTU-distance and PAN index differences was observed at phylogenetic distances >2 units (Fig. 3a), corresponding to the comparison of pairs of OTUs belonging to different domains (i.e. Bacteria and Archaea). This evidence of coherence in lifestyles across phylogenetic levels represent a pioneer effort at testing the phylogenetic signal of a specific prokaryotic niche, consistent with the scattered observations of a significant phylogenetic conservation of other niche descriptors, such as abundance profiles through time-series for the Baltic Sea bacterioplankton (Andersson *et al*. 2010). Our result supports the hypothesis that high prokaryotic taxonomic ranks could be ecologically coherent (Philippot *et al*. 2010; Koeppel & Wu 2012), at least for some niche axes, and identifies the free-living/particle-attached axis as one of those showing phylogenetic coherence for bathypelagic prokaryotes at the global scale.

The fact that FL and PA prokaryotic communities are consistently composed of distinct members across a worldwide survey, together with the observation that these differences in composition are phylogenetically conserved at a Class/Phylum level suggests that the deep-ocean's particles and the water surrounding them are two highly distinct environments that impose a trade-off for a majority of the bathypelagic prokaryotes, which seem to rarely be able to adapt to both environments. From an evolutionary point of view, transitions from one lifestyle to the other, thus, seem to have been rare in the extended evolutionary history of deep-sea prokaryotes, as has also occurred for other ecological barriers, such as the marine-freshwater transitions (Logares *et al*. 2009). Our results are also consistent with the fact that complex traits are more deeply conserved in the phylogeny than traits that depend on a few genes (Martiny *et al*. 2013), as a complex set of functions seems to be responsible for the general trophic strategy of marine bacteria (Lauro *et al*. 2009) and specifically for attachment to particles (González *et al*. 2008; Ivars-Martinez *et al*. 2008; Fernández-Gómez *et al*. 2013). This metabolic complexity associated to the PA lifestyle, jointly with the fact that marine particles are embedded into the water where the FL prokaryotes inhabit, suggests the predominance of selective pressures in the maintenance of the phylogenetic pattern described here. That is, reduced dispersion between the FL and PA habitat seems unlikely to have maintained the long-

term isolation necessary for the phylogenetic conservation of the two lifestyles. Thus, the most likely mechanism maintaining such a pattern would be the existence of strong selection for the FL and PA populations within their respective habitats. The capacity of PA lifestyle, seems to depend on relatively complex metabolic machineries, and thus, would not be easily transferred by horizontal gene transfer. This combination would explain the maintenance of two pools of phylogenetically distant prokaryotes adapted to either a FL or a PA lifestyle.

Taxonomic affiliation of the PA and FL members in the bathypelagic ocean

The existence of 7 out of 15 Phyla with a clear and consistent association either to the FL or PA mode of life supports the conclusion that particle-associated lifestyle is a deeply conserved trait, as proved before. The archaeal domain was highly restricted to a FL lifestyle: both Thaumarchaeota and Euryarchaeota at a Phylum level had a significant association with the FL lifestyle and representative OTUs of Marine Group I and II dominated the indicator OTU list for the FL fraction. A lifestyle associated with particles should provide access to organic substrates supporting the microbe's requirements, compared to the diluted pool of organic carbon that limits growth of free-living prokaryotes (Arrieta *et al*. 2015). The FL lifestyle of bathypelagic Archaea is consistent with their proven capacity to grow autotrophically, presumably linked to the oxidation of ammonia (Könneke *et al*. 2005; Swan *et al*. 2014) or through the incorporation of simple organic compounds such as amino acids (Ouverney & Fuhrman 2000) or urea (Alonso-Saez *et al*. 2012; Swan *et al*. 2014). Although the primers used in this study are known to underestimate some Thaumarchaeota lineages (Parada *et al*. 2015) we did not observe such underestimation when metagenomic data of these same samples (and thus, without primer biases) was used for comparison (Salazar *et al*. 2016). However all the analyses in the current study are based on abundance comparisons between FL and PA samples and the possible overestimation/underestimation of some lineage abundances should occur in both sets of samples and thus, would not affect our conclusions.

The bacterial OTUs associated with the FL fraction corresponded to those found to be abundant and specific of bathypelagic waters in the few studies where different sampling depths have been analyzed (but without size fractionation), such as representatives of the SAR86, SAR324, SAR406 and SAR202 clades (Agogué *et al*. 2011; Ghiglione *et al*. 2012). This suggests that FL prokaryotes would make up the bulk of the microbial populations in bathypelagic waters while cells attached to particles would constitute a minor fraction of total abundances, as reported for mesopelagic environments (Kirchman & Mitchell 1982; Turley & Mackie 1995; Ghiglione *et al*. 2007). However, PA prokary-

otes seem to be highly active (Kirchman 1993; Ghiglione *et al*. 2007) and thus play a very relevant ecological role, in spite of their lower abundance. In the present study, Bacteroidetes, Firmicutes and Planctomycetes clearly exhibited a consistent PA lifestyle. The preference of Bacteroidetes for the degradation of polymers (Cottrell & Kirchman 2000) and for a PA lifestyle had been also reported before for surface waters (Delong *et al*. 1993).

The differentiation between the particle-associated and free-living lifestyles is consistent with the differences in the biochemical composition between deep particulate and dissolved organic materials. Marine particles are concentrated sources of polymeric material (Minor *et al*. 2003) while the deep oceanic dissolved organic carbon available to free-living bacteria consists mainly of small molecular-size, very diluted molecules (Arrieta *et al*. 2015; Hansman *et al*. 2015). Planctomycetes members are known to be specialized degraders of marine snow and thus play a key role in global carbon turnover (Woebken *et al*. 2007). In fact, most of the indicator OTUs from this Phylum belonged to the well-known *Rhodopirellula* genus, whose genome sequence revealed a large number of genes involved in the breakdown of sulfated polysaccharides (Glöckner *et al*. 2003). The preference for a PA lifestyle for these three Phyla had been previously detected in a single deep sample from the Mediterranean Sea (Crespo *et al*. 2013) and another sample from the Puerto Rico Trench (Eloe *et al*. 2011). Here we report that the association to particles of these three Phyla, jointly with the hitherto unknown association of Deltaproteobacteria clades OM27 and Desulfuromonadales with particles, seems to be a globally consistent feature of the bathypelagic ocean.

References

Acinas SG, Antón J, Rodríguez-Valera F (1999) Diversity of Free-Living and Attached Bacteria in Offshore Western Mediterranean Waters as Depicted by Analysis of Genes Encoding 16S rRNA. *Applied and Environmental Microbiology*, **65**:514–522.

Agogué H, Lamy D, Neal PR, Sogin ML, Herndl GJ (2011) Water mass-specificity of bacterial communities in the North Atlantic revealed by massively parallel sequencing. *Molecular Ecology*, **20**:258–274.

Allen LZ, Allen EE, Badger JH *et al.* (2012) Influence of nutrients and currents on the genomic composition of microbes across an upwelling mosaic. *The ISME Journal*, **6**:1403–1414.

Alonso-Saez L, Waller AS, Mende DR *et al.* (2012) Role for urea in nitrification by polar marine Archaea. *Proceedings of the National Academy of Sciences*, **109**:17989–17994.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**:403–10.

Anderson MJ (2001) A new method for non-parametric multivariate analysis of variance. *Austral Ecology*, **26**:32–46.

Andersson AF, Riemann L, Bertilsson S (2010) Pyrosequencing reveals contrasting seasonal dynamics of taxa within Baltic Sea bacterioplankton communities. *The ISME Journal*, **4**_171–81.

Apprill a, McNally S, Parsons R, Weber L (2015) Minor revision to V4 region SSU rRNA 806R gene primer greatly increases detection of SAR11 bacterioplankton. *Aquatic Microbial Ecology*, **75**:129–137.

Arístegui J, Gasol JM, Duarte CM, Herndl GJ (2009) Microbial oceanography of the dark ocean's pelagic realm. *Limnology and Oceanography*, **54**:1501–1529.

Arístegui J, Duarte CM, Agustí S *et al.* (2002) Dissolved organic carbon support of respiration in the dark ocean. *Science*, **298** :1967.

Arrieta JM, Mayol E, Hansman RL *et al.* (2015) Dilution limits dissolved organic carbon utilization in the deep ocean. *Science*, **348**:331–333.

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, **57**:289–300.

Berger SA, Krompass D, Stamatakis A (2011) Performance, accuracy, and Web server for evolutionary placement of short sequence reads under maximum likelihood. *Systematic Biology*, **60**:291–302.

Brazelton WJ, Morrill PL, Szponar N, Schrenk MO (2013) Bacterial communities associated with subsurface geochemical processes in continental serpentinite springs. *Applied*

and Environmental Microbiology, **79**:3906–3916.

De Cáceres M, Legendre P (2009) Associations between species and groups of sites: indices and statistical inference. *Ecology*, **90**:3566–74.

Caporaso JG, Lauber CL, Walters W a *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 USA*, **108:**4516–22.

Caron DA, Davis PG, Madin LP, Sieburth, McN. J (1982) Heterotrophic bacteria and bacterivorous protozoa in oceanic macroaggregates. *Science*, **218**:795–797.

Colwell RK, Coddington JA (1994) Estimating terrestrial biodiversity through extrapolation. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, **345**:101–18.

Cottrell MT, Kirchman DL (2000) Natural assemblages of marine proteobacteria and members of the Cytophaga-Flavobacter cluster consuming low- and high-molecular-weight dissolved organic matter. *Applied and Environmental Microbiology,* 66:1692-1697.

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**:541–52.

Crump BC, Armbrust E V, Baross JA (1999) Phylogenetic analysis of particle-attached and free-living bacterial communities in the Columbia river, its estuary, and the adjacent coastal ocean. *Applied and environmental microbiology*, **65**:3192–204.

DeLong EF, Franks DG, Alldredge AL (1993) Phylogenetic diversity of aggregate-attached vs. free-living marine bacterial assemblages. *Limnology and Oceanography*, **38:**924–934.

Duarte CM (2015) Seafaring in the 21St Century: The Malaspina 2010 Circumnavigation Expedition. *Limnology and Oceanography Bulletin*, **24**:11–14.

Ducklow HW, Steinberg DK, Buesseler KO (2001) Upper ocean carbon export and the biological pump. *Oceanography*, **14**:50–58.

Dufrene M, Legendre P (1997) Species Assemblages and Indicator Species: The Need for a Flexible Asymmetrical Approach. *Ecological Monographs*, **67**:345–366.

Dunthorn M, Otto J, Berger SA *et al.* (2014) Placing environmental next-generation sequencing amplicons from microbial Eukaryotes into a phylogenetic context. *Molecular Biology and Evolution*, **31**:993–1009.

Eloe EA, Shulse CN, Fadrosh DW *et al.* (2011) Compositional differences in particle-associated and free-living microbial assemblages from an extreme deep-ocean environment. *Environmental Microbiology Reports*, **3**:449–58.

Faith DP (1992) Conservation evaluation and phylogenetic diversity. *Biological Conservation*, **61**:1–10.

Fernández-Gómez B, Richter M, Schüler M *et al.* (2013) Ecology of marine Bacteroidetes: a comparative genomics approach. *The ISME Journal*, **7**:1026–37.

Ghiglione J-F, Conan P, Pujo-Pay M (2009) Diversity of total and active free-living vs. particle-attached bacteria in the euphotic zone of the NW Mediterranean Sea. *FEMS Microbiology Letters*, **299**:9–21.

Ghiglione J-F, Galand PE, Pommier T *et al.* (2012) Pole-to-pole biogeography of surface and deep marine bacterial communities. *Proceedings of the National Academy of SciencesUSA*, **109**:17633–8.

Ghiglione JF, Mevel G, Pujo-Pay M *et al.* (2007) Diel and seasonal variations in abundance, activity, and community structure of particle-attached and free-living bacteria in NW Mediterranean Sea. *Microbial Ecology*, **54**:217–31.

Glöckner FO, Kube M, Bauer M *et al.* (2003) Complete genome sequence of the marine planctomycete Pirellula sp. strain 1. *Proceedings of the National Academy of Sciences USA*, **100**:8298–303.

González JM, Fernández-Gómez B, Fernàndez-Guerra A *et al.* (2008) Genome analysis of the proteorhodopsin-containing marine bacterium Polaribacter sp. MED152 (Flavobacteria). *Proceedings of the National Academy of Sciences USA*, **105**:8724–9.

Hansman RL, Dittmar T, Herndl GJ (2015) Conservation of dissolved organic matter molecular composition during mixing of the deep water masses of the northeast Atlantic Ocean. *Marine Chemistry*, **177**:288-297.

Harmon LJ, Weir JT, Brock CD, Glor RE, Challenger W (2008) GEIGER: investigating evolutionary radiations. *Bioinformatics*, **24**_129–31.

Herndl GJ, Reinthaler T (2013) Microbial control of the dark end of the biological pump. *Nature Geoscience*, **6**:718–724.

Hollibaugh JT, Wong PS (2000) Similarity of particle-associated and free-living bacterial communities in northern San Francisco Bay, California. *Aquatic Microbial Ecology*, **21**:103–114.

Ivars-Martinez E, Martin-Cuadrado A-B, D'Auria G *et al.* (2008) Comparative genomics of two ecotypes of the marine planktonic copiotroph Alteromonas macleodii suggests alternative lifestyles associated with different kinds of particulate organic matter. *The ISME Journal*, **2**:1194–212.

Kellogg C, Deming J (2009) Comparison of free-living, suspended particle, and aggregate-associated Bacterial and Archaeal communities in the Laptev Sea. *Aquatic Microbial Ecology*, **57**:1–18.

Kembel SW, Cowan PD, Helmus MR *et al.* (2010) Picante: R tools for integrating phylogenies and ecology. *Bioinformatics*, **26**:1463–4.

Kirchman DL (1993) Leucine incorporation as a measure of biomass production by heterotrophic bacteria. In: *Handbook of methods in aquatic microbial ecology. Lewis* , pp. 509–512.

Kirchman D, Mitchell R (1982) Contribution of particle-bound Bacteria to total microheterotrophic activity in five ponds and two marshes. *Applied and Environmental Microbiology*, **43**:200-209

Koeppel AF, Wu M (2012) Lineage-dependent ecological coherence in bacteria. *FEMS Microbiology Ecology*, **81**:574–82.

Könneke M, Bernhard AE, de la Torre JR *et al.* (2005) Isolation of an autotrophic ammonia-oxidizing marine archaeon. *Nature*, **437**:543–546.

Lande R, Wood AM (1987) Suspension times of particles in the upper ocean. *Deep Sea Research Part A. Oceanographic Research Papers*, **34**:61–72.

Larsson J, Celepli N, Ininbergs K *et al.* (2014) Picocyanobacteria containing a novel pigment gene cluster dominate the brackish water Baltic Sea. *The ISME Journal*, **8**:1892–1903.

Lauro FM, McDougald D, Thomas T *et al.* (2009) The genomic basis of trophic strategy in marine bacteria. *Proceedings of the National Academy of Sciences USA*, **106**:15527–33.

Logares R, Bråte J, Bertilsson S *et al.* (2009) Infrequent marine-freshwater transitions in the microbial world. *Trends in Microbiology*, **17**:414–22.

Logares R, Sunagawa S, Salazar G *et al.* (2013) Metagenomic 16S rDNA Illumina tags are a powerful alternative to amplicon sequencing to explore diversity and structure of microbial communities. *Environmental Microbiology*, **16**:2659–71.

Martiny AC, Treseder K, Pusch G (2013) Phylogenetic conservatism of functional traits in microorganisms. *The ISME Journal*, **7**:830–8.

Matsen FA, Kodner RB, Armbrust EV (2010) pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics*, **11**:538.

McArdle BH, Anderson MJ (2001) Fitting multivariate models to community data: a comment on distance-based redundancy analysis. *Ecology*, **82**:290–297.

MEDOC Group (1970) Observation of Formation of Deep Water in the Mediterranean Sea. *Nature*, **227**:1037–1040.

Mengual-Chuliá B, García-Pérez R, Gottschling M, Nindl I, Bravo IG (2012) Novel animal papillomavirus sequences and accurate phylogenetic placement. *Molecular Phylogenetics and Evolution*, **65**:883–91.

Miklós I, Podani J (2004) Randomization of presence–absence matrices: comments and new algorithms. *Ecology*, **85**:86–92.

Minchin PR (1987) An evaluation of the relative robustness of techniques for ecological ordination. *Vegetatio*, **69**:89–107.

Minor EC, Wakeham SG, Lee C (2003) Changes in the molecular-level characteristics of sinking marine particles with water column depth. *Geochimica et Cosmochimica Acta*, **67**:4277–4288.

Monier A, Comte J, Babin M *et al.* (2014) Oceanographic structure drives the assembly processes of microbial eukaryotic communities. *The ISME Journal*, **9**:990–1002.

Moret BE, Roshan U, Warnow T (2002) Sequence-length requirements for phylogenetic methods. In: *Algorithms in Bioinformatics SE - 26* Lecture Notes in Computer Science. (eds Guigó R, Gusfield D), pp. 343–356. Springer Berlin Heidelberg.

Münkemüller T, Lavergne S, Bzeznik B *et al.* (2012) How to measure and test phylogenetic signal. *Methods in Ecology and Evolution*, **3**:743–756.

Oksanen J, Blanchet FG, Kindt R *et al.* (2015) vegan: Community Ecology Package.

Ouverney CC, Fuhrman JA (2000) Marine planktonic Archaea take up amino acids. *Applied and Environmental Microbiology*, **66**:4829–4833.

Pagel M (1999) Inferring the historical patterns of biological evolution. *Nature*, **401**:877–884.

Parada A, Needham DM, Fuhrman JA (2015) Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time-series and global field samples. *Environmental microbiology*, **18**:1403-1414.

Paradis E, Claude J, Strimmer K (2004) APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics*, **20**:289–290.

Pedrós-Alió C, Brock TD (1983) The importance of attachment to particles for planktonic bacteria. *Archiv für Hydrobiologie*, **98**:354–379.

Philippot L, Andersson SGE, Battin TJ *et al.* (2010) The ecological coherence of high bacterial taxonomic ranks. *Nature reviews. Microbiology*, **8**:523–9.

R Core Team (2015) R: A language and environment for statistical computing.

Salazar G, Cornejo-Castillo FM, Benítez-Barrios V *et al.* (2016) Global diversity and biogeography of deep-sea pelagic prokaryotes. *The ISME Journal*, **10**:596-608.

Schloss PD, Westcott SL, Ryabin T *et al.* (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, **75**,:537–41.

Shannon CE (1948) A Mathematical Theory of Communication. *Bell System Technical Journal*, **27**:379–423.

Smith MW, Zeigler Allen L, Allen AE, Herfort L, Simon HM (2013) Contrasting genomic properties of free-living and particle-attached microbial assemblages within a coastal ecosystem. *Frontiers in Microbiology*, **4**:1–20.

Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics (Oxford, England)*, **30**:1312–3.

Stegen JC, Lin X, Fredrickson JK *et al.* (2013) Quantifying community assembly processes and identifying features that impose them. *The ISME Journal*, **7**:2069–2079.

Stegen JC, Lin X, Konopka AE, Fredrickson JK (2012) Stochastic and deterministic assembly processes in subsurface microbial communities. *The ISME Journal*, **6**:1653–64.

Swan BK, Chaffin MD, Martinez Garcia M *et al.* (2014) Genomic and Metabolic Diversity of Marine Group I Thaumarchaeota in the Mesopelagic of Two Subtropical Gyres (L Randau, Ed,). *PLoS ONE*, **9**:e95380.

Turley CM, Mackie PJ (1995) Biogeochemical significance of attached and free-living bacteria and the flux of particles in the NE Atlantic Ocean. *Marine Ecology Progress Series*, **115**:191–203.

Venables WN, Ripley BD (2002) *Modern Applied Statistics with S*. Springer, New York.

Wang Q, Garrity GM, Tiedje JM, Cole JR (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology*, **73**:5261–7.

Wang J, Shen J, Wu Y *et al.* (2013) Phylogenetic beta diversity in bacterial assemblages across ecosystems: deterministic versus stochastic processes. *The ISME Journal*, **7**:1310-1321.

Webb CO, Ackerly DD, Kembel SW (2008) Phylocom: software for the analysis of phylogenetic community structure and trait evolution. *Bioinformatics*, **24**:2098–100.

Webb CO, Ackerly DD, McPeek MA, Donoghue MJ (2002) Phylogenies and community ecology. *Annual Review of Ecology and Systematics*, **33**:475–505.

Woebken D, Teeling H, Wecker P *et al.* (2007) Fosmids of novel marine Planctomycetes from the Namibian and Oregon coast upwelling systems and their cross-comparison with planctomycete genomes. *The ISME Journal*, **1**:419–35.

The activity of deep ocean prokaryotes is driven by their particle-association lifestyle

Chapter III

Abstract

The deep ocean is inhabited by, recently characterized, bacterial and archaeal communities. Yet the presence of the DNA of the microbes does not necessarily imply that they are actively growing, and the low temperatures and the scarcity of organic substrates suggest that a large share of the present prokaryotes may be nongrowing. However, a description of the active and inactive members of prokaryotic communities remains unexplored in the bathypelagic realm. Here we use high-troughput sequencing to study the 16S ribosomal RNA (rRNA) and the rRNA gene (rDNA) of the free-living and particle-attached prokaryotic communities of a total of 101 samples from 30 worldwide-distributed stations within the bathypelagic (2,150 – 4,000 m depth). A considerable proportion (22.5 - 56.4%) of the rDNA-detected *operational taxonomic units* (OTUs) were not found in any of the rRNA samples. These globally potentially inactive OTUs corresponded to low-abundance organisms, and are enriched in phototrophic and anaerobic organisms. Euryarchaeota, Thaumarchaeota and the SAR406 clade were the Phyla with lower rRNA:rDNA ratios while SHA-109, Chlamydiae, Actinobacteria and Betaproteobacteria were the ones with higher ratios. We describe a positive relation between the rRNA:rDNA ratio of OTUs and their particle attachment preference. This hints to a global relationship between attachment to particulate matter and rates of growth for the bathypelagic prokaryotes. Our data add to the increasing knowledge of the deep ocean microbiota and their potential impact on the biogeochemical cycles by identifying a globally consistent relationship between their activity and preference to live attached to particulate organic matter.

Introduction

The deep ocean contains ca. 70% of the ocean's microbes (Arístegui *et al.* 2009) and our knowledge about the biological diversity and distribution of these microbes has increased considerably in the last years. Most of this knowledge has been obtained through the characterization of microbial communities by metagenomics (DeLong *et al.* 2006; Martín-Cuadrado *et al.* 2007) or massive sequencing of the ribosomal 16S RNA gene (rDNA) of deep water samples collected in regional and global surveys (Brown *et al.* 2009; Galand *et al.* 2010; Agogué *et al.* 2011; Quaiser *et al.* 2011; Eloe *et al.* 2011; Smedile *et al.* 2012; Wilkins *et al.* 2013; Ganesh *et al.* 2014; Salazar *et al.* 2016). However, rDNA-based techniques only identify populations whose ribosomal genes are present and sufficiently intact to be amplified, but do not discern the active from the inactive populations, which is fundamental to understand the role of different members of the bacterial communities of the bathypelagic ocean. The active and inactive members of bacterial communities have been lately approached through the combined study of their ribosomal RNA (rRNA) and DNA (rDNA). The rRNA has been proved very useful for delineating the active members of microbial communities, yet the detection of rRNA from a given microbe is strictly indicative of its "capacity for protein synthesis" (Blazewicz *et al.* 2013) and thus of its potential activity. Hereafter that is what we will refer as "active" prokaryotes. This approach has been applied in aquatic environments, such as lakes (Jones & Lennon 2010; Denef *et al.* 2016) or the surface ocean (Campbell *et al.* 2009, 2011; Ghiglione *et al.* 2009; Campbell & Kirchman 2012; Hugoni *et al.* 2013; Hunt *et al.* 2013; Zhang *et al.* 2014). These analyses showed that a considerable proportion (10% - 40%) of the operational taxonomic units (OTUs) within a community are potentially inactive or dormant in these environments (Jones & Lennon 2010; Campbell *et al.* 2011). In addition, a negative relation was observed between the OTU's abundance and the probability of being active, which results in that the rare bacterial OTUs are more likely to be active than the abundant ones (Jones & Lennon 2010; Campbell *et al.* 2011; Zhang *et al.* 2014). However, the proportion of active and inactive OTUs, their identity and their relative abundances remains unexplored in the bathypelagic ocean and one could expect that many microbes are inactive given the low temperatures and the scarcity of organic C in this environment.

Particle-attached (PA) and free-living (FL) prokaryotes have been shown to compose two distinct pools of organisms that, at least in the bathypelagic ocean, are structured by different ecological processes (Salazar *et al.* 2016). These two pools are composed of different Phyla and/or Classes, as the particle-association lifestyle seems to have been strongly conserved through the evolutionary history of the deep-sea Bacteria and Ar-

chaea (Salazar *et al.* 2015). The bathypelagic prokaryotes highly depend on the organic matter produced in the photic layer that reaches the deep ocean in the form of sinking particles (Herndl & Reinthaler 2013). While the FL prokaryotes in the deep ocean face with an environment composed of diluted organic compounds (Arrieta *et al.* 2015) the PA prokaryotes are likely to have access to a more concentrated organic pool composed of polymeric materials (Minor *et al.* 2003) which could more easily provide access to organic substrates supporting their growth requirements. This brings us to hypothesize that PA prokaryotes should be more active than the FL prokaryotes in the bathypelagic ocean.

In the present work we aim to i) quantify and identify the active and inactive members of the deep oceanic prokaryotic communities, ii) test whether the rare members are more likely to be active, as observed in surface waters and iii) test whether the PA prokaryotes are more likely to be active than the FL prokaryotes.

Methods

A total of 101 water samples were collected during the Malaspina 2010 expedition corresponding to 30 different sampling stations globally distributed across the subtropical and tropical region of the world's oceans between 2,150 and 4,000 m depth (Table S1). Samples were obtained for 16S rDNA and 16S rRNA sequencing and targeted the free-living (FL, 0.2-0.8 μm) and particle-attached (PA, 0.8-20 μm) prokaryotes (Acinas *et al.* 1999; Crespo *et al.* 2013). Details of the rDNA dataset have been used in previous publications (Salazar *et al.* 2015, 2016) and consist of 60 samples from 30 different stations for which both the PA and FL samples are available. The 16S rRNA dataset consisted of 41 samples (27 from the FL fraction and 14 from the PA fraction). A total of 12 stations in the dataset contained the 4 possible options (i.e. both the rDNA and rRNA sample in both the FL and PA fraction, see below).

Sample collection

All the samples were pre-filtered through a 200 μm and a 20 μm mesh to remove large plankton. Further filtering was done by pouring 120 L of water serially through a 142 mm polycarbonate membrane filter of 0.8 μm (Merk Millipore, Isopore polycarbonate) and through a 0.2 μm (Merk Millipore, Express Plus) pore size filter for the rDNA analyses, as detailed before (Salazar *et al.* 2015, 2016). For the rRNA analyses the filtration was done by pouring 12 L during 15 minutes through a 142 mm polycarbonate membrane filter of 0.8 μm (Merk Millipore, Isopore polycarbonate) and a 0.2 μm pore size (Merk Millipore, Isopore polycarbonate) in order to minimize filtration time. The filtration was done in both cases with a peristaltic pump (Masterflex, EW-77410-10). The filters were

then flash-frozen in liquid N_2 and stored at -80ºC until DNA/RNA extraction. The filters for rDNA sequencing were cut in small pieces with sterile razor blades and half of each filter was used for DNA extractions, which were performed using the standard phenol-chloroform protocol with slight modifications (Logares *et al.* 2013). Details regarding the DNA extraction have been presented before (Salazar *et al.* 2016). RNA was extracted with the RNEasy kit (Qiagen). Residual DNA was removed using the Turbo DNA-free kit (Applied Biosystems, Austin, TX, USA) and the absence of rDNA in the RNA sample pool was verified through PCR. RNA was reverse transcribed using random hexamers and the SuperScriptIII kit (Invitrogen) according to the manufacturer's instructions.

Sample sequencing and processing

Prokaryotic diversity was assessed using amplicon sequencing of the V4 region of the 16S rDNA gene with the Illumina MiSeq platform using paired-end reads (2 x 250 bp). All library construction and sequencing was carried out at the JGI (www.jgi.doe.gov) following an standard protocol (Caporaso *et al.* 2011). The variable region V4 of the 16S rDNA gene was amplified using primers F515/R806 (5'-GTG CCA GCM GCC GCG GTA A-3' / 5'-GGA CTA CHV GGG TWT CTA AT-3'). Before sequencing, PhiX spike-in shotgun library reads were added to the amplicons pool for a final concentration of about 20-25% of the pair-end reads library as an internal standard. Although primer 806R has been shown to underestimate the abundance of SAR11 (Apprill *et al.* 2015), a good agreement was previously observed for the rDNA dataset with data derived from metagenomes, and thus not dependent on primer choice (Salazar *et al.* 2016).

The same sequence processing protocol was used for both datasets (rDNA and rRNA), including filtration of contaminants, disrupted pair-end reads and PhiX spike-in shotgun library reads included as internal standards, trimming and assembling of remaining pair-end reads, removal of primer sequences, quality control using sliding window, and clustering at 97% identity for the construction of Operational Taxonomic Units (OTUs). Singletons (i.e. OTUs occurring once in just one sample) and chimerical OTUs were removed. The remaining OTUs were taxonomically annotated using BLAST against the SILVA database (v.123, Ref NR99) as reference. Two OTU abundance tables were constructed (for the rDNA and the rRNA datasets) containing the number of reads belonging to every OTU in each sample. Details on the sequence processing have also been described before (Salazar *et al.* 2016). To maintain the consistency with the precedent work (Salazar *et al.* 2015, 2016), the rRNA dataset was analyzed independently and merged to the rDNA dataset by comparing both sets of representative OTU sequences: the two OTU tables were merged by comparing through BLAST the representative sequences of the OTUs

from the rRNA dataset with the OTUs from the rDNA dataset. All the OTUs from the two datasets with an identity greater than 97% and coverage greater than 90% were considered to be the same OTU.

All raw sequences used in this study are publicly available at the NCBI Sequence Read Archive (SRA, http://www.ncbi.nlm.nih.gov/Traces/sra) under SRA Study SRP031469 (for the rDNA dataset) and SRP079340 (for the rRNA dataset).

Statistical analyses

All data treatment and statistical analyses were conducted with the R Statistical Software (R Core Team 2016) using version 3.2.4. Three different OTU tables were constructed: a) a "*complete*" OTU table where all the reads in all the 101 samples are used, b) a "*sub-sampled*" OTU table which is the result of sampling down the complete OTU table to the minimum number of reads per sample in order to avoid artifacts due to an uneven sequencing effort among samples and c) a "*balanced*" OTU table which is the result of selecting from the sub-sampled OTU table only the stations for which all the possible samples were available, i.e. for which the four combinations of FL/PA and rDNA/rRNA did exist (stations 10, 17, 20, 26, 41, 53, 59, 74, 77, 88, 97 and 121). The sub-sampled OTU table was obtained by rarefying the complete OTU table to the minimum number of reads (10,617 reads/sample) using the *rrarefy* function in the *vegan* package (Oksanen *et al.* 2015). This process was repeated 100 times and the mean number of reads (rounded to integers) from the 100 rarefactions was used. The *complete* OTU table was only used for the definition of the "active" and "inactive" OTUs. Each OTU was defined either as "inactive", if it was detected in one or more rDNA sample but was not detected in any rRNA sample, or as "active", if it was detected in one or more rRNA sample, regardless wheter it was also detected or not in any rDNA sample. In this way, the higher number of sequences of the rRNA dataset compared to the rDNA one (~5-fold, see Table S1) minimizes the risk of overestimating the number of inactive OTUs due to an incomplete sequencing effort. The rest of the analyses were performed with the *sub-sampled* or the *complete* OTU tables.

The variation of community structure along samples was approached by computing the pairwise Bray-Curtis dissimilarity and we used for visualization non-metric multidimensional scaling (NMDS) (Minchin 1987) analysis using random starts. Permutational MANOVA (McArdle & Anderson 2001; Anderson 2001) using 1000 permutations was used to test for significant differences between the four groups of samples created by the combination of FL/PA and rDNA/rRNA (i.e. FL - rDNA, PA - rDNA, FL - rRNA and PA - rRNA). The dissimilarity within and between these four sets of samples was tested to be significantly different to the mean dissimilarity between samples with one-sample

Wilcoxon signed rank tests, as data normality was not assured.

For each OTU we defined its "attachment preference" and its "activity". The attachment preference was defined as the quotient between the number of reads in the PA-rDNA and the FL-rDNA set of samples. Thus, an OTU having a high attachment preference value would dominate in the PA fraction while a low value would correspond to an OTU dominating in the FL fraction. The activity was defined as the quotient between the number of reads in the rRNA and the rDNA set of samples. This rRNA:rDNA ratio provides an indication of the potential metabolic activity (Campbell *et al.* 2011; Blazewicz *et al.* 2013). The attachment preference and activity were also computed for the Phyla (except the Proteobacteria, that were split in its constituent Classes) using the SILVA-based OTU's taxonomical annotation. Only the Phyla containing at least 500 reads in the whole dataset were considered.

The difference in the proportion of inactive OTUs between size fraction was statistically tested using the Wilcoxon - Mann Whitney test. The probability of an OTU being inactive as a function of their rank was tested using logistic regression analysis. This analysis was repeated defining as inactive OTUs those with less rRNA reads than rDNA reads in order to exactly compare the analysis with previous works (i.e. Jones & Lennon 2010). The correlation between the attachment preference and the activity was tested using a Pearson correlation test of the log-transformed variables. Two linear models were also built for the prediction of the number of rRNA reads for the main Phyla: one with the number of reads of the rDNA as the only predictor (Model 1), and a second model with the attachment preference as a second predictor (Model 2). Both models were compared through their Akaike Information Criteria (AIC) and a likelihood-ratio test.

Results

The merged dataset containing the rDNA and the rRNA samples (Table S1) resulted in a total of 9,825,311 clean reads and a total of 6,441 OTUs. When the sub-sampled table was used (Table S1) a total of 1,065,973 clean reads and a total of 4,252 OTUs remained.

The proportion of inactive OTUs varied between 3% and 30.8% (mean: 20.5%) when the total OTU number is considered (i.e. OTUs detected in both the rDNA and rRNA samples) and from a 22.5% to a 56.4% (mean: 37.8%) when only the number of OTUs detected in the rDNA samples is considered (Table S1). Only moderately higher proportions were obtained when using the sub-sampled dataset (Table S2). The proportion of inactive OTUs significantly differed between the PA and FL samples, when computed with both the total number of OTUs (Wilcoxon-Mann Whitney test: W = 321, P-value =

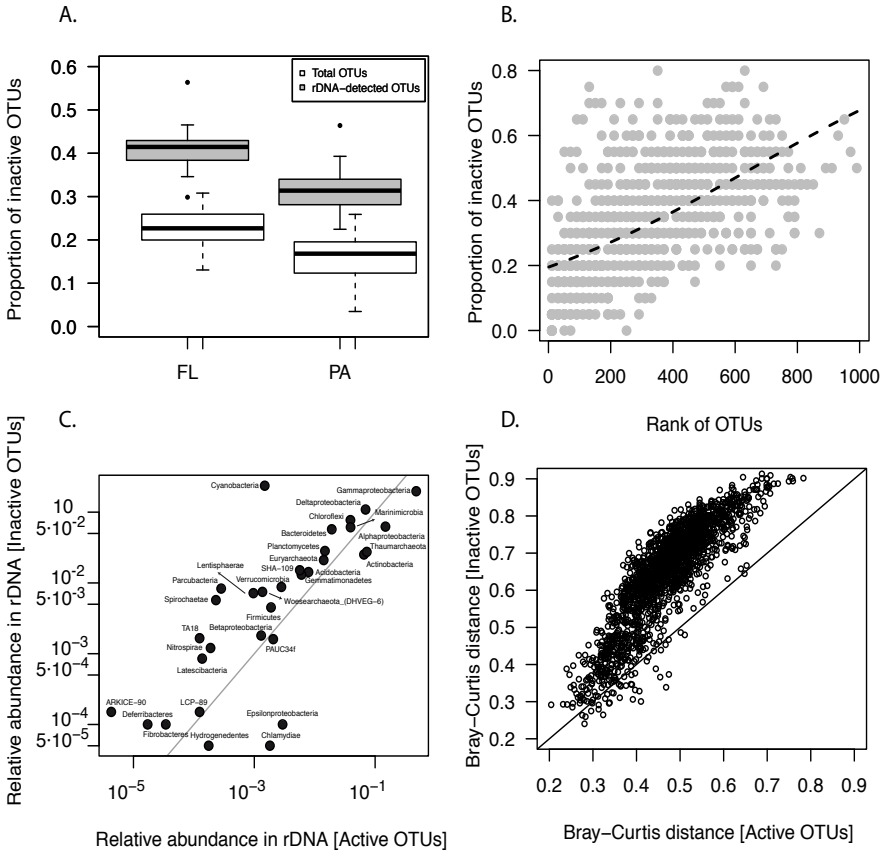


Figure 1: **Inactive OTUs**. A) Proportion of inactive OTUs (i.e. OTUs observed in the rDNA set but not in the rRNA set) in the FL and PA fraction computed as the number of inactive OTUs divided by either the total number of OTUs (in grey) or the total number of OTUs only in the rDNA dataset (in white). B) The proportion of inactive OTUs as a function of their rank for the 24 samples corresponding to the 12 balanced stations. The proportion of inactive OTUs was computed in bins of 20 rank units for each sample. The line corresponds to the equation of the logistic model [$\pi(x) = 1/(e^{-(\beta 0+\beta 1*x)}+1)$] using the mean parameters of the 24 fits ($\beta 0$: -1.3992, $\beta 1$: 0.0021). See Table S3 for the associated logistic regression analyses. C) Relative abundance of the Phyla in the active and inactive pool of OTUs computed from the rDNA dataset. D) Bray-Curtis distance between pairs of samples based on the active and the inactive pool of OTUs.

0.0001) and the rDNA-detected OTUs (Wilcoxon-Mann Whitney test: W = 335, P-value = 0.00001). In both cases the proportion of inactive OTUs was significantly higher in the FL fraction (Fig. 1A). The mean contribution of the inactive OTUs to the total abundance (i.e. the proportion of reads corresponding to inactive OTUs) differed between fractions

and was also significantly higher in the FL fraction (mean relative abundance: PA = 0.062, FL = 0.083; Wilcoxon-Mann Whitney test: W = 293, P-value = 0.004). The proportion of inactive OTUs (among those detected in the rDNA dataset) increased as the rank of the OTUs increased, implying that the less abundant OTUs were more likely to be inactive (Fig. 1B). The mean increase in the odds of an OTU being inactive for a one-unit increase in rank was 0.2% (1.6% - 0.37%). This relation was significant for the whole dataset and for 22 out of the 24 samples within the balanced dataset (Table S3). When inactive OTUs were defined as those with less rRNA reads than rDNA reads, this relation kept being significant for 19 samples (Table S3). The taxonomical composition of the active and inactive members of the communities did not greatly differ at the Phyla level: the relative abundance of the active and inactive members of each Phylum was correlated (Fig. 1C). Only some Phyla, such as Cyanobacteria, Spirochaetae or Parcubacteria, were overrepresented in the inactive pool. The Bray-Curtis distance between communities based on the inactive OTUs was consistently higher than the distance based on the active OTUs (Fig. 1D), i.e. "active" communities were more similar among them than "inactive" communities.

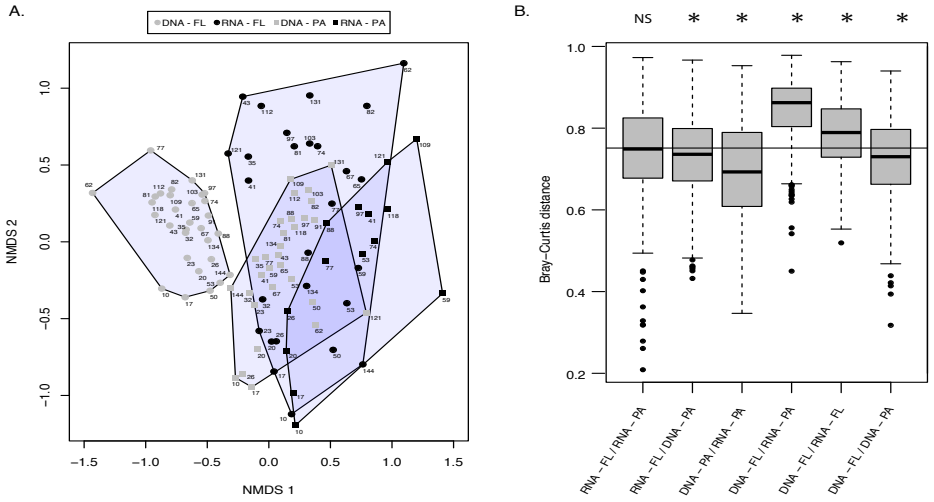


Figure 2: Bray-Curtis distance of rDNA and rRNA samples. A) Distance between samples visualized using Non-Metric Multidimensional Scaling (NMDS). Samples belonging to the particle-attached (PA) or free-living (FL) fraction and to the rDNA or rRNA samples are coded in the upper legend. The number close to each sample corresponds to the sampling station. B) Bray-Curtis distances between each combination of PA/FL and rDNA/rRNA sample. The horizontal line corresponds to the mean distance between samples. Significant diff erences from the mean ($P<0.05$) are labeled with an asterisk. Values above/below the mean indicate more/less different communities than the mean difference between pairs of samples.

The similarity of the 101 samples was described by placing these samples in a two dimensional space based on the NMDS (Fig. 2A). The rDNA samples formed two non-overlapping clusters corresponding to the FL and PA groups of samples. The rRNA samples, both the FL and PA, overlapped with the PA rDNA samples but never with the FL rDNA samples. Differences between these four groups of samples (the PA and FL in both the rDNA and rRNA datasets) were significant and explained 30% of the variance in community composition (Permutational MANOVA: F=13.583, R^2=0.295, P-value<0.001). Only the comparison between the FL communities and the active (i.e. rRNA) communities exceeded the mean difference between samples. The rest of comparisons resulted in

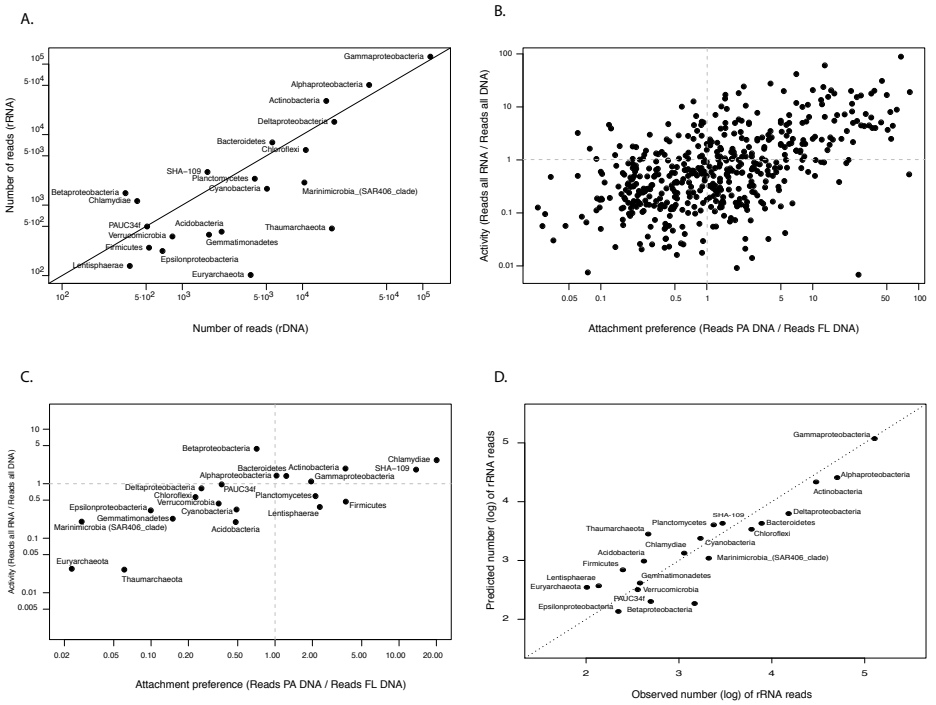


Figure 3: **Attachment preference and activity.** A) Abundance in the rDNA and rRNA datasets of the main Phyla. B) Attachment preference (reads in rDNA-PA / reads in rDNA-FL) and activity (reads in rRNA / reads in rDNA) of the OTUs. C) Attachment preference and activity of the main Phyla (i.e. ratios computed after adding up all the reads for each Phyla). D) Predicted vs. observed rRNA reads (log) of the main Phyla based on Model 2: *log(rRNA) = 1.015*log(rDNA) + 0.523*log(AttPref) – 0.22* (see details in the Table 2). The dotted line corresponds to the 1:1 line. Only the OTUs present in all the four combinations of PA/FL and rDNA/rRNA have been considered for B) and only the Phyla containing a total of more than 500 reads have been considered for A) and C). Proteobacteria have been split in its constituent Classes. Only the 12 balanced stations (i.e. for which both fractions and both the rDNA and rRNA samples do exist) have been used for C).

Table 1: Correlation test between the attachment preference (reads in rDNA-PA / reads in rDNA-FL) and activity (reads in rRNA / reads in rDNA) for the main Phyla in each of the balanced stations (i.e. for which both fractions and both the rDNA and rRNA data exist). Both variables have been log-transformed and only the Phyla containing a total of more than 500 reads have been considered. Significant P-values (<0.05) are in bold.

	r	t	d.f.	P-value
Station 10	0.377	1.68	17	0.1115
Station 17	0.634	3.28	16	**0.0047**
Station 20	0.598	2.89	15	0.0112
Station 26	0.623	3.19	16	**0.0057**
Station 41	0.123	0.53	18	0.6045
Station 53	0.641	3.45	17	**0.0031**
Station 59	0.554	2.74	17	0.0139
Station 74	0.526	2.55	17	**0.0208**
Station 77	0.594	3.13	18	**0.0058**
Station 88	0.646	3.49	17	**0.0028**
Station 97	0.777	5.08	17	**0.0001**
Station 121	0.127	0.51	16	0.6168

non-significant or significantly lower mean BC distances (Fig. 2B).

The abundance of each Phylum in the rDNA dataset tended to be positively related to its abundance in the rRNA dataset (Fig. 3A). However, a considerable variation existed, being the Archaea (both Thaumarchaeota and Euryarchaeota), Marinimicrobia and Acidobacteria overrepresented in the rDNA dataset and the SHA-109 clade, Chlamydiae, Actinobacteria and Betaproteobacteria overrepresented in the rRNA dataset. The attachment preference (reads in rDNA-PA / reads in rDNA-FL) was significantly correlated to the activity (reads in rRNA / reads in rDNA) for the OTUs (Fig. 3B; $r = 0.461$, P-value < 0.0001) and for the Phyla (Fig. 3C; $r = 0.459$, P-value = 0.03). This correlation was significant for 9 out of the 12 balanced stations for which the comparison was possible (all except station 10, 41 and 121; Table 1, Fig. S1). The simple linear model using the log of rDNA read counts of each Phylum for predicting its log rRNA read counts (Model 1 in Table 2) resulted in a significant relation between both variables explaining 55% of the variance. The addition of the log attachment preference as a second predictor improved the fit, both in terms of the AIC and the likelihood-ratio test. This model explained 78% of the variance (Fig. 3D and Model 2 in Table 2).

Table 2: Two linear models for the prediction of the logarithm of the rRNA read counts for the main Phyla (*rDNA*: rDNA read counts; *AttPref*: attachment preference). Only the Phyla containing at least 500 reads were included (see Fig. 3D).

	Model summary								Likelyhood ratio test		
	Predictor	Estimate	t	Beta	P-value	Adjusted R^2	AIC	N	log-likeli-hood	χ^2	P-value
Model 1	Intercept	0.061	0.096	-	0.9250	0.555	40.47	21	-9.398	15.677	**<0.0001**
	log(rDNA)	0.903	5.093	-	**0.0001**						
Model 2	Intercept	-0.220	-0.486	-	0.6326	0.777	26.80	21	-17.236		
	log(rDNA)	1.015	7.936	0.85	**<0.0001**						
	log(AttPref)	0.523	4.469	0.48	**0.0003**						

Discussion

While the analysis of rRNA and rDNA is being increasingly used for the description of active and dormant communities, several issues in the sampling, sample processing and data analysis need to be taken under consideration. The bathypelagic microbes sampled in here were in transit between *in situ* sampling and retrieval on board for 30' (1,000 m samples) to 2 h (4,000 m samples). Once on board, the samples were kept close to their original temperature (at 4º C) and were immediately filtered. Although we are aware that long sampling time may alter the original RNA pool from the community, this study is only focused on the 16S rRNA pool, which is know for its long preservation time, in the order of hours, compared to the half-lives of mRNA, in the order of minutes (Bernstein *et al.* 2004; Steglich *et al.* 2010). The volume of water filtered for rDNA and rRNA was different, what may alter the estimation of relative abundances, although this has been described for a range of volumes (0.05 to 5 L) different than the ones used in this study (Padilla *et al.* 2015). While high volumes of water (120 L) were chosen for the DNA samples due to the low cell abundances in the bathypelagic, this was not applied for the RNA samples (12 L were filtered) in order to avoid biases due to longer sampling times. Additionally, the bias described, i.e. an overrepresentation of PA taxa in high-volume samples (Padilla *et al.* 2015), does not fit our results as it consists in the exact opposite pattern as the one found in here: higher abundance of PA taxa in the RNA samples, which corresponds to the low-volume sample set. Differences in the extraction protocols for both DNA and RNA may also introduce biases in the relative abundance of specific taxa within each set of samples (Tsementzi *et al.* 2014; McCarthy *et al.* 2015). However, although minor effects on the relative abundances cannot be ruled out, the potential biases described above do not seem

to be relevant compared to the natural variation between PA and FL communities. If this were the case, the ordination of samples based on their taxonomical composition would separate the RNA and DNA datasets, instead of grouping communities from both DNA and RNA datasets together (Fig. 2A).

The delineation of active and dormant OTUs has been previously addressed by setting an arbitrary minimum cutoff for the rRNA:rDNA ratio (Jones & Lennon 2010; Campbell *et al.* 2011; Lennon & Jones 2011; Hugoni *et al.* 2013) or a set of cutoffs (Aanderud *et al.* 2016; Kearns *et al.* 2016). However, dormant cells may still contain detectable amounts of rRNA (Chambon *et al.* 1968). Also, some microorganisms can contain in some cases significantly more rRNA in dormancy than in a vegetative state (Sukenik *et al.* 2012). In addition to that, the detectability of an OTU when using high-throughput sequencing is affected by both its relative abundance and the sequencing effort used, rarely sufficient to saturate the complete richness in natural microbial communities. The change in the rRNA content with changing metabolic state in prokaryotes, for which we still lack a general understanding, has been suggested to impact the classification of inactive populations (Steven *et al.* 2017). Given this, we adopted a conservative definition for an OTU being inactive. We considered as inactive every OTU detected in one or more rDNA sample but never detected in any of the rRNA samples. Additionally, the considerably higher number of sequences in the rRNA dataset compared to the rDNA dataset (5-fold) should minimize the chance of overestimating the number of inactive OTUs due to a lack of detectability in the rRNA dataset. Even with such a restrictive definition, about 20.5 - 37.8% of the OTUs were found to be inactive among the samples (range depending on whether we used the total number of OTUs or only those detected in the rDNA samples for the calculation). The proportion of inactive OTUs was higher in the FL samples than in the PA samples, however, their contribution in terms of reads was low in both cases (6.2% and 8.3% respectively of the total number of reads). This indicates that inactive OTUs correspond to rare (i.e. not abundant) OTUs within the bathypelagic communities. In fact, the odds of an OTU being inactive significantly increased with its rank, i.e. as its abundance decreased (Fig. 1B, Table S3). This pattern was maintained when, for comparative reasons, inactive OTUs were defined as in previous studies (Jones & Lennon 2010), i.e. as those with a lower number of rRNA reads than rDNA reads (Table S3). This contrasts with the inverse relationship that has been described for lakes (Jones & Lennon 2010) and surface ocean communities (Campbell *et al.* 2011; Zhang *et al.* 2014) where the rare OTUs appeared to be proportionally more active than the abundant OTUs. This observation was interpreted as the result of the dynamic nature of surface communities in which, probably due to the relatively rapid changing of environmental conditions, OTUs

can easily transit from rare to abundant and vice versa. Thus, the rare biosphere of surface waters would constitute a "seed bank" from which inactive diversity could be recurrently recruited (Lennon & Jones 2011). Our results, however, suggest that the deep oceanic communities behave differently and contain a considerable proportion of rare OTUs not detectably active in any of the globally distributed samples analyzed here.

The taxonomical composition of the active and inactive members of the communities did not greatly differ at the Phyla level (Fig. 1C), with some exceptions such as the overrepresentation of Cyanobacteria, Spirochaetae or Parcubacteria in the inactive pool. The phototrophic or anaerobic nature of these Phyla would suggest that at least a portion of this inactive pool is composed of allochthonous organisms subject to an unfavorable environment, which would explain their inactive state in the deep aphotic and well-oxygenated waters. This pool of OTUs would not be thus subjected to environmental filtering, which is consistent with the more variable composition of the inactive pool compared to the active pool (Fig. 1D). Thus, the bathypelagic prokaryotic communities seem to contain a less dynamic rare biosphere compared to what has been described for surface waters: a considerable proportion of its members seem to be inactive under the whole spectrum of environmental conditions sampled in this study, suggesting that a fraction of the bathypelagic "seed bank" is hardly recruited to become abundant members of their communities. The less dynamic nature of the bathypelagic seed bank may be caused by the more stable environmental conditions of the deep ocean compared to the surface ocean, where the seasonality is extremely important for the composition of microbial communities (Brown *et al.* 2005; Fuhrman *et al.* 2006; Giovannoni & Vergin 2012). Additionally, the strict definition used here for an inactive OTU, especially their absence in all the rRNA samples across different oceans, and the likely allochthonous nature of some of its components lead us to hypothesize that, at least a fraction, of the inactive pool of OTUs may correspond to either "relic DNA", i.e. DNA from dead cells (Carini *et al.* 2016), or to "deceased" cells, i.e. cells not measurably dividing or metabolizing or cells incapable of become metabolically active in the future but from which intact macromolecules still persist (Blazewicz *et al.* 2013). A better characterization of the metabolic state of these members would be next logical step to better understand the rare biosphere of the deep ocean.

Recent empirical evidence has been presented in favor of the "dilution hypothesis", which states that most of the DOC pool in the deep ocean is labile but can not be used by prokaryotes at the very low concentrations at which individual compounds in that pool are found, which are below the levels matching the energetic investment required for their uptake and degradation (Jannasch 1967, 1994). Dilution rather than recalcitrance has

been shown to preclude consumption of a substantial fraction of DOC by the bathypelagic free-living prokaryotes (Arrieta *et al.* 2015). In contrast, particle-attached prokaryotes are likely to have access to a more concentrated organic matter pool (Minor *et al.* 2003). As a consequence, PA prokaryotes would exhibit a more active metabolic state or higher growth rates, which would be translated into higher proportions of rRNA molecules per rDNA gene. This constitutes one of the hypotheses that we test in the present work by jointly studying the rRNA and rDNA of FL and PA bathypelagic prokaryotic communities. The four subsets of samples defined by the targeted molecule (rDNA or rRNA) and the size-fraction analyzed (FL or PA), statistically differed in their OTU composition. However, the rRNA samples (both the FL and PA) clustered together with the PA-rDNA set of samples (Fig. 2A) and apart from the FL-rDNA samples. This suggests that active members of the bathypelagic communities are more abundant in the PA than in the FL communities, supporting the hypothesis tested here. When comparing the rDNA and rRNA sequences from the same samples (Jones & Lennon 2010; Campbell *et al.* 2011; Hunt *et al.* 2013), the quotient of the reads recruited by an OTU in the rRNA and in the rDNA sample (i.e. the rRNA:rDNA ratio) is generally used as a proxy for activity or growth rate, although some limitations to this approach have been identified (Blazewicz *et al.* 2013) and are discussed below. When computing these values, a considerable degree of variability existed in the rRNA:rDNA ratio of the OTUs and Phyla, spanning four and three orders of magnitude, respectively. Almost half of this variability for the OTUs could be explained by their attachment preference (Fig. 3B), indicating that the activity of the bathypelagic prokaryotes is determined to a great extent by their trophic strategy, that is, by their preference on living attached to organic particles or freely in the water surrounding these particles. Coherently, the number of rRNA reads for the main Phyla could be predicted with a considerably precision ($R^2 = 0.78$, $N = 21$) using a linear model that included both the log of the number of rDNA reads and its attachment preference (Fig. S2, Table 2). Marine Archaea, both Euryarchaeota and Thaumarchaeota, and Marinimicrobia (SAR406 clade) were the phyla with a lower rRNA:rDNA ratio and some of the main FL members of the bathypelagic communities. Archaea, in fact, have been proven to be able to grow autotrophically or through the incorporation of simple organic compounds such as amino acids or urea (Ouverney & Fuhrman 2000; Könneke *et al.* 2005; Swan *et al.* 2014). On the other hand, the prokaryotes with higher attachment preference corresponded to Actinobacteria, SHA-109, for which no metabolic information exists, or Chlamydiae, presumably intracellular symbionts/pathogens of eukaryotic hosts and whose importance in marine environments has only recently been noted (Lagkouvardos *et al.* 2014; Viana & Buchrieser 2016). In fact, the presence in the PA samples of some

endosymbiotic, mutualistic or parasitic prokaryotes associated to small protists (<20 μm) cannot be discarded (Salazar *et al.* 2015). In contrast, Betaproteobacteria exhibited the highest rRNA:rDNA ratio although they did not have a clear attachment preference.

The use of the rRNA and rDNA to characterize the growth state of bacteria is supported by the experimental evidence of a correlation between the rRNA content and growth rate of bacteria in pure culture conditions (Kjeldgaard *et al.* 1958; Harvey 1970; Poulsen *et al.* 1993; Pang & Winkler 1994; Lankiewicz *et al.* 2015) and between the rRNA:rDNA ratio and the uptake of ^{3}H-leucine by SAR11 (Salter *et al.* 2015). The use of the rRNA:rDNA ratio as a proxy for growth rate in community-level studies may also be affected by the fact that this relationship varies among taxa (Kemp *et al.* 1993). A higher ribosomal efficiency of marine oligotrophs (based on two cultures of SAR11 and SAR92) compared to previously studied copiotrophs has been proposed as an explanation for such variation in marine bacteria (Lankiewicz *et al.* 2015). That is, oligotrophs would need a relatively lower number of ribosomes per cell in order to sustain a certain growth rate, compared to copiotrophs. The strong relationship between the attachment preference of the OTUs or Phyla and their rRNA:rDNA ratios found here is consistent for all the prokaryotic taxa inhabiting the bathypelagic ocean and across the global-scale of the study. This implies a positive relationship between the particle-related lifestyle of bathypelagic prokaryotes and their growth rate. However a negative relationship between the attachment preference and the ribosomal efficiency of bathypelagic prokaryotes may not be discarded as an alternative explanation, although no additional evidences exist to sustain it. The PA and FL life strategies have been proven to be highly conserved through the prokaryotic life history and thus transitions from one lifestyle to the opposite have been rare at evolutionary timescales (Salazar *et al.* 2015). Here we propose that these two lifestyles also define two distinct growth strategies, as shown by the divergent number of rRNA molecules per rDNA molecule. FL prokaryotes, using a highly diluted pool of diverse dissolved organic substrates, exhibit low rRNA:rDNA ratios indicative of low growth rates (or alternatively, higher ribosomal efficiencies), while PA prokaryotes would use concentrated pools of particulate organic matter, resulting in higher growth rates indicated by higher rRNA:rDNA ratios. Future studies will be needed to characterize the functional and metabolic basis of these two life strategies for a better understanding of the main players in the biogeochemical cycles of the deep ocean.

References

Aanderud ZT, Vert JC, Lennon JT *et al.* (2016) Hypersaline lakes harbor more active bacterial communities. *PeerJ Preprints 4:e1922v1*.

Acinas SG, Antón J, Rodríguez-Valera F (1999) Diversity of free-living and attached bacteria in offshore western Mediterranean waters as depicted by analysis of genes encoding 16S rRNA. *Applied and Environmental Microbiology*, **65**:514–522.

Agogué H, Lamy D, Neal PR, Sogin ML, Herndl GJ (2011) Water mass-specificity of bacterial communities in the North Atlantic revealed by massively parallel sequencing. *Molecular Ecology*, **20**:258–274.

Anderson MJ (2001) A new method for non-parametric multivariate analysis of variance. *Austral Ecology*, **26**:32–46.

Apprill a, McNally S, Parsons R, Weber L (2015) Minor revision to V4 region SSU rRNA 806R gene primer greatly increases detection of SAR11 bacterioplankton. *Aquatic Microbial Ecology*, **75**:129–137.

Arístegui J, Gasol JM, Duarte CM, Herndl GJ (2009) Microbial oceanography of the dark ocean's pelagic realm. *Limnology and Oceanography*, **54**:1501–1529.

Arrieta JM, Mayol E, Hansman RL *et al.* (2015) Dilution limits dissolved organic carbon utilization in the deep ocean. *Science*, **348**:331–333.

Bernstein JA, Lin P-H, Cohen SN, Lin-Chao S (2004) Global analysis of Escherichia coli RNA degradosome function using DNA microarrays. *Proceedings of the National Academy of Sciences USA*, **101**:2758–2763.

Blazewicz SJ, Barnard RL, Daly RA, Firestone MK (2013) Evaluating rRNA as an indicator of microbial activity in environmental communities: limitations and uses. *The ISME Journal*, **7**:2061–2078.

Brown M V, Philip GK, Bunge JA *et al.* (2009) Microbial community structure in the North Pacific ocean. *The ISME Journal*, **3**:1374–86.

Brown M V, Schwalbach MS, Hewson I, Fuhrman JA (2005) Coupling 16S-ITS rDNA clone libraries and automated ribosomal intergenic spacer analysis to show marine microbial diversity: development and application to a time series. *Environmental microbiology*, **7**:1466–79.

Campbell BJ, Kirchman DL (2012) Bacterial diversity, community structure and potential growth rates along an estuarine salinity gradient. *The ISME Journal*, **7**, 210–220.

Campbell BJ, Yu L, Heidelberg JF, Kirchman DL (2011) Activity of abundant and rare bacteria in a coastal ocean. *Proceedings of the National Academy of Sciences USA*, **108**:12776–12781.

Campbell BJ, Yu L, Straza TRA, Kirchman DL (2009) Temporal changes in bacterial rRNA and rRNA genes in Delaware (USA) coastal waters. *Aquatic Microbial Ecology*, **57**:123–135.

Caporaso JG, Lauber CL, Walters W a *et al.* (2011) Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proceedings of the National Academy USA*, **108:**4516–22.

Carini P, Marsden PJ, Leff JW *et al.* (2016) Relic DNA is abundant in soil and obscures estimates of soil microbial diversity. *Nature Microbiology*, **2**:16242.

Chambon P, DuPraw EJ, Kornberg A (1968) Biochemical studies of bacterial sporulation and germination. IX. Ribonucleic acid and deoxyribonucleic acid polymerases in nuclear fractions of vegetative cells and spores of Bacillus megaterium. *Journal of Biological Chemistry*, **243**:5101–5109.

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**:541–52.

DeLong EF, Preston CM, Mincer T *et al.* (2006) Community genomics among stratified microbial assemblages in the ocean's interior. *Science*, **311**:496–503.

Denef VJ, Fujimoto M, Berry MA, Schmidt ML (2016) Seasonal succession leads to habitat-dependent differentiation in ribosomal RNA:DNA ratios among freshwater lake bacteria. *Frontiers in Microbiology*, **7**.

Eloe EA, Shulse CN, Fadrosh DW *et al.* (2011) Compositional differences in particle-associated and free-living microbial assemblages from an extreme deep-ocean environment. *Environmental Microbiology Reports*, **3**:449–58.

Fuhrman JA, Hewson I, Schwalbach MS *et al.* (2006) Annually reoccurring bacterial communities are predictable from ocean conditions. *Proceedings of the National Academy of Sciences USA*, **103**:13104–9.

Galand PE, Potvin M, Casamayor EO, Lovejoy C (2010) Hydrography shapes bacterial biogeography of the deep Arctic Ocean. *The ISME Journal*, **4**:564–76.

Ganesh S, Parris DJ, DeLong EF, Stewart FJ (2014) Metagenomic analysis of size-fractionated picoplankton in a marine oxygen minimum zone. *The ISME Journal*, **8**:187–211.

Ghiglione J-F, Conan P, Pujo-Pay M (2009) Diversity of total and active free-living vs. particle-attached bacteria in the euphotic zone of the NW Mediterranean Sea. *FEMS Microbiology Letters*, **299**:9–21.

Giovannoni SJ, Vergin KL (2012) Seasonality in ocean microbial communities. *Science*, **335**:671–676.

Harvey RJ (1970) Metabolic regulation in glucose-limited chemostat cultures of *Escherichia coli*. *J. Bacteriol.*, **104**:698–706.

Herndl GJ, Reinthaler T (2013) Microbial control of the dark end of the biological pump. *Nature Geoscience*, **6**:718–724.

Hugoni M, Taib N, Debroas D *et al.* (2013) Structure of the rare archaeal biosphere and seasonal dynamics of active ecotypes in surface coastal waters. *Proceedings of the National Academy of Sciences USA*, **110**:6004–9.

Hunt DE, Lin Y, Church MJ *et al.* (2013) Relationship between abundance and specific activity of bacterioplankton in open ocean surface waters. *Applied and Environmental Microbiology*, **79**:177–184.

Jannasch HW (1967) Growth of marine bacteria at limiting concentrations of organic carbon in seawater. *Limnology and Oceanography*, **12**:264–271.

Jannasch HW (1994) The microbial turnover of carbon in the deep-sea environment. *Global and Planetary Change*, **9**:289–295.

Jones SE, Lennon JT (2010) Dormancy contributes to the maintenance of microbial diversity. *Proceedings of the National Academy of Sciences USA*, **107**:5881–6.

Kearns PJ, Angell JH, Howard EM *et al.* (2016) Nutrient enrichment induces dormancy and decreases diversity of active bacteria in salt marsh sediments. *Nature Communications*, **7**:12881.

Kemp PF, Lee S, Laroche J (1993) Estimating the growth rate of slowly growing marine bacteria from RNA content. *Applied Environmental Microbiology*, **59**:2594–2601.

Kjeldgaard NO, MaalOe O, Schaechter M (1958) The transition between different physiological states during balanced growth of Salmonella typhimurium. *Journal of General Microbiology*, **19**:607–616.

Könneke M, Bernhard AE, de la Torre JR *et al.* (2005) Isolation of an autotrophic ammonia-oxidizing marine archaeon. *Nature*, **437**:543–546.

Lagkouvardos I, Weinmaier T, Lauro FM *et al.* (2014) Integrating metagenomic and amplicon databases to resolve the phylogenetic and ecological diversity of the *Chlamydiae*. *The ISME Journal*, **8**:115–125.

Lankiewicz TS, Cottrell MT, Kirchman DL (2015) Growth rates and rRNA content of four marine bacteria in pure cultures and in the Delaware estuary. *The ISME Journal*, **10**:1–10.

Lennon JT, Jones SE (2011) Microbial seed banks: the ecological and evolutionary implications of dormancy. *Nature Reviews Microbiology*, **9**:119–30.

Logares R, Sunagawa S, Salazar G *et al.* (2013) Metagenomic 16S rDNA Illumina tags are a powerful alternative to amplicon sequencing to explore diversity and structure of

microbial communities. *Environmental Microbiology*, **16**:2659–71.

Martín-Cuadrado A-B, López-García P, Alba J-C *et al.* (2007) Metagenomics of the deep Mediterranean, a warm bathypelagic habitat. *PloS one*, **2**Oe914.

McArdle BH, Anderson MJ (2001) Fitting multivariate models to community data: a comment on distance-based redundancy analysis. *Ecology*, **82**:290–297.

McCarthy A, Chiang E, Schmidt ML, Denef VJ (2015) RNA preservation agents and nucleic acid extraction method bias perceived bacterial community composition. *PLOS ONE*, **10**:e0121659.

Minchin PR (1987) An evaluation of the relative robustness of techniques for ecological ordination. *Vegetatio*, **69**:89–107.

Minor EC, Wakeham SG, Lee C (2003) Changes in the molecular-level characteristics of sinking marine particles with water column depth. *Geochimica et Cosmochimica Acta*, **67**:4277–4288.

Oksanen J, Blanchet FG, Kindt R *et al.* (2015) vegan: Community Ecology Package.

Ouverney CC, Fuhrman JA (2000) Marine planktonic Archaea take up amino acids. *Applied and Environmental Microbiology*, **66**:4829–4833.

Padilla CC, Ganesh S, Gantt S *et al.* (2015) Standard filtration practices may significantly distort planktonic microbial diversity estimates. *Frontiers in Microbiology*, **6**:547.

Pang H, Winkler HH (1994) The concentrations of stable RNA and ribosomes in Rickettsia prowazekii. *Molecular Microbiology*, **12**:115–120.

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

Quaiser A, Zivanovic Y, Moreira D, López-García P (2011) Comparative metagenomics of bathypelagic plankton and bottom sediment from the Sea of Marmara. *The ISME Journal*, **5**:285–304.

R Core Team (2016) R: A language and environment for statistical computing.

Salazar G, Cornejo-Castillo FM, Benítez-Barrios V *et al.* (2016) Global diversity and biogeography of deep-sea pelagic prokaryotes. *The ISME Journal*, **10**:596–608.

Salazar G, Cornejo-Castillo FM, Borrull E *et al.* (2015) Particle-association lifestyle is a phylogenetically conserved trait in bathypelagic prokaryotes. *Molecular Ecology*, **24**:5692–5706.

Salter I, Galand PE, Fagervold SK *et al.* (2015) Seasonal dynamics of active SAR11 ecotypes in the oligotrophic Northwest Mediterranean Sea. *The ISME Journal*, **9**:347–360.

Smedile F, Messina E, La Cono V *et al.* (2012) Metagenomic analysis of hadopelagic microbial assemblages thriving at the deepest part of Mediterranean Sea, Matapan-Vavilov

Deep. *Environmental Microbiology*, **15**:167–82.

Steglich C, Lindell D, Futschik M *et al.* (2010) Short RNA half-lives in the slow-growing marine cyanobacterium *Prochlorococcus*. *Genome Biology*, **11**:R54.

Steven B, Hesse C, Soghigian J, Gallegos-Graves LV, Dunbar J (2017) Simulated rRNA/DNA ratios show potential to misclassify active populations as dormant. *Applied and Environmental Microbiology*, **83**:e00696-17.

Sukenik A, Kaplan-Levy RN, Welch JM, Post AF (2012) Massive multiplication of genome and ribosomes in dormant cells (akinetes) of *Aphanizomenon ovalisporum* (Cyanobacteria). *The ISME Journal*, **6**:670–679.

Swan BK, Chaffin MD, Martinez-Garcia M *et al.* (2014) Genomic and metabolic diversity of Marine Group I Thaumarchaeota in the mesopelagic of two subtropical gyres. *PLoS ONE*, **9**:e95380.

Tsementzi D, Poretsky R, Rodriguez-R LM, Luo C, Konstantinidis KT (2014) Evaluation of metatranscriptomic protocols and application to the study of freshwater microbial communities. *Environmental Microbiology Reports*, **6**:640–655.

Viana F, Buchrieser C (2016) Environmental treasures: co-isolation of the first marine *Chlamydiae* and its protozoan host. *Environmental Microbiology*, **18**:2295–2297.

Wilkins D, van Sebille E, Rintoul SR, Lauro FM, Cavicchioli R (2013) Advection shapes Southern Ocean microbial assemblages independent of distance and environment effects. *Nature communications*, **4**:2457.

Zhang Y, Zhao Z, Dai M, Jiao N, Herndl GJ (2014) Drivers shaping the diversity and biogeography of total and active bacterial communities in the South China Sea. *Molecular Ecology*, **23**:2260–2274.

mtagger: an R package for the characterization of microbial communities through rDNA metagenomic fragments

Chapter IV

Abstract

The accelerated decrease in high throughput sequencing costs has allowed the developement of a variety of methods for the taxonomical characterization of microbial communities through the use of metagenomic reads of ribosomal genes (rDNA). Among them, the miTags approach (Logares *et al.* 2013) has proven to be a valid alternative to the amplicon sequencing of the 16S/18S gene. This approach is based on the extraction of rDNA fragments directly from the metagenomic reads and their binning into operational taxonomic units (OTUs) through a similarity search against a full-length rDNA reference database. However a user-friendly and reproducible pipeline for this method has not been provided yet. Here we make available the *mtagger* R package, which incorporates the originally proposed miTag method and a modification (the "Strict" method version) that consists on the exclusion from the analysis of the reads that could not be unambiguously assigned to a single reference sequence and keep them for an alternative last common ancestor (LCA) assignation. We evaluate the performance of both the Original and the Strict miTags method with mock and natural communities. The newly Strict proposed method outperformed the original one in mock communities by drastically reducing the binning process error rate and provided much more accurate estimates of OTU richness. Both methods resulted in correlated estimates of richness and community similarity when used with a real dataset of 139 marine microbial metagenomes. We also identify the heuristic nature of the similarity search as a secondary source of errors in the binning process and provide tools and recomend parameter's choice for the correct estimation of microbial community composition from metagenomic data.

Introduction

The information contained in the metagenomic reads corresponding to ribosomal genes (rDNA) has been proposed as an alternative to the commonly used amplicon sequencing of the 16S/18S rRNA gene for the characterization of microbial communities (Logares *et al.* 2013). Different methods have been developed for this last purpose, such as MEGAN (Huson *et al.* 2007), PhylOTU (Sharpton *et al.* 2011), miTags (Logares *et al.* 2013), PhyloSift (Darling *et al.* 2014) and SSUsearch (Guo *et al.* 2016). Among them, the miTags approach has been proven to circumvent biases introduced by the use of the PCR primers (Logares *et al.* 2013) and has also been satisfactorily applied to large metagenomic datasets (Lima-Mendez *et al.* 2015; Sunagawa *et al.* 2015). However, a user-friendly and reproducible pipeline for this method is lacking. Moreover, although it was tested against the amplicon sequencing of the 16S rRNA gene of the same natural communities (Logares *et al.* 2013), a thorough test has never been conducted with synthetic communities, for which the diversity and structure are known and thus, a quantitative evaluation of the method performance is possible.

The miTags approach is based on the extraction of the fraction of metagenomic reads that correspond to rDNA genes through hidden markov models (HMM) and its binning into operational taxonomic units (OTUs) by performing a similarity search of each read against a rDNA reference database. Thus, it is a *reference-based* approach, in opposition to methods that are based on a *de novo* clustering of overlapping sequences, such as the ones used with amplicon sequencing. The miTag approach, thus, allows the exploitation of the whole variability within the rDNA gene for the characterization of microbial communities, instead of restricting to a specific variable region, thus maximizing the useful reads contained in the metagenome. However, the heterogeneity in the sequence variability of the ribosomal genes, which are organized into conserved and variable regions (Van de Peer *et al.* 1996), may be problematic for the correct assignation of the reads to closely related reference sequences. Additionally, the heuristic nature of the similarity search algorithms may also introduce errors in the binning of reads into OTUs.

Here we perform a test of the original miTags approach in order to quantitatively evaluate its performance against synthetic communities of known composition. A modified pipeline is proposed and tested, which outperformed the original one in terms of its error rate in the binning process. We also make available the *mtagger* R package, which incorporates the original and the modified miTags method, and recommend the set of acceptable parameters for a correct characterization of microbial communities from metagenomic data using this method.

Material and Methods

We developed an R package (available at https://github.com/GuillemSalazar/mtagger) for the extraction of ribosomal fragments from metagenomic datasets and their assignation to reference OTUs and higher taxonomic levels. The package depends on the execution of two external softwares, HMMER (Eddy 2011) and USEARCH (Edgar 2010). The proposed pipeline consists on the sequential execution of 7 functions implemented within the package that perform 4 main steps: i) the download and processing of the reference rDNA database, ii) the extraction of rDNA fragments from the metagenomic reads, iii) the mapping step, i.e. the assignation of the rDNA fragments to reference OTUs and iv) the construction of abundance tables. The database is downloaded from the SILVA source site (https://www.arb-silva.de/download/archive/) and is stored in the computer with the '*mtagger.db*' function. The database we suggest to use as a reference is the SSU Ref Nr99 SILVA database (Pruesse *et al.* 2007) although other databases may be used. A taxonomy file that links the accession number for each sequence to its taxonomical assignation is built with the function '*mtagger.buildtaxfile*'. The reference database is reduced by clustering it to the desired similarity cutoff (97% is implemented as the standard but this can be varied) by executing USEARCH (Edgar 2010) through the '*mtagger.clusterdb*' function. This clustered database will serve as the reference against each rDNA fragment will be searched, i.e. each sequence in this database will be considered an OTU to which the rDNA fragments will be assigned by performing a similarity search. The extraction of rDNA fragments from metagenomic read's files is performed by executing HMMER (Eddy 2011) with the function '*mtagger.extract.all*'. The extracted rDNA fragments are mapped to the reference database by executing USEARCH through the function '*mtagger.mapping.all*'. This mapping may be performed in two modes: the "original" and the "strict" mode. The original mode corresponds to the previously published "miTags approach" (Logares *et al.* 2013) where each rDNA fragment is searched against the reference database and the best hit satisfying the similarity cutoff is used to assign the fragment to a reference OTU. The strict mode, introduced in this work, will only assign to a reference OTU those rDNA fragments with unambiguous hits. An rDNA fragment having an unambiguous hit is defined as the one for which only one hit satisfying the similarity cutoff is found, i.e. an rDNA fragment that can be unambiguously assigned to and only to one single reference sequence. The rDNA fragments with more than one hit will be discarded for the assignation at the OTU level, but will be used for the assignation to higher taxonomical levels through a least-common-ancestor (LCA) approach. This is performed by scanning all the hits for a single rDNA fragment in order to find the lower taxonomical

rank common to all the hits. The SILVA tax_slv file, which contains the taxonomic rank designations of the database, is used for the LCA approach. Finally, abundance tables containing the number of rDNA fragments that could be assigned at the OTU level and at all the possible higher levels are built for each sample with the '*mtagger.table.all*' function. A merged table for a collection of samples analyzed through the previously described steps may be constructed with the '*mtagger.merge.tables*' function.

The performance of the two pipelines implemented within the mtagger package (the original and strict mode) was tested with in-silico created mock communities. Mock communities were simulated by randomly selecting 100-200 bp fragments of sequences from the SILVA SSU Ref Nr99 v.123 database (after a pre-clustering at 97% similarity). The use of the same database for the construction and analysis of mock communities allowed to know the sequence from which each fragment was generated and thus evaluate the error rate of the two pipelines. Two sets of mock communities were generated (Fig. 1). The first set (named the *uniform* dataset hereafter) consisted on 100 mock communities (with 100,000 rDNA fragments each), generated by randomly sampling 10,000 sequences from the SILVA database and 10 randomly selected fragments per sequence. This sampling was performed using a uniform distribution. The second set of mock communities

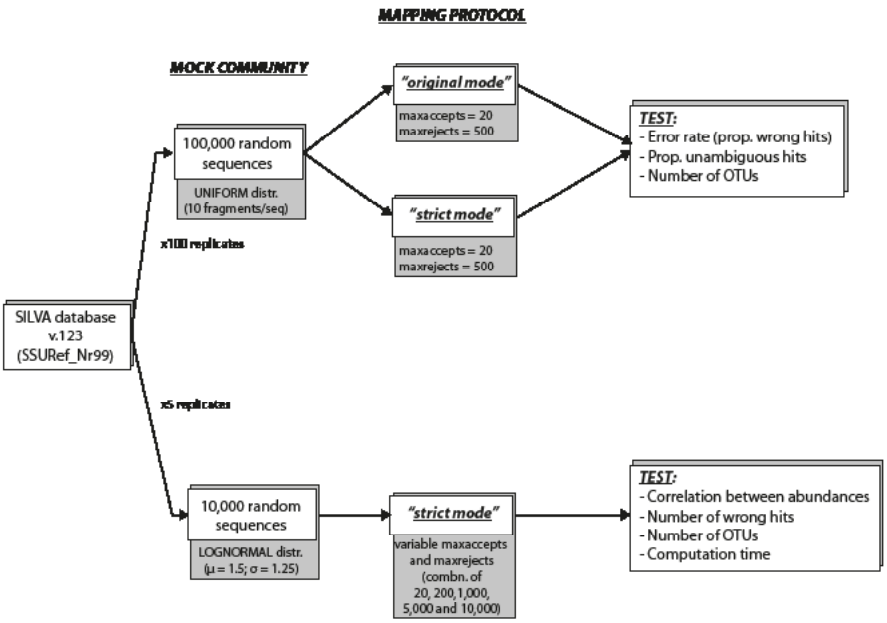


Figure 1: **Schematic representation of the process of evaluation of the original (Original) and modified (Strict) miTags approach with mock communities.**

(named the *lognormal* dataset) consisted on 5 mock communities with 100,000 rDNA fragments each, that were generated by sampling the same SILVA database by using a log-normal distribution ($\mu = 1.5$, $\sigma = 1.25$). This was done in order to simulate communities with realistic abundance distributions. The two sets of mock communities were generated twice using fragments random of 100 bp and 200 bp length. The mock communities were generated with two R functions ('*fasta.sample*' and '*fasta.cutter*') created for that purpose, which are available as part of the *FastaUtils* package (https://github.com/GuillemSalazar/FastaUtils).

The mock communities in the uniform dataset were analyzed with the mtagger R package and abundances at the OTU level were estimated with both the Original and Strict mode with the parameter combination *maxaccept* = 20 and *maxreject* = 500. The error rate (i.e. the proportion of rDNA fragments assigned to a sequence different than the one they were generated from), the proportion of unambiguous hits (for the Strict mode) and the number of OTUs were computed for each community. The lognormal dataset was used to evaluate the performance of the Strict method when the heuristic nature of the USEARCH algorithm is varied. This was done by testing different values of the two parameters that control its heuristic behavior: *maxaccepts* and *maxrejects*. The 5 mock communities were analyzed under the Strict mode using all the combinations of the values 20, 200, 1000, 5000 and 10000 for these two parameters. The number of wrong hits and the number of real and estimated OTUs were computed for each community.

The Original and Strict mode of the mtagger approach were also compared with a real dataset containing 16S rDNA reads from 139 metagenomic samples of marine planktonic prokaryotes from the *Tara* Ocean Expedition (Sunagawa *et al.* 2015) which are publically available (http://ocean-microbiome.embl.de/companion.html). The similarity between samples was compared for the two modes by computing the Bray-Curtis dissimilarity (Bray & Curtis 1957) and tested for correlation with a Mantel test.

Results

When *mtagger* was run using the uniform set of mock communities and the parameter combination *maxaccept* = 20 and *maxreject* = 500, the mean error rate (the proportion of reads in a mock community that was assignated to an erroneous reference sequence) of the original pipeline was of 0.468 and 0.285 for mock communities of 100 bp and 200 bp, respectively. This error rate was unevenly distributed along the 16S rRNA gene position (Fig. 2). It tended to accumulate in some conserved regions (between the variable region V2-V3 and V8-V9) and was low in the V4 and V6 regions. The mean error rate was dramatically lowered to 0.017 and 0.032 respectively when the strict mode was used (Table

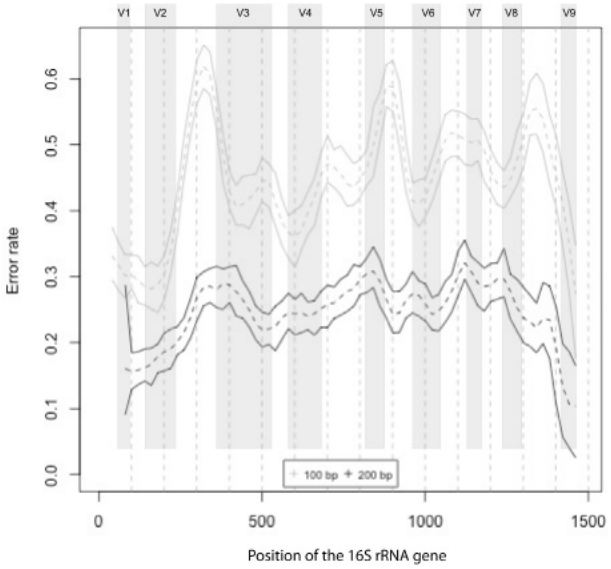


Figure 2: **Error rate of the original miTags approach through the 16S rDNA nucleotidic position. The error rate is compu**ted as the proportion of reads in a mock community that was assignated to an erroneous reference sequence (i.e. a sequence different to the one the synthetic read originated from). The mean (dashed line) and minimim and maximum values (solid lines) are represented for 100 mock communities composed of 100,000 reads of 100 bp (gray) and 200 bp (black) length. Variable regions (V1-V9) are indicated as gray columns.

Table 1: Error rate and estimated number of OTUs for 100 mock communities for the original and strict miTags method.

			mean	**min**	**max**
Error rate	100 bp	*original mode*	0.468	0.46	0.475
		strict mode	0.017	0.015	0.018
	200 bp	*original mode*	0.285	0.278	0.291
		strict mode	0.032	0.03	0.034
Number of OTUs	100 bp	*original mode*	38231.2	37883	38499
		strict mode	9773.8	9704	9864
	200 bp	*original mode*	26835.9	26503	27155
		strict mode	11271.6	11182	11366

The error rate, i.e. the proportion of synthetic reads that were not correctly assigned to its reference sequence, and the total number of estimated OTUs for 100 mock communities of 100bp and 200bp read's length estimated through the original and strict miTags methods. The real number of OTUs in all mock communities was 10,000 (see Material and Methods).

Table 2: Effect of the heuristic nature of the binning process for 5 lognormal mock communities.

		100 bp rDNA fragments				200 bp rDNA fragments			
maxaccepts	maxrejects	Real num. of otus	Estimated num. of otus	Erroneous mappings	r	Real num. of otus	Estimated num. of otus	Erroneous mappings	r
20	20	288	361.6	147.4	0.8802	289.4	447.2	197.4	0.9142
20	200	288	361.6	147.4	0.8802	289.4	447.2	197.4	0.9142
20	2000	288	361.6	147.4	0.8802	289.4	447.2	197.4	0.9142
20	1000	288	359.6	146.2	0.8804	289.4	446.2	196.6	0.914
20	5000	288	360	146.2	0.8808	289.4	446.4	196.6	0.9142
20	10000	288	359.6	146.2	0.8804	289.4	446.2	196.6	0.914
200	20	288	234.8	19.8	0.8782	289.4	272	19.4	0.922
200	200	288	234.8	19.8	0.8782	289.4	272	19.4	0.922
200	2000	288	234.8	19.8	0.8782	289.4	272	19.4	0.922
200	5000	288	232.2	18	0.8778	289.4	270.4	17.8	0.9216
200	1000	288	231.6	17.8	0.877	289.4	269.8	17.4	0.9212
200	10000	288	231.6	17.8	0.877	289.4	269.8	17.4	0.9212
1000	20	288	220.6	5.8	0.8778	289.4	256.6	3.4	0.9256
1000	200	288	220.6	5.8	0.8778	289.4	256.6	3.4	0.9256
1000	2000	288	220.6	5.8	0.8778	289.4	256.6	3.4	0.9256
1000	5000	288	218.6	4.6	0.877	289.4	255	1.8	0.9248
2000	20	288	219.2	4.2	0.8778	289.4	255.4	2.4	0.9254
2000	200	288	219.2	4.2	0.8778	289.4	255.4	2.4	0.9254
2000	2000	288	219.2	4.2	0.8778	289.4	255.4	2.4	0.9254
1000	1000	288	217.2	3.6	0.8762	289.4	254.4	1.4	0.9244

1000	10000	288	217.2	3.6	0.8762	289.4	254.4	1.4	0.9244
5000	20	288	218.6	3.6	0.8776	289.4	255.2	2.2	0.9252
5000	200	288	218.6	3.6	0.8776	289.4	255.2	2.2	0.9252
5000	2000	288	218.6	3.6	0.8776	289.4	255.2	2.2	0.9252
10000	20	288	218.6	3.6	0.8776	289.4	255.2	2.2	0.9252
10000	200	288	218.6	3.6	0.8776	289.4	255.2	2.2	0.9252
10000	2000	288	218.6	3.6	0.8776	289.4	255.2	2.2	0.9252
2000	5000	288	217	2.6	0.877	289.4	253.8	0.8	0.9248
5000	5000	288	216.4	2	0.877	289.4	253.6	0.6	0.9248
10000	5000	288	216.4	2	0.877	289.4	253.6	0.6	0.9248
2000	1000	288	215.6	1.6	0.8762	289.4	253.2	0.4	0.9242
2000	10000	288	215.6	1.6	0.8762	289.4	253.2	0.4	0.9242
5000	1000	288	214.8	0.8	0.8762	289.4	253	0.2	0.9242
5000	10000	288	214.8	0.8	0.8762	289.4	253	0.2	0.9242
10000	1000	288	214.8	0.8	0.8762	289.4	253	0.2	0.9242
10000	10000	288	214.8	0.8	0.8762	289.4	253	0.2	0.9242

Mean values for the estimated number of OTUs, erroneous mappings (i.e. the number of reads assigned to an erroneous reference sequence) and the correlation between the real and estimated abundances within each community of 5 mock communities of 100bp and 200bp read's length. All mock communities consisted in 10,000 synthetic reads sampled from the reference database with a lognormal distribution (see Material and Methods).

1). The number of OTUs estimated with the original mode was between 3.8 and 2.7 times the real OTU richness while these values were of 0.97 and 1.12 for the strict mode (Table 1). The strict mode was able to unambiguously assign an average of 45.9% (min – max: 45.2% – 46.9%) and 67.4% (min – max: 66.7% – 68.1%) of the rDNA fragments at the OTU level, for reads of 100 bp and 200 bp respectively.

When different combination of values of the *maxaccept* and *maxreject* parameters were tested on the lognormal (Fig. 1) dataset, the number of OTUs was overestimated for low values of these parameters and underestimated for the higher rank of values tested (Table 2). The overestimation of OTUs was created by the wrong estimation of OTUs not present in the real data but present at low abundances in the extracted data when low values of *maxaccept* and *maxreject* were used (Fig 3A-B). These unreal low-abundance OTUs were not estimated when higher values of the parameters were used (Fig 3C-D). The number of wrong mappings decreased as the parameter's values increased, while the correlation between the real and estimated OTU's abundances was very insensitive to the variation in values tested (Table 2). These patterns were consistent for both, the 100 bp and the 200 bp datasets.

The compositional dissimilarities (Bray-Curtis) between pairs of communities in the mock and reconstructed dataset were significantly correlated (Mantel r = 0.548, P-value < 0.001) for the original and the strict mode, being always lower for the strict mode (Fig. 4A). The number of OTUs observed with the two modes significantly fitted to a linear model (P-value <0.0001, $R^2 = 0.811$) yet were always much lower for the Strict mode (Fig. 4B).

Discussion

The use of metagenomic rDNA reads has been proposed as an efficient tool for the description of microbial communities and different methods have been developed for this purpose (Huson *et al.* 2007; Sharpton *et al.* 2011; Logares *et al.* 2013; Darling *et al.* 2014; Guo *et al.* 2016). We provide here an R package that implements the previously described miTags approach (Logares *et al.* 2013). This method is based on the extraction of rDNA reads by using HMM and the binning of these reads by performing a search against a reference database of full-length rDNA sequences. We also evaluate, through the use of mock communities, a modification of the original method for dealing with the uncertainty in the taxonomical assignation of rDNA reads at the OTU level.

The heterogeneous variability at the sequence level of the ribosomal genes, which are composed of variable and conserved regions (Van de Peer *et al.* 1996), may be problematic when binning rDNA reads into OTUs by similarity searches against reference databases.

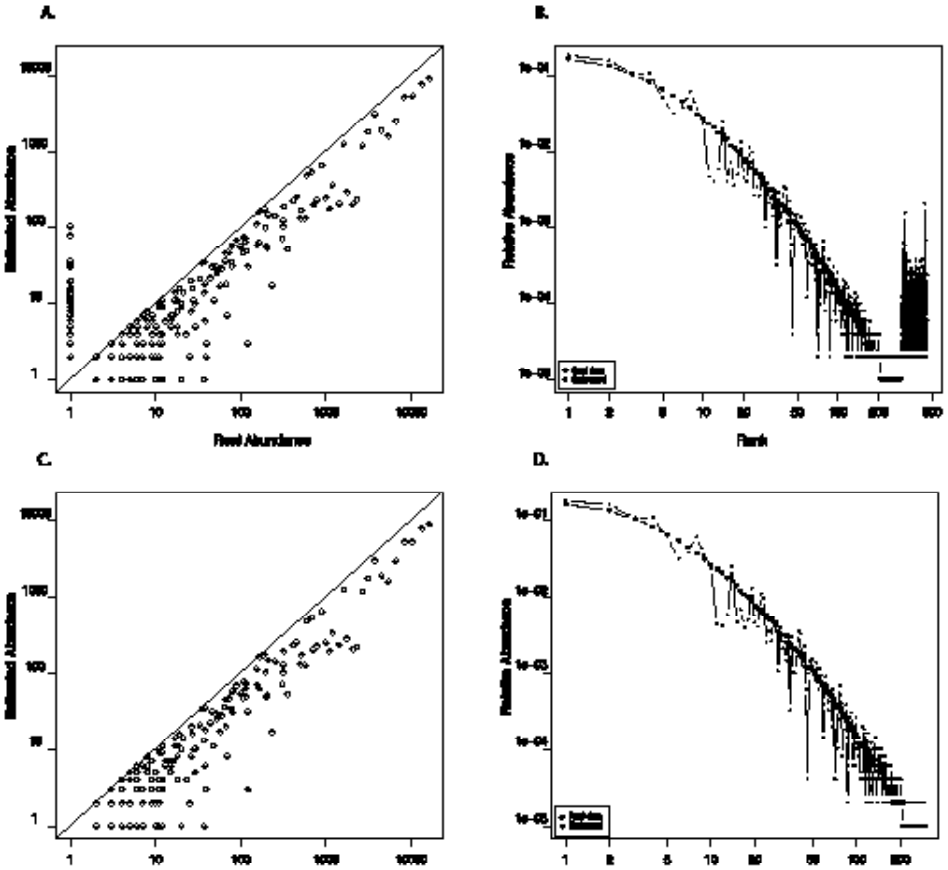


Figure 3: **Real and estimated OTU abundances within a lognormal mock community reconstructed through the Strict mode**. Real versus estimated abundances (A and C). The rank-abundance relationships (B and D) based on the real community (large points) with the estimated abundances over imposed (smaller points). The reconstruction of abundances was performed using two parameter combinations: *maxaccepts = 20, maxrejects = 2000* (A and B) and *maxaccepts = 5000, maxrejects = 5000* (C and D).

Specifically, reads that correspond to conserved regions of the rDNA gene may align to several reference sequences introducing uncertainty in OTU assignation. We evaluated this uncertainty through the construction of mock communities obtained by generating in-silico rDNA fragments from the same database used as a reference in the miTags pipeline. In this manner, the correct assignation of each rDNA fragment was known and the error rate of the method, i.e. the proportion of rDNA fragments assigned to an erroneous reference sequence, could be computed. The mean error rate of the original method, which assigned each rDNA read to a reference sequence by using the best hit in

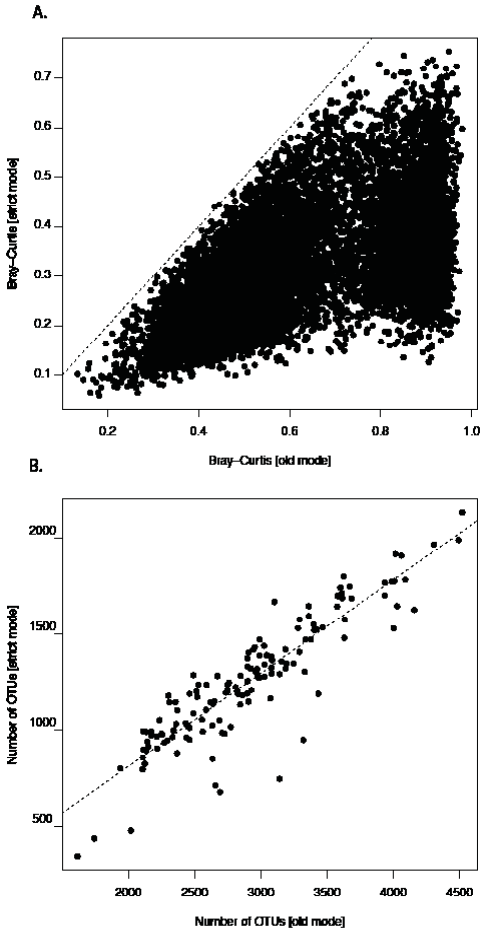


Figure 4: Scatterplot of Bray-Curtis dissimilarities (A) between pairs of communities for 139 marine metagenomes (see Material and Methods) analyzed with the Original and the Strict mode. The number of OTUs from the same dataset estimated with the Original and Strict modes (B). The original mode was run with the parameter combination *maxaccepts = 20, maxrejects = 500*. The strict mode with the combination *maxaccepts = 1000, maxrejects = 1000*. The dashed line in B represents the best fit to a linear model (y = 0.484x – 152.64).

a similarity search, ranged between 0.285 and 0.468 for mock communities of 200 bp and 100 bp length (Table 1). Thus, a high proportion of the rDNA fragments were assigned to erroneous OTUs. This resulted in the overestimation of OTU richness by a factor of 2.7 and 3.8, for the 200 bp and 100 bp mock communities, respectively (Table 1). The new method proposed in this work (i.e. the Strict mode within the *mtagger* R package) consists on restricting the analysis to those rDNA fragments that result in a hit to a single reference sequence, and thus the rDNA fragments that may not be assigned unambiguously to a single reference sequence are discarded. These reads may however be later used for higher taxonomical ranks assignation through a LCA approach. This new pipeline substantially lowered the error rate to mean values of 0.033 and 0.017 (Table 1), at the expenses of assigning satisfactorily a mean of 45.9% and 67.4% of the rDNA fragments from the mock

communities (100 bp and 200 bp dataset, respectively). With this new method the mean estimated number of OTUs was 9,773.8 and 11,271.6 (Table 1), much more accurate compared to the real value in the mock community, 10,000 OTUs. The new method proposed outperformed the original one in both its error rate and in OTU richness estimation by discarding the reads with ambiguous hits, i.e. the reads that could not be mapped to a single reference sequence. The error rate of the original method was unevenly distributed along the 16S rRNA gene: it was especially high in some conserved regions, especially the ones between the V2 and V3 and V8 and V9 and especially low in the variable regions V4 and V6 (Fig 2), the latter those V regions commonly used for amplicon sequencing (e.g. Ghiglione *et al.* 2012; Salazar *et al.* 2016). However, the modified method proposed here (the Strict mode), instead of restricting the analysis to a particular V regions as other methods have proposed (Guo *et al.* 2016), takes profit from the natural variability of the whole rDNA sequence length.

The *mtagger* R package uses USEARCH (Edgar 2010) for mapping the rDNA reads from metagenomic datasets to the reference database. USEARCH is a heuristic search method, i.e. it does not search each query sequence to all the target sequence in the reference database. The heuristic nature of this search may result in a low proportion of erroneous hits that would explain the low but positive error rates of the Strict mode (see above). The degree of sensitivity in the USEARCH algorithm may be modulated by the *maxaccepts* and *maxrejects* parameters. However, a thorough test of the effects of these parameters on the miTags approach had not been performed before. Here we created 5 mock communities with realistic abundances by sampling the reference database with a lognormal distribution. Each of these mock communities was analyzed with the mtagger package by using the Strict method and testing combinations of increasing values of the *maxaccepts* and *maxrejects* parameters. The error rate decreased with increasing the values of both parameters from 0.014 - 0.02 to virtually 0 (Table 2). An error rate between 0.14 and 0.2 corresponds to ~150 – 200 erroneous mappings, which, despite being a small proportion of the 10,000 rDNA fragments analyzed per sample, considerably inflates the number of total OTUs (Table 2). These erroneous mappings tend to bias the estimation of the total number of OTUs by adding a high number of low-abundance erroneous OTUs (Fig. 3A-B). This bias is corrected when higher values for the *maxaccepts* and *maxrejects* parameters are used (Fig. 3C-D). However, for high values of these parameters, the total estimate of OTUs is lower than the real OTU richness (Table 2). This is due to a different process: as the precision of the strict method comes at the expenses of discarding a considerable proportion of the rDNA fragments, and their abundance distribution is skewed to low-abundance OTUs, some of the discarded rDNA fragments are the only

representatives of their OTU. However, this is a general obstacle for any method characterizing microbial communities through sequencing technologies: as microbial community diversity is rarely fully sampled and sequenced, there is always a strong dependence between total OTU richness and the sequencing effort (i.e. the number of sequences per sample) that can only be solved by increasing the latter. Based on the tests performed here, we recommend, as a standard pipeline, the use of the mtagger package in the Strict mode and a minimum value of 1,000 for both the *maxaccepts* and *maxrejects* parameters. This combination of values results in accurate estimates of the community composition within a sample, with a mean error rate of $3.6 \cdot 10^{-4}$.

Finally, both approaches were compared with a real dataset consistent on the rDNA fragments of 139 publically available marine metagenomes in which the original miTags protocol were used (Sunagawa *et al.* 2015). Consistently with the in-silico tests, the similarities between communities (i.e. the Bray-Curtis index) estimated by the two approaches were correlated and lower for the Strict mode (Fig. 4A). The richness estimates were also lower for the strict mode when applied to this real dataset and about half of the total OTUs detected with the original method, consistently with the result of the in-silico tests. However, both richness measures were highly correlated too (Fig. 4B).

Therefore, the newly proposed pipeline outperformed the original method in both the estimation of the OTU abundances (especially for the low-abundance OTUs) and in the estimation of the total OTU richness. We also make available a software package for the estimation of microbial community composition from metagenomic data with an improved method, a standard and reproducible pipeline, and an assessment of its error rate and limitations.

References

Bray JR, Curtis JT (1957) An ordination of the upland forest communities of southern Wisconsin. *Ecological Monographs*, **27**:325–349.

Darling AE, Jospin G, Lowe E *et al.* (2014) PhyloSift: phylogenetic analysis of genomes and metagenomes. *PeerJ*, **2**:e243.

Eddy SR (2011) Accelerated profile HMM searches. *PLoS Computational Biology*, **7**:e1002195.

Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, **26**:2460–2461.

Ghiglione J-F, Galand PE, Pommier T *et al.* (2012) Pole-to-pole biogeography of surface and deep marine bacterial communities. *Proceedings of the National Academy of Sciences USA*, **109**:17633–17638.

Guo J, Cole JR, Zhang Q, Brown CT, Tiedje JM (2016) Microbial community analysis with ribosomal gene fragments from shotgun metagenomes. *Applied and Environmental Microbiology*, **82**:157–166.

Huson D, Auch A, Qi J, Schuster S (2007) MEGAN analysis of metagenome data. *Genome Res.*, **17**:377–386.

Lima-Mendez G, Faust K, Henry N *et al.* (2015) Determinants of community structure in the global plankton interactome. *Science*, **348**:1262073–1262073.

Logares R, Sunagawa S, Salazar G *et al.* (2013) Metagenomic 16S rDNA Illumina tags are a powerful alternative to amplicon sequencing to explore diversity and structure of microbial communities. *Environmental Microbiology*, **16**:2659–2671.

Van de Peer Y, Chapelle S, De Wachter R (1996) A quantitative map of nucleotide substitution rates in bacterial rRNA. *Nucleic acids research*, **24**:3381–3391.

Pruesse E, Quast C, Knittel K *et al.* (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Research*, **35**:7188–196.

Salazar G, Cornejo-Castillo FM, Benítez-Barrios V *et al.* (2016) Global diversity and biogeography of deep-sea pelagic prokaryotes. *The ISME Journal*, **10**:596–608.

Sharpton TJ, Riesenfeld SJ, Kembel SW *et al.* (2011) PhylOTU: A high-throughput procedure quantifies microbial community diversity and resolves novel taxa from metagenomic data. *PLoS Computational Biology*, **7**:e1001061.

Sunagawa S, Coelho LP, Chaffron S *et al.* (2015) Structure and function of the global ocean microbiome. *Science*, **348**:1261359–1261359.

Vertical microbial connectivity in the global ocean

Chapter V

Abstract

The vertical structure of the ocean is characterized by steep gradients of major abiotic factors, such as light, temperature, and nutrients. As a result, considerable vertical variations in the composition of microbial communities are expected and have been reported, showing low connectivity between the different depths. However, processes such as the sinking of particles or the vertical movement of water masses, may increase the vertical connectivity of the ocean. Here we evaluate the vertical structure and/or connectivity of the prokaryotic communities at various depth in the ocean by modeling the effects of dispersal events between communities with metagenomic data from the *Tara Oceans* expedition, which includes the surface, deep chlorophyll maximum (DCM) and mesopelagic of 66 world-wide distributed locations. We observed and defined a quasi-universal fast decay of community similarity with depth which fits to a power-law equation. The prokaryotic communities are clearly structured into two realms (the photic and the mesopelagic) with little diversity in common. This modeling of community compositional data suggests mixing as the process governing the vertical similarity within the photic ocean, i.e. between the surface and the DCM. We also identified a particularly increased connectivity between the photic and the aphotic ocean in 5 stations, which is compatible with events of whole community export from the photic ocean to the mesopelagic. The approach applied here allowed the incorporation of directionality in the measurement of similarity between microbial communities, which may be useful for a better understanding of other ecosystems and identifies instances of strong vertical mixing, not apparent from the inspection of other data.

Introduction

A substantial amount of studies have addressed the structure of microbial communities along the vertical gradient of the ocean. Most of them are based on vertical profiles in a single station or in several stations in the same ocean basin (Moeseneder *et al.* 2001; Pham *et al.* 2008; Treusch *et al.* 2009; Brown *et al.* 2009; Galand *et al.* 2010; Eiler *et al.* 2011; Friedline *et al.* 2012; Wilkins *et al.* 2013; Cram *et al.* 2015). Only recently, a global-scale description of the vertical structure of microbial diversity with standardized sampling efforts has covered simultaneously extensive sampling areas and several depths (Sunagawa *et al.* 2015). Consequently, basic questions regarding the vertical organization of microbial diversity are still to be answered. The aphotic ocean accounts for more than half of the prokaryotic biomass and production of the global ocean (Arístegui *et al.* 2009). However, the known sources of organic matter feeding the dark ocean are not sufficient to meet the energy demands the deep ocean microbes (Herndl & Reinthaler 2013). Understanding the vertical structure of microbial communities and the processes linking the photic and aphotic microbial plankton might help in deciphering the ecological functioning of the ocean.

The vertical stratification of the dominant planktonic microbes has been described since genetic tools were first used for the characterization of ocean microbial communities (Giovannoni *et al.* 1996; Fuhrman & Davis 1997; Field *et al.* 1997; Lovejoy *et al.* 2006; Countway *et al.* 2007). Steep vertical gradients of light quality and intensity, temperature, and nutrient concentrations are present in all the oceans and have been shown to influence species distributions (Rocap *et al.* 2003; Sampayo *et al.* 2007; Hu *et al.* 2011; Vergin *et al.* 2013). However, several processes, such as vertical diffusive or convective mixing (Carlson *et al.* 1994; Hansell 2002), the overturning circulation where intermediate- and deep water formation take place (Hansell 2002; Hansell *et al.* 2004), or the flux of sinking particles (Arístegui *et al.* 2005) have been described as mechanisms exporting organic matter to the deep ocean or bringing deep water to the surface, and thus connecting upper with deeper layers. Planktonic microbes are incapable of active movement through long distances and might be trapped in water masses, as has been repeatedly suggested (Galand *et al.* 2010; Agogué *et al.* 2011) or might be carried by sinking particles. Consequently, some or all of these processes exporting organic matter to the deep ocean may export as well entire microbial communities. Although theoretically proposed (Rillig *et al.* 2015), the existence, incidence and detectability of whole microbial community export events to the deep ocean has not been explored. Yet such processes should be detectable by studying

the composition of microbial communities through the water column.

Here we analyzed the free-living pelagic prokaryotes of 139 globally distributed samples from the surface, DCM and mesopelagic ocean layers retrieved during the Tara Oceans Expedition (Sunagawa *et al.* 2015) to derive general patterns of the vertical structure of prokaryotic communities. Prokaryotic diversity resulted to be organized into two distinct photic and aphotic realms through a globally consistent fast decay of community similarity with depth. Vertical mixing seemed to govern the variability in the patterns of similarity between the surface and the DCM, within the photic layer. We, however, detected some sites with a high connectivity between the photic and aphotic realms that suggests the existence of events of fast export of photic prokaryotes into the deep ocean.

Material and Methods

Sample collection and sequencing

Water samples were collected during the Tara Oceans Expedition (Karsenti *et al.* 2011; Bork *et al.* 2015) from 66 stations along the Mediterranean Sea and the Atlantic, Indian and Pacific Oceans (Table S1, Fig. S1). Water was sampled from a maximum of three depths in each station, comprising a total number of 139 samples: i) the surface (63 samples), i.e. between 3 and 7 m depth, ii) the deep chlorophyll maximum (DCM; 46 samples) determined from the chlorophyll fluorescence as detected by the CTD sensor, and iii) the mesopelagic zone (30 samples), i.e. the layer between 200 and 1000 m. The sampling depth within the mesopelagic zone, which varied from station to station, was selected based on the vertical profiles of temperature, salinity, fluorescence and oxygen. Detailed sampling protocols of the whole expedition have been previously described (Pesant *et al.* 2015). Briefly, for each sample in this study 100 L of water were collected and on-board pre-filtered successively with a mesh of 200 µm, 20 µm and a membrane of either 1.6 or 3 µm pore size and the cells were retained using a membrane with a pore size of 0.2 µm, as previously described (Sunagawa *et al.* 2015). Thus the samples here considered correspond to the plankton comprised in the 1.6 – 0.2 µm or the 3 – 0.2 µm size fraction. The 0.2-µm membranes were flash frozen in liquid nitrogen and stored at -80º C.

Metagenomic DNA was extracted as described before (Logares *et al.* 2013). DNA was sequenced using pair-end Illumina sequencing technology (Illumina, USA) and high-quality (HQ) reads were obtained using MOCAT (v. 1.2) (Kultima *et al.* 2012) as previously described (Sunagawa *et al.* 2015).

Taxonomical community composition

The taxonomical description of the prokaryotic communities was performed using

miTags, i.e. 16S gene's fragments extracted from Illumina-derived metagenomes (Logares *et al.* 2013). The HQ reads corresponding to 16S genes were detected using HMMER v.3.0 (www.hmmer.org). All reads detected as part of an rRNA gene with length >100 bp were aligned against the SILVA database (v.123) (Pruesse *et al.* 2007) using USEARCH (Edgar 2010). The alignment was done using a 97% similarity cut-off and thus the SILVA database was pre-clustered, also at a 97% similarity, before the alignment. All the ambiguous hits (i.e. reads with a successful hit to more than one sequence of the reference database) were excluded. In this way the miTags were binned into Operational Taxonomic Units (OTUs): each sequence of the pre-clustered SILVA database that recruited a read was considered as an OTU and their abundance was obtained as the number of reads that aligned to this sequence. An OTU table was constructed containing the number of reads that belonged to each OTU in each of the samples. In order to avoid artifacts due to the uneven sequencing effort among samples, this OTU table was rarefied to the minimum number of total miTags per sample (8,658) using the *rrarefy* function in the *vegan* package (Oksanen *et al.* 2015) within R Statistical Software (R Core Team 2016). This process was repeated 100 times and the mean number of reads (rounded to integers) from the 100 rarefactions was used. The taxonomy from the SILVA sequences was used as the taxonomical classification of the OTUs.

Abundance of PSII genes

The abundance of the genes involved in the photosystem II protein complex were obtained from functional profiles tables of this dataset previously analyzed (Sunagawa *et al.* 2015) based on KEGG orthologous groups (KOs) assignments. Abundances were calculated as the sum of the relative abundances of reference genes, or key marker genes, annotated as KOs. This data is available at http://ocean-microbiome.embl.de/companion.html. Abundances of psbA-F genes were compiled by searching for their corresponding KO terms (K02703, K02704, K02705, K02706, K02707, K02708).

Statistical analyses

Species diversity for each sample was estimated as the Shannon index with the *diversity* functions in the *vegan* R package. Community composition was compared between samples by using the Bray-Curtis dissimilarity index and the resulting dissimilarity matrix was used to perform a non-metric multidimensional scaling analysis (NMDS) (Minchin 1987) using random starts. Differences between layers in community composition were tested using Permutational MANOVA (Anderson 2001) with 1,000 permutations. Additionally, the Euclidean distance (after Hellinger transformation of the OTU table), the

Canberra, Kulczynski, Morista and Gower distances were also computed as implemented in the *vegdist* function of the *vegan* R package.

Disperflux model

A simulation model, referred hereafter as the "disperflux model", was developed in order to in silico reproduce the effect of a directional single event of whole community export from one ecological community (source community) to another (sink community). Whole community export is defined here as a single event that transfers a fraction of the entire source community (i.e. of its individuals) to the sink community, proportionally to the species' abundances in the source community. Given two communities, the model also allows estimating from the real data the magnitude of such export event (referred hereafter as "flux") in any of the two directions (i.e. using sequentially both communities as source and as a sink community). These estimated fluxes may be interpreted as a directional measure of similarity between two communities (see SI for a detailed explanation of the model architecture and output interpretation).

The *disperflux* model was run for every pair of samples in both directions and using 1,000 iterations. Three matrices were constructed representing the estimated mean flux between samples, and the upper and lower 95% CI from every sample to every other. Downward fluxes (i.e. fluxes from surface/DCM to the mesopelagic) were related to the Shannon index of mesopelagic communities and to the vertical distance separating the samples by fitting a linear and potential equation, respectively, with a least-square linear regression (of the log-transformed variables in the latter case). Downward fluxes were related to previously derived particle fluxes at 150 m (Guidi *et al.* 2016), the absolute difference in depth, temperature, oxygen concentration and salinity between samples and to the strain submesoscale index (the intensity of geostrophic currents and the SST gradient, Pesant *et al.* 2015) by using a linear model.

The disperflux model is made available as an R package and is accessible at https://github.com/GuillemSalazar/disperflux.

Results

A total of 14,129,971 of 16S rDNA gene reads were extracted from the metagenomes from which a 23.5% could be unambiguously mapped to the SILVA database (Table S1). An OTU table was constructed and rarefied to the minimum reads/sample (see Material and Methods) obtaining a final dataset containing 11,565 OTUs and 1,191,397 counts.

Alpha and beta-diversity of prokaryotic and eukaryotic communities

The NMDS based on the prokaryotic data resulted in two nearly non-overlapping clusters of samples that corresponded to the mesopelagic samples and the surface/DCM samples (Fig. 1A). Bray-Curtis similarity was high within surface samples, DCM samples, mesopelagic samples and when comparing surface samples with DCM samples. Similarities were lower when comparing either surface or DCM samples with mesopelagic samples (Fig. 1B). Differences in community composition between layers were significant for all the pairwise comparisons, although the difference between surface and DCM explained a minor proportion of the variance (surface-DCM: P-value < 0.001, R^2 = 6.4%;

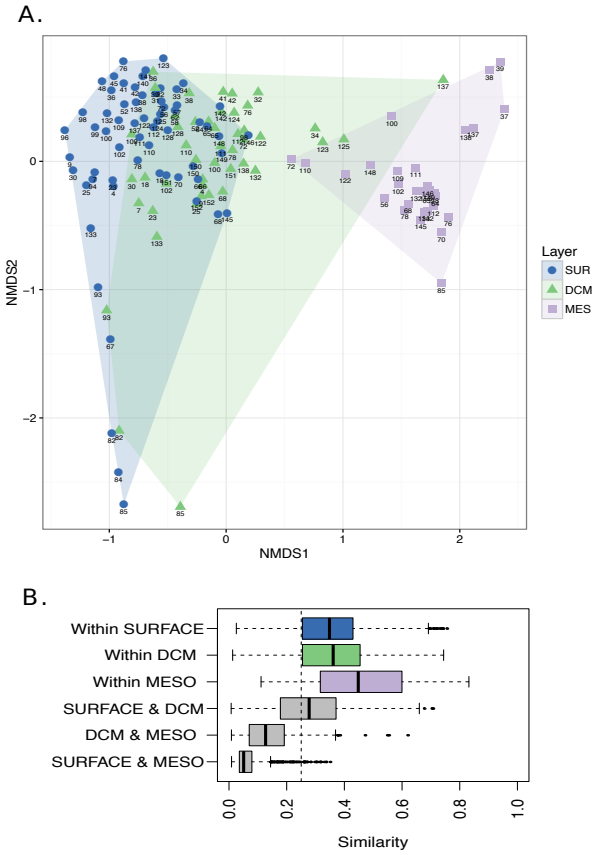


Figure 1: **Beta-diversity of prokaryotic communities**. Non-metric multidimensional scaling (A) based on Bray-Curtis dissimilarities of prokaryotic abundances. Surface, the depth of the deep chlorophyll maxima (DCM) and mesopelagic samples are color-coded. Distribution of Bray-Curtis similarities within and between layers (B). The mean similarity is denoted with a vertical dashed line.

surface-mesopelagic: P-value < 0.001, R^2 = 33.5%; DCM-mesopelagic: P-value < 0.001, R^2 = 31.1%)

For the 22 stations for which the three layers were sampled, there was on average a 13%, 11.4% and 37.3% of OTUs that were unique to the surface, DCM and mesopelagic respectively. An average 9.7% of the OTUs were shared between the three layers. A 2%, 16.9% and 8.7% of the OTUs were shared, on average, between the surface and the mesopelagic, the surface and DCM and the DCM and the mesopelagic, respectively (Fig. S2).

Pairwise comparison of OTU's abundances

The pairwise comparison of the OTU abundances between samples within the same station resulted in three distinctive patterns: a) For most of the surface-DCM comparisons the OTUs abundances were correlated in both samples and laid along the 1:1 line (Fig. 2A, D). b) For most of the surface–mesopelagic or DCM-mesopelagic comparisons the OTUs present in one sample were absent or very rare in the other sample (Fig. 2B,

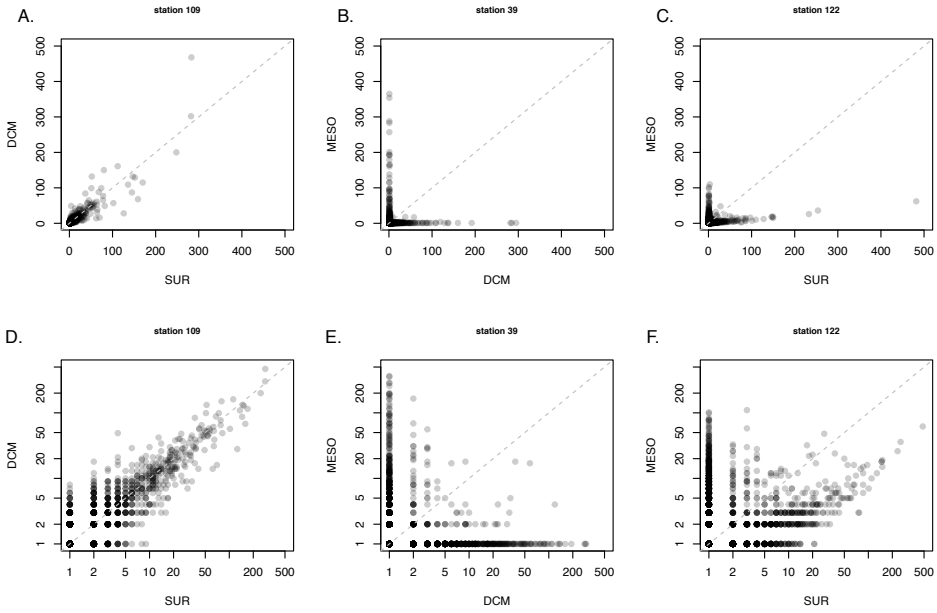


Figure 2: **Pairwise comparison of OTU abundances**. Scatterplots of the abundances (reads+1) of all the OTUs in pairs of samples. One example of the "Mixing" (A, D), "No flux" (B, E) and "Directional flux" (C, F) scenarios are presented (see Results and SI for explanation), corresponding to the stations 109, 39 and 122, respectively. The layers compared in each panel are denoted in the axes. The lower panels correspond to the log-log version of the upper panels and facilitate visualization of some patterns described in the text.

E). c) For eight of the later comparisons (corresponding to stations 56, 72, 110, 122 and 148) the OTUs present in the mesopelagic sample were virtually absent from the surface/DCM sample while the abundant OTUs in the surface/DCM sample were present and proportionally represented in the mesopelagic sample, although with lower abundances. Thus, this resulted in a set of points that lay parallel and below the 1:1 line. (Fig. 2C, F and Fig. S3).

Estimation of fluxes through the disperflux model

Vertical fluxes (% of community) estimated from prokaryotic data within each station were high although with a considerable variability for most of the surface-DCM comparisons in both directions and close to 0 for most of the surface/DCM-mesopelagic comparisons (Fig. 3, Fig. S5). Fluxes for stations 56, 72, 110, 122 and 148 were higher for the downward direction (from surface/DCM to mesopelagic) than for the upward direction (from mesopelagic to surface/DCM). Station 137 had high fluxes in both directions for the DCM-mesopelagic but not for the surface-mesopelagic comparison.

The magnitude of the downward fluxes decayed as a potential function of the vertical distance separating the pairs of communities within each station. A statistically different best fit was found for stations with an increased downward flux as identified in Fig 2 (stations 56, 72, 110, 122 and 148)

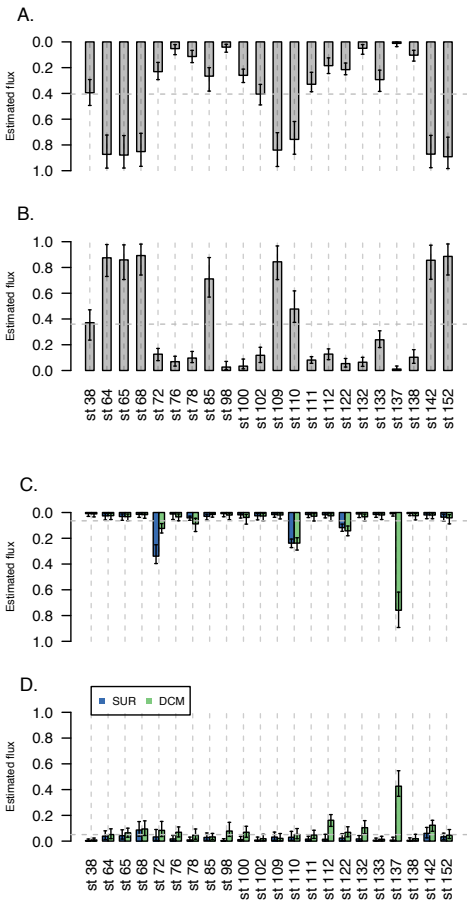


Figure 3: **Estimated fluxes based on the disperflux model**. Mean and 95% CI of the optimal fluxes (% of community) estimated for the stations for which the three samples (surface, DCM and mesopelagic) were available. Estimated fluxes from surface to DCM (A), from DCM to surface (B), from surface/DCM to mesopelagic (C) and from mesopelagic to surface/DCM (D). Estimations are based on 1,000 iterations of the disperflux model.

compared to the rest of stations (Fig. 4A). A similar pattern was observed when using Bray-Curtis dissimilarities (Fig S6). The Shannon diversity index for mesopelagic communities and the number of OTUs were found to be a linear function of the flux values from the surface to the mesopelagic (Fig. 4B and C). Downward fluxes from surface to DCM were statistically related to both, the (log) vertical separation between samples, and the absolute difference in temperature between the sampled depths (Fig. S7).

The fluxes estimated by applying the disperflux model to all pairwise comparisons were correlated to the OTU-based dissimilarity measures tested (Bray-Curtis, Euclidean, Canberra, Kulczynski, Morista and Gower). Pairs of samples with a high dissimilarity cor-

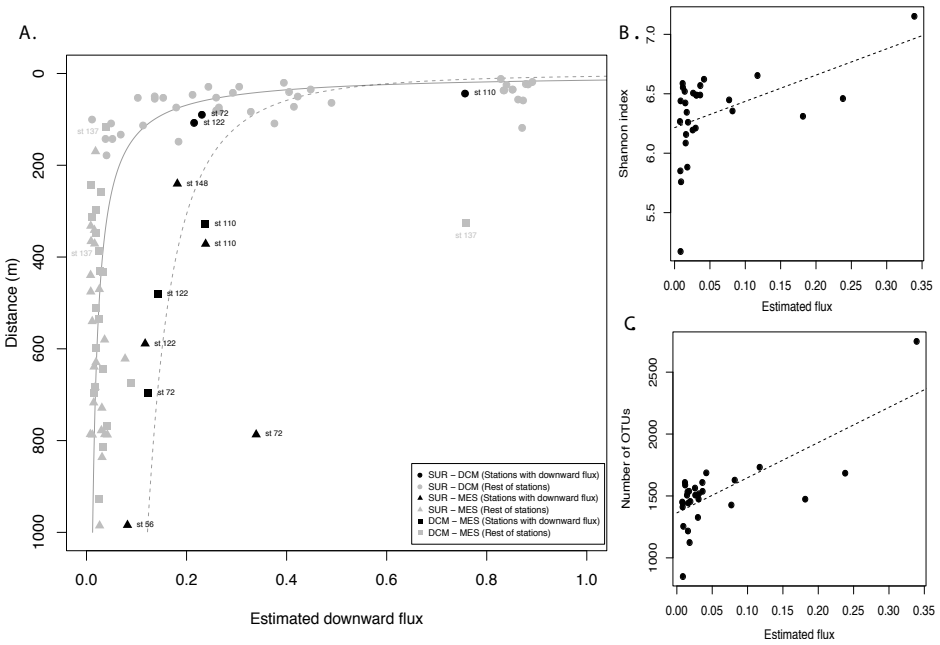


Figure 4: **Relation of fluxes to depth and diversity**. Downward mean fluxes (i.e. from surface to DCM and from surface/DCM to mesopelagic) in relation to the vertical distance (m) separating each pair of samples (A). Stations with increased downward flux (stations 56, 72, 110, 122 and 148) are in black and the rest are in grey. Lines correspond to the best fit of a potential equation for each of the two groups of points (grey points: $y = 16.75 * x^{-1.046.}$, Adjusted $R^2 = 0.742$; black points: $y = 2.2 * x^{-0.419}$, Adjusted $R^2 = 0.440$). Station 137 was excluded from the analysis (see Discussion). Both equations were significantly different (P-value < 0.05; tested through the interaction's significance of an ANCOVA model with log[flux] as independent variable, log[distance] as dependent variable and a factor differentiating stations with increased downward flux from the rest of the stations). Mean fluxes from surface to mesopelagic in relation to the Shannon index (B) and the richness (number of OTUs) of mesopelagic communities (C). Lines correspond to the best fit of a linear equation (Shannon index: $y = 6.21 + 2.21*x$, Adjusted $R^2 = 0.201$; Number of OTUs: $y = 1363.25 + 2841.37*x$, Adjusted $R^2 = 0.514$).

responded to low fluxes and those with a low dissimilarity corresponded to high fluxes. Pairs of samples with intermediate dissimilarity values exhibited a high dispersion of the estimated flux values (Fig. S4).

Photosynthetic organisms and genes in the mesopelagic ocean

The abundance (number of reads) of genes involved in the photosystem II protein complex was high in most of the surface and DCM samples and low in the mesopelagic, except for some stations including the stations with an increased downward flux, i.e. stations 56, 72, 110, 122 and 148 (Fig. S8).

Discussion

Beta-diversity of microbial communities along the vertical gradient

Mesopelagic prokaryotic communities have been shown to differ highly in their taxonomical composition from photic communities at global scales (Zinger *et al.* 2011; Sunagawa *et al.* 2015). Here, the photic communities clustered together and apart from mesopelagic communities in the multidimensional scaling (Fig. 1A) and a decrease in community similarity was observed when comparing mesopelagic communities to any of the photic layers (surface or DCM) while similarity was relatively higher for intra-layer comparisons (Fig. 1B). Although some DCM samples differed from their surface counterparts in the NMDS the similarity between all surface and DCM samples was only slightly lower than the similarity within each of the two layers (Fig. 1B). Differences in community composition between layers were significant for all the pairwise comparisons, although the difference between surface and DCM explained a minor proportion of the variance ($R^2 = 0.06$) compared to the differences between surface or DCM and the mesopelagic ($R^2 = 0.31$ and $R^2 = 33$, respectively). Other studies, however, have detected clearer differences between DCM and surface layers although they have explored the temporal variability in one site through the water column (Treusch *et al.* 2009; Chow *et al.* 2013). Thus, at a global spatial scale and integrating the temporal variability, our results suggest a two-realm structure of ocean's microbial diversity: a photic realm (i.e. including surface and DCM with minor differences between them) and a mesopelagic realm below 200 m depth. However, the boundary between these two realms, in terms of community composition, is not perfect as a considerable proportion of diversity (20.4% of the OTUs) is shared between them. This shared diversity is not symmetrically distributed, as an average of 37.3% of the OTUs within a station were unique to the mesopelagic while only a 13% and 11.4% were unique to the surface or DCM. This indicates directionality in the vertical

structure of the oceanic microbial communities: a high proportion of the photic diversity is contained within the mesopelagic while the opposite is not true. A similar pattern has been described for these same stations when analyzing their functional core microbiome (Sunagawa *et al.* 2015).

Estimation of whole community export fluxes

In order to analyze this asymmetry along the vertical gradient we performed pairwise comparisons of samples within each station (i.e. in the vertical gradient) by confronting the OTUs abundance by pairs of samples. Three distinctive patterns emerged: a) a pattern characterized by correlated abundances of OTUs in both samples (Fig. 2A, D), found for most of the surface-DCM comparisons, b) a pattern characterized by the virtual absence of common diversity between samples (Fig. 2B, E), found for most of the photic-mesopelagic comparisons, and c) a more complex pattern where the low abundant OTUs within a community resembled a subsample of the dominant OTUs in the other community being compared. This pattern results in a characteristic distribution of points across a straight line parallel to, but below, the 1:1 line (Fig. 2C, F). This pattern was found when comparing photic samples to mesopelagic samples in 5 stations (Fig. S3). The fact that the rare members in these mesopelagic communities correspond to the dominant OTUs in the photic zone and that their relative abundances are correlated in both communities points to some sort of directional community exchange. In fact, aquatic environments have been proposed as likely environments where the interchange of entire communities may be a relevant process shaping community composition (Rillig *et al.* 2015). Moreover although poorly studied in marine environments (but see Wilkins *et al.* 2013) mass effects, i.e. the continuous or massive immigration of organisms that are not self-maintaining in the target environment (Shmida & Wilson 1985), have been proved to be a relevant process shaping the community composition in experimental (Livingston *et al.* 2013; Souffreau *et al.* 2014) and natural (Crump *et al.* 2007; Adams *et al.* 2014; Ruiz-González *et al.* 2015) freshwater environments.

Here we propose that the reconstruction of the community composition through the modeling of events of directional interchange between communities may result in a better characterization of beta-diversity. The in-silico simulation of a whole community export event from the photic to the mesopelagic ocean (i.e. the disperflux model; see Material and Methods and SI) was able to reproduce the pattern observed in the real data from stations 56, 72, 110, 122 and 148 (Fig. S9): the addition of a random sample of individuals from the photic community was able to reproduce the rare biosphere of the mesopelagic sample, i.e. the straight line parallel to and below the 1:1 line observed for the real data.

Thus, a single massive event of directional export of the complete photic community into the mesopelagic community, although being probably a simplified scenario, is sufficient to explain the abundance patterns observed for the 5 stations described above. In fact, some evidences exist of fast export events of photic microbes to the deep ocean: the presence of healthy photosynthetic cells down to 4,000 m has been previously reported and attributed to fast particle's sinking rates (Kimball Jr. *et al.* 1963; Lochte & Turley 1988; Agusti *et al.* 2015). The presence of *Prochlorococcus* populations in the Pacific Ocean down to 1,500 m has also been reported and attributed to physical transport processes (Jiao *et al.* 2014). The disperflux model also allowed estimating the magnitude of such potential export events from the real data. The "flux" (the only parameter to be estimated within the model) is defined as the proportion of the sink community (i.e. of individuals) that has to be transferred from the source community in order to better reproduce the sink community structure. This only parameter, estimated in both directions, was able to reproduce any of the three pairwise patterns observed for the whole dataset (Fig. 2). It is also correlated to the commonly used measures of community similarity (Fig. S4) but incorporates a directional component. Consequently, the flux may be used as a "directional similarity measure" between pairs of communities. The flux is a dispersal-centered measure, as the process simulated for its estimation is the interchange of individuals between communities. Thus it may be interpreted as a measure ranging from a dispersal limitation situation (low flux values) to a mass effect situation (high flux values), although other processes, such as environmental filtering, may result in similar patterns in the cases when the two possible fluxes between pairs of communities are of the similar magnitude (see SI for further discussion).

The flux estimates for all the vertical comparisons resulted in high fluxes between the surface and the DCM in both directions, although a high variability existed (Fig. 3A, B and Fig. S5A, B). This indicates that the community composition of a surface sample can be reconstructed to a high extent by randomly sampling the DCM community in the same station and vice versa. This may indicate either the presence of a common environmental factor exerting a filtering effect and resulting in similar microbial assemblages in both layers and/or a high interchange rate of communities between the surface and the DCM layer. The estimated fluxes between surface/DCM and the mesopelagic were virtually null for the majority of the stations (Fig. 3C and Fig. S5C), i.e. no addition of individuals from the photic layer was able to reproduce the abundances of mesopelagic communities and vice versa. That stresses the generality of the two-realm structure of the ocean described above. Yet, however, stations 56, 72, 110, 122 and 148 resulted in low but positive flux estimates, which were higher in the downward direction (from surface/DCM

to the mesopelagic) than in the upward direction (from the mesopelagic to the surface/ DCM) (Fig. 3D and Fig. S5D). This indicates a directional asymmetry in the similarity of these communities: the low-abundance members of the mesopelagic community could be reconstructed by taking a random sample of the surface and DCM community within the same station. Thus, an export event of the whole photic community to the deep ocean is likely to have occurred in these locations, transferring a non-negligible proportion of individuals. The calculated fluxes are our best estimates for the magnitude of such event, even though our model doesn't temporally frame the events ("flux" has no units of time). In fact, genes from the photosystem II complex extracted from metagenomes, mainly present in most of the photic samples, were also observed in the mesopelagic samples from the 5 stations with high downward fluxes (Fig. S8). Additionally, for station 137 high fluxes were estimated between the DCM and the mesopelagic communities but not with the surface community. In fact, the DCM sample in this station does not correspond to a "canonical" DCM at the nitracline and the base of the photic layer, but corresponds to a secondary DCM peak well within an oxygen minimum zone that also includes the mesopelagic sample but not the surface one (Fig. S10). Thus, the high bi-directional flux estimated for this situation reflects the fact that both the DCM and the mesopelagic sample belonged to a water-mass with similar characteristics and possibly with high rates of mixing between them.

A quasi-universal vertical structure of microbial diversity

Although a considerably amount of studies have characterized the structure of microbial communities along the vertical gradient (Moeseneder *et al.* 2001; Pham *et al.* 2008; Treusch *et al.* 2009; Brown *et al.* 2009; Galand *et al.* 2010; Eiler *et al.* 2011; Friedline *et al.* 2012; Wilkins *et al.* 2013; Cram *et al.* 2015), an effort for testing the existence of vertical biogeographical patterns that are universal in all oceans has not been conducted. Here, the downward estimated fluxes for all vertical comparisons (excluding the 5 stations with high downward fluxes and the station 137) fitted to a common potential decay function of the vertical distance separating pairs of samples (Fig. 4A). The 5 stations with increased downward fluxes from the photic to the mesopelagic also fitted to a decay function, but with a statistically different decay rate (Fig. 4A). A similar fit to two different potential decay functions was also found for the Bray-Curtis similarity (Fig. S6). Thus, this indicates that a universal relation exists for the major oceans describing the loss rate of "photic microbes" and associated change in community similarity along the vertical gradient. This universal relation is only modified in scattered locations and can be used as a null hypothesis against which to detect these particular locations. Consequently, the estimated downward flux

from the photic layer was a good predictor of mesopelagic richness and diversity (Fig. 4B, C). The mesopelagic communities, thus, seem to be composed of a pool of deep microbial diversity to which the photic microbes are added when significant export events occur. This suggests that one or several transport processes may act at specific locations and/or situations injecting (likely in a non-continuous way) a considerable proportion of the photic members into the deep ocean.

Processes increasing vertical connectivity of the ocean

Several processes, both physical and biological, are known to link vertically the ocean. Particle formation in the photic layer and their sinking (i.e. the biological pump) has been recognized for long as the dominant process exporting biological material into the deep ocean (Ducklow *et al.* 2001; Arístegui 2002). Particles are known to be colonized by bacterial communities (Pedrós-Alió & Brock 1983; DeLong *et al.* 1993; Acinas *et al.* 1999) and thus particle export may drive the injection of photic microbes into the deep ocean. There are also a variety of physical processes capable of exporting photic organisms to deeper layers. Oceanic currents and advection have been shown to drive similarity patterns of pelagic bacterial communities through transport processes (Ghiglione *et al.* 2012; Wilkins *et al.* 2013). This large-scale circulation acts at relatively large spatial (100-1,000 km) and temporal (months to years) scales and would be responsible mainly for horizontal transport of microbial communities, except in the deep-water formation zones where vertical sinking of water masses does occur. Other transport processes, acting at shorter spatial scales, such as mesoscale eddies or convective mixing, may also act. In fact, numerical simulations have recently revealed that phytoplankton diversity and community structure is influenced by dispersion arising from mesoscale and sub-mesoscale transport processes (Lévy *et al.* 2014). Additionally, the presence of *Prochlorococcus* in the aphotic waters of the western Pacific Ocean has been attributed to vertical transport associated to eddies (Jiao *et al.* 2014). We tested the potential effect of particle sinking and mesoscale transport processes at explaining the detected flux patterns by relating the estimated downward fluxes to environmental variables (see Material and Methods). A linear model using the absolute difference in depth and temperature between surface and mesopelagic as significant explanatory variables explained 55% of the variance in the surface-to-DCM fluxes (Fig. S7). Thus, the similarity between the DCM and the surface ocean, at a global scale, highly depends on their vertical proximity and temperature difference. Although the effect of temperature exerting an environmental filtering effect over the members of surface and DCM communities may not be discarded, the joint effect of temperature and vertical distance as explanatory variables suggests mixing as the main force structuring surface

and DCM similarity. On the contrary, particle export seems not to be affecting community similarity, although its influence on specific taxa has been proved before (Guidi *et al.* 2016). For the surface-to-mesopelagic and DCM-to-mesopelagic fluxes no variable was found to be significant. The 5 stations with increased downward fluxes between the photic layers and the mesopelagic do not clearly correspond to zones with increased rates of particle export or an evidence of vertical transport processes. Thus, it doesn't seem to exist a single common process, out of those considered above, responsible for the pattern observed in these 5 stations. Station 148 is one of the shallower mesopelagic samples (250 m depth) and its temperature and salinity difference to the corresponding surface sample was low (2.2º C and 0.035 PSU; i.e. the station was very weekly stratified). In this case mixing of the shallow mesopelagic with the surface ocean may be a plausible explanation for the apparent directional flow described. However, this is not the case for the remaining four stations, which were very well stratified. Further studies will be needed to decipher the mechanisms responsible for the presence of a not negligible amount of typically photic taxa in the mesopelagic layer in these locations.

This work constitutes a first attempt to understand the effects of ocean (vertical) connectivity on its microbial diversity by modeling the composition of communities as a result of export events between communities. It represents an effort to move from studies based on ecological dissimilarity indices towards a more explicit modeling of processes capable of shaping community composition. The methods developed here revealed useful for the study of the vertical structure of marine prokaryotic communities but may apply to other ecosystems with high connectivity and clear directionality that are inhabited by microbes, such as inland water systems or the human gut.

References

Acinas SG, Antón J, Rodríguez-Valera F (1999) Diversity of free-living and attached Bacteria in offshore western Mediterranean waters as depicted by analysis of genes encoding 16S rRNA. *Applied and Environmental Microbiology*, **65**:514–522.

Adams HE, Crump BC, Kling GW (2014) Metacommunity dynamics of bacteria in an arctic lake: The impact of species sorting and mass effects on bacterial production and biogeography. *Frontiers in Microbiology*, **5**:1–10.

Agogué H, Lamy D, Neal PR, Sogin ML, Herndl GJ (2011) Water mass-specificity of bacterial communities in the North Atlantic revealed by massively parallel sequencing. *Molecular Ecology*, **20**:258–274.

Agustí S, González-Gordillo JI, Vaqué D *et al.* (2015) Ubiquitous healthy diatoms in the deep sea confirm deep carbon injection by the biological pump. *Nature Communications*, **6**:7608.

Anderson MJ (2001) A new method for non-parametric multivariate analysis of variance. *Austral Ecology*, **26**:32–46.

Aristegui J (2002) Dissolved organic carbon support of respiration in the dark ocean. *Science*, **298**:1967–1967.

Arístegui J, Duarte CM, Gasol JM, Alonso-Sáez L (2005) Active mesopelagic prokaryotes support high respiration in the subtropical northeast Atlantic Ocean. *Geophysical Research Letters*, **32**:3608.

Arístegui J, Gasol JM, Duarte CM, Herndl GJ (2009) Microbial oceanography of the dark ocean's pelagic realm. *Limnology and Oceanography*, **54**:1501–1529.

Bork P, Bowler C, de Vargas C *et al.* (2015) Tara Oceans studies plankton at planetary scale. *Science*, **348**:873–873.

Brown M V, Philip GK, Bunge JA *et al.* (2009) Microbial community structure in the North Pacific ocean. *The ISME Journal*, **3**:1374–86.

Carlson C, Ducklow H, Michaels A (1994) Annual flux of dissolved organic carbon from the euphotic zone in the northwestern Sargasso Sea. *Nature*, **371**:405–408

Chow C-ET, Sachdeva R, Cram J a *et al.* (2013) Temporal variability and coherence of euphotic zone bacterial communities over a decade in the Southern California Bight. *The ISME Journal*, **7**:1–15.

Countway PD, Gast RJ, Dennett MR *et al.* (2007) Distinct protistan assemblages characterize the euphotic zone and deep sea (2500 m) of the western North Atlantic (Sargasso Sea and Gulf Stream). *Environmental microbiology*, **9**:1219–1232.

Cram J a, Xia LC, Needham DM *et al.* (2015) Cross-depth analysis of marine bacte-

rial networks suggests downward propagation of temporal changes. *The ISME Journal*, **9**: 2573-2586.

Crump BC, Adams HE, Hobbie JE, Kling GW (2007) Biogeography of bacterioplankton in lakes and streams of an arctic tundra catchment. *Ecology*, **88**:1365–1378.

DeLong EF, Franks DG, Alldredge AL (1993) Phylogenetic diversity of aggregate-attached vs. free-living marine bacterial assemblages. *Limnology and Oceanography*, **38**:924–934.

Ducklow HW, Steinberg DK, Buesseler KO (2001) Upper ocean carbon export and the biological pump. *Oceanography*, **14**:50–58.

Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, **26**:2460–1.

Eiler A, Heinrich F, Bertilsson S (2011) Coherent dynamics and association networks among lake bacterioplankton taxa. *The ISME Journal*, **6**:330–342.

Field K, Gordon D, Wright T *et al.* (1997) Diversity and depth-specific distribution of SAR11 cluster rRNA genes from marine planktonic bacteria. *Applied and Environmental Microbiology*, **63**:63–70.

Friedline CJ, Franklin RB, McCallister SL, Rivera MC (2012) Bacterial assemblages of the eastern Atlantic Ocean reveal both vertical and latitudinal biogeographic signatures. *Biogeosciences*, **9**:2177–2193.

Fuhrman JA, Davis AA (1997) Widespread Archaea and novel bacteria from the deep sea as shown by 16S rRNA gene sequences. *Oceanographic Literature Review* **9**:1025.

Galand PE, Potvin M, Casamayor EO, Lovejoy C (2010) Hydrography shapes bacterial biogeography of the deep Arctic Ocean. *The ISME Journal*, **4**:564–76.

Ghiglione J-F, Galand PE, Pommier T *et al.* (2012) Pole-to-pole biogeography of surface and deep marine bacterial communities. *Proceedings of the National Academy of Sciences USA*, **109**:17633–17638.

Giovannoni SJ, Rappé MS, Vergin KL, Adair NL (1996) 16S rRNA genes reveal stratified open ocean bacterioplankton populations related to the Green Non-Sulfur bacteria. *Proceedings of the National Academy of Sciences USA*, **93**:7979–7984.

Guidi L, Chaffron S, Bittner L *et al.* (2016) Plankton networks driving carbon export in the oligotrophic ocean. *Nature*, **532**:465–470.

Hansell DA (2002) DOC in the Global Ocean Carbon Cycle. In: Biogeochemistry of Marine Dissolved Organic Matter , pp. 685–715. Elsevier.

Hansell DA, Ducklow HW, Macdonald AM, O-Neil Baringer M (2004) Metabolic poise in the North Atlantic Ocean diagnosed from organic matter transports. *Limnology and Oceanography*, **49**:1084–1094.

Herndl GJ, Reinthaler T (2013) Microbial control of the dark end of the biological pump. *Nature Geoscience*, **6**:718–724.

Hu A, Jiao N, Zhang R, Yang Z (2011) Niche partitioning of marine group I Crenarchaeota in the euphotic and upper mesopelagic zones of the East China Sea. *Applied and Environmental Microbiology*, **77**:7469–78.

Jiao N, Luo T, Zhang R *et al.* (2014) Presence of *Prochlorococcus* in the aphotic waters of the western Pacific Ocean. *Biogeosciences*, **11**:2391–2400.

Karsenti E, Acinas SG, Bork P *et al.* (2011) A holistic approach to marine eco-systems biology. *PLoS Biology*, **9**:e1001177.

Kimball Jr. JF, Corcoran EF, Wood FEJ (1963) Chlorophyll-containing microorganisms in the aphotic zone of the oceans. *Bulletin of Marine Science*, **13**:574–577.

Kultima JR, Sunagawa S, Li J *et al.* (2012) MOCAT: A Metagenomics Assembly and Gene Prediction Toolkit (JA Gilbert, Ed,). *PLoS One*, **7**:e47656.

Lévy M, Jahn O, Dutkiewicz S, Follows MJ (2014) Phytoplankton diversity and community structure affected by oceanic dispersal and mesoscale turbulence. *Limnology and Oceanography*, **4**:67–84..

Livingston G, Jiang Y, Fox JW, Leibold M a. (2013) The dynamics of community assembly under sudden mixing in experimental microcosms. *Ecology*, **94**:2898–2906.

Lochte K, Turley CM (1988) Bacteria and cyanobacteria associated with phytodetritus in the deep sea. *Nature*, **333**:67–69.

Logares R, Sunagawa S, Salazar G *et al.* (2013) Metagenomic 16S rDNA Illumina tags are a powerful alternative to amplicon sequencing to explore diversity and structure of microbial communities. *Environmental Microbiology*, **16**:2659–2671.

Lovejoy C, Massana R, Pedrós-Alió C (2006) Diversity and distribution of marine microbial eukaryotes in the Arctic Ocean and adjacent seas. *Applied and environmental microbiology*, **72**:3085–3095.

Minchin PR (1987) An evaluation of the relative robustness of techniques for ecological ordination. *Vegetatio*, **69**:89–107.

Moeseneder MM, Winter C, Herndl GJ (2001) Horizontal and vertical complexity of attached and free-living bacteria of the eastern Mediterranean Sea, determined by 16S rDNA and 16S rRNA fingerprints. *Limnology and Oceanography*, **46**:95–107.

Oksanen J, Blanchet FG, Kindt R *et al.* (2015) vegan: Community Ecology Package.

Pedrós-Alió C, Brock TD (1983) The importance of attachment to particles for planktonic bacteria. *Archiv für Hydrobiologie*, **98**:354–379.

Pesant S, Not F, Picheral M *et al.* (2015) Open science resources for the discovery and analysis of Tara Oceans data. *Scientific Data*, **2**:150023.

Pham VD, Konstantinidis KT, Palden T, DeLong EF (2008) Phylogenetic analyses of ribosomal DNA-containing bacterioplankton genome fragments from a 4000 m vertical profile in the North Pacific Subtropical Gyre. *Environmental microbiology*, **10**:2313–30.

Pruesse E, Quast C, Knittel K *et al.* (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic acids research*, **35**:7188–96.

R Core Team (2016) R: A language and environment for statistical computing.

Rillig MC, Antonovics J, Caruso T et al. (2015) Interchange of entire communities: microbial community coalescence. *Trends in Ecology & Evolution*, **30**:470–476.

Rocap G, Larimer FW, Lamerdin J *et al.* (2003) Genome divergence in two Prochlorococcus ecotypes reflects oceanic niche differentiation. *Nature*, **424**:1042–7.

Ruiz-González C, Niño-García JP, del Giorgio P a. (2015) Terrestrial origin of bacterial communities in complex boreal freshwater networks. *Ecology Letters*, **18**:1198–1206.

Sampayo EM, Franceschinis L, Hoegh-Guldberg O, Dove S (2007) Niche partitioning of closely related symbiotic dinoflagellates. *Molecular Ecology*, **16**:3721–33.

Shmida A, Wilson M V. (1985) Biological determinants of species diversity. *Journal of Biogeography*, **12**:1.

Souffreau C, Pecceu B, Denis C, Rummens K, De Meester L (2014) An experimental analysis of species sorting and mass effects in freshwater bacterioplankton. *Freshwater Biology*, **59**:2081–2095.

Sunagawa S, Coelho LP, Chaffron S *et al.* (2015) Structure and function of the global ocean microbiome. *Science*, **348**:1261359–1261359.

Treusch AH, Vergin KL, Finlay L a *et al.* (2009) Seasonality and vertical structure of microbial communities in an ocean gyre. *The ISME Journal*, **3**:1148–1163.

Vergin KL, Beszteri B, Monier A *et al.* (2013) High-resolution SAR11 ecotype dynamics at the Bermuda Atlantic Time-series Study site by phylogenetic placement of pyrosequences. *The ISME Journal*, **7**:1322–32.

Wilkins D, van Sebille E, Rintoul SR, Lauro FM, Cavicchioli R (2013) Advection shapes Southern Ocean microbial assemblages independent of distance and environment effects. *Nature communications*, **4**:2457.

Zinger L, Amaral-Zettler L a., Fuhrman J a. *et al.* (2011) Global patterns of bacterial beta-diversity in seafloor and seawater ecosystems. *PLoS One*, **6**:e24570.

General discussion

The main objective of this thesis has been to describe the spatial organization of the prokaryotic communities in the ocean to ultimately provide a global understanding of how are organized and which are the processes that structure them. For this purpose, we used high-throughput sequencing techniques on the samples collected during two worldwide sampling expeditions that had the aim of covering the two main spatial axes of variation in the ocean, that is the vertical and horizontal axes. Our approach has allowed the drawing of a global picture of the bathypelagic realm, and the establishment of general patterns in the vertical organization of the epipelagic and mesopelagic environments. We compile below the sets of patterns that we have identified.

The deep ocean microbes unveiled

Community structure of bathypelagic prokaryotes

This thesis incorporates a worldwide dimension in the description of the prokaryotic diversity of the bathypelagic free-living and particle-attached Bacteria and Archaea, which did not exist when I started this work. I provide in Chapter 1 the first local and global estimates of bathypelagic prokaryotic species richness of this realm: the communities studied had an average richness of ~ 650 OTUs per site but only ~3,500 unique OTUs were found for the whole dataset (that covers the tropical and subtropical oceans reasonably well, but leaves aside the colder marine sites). Although this is a minimum estimate of bathypelagic richness, a richness plateau was reached as more samples were included (Fig. 1A & 1B), indicating that the sites sampled during the Malaspina 2010 Expedition seem to correctly cover the geographical variation within the sampled bathypelagic ocean. Such a plateau was not observed in the case of the epipelagic and mesopelagic oceans studied in Chapter 5 through the *Tara* Oceans Expedition (Fig. 1C and 1D). Although a fair comparison is not possible between these two datasets because different methodologies were used (amplicon 16S vs 16 miTAGs; but see the section below) the bathypelagic communities seem to be much more homogeneous as compared to those from the upper layers. This is possibly related to the much more temporally stable environmental conditions of the deep-ocean (observed range of temperature: 0.5 – 2.2 ºC), compared to the surface ocean (observed range of temperature: 0.5 – 30.5 ºC), where seasonality has been demonstrated to be a strong determinant of bacterial community structure (e.g. Fuhrman *et al.* 2006; Brown *et al.* 2009).

We also describe in Chapter 1 the spatial distribution of the main prokaryotic inhabitants of the bathypelagic ocean by using a framework that, in the absence of a consensus denomination, we will call the *classical biogeographical framework* (for a review of this ap-

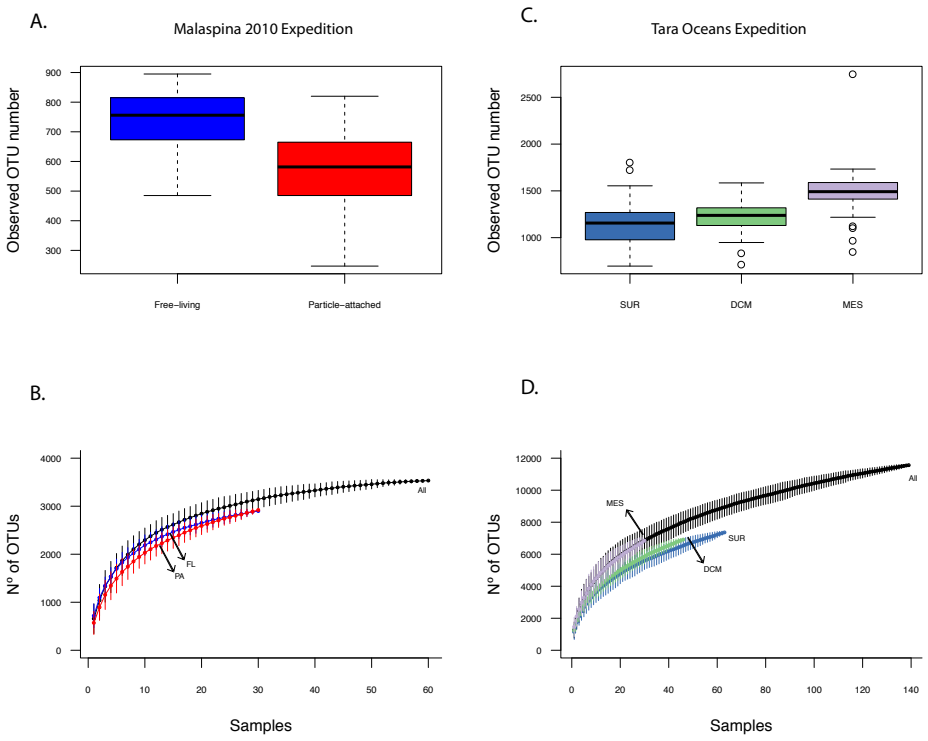


Figure 1: Prokaryotic richness observed in the Malaspina and *Tara* Ocens datasets. The number of OTUs of the free-living and particle attached communities in the bathypelagic Malaspina dataset (A, corresponding to Chapters 1, 2 and 3) and the surface, DCM and mesopelagic samples from the *Tara* Oceans' dataset (C, corresponding to Chapters 4 and 5). Sample-based accumulation curves for the same two datasets (B and D). The Malaspina dataset corresponds to amplicon sequencing while the *Tara* Oceans dataset correspond to miTags. Note the differences in the vertical scales (see corresponding chapters for details).

proach see Martiny *et al.* 2006; Hanson *et al.* 2012). It is a community-centered approach: it considers the sampled communities as the units of study and is based on the description of the composition and the degree of similarity between the different communities. Its main goal is to test for non-random patterns in the composition of communities, i.e. to demonstrate the existence of biogeography. The biogeographical patterns are then related to the geographical distance between sites and to their differences in environmental characteristics in order to describe what are the ecological factors driving variations in the communities. This is especially useful as practical statistical procedures exists in order to link the geographical variation to the potential ecological processes generating them (Hanson *et al.* 2012), which, at the same time, are well-rooted in theoretical constructs

such as the metacommunity concept (Leibold *et al.* 2004) or the neutral theory of biodiversity (Hubbell 2008). By applying this approach, we demonstrate that the communities belonging to the same water masses tended to be similar and different to the communities belonging to different water masses. This corroborates, at a world-wide scale, the previously proposed characteristics of the deep-ocean's water masses as bio-oceanographical islands (Agogué *et al.* 2011). However, we noticed different geographical patterns for the free-living (FL) and the particle-attached (PA) communities within the bathypelagic: only the PA communities were different between "deep-ocean basins". The basins are defined as the semi-isolated areas originated by the presence of submarine mountains (Chapter 1, Fig. 5). This is the first time that the isolation of different areas of the ocean by the submerged terrain is proven to drive the composition of marine microbial communities. These differences in the composition between basins could be related to differences in the apparent oxygen utilization (a proxy for the mixing and aging of water masses) and some degree of dispersal limitation was proven to affect the similarity between PA communities. On the contrary, neither dispersal limitation nor "basin effect" was observed for the FL communities. Indeed, the FL communities had a much weaker biogeographical signal, consistent with the main finding of Chapter 3, i.e. the positive relationship between the attachment preference of the bathypelagic prokaryotes and their potential metabolic activity (measured as the rRNA:rDNA ratio). It can be hypothesized that the FL prokaryotes have lower levels of metabolic activity (and presumably lower growth rates) and thus the effect of environmental filtering, which would lead to the biogeographical differentiation between communities, is consequently also weaker.

Two prokaryotic life strategies

In Chapter 1 we demonstrated a clear difference in the taxonomical composition of the FL and PA communities, as well as the existence of differences in their biogeographical organization. This was done within the *classical biogeographical framework* (see above), which is based on the description of communities as the unit of study. However, this community-centered approach does not exploit the taxonomical identity and phylogenetic relatedness of the members of the communities that are responsible for the observed biogeographical differences. In Chapters 2 and 3 we adopted a different approach, which is centered in the organisms composing such communities: we use the operational taxonomic units (OTUs) as our unit of study. This OTU-based perspective allows us to address different ecological and evolutionary questions not possible with the community-based approach. We demonstrate in Chapter 2 that two different community types according to the attachment to particles do exist for the bathypelagic prokaryotes, i.e. the FL and the PA

communities that have different lifestyles. Most bathypelagic prokaryotes exhibited non-random distributions in respect to the FL and PA fractions, indicating that the preference for either a FL or PA lifestyle was general in the bathypelagic realm. We also demonstrated that these two lifestyles are highly conserved from a phylogenetic perspective: Classes or Phyla are coherent in their particle-related lifestyle, which indicates that transitions between the FL and PA lifestyle have been rare at an evolutionary scale. These findings depict a bathypelagic realm where the axis defined by the dissolved/particulate nature of the organic matter has a key role in understanding the existing prokaryotic diversity and its structure.

The bathypelagic prokaryotes lifestyle is also linked with their activity

We repeated in Chapter 3 the OTU-based approach incorporating the study of the ribosomal RNA (rRNA) from the same set of samples to the previous study of the rRNA gene (Chapters 1 and 2). The rRNA has been used to estimate the potential metabolic activities of the prokaryotes, although not without discussion (Blazewicz *et al.* 2013; Lankiewicz *et al.* 2015). We described for the first time a positive relationship between the particle-attachment preference of the bathypelagic prokaryotes and their potential activity (Chapter 3, Fig. 3) that appeared to be globally consistent, at least for the bathypelagic of the tropical and subtropical ocean. This indicates that the two particle-related community types described in Chapter 2 in fact correspond to two growth strategies: highly active prokaryotes that live preferentially attached to particulate organic matter, and less active prokaryotes that make a living on the dissolved organic pool. The key role of the various dissolved (DOM) and particulate organic matter (POM) pools in the deep ocean has been largely recognized and its biogeochemical implications are still a matter under discussion (Herndl & Reinthaler 2013). With this thesis we incorporate a microbial-centered perspective to this discussion and provide evidence indicating that the dissolved or particulate nature of the bathypelagic organic matter (or at least the pools that can be separated by 0.8 μm) has clearly influenced the evolutionary history of Bacteria and Archaea. As a result, the present-day prokaryotic communities in the deep ocean are composed of two pools of organisms corresponding to two life history strategies. On the one hand, free-living prokaryotes with few ribosomes per cell that probably correspond to slow-growers. We hypothesize here that this life strategy is an adaptation to the very diluted DOM compounds in the bathypelagic which, in fact, have been proven to limit microbial growth (Arrieta *et al.* 2015). On the other hand, the particle-attached prokaryotes would contain a higher number of ribosomes per cell. The much higher number of ribosomes per cell of the PA prokaryotes would be necessary for the maintenance of

relatively higher growth rates supported by the much higher C concentration of the POM (Minor *et al.* 2003), compared to the DOM or, alternatively, could also be an adaptation to the higher variability of the POM quantity and quality supply to the deep ocean, compared to the rather stable concentrations of the bathypelagic DOM (Hansell 2002). The presence of a high basal number of ribosomes would allow a fast metabolic response for the consumption of transient particles and its investment into growth. Similarly to the life history strategies proposed here, other strategies such as the *defense* or *competition specialists* have been proposed for marine microbes based on the trade-off between defense and growth capacity (Winter *et al.* 2010). Functional and experimental analyses, which are out of the scope of this thesis, would be necessary in the future for elucidating the metabolic basis of such life history strategies.

A process-based approach to the vertical structure of the ocean

While Chapters 1 to 3 of this thesis dealt with the horizontal spatial variability of the prokaryotic communities within the bathypelagic depth layer, the rest of this thesis addresses the vertical variability of prokaryotic communities in the epipelagic and mesopelagic ocean. In Chapter 4 we evaluate and extend the *miTags* approach, a method previously developed in our group for the utilization of rDNA reads from metagenomes to characterize microbial communities (Logares *et al.* 2013). This methodological chapter evaluates the performance of the original miTags approach with mock and real communities and proposes an improvement based on the unambiguous binning of the rDNA reads that corrects the limitations introduced by the original version of the method. As a result, we developed and made public the *mtagger* R package, which implements both the original and the modified miTags method. This later method is applied to the *Tara* Oceans metagenomes for extracting the data used in the last chapter of the thesis, Chapter 5.

In Chapter 5 the vertical structure of the prokaryotic communities at a global scale is analyzed through the extraction and binning of the 16S rDNA reads from the metagenomes sequenced within the *Tara* Oceans expedition by using the methods developed and tested in the previous chapter. This analysis depicts a two-realm ocean where the composition of communities is organized in two very segregated clusters of samples: the photic communities, including the samples from the surface ocean and from the deep-chlorophyll maximum (DCM), and the aphotic communities, i.e. the mesopelagic samples. The consistence in the vertical clustering of the prokaryotic diversity in a dataset including samples from the main oceans indicates that the vertical axis is probably the most important one along which prokaryotic diversity is structured in the ocean. This pattern

had already been observed in the International Census of Marine Microbes (ICoMM) dataset (Zinger *et al.* 2011) and we also described it in the *Tara* Oceans analysis (Sunagawa *et al.* 2015). However, the vertical segregation of prokaryotic diversity into photic and aphotic communities is not perfect, as a considerable proportion of the diversity is shared between layers. Moreover, this shared diversity is not symmetrically distributed between the photic and aphotic layers: a high proportion of OTUs within a single station are unique to the mesopelagic ocean, while the proportion of OTUs that are unique to the surface or DCM samples is much lower. This indicates that most of the prokaryotes present in the photic waters are also found in the mesopelagic. However, a considerable proportion of the mesopelagic prokaryotes are never found in photic waters, indicating that directionality exists in the vertical organization of the marine prokaryotic communities. Both the vertical segregation of the communities and this asymmetrical distribution of the shared diversity between the photic and aphotic ocean are consistent with current knowledge on the ocean ecological functioning, basically defined by the vertical segregation of the main ecological processes involved in the carbon cycle: primary production through photosynthesis restricted to the photic ocean, export of the fixed carbon to the deep ocean by the biological pump and its ultimate remineralization through the whole water column by the heterotrophs.

In Chapter 5 we adopt a process-based approach to explore the biogeography of prokaryotic communities in the vertical gradient. Process-based approaches in community ecology consist in the assumption of a few simple ecological process acting on the organization of communities, which are formulated generally as mathematical models and whose predictions may be tested against real data of the structure of ecological communities. We developed in Chapter 5 a model framework (the *disperflux model*) where a directional whole-community transport event between communities is explicitly assumed as the process potentially generating the similarity patterns observed, a process that although had been proposed theoretically (Rillig *et al.* 2015), had never been tested. The model allows us to interpret the similarity of the communities in terms of connectivity and to incorporate directionality in the measure of pairwise similarities between ecological communities. We observe and describe a quasi-universal decay relationship between the connectivity of communities and the vertical distance separating them. This suggests that a common process is responsible for the vertical segregation of prokaryotic communities in the vertical gradient throughout the main oceans. Our model does not require of particle export as the process explaining connectivity between communities, although this has been proven to explain the distribution of specific taxa (Guidi *et al.* 2016) as, unlike in the previous chapters, the dataset analyzed in Chapter 5 comprised

only of free-living prokaryotes (prokaryotes retained between the 0.2 μm and 3 μm size filters), which may explain the absence of a correlation between vertical particle fluxes and the connectivity estimated through the *disperflux model*, which would be expected based on the assumed central role of the carbon pump in the general functioning of the ocean. We identified, however, vertical mixing as a process likely driving the connectivity between the surface ocean and the DCM. This decay relationship failed to explain the data in 5 locations, in which we could prove that an event of massive transport of individuals from the surface ocean to the mesopelagic was capable of generating the similarity patterns observed between the communities of these locations.

The future goal: a 3D perspective of the ocean's microbial diversity

The introduction of high-throughput sequencing (HTS) techniques applied to the 16S ribosomal gene has represented a key step in the taxonomical characterization of marine microbial communities by making possible quantitative estimates of the abundance of thousands of fine-scale taxonomical units within a sample. However, most of the steps of an HTS-based study lack standardized procedures and introduce a considerable degree of variation in these estimates, something that makes the combination of datasets originated from different studies virtually impossible. As a consequence, a worldwide study of marine microbial diversity has been hampered by the impossibility of merging the data from different studies. This limitation has been partially solved by the development of global initiatives to survey the marine diversity with coherent sampling protocols and sequencing techniques, such as the ICoMM, OSD, Malaspina 2010 and the *Tara* Oceans expeditions among others. Two of of them constitute the data sources of this thesis. However, the impossibility of directly combining the data from both expeditions, which cover two complementary portions of the ocean (the bathypelagic ocean and the surface-to-mesopelagic ocean), still limits a comprehensive view of the microbial diversity in the whole ocean.

In Chapter 4 we extend a previously developed method that extracts the rDNA reads from metagenomes for the description of the taxonomical diversity of microbial communities. This method circumvents the main bias-introducing steps in amplicon-based studies, by in-silico selecting the rDNA genes after the sequencing of a metagenome, instead of using the amplification of the 16S gene with a PCR before sequencing. Thus, this approach, as well as similar approaches recently developed (Bengtsson-Palme *et al.* 2015; Ramazzotti *et al.* 2015; Guo *et al.* 2016; Xie *et al.* 2016) are alternatives to the amplicon sequencing. The lack of a PCR eliminates some of the steps that prevent the combination

of amplicon datasets, mainly the biases introduced by PCR amplification dynamics and the choice of different PCR primers and different variable regions to be amplified within a gene. Thus, it converts the miTags into a feasible approach for successfully combining metagenomic datasets. In fact, although not developed in Chapter 4, we have attempted to use this approach for merging two metagenomic datasets from the *Tara* Oceans and Malaspina expeditions, as a preliminary exercise within this Thesis. These datasets correspond to the metagenomes analyzed in Chapter 5 from the *Tara* Oceans Expedition and 7 vertical profiles (not previously used in the chapters) from the Malaspina Expedition, ranging from the surface ocean to the bathypelagic ocean. In this way, these two datasets combine samples from the epipelagic and mesopelagic ocean coming from both projects. The reads corresponding to the 16S gene from these two datasets were extracted and processed using the *mtagger package* described in Chapter 4. A linear relationship exists for the three layers between the number of sequences analyzed and the number of OTUs detected, which is consistent between datasets (Fig. 2). This indicates that the amount of retrieved diversity as sequencing effort increases is comparable for both datasets when analyzed using the miTags approach. This linear relationship seems to be specific for the epipelagic, mesopelagic and bathypelagic ocean, that is, differences in richness and diversity exist between these three oceanic layers, as expected. Moreover, the samples clustered based on their taxonomical similarity along a vertical gradient, irrespectively of the project of origin (Fig. 3). This indicates that the miTags technique seems to sufficiently bypass the potential biases of each study (i.e. volume sampled, ammount of DNA extracted, etc) and reproduce the vertical segregation of marine prokaryotic communities described in Chapter 5, when merging metagenomes from different projects. We use this preliminary exercise as a corollary of this Thesis and as a proof of concept for the possibility of jointly analyzing datasets that have been until now analyzed separately. This analysis allows to answer questions that were impossible before a successful merging of the two *Tara* Oceans and Malaspina datasets could be done.

The sustained decrease of sequencing costs in the past decade is making now feasible the generalized use of metagenomics as the first approximation to the description of microbial communities. miTags extraction allows the taxonomical description of these communities in terms of OTUs directly from metagenomic data and circumvents some of the biases introduced by the use of PCR and primer biases. However, besides being an alternative to amplicon sequencing, as was originally proposed (Logares *et al.* 2013), we demonstrate that it also allows the joint analysis of new datasets with those already existing. As a general rule, we would recommend the use of metagenomics as the first approach to characterize new microbial communities both functionally and taxonomically

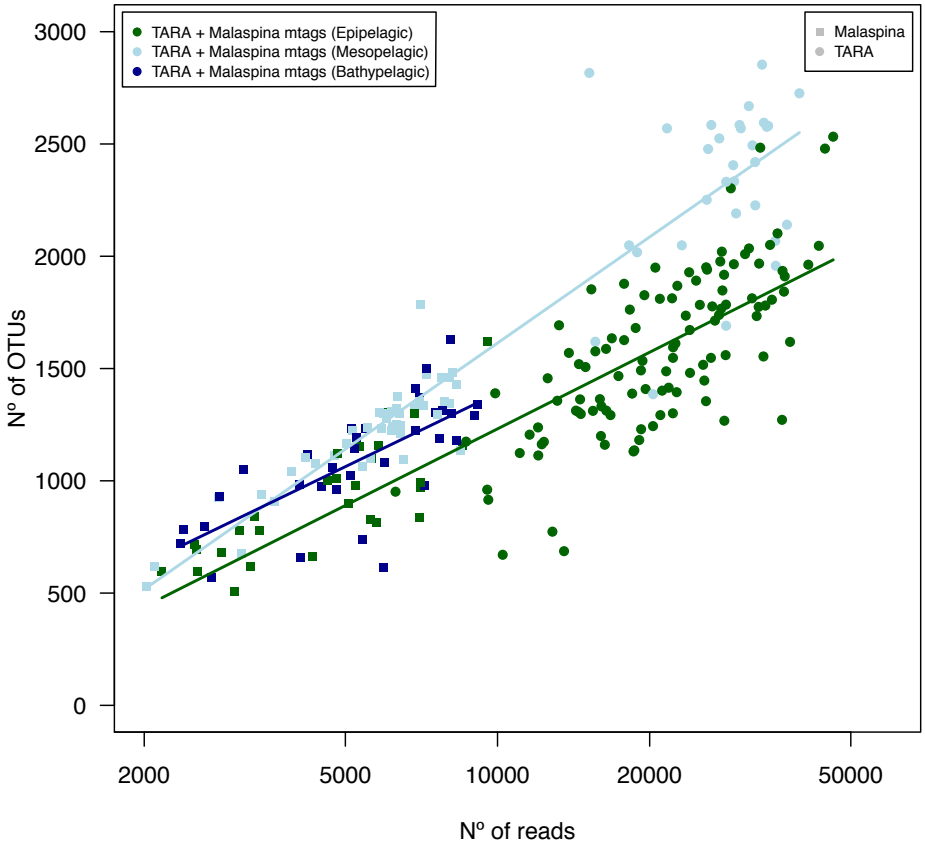


Figure 2: Sequencing effort vs. estimated OTU richness for the *Tara* Oceans-Malaspina combined dataset. Relationship between the number of reads and the number of OTUs for a dataset merging the *Tara* Ocenas dataset (see Chapter 5) and 7 epipelagic-to-bathypelagic vertical profiles from the Malaspina Expedition (not included in any chapter; see discussion). Both datasets are analyzed and merged through the technique explained in Chapter 4.

(through miTags), as has been in fact the case for the *Tara* Oceans Expedition (Sunagawa *et al.* 2015). The miTags approach, however, relies on the extraction of the reads corresponding to the rDNA gene, which constitutes a minor fraction (~0.1%) of a metagenome. A sufficiently deep sequencing of the metagenomes is thus a pre-requisite for a successful use of this approach as an alternative to amplicon sequencing. This may preclude its use for specific cases, such as when a low sequencing effort has been used, or when the object of study is the low-abundant members of microbial communities, that may be better characterized by amplicon sequencing.

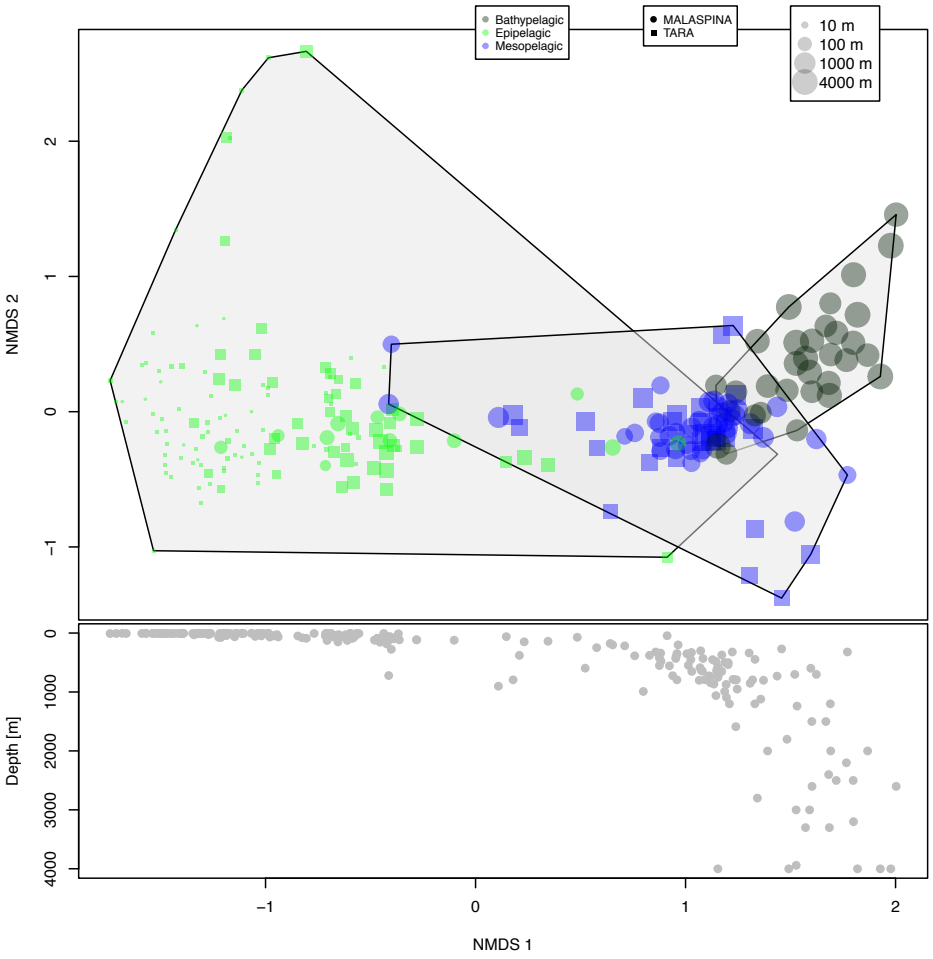


Figure 3: NMDS for the *Tara* Oceans-Malaspina combined dataset. Non–metric multidimensional scaling (NMDS) of the combined dataset described in Figure 2. The epipelagic, mesopelagic and bathypelagic samples are color-coded. The depth (m) of each sample is plotted in the bottom panel and represented in the upper panel as the size of the dots in a log scale.

I am convinced that the ultimate goal of understanding the geographical distribution of marine microbial communities along the vertical and horizontal ocean axes (i.e. a 3D perspective of the ocean's microbial biogeography) will be advanced not only by the recent arrival/design of large-scale sampling efforts but through the widespread use of techniques, such as the one developed here, for the extraction of comparable information of datasets hitherto incomparable.

References

Agogué H, Lamy D, Neal PR, Sogin ML, Herndl GJ (2011) Water mass-specificity of bacterial communities in the North Atlantic revealed by massively parallel sequencing. *Molecular Ecology*, **20**:258–274.

Arrieta JM, Mayol E, Hansman RL *et al.* (2015) Dilution limits dissolved organic carbon utilization in the deep ocean. *Science*, **34:**331–333.

Bengtsson-Palme J, Hartmann M, Eriksson KM *et al.* (2015) metaxa 2: improved identification and taxonomic classification of small and large subunit rRNA in metagenomic data. *Molecular Ecology Resources*, **15**:1403–1414.

Blazewicz SJ, Barnard RL, Daly RA Firestone MK (2013) Evaluating rRNA as an indicator of microbial activity in environmental communities: limitations and uses. *The ISME Journal*, **7**:2061–8.

Brown M V, Philip GK, Bunge JA *et al.* (2009) Microbial community structure in the North Pacific ocean. *The ISME journal*, **3**:1374–86.

Fuhrman JA, Hewson I, Schwalbach MS *et al.* (2006) Annually reoccurring bacterial communities are predictable from ocean conditions. *Proceedings of the National Academy of Sciences of the United States of America*, **103**:13104–13109.

Guidi L, Chaffron S, Bittner L *et al.* (2016) Plankton networks driving carbon export in the oligotrophic ocean. *Nature*, **532**:465–470.

Guo J, Cole JR, Zhang Q, Brown CT, Tiedje JM (2016) Microbial Community Analysis with Ribosomal Gene Fragments from Shotgun Metagenomes. , **82**:157–166.

Hansell DA (2002) DOC in the global ocean carbon cycle. In: Biogeochemistry of Marine Dissolved Organic Matter , pp. 685–715. Elsevier.

Hanson CA, Fuhrman JA, Horner-Devine MC, Martiny JBH (2012) Beyond biogeographic patterns: processes shaping the microbial landscape. *Nature Reviews Microbiology*, **10**:1–10.

Herndl GJ, Reinthaler T (2013) Microbial control of the dark end of the biological pump. *Nature Geoscience*, **6**:718–724.

Hubbell SP (2008) *The Unified Neutral Theory of Biodiversity and Biogeography*.

Lankiewicz TS, Cottrell MT, Kirchman DL (2015) Growth rates and rRNA content of four marine bacteria in pure cultures and in the Delaware estuary. *The ISME Journal*, **10**:1–10.

Leibold MA., Holyoak M, Mouquet N *et al.* (2004) The metacommunity concept: a framework for multi-scale community ecology. *Ecology Letters*, **7**:601–613.

Logares R, Sunagawa S, Salazar G *et al.* (2013) Metagenomic 16S rDNA Illumina tags

are a powerful alternative to amplicon sequencing to explore diversity and structure of microbial communities. *Environmental Microbiology*, **16**:2659–2671.

Martiny JBH, Bohannan BJM, Brown JH *et al.* (2006) Microbial biogeography: putting microorganisms on the map. *Nature reviews. Microbiology*, **4**:102–112.

Minor EC, Wakeham SG, Lee C (2003) Changes in the molecular-level characteristics of sinking marine particles with water column depth. *Geochimica et Cosmochimica Acta*, **67**:4277–4288.

Ramazzotti M, Berná L, Donati C, Cavalieri D (2015) riboFrame: an improved method for microbial taxonomy profiling from non-targeted metagenomics. *Frontiers in Genetics*, **6**:329.

Rillig MC, Antonovics J, Caruso T et al. (2015) Interchange of entire communities: microbial community coalescence. *Trends in Ecology & Evolution*, **30**:470–476.

Sunagawa S, Coelho LP, Chaffron S *et al.* (2015) Structure and function of the global ocean microbiome. *Science*, **348**:1261359–1261359.

Winter C, Bouvier T, Weinbauer MG, Thingstad TF (2010) Trade-offs between competition and defense specialists among unicellular planktonic organisms: the "killing the winner" hypothesis revisited. *Microbiology and Molecular Biology Reviews*, **74**:42–57.

Xie C, Goi CLW, Huson DH, Little PFR, Williams RBH (2016) RiboTagger: fast and unbiased 16S/18S profiling using whole community shotgun metagenomic or metatranscriptome surveys. *BMC Bioinformatics*, **17**:508.

Zinger L, Amaral-Zettler L a., Fuhrman J a. *et al.* (2011) Global patterns of bacterial beta-diversity in seafloor and seawater ecosystems. *PLoS One*, **6**:e24570.

Supplementary information

SUPPLEMENTARY INFORMATION CHAPTER 1

Supplementary Material and Methods

DNA extraction

The filters were cut in small pieces with sterile razor blades and half of each filter was resuspended in 3 ml of lysis buffer (40 mM EDTA, 50 mM Tris-HCl, 0.75 M sucrose). Lysozyme (1 mg ml^{-1} final concentration) was added and the samples were incubated at 37ºC for 45 min with slight movement. Then, sodium dodecyl sulfate (SDS, 1% final concentration) and proteinase K (0.2 mg ml^{-1} final concentration) were added and the samples were incubated at 55ºC for 60 min under slight movement. The lysate was collected and processed with the standard phenol-chloroform extraction procedure: an equal volume of Phenol:$CHCl_3$:IAA (25:24:1, vol:vol:vol) was added to the lysate, carefully mixed and centrifuged 10 min at 3,000 rpm. Then the aqueous phase was recovered and the procedure was repeated. Finally, an equal volume of $CHCl_3$:IAA (24:1, vol:vol) was added to the recovered aqueous phase in order to remove residual phenol. The mixture was centrifuged and the aqueous phase was recovered for further purification. The aqueous phase was then concentrated by centrifugation with a Centricon concentrator (Millipore, Amicon Ultra-4 Centrifugal Filter Unit with Ultracel-100 membrane). Once the aqueous phase was concentrated, this step was repeated three times adding 2 ml of sterile MilliQ water each time in order to purify the DNA. After the third wash, between 100 and 200 µl of purified total genomic DNA product per sample could be recovered. Extracted DNA was quantified using a Nanodrop ND-1000 spectrophotometer (NanoDrop Technologies Inc, Wilmington, DE, USA) and the Quant_iT dsDNA HS Assay Kit with a Qubit fluorometer (Life Technologies, Paisley, UK).

Amplicon sequencing and sequence data processing

All library construction and sequencing was carried out at the JGI (www.jgi.doe.gov) following a pipeline previously published (Caporaso *et al.*, 2011). Briefly, the variable region V4 of the 16S rDNA gene was amplified using primers F515/R806 (5'-GTGCCAGCMGCCGCGGTAA-3' / 5'-GGACTACHVGGGTWTCTAAT-3'). The amplicons were sequenced using Illumina MiSeq with 2x250 bp reads configuration. Before sequencing, PhiX spike-in shotgun library reads were added to the amplicons pool for a final concentration of about 20-25% of the pair-end reads library as an internal standard.

The reads were first scanned for PhiX reads and contaminants (e.g. Illumina adapter sequences) and all disrupted pair-end reads (every read pair for which one read has been

lost due to the screening) were discarded. The remaining reads were trimmed to 165 bp and assembled using FLASH software (Magoč & Salzberg, 2011) and primer sequences were removed from the assembled reads. The minimum overlap length was set to 20bp and the rest of parameters were used as default. Assembled reads were trimmed from both 5' and 3' ends using a 20 bp sliding window (mean quality threshold >30). Trimmed reads with more than 5 Ns or 10 nucleotides below quality 15 were discarded. Filtered reads were then clustered using USEARCH (Edgar, 2010) at 99% identity and clusters having abundances less than 3 reads were discarded. An extra clustering step at 97% identity was performed on the remaining clusters providing a final OTU dataset from which the most abundant sequence of each OTU was considered as its representative sequence. Finally, these representative OTU sequences were checked for chimeras using both the Chimera Slayer algorithm as implemented in software MOTHUR and the UCHIME *de novo* and reference-based algorithms (Edgar *et al.*, 2011). The OTUs identified as chimeric sequences by any of these methods were removed. Non-chimeric OTUs were taxonomically annotated using the BLAST-based classifier within the QIIME pipeline using the SILVA database (release 111) as reference.

Primer coverage of F515 and R806 primers and others primers used in previous amplicon sequencing studies were checked using Test Probe 3.0 software (http://www.arb-silva.de/?id=650) allowing 0 and 1 mismatch for every primer. Coverage values for Archaea and Bacteria were collected for every primer using the SILVA reference database.

Automated Ribosomal Intergenic Spacer Analysis (ARISA)

Intergenic Transcribed Spacers (ITS) from DNA samples were amplified using PCR with primers ITSF/ITSReub (5'-GTC GTA ACA AGG TAG CCG TA-3' / 5'-GCC AAG GCA TCC ACC-3') and using the fluorescently labeled forward primer (5-FAM). The PCR mixture (40 µl) contained a final concentration of 0.25 ng μl^{-1} of DNA template, 250 nM of each primer, 250 µM of each dNTP, 2.5 mM $MgCl_2$, 3 units of a Taq DNA polymerase (Invitrogen-Life Technologies), 40 ng μl^{-1} of BSA and the enzyme buffer. PCR cycling, carried out in an automated thermocycler (BioRad), was: initial denaturation at 94ºC for 2 min; 32 cycles with denaturation at 94ºC for 15 sec, annealing at 55ºC for 30 sec and extension at 72ºC for 3 min; and a final extension at 72ºC for 9 min. PCR products, stored at 4ªC, were purified with the QIAquick PCR Purification Kit (Qiagen) and quantified with NanoDrop 1000 (Thermo Fisher Scientific Inc., Wilmington, DE).

Each purified PCR product was added to a mix composed of 10 µl of Hi-Di formamide, 0.3 µl of the internal size standard X-Rhodamine MapMarker 1000 (ROX) (BioVentures). The PCR product final concentration was 1 ng μl^{-1}. The samples were run using a genetic

analyzer with 36 cm Capillary Array and 3130 POP-7 Polymer (Applied Biosystems). The electropherograms were then analyzed using the GeneMarker analysis software (Softgenetics) for size calibration. Binning of the peaks into OTUs was done using R scripts as in (Ramette, 2009) and available at http://www.mpi-bremen.de/en/Software_2.html. A minimum RFI cutoff value of 0.01% and a windows size of 2 bp were used.

Illumina metagenomes and data analyses

We used the taxonomic assignation of the metagenomic reads from 46 metagenomes from those samples also analyzed by amplicon 16S sequencing to compare the relative abundances at Phylum-level. Metagenomic sequencing and analyses was performed at the JGI. Samples were sent in a 96-well plate and unamplified libraries were generated using a modified version of Illumina's TruSeq DNA sample preparation protocol and KAPA Biosystem's Library Preparation kit for Illumina. Sample preparation was performed on a PerkinElmer Sciclone NGS G3 Liquid Handling Workstation capable of processing 96 plate-based samples in parallel. Two hundred nanograms of genomic DNA were used for each sample and the DNA was sheared using a Covaris LE220 focused-ultrasonicator to generate sheared fragments of 270 bp in length. The sheared DNA fragments were size selected by SPRI to 270 bp and the selected fragments were then end-repaired, A-tailed, and ligated with Illumina compatible sequencing adaptors containing a unique molecular barcode (index) for each sample library.

The prepared sample libraries were quantified using KAPA Biosystem's next-generation sequencing library qPCR kit and run on a Roche LightCycler 480 real-time PCR instrument. The quantified sample libraries were then pooled together into pools of 12 libraries each. These pools were prepared for sequencing on the Illumina HiSeq sequencing platform utilizing a TruSeq paired-end cluster kit, v3, and Illumina's cBot instrument to generate clustered flowcells for sequencing. Sequencing of the flowcells was performed on the Illumina HiSeq2000 sequencer using Illumina TruSeq SBS sequencing kits, v3, following a 2x150 indexed high-output run recipe. Data processing was done using the standard metagenomic annotation pipeline within the IMG/M platform (information available at https://img.jgi.doe.gov/m/doc/MetagenomeAnnotationSOP.pdf). The composition at a Phylum level of the 46 metagenomes was assessed using the Phylogenetic Distribution of Genes online tool, which allows assessing the phylogenetic composition of a metagenomic sample based on the distribution of best BLAST hits of protein-coding genes in the genomic IMG dataset. A 60% identity was used as a minimum cutoff and the estimated gene copies were used instead of the raw gene count.

Primer coverage analysis and inter-technique validation

It is well-known that the use of PCR in the 16S rRNA amplicon sequencing data may introduce potential biases in both diversity and relative abundance estimation (Acinas *et al.*, 2005; Hong *et al.*, 2009; Engelbrektson *et al.*, 2010) and we decided to investigate the degree of coverage of the primers used in this study with an in-silico a primer analysis in comparison with other primers used previously in studies of the deep ocean (see *SI*). Our primers (515F-806R) exhibited comparable diversity coverage for both Bacteria and Archaea to those used in previous Next Generation Sequencing (NGS) studies of bathypelagic samples (Table S7). Moreover, the coverage values for both domains were comparable to those obtained using primers that specifically target each domain, either Bacteria only or Archaea only. Similarly, coverage values were very high when computed for lower rank taxa (details not presented) with the exception of SAR11 and Propionibacteriales clades, whose abundances must be taken with care as most of its members have one mismatch with the primer set used. However when allowing for one mismatch, the coverage was particularly good for both domains (both with coverage values >97%) and for lower rank taxa (all with coverage values >80%) (Table S7 and details not presented). In fact, a perfect match is not required for an efficient amplification, specially when melting temperature specificities are taken into account (Sommer & Tautz, 1989; Kwok *et al.*, 1990).

Two independent techniques were additionally applied to the same samples in order to corroborate the abundance of taxa and community patterns found with the iTags. Firstly, a PCR- independent approach (metagenomes) was used to evaluate the effect of PCR amplification biases in the estimation of abundance. Although 16S rRNA gene fragments can be efficiently extracted from illumina metagenomes (miTags) to explore microbial diversity and community structure patterns (Logares *et al.*, 2014), we did not use them in this particular study because the number of miTAGs per sample was too low for obtaining significant values. We used instead the taxonomic assignment of all metagenomic reads of the different samples to estimate the relative abundance of the dominant Phyla and this was compared to the iTag-based data (see SI). Phylum level relative abundances based on iTags were highly correlated with metagenomic-based relative abundances (N= 46, Pearson r = 0.972, P-value < 0.0001; Fig. S3a). However, when analyzing the correlation group by group, we observed that some Phyla such as Deferribacteres and Gemmatimonadetes / Bacteroidetes, even though correlated, were consistently underestimated / overestimated with iTags compared to metagenomic data (Fig. S4). In addition to a PCR bias, this could also be due to an uneven distribution of reference genomes among different Phyla, which would affect the annotation of metagenomic data. Although the values of relative abundances for some Phyla may differ between techniques, differences in abundances between

samples should not be affected and thus inter-sample comparisons are robust to the sequencing technology used. Finally we used ARISA as a second approach to compare with our iTAGs results. ARISA involves also a PCR reaction but of a different phylogenetic marker -the ITS- and with different primers and therefore allowed us to compare community structure using two different gene markers with distinct PCR primer sets. ARISA-based community dissimilarities were highly correlated to iTags-based community dissimilarities (N=60, Mantel r = 0.69, P = 0.001; Fig. S3b), extending the consistency of the relative abundances found when comparing OTUs at a finer-scale and coherently what with has already been described when comparing HTS with ARISA (Gobet *et al.*, 2013).

Environmental and geographical data

The ocean's bathymetry and a set of 4 variables (depth, potential temperature, salinity and dissolved oxygen) were used to relate prokaryote's biogeography to *deep-water clusters* and *deep oceanic basins* (both defined below). Depth, potential temperature and salinity, jointly with the Apparent Oxygen Utilization (AOU) and 5 extra biotic variables (all described in Table S6) were used for the variance partitioning of beta-diversity.

Depth, potential temperature (θ), salinity (S) and dissolved oxygen concentration data were recorded by a SeaBird 911+ CTD, which was equipped with a redundant temperature and salinity sensor for intercomparison during the circumnavigation. Temperature and pressure sensors were calibrated at the SeaBird laboratory before the cruise. Salinity and dissolved oxygen sensors were calibrated against water samples measured on board with a Guild-line AUTOSAL model 8410A salinometer with a precision better than 0.002 for single samples and the potentiometric end-point Winkler method, respectively. AOU for each water sample was calculated as the difference between the saturation and measured dissolved oxygen. Oxygen saturation was obtained from θ and S using the equation in Benson & Krause, 1984. AOU values were included as an extra environmental variable and used as a proxy for water mass ageing. Additionally, prokaryotic heterotrophic activity, abundance (in situ and integrated from 0 to 200 m depth), prokaryotic biomass duplication time and the % of (High Nucleic Acid) HNA-content prokaryotes were included as variables in the analysis. Prokaryote abundance was determined by flow cytometry using standard protocols (Gasol & Giorgio, 2000). A previously-published (Calvo-Díaz & Morán, 2006) calibration curve was used to transform the cytometric signal into cell size, and cell size was converted to biomass with a standard conversion. This resulted in an average prokaryotic size of 7.1 fgC cell-1 for 4,000 meters prokaryotes. Prokaryotic heterotrophic production (PHP) was estimated from the incorporation of undiluted ^{3}H-leucine at 5 nM during 7-10 hours and collected on a 0.22 μm filter, rinsed with TCA

and the radioactivity measured in a Beckman scintillation counter after addition of cocktail. A standard 1.5 kgC mol leucine^{-1} conversion factor was used to transform leucine incorporation (activity) into biomass production. Specific growth rate was calculated as $SGR=\ln(1+{}^{Prok.Het.Prod}/_{Prok.Biomass})$. Biomass duplication time is defined as ln(2)/SGR. The environmental variables used are described in Table S6.

θ, S and AOU have been used to classify the water samples collected in this work according to their Arctic, Circumpolar or Antarctic origin. The bathypelagic waters of the World Ocean occupied during the Malaspina 2010 circumnavigation are mainly composed of North Atlantic Deep Water (NADW), Circumpolar Deep Water (CDW) and Antarctic Bottom Water (AABW). NADW is the warmest (2–4 ºC) and saltiest (34.9–35.0) of the deep ocean waters. By contrast, a potential temperature of 1.7ºC and a salinity of 34.7 characterize the CDW at Drake Passage. Finally, any sample collected during the Malaspina 2010 with a potential temperature <1.7ºC and a salinity <34.7 contains variable volumes of AABW derived from the shores of the Weddell and Ross seas, which enter the Atlantic, Indian and Pacific oceans with θ <0 ºC. As a result, the θ–S diagram of the samples collected during the circumnavigation (Fig. S2a) exhibits the V–shape characteristic of the NADW–CDW–AABW mixing triangle. These water masses also present contrasting AOU levels, with NADW being the most ventilated (lowest AOU) and the CDW the most aged (highest AOU) of the bathypelagic waters of the World Ocean (Fig. S2b).

On basis of their θ, S and AOU, the samples have been grouped into six *deep-water clusters* (Table S1) characterized by variable contributions of NADW, CDW and AABW. The deep-water cluster 1 comprises the samples with the highest θ and S (Fig. S2a) and the lowest AOU (Fig. S2b), characteristic of unmixed NADW. Note that station 32 belongs to this cluster because this sample was collected at about 3,000 m, the core of the NADW in the South Atlantic (Álvarez *et al.*, 2014) rather than at about 4,000 m. The deep-water cluster 2 is composed of samples whose lower θ and S and higher AOU point to a progressive dilution of NADW with CDW and AABW. The deep-water cluster 3 and 4 consist of the samples with the largest contribution of AABW from either the Weddell Sea (deep-water cluster 3) or the Ross Sea (deep-water cluster 4), characterized by their low potential temperature (generally <1ºC) and intermediate AOU (about 120 μmol kg^{-1} for deep-water cluster 3 and 140 for deep-water cluster 4 μmol kg^{-1}). Despite the high potential temperature of station 53 (1.3ºC), which was sampled at 3,500 m instead of at 4,000 m, this sample was included in deep-water cluster 4 (Fig. S2a). The deep-water cluster 5 is made of the samples with a potential temperature, salinity and AOU intermediate between AABW and CDW (Fig. S2). Finally, the deep-water cluster 6 comprises the purest

CDW samples. Stations 62 and 82 (white dots in Fig. S2) are difficult to classify within a deep-water cluster. They were the shallowest of all samples, being collected at 2,400 and 2,150 m, respectively. These samples are mainly composed of CDW, but mixed with small volumes of NADW, responsible for the relatively high salinity of station 62, and Antarctic Intermediate Water (AAIW), responsible for its relatively low salinity of station 82. For the purposes of this work, stations 62 and 82 were included in deep-water cluster 6.

A geographical distance matrix between sampling stations was constructed by computing the shortest distance between two sampling stations avoiding landmasses using the geographical coordinates of each sampling station. For that purpose the bathymetry across the globe at one degree intervals from (Andersson, 2004) (available in *marelac* package from R Statistical Softawre) was used to construct a raster object considering only the coordinates corresponding to elevation below 100 m, i.e. excluding land masses. This raster object was used for computing the shortest distance between sampling stations using the Dijkstra algorithm implemented as the *costDistance* function within the *gdistance* R package (Dijkstra, 1959). The same bathymetric information was used to define deep oceanic basins and locate sampling stations within one of these basins, when possible. Basins were defined as the completely or nearly completely enclosed water bodies placed below 3500 m depth (represented in Fig. 1). The few samples above 3500 m were considered to be out of any defined basin.

Variance partitioning of beta-diversity

Out of the four ecological processes involved in shaping composition and diversity within and among microbial communities (speciation, selection, dispersal and ecological drift) (Vellend, 2010), we focused only in those with potential for the generation of biogeographical patterns in prokaryotic communities, which are usually grouped into "present environmental selection" and "historical processes" (Hanson *et al.*, 2012). Permutation-based multiple regression on matrices (MRM, using *MRM* function from *ecodist* package) (Lichstein, 2006) was used to quantify the relative contribution of "present environmental selection" and "historical processes" on the biogeography of prokaryotic community composition. Present environmental selection corresponds to the influence of the current environment on the current distribution of microbial diversity and is detected when finding a significant correlation between community composition and the variables that define the environment (i.e. that define an ecological niche). Historical processes correspond to the past action of environmental selection or ecological drift in combination with some degree of dispersal limitation: if dispersal is not completely efficient, drift or past environmental selection will leave a legacy on the current distribution of microbial

communities. Thus, a significant correlation between geographical distance and community composition, after controlling for the present environmental effect, would be indicative of the action of historical processes, i.e. implying some degree of dispersal limitation (Martiny *et al.*, 2006; Nekola *et al.*, 1999). As proposed by Duivenvoorden et al. (Duivenvoorden *et al.*, 2002) the beta-diversity matrix was partitioned into four components: (i) variation explained by pure environmental heterogeneity, (ii) variation explained by pure geographical distance, (iii) variation explained by both environmental heterogeneity and geographical distance (or spatially structured environmental variation) and (iv) unexplained variation. First, environmental variables contributing to the variation in prokaryotic communities were selected using the BIOENV approach implemented as the *bioenv* function within the *vegan* R package (Clarke & Ainsworth, 1993). It consists on the selection of the best subset of environmental variables so that the Euclidean distances based on scaled environmental variables have the maximum correlation with community dissimilarities. Secondly, the R-squared of the selected environmental variables as an independent matrix (R^2_E), geographical distance as independent matrix (R^2_G), and both matrices (R^2_T) were used to compute the four components of variation above mentioned as suggested by Jones et al. (Jones *et al.*, 2006) : (i) pure environmental variation = R^2_T-R^2_G, (ii) pure geographical distance = R^2_T-R^2_E, (iii) spatially structured environmental variation = R^2_G+R^2_E-R^2_T and (iv) unexplained variation = 1-R^2_T. The geographical distance matrix was ln-transformed prior to MRM analysis, as suggested in (Martiny *et al.*, 2011).

References

Acinas SG, Sarma-Rupavtarm R, Klepac-Ceraj V, Polz MF. (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**:8966–9.

Álvarez M, Brea S, Mercier H, Álvarez-Salgado XA (2014) Mineralization of biogenic materials in the water masses of the South Atlantic Ocean. I: Assessment and results of an optimum multiparameter analysis. *Progress in Oceanography*, **123**:1–23.

Andersson JH. (2004). Respiration patterns in the deep ocean. *Geophys Research Lett* ers, **31**:L03304.

Benson BB, Krause D. (1984). The concentration and isotopic fractionation of oxygen dissolved in freshwater and seawater in equilibrium with the atmosphere. *Limnol Oceanography,* **29**:620–632.

Calvo-Díaz A, Morán X. (2006). Seasonal dynamics of picoplankton in shelf waters of the southern Bay of Biscay. *Aquatic Microbial Ecology,* **42**:159–174.

Caporaso JG, Lauber CL, Walters W a, Berg-Lyons D, Lozupone C a, Turnbaugh PJ, *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 USA,* **108**:4516–22.

Clarke K, Ainsworth M. (1993). A method of linking multivariate community structure to environmental variables. *Marine Ecological Progress Series,* **92**:205–219.

Dijkstra EW. (1959). A Note on Two Problems in Connexion with Graphs. *Numerical Mathematics,* **1**:269–271.

Duivenvoorden JF, Svenning JC, Wright SJ. (2002). Ecology. Beta diversity in tropical forests. *Science* **295**:636–7.

Edgar RC. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics,* **26**:2460–1.

Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R. (2011). UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics,* **27**:2194–200.

Engelbrektson A, Kunin V, Wrighton KC, Zvenigorodsky N, Chen F, Ochman H, *et al.* (2010). Experimental factors affecting PCR-based estimates of microbial species richness and evenness. *The ISME Journal,* **4**:642–7.

Gasol JM, Giorgio PA del. (2000). Using flow cytometry for counting natural planktonic bacteria and understanding the structure of planktonic bacterial communities. *Scientia Marina,* **64**:197–224.

Gobet A, Boetius A, Ramette A (2014) Ecological coherence of diversity patterns derived from classical fingerprinting and Next Generation Sequencing techniques. *Environmental Microbiology*, **16**:2672–2681.

Hanson CA, Fuhrman JA, Horner-Devine MC, Martiny JBH. (2012). Beyond biogeographic patterns: processes shaping the microbial landscape. *Nature Review Microbiology,* **10**:1–10.

Hong S, Bunge J, Leslin C, Jeon S, Epstein SS. (2009). Polymerase chain reaction primers miss half of rRNA microbial diversity. *The ISME Journal* **3**:1365–73.

Jones MiM, Tuomisto H, Clark DB, Olivas P. (2006). Effects of mesoscale environmental heterogeneity and dispersal limitation on floristic variation in rain forest ferns. *Journal of Ecology,* **94**:181–195.

Kwok S, Kellogg DE, McKinney N, Spasic D, Goda L, Levenson C, *et al.* (1990). Effects of primer-template mismatches on the polymerase chain reaction: human immunodeficiency virus type 1 model studies. *Nucleic Acids Research,* **18**:999–1005.

Lichstein JW. (2006). Multiple regression on distance matrices: a multivariate spatial analysis tool. *Plant Ecology,* **188**:117–131.

Logares R, Sunagawa S, Salazar G, Cornejo-Castillo FM, Ferrera I, Sarmento H, *et*

al. (2014). Metagenomic 16S rDNA Illumina tags are a powerful alternative to amplicon sequencing to explore diversity and structure of microbial communities. *Environmental Microbiology,* **16**:2659–71.

Magoč T, Salzberg SL. (2011). FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics,* **27**:2957–63.

Martiny JBH, Bohannan BJM, Brown JH, Colwell RK, Fuhrman J a, Green JL, *et al.* (2006). Microbial biogeography: putting microorganisms on the map. *Nature Reviews Microbiology,* **4**:102–12.

Martiny JBH, Eisen JA, Penn K, Allison SD, Horner-Devine MC. (2011). Drivers of bacterial beta-diversity depend on spatial scale. *Proceedings of the National Academy of Sciences U S A,* **108**:7850–4.

Nekola JC, White PS, Carolina N, Hill C. (1999). The distance decay of similarity in biogeography and ecology. *Journal of Biogeography,* **26**:867–878.

Ramette A. (2009). Quantitative community fingerprinting methods for estimating the abundance of operational taxonomic units in natural microbial communities. *Appl ied and Environmental Microbiology,* **75**:2495–505.

Sommer R, Tautz D. (1989). Minimal homology requirements for PCR primers. *Nucleic Acids Research,* **17**:6749–6749.

Vellend M (2010) Conceptual Synthesis in Community Ecology. *The Quarterly Review of Biology*, **85**:183–206.

Supplementary Figures and Tables

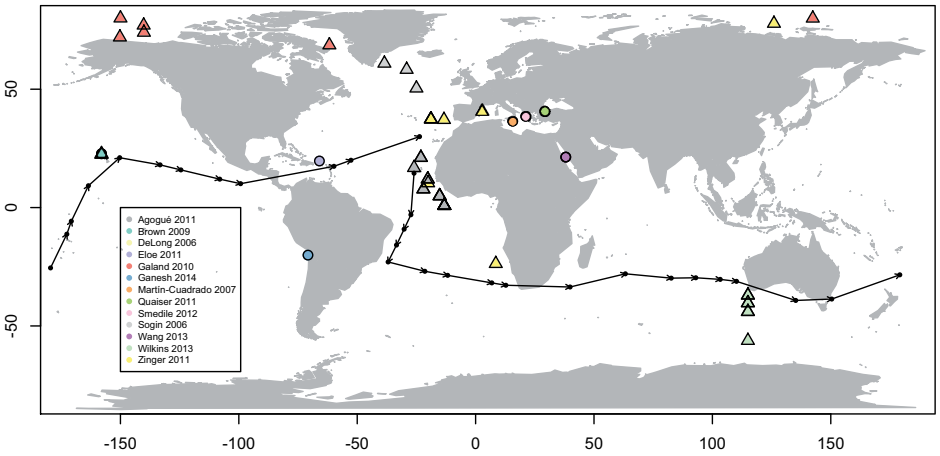


Fig. S1 World map with the position of samples described in previous studies using High-Throughput Sequencing (HTS) of microbial pelagic communities below 1,000 m depth (vents and subsurface communities excluded). Studies correspond to references Agogué *et al*., 2011; Brown *et al*., 2009; DeLong *et al*., 2006; Eloe *et al*., 2011; Galand *et al*., 2010; Ganesh *et al*., 2014; Martín-Cuadrado *et al*., 2007; Quaiser *et al*., 2011; Smedile *et al*., 2012; Sogin *et al*., 2006a; Wang *et al*., 2013; Wilkins *et al*., 2013; Zinger *et al*., 2011 in the main text. Triangles correspond to amplicon-based studies and circles to metagenomic studies.

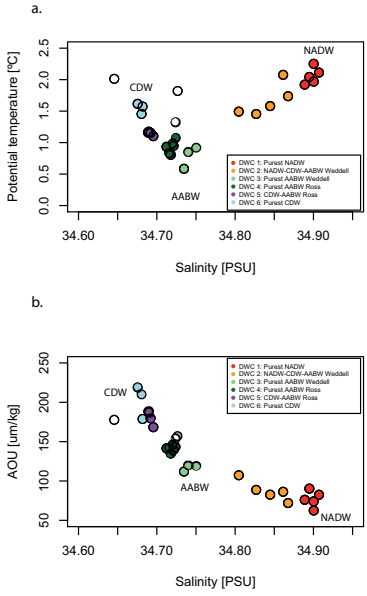


Fig. S2 T-S and AOU-S plots (A and B, respectively) of the sampling stations colored by the Water Mass-based Station Cluster (WMb-SC) they belong. Sampling stations that were difficult to classify because of a shallower sampling depth are in white.

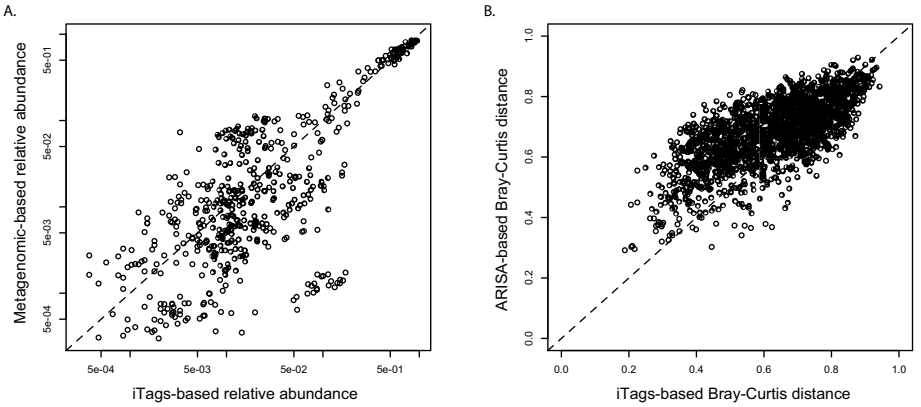


Fig. S3 Inter-technique comparison: (A) comparison of the relative abundance of dominant Phyla estimated through amplicon 16S rDNA sequencing (iTags) and metagenomic data (N= 46, Pearson r = 0.972, P-value < 0.0001), (B) comparison of Bray-Curtis distances computed with tag-based and ARISA-based OTUs for the whole dataset (N=60, Mantel r = 0.69, P = 0.001).

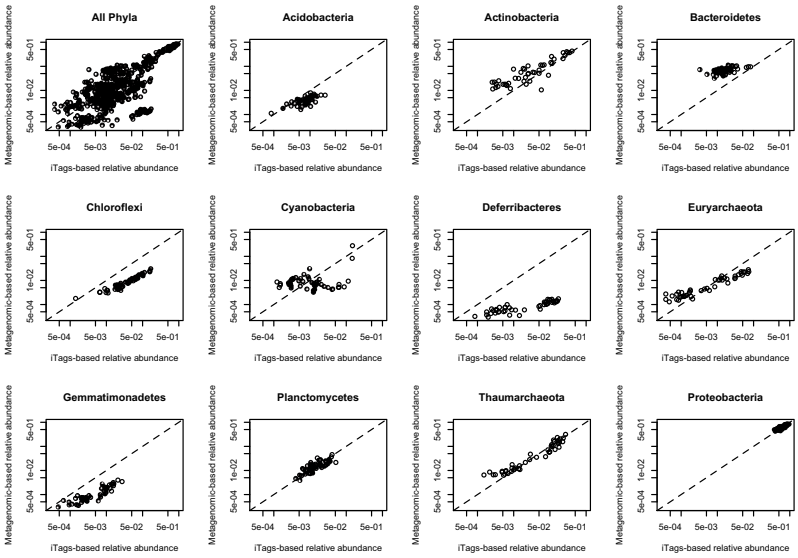


Fig. S4 Comparison of the relative abundance of each of the dominant Phyla estimated through amplicon sequencing of the 16S rDNA (iTags) and metagenomic data for 46 samples in common. Pearson correlation values: Acidobacteria, r = 0.68; Actinobacteria, r = 0.95, Bacteroidetes, r = 0.5; Chloroflexi, r = 0.97; Cuanobacteria, r = 0.62; Deferribacteres, r = 0.88; Euryarchaeota, r = 0.93; Gemmatimonadetes, r = 0.91; Planctomycetes, r = 0.73; Thaumarchaeota, r = 0.93; Proteobacteria, r = 0.88. All P-values < 0.001.

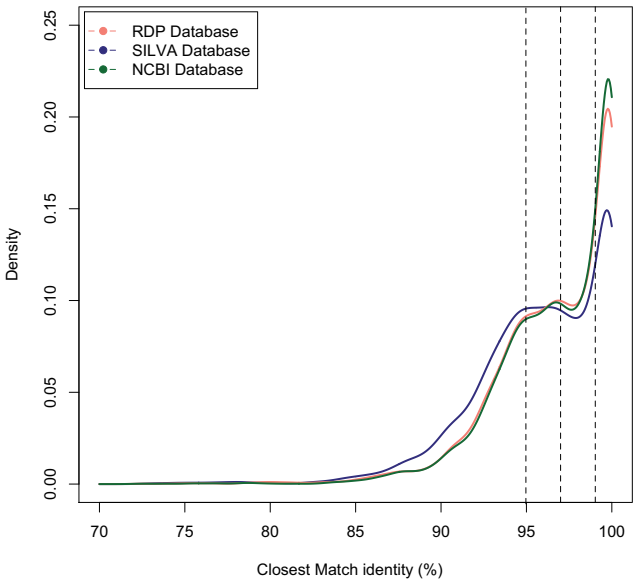


Fig. S5 Density plot of the closest match between the OTU's representative sequences and each of the three rDNA databases compared (NCBI, RDP and SILVA). The vertical dashed lines indicate 95%, 97% and 99% identity.

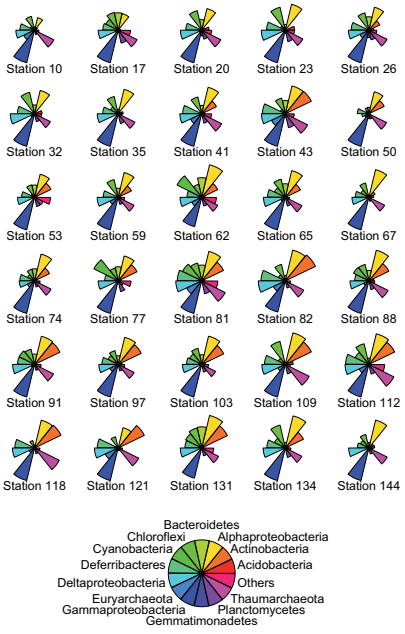


Fig. S6 Abundance (read number) along the 30 sampling stations of the dominant Phyla. Abundances were log-transformed before plotting and only the Phyla representing more than a 0.5% of the reads in the whole dataset were included. Data from the two size-fractions within a station was added after subsampling. OTUs that could not be assigned to any Phylum are included into the "Others" category.

Table S1. Composition of the six defined Water Mass-based Station Cluster (WMbSC).

	Water mass composition	**Stations**	Θ (°C)	**S**	**AOU (µmol kg^{-1})**
DWC 1	Purest NADW	10, 32, 131, 134, 144	1.92 – 2.25	34.89 – 34.91	88 – 114
DWC2	NADW–CDW–AABW Weddell	17, 20, 23, 35, 41	1.45 – 2.08	34.81 – 34.87	97 – 130
DWC3	Purest AABW Weddell	26, 43, 50	0.58 – 0.92	34.74 – 34.75	135 – 142
DWC4	Purest AABW Ross	(53), 59, 65, 67, 74, 77, 88	(1.32) 0.80 – 1.08	34.71 – 34.72	(172) 155 – 163
DWC5	CDW–AABW Ross	91, 97, 103, 109, 112	1.10 – 1.17	34.69 – 34.70	186 – 204
DWC6	Purest CDW	(62), (82), 81, 118, 121	(1.82) (2.01) 1.46 – 1.62	(34.73) (34.65) 34.68	(174) (192) 195 – 231

Water mass composition, potential temperature (Θ), salinity (S) and Apparent Oxygen Utilization (AOU) of each of the six deep-water clusters (DWC) defined. Sampling stations that were difficult to classify because of a shallower sampling depth are between parentheses (see *Methods* and Fig. S1 for details).

Table S2. Sample summary and iTags sequence processing results.

									Trimmed, cleaned and assembled reads		**Chimerical sequences excluded**	
Sample Name	**Station**	**Filter-size**	**Date (D/M/Y)**	**Depth**	**Longitude (E)**	**Latitude (N)**	**Ocean**	**Raw unassembled pairs**	**Reads**	**Proportion**	**Reads**	**Proportion**
MP0144	10	0.8	26/12/10	-4002.2	-26.000	14.520	Atlantic	44813	38520	0.860	32327	0.721
MP0145	10	0.2	26/12/10	-4002.2	-26.000	14.520	Atlantic	20917	18089	0.865	15271	0.730
MP0261	17	0.8	02/01/11	-4002.0	-27.330	-3.030	Atlantic	52982	43592	0.823	29092	0.549
MP0262	17	0.2	02/01/11	-4002.0	-27.330	-3.030	Atlantic	53105	45735	0.861	38485	0.725
MP0326	20	0.8	05/01/11	-4001.5	-30.190	-9.120	Atlantic	77580	57379	0.740	42764	0.551
MP0327	20	0.2	05/01/11	-4001.5	-30.190	-9.120	Atlantic	54372	45843	0.843	37782	0.695
MP0371	23	0.8	08/01/11	-4003.2	-33.410	-15.830	Atlantic	20573	17147	0.833	12816	0.623
MP0372	23	0.2	08/01/11	-4003.2	-33.410	-15.830	Atlantic	74895	63456	0.847	53167	0.710
MP0440	26	0.8	11/01/11	-3906.8	-36.950	-22.970	Atlantic	59389	45839	0.772	37358	0.629
MP0441	26	0.2	11/01/11	-3906.8	-36.950	-22.970	Atlantic	36007	30501	0.847	24221	0.673
MP0555	32	0.8	24/01/11	-3198.8	-21.430	-26.910	Atlantic	36044	30709	0.852	24233	0.672
MP0556	32	0.2	24/01/11	-3198.8	-21.430	-26.910	Atlantic	39296	33618	0.856	27340	0.696
MP0626	35	0.8	27/01/11	-3661.7	-11.800	-28.620	Atlantic	54590	40654	0.745	31829	0.583
MP0627	35	0.2	27/01/11	-3661.7	-11.800	-28.620	Atlantic	52334	44416	0.849	35326	0.675
MP0739	41	0.8	02/02/11	-4001.3	6.840	-31.810	Atlantic	62734	53327	0.850	42549	0.678
MP0740	41	0.2	02/02/11	-4001.3	6.840	-31.810	Atlantic	59262	50827	0.858	40266	0.679
MP0758	43	0.8	04/02/11	-3901.6	12.769	-32.813	Atlantic	50248	37859	0.753	29512	0.587
MP0759	43	0.2	04/02/11	-3901.6	12.769	-32.813	Atlantic	68283	57918	0.848	44816	0.656
MP0900	50	0.8	18/02/11	-4002.1	39.890	-33.550	Indian	47668	40587	0.851	34961	0.733
MP0901	50	0.2	18/02/11	-4002.1	39.890	-33.550	Indian	45427	38946	0.857	32846	0.723
MP0959	53	0.8	25/02/11	-3500.5	63.248	-27.978	Indian	54346	40507	0.745	30582	0.563
MP0960	53	0.2	25/02/11	-3500.5	63.248	-27.978	Indian	41892	35069	0.837	26901	0.642
MP1091	59	0.8	03/03/11	-4000.3	82.620	-29.810	Indian	45655	38706	0.848	29616	0.649
MP1092	59	0.2	03/03/11	-4000.3	82.620	-29.810	Indian	24139	20502	0.849	15958	0.661
MP1140	62	0.8	06/03/11	-2399.6	92.985	-29.653	Indian	22591	16821	0.745	10617	0.470

MP1141	62	0.2	06/03/11	-2399.6	92.985	-29.653	Indian	26434	22281	0.843	16035	0.607
MP1201	65	0.8	09/03/11	-4000.8	103.308	-30.333	Indian	64131	54243	0.846	42546	0.663
MP1202	65	0.2	09/03/11	-4000.8	103.308	-30.333	Indian	33907	28901	0.852	23204	0.684
MP1241	67	0.8	11/03/11	-4004.2	110.180	-31.160	Indian	60196	43482	0.722	33839	0.562
MP1242	67	0.2	11/03/11	-4004.2	110.180	-31.160	Indian	84095	71677	0.852	56843	0.676
MP1373	74	0.8	23/03/11	-3995.6	135.190	-39.230	Indian	56897	48416	0.851	34030	0.598
MP1374	74	0.2	23/03/11	-3995.6	135.190	-39.230	Indian	56581	48037	0.849	35975	0.636
MP1434	77	0.8	27/03/11	-4001.1	150.410	-38.640	Indian	29305	21489	0.733	15702	0.536
MP1435	77	0.2	27/03/11	-4001.1	150.410	-38.640	Indian	26640	22166	0.832	16641	0.625
MP1482	81	0.8	18/04/11	-3501.1	179.141	-28.406	Pacific	59642	49889	0.836	35620	0.597
MP1483	81	0.2	18/04/11	-3501.1	179.141	-28.406	Pacific	78761	66991	0.851	51097	0.649
MP1493	82	0.8	19/04/11	-2150.1	-179.520	-25.490	Pacific	88121	65900	0.748	48425	0.550
MP1494	82	0.2	19/04/11	-2150.1	-179.520	-25.490	Pacific	65789	55431	0.843	41438	0.630
MP1604	88	0.8	25/04/11	-4000.9	-172.640	-11.230	Pacific	21953	18392	0.838	12318	0.561
MP1605	88	0.2	25/04/11	-4000.9	-172.640	-11.230	Pacific	84369	72416	0.858	49856	0.591
MP1648	91	0.8	28/04/11	-4017.7	-170.741	-5.750	Pacific	21819	16226	0.744	12174	0.558
MP1649	91	0.2	28/04/11	-4017.7	-170.741	-5.750	Pacific	25295	21485	0.849	15261	0.603
MP1787	97	0.8	04/05/11	-3818.1	-163.530	9.220	Pacific	78218	66377	0.849	48839	0.624
MP1788	97	0.2	04/05/11	-3818.1	-163.530	9.220	Pacific	76087	65044	0.855	48663	0.640
MP1896	103	0.8	16/05/11	-4012.8	-150.319	21.064	Pacific	22448	16968	0.756	12317	0.549
MP1897	103	0.2	16/05/11	-4012.8	-150.319	21.064	Pacific	63809	54091	0.848	40447	0.634
MP2015	109	0.8	22/05/11	-4004.0	-133.260	18.040	Pacific	23790	20105	0.845	14933	0.628
MP2016	109	0.2	22/05/11	-4004.0	-133.260	18.040	Pacific	25264	21480	0.850	17439	0.690
MP2052	112	0.8	25/05/11	-4002.4	-124.474	15.909	Pacific	22943	17087	0.745	13326	0.581
MP2053	112	0.2	25/05/11	-4002.4	-124.474	15.909	Pacific	22321	18978	0.850	15683	0.703
MP2158	118	0.8	31/05/11	-3103.1	-108.060	12.000	Pacific	86189	65632	0.761	49161	0.570
MP2159	118	0.2	31/05/11	-3103.1	-108.060	12.000	Pacific	24025	20331	0.846	16399	0.683
MP2252	121	0.8	03/06/11	-3007.9	-99.246	10.093	Pacific	19111	16376	0.857	13812	0.723
MP2253	121	0.2	03/06/11	-3007.9	-99.246	10.093	Pacific	75463	64801	0.859	51225	0.679
MP2497	131	0.8	25/06/11	-4003.3	-59.830	17.430	Atlantic	31242	23290	0.745	16655	0.533
MP2498	131	0.2	25/06/11	-4003.3	-59.830	17.430	Atlantic	27416	23349	0.852	17806	0.649
MP2633	134	0.8	28/06/11	-4002.7	-52.637	19.990	Atlantic	23235	19474	0.838	15346	0.660
MP2634	134	0.2	28/06/11	-4002.7	-52.637	19.990	Atlantic	26094	22166	0.849	17955	0.688
MP2913	144	0.8	08/07/11	-4003.5	-23.690	29.970	Atlantic	56337	42092	0.747	34312	0.609
MP2914	144	0.2	08/07/11	-4003.5	-23.690	29.970	Atlantic	34797	29767	0.855	23470	0.674
							Mean	47031.1	38689.95	0.823161015	29823.78333	0.635285478
							SD	20972.83364	17302.38397	0.045388632	13257.33497	0.060039884
							Range	19111 - 88121	16226 - 72416	0.722 - 0.865	10617 - 56843	0.47 - 0.733
							TOTAL	2821866	2321397		1789427	

Sample code, location, date and depth of sampling and raw unassembled pairs obtained for each sample and reads obtained after sequence processing and chimerical sequence removal. Raw unassembled pairs include all paired reads obtained after sequencing excluding contaminant reads (42,254 reads in total) and PhiX reads (1,526,330 reads in total). See *SI* for the processing details. Proportions are computed with the raw unassembled pairs.

Table S3. Relative abundance (%) of the 30 most abundant OTUs.

	St.10	St. 17	St .20	St.23	St.26	St.32	St.35	St.41	St .43	St.50	St.53	St.59	St.62	St .65	St.67	St.74	St.77	St .81	St.82	St.88	St.91	St.97	St .103	St.109	St.112	St .118	St.121	St.131	St .134	St.144
OTU 0: Bacteria;Proteobacteria;Gammaproteobacteria;Alteromonadales;Alteromonadaceae;Alteromonas;uncultured gamma proteobacterium	19.657	19.784	16.379	16.262	11.661	5.482	5.213	8.887	8.077	19.568	24.348	13.153	3.998	7.413	11.326	5.736	9.956	3.739	2.755	10.427	7.399	5.515	1.705	1.234	2.609	5.491	25.902	3.023	12.414	18.329
OTU 8: Archaea;Thaumarchaeota;Marine Group I;uncultured marine group I crenarchaeote	8.637	7.898	7.314	5.632	3.650	3.815	6.000	7.248	4.757	5.637	1.347	5.689	3.410	5.458	6.475	5.491	1.300	8.990	5.044	5.562	6.433	4.563	3.207	10.469	11.270	11.840	6.946	3.862	4.757	6.518
OTU 1: Bacteria;Actinobacteria;Actinobacteria;Corynebacteriales;uncultureduncultured bacterium	0.000	0.000	0.000	0.000	0.014	0.085	0.235	2.449	10.196	4.978	5.515	3.179	1.041	3.400	1.380	3.381	3.033	5.590	19.403	6.268	14.270	12.193	10.328	12.499	10.502	10.398	17.745	11.058	4.012	0.170
OTU 6: Bacteria;Proteobacteria;Gammaproteobacteria;Oceanospirillales;Alcanivoraceae;Alcanivorax;uncultured Alcanivorax sp.	0.019	0.042	0.122	0.471	0.476	18.715	18.442	11.840	1.926	0.099	8.755	8.378	1.507	3.617	3.400	22.907	3.702	2.736	3.056	8.995	7.413	13.130	16.676	6.932	1.064	0.471	0.198	0.683	0.786	2.760
OTU 5: Bacteria;Proteobacteria;Deltaproteobacteria;SAR324 clade(Marine group B);uncultured organism	6.965	5.764	6.381	9.273	4.083	6.410	6.231	6.650	5.868	0.297	2.925	6.202	3.245	6.692	5.030	2.675	4.003	7.497	9.301	3.193	4.738	3.678	4.140	6.442	6.438	6.767	8.458	5.049	6.311	4.281
OTU 2: Bacteria;Proteobacteria;Gammaproteobacteria;Oceanospirillales;Halomonadaceae;Halomonas;uncultured bacterium	5.331	11.849	13.398	11.369	7.253	3.777	3.838	6.687	6.155	3.551	4.243	3.702	2.684	7.738	3.480	4.540	3.353	8.213	4.888	6.311	5.124	4.940	5.227	2.562	2.868	4.441	0.640	3.838	6.687	3.961
OTU 7: Bacteria;Proteobacteria;Alphaproteobacteria;Sphingomonadales;Erythrobacteraceae;uncultured;uncultured organism	0.042	0.645	2.896	4.300	3.852	3.297	5.331	5.044	5.251	3.217	1.851	2.378	1.097	3.881	2.501	5.077	1.253	4.394	3.852	9.207	6.235	6.377	6.273	6.306	5.651	6.998	2.199	9.923	9.475	3.438
OTU 4: Bacteria;Proteobacteria;Gammaproteobacteria;Alteromonadales;Pseudoalteromonadaceae;Pseudoalteromonas;uncultured Pseudoalteromonas sp.	5.030	2.190	2.067	0.862	4.667	11.100	9.692	2.487	4.648	31.191	12.249	0.834	0.099	2.755	1.465	0.273	0.989	0.226	0.146	0.556	0.297	0.560	1.370	0.221	1.314	0.193	0.268	0.494	1.418	17.453
OTU 3: Bacteria;Proteobacteria;Alphaproteobacteria;Rhodobacterales;Rhodobacteraceae;Sulfitobacter;Sulfitobacter sp. NAS-14.1	0.009	0.221	0.852	3.490	3.132	3.066	2.746	5.595	4.521	4.620	4.243	2.901	0.815	4.125	13.201	8.854	3.805	2.006	4.060	5.877	3.744	4.822	4.074	2.345	1.305	0.824	0.193	1.111	2.642	14.562
OTU 10: Bacteria;Proteobacteria;Gammaproteobacteria;Pseudomonadales;Moraxellaceae;Acinetobacter;uncultured prokaryote	18.758	5.534	2.821	0.292	15.824	4.615	4.050	3.871	0.796	0.188	0.141	0.042	0.000	0.019	0.005	0.000	0.019	0.033	0.014	0.005	0.028	0.019	0.429	0.033	0.009	0.009	0.000	0.005	0.019	0.028
OTU 9: Bacteria;Proteobacteria;Gammaproteobacteria;Thiotrichales;Piscirickettsiaceae;Methylophaga;Methylophaga alcalica	0.005	0.160	0.485	0.970	0.344	0.688	0.259	0.250	0.396	0.066	0.546	6.254	1.220	1.201	11.722	6.367	4.215	0.895	1.361	1.272	0.739	1.168	0.881	1.653	0.975	0.245	0.085	0.113	0.320	0.636
OTU 16: Bacteria;Proteobacteria;Alphaproteobacteria;Rhodobacterales;Rhodobacteraceae;Marinovum;Marinovum algicola	0.009	0.170	0.490	0.895	0.457	1.088	2.369	0.565	0.471	3.188	1.502	1.436	0.565	2.082	0.862	1.846	1.328	1.144	2.322	3.791	1.347	2.034	3.099	2.312	1.446	0.367	0.193	1.262	4.865	0.862
OTU 12: Bacteria;Chloroflexi;SAR202 clade;uncultured bacterium	1.441	1.549	1.300	2.774	0.857	1.606	1.229	1.088	0.885	0.221	0.551	1.658	0.268	1.724	1.121	0.645	1.045	1.224	0.951	1.069	0.970	0.862	1.060	2.030	1.983	0.890	0.678	1.281	2.308	0.137
OTU 11: Bacteria;Deferribacteres;Deferribacteres;Deferribacteres(SAR406 clade(Marine group A);uncultured bacterium	1.921	1.069	0.608	1.003	0.805	0.518	0.476	1.300	2.435	0.631	0.476	1.130	0.198	1.469	0.989	0.932	1.191	2.133	0.885	0.339	0.947	0.871	0.551	0.989	1.752	1.874	1.992	0.810	1.022	0.334
OTU 40: Bacteria;Proteobacteria;Alphaproteobacteria;Rickettsiales;S25-593;uncultured bacterium	0.980	0.711	0.923	1.535	0.608	0.852	0.838	1.587	0.909	0.198	0.542	1.761	0.706	1.135	0.655	0.782	0.890	1.286	0.702	0.636	0.900	0.593	0.867	1.620	1.210	0.688	0.626	0.542	1.229	0.711
OTU 22: Bacteria;Proteobacteria;Gammaproteobacteria;Pseudomonadales;Moraxellaceae;Psychrobacter;uncultured gamma proteobacterium	0.009	0.033	0.344	0.042	0.231	0.052	0.052	0.047	0.170	0.174	0.188	0.344	0.301	0.560	0.471	0.198	0.367	0.593	0.476	0.278	1.154	1.747	2.030	3.513	5.435	5.581	0.381	0.716	1.564	0.122
OTU 13: Bacteria;Actinobacteria;Acidimicrobiia;Acidimicrobiales;Sva0996 marine group;uncultured bacterium	0.951	0.188	0.165	0.113	1.041	0.009	0.146	1.328	4.921	0.603	0.245	0.805	0.250	1.276	0.320	0.871	4.154	1.677	0.735	0.419	0.410	0.316	0.132	0.240	0.640	1.785	1.248	0.226	0.155	0.009
OTU 18: Bacteria;Cyanobacteria;Subsection III;Microcoleus;Microcoleus chthonoplastes PCC 7420	1.055	1.978	0.141	0.612	0.047	0.254	1.319	0.278	0.551	0.589	0.278	0.414	0.320	1.121	1.653	0.391	9.598	0.377	0.250	0.104	0.311	0.104	0.160	0.325	0.438	0.424	0.085	0.513	0.706	0.377
OTU 14: Bacteria;Proteobacteria;Gammaproteobacteria;Pseudomonadales;Pseudomonadaceae;Pseudomonas;Pseudomonas sp. E505-11	0.278	0.396	2.995	1.361	1.507	0.273	0.330	0.066	0.301	1.700	0.989	1.615	0.542	1.841	2.491	0.650	0.598	0.160	0.132	0.057	0.019	0.014	0.005	0.033	0.033	2.110	2.275	0.932	0.127	0.782
OTU 17: Bacteria;Deferribacteres;Deferribacteres;Deferribacteres(SAR406 clade(Marine group A);uncultured bacterium	0.311	0.626	0.796	1.352	0.608	0.895	1.022	1.319	1.074	0.141	0.589	0.702	0.358	1.201	0.900	0.782	0.589	1.135	0.975	0.645	0.758	0.688	0.584	1.069	1.314	1.107	0.956	0.683	1.064	0.306
OTU 24: Bacteria;Cyanobacteria;SHA-109;uncultured bacterium	0.000	0.000	0.000	0.000	0.000	0.000	0.009	0.009	0.005	0.009	0.071	0.071	0.047	0.339	0.137	0.589	0.452	4.384	0.932	3.023	3.292	1.342	1.926	0.796	2.270	0.386	0.316	2.543	0.297	0.268
OTU 21: Bacteria;Proteobacteria;Gammaproteobacteria;Alteromonadales;Alteromonadaceae;Marinobacter;Marinobacter hydrocarbonoclasticus ATCC 49840	0.165	7.191	1.493	0.372	2.411	0.928	0.424	0.650	0.900	0.490	0.424	0.306	0.127	0.179	0.386	0.203	0.358	0.758	0.203	0.725	0.273	0.683	0.471	0.207	0.118	0.137	0.396	0.122	0.363	0.438
OTU 25: Archaea;Euryarchaeota;Thermoplasmata;Thermoplasmatales;Marine Group II;uncultured euryarchaeote	2.162	0.212	0.061	0.240	0.537	0.033	0.137	1.253	2.171	0.542	0.099	0.871	0.221	1.017	0.217	0.457	0.730	0.871	0.683	0.443	0.240	0.527	0.094	0.334	0.876	0.989	1.846	0.174	0.122	0.009
OTU 26: Bacteria;Actinobacteria;Actinobacteria;Corynebacteriales;uncultured;uncultured bacterium	0.024	0.000	0.000	0.000	0.000	0.009	0.019	0.038	0.207	0.047	0.085	0.052	0.028	0.099	0.028	0.188	0.113	0.254	0.829	0.711	1.413	1.229	1.003	1.078	1.356	3.287	0.429	3.490	1.790	0.240
OTU 28: Bacteria;Chloroflexi;SAR202 clade;uncultured Chloroflexi bacterium	0.527	0.655	0.716	1.224	0.405	0.885	0.664	0.556	0.391	0.075	0.358	0.834	0.372	1.045	0.598	0.325	0.471	0.570	0.523	0.353	0.471	0.358	0.560	1.262	0.961	0.320	0.311	0.518	1.144	0.429
OTU 35: Archaea;Thaumarchaeota;Marine Group I;uncultured marine group I crenarchaeote	2.491	0.400	0.363	0.297	1.573	0.028	0.080	0.862	4.022	2.458	0.132	0.226	0.009	0.655	0.203	0.283	0.358	0.170	0.014	0.603	0.245	0.155	0.057	0.094	0.141	0.367	0.494	0.198	0.085	0.033

OTU 20: Bacteria;Proteobacteria;Gammaproteobacteria;Pseudomonadales;Moraxellaceae;Psychobacter;Psychrobacter sp. ARCTIC-P41	0.066	0.033	0.212	0.151	10.474	1.281	0.640	0.706	0.655	0.146	0.052	0.075	0.009	0.038	0.014	0.000	0.005	0.028	0.000	0.014	0.085	0.094	0.014	0.038	0.462	0.655	0.028	0.287	0.047	0.014
OTU 43: Bacteria;Deferribacteres;Deferribacteres;Deferribacterales;SAR406 clade(Marine group A);uncultured SAR406 cluster bacterium HF4000_22B16	0.348	0.386	0.320	0.598	0.532	0.264	0.537	0.414	0.768	0.306	0.259	0.447	0.137	0.669	0.518	0.348	0.546	0.739	0.579	0.273	0.424	0.466	0.424	0.711	1.272	0.622	0.457	0.250	0.518	0.104
OTU 46: Bacteria;Deferribacteres;Deferribacteres;Deferribacterales;SAR406 clade(Marine group A);uncultured bacterium	0.725	0.405	0.457	0.589	0.344	0.396	0.391	0.476	0.829	0.268	0.207	0.466	0.184	0.527	0.330	0.292	0.457	0.918	0.466	0.207	0.405	0.311	0.268	0.716	0.706	0.542	0.617	0.424	0.575	0.268
OTU 44: Bacteria;Proteobacteria;Deltaproteobacteria;Desulfuromonadales;GR-WP33-58;uncultured bacterium	0.000	0.000	0.009	0.019	0.019	4.676	1.111	0.391	0.160	0.358	0.259	0.532	0.028	0.485	0.137	0.118	1.126	0.250	0.203	0.490	1.516	0.895	0.673	0.038	0.000	0.005	0.042	0.033	0.024	0.104

Relative abundance in each station (%) of the 30 most abundant OTUs in the dataset. Abundances from the free-living and particle-attached sample from each station are combined. The total number of reads for each station was normalized using a subsample of 10,617 reads/sample (thus 21,234 reads/station)

Table S4. Permutational MANOVA results.

Whole dataset				
	df	F statistic	R^2	P-value
Lifestyle (PA/FL)	1	40.893	0.31196	**<0.0001**
WMbSC	5	4.711	0.17969	**<0.0001**
Deep Ocean Basin	10	2.364	0.18031	**<0.0001**
Residuals	43	-	0.32804	-
Total	59	-	1	-
Free-living fraction				
	df	F statistic	R^2	P-value
WMbSC	5	2.1279	0.29895	**<0.0001**
Deep Ocean Basin	10	1.0951	0.30769	0.3412
Residuals	14	-	0.39336	-
Total	29	-	1	-
Particle-attached fraction				
	df	F statistic	R^2	P-value
WMbSC	5	3.2291	0.34679	**<0.001**
Deep Ocean Basin	10	1.6411	0.35250	**0.004**
Residuals	14	-	0.30071	-
Total	29	-	1	-

Permutational MANOVA testing for the effect of size fraction, Water Mass-based Station Clusters (WMbSC) and oceanic basins on prokaryote community structure. Significant P-values (<0.05) are in bold. Since the effect of the size fraction resulted to be extremely important, we present the results for the whole dataset and for each size-fraction separately.

Table S5. Permutational MANOVA results including the date of sampling.

Whole dataset				
	df	F statistic	R^2	P-value
Date of sampling	1	9.360	0.06926	**<0.0001**
Fraction (PA/FL)	1	42.160	0.31196	**<0.0001**
WMbSC	5	3.735	0.13820	**<0.0001**
Deep Ocean Basin	10	2.295	0.16981	**<0.0001**
Residuals	42	-	0.31077	-
Total	59	-	1	-
Free-living fraction				
	df	F statistic	R^2	P-value
Date of sampling		0.216	0.05937	**0.013**
WMbSC	5	2.031	0.27426	**<0.0001**
Deep Ocean Basin	10	1.168	0.31539	0.239
Residuals	14	-	0.35098	-
Total	29	-	1	-
Particle-attached fraction				
	df	F statistic	R^2	P-value
Date of sampling		9.235	0.18759	**<0.001**
WMbSC	5	2.304	0.23406	**<0.001**
Deep Ocean Basin	10	1.5474	0.31430	**0.009**
Residuals	14	-	0.26405	-
Total	29	-	1	-

Permutational MANOVA testing for the effect of the date of sampling, size fraction, Water Mass-based Station Clusters (WMbSC) and oceanic basins on prokaryote community structure. Significant P-values (<0.05) are in bold. Since the effect of the size fraction resulted to be extremely important, we present the results for the whole dataset and for each size-fraction separately.

Table S6. Environmental variables used for the BIOENV+MRM variance decomposition analysis.

Environmental variable	Units	mean	sd	Range	N	Measurement method
Depth	m	3740	489	2150 - 4018	30	CTD Data
Salinity	PSU	34.76	0.083	34.65 - 34.91	30	CTD Data
Potential Temperature	^{o}C	1.404	0.465	0.58 - 2.25	30	CTD Data
Apparent Oxygen Utilization	µM	154.25	41.08	88.22 - 230.84	30	Derived from CTD Data
Prokaryotic heterotrophic activity	pmol Leu L^{-1} h^{-1}	0.291	0.396	0.007 - 1.530	26	Leucine incorporation
Prokaryote Abundance	cells mL^{-1}	5.55E+04	6.72E+04	1.24E+04 - 31.34E+4	24	Cytometry
Percentage of HNA-content prokaryotes	-	63.25	7.38	55.15 - 90.48	24	Cytometry
Prok. Biomass Duplication Time (DT)	days	96.98	87.25	5.880 - 316.62	22	Derived from others
Prokaryote Abundance (integrated 0-200m)	cells	6.59E+05	2.40E+05	2.48E+5 - 12.56E+5	26	Derived from others

Table S7. Primer coverage for Bacteria and Archaea of the primer set used for 16S iTAGs (515F-806R) and comparison with previously used primer pairs.

				0 mismatch			1 mismatch		
Study	**Target Group**	**Primer**	**Sequence (5' – 3')**	**Archaea**	**Bacteria**	**Whole SILVA**	**Archaea**	**Bacteria**	**Whole SILVA**
ICOMM*	Archaea	1048R	CGRCRGCCATGYACCWC	82.7	9.1	10	93.9	86.7	78.5
ICOMM*	Archaea	517F	GYYTAAARNRYYYGTAGC	---	---	---	---	---	---
ICOMM*	Archaea	958F	AATTGGANTCAACGCCGG	71.2	0	2.4	88.8	0	3
ICOMM*	Archaea	958R	CGRCRGCCATGYACCWC	71.2	0	2.4	88.8	0	3
Brown 2009	Bacteria	1392Fmod	TACACACCGCCCGT	3.7	79.4	79.3	67.5	88.5	88.9
Brown 2009	Bacteria	1492R	RGMAACCTTGTACGACTT	0	0	0	0	0.2	0.2
ICOMM*	Bacteria	1064R	CGACRRCCATGCANCACCT	18	93.6	82.6	56	98.6	87.9
ICOMM*	Bacteria	341F	YCTACGGRNGGCWGCAG	0.7	95.3	83.4	70.5	98.6	90.4
ICOMM*	Bacteria	518F	CCAGCAGCYGCGGTAAN	1.1	94.7	91.3	59.5	97.9	96.5
ICOMM*	Bacteria	926R	CCGTCWATTYNTTTRANT	---	---	---	---	---	---
ICOMM*	Bacteria	967F	MNAMSCGMNRAACCTYANC	---	---	---	---	---	---
Caporaso 2011	Bacteria and Archaea	515F	GTGCCAGCMGCCGCGGTAA	58.4	94.1	92.6	96.7	97.7	97.6
Caporaso 2011	Bacteria and Archaea	806R	GGACTACHVGGGTWTCTAAT	89.7	90.9	82	97	97.1	87.7
Wilkins 2013	Bacteria and Archaea	1392R	ACGGGCGGTGTGTRC	66.8	84	84.6	71	91.3	91.4
Wilkins 2013	Bacteria and Archaea	926wF	AAACTYAAAKGAATT-GRCGG	87.7	90.8	89	94.3	97.5	97.3

Primer coverage (%) against SILVA database (release 119) was checked using Test Probe 3.0 software (http://www.arb-silva.de/?id=650) allowing 0 and 1 mismatch for every primer. Primers were obtained for the amplicon-based studies compiled for Fig. S1. Prim-

ers for which data is missing had too many ambiguities to be checked. Studies indicated with asterisks belong to the International Census of Marine Microbes project (references 2,5,6,11 and 14) and the primer sequences were obtained from http://vamps.mbl.edu/resources/primers.php. The rest of the primer sequences were obtained from the publication (Brown 2009; Caporaso 2011; Wilkins 2013).

SUPPLEMENTARY INFORMATION CHAPTER 2

Supplementary Figures and Tables

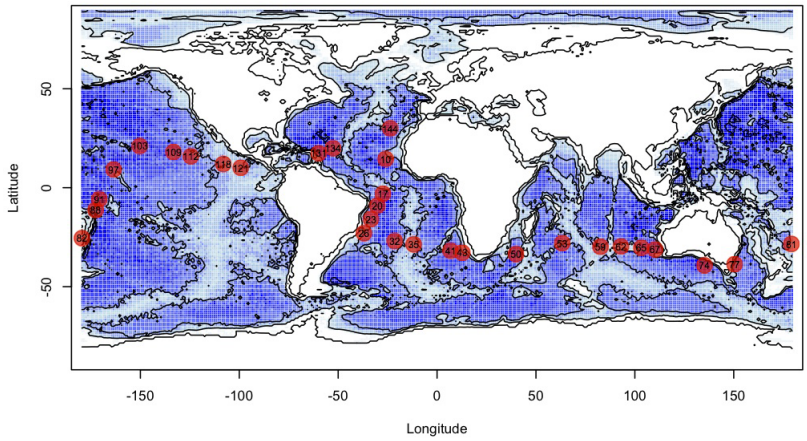


Figure S1: Location of each sampling station. Station number matches information on Table S1. Ocean's bathymetry is color-coded as a dark-to-light blue gradient color.

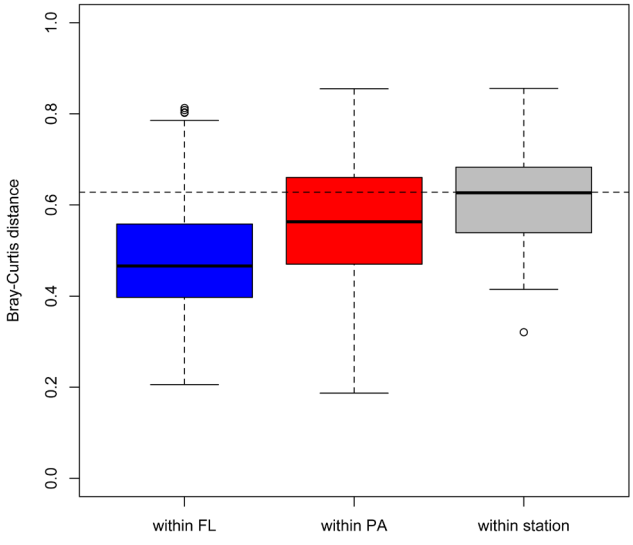


Figure S2: Boxplot of Bray-Curtis distances (BC) of all the samples within the FL fraction (blue), the PA fraction (red) and within each station, i.e. BC distances between the two size fractions for each station (gray). Differences for the mean BC distance were significant for the three pairs (see main text).

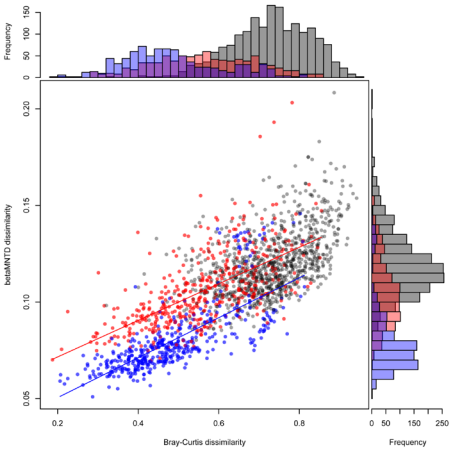


Figure S3: OTU-based betadiversity (Bray-Curtis distances) vs. phylogenetic betadiversity (betaMNTD distance). Top and right panel show the distribution of each distance (BC and betaMNTD respectively) for the pair of samples belonging to the FL fraction (blue), the PA fraction (red) and between fractions (gray). The same color code is used for the dot plot.

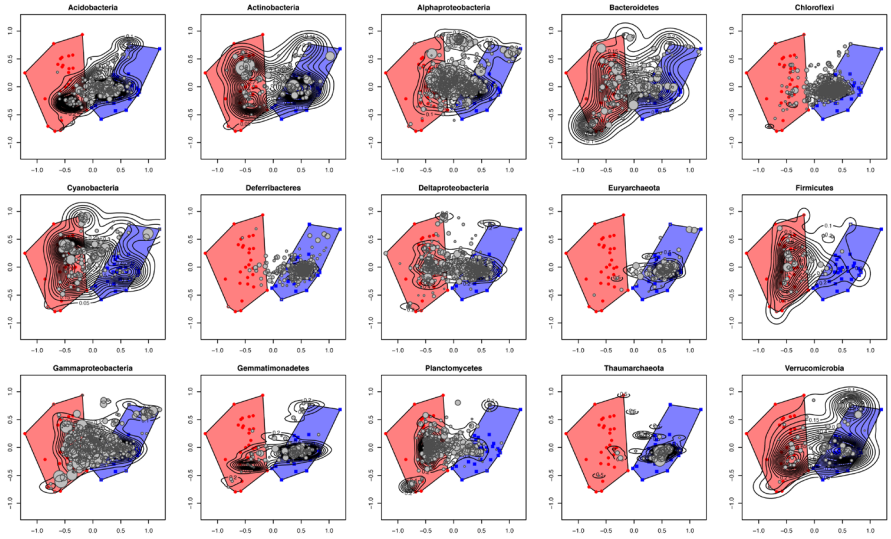


Figure S4: Non-metric Multidimensional Scaling (NMDS) showing the ordination of samples (FL: blue dots; PA: red dots) identical to Fig 1a. For each panel, weighted scores for each OTU belonging to the target Philum are depicted as points. Point diameter is proportional to the log-transformed abundance of each OTU. A contour plot for all the OTUs in each panel is added in order to better visualize the zones with higher and lower density of OTUs.

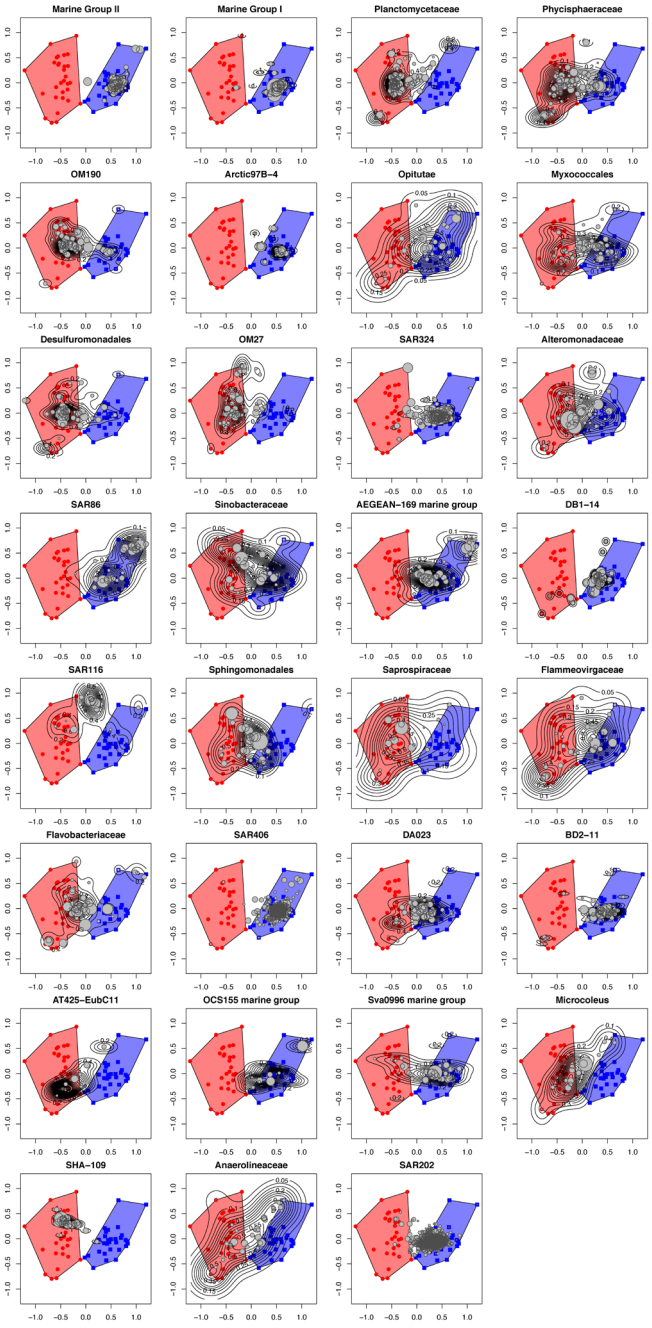
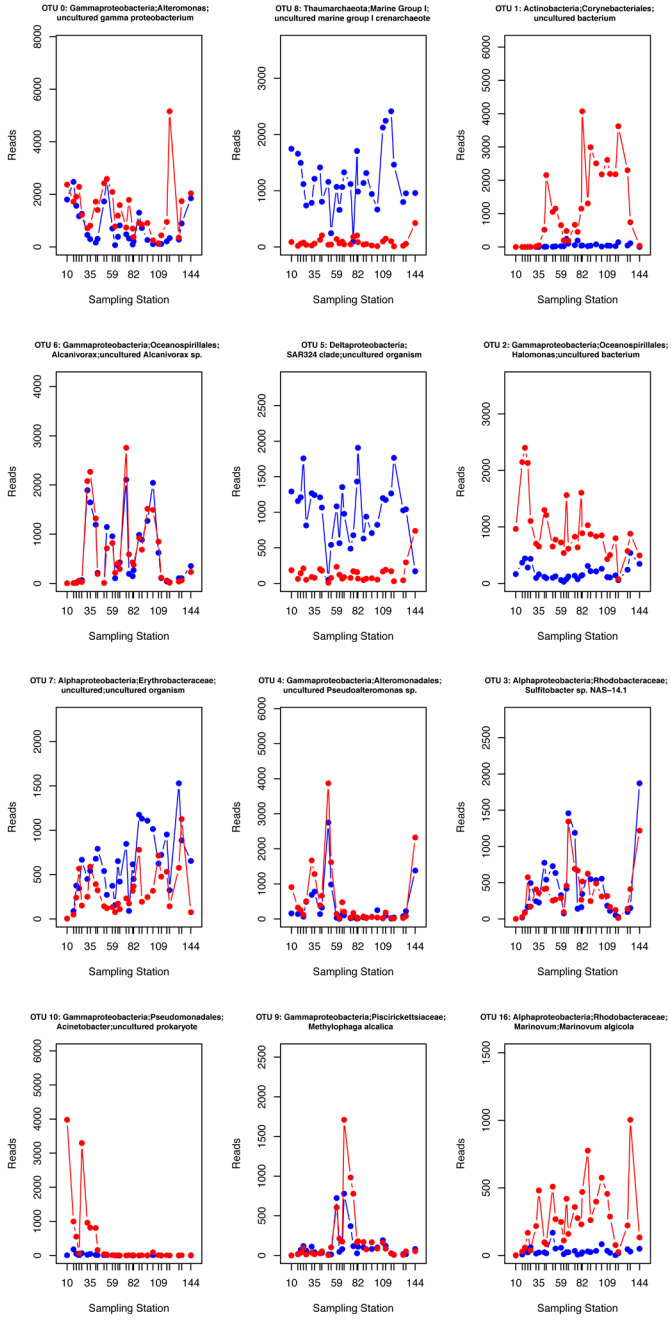
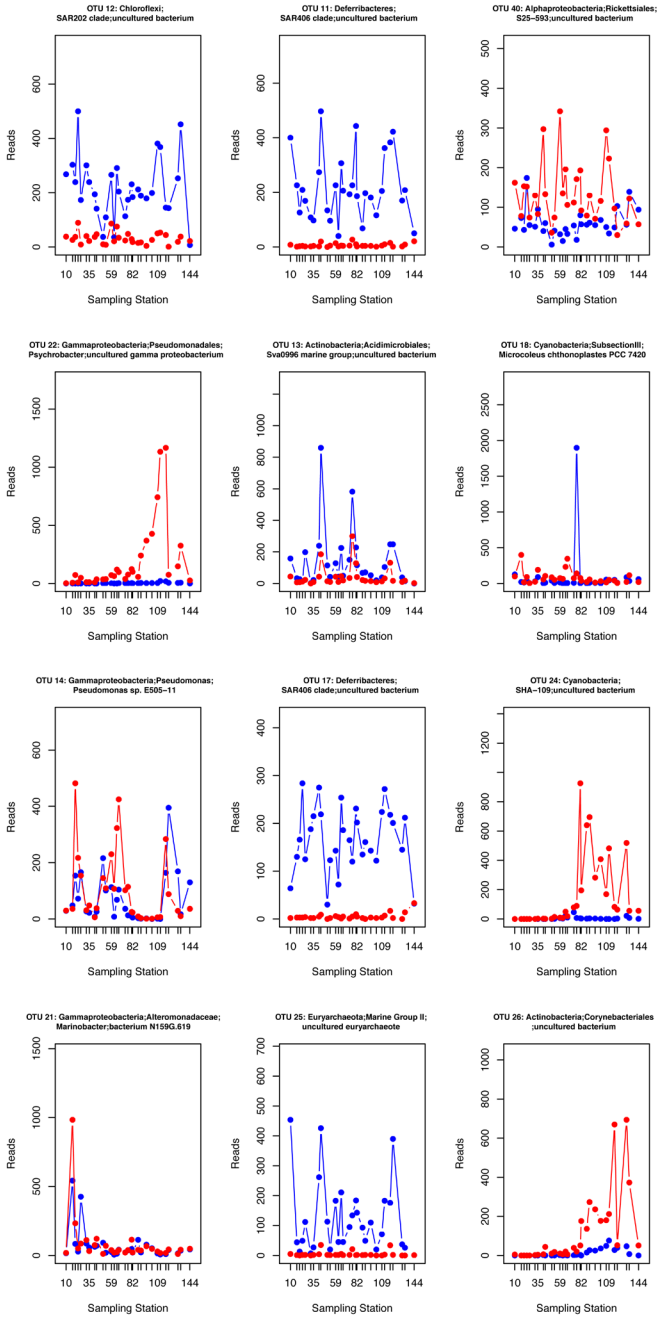


Figure S5: Non-metric Multidimensional Scaling (NMDS) showing the ordination of samples (FL: blue dots; PA: red dots) identical to Fig 1a. For each panel, weighted scores for each OTU belonging to the target lineage are depicted as points. Point diameter is proportional to the log-transformed abundance of each OTU. A contour plot for all the OTUs in each panel is added in order to better visualize the zones with higher and lower density of OTUs.


OTU 0: Gammaproteobacteria;Alteromonas; uncultured gamma proteobacterium
OTU 8: Thaumarchaeota;Marine Group I; uncultured marine group I crenarchaeote
OTU 1: Actinobacteria;Corynebacteriales; uncultured bacterium
OTU 6: Gammaproteobacteria;Oceanospirillales; Alcanivorax;uncultured Alcanivorax sp.
OTU 5: Deltaproteobacteria; SAR324 clade;uncultured organism
OTU 2: Gammaproteobacteria;Oceanospirillales; Halomonas;uncultured bacterium
OTU 7: Alphaproteobacteria;Erythrobacteraceae; uncultured;uncultured organism
OTU 4: Gammaproteobacteria;Alteromonadales; uncultured Pseudoalteromonas sp.
OTU 3: Alphaproteobacteria;Rhodobacteraceae; Sulfitobacter sp. NAS−14.1
OTU 10: Gammaproteobacteria;Pseudomonadales; Acinetobacter;uncultured prokaryote
OTU 9: Gammaproteobacteria;Piscirickettsiaceae; Methylophaga alcalica
OTU 16: Alphaproteobacteria;Rhodobacteraceae; Marinovum;Marinovum algicola
Reads
Sampling Station
10 35 59 82 109 144


OTU 12: Chloroflexi; SAR202 clade;uncultured bacterium
OTU 11: Deferribacteres; SAR406 clade;uncultured bacterium
OTU 40: Alphaproteobacteria;Rickettsiales; S25−593;uncultured bacterium
OTU 22: Gammaproteobacteria;Pseudomonadales; Psychrobacter;uncultured gamma proteobacterium
OTU 13: Actinobacteria;Acidimicrobiales; Sva0996 marine group;uncultured bacterium
OTU 18: Cyanobacteria;SubsectionIII; Microcoleus chthonoplastes PCC 7420
OTU 14: Gammaproteobacteria;Pseudomonas; Pseudomonas sp. E505−11
OTU 17: Deferribacteres; SAR406 clade;uncultured bacterium
OTU 24: Cyanobacteria; SHA−109;uncultured bacterium
OTU 21: Gammaproteobacteria;Alteromonadaceae; Marinobacter;bacterium N159G.619
OTU 25: Euryarchaeota;Marine Group II; uncultured euryarchaeote
OTU 26: Actinobacteria;Corynebacteriales ;uncultured bacterium
Reads
Sampling Station
10 35 59 82 109 144

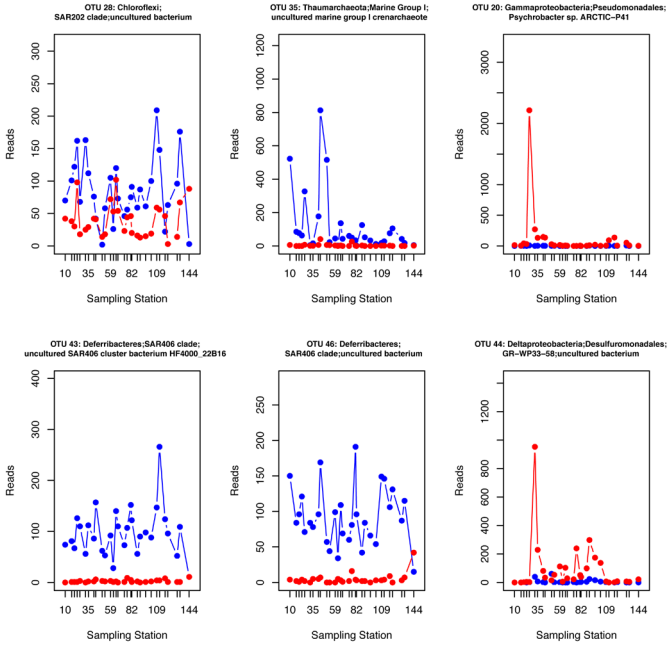


Figure S6: Distribution of abundances (reads) for each of the 30 most abundant OTUs across stations for the FL (blue) and PA (red) fractions. OTU taxonomical annotation is based on SILVA taxonomy.

Table S1: Date, depth and coordinates of sampling stations. See the location in Fig. S1.

station	**Date** (D/M/Y)	**Depth** (m)	**Longitude** (E)	**Latitude** (N)
10	26/12/10	-4002	-26.00	14.52
17	02/01/11	-4002	-27.33	-3.03
20	05/01/11	-4001	-30.19	-9.12
23	08/01/11	-4003	-33.41	-15.83
26	11/01/11	-3907	-36.95	-22.97
32	24/01/11	-3199	-21.43	-26.91
35	27/01/11	-3662	-11.80	-28.62
41	02/02/11	-4001	6.84	-31.81
43	04/02/11	-3902	12.77	-32.81
50	18/02/11	-4002	39.89	-33.55
53	25/02/11	-3500	63.25	-27.98
59	03/03/11	-4000	82.62	-29.81
62	06/03/11	-2400	92.99	-29.65
65	09/03/11	-4001	103.31	-30.33
67	11/03/11	-4004	110.18	-31.16
74	23/03/11	-3996	135.19	-39.23
77	27/03/11	-4001	150.41	-38.64
81	18/04/11	-3501	179.14	-28.41
82	19/04/11	-2150	-179.52	-25.49
88	25/04/11	-4001	-172.64	-11.23
91	28/04/11	-4018	-170.74	-5.75
97	04/05/11	-3818	-163.53	9.22
103	16/05/11	-4013	-150.32	21.06
109	22/05/11	-4004	-133.26	18.04
112	25/05/11	-4002	-124.47	15.91
118	31/05/11	-3103	-108.06	12.00
121	03/06/11	-3008	-99.25	10.09
131	25/06/11	-4003	-59.83	17.43
134	28/06/11	-4003	-52.64	19.99
144	08/07/11	-4003	-23.69	29.97

Table S2: Mean particle-association niche index (PAN index) for each Phylum and one-sample Wilcoxon signed rank test (see Material and Methods). Significant corrected P-values (<0.05) are in bold.

Phylum	Mean PAN index	sd	Statistic	**P-value**	**corrected P-value**
Acidobacteria	0.490	0.337	6431.5	0.7761	1.0000
Actinobacteria	0.531	0.399	4273.5	0.3157	1.0000
Alphaproteobacteria	0.492	0.319	28889.5	0.5449	1.0000
Bacteroidetes	0.609	0.365	16170.5	0.0000	**0.0001**
Chloroflexi	0.277	0.262	21378.0	0.0000	**0.0000**
Cyanobacteria	0.568	0.370	5211.5	0.0875	1.0000
Deferribacteres	0.123	0.210	1077.0	0.0000	**0.0000**
Deltaproteobacteria	0.524	0.402	40403.0	0.0370	0.5555
Euryarchaeota	0.141	0.207	104.0	0.0000	**0.0000**
Firmicutes	0.761	0.318	1351.5	0.0000	**0.0001**
Gammaproteobacteria	0.469	0.343	24328.5	0.0699	1.0000
Gemmatimonadetes	0.380	0.401	1705.5	0.0210	0.3143
Planctomycetes	0.757	0.311	33358.0	0.0000	**0.0000**
Thaumarchaeota	0.217	0.310	139.0	0.0000	**0.0005**
Verrucomicrobia	0.448	0.395	1661.0	0.2115	1.0000

Table S3: Mean particle-association niche index (PAN index) for each selected lineage within a Phylum and one-sample Wilcoxon signed rank test (see Material and Methods). Significant corrected P-values (<0.05) are in bold.

Lineage	Mean	sd	Statistic	P-value	corrected P-value
Marine Group II	0.044	0.083	0.0	0.0000	**0.0000**
Marine Group I	0.170	0.262	29.0	0.0001	**0.0039**
Planctomycetaceae	0.784	0.302	3218.5	0.0000	**0.0000**
Phycisphaeraceae	0.743	0.297	2144.5	0.0000	**0.0000**
OM190	0.736	0.322	1036.0	0.0001	**0.0035**
Arctic97B-4	0.111	0.160	0.0	0.0003	**0.0084**
Opitutae	0.414	0.334	140.5	0.2478	1.0000
Myxococcales	0.513	0.406	1917.5	0.5473	1.0000
Desulfuromonadales	0.827	0.272	1467.5	0.0000	**0.0000**
OM27	0.844	0.261	1090.0	0.0000	**0.0000**
SAR324	0.211	0.247	442.0	0.0000	**0.0000**
Alteromonadaceae	0.490	0.311	399.0	0.8875	1.0000
SAR86	0.146	0.158	1.0	0.0000	**0.0001**
Sinobacteraceae	0.457	0.338	163.5	0.5463	1.0000
AEGEAN-169 marine group	0.271	0.205	21.0	0.0000	**0.0011**
DB1-14	0.227	0.259	110.0	0.0001	**0.0029**
SAR116	0.665	0.316	78.0	0.1161	1.0000
Sphingomonadales	0.519	0.311	221.5	0.6815	1.0000
Saprospiraceae	0.667	0.398	216.0	0.0553	1.0000
Flammeovirgaceae	0.597	0.372	405.5	0.0646	1.0000
Flavobacteriaceae	0.617	0.296	619.0	0.0147	0.4562
SAR406	0.070	0.142	353.0	0.0000	**0.0000**
DA023	0.486	0.316	421.0	0.7071	1.0000
BD2-11	0.179	0.267	146.0	0.0000	**0.0000**
AT425-EubC11	0.815	0.261	99.0	0.0032	0.0987
OCS155 marine group	0.226	0.229	2.0	0.0106	0.3293
Sva0996 marine group	0.360	0.357	204.0	0.0256	0.7929
Microcoleus	0.627	0.267	113.0	0.0881	1.0000
SHA-109	0.854	0.166	275.0	0.0000	**0.0010**
Anaerolineaceae	0.604	0.368	577.5	0.0551	1.0000
SAR202	0.224	0.190	4480.5	0.0000	**0.0000**

Table S4: Significant indicator OTUs for the free-living (FL) size-fraction. Mean relative abundance in the FL samples (A), relative frequency of occurrences in the FL samples (B), indicator value (IndVal, the statistic), randomization-based P-value and taxonomical annotation for each OTU. Only OTUs with P-value<0.05 and those with A and B higher than 0.8 were considered as valid indicator OTUs (see Material and Methods).

OTU identity	A	B	IndVal	P-value	Taxonomy
47	0.9766	1.0000	0.9882	0.0010	Archaea;Euryarchaeota;Thermoplasmata;Thermoplasmatales;Marine Group II;uncultured marine group II euryarchaeote DeepAnt-15E7
49	0.9761	1.0000	0.9880	0.0010	Bacteria;Deferribacteres;Deferribacteres;Deferribacterales;SAR406 clade(Marine group A);uncultured bacterium
43	0.9742	1.0000	0.9870	0.0010	Bacteria;Deferribacteres;Deferribacteres;Deferribacterales;SAR406 clade(Marine group A);uncultured SAR406 cluster bacterium HF4000_22B16
17	0.9703	1.0000	0.9850	0.0010	Bacteria;Deferribacteres;Deferribacteres;Deferribacterales;SAR406 clade(Marine group A);uncultured bacterium
35	0.9675	1.0000	0.9836	0.0010	Archaea;Thaumarchaeota;Marine Group I;uncultured marine group I crenarchaeote
90	0.9657	1.0000	0.9827	0.0010	Bacteria;Deferribacteres;Deferribacteres;Deferribacterales;SAR406 clade(Marine group A);uncultured bacterium
25	0.9650	1.0000	0.9823	0.0010	Archaea;Euryarchaeota;Thermoplasmata;Thermoplasmatales;Marine Group II;uncultured euryarchaeote
158	0.9573	1.0000	0.9784	0.0010	Bacteria;Chloroflexi;SAR202 clade;uncultured deep-sea bacterium
157	0.9560	1.0000	0.9778	0.0010	Archaea;Thaumarchaeota;Marine Benthic Group A;uncultured archaeon
46	0.9524	1.0000	0.9759	0.0010	Bacteria;Deferribacteres;Deferribacteres;Deferribacterales;SAR406 clade(Marine group A);uncultured bacterium
352	0.9771	0.9667	0.9719	0.0010	Bacteria;Deferribacteres;Deferribacteres;Deferribacterales;SAR406 clade(Marine group A);uncultured bacterium
1182	0.9356	1.0000	0.9673	0.0010	Archaea;Thaumarchaeota;Marine Group I;uncultured marine archaeon
173	0.9654	0.9667	0.9660	0.0010	Archaea;Thaumarchaeota;Marine Group I;uncultured archaeon
208	0.9615	0.9667	0.9641	0.0010	Bacteria;Deferribacteres;Deferribacteres;Deferribacterales;SAR406 clade(Marine group A);uncultured SAR406 cluster bacterium
1898	0.9929	0.9333	0.9626	0.0010	Bacteria;Deferribacteres;Deferribacteres;Deferribacterales;SAR406 clade(Marine group A);uncultured bacterium
5456	0.9252	1.0000	0.9619	0.0010	Archaea;Thaumarchaeota;Marine Group I;uncultured marine group I crenarchaeote
271	0.9885	0.9333	0.9605	0.0010	Bacteria;Deferribacteres;Deferribacteres;Deferribacterales;SAR406 clade(Marine group A);uncultured SAR406 cluster bacterium
58	0.9210	1.0000	0.9597	0.0010	Bacteria;Gemmatimonadetes;Gemmatimonadetes;BD2-11 terrestrial group;uncultured bacterium
127	0.9515	0.9667	0.9591	0.0010	Bacteria;Deferribacteres;Deferribacteres;Deferribacterales;SAR406 clade(Marine group A);uncultured bacterium
73	0.9441	0.9667	0.9553	0.0010	Bacteria;Deferribacteres;Deferribacteres;Deferribacterales;SAR406 clade(Marine group A);uncultured bacterium
169	0.9757	0.9333	0.9543	0.0010	Bacteria;Verrucomicrobia;Arctic97B-4 marine group;uncultured organism
111	0.9404	0.9667	0.9534	0.0010	Archaea;Thaumarchaeota;Marine Group I;uncultured marine crenarchaeote KM3-34-D9
430	0.9723	0.9333	0.9526	0.0010	Archaea;Euryarchaeota;Halobacteria;Halobacteriales;Deep Sea Hydrothermal Vent Gp 6(DHVEG-6);uncultured euryarchaeote
148	0.9339	0.9667	0.9501	0.0010	Bacteria;Deferribacteres;Deferribacteres;Deferribacterales;SAR406 clade(Marine group A);uncultured bacterium
149	0.9649	0.9333	0.9490	0.0010	Archaea;Euryarchaeota;Thermoplasmata;Thermoplasmatales;Marine Group II;uncultured archaeon
2138	0.9647	0.9333	0.9489	0.0010	Archaea;Euryarchaeota;Thermoplasmata;Thermoplasmatales;Marine Group III;uncultured archaeon
738	0.8984	1.0000	0.9478	0.0010	Bacteria;Deferribacteres;Deferribacteres;Deferribacterales;SAR406 clade(Marine group A);uncultured bacterium
632	0.9618	0.9333	0.9475	0.0010	Bacteria;Deferribacteres;Deferribacteres;Deferribacterales;SAR406 clade(Marine group A);uncultured bacterium
4549	0.9455	0.9333	0.9394	0.0010	Archaea;Euryarchaeota;Thermoplasmata;Thermoplasmatales;Marine Group III;uncultured marine group III euryarchaeote SAT1000-53-B3

119	0.9123	0.9667	0.9391	0.0010	Bacteria;Deferribacteres;Deferribacteres;Deferribacterales;PAUC34f;uncultured bacterium
3646	0.9441	0.9333	0.9387	0.0010	Bacteria;Deferribacteres;Deferribacteres;Deferribacterales;SAR406 clade(Marine group A);uncultured bacterium
87	0.9072	0.9667	0.9364	0.0040	Bacteria;Cyanobacteria;SubsectionI;Prochlorococcus;uncultured bacterium
27	0.8758	1.0000	0.9358	0.0010	Bacteria;Proteobacteria;Gammaproteobacteria;Oceanospirillales;ZD0405;uncultured bacterium
508	0.9718	0.9000	0.9352	0.0010	Bacteria;Deferribacteres;Deferribacteres;Deferribacterales;SAR406 clade(Marine group A);uncultured marine bacterium
2729	0.9012	0.9667	0.9334	0.0010	Archaea;Thaumarchaeota;Marine Group I;uncultured marine group I crenarchaeote
4580	0.9675	0.9000	0.9331	0.0010	Bacteria;Deferribacteres;Deferribacteres;Deferribacterales;SAR406 clade(Marine group A);uncultured bacterium
2137	0.9903	0.8667	0.9264	0.0010	Archaea;Thaumarchaeota;Marine Group I;uncultured marine group I crenarchaeote
1846	0.8862	0.9667	0.9255	0.0010	Bacteria;Deferribacteres;Deferribacteres;Deferribacterales;SAR406 clade(Marine group A);uncultured bacterium
392	0.8803	0.9667	0.9225	0.0010	Bacteria;Proteobacteria;Deltaproteobacteria;SAR324 clade(Marine group B);uncultured bacterium
618	0.9815	0.8667	0.9223	0.0010	Bacteria;Proteobacteria;Deltaproteobacteria;SAR324 clade(Marine group B);uncultured bacterium
864	0.9810	0.8667	0.9220	0.0010	Archaea;Thaumarchaeota;Marine Group I;uncultured marine group I crenarchaeote
671	0.9780	0.8667	0.9207	0.0010	Bacteria;Proteobacteria;Deltaproteobacteria;SAR324 clade(Marine group B);uncultured bacterium
867	0.9769	0.8667	0.9201	0.0010	Bacteria;Deferribacteres;Deferribacteres;Deferribacterales;SAR406 clade(Marine group A);uncultured bacterium
252	0.9375	0.9000	0.9186	0.0010	Bacteria;Bacteroidetes;Cytophagia;Cytophagales;Flammeovirgaceae;Marinoscillum;uncultured bacterium
5511	0.8361	1.0000	0.9144	0.0010	Bacteria;Chloroflexi;SAR202 clade;uncultured bacterium
396	0.9259	0.9000	0.9129	0.0010	Bacteria;Planctomycetes;OM190;uncultured bacterium
1923	1.0000	0.8333	0.9129	0.0010	Bacteria;Deferribacteres;Deferribacteres;Deferribacterales;SAR406 clade(Marine group A);uncultured SAR406 cluster bacterium HF4000_22B16
2002	1.0000	0.8333	0.9129	0.0010	Bacteria;Deferribacteres;Deferribacteres;Deferribacterales;SAR406 clade(Marine group A);uncultured bacterium
425	0.8929	0.9333	0.9129	0.0010	Bacteria;Chloroflexi;SAR202 clade;uncultured deep-sea bacterium
424	0.8868	0.9333	0.9098	0.0010	Bacteria;Chloroflexi;SAR202 clade;uncultured Chloroflexi bacterium
774	0.9915	0.8333	0.9090	0.0010	Bacteria;Deferribacteres;Deferribacteres;Deferribacterales;SAR406 clade(Marine group A);uncultured bacterium
446	0.9882	0.8333	0.9075	0.0010	Bacteria;Deferribacteres;Deferribacteres;Deferribacterales;SAR406 clade(Marine group A);uncultured bacterium
238	0.9430	0.8667	0.9040	0.0010	Bacteria;Deferribacteres;Deferribacteres;Deferribacterales;SAR406 clade(Marine group A);uncultured SAR406 cluster bacterium
247	0.9245	0.8667	0.8951	0.0010	Archaea;Euryarchaeota;Halobacteria;Halobacteriales;Deep Sea Hydrothermal Vent Gp 6(DHVEG-6);uncultured crenarchaeote
814	1.0000	0.8000	0.8944	0.0010	Bacteria;Deferribacteres;Deferribacteres;Deferribacterales;SAR406 clade(Marine group A);uncultured bacterium
1427	1.0000	0.8000	0.8944	0.0010	Bacteria;Deferribacteres;Deferribacteres;Deferribacterales;SAR406 clade(Marine group A);uncultured bacterium
2267	0.8852	0.9000	0.8926	0.0010	Bacteria;Chloroflexi;SAR202 clade;uncultured bacterium
264	0.9551	0.8333	0.8921	0.0010	Bacteria;Actinobacteria;Acidimicrobiia;Acidimicrobiales;uncultured;uncultured bacterium
571	0.9524	0.8333	0.8909	0.0010	Bacteria;Proteobacteria;Gammaproteobacteria;Alteromonadales;Alteromonadaceae;Candidatus Endobugula;uncultured bacterium
261	0.8805	0.9000	0.8902	0.0010	Bacteria;Actinobacteria;Acidimicrobiia;Acidimicrobiales;OCS155 marine group;uncultured bacterium
302	0.9099	0.8667	0.8880	0.0010	Archaea;Thaumarchaeota;Marine Benthic Group A;uncultured marine crenarchaeote KM3-153-F8
949	0.9851	0.8000	0.8877	0.0010	Archaea;Thaumarchaeota;Marine Group I;Candidatus Nitrosopumilus;uncultured archaeon
2253	0.8704	0.9000	0.8851	0.0010	Bacteria;Proteobacteria;Deltaproteobacteria;SAR324 clade(Marine group B);uncultured delta proteobacterium
798	0.9714	0.8000	0.8816	0.0010	Bacteria;Chloroflexi;SAR202 clade;uncultured Chloroflexi bacterium
5170	0.9316	0.8333	0.8811	0.0010	Bacteria;Chloroflexi;SAR202 clade;uncultured bacterium

315	0.9702	0.8000	0.8810	0.0010	Bacteria;Deferribacteres;Deferribacteres;Deferribacterales;SAR406 clade(Marine group A);uncultured marine bacterium
1907	0.9697	0.8000	0.8808	0.0010	Bacteria;Deferribacteres;Deferribacteres;Deferribacterales;SAR406 clade(Marine group A);uncultured bacterium
357	0.9672	0.8000	0.8796	0.0010	Bacteria;Verrucomicrobia;Arctic97B-4 marine group;uncultured organism
726	0.9655	0.8000	0.8789	0.0010	Bacteria;Chloroflexi;SAR202 clade;uncultured bacterium
961	0.9647	0.8000	0.8785	0.0010	Bacteria;Chloroflexi;SAR202 clade;uncultured Chloroflexi bacterium
5708	0.8889	0.8667	0.8777	0.0010	Bacteria;Proteobacteria;Deltaproteobacteria;SAR324 clade(Marine group B);uncultured bacterium
203	0.8830	0.8667	0.8748	0.0010	Bacteria;Chloroflexi;SAR202 clade;uncultured Chloroflexi bacterium
3541	0.8824	0.8667	0.8745	0.0010	Bacteria;Proteobacteria;Gammaproteobacteria;Oceanospirillales;SAR86 clade;uncultured gamma proteobacterium
2404	0.9167	0.8333	0.8740	0.0010	Archaea;Thaumarchaeota;Marine Group I;uncultured marine group I crenarchaeote
211	0.9534	0.8000	0.8733	0.0010	Bacteria;Actinobacteria;Acidimicrobiia;Acidimicrobiales;OCS155 marine group;uncultured bacterium
234	0.8789	0.8667	0.8728	0.0010	Bacteria;Chloroflexi;SAR202 clade;uncultured Chloroflexi bacterium
483	0.9481	0.8000	0.8709	0.0010	Archaea;Euryarchaeota;Halobacteria;Halobacteriales;Deep Sea Hydrothermal Vent Gp 6(DHVEG-6);uncultured crenarchaeote
5714	0.9412	0.8000	0.8677	0.0010	Bacteria;Proteobacteria;Deltaproteobacteria;SAR324 clade(Marine group B);marine metagenome
142	0.9257	0.8000	0.8605	0.0010	Bacteria;Proteobacteria;AIphaproteobacteria;DB1-14;uncultured bacterium
988	0.9245	0.8000	0.8600	0.0010	Bacteria;Chloroflexi;TK10;uncultured bacterium
284	0.8176	0.9000	0.8578	0.0010	Bacteria;Chloroflexi;SAR202 clade;uncultured bacterium
477	0.9167	0.8000	0.8563	0.0010	Bacteria;Gemmatimonadetes;Gemmatimonadetes;PAUC43f marine benthic group;uncultured sponge symbiont PAUC43f
222	0.8452	0.8667	0.8559	0.0010	Bacteria;Proteobacteria;Gammaproteobacteria;KI89A clade;uncultured bacterium
3951	0.9157	0.8000	0.8559	0.0010	Archaea;Euryarchaeota;Thermoplasmata;Thermoplasmatales;Marine Group III;uncultured marine group III euryarchaeote SAT1000-53-B3
163	0.9114	0.8000	0.8539	0.0010	Bacteria;Proteobacteria;Deltaproteobacteria;Sh765B-TzT-29;uncultured bacterium AD264-E4
274	0.8727	0.8333	0.8528	0.0020	Bacteria;Acidobacteria;Holophagae;TK85;uncultured bacterium
481	0.9024	0.8000	0.8497	0.0010	Bacteria;Acidobacteria;Acidobacteria;BPC102;uncultured Acidobacterium sp.
459	0.8889	0.8000	0.8433	0.0010	Bacteria;Chloroflexi;SAR202 clade;uncultured Chloroflexus sp.
505	0.8866	0.8000	0.8422	0.0010	Bacteria;Chloroflexi;S085;uncultured Chloroflexus sp.
614	0.8444	0.8333	0.8389	0.0010	Bacteria;Acidobacteria;Acidobacteria;DA023;uncultured Acidobacterium sp.
336	0.8089	0.8667	0.8373	0.0010	Bacteria;Chloroflexi;SAR202 clade;uncultured Chloroflexi bacterium
381	0.8750	0.8000	0.8367	0.0010	Bacteria;Proteobacteria;Gammaproteobacteria;Xanthomonadales;Sinobacteraceae;JTB255 marine benthic group;uncultured bacterium
1131	0.8235	0.8333	0.8284	0.0010	Bacteria;Gemmatimonadetes;Gemmatimonadetes;BD2-11 terrestrial group;uncultured bacterium
5432	0.8548	0.8000	0.8270	0.0010	Bacteria;Proteobacteria;Deltaproteobacteria;SAR324 clade(Marine group B);uncultured organism
347	0.8413	0.8000	0.8204	0.0010	Bacteria;Chloroflexi;SAR202 clade;uncultured Chloroflexales bacterium
556	0.8279	0.8000	0.8138	0.0010	Bacteria;Chloroflexi;SAR202 clade;uncultured bacterium
3318	0.8272	0.8000	0.8135	0.0010	Bacteria;Chloroflexi;SAR202 clade;uncultured bacterium
1044	0.8224	0.8000	0.8111	0.0010	Bacteria;Chloroflexi;SAR202 clade;uncultured Chloroflexi bacterium
474	0.8141	0.8000	0.8070	0.0010	Bacteria;Proteobacteria;Deltaproteobacteria;SAR324 clade(Marine group B);uncultured bacterium
617	0.8000	0.8000	0.8000	0.0010	Bacteria;Chloroflexi;SAR202 clade;uncultured Chloroflexus sp.

Table S5: Significant indicator OTUs for the particle-attached (PA) size-fraction. Mean relative abundance in the PA samples (A), relative frequency of occurrences in the PA samples (B), indicator value (IndVal, the statistic), randomization-based P-value and taxonomical annotation for each OTU. Only OTUs with P-value<0.05 and those with A and B higher than 0.8 were considered as valid indicator OTUs (see Material and Methods).

OTU identity	A	B	IndVal	P-value	Taxonomy
22	0.9842	1.0000	0.9921	0.0010	Bacteria;Proteobacteria;Gammaproteobacteria;Pseudomonadales;Moraxellaceae;Psychrobacter;uncultured gamma proteobacterium
75	0.9699	0.9667	0.9683	0.0010	Bacteria;Proteobacteria;Gammaproteobacteria;Pseudomonadales;Moraxellaceae;Acinetobacter;uncultured bacterium
245	0.9568	0.9667	0.9617	0.0010	Bacteria;Planctomycetes;Planctomycetacia;Planctomycetales;Planctomycetaceae;Rhodopirellula;uncultured planctomycete
74	0.9228	1.0000	0.9606	0.0010	Bacteria;Planctomycetes;Pla3 lineage;uncultured bacterium
20	0.9864	0.9333	0.9595	0.0010	Bacteria;Proteobacteria;Gammaproteobacteria;Pseudomonadales;Moraxellaceae;Psychrobacter;Psychrobacter sp. ARCTIC-P41
285	0.9834	0.9333	0.9581	0.0010	Bacteria;Planctomycetes;Planctomycetacia;Planctomycetales;Planctomycetaceae;Rhodopirellula;uncultured bacterium
293	0.9796	0.9333	0.9562	0.0010	Bacteria;Planctomycetes;Phycisphaerae;Phycisphaerales;Phycisphaeraceae;CL500-3;uncultured planctomycete
177	0.9036	1.0000	0.9506	0.0010	Bacteria;Bacteroidetes;Flavobacteria;Flavobacteriales;Flavobacteriaceae;Flavobacterium;uncultured bacterium
102	0.9586	0.9333	0.9459	0.0010	Bacteria;Firmicutes;Bacilli;Bacillales;Staphylococcaceae;Staphylococcus;Staphylococcus caprae
893	0.9815	0.9000	0.9399	0.0010	Bacteria;Planctomycetes;Planctomycetacia;Planctomycetales;Planctomycetaceae;Planctomyces;uncultured bacterium
270	0.9677	0.9000	0.9333	0.0010	Bacteria;Proteobacteria;Deltaproteobacteria;Sh765B-TzT-29;uncultured delta proteobacterium
10	0.9662	0.9000	0.9325	0.0010	Bacteria;Proteobacteria;Gammaproteobacteria;Pseudomonadales;Moraxellaceae;Acinetobacter;uncultured prokaryote
115	0.9630	0.9000	0.9309	0.0010	Bacteria;Actinobacteria;Actinobacteria;Pseudonocardiales;Pseudonocardiaceae;Pseudonocardia;Pseudonocardia sp. 13630D
79	0.9268	0.9333	0.9300	0.0010	Bacteria;Proteobacteria;Alphaproteobacteria;Rhizobiales;Aurantimonadaceae;Fulvimarina;Rhizobiales bacterium 8.047
44	0.9398	0.9000	0.9197	0.0010	Bacteria;Proteobacteria;Deltaproteobacteria;Desulfuromonadales;GR-WP33-58;uncultured bacterium
1	0.9641	0.8667	0.9141	0.0010	Bacteria;Actinobacteria;Actinobacteria;Corynebacteriales;uncultured;uncultured bacterium
420	1.0000	0.8333	0.9129	0.0010	Bacteria;Planctomycetes;Planctomycetacia;Planctomycetales;Planctomycetaceae;Rhodopirellula;uncultured planctomycete
56	0.9705	0.8333	0.8993	0.0010	Bacteria;Chlamydiae;Chlamydiae;Chlamydiales;Parachlamydiaceae;Candidatus Protochlamydia;Parachlamydiaceae bacterium CRIB39
579	1.0000	0.8000	0.8944	0.0010	Bacteria;Planctomycetes;Pla3 lineage;uncultured bacterium
337	0.9565	0.8333	0.8928	0.0010	Bacteria;Proteobacteria;Gammaproteobacteria;Pseudomonadales;Moraxellaceae;Acinetobacter;uncultured bacterium
6242	0.8198	0.9667	0.8902	0.0010	Bacteria;Proteobacteria;Gammaproteobacteria;Oceanospirillales;Halomonadaceae;Halomonas;uncultured bacterium
1120	0.9508	0.8333	0.8901	0.0010	Bacteria;Lentisphaerae;Lentisphaeria;LD1-PA26;uncultured bacterium AD45-G12
432	0.9452	0.8333	0.8875	0.0010	Bacteria;Planctomycetes;Planctomycetacia;Planctomycetales;Planctomycetaceae;Planctomyces;uncultured bacterium
408	0.9817	0.8000	0.8862	0.0010	Bacteria;Planctomycetes;Phycisphaerae;Phycisphaerales;Phycisphaeraceae;I-8;uncultured bacterium
26	0.9058	0.8667	0.8860	0.0010	Bacteria;Actinobacteria;Actinobacteria;Corynebacteriales;uncultured;uncultured bacterium
470	0.9773	0.8000	0.8842	0.0010	Bacteria;Proteobacteria;Deltaproteobacteria;Sh765B-TzT-29;uncultured bacterium
24	0.9710	0.8000	0.8813	0.0010	Bacteria;Cyanobacteria;SHA-109;uncultured bacterium
1071	0.9296	0.8333	0.8801	0.0010	Bacteria;Planctomycetes;OM190;uncultured bacterium
650	0.8295	0.9333	0.8799	0.0010	Bacteria;Proteobacteria;Alphaproteobacteria;Rhodospirillales;Rhodospirillaceae;uncultured;uncultured alpha proteobacterium

296	0.8556	0.9000	0.8775	0.0010	Bacteria;Proteobacteria;Alphaproteobacteria;Rhodospirillales;Rhodospirillaceae;uncultured;uncultured bacterium
639	0.8721	0.8667	0.8694	0.0010	Bacteria;Planctomycetes;Planctomycetacia;Planctomycetales;Planctomycetaceae;uncultured;uncultured deep-sea bacterium
4567	0.8879	0.8333	0.8602	0.0010	Bacteria;Proteobacteria;Alphaproteobacteria;Rickettsiales;S25-593;uncultured Rickettsiales bacterium
29	0.9142	0.8000	0.8552	0.0010	Bacteria;Bacteroidetes;Sphingobacteria;Sphingobacteriales;Saprospiraceae;Lewinella;uncultured bacterium
561	0.8087	0.8667	0.8372	0.0010	Bacteria;Proteobacteria;Deltaproteobacteria;Bdellovibrionales;Bdellovibrionaceae;OM27 clade;uncultured bacterium
250	0.8137	0.8000	0.8068	0.0010	Bacteria;Proteobacteria;Alphaproteobacteria;Rhodospirillales;Rhodospirillaceae;uncultured;uncultured deep-sea bacterium

SUPPLEMENTARY INFORMATION CHAPTER 3

Supplementary Figures and Tables

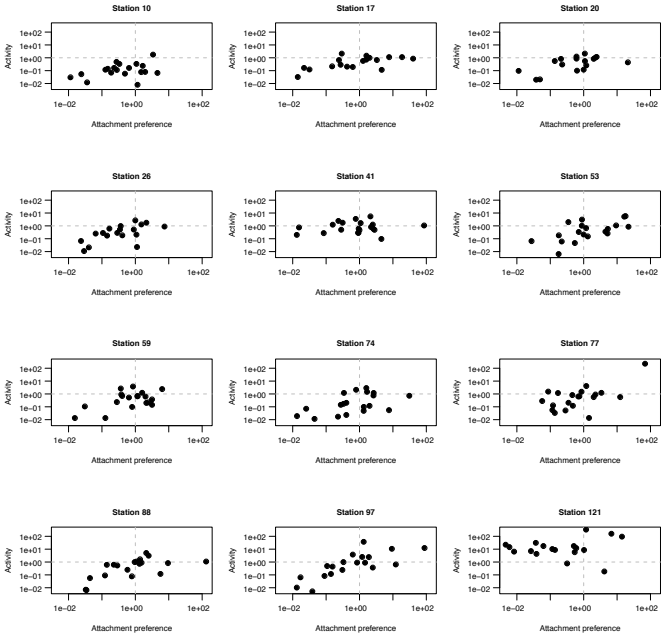


Figure S1: Attachment preference (reads in rDNA-PA / reads in rDNA-FL) and activity (reads in rRNA / reads in rDNA) of the main Phyla in the 12 balanced stations. Only the Phyla containing a total of more than 500 reads have been considered.

Table S1: Sampling and sequencing information of the rDNA and rRNA samples for the complete (i.e. without sub-sampling) dataset. The percentage contribution of inactive and active OTUs has been computed with the total reads from the rDNA dataset.

							rDNA		rRNA											
Station	Size fraction	Date (d/m/y)	Ocean	Depth (m)	Lon	Lat	Number of reads	Number of OTUs	Number of reads	Number of OTUs	Total OTUs	Shared OTUs	Inactive OTUs	Inactive OTUs (%)	Active OTUs	Active OTUs (%)	Contribution of inactive OTUs	Contribution of inactive OTUs (%)	Contribution of active OTUs	Contribution of active OTUs (%)
10	FL	26/12/10	North Atlantic	-4002	-26	14.52	15271	840	223184	1182	1698	324	356	20.97	1342	79.03	1310	8.58	13961	91.42
17	FL	2/1/11	South Atlantic	-4002	-27.33	-3.03	38485	1206	124121	1350	2015	541	500	24.81	1515	75.19	2553	6.63	35932	93.37
20	FL	5/1/11	South Atlantic	-4001	-30.19	-9.12	37782	1274	189025	1147	1803	618	481	26.68	1322	73.32	2495	6.60	35287	93.40
23	FL	8/1/11	South Atlantic	-4003	-33.41	-15.83	53167	1467	235026	1654	2379	742	609	25.60	1770	74.40	4002	7.53	49165	92.47
26	FL	11/1/11	South Atlantic	-3907	-36.95	-22.97	24221	891	165854	1302	1793	400	361	20.13	1432	79.87	1592	6.57	22629	93.43
32	FL	24/1/11	South Atlantic	-3199	-21.43	-26.91	27340	1158	147674	1511	2056	613	408	19.84	1648	80.16	1449	5.30	25891	94.70
35	FL	27/1/11	South Atlantic	-3662	-11.8	-28.62	35326	1302	260265	1910	2533	679	524	20.69	2009	79.31	2683	7.59	32643	92.41
41	FL	2/2/11	South Atlantic	-4001	6.84	-31.81	40266	1275	202110	2043	2659	659	532	20.01	2127	79.99	2915	7.24	37351	92.76
43	FL	4/2/11	South Atlantic	-3902	12.7692	-32.8128	44816	1175	214227	1389	2035	529	529	26.00	1506	74.00	3752	8.37	41064	91.63
50	FL	18/2/11	Indian	-4002	39.89	-33.55	32846	809	195949	1019	1469	359	305	20.76	1164	79.24	1391	4.23	31455	95.77
53	FL	25/2/11	Indian	-3500	63.2478	-27.9783	26901	943	190412	1301	1820	424	340	18.68	1480	81.32	1170	4.35	25731	95.65
59	FL	3/3/11	Indian	-4000	82.62	-29.81	15958	1001	228707	1246	1751	496	392	22.39	1359	77.61	1065	6.67	14893	93.33
62	FL	6/3/11	Indian	-2400	92.9852	-29.6525	16035	737	128081	802	1278	261	255	19.95	1023	80.05	827	5.16	15208	94.84
65	FL	9/3/11	Indian	-4001	103.3075	-30.3327	23204	1111	205819	1495	2145	461	507	23.64	1638	76.36	2088	9.00	21116	91.00
67	FL	11/3/11	Indian	-4004	110.18	-31.16	56843	1485	233796	1658	2495	648	646	25.89	1849	74.11	3749	6.60	53094	93.40
74	FL	23/3/11	Southern	-3996	135.19	-39.23	35975	1159	217777	1382	2080	461	480	23.08	1600	76.92	2061	5.73	33914	94.27

77	FL	27/3/11	South Pacific	-4001	150.41	-38.64	16641	667	215436	983	1444	206	376	26.04	1068	73.96	7826	47.03	8815	52.97
81	FL	18/4/11	South Pacific	-3501	179.1413	-28.4063	51097	1551	183044	1442	2347	646	683	29.10	1664	70.90	4570	8.94	46527	91.06
82	FL	19/4/11	South Pacific	-2150	-179.52	-25.49	41438	1426	214227	1587	2468	545	623	25.24	1845	74.76	3765	9.09	37673	90.91
88	FL	25/4/11	South Pacific	-4001	-172.64	-11.23	49856	1443	211357	1201	2072	572	609	29.39	1463	70.61	3853	7.73	46003	92.27
91	FL	28/4/11	South Pacific	-4018	-170.7407	-5.75	15261	928	-	-	-	-	-	-	-	-	-	-	-	-
97	FL	4/5/11	North Pacific	-3818	-163.53	9.22	48663	1427	128113	1042	1925	544	593	30.81	1332	69.19	2822	5.80	45841	94.20
103	FL	16/5/11	North Pacific	-4013	-150.3192	21.0638	40447	1288	167943	1131	1901	518	519	27.30	1382	72.70	2241	5.54	38206	94.46
109	FL	22/5/11	North Pacific	-4004	-133.26	18.04	17439	1079	-	-	-	-	-	-	-	-	-	-	-	-
112	FL	25/5/11	North Pacific	-4002	-124.4738	15.9087	15683	963	203029	1659	2155	467	404	18.75	1751	81.25	1186	7.56	14497	92.44
118	FL	31/5/11	North Pacific	-3103	-108.06	12	16399	843	-	-	-	-	-	-	-	-	-	-	-	-
121	FL	3/6/11	North Pacific	-3008	-99.2462	10.0927	51225	1285	194178	1973	2637	621	598	22.68	2039	77.32	4468	8.72	46757	91.28
131	FL	25/6/11	North Atlantic	-4003	-59.83	17.43	17806	1023	240589	1524	2046	501	399	19.50	1647	80.50	1226	6.89	16580	93.11
134	FL	28/6/11	North Atlantic	-4003	-52.6367	19.9897	17955	1123	206058	1955	2520	538	396	15.71	2124	84.29	1357	7.56	16598	92.44
144	FL	8/7/11	North Atlantic	-4003	-23.69	29.97	23470	740	204870	1298	1692	346	221	13.06	1471	86.94	1035	4.41	22435	95.59
10	PA	26/12/10	North Atlantic	-4002	-26	14.52	32327	777	169701	927	1397	307	249	17.82	1148	82.18	1639	5.07	30688	94.93
17	PA	2/1/11	South Atlantic	-4002	-27.33	-3.03	29092	1058	170025	1229	1896	391	491	25.90	1405	74.10	6697	23.02	22395	76.98
20	PA	5/1/11	South Atlantic	-4001	-30.19	-9.12	42764	953	235512	1194	1624	523	261	16.07	1363	83.93	2300	5.38	40464	94.62
23	PA	8/1/11	South Atlantic	-4003	-33.41	-15.83	12816	894	-	-	-	-	-	-	-	-	-	-	-	-
26	PA	11/1/11	South Atlantic	-3907	-36.95	-22.97	37358	682	166800	1331	1592	421	164	10.30	1428	89.70	1253	3.35	36105	96.65
32	PA	24/1/11	South Atlantic	-3199	-21.43	-26.91	24233	666	-	-	-	-	-	-	-	-	-	-	-	-
35	PA	27/1/11	South Atlantic	-3662	-11.8	-28.62	31829	950	-	-	-	-	-	-	-	-	-	-	-	-

41	PA	2/2/11	South Atlantic	-4001	6.84	-31.81	42549	1038	164035	1204	1750	492	353	20.17	1397	79.83	2317	5.45	40232	94.55
43	PA	4/2/11	South Atlantic	-3902	12.7692	-32.8128	29512	952	-	-	-	-	-	-	-	-	-	-	-	-
50	PA	18/2/11	Indian	-4002	39.89	-33.55	34961	677	-	-	-	-	-	-	-	-	-	-	-	-
53	PA	25/2/11	Indian	-3500	63.2478	-27.9783	30582	832	216394	1507	1877	462	234	12.47	1643	87.53	1330	4.35	29252	95.65
59	PA	3/3/11	Indian	-4000	82.62	-29.81	29616	1117	197520	1016	1719	414	394	22.92	1325	77.08	1779	6.01	27837	93.99
62	PA	6/3/11	Indian	-2400	92.9852	-29.6525	10617	715	-	-	-	-	-	-	-	-	-	-	-	-
65	PA	9/3/11	Indian	-4001	103.3075	-30.3327	42546	1128	-	-	-	-	-	-	-	-	-	-	-	-
67	PA	11/3/11	Indian	-4004	110.18	-31.16	33839	1221	-	-	-	-	-	-	-	-	-	-	-	-
74	PA	23/3/11	South-ern	-3996	135.19	-39.23	34030	866	185135	1156	1662	360	292	17.57	1370	82.43	2026	5.95	32004	94.05
77	PA	27/3/11	South Pacific	-4001	150.41	-38.64	15702	811	224812	1667	2018	460	249	12.34	1769	87.66	1197	7.62	14505	92.38
81	PA	18/4/11	South Pacific	-3501	179.1413	-28.4063	35620	1231	-	-	-	-	-	-	-	-	-	-	-	-
82	PA	19/4/11	South Pacific	-2150	-179.52	-25.49	48425	1102	-	-	-	-	-	-	-	-	-	-	-	-
88	PA	25/4/11	South Pacific	-4001	-172.64	-11.23	12318	535	168100	1224	1459	300	155	10.62	1304	89.38	606	4.92	11712	95.08
91	PA	28/4/11	South Pacific	-4018	-170.7407	-5.75	12174	533	-	-	-	-	-	-	-	-	-	-	-	-
97	PA	4/5/11	North Pacific	-3818	-163.53	9.22	48839	925	220595	1249	1710	464	308	18.01	1402	81.99	2097	4.29	46742	95.71
103	PA	16/5/11	North Pacific	-4013	-150.3192	21.0638	12317	526	-	-	-	-	-	-	-	-	-	-	-	-
109	PA	22/5/11	North Pacific	-4004	-133.26	18.04	14933	727	168865	1067	1416	378	207	14.62	1209	85.38	685	4.59	14248	95.41
112	PA	25/5/11	North Pacific	-4002	-124.4738	15.9087	13326	701	-	-	-	-	-	-	-	-	-	-	-	-
118	PA	31/5/11	North Pacific	-3103	-108.06	12	49161	1219	195283	1777	2450	546	479	19.55	1971	80.45	2534	5.15	46627	94.85
121	PA	3/6/11	North Pacific	-3008	-99.2462	10.0927	13812	285	222236	1755	1841	199	64	3.48	1777	96.52	291	2.11	13521	97.89
131	PA	25/6/11	North Atlantic	-4003	-59.83	17.43	16655	566	-	-	-	-	-	-	-	-	-	-	-	-

134	PA	28/6/11	North Atlantic	-4003	-52.6367	19.9897	15346	759	-	-	-	-	-	-	-	-	-	-	-
144	PA	8/7/11	North Atlantic	-4003	-23.69	29.97	34312	1022	-	-	-	-	-	-	-	-	-	-	-

Table S2: Sampling and sequencing information of the rDNA and rRNA samples for the sub-sampled dataset. The percentage contribution of inactive and active OTUs has been computed with the total reads from the rDNA dataset.

							rDNA		rRNA											
Station	Size fraction	Date (d/m/y)	Ocean	Depth (m)	Lon	Lat	Number of reads	Number of OTUs	Number of reads	Number of OTUs	Total OTUs	Shared OTUs	Inactive OTUs	**Inactive OTUs (%)**	**Active OTUs**	**Active OTUs (%)**	**Contribution of inactive OTUs**	**Contribution of inactive OTUs (%)**	**Contribution of active OTUs**	**Contribution of active OTUs (%)**
10	FL	26/12/10	North Atlantic	-4002	-26	14.52	10691	840	10510	229	969	100	356	36.74	613	63.26	951	8.90	9740	91.10
17	FL	2/1/11	South Atlantic	-4002	-27.33	-3.03	10569	767	10491	420	940	247	289	30.74	651	69.26	671	6.35	9898	93.65
20	FL	5/1/11	South Atlantic	-4001	-30.19	-9.12	10562	848	10513	318	947	219	289	30.52	658	69.48	676	6.40	9886	93.60
23	FL	8/1/11	South Atlantic	-4003	-33.41	-15.83	10523	885	10500	591	1131	345	304	26.88	827	73.12	745	7.08	9778	92.92
26	FL	11/1/11	South Atlantic	-3907	-36.95	-22.97	10498	574	10505	291	718	147	221	30.78	497	69.22	647	6.16	9851	93.84
32	FL	24/1/11	South Atlantic	-3199	-21.43	-26.91	10497	774	10510	583	1074	283	222	20.67	852	79.33	503	4.79	9994	95.21
35	FL	27/1/11	South Atlantic	-3662	-11.8	-28.62	10557	858	10479	512	1085	285	299	27.56	786	72.44	773	7.32	9784	92.68
41	FL	2/2/11	South Atlantic	-4001	6.84	-31.81	10551	798	10444	574	1079	293	322	29.84	757	70.16	747	7.08	9804	92.92
43	FL	4/2/11	South Atlantic	-3902	12.7692	-32.8128	10519	661	10505	314	788	187	278	35.28	510	64.72	836	7.95	9683	92.05
50	FL	18/2/11	Indian	-4002	39.89	-33.55	10551	477	10534	242	592	127	142	23.99	450	76.01	412	3.90	10139	96.10
53	FL	25/2/11	Indian	-3500	63.2478	-27.9783	10501	568	10511	330	758	140	176	23.22	582	76.78	410	3.90	10091	96.10
59	FL	3/3/11	Indian	-4000	82.62	-29.81	10702	1001	10518	330	1116	215	392	35.13	724	64.87	756	7.06	9946	92.94
62	FL	6/3/11	Indian	-2400	92.9852	-29.6525	10694	737	10540	209	841	105	255	30.32	586	69.68	587	5.49	10107	94.51
65	FL	9/3/11	Indian	-4001	103.3075	-30.3327	10509	807	10491	340	960	187	341	35.52	619	64.48	890	8.47	9619	91.53
67	FL	11/3/11	Indian	-4004	110.18	-31.16	10517	816	10490	368	980	204	322	32.86	658	67.14	656	6.24	9861	93.76
74	FL	23/3/11	Southern	-3996	135.19	-39.23	10574	737	10502	287	862	162	268	31.09	594	68.91	581	5.49	9993	94.51

77	FL	27/3/11	South Pacific	-4001	150.41	-38.64	10646	667	10540	292	848	111	376	44.34	472	55.66	5008	47.04	5638	52.96
81	FL	18/4/11	South Pacific	-3501	179.1413	-28.4063	10504	887	10511	458	1107	238	384	34.69	723	65.31	886	8.43	9618	91.57
82	FL	19/4/11	South Pacific	-2150	-179.52	-25.49	10528	880	10481	321	1046	155	352	33.65	694	66.35	917	8.71	9611	91.29
88	FL	25/4/11	South Pacific	-4001	-172.64	-11.23	10529	832	10535	305	953	184	336	35.26	617	64.74	781	7.42	9748	92.58
91	FL	28/4/11	South Pacific	-4018	-170.7407	-5.75	10696	928	-	-	-	-	-	-	-	-	-	-	-	-
97	FL	4/5/11	North Pacific	-3818	-163.53	9.22	10465	713	10518	340	857	196	265	30.92	592	69.08	548	5.24	9917	94.76
103	FL	16/5/11	North Pacific	-4013	-150.3192	21.0638	10562	765	10511	267	885	147	278	31.41	607	68.59	566	5.36	9996	94.64
109	FL	22/5/11	North Pacific	-4004	-133.26	18.04	10750	1071	-	-	-	-	-	-	-	-	-	-	-	-
112	FL	25/5/11	North Pacific	-4002	-124.4738	15.9087	10689	963	10483	394	1128	229	404	35.82	724	64.18	842	7.88	9847	92.12
118	FL	31/5/11	North Pacific	-3103	-108.06	12	10700	842	-	-	-	-	-	-	-	-	-	-	-	-
121	FL	3/6/11	North Pacific	-3008	-99.2462	10.0927	10530	753	10471	783	1190	346	315	26.47	875	73.53	877	8.33	9653	91.67
131	FL	25/6/11	North Atlantic	-4003	-59.83	17.43	10751	1013	10490	338	1157	194	396	34.23	761	65.77	801	7.45	9950	92.55
134	FL	28/6/11	North Atlantic	-4003	-52.6367	19.9897	10736	1100	10440	435	1317	218	385	29.23	932	70.77	854	7.95	9882	92.05
144	FL	8/7/11	North Atlantic	-4003	-23.69	29.97	10516	478	10507	235	606	107	116	19.14	490	80.86	430	4.09	10086	95.91
10	PA	26/12/10	North Atlantic	-4002	-26	14.52	10556	475	10532	134	544	65	116	21.32	428	78.68	507	4.80	10049	95.20
17	PA	2/1/11	South Atlantic	-4002	-27.33	-3.03	10528	669	10511	337	863	143	330	38.24	533	61.76	2411	22.90	8117	77.10
20	PA	5/1/11	South Atlantic	-4001	-30.19	-9.12	10541	498	10522	290	621	167	101	16.26	520	83.74	546	5.18	9995	94.82
23	PA	8/1/11	South Atlantic	-4003	-33.41	-15.83	10712	894	-	-	-	-	-	-	-	-	-	-	-	-
26	PA	11/1/11	South Atlantic	-3907	-36.95	-22.97	10556	343	10501	399	578	164	59	10.21	519	89.79	333	3.15	10223	96.85
32	PA	24/1/11	South Atlantic	-3199	-21.43	-26.91	10514	391	-	-	-	-	-	-	-	-	-	-	-	-
35	PA	27/1/11	South Atlantic	-3662	-11.8	-28.62	10551	602	-	-	-	-	-	-	-	-	-	-	-	-
41	PA	2/2/11	South Atlantic	-4001	6.84	-31.81	10524	532	10521	359	727	164	130	17.88	597	82.12	531	5.05	9993	94.95
43	PA	4/2/11	South Atlantic	-3902	12.7692	-32.8128	10528	597	-	-	-	-	-	-	-	-	-	-	-	-

50	PA	18/2/11	Indian	-4002	39.89	-33.55	10578	395	-	-	-	-	-	-	-	-	-	-	-	-	-
53	PA	25/2/11	Indian	-3500	63.2478	-27.9783	10539	496	10488	404	722	178	113	15.65	609	84.35	427	4.05	10112	95.95
59	PA	3/3/11	Indian	-4000	82.62	-29.81	10527	719	10528	161	782	98	202	25.83	580	74.17	591	5.61	9936	94.39
62	PA	6/3/11	Indian	-2400	92.9852	-29.6525	10617	715	-	-	-	-	-	-	-	-	-	-	-	-	-
65	PA	9/3/11	Indian	-4001	103.3075	-30.3327	10560	731	-	-	-	-	-	-	-	-	-	-	-	-	-
67	PA	11/3/11	Indian	-4004	110.18	-31.16	10573	884	-	-	-	-	-	-	-	-	-	-	-	-	-
74	PA	23/3/11	Southern	-3996	135.19	-39.23	10566	535	10508	200	627	108	142	22.65	485	77.35	607	5.74	9959	94.26
77	PA	27/3/11	South Pacific	-4001	150.41	-38.64	10675	811	10469	359	966	204	249	25.78	717	74.22	840	7.87	9835	92.13
81	PA	18/4/11	South Pacific	-3501	179.1413	-28.4063	10563	780	-	-	-	-	-	-	-	-	-	-	-	-	-
82	PA	19/4/11	South Pacific	-2150	-179.52	-25.49	10495	515	-	-	-	-	-	-	-	-	-	-	-	-	-
88	PA	25/4/11	South Pacific	-4001	-172.64	-11.23	10678	535	10513	348	702	181	155	22.08	547	77.92	539	5.05	10139	94.95
91	PA	28/4/11	South Pacific	-4018	-170.7407	-5.75	10685	533	-	-	-	-	-	-	-	-	-	-	-	-	-
97	PA	4/5/11	North Pacific	-3818	-163.53	9.22	10514	437	10518	313	594	156	118	19.87	476	80.13	415	3.95	10099	96.05
103	PA	16/5/11	North Pacific	-4013	-150.3192	21.0638	10677	526	-	-	-	-	-	-	-	-	-	-	-	-	-
109	PA	22/5/11	North Pacific	-4004	-133.26	18.04	10672	727	10527	250	822	155	207	25.18	615	74.82	512	4.80	10160	95.20
112	PA	25/5/11	North Pacific	-4002	-124.4738	15.9087	10704	701	-	-	-	-	-	-	-	-	-	-	-	-	-
118	PA	31/5/11	North Pacific	-3103	-108.06	12	10505	630	10450	353	789	194	150	19.01	639	80.99	473	4.50	10032	95.50
121	PA	3/6/11	North Pacific	-3008	-99.2462	10.0927	10662	285	10493	464	610	139	64	10.49	546	89.51	234	2.19	10428	97.81
131	PA	25/6/11	North Atlantic	-4003	-59.83	17.43	10703	566	-	-	-	-	-	-	-	-	-	-	-	-	-
134	PA	28/6/11	North Atlantic	-4003	-52.6367	19.9897	10675	759	-	-	-	-	-	-	-	-	-	-	-	-	-
144	PA	8/7/11	North Atlantic	-4003	-23.69	29.97	10567	624	-	-	-	-	-	-	-	-	-	-	-	-	-

Table S3: Logistic regression analysis using the rank of each OTU as a predictor variable for the probability of being dormant.

	Inactive OTUs defined by this work					Inactive OTUs defined as in Jones & Lennon, 2010.				
	Intercept	Slope	exp(Slope)	z value	P-value	Intercept	Slope	exp(Slope)	z value	P-value
FL 10	-0.9940	0.0016	1.0016	5.393	**<0.001**	3.0916	0.0024	1.0024	2.095	**0.036**
FL 17	-1.1840	0.0017	1.0017	4.964	**<0.001**	1.9133	0.0020	1.0020	2.960	**0.003**
FL 20	-1.3554	0.0016	1.0016	5.169	**<0.001**	2.2352	0.0019	1.0019	2.801	**0.005**
FL 26	-1.0692	0.0020	1.0020	3.848	**<0.001**	1.9695	0.0051	1.0051	3.384	**0.001**
FL 41	-1.2096	0.0020	1.0020	6.089	**<0.001**	1.6074	0.0012	1.0012	2.392	**0.017**
FL 53	-1.6348	0.0028	1.0028	4.811	**<0.001**	2.5178	-0.0001	0.9999	-0.057	0.955
FL 59	-1.1526	0.0014	1.0014	5.970	**<0.001**	1.4372	0.0036	1.0036	5.882	**<0.001**
FL 74	-1.4286	0.0023	1.0023	5.986	**<0.001**	1.4733	0.0040	1.0040	4.757	**<0.001**
FL 77	-0.2743	0.0016	1.0016	3.879	**<0.001**	2.2516	0.0034	1.0034	2.940	**0.003**
FL 88	-1.2141	0.0019	1.0019	6.258	**<0.001**	2.1569	0.0029	1.0029	3.510	**<0.001**
FL 97	-1.6058	0.0029	1.0029	7.126	**<0.001**	1.5938	0.0029	1.0029	3.868	**<0.001**
FL 121	-1.0412	0.0019	1.0019	5.258	**<0.001**	0.9485	0.0016	1.0016	3.554	**<0.001**
PA 10	-1.7657	0.0025	1.0025	3.170	**0.002**	2.9162	0.0054	1.0054	1.893	0.058
PA 17	0.4012	-0.0013	0.9987	-3.151	**0.002**	2.3292	0.0006	1.0006	0.751	0.453
PA 20	-1.9279	0.0021	1.0021	2.684	**0.007**	1.6997	0.0025	1.0025	2.223	**0.026**
PA 26	-1.9360	0.0020	1.0020	1.398	0.162	0.9908	0.0012	1.0012	0.912	0.362
PA 41	-1.7569	0.0022	1.0022	3.319	**0.001**	1.6559	0.0034	1.0034	3.047	**0.002**
PA 53	-2.2043	0.0037	1.0037	4.588	**<0.001**	1.7478	0.0019	1.0019	1.779	0.075
PA 59	-1.8760	0.0024	1.0024	5.765	**<0.001**	2.5812	0.0033	1.0033	2.653	**0.008**
PA 74	-1.6940	0.0024	1.0024	3.669	**<0.001**	1.7020	0.0052	1.0052	3.619	**<0.001**
PA 77	-1.9430	0.0026	1.0026	7.374	**<0.001**	2.3619	0.0018	1.0018	2.411	**0.016**
PA 88	-1.6495	0.0027	1.0027	4.168	**<0.001**	1.4642	0.0029	1.0029	2.999	**0.003**
PA 97	-1.4859	0.0022	1.0022	2.491	**0.013**	1.3097	0.0043	1.0043	3.232	**0.001**
PA 121	-1.5784	0.0023	1.0023	1.323	0.186	0.5445	0.0059	1.0059	3.159	**0.002**

Logistic regression was performed for the inactive OTUs defined in two ways: 1) as defined for this work, i.e. as OTUs detected in one or more rDNA sample but not detected in any rRNA sample, and 2) as defined before (Jones and Lennon, 2010), i.e. as the OTUs with a higher number of rRNA reads than rDNA reads. Significant P-values (<0.05) are in bold.

SUPPLEMENTARY INFORMATION CHAPTER 5

Supplementary Information

Model architecture

The motivation of the model's architecture is based on the fact that if a directional whole community export event occurs from one community (the source) to another one (the sink) with a different composition this should leave a characteristic mark on the sink community: the rare (i.e. low abundant) members of the sink community should be a proportional representation of the source community, and thus, the identity and abundances of this rare fraction of the sink community should be reproducible by taking a random sample of the source community. Therefore, the disperflux model lays on two assumptions: i) The existence of two different ecological communities (source and sink) whose species' abundances are known (corresponding in this case to two different samples for which we have an estimate of their OTU's abundances) and ii) that the dominant species in the source community (i.e. all the individuals belonging to OTUs with higher abundances in the source community compared to the sink community) are also present in the sink community only because an event of whole community export has occurred. Given these assumptions the disperflux model is built as a sequential simulation with the following steps:

The dominant species in the source community are removed from the sink community (i.e. the read count of the OTUs that are more abundant in the source community than in the sink community are set to 0 in the latter) creating thus a "hypothetical original sink community".

A random sample of individuals from the source community is taken: a random sample of the OTUs of the source community is taken using the read count as probability weights. The sample size, i.e. the proportion of the source community exported to the sink community, is the only parameter of the model and is referred hereafter as the "*flux*".

The sampled individuals are added to the sink community, creating thus a new "simulated sink community".

The three former steps constitute the core of the disperflux model: the simulation of a whole community export event from a source to a sink community. Two additional steps were developed in order to estimate the optimal *flux* between two communities:

One dimension optimization procedure (the *optimize* function in the *stats* package within R software) is used to find the optimal *flux* value. The Pearson correlation between the real sink (i.e. composition of the sink community) and the simulated sink community (from the process in 1) is used as the optimization function. That is, values of the flux

parameter are tested in the interval from 0 to 1 and the disperflux model is applied in order to find the value that yields the maximum Pearson correlation. Thus the estimated optimal flux corresponds to the fraction of the sink community that needs to be transferred from the source community in order to better reproduce its composition.

As the disperflux model involves a stochastic step (choosing a random sample of individuals from the source community) the optimization procedure (step 4) is repeated a high number of times and a mean *flux* value and their confidence interval (CI) are computed.

Interpretation of model results

Given two communities (community A and B), the disperflux model can be applied in two directions: by transferring a proportion of the individuals (i.e. read counts) of the community A to the community B and vice versa. That is, by using each of the two communities as a source and a sink community. Similarly, when the model is used for the estimation (not only the simulation) of the optimal flux between two communities, two fluxes can be estimated for a pair of samples. Thus, three scenarios may result: a) A high flux, i.e. close to 1, in both directions, which results in correlated abundances between communities (named as *Mixing scenario*, Fig. 3A), b) a low flux (close to 0) in both directions, which results in the virtual absence of common diversity (named as *No flux scenario*, Fig 3B) and c) the most interesting scenario (named as *Directional flux scenario*, Fig. 3C), which arises when an asymmetry in the fluxes estimates exists, i.e. a moderate flux in one direction and a nearly 0 flux in the opposite direction. Thus, the disperflux model allows differentiating the three patterns observed when comparing OTU's abundances between pairs of samples (Fig. 2). In fact, the fluxes estimates for all the pairwise comparisons were related to OTU-based dissimilarity measures that are commonly used, such as Bray-Curtis and others (Fig. S6): the Mixing scenario and the No flux scenario corresponded to low and high dissimilarity values, respectively. More interestingly, a high variability of flux estimates existed for intermediate dissimilarity values. This indicates that the disperflux model is able to discriminate better situations where traditional dissimilarity metrics fail. This is due to the fact that the disperflux approach introduces directionality in the measure of community similarity. That is, two communities with a Bray-Curtis value of 0.4 will correspond to two flux estimates (one in each direction), that may be in the range from 0 to 0.8 (Fig. S6). If this intermediate Bray-Curtis value is due to a whole-community export event between two originally distinct communities, the pattern observed in Fig. 2C will be found and the Bray-Curtis value will correspond to a pair of assymmetrical flux values, one close to 0 and a second higher one. If not such a pattern is present, the

Bray-Curtis value will correspond to a pair of nearly 0 flux estimates. Thus, the disperflux model, as a tool, produces a similarity measure between communities with a directionality component that allows discriminating the effect of massive dispersion from other mechanisms maintaining intermediate levels of community similarity.

The foundational assumption of the disperflux model is that, given a source and a sink community, the OTUs dominating in the source community, if present also in the sink community, are there only because a whole community export event occurred and transferred them. Thus, the three scenarios that the model is able to differentiate should, a priori, be interpreted in terms of dispersion or immigration between communities. Thus, the disperflux model is inserted in the long developed metacommunity framework (Leibold *et al.* 2004). It consists on an attempt to model the effect of dispersal rate in community composition. Dispersal rate is an essential ingredient of the metacommunity framework and other similar theoretical approaches to biogeography (Martiny *et al.* 2006) as it may decouple beta-diversity from habitat environmental variability either by means of high (i.e. mass effects) or low dispersal rates (i.e. dispersal limitation) (Lindström & Langenheder 2012). If dispersal rate is low enough between two communities, differences in the environment will result in the selection of different assemblages through species sorting (or environmental filtering). Even when assuming neutral dynamics (that is, all species are similar in their competitive ability and in dispersal) dispersal limitation may rule community differentiation (Condit *et al.* 2002; Hanson *et al.* 2012). Thus, for two communities to diverge in their composition, the dispersal rate needs to be low enough to not counteract the effect of either environmental filtering. Thus, the detection through our model of low fluxes in both directions between two communities (i.e. the No flux scenario) has to be interpreted as an indication of low dispersal rates between communities, irrespectively of the existence of an additional environmental filtering effect. On the contrary, two communities exhibiting correlated OTU abundances (i.e. the Mixing scenario) may be achieved by two different processes: either because a) relevant environmental drivers are similar in both habitats and thus similar assemblages are selected through environmental filtering or because b) dispersal rates between both communities are high enough to erode the environmental filtering exerted by differences that may exist between the two habitats. Finally, when asymmetrical fluxes are estimated for a pair of communities (i.e. the Directional flux scenario) the only possible interpretation is the existence of an event that disperses a fraction of one community into a different one, with a sufficient magnitude to be detected in their low-abundant members (that is, presenting evidence of mass effects). In fact, the disperflux model itself is the proof that such an event leaves a detectable pattern in OTU abundances, such as the one described before (Fig. 2C).

References

Condit R, Pitman N, Leigh EG *et al.* (2002) Beta-diversity in tropical forest trees. *Science*, **295**:666–9.

Hanson CA, Fuhrman JA, Horner-Devine MC, Martiny JBH (2012) Beyond biogeographic patterns: processes shaping the microbial landscape. *Nature reviews. Microbiology*, **10**:1–10.

Leibold M a., Holyoak M, Mouquet N *et al.* (2004) The metacommunity concept: a framework for multi-scale community ecology. *Ecology Letters*, **7**:601–613.

Lindström ES, Langenheder S (2012) Local and regional factors influencing bacterial community assembly. *Environmental Microbiology Reports*, **4**:1–9.

Martiny JBH, Bohannan BJM, Brown JH *et al.* (2006) Microbial biogeography: putting microorganisms on the map. *Nature reviews. Microbiology*, **4**:102–12.

Supplementary Figures and Tables

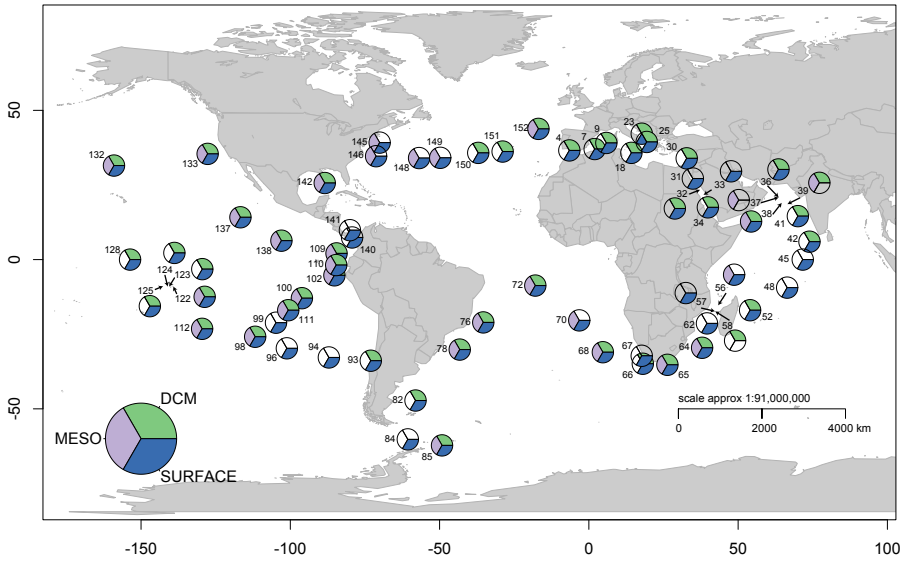


Figure S1: Location of the Tara Ocean's stations used in this study and the availability in each station of samples from the three layers (surface, DCM and mesopelagic).

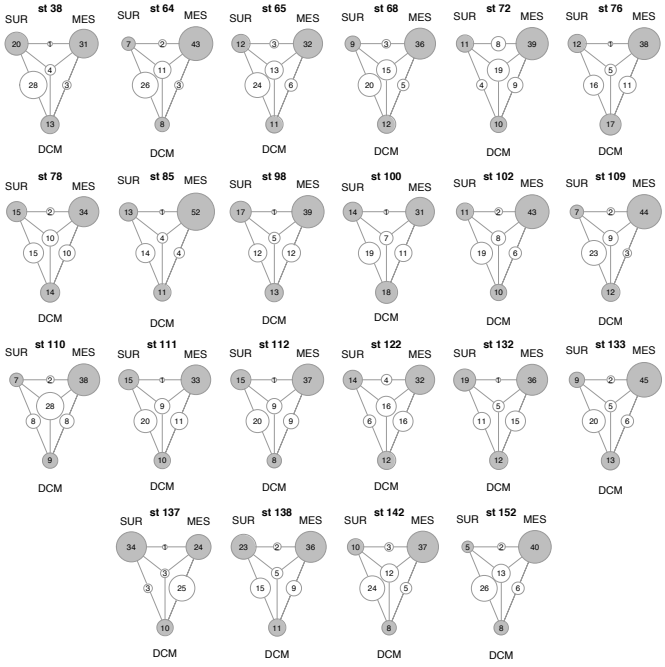


Figure S2: Percentage of OTUs unique (in grey) and shared (in white) between the three layers (surface, DCM and mesopelagic) in each station. Only the stations for which the three samples were available have been used.

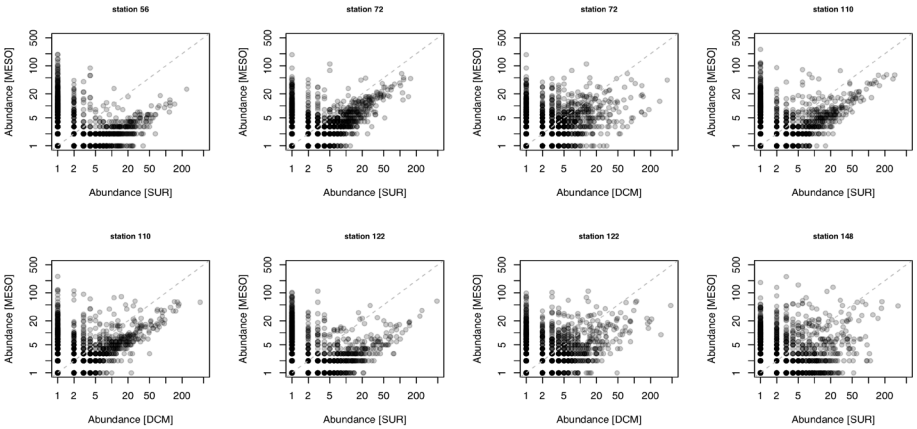


Figure S3: Scatterplots of the abundances (reads+1) of all the OTUs for photic (i.e. surface or DCM) vs. mesopelagic samples in stations 56, 72, 110, 122 and 148. The layers compared in each panel are denoted in the axes. Both axis are in log scale.

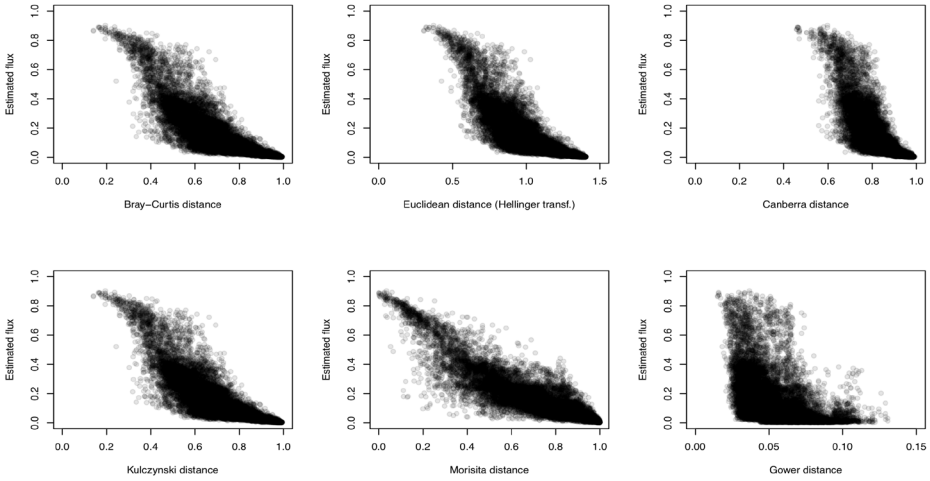


Figure S4: Comparison of all the pairwise dissimilarity values to the estimated fluxes. Bray-Curtis, Euclidean distance (after Hellinger transformation of the OTU table), the Canberra, Kulczynski, Morista and Gower dissimilarities are presented.

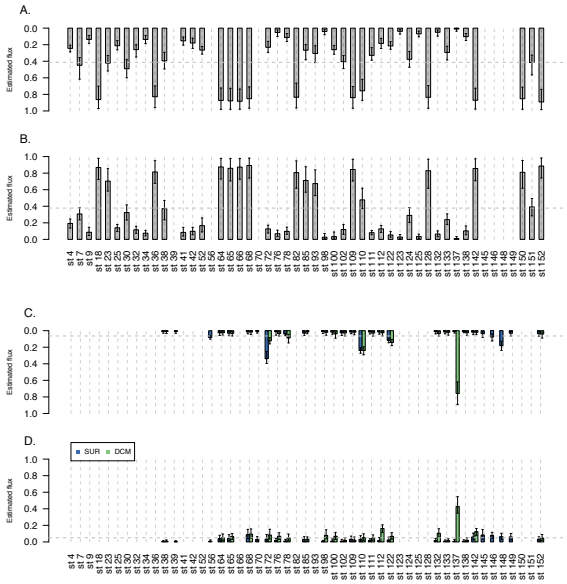


Figure S5: Mean and 95% CI of the optimal fluxes estimated within all the stations. Estimated fluxes from surface to DCM (A), from DCM to surface (B), from surface/DCM to mesopelagic (C) and from mesopelagic to surface/DCM (D). Estimations are based on 1,000 iterations of the disperflux model.

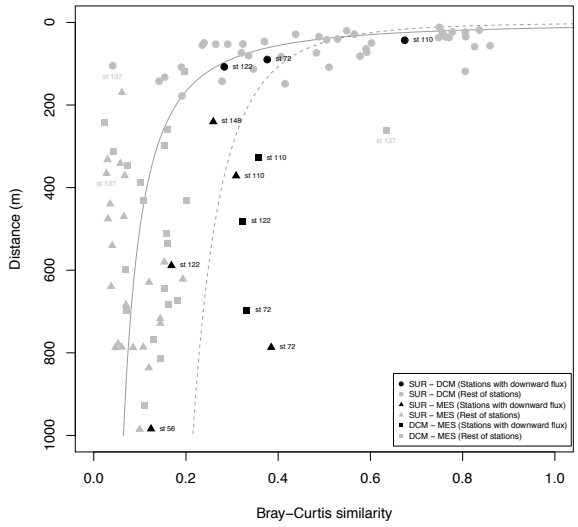


Figure S6: Bray-Curtis similarity in relation to the vertical distance separating each pair of samples. Stations with increased downward flux (stations 56, 72, 110, 122 and 148) are in black and the rest are in grey. Lines correspond to the best fit of a potential equation for each of the two groups of points (grey points: $y = 5.11 * x^{-0.633}$, Adjusted $R^2 = 0.648$; black points: $y = 1.46 * x^{-0.277}$, Adjusted $R^2 = 0.332$). Station 137 was excluded from the analysis (see Discussion). Both equations were significantly different (P-value < 0.05; tested through the interaction's significance of an ANCOVA model with log[Bray-Curtis] as independent variable, log[distance] as dependent variable and a factor differentiating stations with increased downward flux from the rest of the stations)

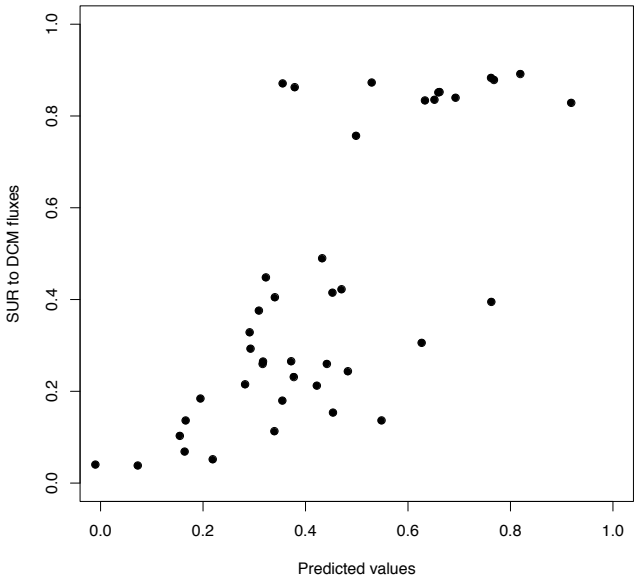


Figure S7: Comparison between the observed linear model for surface to DCM fluxes vs. those predicted flux based on a linear model containing the log of the difference in depth (x_1) and the absolute difference in temperature (x_2) between the surface and DCM within each station. The model corresponds to: Flux = 1.549 + -0.574*x_1 - 0.053*x_2; Adjusted R^2 = 0.529; P-values of slopes < 0.001 in both cases).

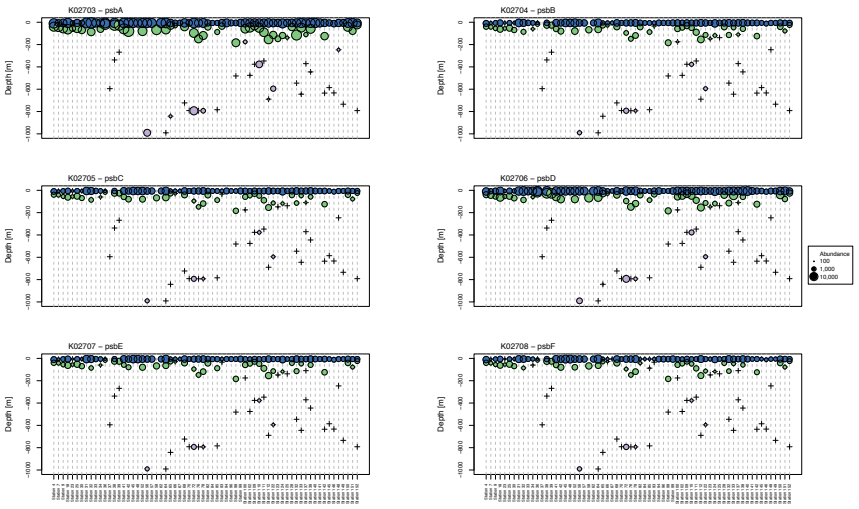


Figure S8: Photosynthetic genes. Abundance of the genes involved in the photosystem II protein complex. The diameter of the bubbles is proportional to the log of the abundance. Samples are denoted with crosses. Surface, DCM and mesopelagic samples are color-coded (green, blue and purple, respectively).

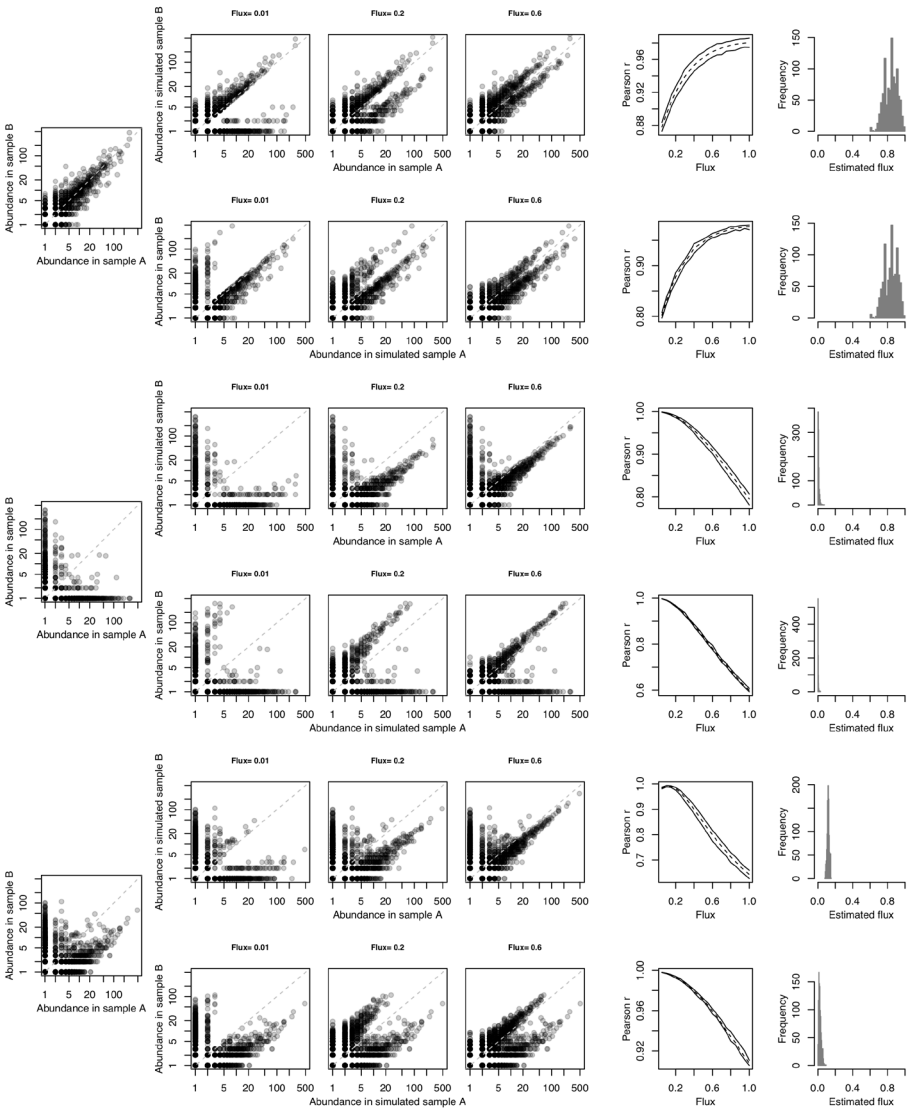


Figure S9: Disperflux model architecture. Three examples of the disperflux model applied to the same comparisons of samples than Fig. 2 (A, B and C). The left panel corresponds to the pairwise comparison of real data (identical to Fig. S3D, E and F). The three central panels correspond to the simulation of increasing flux values (0.01, 0.2 and 0.6) from sample A to sample B (upper panels) and from sample B to sample A (lower panels). The fourth column of panels corresponds to the Pearson's r-value between the sink and the simulated sink community (see Material and Methods) for increasing flux values. The mean (dashed line) and 95% CI (solid lines) is computed from 1,000 iterations. The panels on the right correspond to the distribution of the flux values estimated from 1,000 iterations.

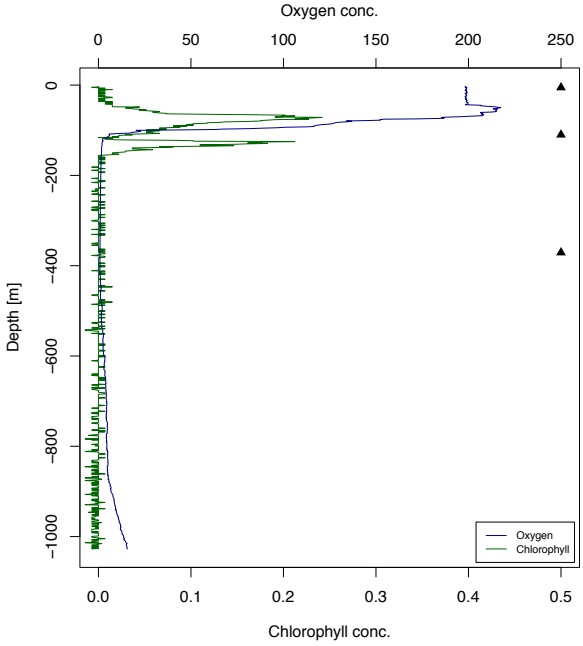


Figure S10: Depth profile of chlorophyll concentration (mg m^{-3}) estimated from fluorescence and oxygen concentration (µmol kg^{-1}) of station 137. The sample depths are indicated with a triangle on the right side of the plot.

Table S1. Sample information

Sample label	INSDC accession number(s)	PANGAEA sample identifier	Latitude	Longitude	Depth [m]	Extracted miTags	Mapped miTags
TARA_004_DCM_0.22-1.6	ERS487936	TARA_X000000368	36.5533	-6.5669	40	143169	36838
TARA_004_SRF_0.22-1.6	ERS487899	TARA_Y200000002	36.5533	-6.5669	5	91124	21154
TARA_007_DCM_0.22-1.6	ERS477953	TARA_A200000159	37.0541	1.9478	42	77642	13822
TARA_007_SRF_0.22-1.6	ERS477931	TARA_A200000113	37.051	1.9378	5	75826	14517
TARA_009_DCM_0.22-1.6	ERS488147	TARA_X000001036	39.0609	5.9422	55	96015	25086
TARA_009_SRF_0.22-1.6	ERS488119	TARA_X000000950	39.1633	5.916	5	147018	33328
TARA_018_DCM_0.22-1.6	ERS488346	TARA_S200000501	35.7528	14.2765	60	115928	32530
TARA_018_SRF_0.22-1.6	ERS488330	TARA_A100000164	35.759	14.2574	5	133643	34889
TARA_023_DCM_0.22-1.6	ERS477998	TARA_E500000081	42.1735	17.7252	55	53084	13089
TARA_023_SRF_0.22-1.6	ERS477979	TARA_E500000075	42.2038	17.715	5	81189	19275
TARA_025_DCM_0.22-1.6	ERS488509	TARA_E500000331	39.3991	19.3997	50	96926	27251
TARA_025_SRF_0.22-1.6	ERS488486	TARA_E500000178	39.3888	19.3905	5	133238	26364
TARA_030_DCM_0.22-1.6	ERS478040	TARA_A100001011	33.9235	32.8118	70	157853	21971
TARA_030_SRF_0.22-1.6	ERS478017	TARA_A100001015	33.9179	32.898	5	95690	20253
TARA_031_SRF_0.22-1.6	ERS488545	TARA_A100001388	27.16	34.835	5	124569	19327
TARA_032_DCM_0.22-1.6	ERS488599	TARA_A100001037	23.4183	37.245	80	127264	41284
TARA_032_SRF_0.22-1.6	ERS488569	TARA_A100001035	23.36	37.2183	5	86601	14563
TARA_033_SRF_0.22-1.6	ERS488621	TARA_A100001234	21.9467	38.2517	5	69314	11553
TARA_034_DCM_0.22-1.6	ERS488685	TARA_B100000029	18.4417	39.8567	60	57414	20976
TARA_034_SRF_0.22-1.6	ERS488649	TARA_B100000003	18.3967	39.875	5	64924	12324
TARA_036_DCM_0.22-1.6	ERS488747	TARA_B100000035	20.8222	63.5133	17	100170	15867
TARA_036_SRF_0.22-1.6	ERS488714	TARA_Y100000022	20.8183	63.5047	5	62530	9482
TARA_037_MES_0.22-1.6	ERS488769	TARA_B100000315	20.8457	63.5851	600	103619	37577
TARA_038_DCM_0.22-1.6	ERS488830	TARA_B100000073	19.0284	64.5126	25	100385	18459
TARA_038_MES_0.22-1.6	ERS488849	TARA_Y100000294	19.0351	64.5638	340	54844	20376
TARA_038_SRF_0.22-1.6	ERS488799	TARA_Y100000287	19.0393	64.4913	5	68985	11056
TARA_039_DCM_0.22-1.6	ERS488916	TARA_B100000085	18.5839	66.4727	25	71132	12170
TARA_039_MES_0.22-1.6	ERS488936	TARA_Y100000031	18.7341	66.3896	270	78278	28408
TARA_041_DCM_0.22-1.6	ERS489074	TARA_B100000287	14.5536	70.0128	60	76328	21807
TARA_041_SRF_0.22-1.6	ERS489043	TARA_B100000282	14.6059	69.9776	5	101365	14480
TARA_042_DCM_0.22-1.6	ERS489134	TARA_B100000131	5.9998	73.9067	80	56690	16053
TARA_042_SRF_0.22-1.6	ERS489087	TARA_B100000123	6.0001	73.8955	5	97737	15297
TARA_045_SRF_0.22-1.6	ERS489236	TARA_B100000161	0.0033	71.6428	5	124540	16647
TARA_048_SRF_0.22-1.6	ERS489315	TARA_B100000242	-9.3921	66.4228	5	139279	19155
TARA_052_DCM_0.22-1.6	ERS489585	TARA_B100000214	-16.9534	53.9601	75	96919	22121
TARA_052_SRF_0.22-1.6	ERS489529	TARA_B100000212	-16.957	53.9801	5	108588	21006
TARA_056_MES_0.22-3	ERS489727	TARA_B100000378	-15.3379	43.2948	1000	95667	26453
TARA_056_SRF_0.22-3	ERS489712	TARA_B000000609	-15.3424	43.2965	5	101902	16323
TARA_057_SRF_0.22-3	ERS489733	TARA_B000000565	-17.0248	42.7401	5	100892	15308
TARA_058_DCM_0.22-3	ERS489846	TARA_B000000557	-17.2855	42.2866	66	95349	17760
TARA_062_SRF_0.22-3	ERS489877	TARA_B000000532	-22.3368	40.3412	5	69585	12524
TARA_064_DCM_0.22-3	ERS490002	TARA_B100000405	-29.5333	37.9117	65	161337	36559

TARA_064_MES_0.22-3	ERS489987	TARA_B100000408	-29.5046	37.9599	1000	63894	18245
TARA_064_SRF_0.22-3	ERS489917	TARA_B100000401	-29.5019	37.9889	5	186898	36790
TARA_065_DCM_0.22-3	ERS490085	TARA_B000000441	-35.2421	26.3048	30	155299	33260
TARA_065_MES_0.22-3	ERS490065	TARA_B000000460	-35.1889	26.2905	850	118035	34127
TARA_065_SRF_0.22-3	ERS490029	TARA_B000000437	-35.1728	26.2868	5	61426	14918
TARA_066_DCM_0.22-3	ERS490163	TARA_B000000477	-34.8901	18.0459	30	39410	9907
TARA_066_SRF_0.22-3	ERS490124	TARA_B000000475	-34.9449	17.9189	5	70388	16817
TARA_067_SRF_0.22-3	ERS490183	TARA_B100000497	-32.2401	17.7103	5	40590	8658
TARA_068_DCM_0.22-3	ERS490296	TARA_B100000482	-31.027	4.6802	50	56160	14282
TARA_068_MES_0.22-3	ERS490230	TARA_B100000470	-31.0198	4.6685	700	115422	28384
TARA_068_SRF_0.22-3	ERS490265	TARA_B100000475	-31.0266	4.665	5	68154	17350
TARA_070_MES_0.22-3	ERS490373	TARA_B100000446	-20.4075	-3.1641	800	114132	18904
TARA_070_SRF_0.22-3	ERS490327	TARA_B100000459	-20.4091	-3.1759	5	40905	12012
TARA_072_DCM_0.22-3	ERS490476	TARA_B100000427	-8.7296	-17.9604	100	49909	15592
TARA_072_MES_0.22-3	ERS490507	TARA_B100000508	-8.7986	-17.9034	800	73884	15156
TARA_072_SRF_0.22-3	ERS490433	TARA_B100000424	-8.7789	-17.9099	5	78627	14435
TARA_076_DCM_0.22-3	ERS490597	TARA_B100000519	-21.0292	-35.3498	150	68102	17826
TARA_076_MES_0.22-3	ERS490633	TARA_B100000749	-20.9315	-35.1794	800	115294	29229
TARA_076_SRF_0.22-3	ERS490542	TARA_B100000513	-20.9354	-35.1803	5	78051	18559
TARA_078_DCM_0.22-3	ERS490691	TARA_B100000530	-30.1484	-43.2705	120	87909	24707
TARA_078_MES_0.22-3	ERS490714	TARA_B100000745	-30.1471	-43.2915	800	83820	21627
TARA_078_SRF_0.22-3	ERS490659	TARA_B100000524	-30.1367	-43.2899	5	78486	15927
TARA_082_DCM_0.22-3	ERS490928	TARA_B100000767	-47.2007	-57.9446	40	161515	19154
TARA_082_SRF_0.22-3	ERS490885	TARA_B100000768	-47.1863	-58.2902	5	72336	9655
TARA_084_SRF_0.22-3	ERS491001	TARA_B100000780	-60.2287	-60.6476	5	134591	12896
TARA_085_DCM_0.22-3	ERS491095	TARA_B100000795	-62.2231	-49.2139	90	128279	13524
TARA_085_MES_0.22-3	ERS491110	TARA_B100000809	-61.9689	-49.5017	790	91144	23154
TARA_085_SRF_0.22-3	ERS491044	TARA_B100000787	-62.0385	-49.529	5	163215	10385
TARA_093_DCM_0.22-3	ERS491463	TARA_B100001059	-33.9116	-73.0537	35	186478	24003
TARA_093_SRF_0.22-3	ERS491421	TARA_B100001063	-34.0614	-73.1066	5	89154	11969
TARA_094_SRF_0.22-3	ERS491492	TARA_B100001057	-32.7971	-87.0693	5	127765	28073
TARA_096_SRF_0.22-3	ERS491525	TARA_B100000989	-29.7238	-101.1604	5	136798	36404
TARA_098_DCM_0.22-3	ERS491740	TARA_B100001029	-25.826	-111.7294	188	73002	22201
TARA_098_MES_0.22-3	ERS491767	TARA_B100001013	-25.8076	-111.6906	488	100451	32380
TARA_098_SRF_0.22-3	ERS491699	TARA_B100001027	-25.8051	-111.7202	5	72312	18472
TARA_099_SRF_0.22-3	ERS491804	TARA_B100000886	-21.146	-104.787	5	96066	22106
TARA_100_DCM_0.22-3	ERS491874	TARA_B100000965	-12.9723	-96.0122	50	103040	27536
TARA_100_MES_0.22-3	ERS491913	TARA_B100000959	-12.9794	-96.0232	177	111240	35594
TARA_100_SRF_0.22-3	ERS491836	TARA_B100000963	-13.0023	-95.9759	5	122482	25776
TARA_102_DCM_0.22-3	ERS492012	TARA_B100000902	-5.2669	-85.2732	40	112336	26793
TARA_102_MES_0.22-3	ERS491980	TARA_B100000953	-5.261	-85.1678	480	56449	15606
TARA_102_SRF_0.22-3	ERS491938	TARA_B100000900	-5.2529	-85.1545	5	74287	16312
TARA_109_DCM_0.22-3	ERS492177	TARA_B100000927	2.0299	-84.5546	30	77265	16407
TARA_109_MES_0.22-3	ERS492205	TARA_B100000929	2.0649	-84.5546	380	110154	29743
TARA_109_SRF_0.22-3	ERS492145	TARA_B100000925	1.9928	-84.5766	5	128823	25503

TARA_110_DCM_0.22-3	ERS492264	TARA_B100001113	-1.9002	-84.6265	50	128677	31720
TARA_110_MES_0.22-3	ERS492294	TARA_B100001079	-1.8902	-84.6141	380	114271	31425
TARA_110_SRF_0.22-3	ERS492228	TARA_B100001109	-2.0133	-84.589	5	106731	21501
TARA_111_DCM_0.22-3	ERS492357	TARA_B100000579	-16.9587	-100.6751	90	112059	33628
TARA_111_MES_0.22-3	ERS492381	TARA_B100000586	-16.9486	-100.6715	350	117215	35497
TARA_111_SRF_0.22-3	ERS492321	TARA_B100000575	-16.9601	-100.6335	5	115161	25641
TARA_112_DCM_0.22-3	ERS492445	TARA_B100000945	-23.2189	-129.4997	155	106315	32679
TARA_112_MES_0.22-3	ERS492471	TARA_B100000949	-23.2232	-129.5986	696	82500	26067
TARA_112_SRF_0.22-3	ERS492408	TARA_B100000941	-23.2811	-129.3947	5	119900	26517
TARA_122_DCM_0.22-3	ERS492699	TARA_B100000700	-9.0063	-139.1394	115	93757	27762
TARA_122_MES_0.22-3	ERS492680	TARA_B100000678	-8.9729	-139.2393	600	114755	33262
TARA_122_SRF_0.22-3	ERS492642	TARA_B100001115	-8.9971	-139.1963	5	108073	23996
TARA_123_MIX_0.22-3	ERS492778	TARA_B100000686	-8.9109	-140.2845	150	125791	46323
TARA_123_SRF_0.22-3	ERS492733	TARA_B100000683	-8.9068	-140.283	5	86324	23899
TARA_124_MIX_0.22-3	ERS492863	TARA_B100000676	-9.0714	-140.5973	120	141432	44409
TARA_124_SRF_0.22-3	ERS492821|ERS492814	TARA_B100000674	-9.1504	-140.5216	5	144234	32775
TARA_125_MIX_0.22-3	ERS492926	TARA_B100001123	-8.8999	-142.5461	140	79206	27837
TARA_125_SRF_0.22-3	ERS492888	TARA_B100001121	-8.9111	-142.5571	5	129135	28162
TARA_128_DCM_0.22-3	ERS493098	TARA_B100000614	0.0222	-153.6858	40	76317	18223
TARA_128_SRF_0.22-3	ERS493044	TARA_B100000609	0.0003	-153.6759	5	102780	22661
TARA_132_DCM_0.22-3	ERS493340	TARA_B100001250	31.5168	-159.046	115	104381	35804
TARA_132_MES_0.22-3	ERS493372	TARA_B100001245	31.528	-159.0224	550	112779	30104
TARA_132_SRF_0.22-3	ERS493300	TARA_B100001248	31.5213	-158.9958	5	106220	19556
TARA_133_DCM_0.22-3	ERS493431	TARA_B100001094	35.4002	-127.7499	45	92255	21994
TARA_133_MES_0.22-3	ERS493460	TARA_B100001105	35.2698	-127.7268	650	123509	30214
TARA_133_SRF_0.22-3	ERS493390	TARA_B100001093	35.3671	-127.7422	5	149404	37848
TARA_137_DCM_0.22-3	ERS493670	TARA_B100001964	14.2075	-116.6468	110	118179	43112
TARA_137_MES_0.22-3	ERS493705	TARA_B100001971	14.2025	-116.6433	375	124039	39570
TARA_137_SRF_0.22-3	ERS493636	TARA_B100001287	14.2035	-116.6261	5	119711	22598
TARA_138_DCM_0.22-3	ERS493788	TARA_B100001996	6.3378	-102.9538	60	68607	22370
TARA_138_MES_0.22-3	ERS493822	TARA_B100002003	6.3559	-103.0598	450	100842	33610
TARA_138_SRF_0.22-3	ERS493752	TARA_B100001989	6.3332	-102.9432	5	111087	18771
TARA_140_SRF_0.22-3	ERS493877	TARA_B100002019	7.4122	-79.3017	5	115773	19492
TARA_141_SRF_0.22-3	ERS493914	TARA_B100001939	9.8481	-80.0454	5	77211	13141
TARA_142_DCM_0.22-3	ERS493981	TARA_B100002052	25.6168	-88.4532	125	100452	25802
TARA_142_MES_0.22-3	ERS494006	TARA_B100002049	25.6236	-88.45	640	111238	29431
TARA_142_SRF_0.22-3	ERS493938	TARA_B100002051	25.5264	-88.394	5	109445	25889
TARA_145_MES_0.22-3	ERS494208	TARA_B100001146	39.2392	-70.0343	590	98337	27437
TARA_145_SRF_0.22-3	ERS494170	TARA_B100001142	39.2305	-70.0377	5	84116	20518
TARA_146_MES_0.22-3	ERS494274	TARA_B100001167	34.6663	-71.2907	640	83957	25957
TARA_146_SRF_0.22-3	ERS494236	TARA_B100001540	34.6712	-71.3093	5	118221	34571
TARA_148_SRF_0.22-3	ERS494332	TARA_B100001741	31.6948	-64.2489	5	95080	34411
TARA_148b_MES_0.22-3	ERS494374	TARA_B100001750	34.1504	-56.9684	250	118053	31322
TARA_149_MES_0.22-3	ERS494431	TARA_B100001765	34.0771	-49.8233	740	109336	32447
TARA_149_SRF_0.22-3	ERS494394	TARA_B100001758	34.1132	-49.9181	5	120367	30813

TARA_150_DCM_0.22-3	ERS494488	TARA_B100001778	35.8427	-37.1526	40	118668	27553
TARA_150_SRF_0.22-3	ERS494445	TARA_B100001769	35.9346	-37.3032	5	123658	28001
TARA_151_DCM_0.22-3	ERS494559	TARA_B100001559	36.1811	-28.9373	80	104170	28925
TARA_151_SRF_0.22-3	ERS494518	TARA_B100001564	36.1715	-29.023	5	122930	28296
TARA_152_MES_0.22-3	ERS494616	TARA_B100001179	43.7182	-16.8714	800	113269	29208
TARA_152_MIX_0.22-3	ERS494628	TARA_B100001175	43.7056	-16.8794	25	109195	31987
TARA_152_SRF_0.22-3	ERS494579	TARA_B100001173	43.6792	-16.8344	5	92320	23463
					Total	14129971	3323839
					Mean	101654.5	23912.51079
					Min	39410	8658
					Max	186898	46323

Sample label (Tara + station number + depth layer + size fraction), identifiers for PANGEA and INSDC, geographical coordinates and sampling depth and miTags' mapping statistics for all the samples analyzed. More information may be acquired by matching identifiers to Ocean Microbiome Companion Site's Tables (http://ocean-microbiome.embl.de/companion.html) from (Sunagawa *et al*. 2015).

Acknowledgements

Estaré sempre agraït a Pep i Silvia, que em van donar la possibilitat de fer aquest doctorat. Agraït per la confiança durant aquests anys en que pogués treure endavant aquest treball, pels reptes compartits, per les coses que han funcionat i les que no, per la paciència. Ells n'eren experts, jo començava escrivint un e-mail des de València, amb tot per aprendre. El que ara en sé de fer ciència és bàsicament el vostre mèrit.

Sempre estaré agraït als incondicionals, als més propers, als companys del viatge. A Fran i Anamari pels milers de vivències viscudes des de l'inici i durant tants anys que son impossibles de resumir, a Massimo, el meu tàndem de tantes altres aventures, a Clara, per la dolçor en cada moment, a Sara, per la dolçor amagadeta, a Juancho, imprescindible en tants moments, a Pablo, quantes converses amb cafè, cervesa o orxata i a Ramiro, indispensable. A Eli, Elena, Mariona. A tots aquests i altres que s'entrecreuen i repeteixen en tants significants: als ping-pong piti, els del volley, l'Acinas dream team, els wassapitos els companys de despatxos. A Jordi, a Sdena, a Caterina, a Isabelita, a Mireia. A Clara i Vane, a Isabel, a Marta. Als incomptables companys de campanyes.

Agraït també a tants de qui he aprés tant i que sàviament mantenen aquest lloc tan especial: Marta Estrada, Carlos Pedrós, Rafel Simó, Celia Marrasé, Montse Sala, Dolors Vaqué, Elisa Berdalet, Esther Garcés, Albert Calbet, Miquel Alcaraz y Enric Saiz i Ramón Massana.

En fi, agraït a la gran i diversa família del ICM. Als seus membres individualment i a allò, difícil de definir, que es una mica més que la suma de les persones que s'han mogut i es mouen pels seus passadissos.

Agraït als meus pares. Agraït infinitament a Marieta, continuem viatge...