MAX-PLANCK-GESELLSCHAFT

Max Planck Institute
for Marine Microbiology

marmic

JACOBS UNIVERSITY

Microbial Community Ecology of Temperate Coastal Sands

By

Angélique Gobet, M. Sc.

A thesis submitted in partial fulfillment of requirements for the degree of

Doctor of Philosophy in Biology

Defense at Jacobs University Bremen, School of Engineering and Science,
on December 13th 2010

Approved Thesis Committee

Prof. Dr. Antje Boetius (Chair)

HGF-MPG Group for Deep Sea Ecology and Technology,

Alfred Wegener Institute for Polar and Marine Research

University of Bremen

Jacobs University

Prof. Dr. Matthias Ullrich

Jacobs University

Dr. Alban Ramette

Max Planck Institute for Marine Microbiology

Statement of Sources

Declaration

I, Angélique Gobet, certify that this thesis is my own work and has not been submitted at another university or other institution of tertiary education for the conferral of a Degree or diploma. Information derived from published or unpublished scientific work has been cited in the text and listed in the references.

Date **Signature**

Thesis Abstract

Classical community ecology concepts are just beginning to poke microbial ecology. Indeed, the development of high-throughput molecular techniques coupled with community ecology theories offers promising opportunities to understand the ecology of microbes. However, we are only starting to cope with the colossal work of explaining the structuring and the ecology of microbial communities. Further analyses are yet to be done to decipher processes structuring communities such as speciation, selection, ecological drift or passive dispersal. Hence, this requires the comprehension of the effect of time, space and the environment on the structuring of microbial communities. The understanding of microbial community ecology first requires the development of time- and cost-effective tools, and pipelines that should be available to most microbial ecologists. These tools must also lead to a high resolution description of microbial communities and variations of their ecological patterns. In the present thesis, patterns of diversity, community structure and ecology were investigated on temperate coastal sandy sediment. Samples taken over a two-year period were obtained from a previous study where changes in bacterial community composition linked to ecosystem dynamics were previously observed (Böer et al. 2009). Automated rRNA intergenic spacer analysis (ARISA) and high-throughput 454 massively tag sequencing (MTPS) were applied on these samples.

Despite being known as a consistent descriptor of microbial community patterns, ARISA may appear obsolete in comparison to emergent high-throughput sequencing technologies. We thus compared ARISA and 454 MPTS to check whether each approach could be better adapted to a specific type of question. Whereas observing high differences in community turnover, both techniques presented similar patterns in community structure. Additionally, the same combination of biogeochemical parameters could explain the resulting microbial ecological patterns. This study validated ARISA as a consistent technique to describe microbial community patterns and also suggested to couple community fingerprinting and high-throughput sequencing techniques for (i) a broad and rapid overview of diversity patterns in many samples as well as (ii) a detailed description of microbial community composition and dynamics.

As next-generation sequencing techniques are emerging, a massive amount of data is accumulating and a more thorough interpretation of high-throughput data sets is needed. By implementing new user-friendly statistical tools (MultiCoLA, www.ecology-research.com), we tested the effect of applying successive definitions of rare and abundant types in complex community data sets on the resulting ecological interpretation. Similar ecological patterns could be observed even after removing a high proportion of the data set (35-40%). This study confirmed the importance of defining different fractions of the microbial community for a consistent ecological interpretation of large community data sets.

Some recent studies using 454 MPTS showed an unprecedented diversity and allowed preliminary conclusions on the distribution patterns of different subsets of microbial communities. For more insights into microbial ecological patterns, we applied 454 MPTS on temperate coastal sands, a highly dynamic marine environment characterized by strong physical mixing and seasonal variation. There were remarkable shifts in community composition over a few centimeters of sediment depth or between any two consecutive sampling times, with up to 70-80% of community turnover. These drastic shifts were not random as most of the variation could be attributed to a combination of biogeochemical parameters (*e.g.* temperature, nutrients, pigments, production of extracellular enzymes), and to specific shifts of the large majority of rare bacterial types. This study thus demonstrates how dynamic microbial diversity may be in coastal sandy sediments. Many microbial niches may be created by strong vertical shifts in nutrient, organic matter and oxygen availability, which may support a high turnover of bacterial types in sandy sediments.

The accomplishments of this PhD thesis allowed improvements to extract the deeper meaning out of complex community data sets. This work shed light on main processes shaping microbial communities, by constructing robust bases in microbial community ecology.

Table of content

1 Introduction

1.1 Global Biodiversity

1.1.1 Importance of Studying Biodiversity

Global warming is currently changing environmental conditions on Earth at an unprecedented rate, inducing dramatic shifts in structure, species composition, and functioning of ecosystems. Such variations may lead to biodiversity loss, or habitat destruction (Sala et al. 2000). Biodiversity is currently being lost at an accelerating rate, with an estimated 50% of all species, including mammals, birds, and reptiles, to be lost in the next 300–400 years (Mace 1995). Such a reduction in biodiversity is of public concern, as it directly affects ecosystem services, and may therefore have significant socio-economical consequences (*e.g.* health, agriculture or economy). An example of how species invasion can directly influence our everyday-life, is the invasive zebra mussel, which induced the extermination of native mussel populations into the Great Lakes. This invasion led to shut down electrical utilities by clogging water intake pipes, costing about 5 billion dollars, according to the U.S. Fish and Wildlife Service (Wilson 2000).

Despite observing high extinction rates [50 to 100 times the average expected natural rate, (Hawksworth & Kalin-Arroyo 1995)], Wilson stated that the whole extent of biological diversity remains far from being completely described [**Fig. 1.1.**, (Wilson 2000)]. Indeed, there is a very wide range of estimation of the biodiversity on Earth (13-14 millions estimated species), and it seems that only about 13% has been scientifically described [**Fig. 1.1.**, (Hawksworth & Kalin-Arroyo 1995)]. In the big picture, Bacteria seem to remain as a real black hole, especially considering the relatively small area they occupy in the pie [**Fig. 1.1.** (Wilson 1992, 2000)]. This confirms the need to thoroughly study biodiversity to engender a better understanding of current environmental changes and its consequences on ecosystem services.

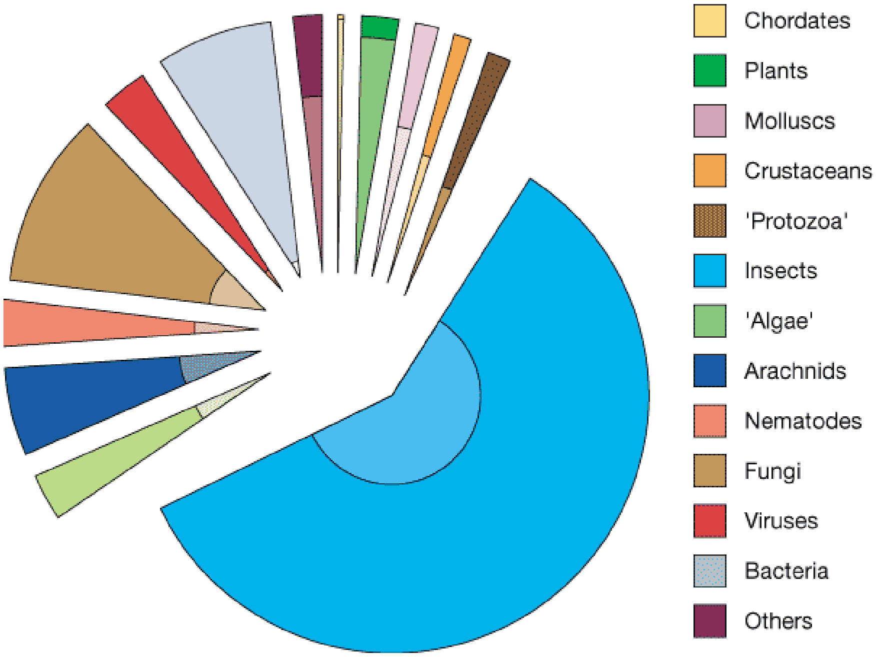


Figure 1.1. Species richness in major groups of organisms. The main 'pie' shows the species estimated to exist in each group; the smaller area within each slice shows the described proportion. Figure from (Purvis & Hector 2000), with data from (Hawksworth & Kalin-Arroyo 1995).

1.1.2 Evaluating the Role of Microbes on Earth

Over the last few decades, compared to macroorganisms, we now know that microorganisms are extremely more diverse than hitherto presumed (**Fig. 1.1.**) and may actually represent most of Earth’s biodiversity (Pace 1997, Torsvik et al. 2002, Venter et al. 2004). For instance, bacterial diversity was estimated to reach 2.10^6 species in the global ocean (Curtis et al. 2002). Microbes may thus be the key to understanding drastic shifts in environmental processes. Although most of them are invisible to the human eye, microbes constitute an essential component of Earth’s biosphere and can be found in all types of habitats, including marine sediments, aquatic systems and living organisms. The impressive microbial world has been estimated to $4\text{-}6.10^{30}$ cells on Earth (Whitman et al. 1998), which represent about 10^9 times the number of stars in the universe (Curtis & Sloan 2004). Their importance can be underlined by the total amount of prokaryotic biomass they represent, which is about 60–100% of the estimated total plant biomass [*i.e.*

350–550 Pg of C, with 1 Pg = 10^{15} g, (Whitman et al. 1998)]. Also, microbes represent the largest living reservoir of nutrients, and contribute to major ecosystem processes by recycling elements (Whitman et al. 1998, Azam & Worden 2004, Falkowski et al. 2008). Hence, determining the extent of microbial diversity, together with the factors shaping it, is of prime concern. Considerable efforts are thus required toward characterizing microbial diversity patterns, which may be enlightened by applying community ecology concepts (**Box 1**, **§ 1.3.1**).

In this introduction, some of the classical concepts used in community ecology are reviewed, followed by a state of the art of the current knowledge on microbial diversity patterns. The applications of community ecology concepts in microbial ecology as well as their efficiency to unveil microbial ecological patterns are also reviewed.

Box 1. Brief Historic of Some Community Ecology Terms.

"*-Oecologie…der Wissenschaft von der Oeconomie, von der Lebensweise, von der äusseren Lebensbeziehungen der Organismem zu einander etc*." Haeckel, 1866.

Ecology. The term Ecology was first introduced in 1866 by the theoretical morphologist and field naturalist Ernst Haeckel in his book *Generelle Morphologie der Organismen* (Stauffer 1957). Haeckel originally defined Ecology as a restraint of the term „biology" which would be the science of the economy, of nature, the external relationships between organisms, etc. Haeckel's enthusiasm to Darwin's *Origin of species* directly influenced his original definition of Ecology, which then evolved through Haeckel's successive statements (Stauffer 1957). According to Begon, a more precise definition of Ecology would be "the scientific study of the distribution and abundance of organisms and the interactions that determine distribution and abundance" (Begon et al. 2006). In other words, Ecology should allow the understanding of the abundance and distribution of organisms in the environment, their succession in ecosystems, their adaptations, and the processes of energy transfers through the living communities.

Community. Concepts of community were first laid within plant Ecology. At first, no consensus definition of community existed, but several definitions were in use. For instance, in the 1920-1930's, Gleason centered the community rather on the species needs, while Clements based it on the species interactions. Nowadays, an agreed definition of the community would be the patterns in comparison and diversity of species living close enough to interact in a specified place and time (Whittaker 1975). A community can be defined by its richness, its evenness, its composition and its functional characteristics.

Biological diversity and biodiversity. In *Measuring Biological Diversity*, Magurran reviewed the historic and successive usages of the term "biological diversity" and could trace it back to 1955 (Gerbilskii & Petrunkevitch 1955), in the context of intraspecific variation in behavior and life history of sturgeons. In 1986, biological diversity was divided into three terms: genetic diversity (within-species diversity), species diversity (number of species) and ecological diversity [diversity of communities, (Norse 1986, Harper & Hawksworth 1995)]. The term "biodiversity" was then introduced by Rosen in 1986 (Harper & Hawksworth 1995). In *The Unified Neutral Theory of Biodiversity and Biogeography*, Hubbell proposed the term biodiversity to be considered as a "synonymous with species richness and relative species abundance in space and time" (Hubbell 2001). Also, it seems that both terms biological diversity and biodiversity are commonly used as synonyms by most authors (Harper & Hawksworth 1995, Magurran 2004).

When studying community Ecology, scientists often describe the species diversity within a context of community and seek to identify the principal processes structuring the community, *i.e.* the factors determining the species diversity, composition and abundance in the community.

1.2 Measuring Microbial Diversity

1.2.1 Species Concept

In microbial ecology, one of the biggest issues in measuring biological diversity is not a methodological or algorithmic one, it is simply coming to an agreement on what a microbial "species" actually is. What "unit" should these tests of diversity measure? For macroorganisms, species are most commonly defined according to Mayr's biological species concept (Mayr 1942). According to this definition, a species is a group of interbreeding individuals isolated from other groups by barriers of recombination (Mayr 1942). This is a genetic definition of species, implying that members of the same species should have genetic exchange sufficiently extensive, thus being genetically homogeneous among themselves and distinct from other species.

However, this species concept cannot be applied to microorganisms. Indeed, the short generation time and clonal reproduction of microbes result in the absence of clear genetic isolation (Acinas et al. 2004). On the other hand, their generally short generation time enhances their genetic mutation rate, and causes a high variability in microbial genomes. This variability is increased by horizontal gene transfer (Ochman et al. 2000), a mechanism allowing microbes to acquire genes from surrounding organisms, genetically related or not. This genetic variability provides microbes with a high adaptability to ever-changing and complex environments (Rosenzweig et al. 1994). Subsequently, some microbiologist even believe that the species concept cannot be applied to microbes [see (Achtman & Wagner 2008)]. Hence, a common agreement among microbial ecologists is to use an arbitrary unit to describe bacterial diversity, by referring to "operational taxonomic units" (OTU), rather than species (**Box 2**).

1.2.2 Classical Community Ecology Tools Applied to Microbial Community Ecology

Measuring alpha-diversity (Box 3). There are several approaches to deal with the daunting task of predicting microbial richness. Among them, rarefaction curves allow an estimation of the expected sampling effort to cover the whole diversity by plotting types, or “units”, against individuals (Gotelli & Colwell 2001). Microbial ecologists commonly use rarefaction curves to estimate whether their sampling effort is sufficient to accurately estimate microbial diversity from 16S rRNA clone libraries [**Fig. 1.2.A** and (Kemp & Aller 2004)]. Unfortunately, the required number of clones needed to fully represent the community diversity (*i.e.* to reach an asymptote with the rarefaction curve) is usually quite high, and additional sampling effort is thus necessary [**Fig. 1.2.A,** (Kemp & Aller 2004)]. Interestingly, whereas next-generation sequencing techniques might appear as an alternative to the limitations of clone library-based studies (**Box 2**), it is becoming obvious that these high-throughput sequencing methods are still usually far from unveiling the full diversity of bacterial communities [**Fig. 1.2.B**, (Sogin et al. 2006)]. For instance, further calculations of diversity estimates have shown that even the better-sampled site from Sogin’s study would need a sample size 280 times larger, requiring 120 million reads, just to reach 90% of the bacterial diversity (Quince et al. 2008). However, these methods do seem to encompass a more representative portion of the taxa-poor archaeal diversity [**Fig. 1.2.C**, (Galand et al. 2009a, Brazelton et al. 2010)].

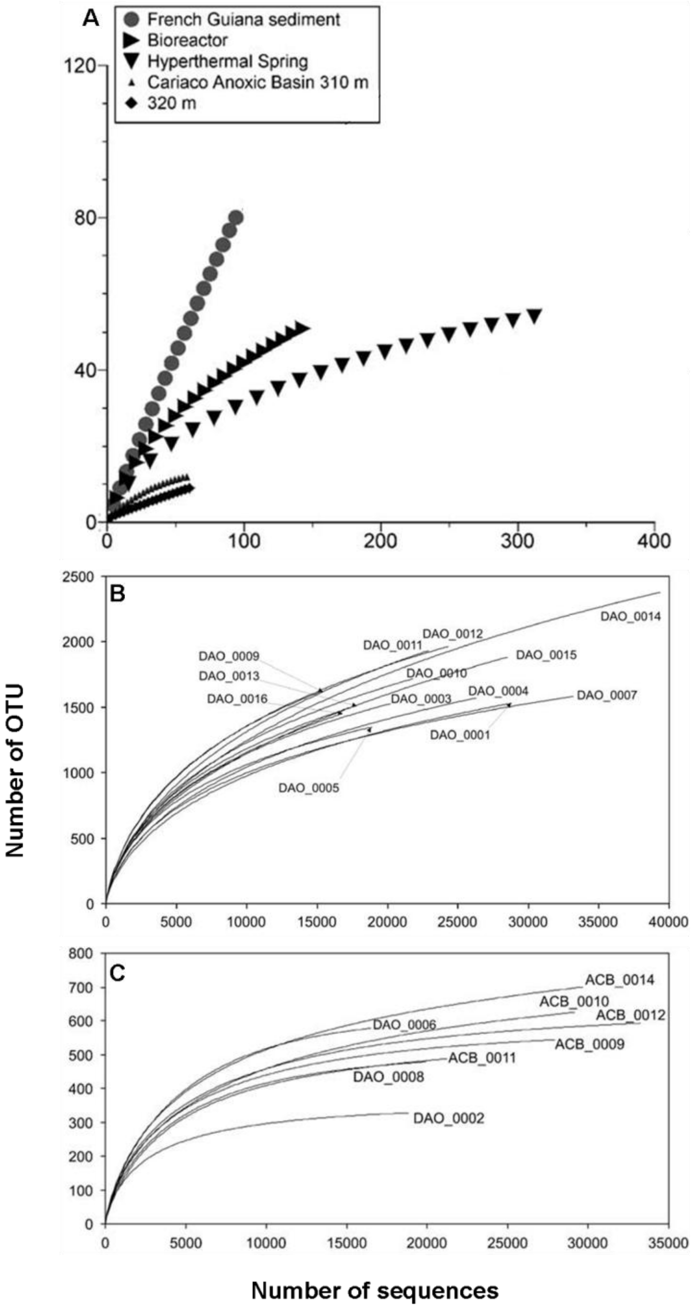


Figure 1.2. Rarefaction curves. (**A**) Clone library-based sequencing on several example of environments, modified from (Kemp & Aller 2004), and (**B**, **C**) 454 massively parallel tag sequencing (MPTS) on water column from the Arctic ocean. (**B**) Rarefaction curves of Bacteria (Galand et al. 2010), (**C**) Rarefaction curves of Archaea (Galand et al. 2009b). ACB, surface waters, DAO, deep water masses. Numbers indicate sample identification.

The rank-abundance curve is another option to estimate microbial diversity. It allows an approximation of the frequency distribution of the sampled species and an estimation of the unsampled ones [see **§ 1.3.2.** and (Curtis et al. 2002, Magurran 2004)].

An additional way to estimate the species richness of a community is by using coverage-based nonparametric estimators, such as Chao or ACE estimators (Chao 1984, Chao & Lee 1992). For a reliable estimation of diversity, Chao and ACE estimators should be applied to large community data sets, with a good coverage of the total diversity (Chao & Bunge 2002). These estimators are often used to estimate microbial species richness (Curtis et al. 2002), but they may not be appropriate for all microbial studies. In the case of 16S rRNA-based libraries and high-throughput sequencing technologies, when data is limiting, these robust estimators may underestimate the true microbial diversity (Hong et al. 2006, Quince et al. 2008).

Measures of beta-diversity (Box 3). In the case of 16S rRNA-and PCR-based technologies, alpha-diversity estimates may not be adequate [see **§ 1.2.3** and (Bent & Forney 2008, Reeder & Knight 2010)]. Studying patterns of the community structure may be better suited to analyze such species data sets (see **§ 1.2.3**). For a better understanding of microbial community ecology (see **§ 1.3.1.**), methods traditionally used in community ecology may allow a better investigation of relationships between the observed microbial diversity and their environment (Ramette 2007). To measure beta-diversity, *i.e.* compare the diversity between two sites or along an environmental gradient, the similarity or dissimilarity between the sites, or samples, has to be calculated. Many indices are available to compare samples, amongst them, asymmetrical coefficients are preferable to deal with the double zero problem (Legendre & Legendre 1998). These types of coefficient do not treat double absence of a species in two samples similarly as a species can be absent for different reasons [*e.g.* it can be because of sampling resolution, (Legendre & Legendre 1998)]. When dealing with abundance data, the Bray-Curtis index [see equation (1), (Bray & Curtis 1957)] may be used.

$$BC_{ij} = \frac{S_i + S_j - 2C_{ij}}{S_i + S_j} \tag{1}$$

*Where **i**, **j** are each sample; **S** the richness in each sample; **C** is the number of species common to both samples.*

It represents the total number of unique species to each sites (or turnover) divided by the total number of species over the two sites. When dealing with presence-absence (binary) data, one may choose the Jaccard similarity coefficient [see equation (2), (Jaccard 1901)], which measures similarity between samples, and is defined as the size of the intersection divided by the size of the union of the samples.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{2}$$

*Where **A** and **B** are each sample.*

Such indices may be applied to compare samples in the species data set before using multivariate analyses [see all chapters of the thesis and (Ramette 2007)].

Also, the advance of sequencing technologies lead to a massive accumulation of microbiological information, along with the accumulation of contextual environmental parameters. There is still a need to systematically study beta-diversity of microbes by interpreting the massive data output and making such pipelines available to a larger range of scientists [(Rothberg & Leamon 2008), (**Chapters 1 and 2**)].

1.2.3 Resolution and Biases of the Techniques

Sampling design. It is important to note that sampling has a major role in the future description of communities. For instance, scientists should make sure to sample as randomly as possible and to take replicates of each similar type of sample (Magurran 2004). First, the sample strategy (*i.e.* where to sample, the number of replicates) is extremely important for robust ecological interpretation (Horner-Devine et al. 2004, Green & Bohannan 2006). In microbial ecology, a common solution to these problems is to work on pooled samples, by mixing extracted DNA from spatial replicates (Schwarzenbach et al. 2007). However, this solution may not apply in the case of spatial studies, where sample pooling would remove spatial variability (Prosser 2010). Also, sample size is an important factor to take into account. Indeed, the number of species observed increases with the sampling effort until the total richness is sampled. As such, the larger the sampling effort, the greater the chance is to obtain the total diversity in the community, especially considering the highly diverse microbial world (Curtis et al. 2002). This has been observed when filtering large volume of water column, a large sample representing a great number of different microbial species (Venter et al. 2004). These sampling issues are similar to those encountered for animals and plants studies but, for molecular-based studies, there are additional technical issues that may induce biases.

Technical limitations in molecular ecology. In molecular-based studies such as the “rRNA approach” (**Box 2**), several biases (here not exhaustively listed) can inflate estimates of microbial diversity. Among them, the PCR-step can be critical. The DNA polymerase enzyme can make copy errors, hop from one fragment to the other and create chimeras (Qiu et al. 2001) but can be corrected by sequence clustering [*e.g.* 99% sequence similarity, (Acinas et al. 2005)]. Also, formation of chimeras or heteroduplex molecules (Qiu et al. 2001) can be avoided by changing PCR conditions (Acinas et al. 2005). It is also known that PCR skew the distribution of PCR products as it amplifies inequally DNA fragments (Acinas et al. 2005). The choice of primers can also greatly influence the observed microbial diversity (Huber et al. 2009). Data output deriving from PCR-based techniques must then be interpreted carefully. In the rRNA approach, the cloning step also presents some bias as heteroduplex molecules may be subjected to *E.coli* DNA repair mechanisms, which result in hybrid plasmid inserts (Parker & Marinus 1992).

The development of high-throughput molecular techniques such as automated rRNA intergenic spacer analysis (ARISA) or 454 massively parallel tag sequencing (MPTS) allows to avoid biases depending on cloning and allows a rapid processing of many samples (see **box 2 and Chapter 1**). However, other type of biases should be taken into account. For instance, ARISA permits a rapid identification of bacterial types by identifying the size of the ITS region of the 16S rRNA gene. As it does not allow the identification of the composition in nucleotide of the ITS fragment but only gives fluorescence profiles, each identified peak may represent several taxa with the same phylotype size (Crosby & Criddle 2003, Yannarell & Triplett 2005). Consequently, ARISA may not be well suited to estimate bacterial alpha-diversity in the sample (Bent &

Forney 2008). For a taxonomic identification, ARISA needs to be coupled with sequencing techniques (Fisher & Triplett 1999, Brown et al. 2005).

As 454 MPTS (for a detailed description of the technique, see **§ 1.4.2.2.**) is becoming of common use, concerns about data reliability are emerging. The accuracy of the massive data output is highly discussed as this technique may be affected by PCR and pyrosequencing errors and may produce chimeric types (Quinlan et al. 2008). Several studies have proposed ways of handling these issues. For instance, tag sequences can be clustered by following 97% sequence similarity (Kunin et al. 2010), or the 454 MPTS output can be corrected at the pre-clustering level, based on the electrophoregram (Quince et al. 2009). The data can also be denoised based on the prefixes of tag sequences, resulting in a lower estimation of alpha-diversity (Reeder & Knight 2010). However, the reliability of 454 MPTS sequences may not be that of an important issue, if the scientific question does not depend on estimating the diversity. Actually, the above sequencing errors may inflate alpha-diversity (**Box 3**) estimates but not the beta-diversity interpretation (Gobet et al. 2010, Reeder & Knight 2010).

Finally, an important point is that many studies compare communities based on no further than the phylum or class level. Even though constantly increasing, the general taxonomic knowledge is obviously limited, and current high-throughput sequencing studies confirm the enormous lack of knowledge considering the amount of microbes not yet identified (*e.g.* in the data set processed for this thesis, 20% of the tag sequences were identified from the phylum to the genus level).

To conclude, in the case of high-throughput molecular techniques (*e.g.* ARISA & 454), the effect of such technical biases on the resulting ecological interpretation is still not well studied. Consequently, there is a need (i) to confirm the consistency of data output from fingerprinting techniques with that of high-throughput sequencing techniques (**Chapter 1**), (ii) to analyze the effect of correcting or truncating 454 MPTS data sets on microbial ecological patterns and see whether the correction of such data sets indeed matters for the ecological interpretation (**Chapter 2**), and to test for the consistency of ecological patterns at different taxonomic levels (**Chapters 1 and 2**).

Box 2. Overview of Microbial Ecology Tools.

Operational taxonomic unit (OTU). OTU are most of the time based on the 16S rRNA gene and can be defined by clustering sequences according to a chosen percentage of sequence similarity (Rossello-Mora & Amann 2001).

Traditional molecular approaches. As cultivation is a labor-intensive way to identify microorganisms, the "ribosomal RNA approach", including culture-independent molecular techniques [PCR, 16S rRNA-based clone libraries and Sanger sequencing (Sanger et al. 1977)], was proposed in the 1980's, to rapidly describe microbial diversity (Olsen et al. 1986). This set of molecular biology tools enables the description of microbial diversity by targeting nucleic acids, a record of microbes' evolution and functional processes (Olsen et al. 1986, Woese 1987). For a good resolution of the microbial community description and estimation of the bacterial diversity, many clones need to be screened (Hughes et al. 2001), thus sequence-library based approaches appear relatively time-consuming and expensive. To cope with those central problems, molecular fingerprinting techniques allow a better resolution of microbial diversity from many samples (Fisher & Triplett 1999). Reproducible patterns can be obtained by separating DNA phylotypes either according to their nucleic acid content [*e.g.* denaturing gradient gel electrophoresis *(*DGGE), (Muyzer et al. 1993)] or to their size [*e.g.* terminal restriction fragment length polymorphism (T-RFLP), (Avaniss-Aghajani et al. 1994), automated rRNA intergenic spacer analysis (ARISA), (Fisher & Triplett 1999)].

New high-throughput opportunities to describe microbial diversity. The quest to describe microbial communities is currently ongoing via the development of high-throughput sequencing techniques, leading toward a high resolution description of the microbial world. Indeed, many studies are currently being held using 454 massively parallel tag sequencing [MPTS, (Margulies et al. 2005)] in a wide range of fields [*e.g.* soils (Lauber et al. 2009), water column (Galand et al. 2009a, Gilbert et al. 2009), deep sea vents (Sogin et al. 2006, Huber et al. 2007), human gut (Zhang et al. 2009) and hand surface (Fierer et al. 2008)]. However, as 454 MPTS becomes cheaper, an increasing number of samples can be processed simultaneously. Consequently, the data output is growing at an unprecedented rate and new issues are emerging. For instance, the accuracy of the huge data output and the handling of data are still limiting factors (Quinlan et al. 2008, Quince et al. 2009, Reeder & Knight 2009, Kunin et al. 2010, Reeder & Knight 2010). Moreover, there is a need to make these new technologies available to a larger range of scientists (Rothberg & Leamon 2008).

1.3 Applying Community Ecology Concepts to Microbes

1.3.1 Basic Concepts in Community Ecology

‘‘Community ecology in particular is about to emerge as one of the most significant intellectual frontiers of the 21st century.’’ (Wilson 2000).

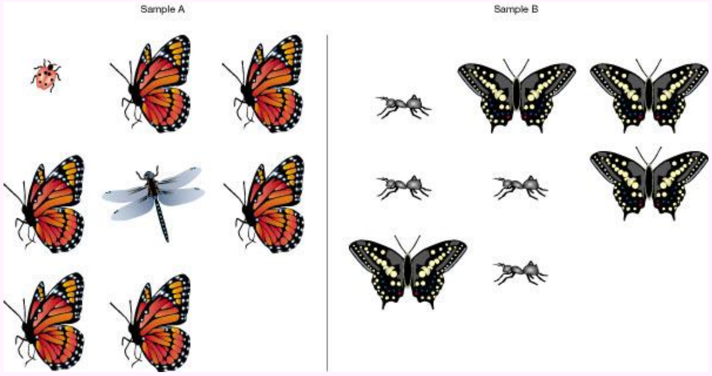
Community Ecology typically consists of studying composition, diversity, and abundance of species in the community, variations in patterns of community structure, as well as functional interactions between species [*e.g.* (Konopka 2009)]. The community structure depends on environmental and historical processes, *i.e.* the combination of local abiotic dynamics and ecological succession of species in the community. Four main classes of processes are involved in shaping community assemblages, namely selection, drift, speciation and dispersal (Vellend 2010). For instance, local abiotic and biotic conditions can lead to the adaptation or even speciation of a given population or set of populations and new species may arise. Additionally, species may migrate into a given ecosystem via passive or active dispersal. Some individuals may become extinct due to stochastic changes in species abundances (*i.e.* ecological drift) or due to deterministic processes (*i.e.* selection), as some species have a fitness advantage over other species for a given niche (Vellend 2010). Notably, all these processes are induced by abiotic (environmental conditions, time and space) and biotic (ecological interactions between populations: coevolution, predation, competition[1], symbiosis; and species intrinsic characteristics: reproduction, niche preferences) factors.

[1] *Life is a jungle, one must fight to succeed,* from the original *“La vie c’est la jungle il faut se battre pour y arriver”,* 1992, Les Inconnus.

Box 3. Community Ecology Definitions.

Alpha-diversity. a component of diversity that considers the total number of species present within a particular area, community or ecosystem, it is usually **Species richness** [**Fig. B.1.1.**, (Whittaker 1972)].
Species evenness. a component of diversity that considers how individuals are distributed among species (**Fig. B.1.1.**). Its estimation permits to complement information obtained from species richness.

Figure B.1.1. Illustration of species richness and species evenness. These are two samples of insects from different locations. With three species, sample A present a higher species richness than sample B. However, species are more evenly distributed in sample B. Sample B may thus be considered more diverse as there is less chance to pick randomly two individuals of the same species than in sample A (Purvis & Hector 2000).



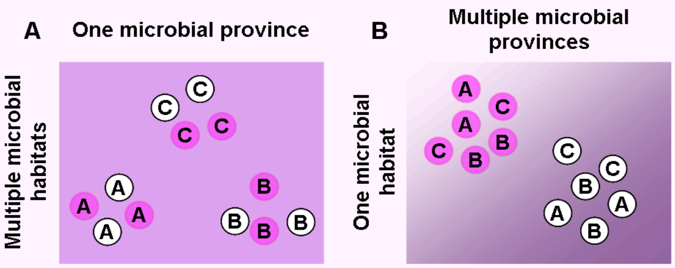
Beta-diversity. Scientists usually compare habitat biodiversity (*i.e.* beta-diversity, **box 1**) to garner insights into organism distribution. Beta-diversity compares the difference in diversity or the species turnover between two given locations. Hence, it informs about the dynamics of the community (Magurran 2004). Studying beta-diversity also consists in deciphering the extent of change of community composition in relation to variations in environmental factors (Whittaker 1960).

Niche's most widely accepted definition is: "The niche is the set of biotic and abiotic conditions in which a species is able to persist and maintain stable population sizes." (Hutchinson, 1957).

Habitat of a species is a related but distinct concept to the Niche and describes the environment over which a species is known to occur and the type of community that is formed as a result (Whittaker et al. 1973).

Biogeography. In classical community ecology, it has been shown that species turnover along spatial gradients has an effect on species-area relationships and total species richness (Hubbell 2001). Species biogeography describes patterns of distribution across geographical areas, and can be explained or predicted through knowledge and understanding of species traits and niche requirements (Pearman et al. 2008). For instance, biogeography can be due to either (i) **sympatric speciation** (*i.e.* the creation of new species due to a strong influence of contemporary environmental parameters, **Fig. B.1.2.A**), implying multiple habitats with different environmental conditions within one province, or to (ii) **allopatric speciation** (*i.e.* the creation of new species due to geographic barriers, with historical influences and a lack of dispersal), implying multiple provinces and one habitat (Martiny et al. 2006).

Figure B.1.2. Two types of speciation that can explain biogeography. (**A**) Sympatric speciation, (**B**) Allopatric speciation. Samples are indicated in circles, their colors (pink vs. white) indicate the location and a letter A, B or C indicates their habitat types. Axes have no dimension, samples closer to another have a more similar species composition than other samples situated further. Modified from (Martiny et al. 2006).



1.3.2 On the Commonness and Rarity of Microbes

One of the most basics and intuitive way to understand biodiversity patterns is to count the species inhabiting various environments. Indeed, when a community is considered at an equilibrium, individuals are often partitioned among few abundant species, moderately common species, and many rare species (Putman 1994, Pachepsky et al. 2001). These observations suggest that diversity should be described by taking into account species richness and evenness (**Box 3**), which can be represented on a species rank-abundance curve (Magurran 2004). This curve forms by plotting species ordered from most to least abundant on the *x* axis, and the abundance of each type observed on the *y* axis [**Fig. 1.3.,** (Magurran 2004)].

Considering the massive extent of microbial diversity (Curtis et al. 2002), assessing organisms' richness may be insufficient for understanding microbial diversity. Likewise, as for macroorganisms, microbial species abundances are not evenly distributed in the community. For instance, similar patterns emerged from the comparison of species abundance distribution of tropical moth communities with that of temperate soil bacteria (Hughes et al. 2001). Few abundant species were observed in the community, while most types were rare, producing a long right-handed tail on the rank-abundance curve [**Fig. 1.3.,** (Hughes et al. 2001, Pedrós-Alió 2006)**]**. This typical hollow curve can thus be used to estimate the total number of microbial species from the total number of individuals and the abundance of the most abundant type. Curtis et al. (2002) applied this method by assuming that the prokaryotic species-abundance distribution was in equilibrium, following a log-normal species abundance curve, with many rare species and relatively few common ones (Curtis et al. 2002). This was further confirmed by several observations, which reported a huge number of rare species, representing most of the total diversity of the community [*e.g.* 63% of the total richness in clone libraries, (Pommier et al. 2007)]. This brought forth the possibility to apply an established, traditional ecological method to microbial communities, paving the road towards addressing questions pertaining to the rare microbial communities (Hughes et al. 2001).

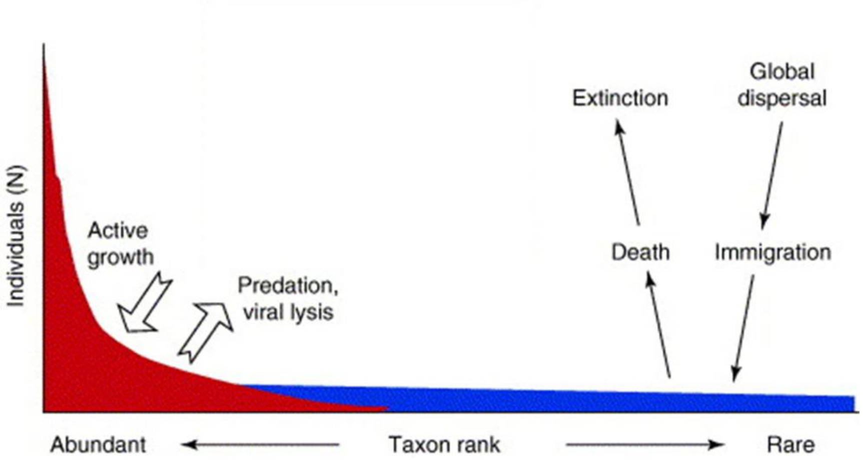


Figure 1.3. (**A**) Rank-abundance curve representing biodiversity and composed of two sections: (i) in red, diversity of the most abundant taxa, (ii) the blue section of the curve corresponds to rare taxa. Modified from (Pedrós-Alió 2006).

1.3.3 Distribution of Rare and Dominant Microbes

In a 21-year survey of estuarine fishes, species were distinguished according to their abundance on the rank-abundance curve. Dominant species identified every year were described as the “core” species, with an active role in carbon and energy flow. In parallel, rarer species were defined as “occasional”, and referred to the transient fish that would stay until their limit of tolerance and disappear again (Magurran & Henderson 2003).

Since microbes display a long right-handed tail on their rank-abundance curve [**Fig. 1.3.,** (Curtis et al. 2002, Pedrós-Alió 2006)], Pedrós-Alió attempted to identify the same two entities, “core” and “occasional”, as seen among the estuarine fishes. He hypothesized that the “core” species are defined as the dominant, active, and persistent microbes that maintain ecosystem functions and induce carbon and energy flow (Cottrell & Kirchman 2003), *e.g.* through predation and viral lysis [**Fig. 1.3.,** (Pedrós-Alió 2006)]. On the other hand, “occasional” species, or rare microbial types, represent many small, slow growing microbes (Fenchel & Finlay 2004), perhaps persisting only in a dormancy stage, or as a spore, representing a “seed-bank” (Finlay 2002, Pedrós-Alió 2006). This seed-bank represents the many rare taxa which may become part of the dominant core zone if appropriate environmental conditions are met. Also, these rare species may result from the easy dispersal of microorganisms, leading to high migration rates (Finlay 2002, Pedrós-Alió 2006). Accordingly, a dominant member from the core zone can be grazed down below a certain threshold, and become a rare taxon [**Fig. 1.3.,** (Pedrós-Alió 2006)]. Furthermore, differential top-down influences emerge between the "core" and "occasional" members of a community. For instance, viruses infect their hosts in a density dependent manner, following a "kill the winner" strategy (Thingstad 2000), and bacterivores mostly seek the largest (Pernthaler 2005) and most active bacteria (Kjelleberg et al. 1987). Accordingly, rare microbial types are likely less affected by viral lysis or predation and a lower loss rate is thus expected. The low extinction rate and the high migration rate of this "rare biosphere" (Sogin et al. 2006, Pedrós-Alió 2007) may thus explain the long right-handed tail of a microbial rank-abundance curve, in which most microbial diversity is contained (Curtis & Sloan 2005, Pedrós-Alió 2006).

However, this is merely an emerging field, ripe with hypotheses. Much work is yet to be done (i) to define dominant and rare microbes in the ecosystem (**Chapters 2 and 3**), (ii) to learn their influence on the structuring of the community (**Chapters 2 and 3**), and (iii) to understand whether their fluctuations are random or driven by time, space and environmental parameters (**Chapter 3**).

1.3.4 Patterns of the "Rare Biosphere"

"The present is a key to the past", early 1830's, Lyell's *Principles of Geology*

Due to methodological limitations, microbial research has until recently mainly focused on the dominant microbial types. Molecular techniques using universal primers, such as clone libraries, most easily amplify DNA from microbes with an abundance of more than 1% of the total community (Casamayor et al. 2000, Pedrós-Alió 2006). More recently, next-generation sequencing techniques, such as pyrosequencing, have allowed the first glimpse of this microbial "rare biosphere", *e.g.* in the deep ocean (Sogin et al. 2006, Huber et al. 2007, Brazelton et al. 2010) and the Arctic Ocean [(Galand et al. 2009a, Kirchman et al. 2010), **box 2**].

More than a decade ago, some marine microbial studies reintroduced the first part of Baas-Becking’s dictum saying "everything is everywhere, but, the environment selects" (Baas-Becking 1934), and supporting the ubiquity of microbes due to high dispersal, thus leading to the colonization of new habitats. This statement would then imply that microbial communities should not present any spatial differentiation (Finlay 2002, Fenchel & Finlay 2004). However, a library-based study of bacterioplankton biogeography have shown that only few abundant operational taxonomic units (OTU) are cosmopolitan, while numerous rare bacteria are endemic in the global ocean (Pommier et al. 2007). Accordingly, a pyrosequencing-based study describing the microbial community composition in different water masses of the Arctic ocean lead to similar conclusions (Galand et al. 2009a). Distinct distributions of the abundant and rare fractions of the community could be observed: the dominant microbes followed a log-normal distribution, while the rare microbes followed a log-series distribution, as already observed in the case of fishes (Magurran & Henderson 2003). Also, rare microbes were found to have a biogeography, thus implying that they do not follow a cosmopolitan distribution governed by stochastic immigration (passive dispersal), refuting what has been proposed earlier (Finlay 2002, Pedrós-Alió 2006). Marine microbes may thus exhibit low dispersal rates, possibly due to existing barriers of dispersal, and rare microbial communities may suffer from active loss (due to predation or viral lysis). The latter study indicated that the “rare biosphere” may be subjected to **selection**, **speciation** and **extinction** (Galand et al. 2009a).

Hypotheses based on the rank-abundance curve suggested that members of the seed bank may become abundant and inversely (Pedrós-Alió 2006). Subsequently, studies conducted in the Arctic ocean over one year showed that most rare organisms were always rare, even during extreme contrasts in environmental conditions [winter vs. summer, (Galand et al. 2009a, Kirchman et al. 2010)]. Indeed, 1% of the rare OTU were found to be abundant in other samples [some OTU abundant in surface waters were rare in deep waters, and inversely (Galand et al. 2009a)]. In addition, high-throughput sequencing of hydrothermal vent microbial communities, with ages of the chimneys differing by thousands of years, supported the above hypothesis proposing that rare microbial types may become abundant when appropriate conditions are occurring

(Brazelton et al. 2010). Indeed, the rare microbial types observed in young chimneys were more abundant in chimneys thousands of years older. Rare types may stay rare for a long period until environmental conditions change and outcompete the once abundant types. Some have hypothesized that rare microbial types may be preadapted to environmental conditions that already occurred in the past and may occur in the future (Brazelton et al. 2010). The second part of Baas-Becking's dictum saying " but, the environment selects" (Baas-Becking 1934), often forgotten or misused (de Wit & Bouvier 2006), may apply here, as the environment seems to be a main factor influencing microbes' fluctuations. Likewise, a **selection** process would rather occur than new **speciation** events, as dominant favorably selected traits already existed as rare types at an earlier time, before the environment changes.

The heterogeneity of these studies [*i.e.* differences in time scale (seasons vs. thousand years) and in extreme environmental conditions (cold water masses vs. hydrothermal environment)] have allowed preliminary insights into the processes shaping microbial communities, but preclude to generalize these observations to the global rare biodiversity due to the particularity of those environments. Indeed, further understanding of dominant vs. rare microbial patterns in, milder, temperate environments and a short time-period, may be of interest (**Chapter 3**). The impact of dominant and rare microbes' fluctuations on the overall community structure, along with the abiotic parameters shaping it, still need deeper insights (**Chapters 2 and 3**).

1.4 Study Site and Methods

1.4.1 Study Site

The Wadden Sea is situated in the southeastern part of the North Sea and stretches from Den Helder in the Netherlands in the southwest, to Blavands Huk in Denkmark. It is one of the largest coherent tidal flat systems of the world and covers an area of roughly 13,000 km² (van Beusekom & de Jonge 2002). Wadden Sea's tidal flats are characterized by a semi-diurnal tidal cycle (ranging up to 3.5 m) and strong seasonal changes in temperature, light availability and phytoplankton biomass (van Beusekom 2005).

In the Wadden Sea, the German Frisian island Sylt (**Fig. 1.4.**), together with the Danish island Romo, define the boundaries of List tidal basin. Its waters are also confined between two causeways connecting both islands with the mainland (Reise & Gatje 1997).

The *Hausstrand* site is a subtidal sandflat of the List tidal basin, at the eastern side of Sylt (**Fig. 1.4.**). The site is characterized by strong hydrodynamic forces (tides and wind-induced waves) with water depth ranging between 0.5 and 2.5 m. Sediments consist of well to moderately well sorted silicate sand [average grain size ~350 µm, permeability in the upper 15 cm = 1-3.10^{-11} m², (Böer 2008)].

Sediment cores were collected at the exposed *Hausstrand* site, in February, April, July and November 2005 and at the beginning and end of March 2006 (**Fig. 1.4.**). Sediment was then stored or processed for DNA extraction and environmental measurements.

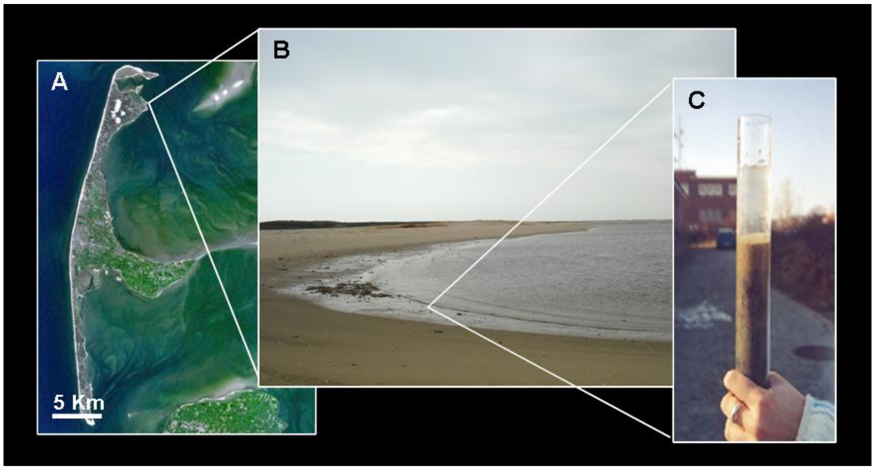


Figure 1.4. Study site. (**A**) The North Sea island Sylt, the white lines indicate the *Hausstrand* site. "Sylt." *Wikipedia, The Free Encyclopedia*, http://en.wikipedia.org/w/index.php?title=Sylt&oldid=299809341, July, 2nd 2009, 02:01 UTC, (**B**) seashore from the North Sea island Sylt, Gobet, A., (**C**) collected sediment core from the North Sea island Sylt (Böer 2008).

All chapters of this PhD thesis are based on the same set of samples, coming from Simone Böer's PhD thesis (Böer 2008) where she gives a detailed description of the study sites, sampling procedures and measurements, see also (Böer et al. 2008, Böer et al. 2009).

1.4.2 Molecular Techniques

The bacterial community structure of the North Sea island Sylt was assessed by applying the two following molecular techniques:

1.4.2.1 Automated rRNA Intergenic Spacer Analysis (ARISA)

ARISA is a time- and cost-effective technique permitting to process many samples and to still obtain robust reproducible patterns (Fisher & Triplett 1999). This fingerprinting technique allows the rapid assessment of microbial diversity and community structure by targeting the Intergenic Transcribed Spacer (ITS), highly variable in nucleotide sequence and length and located between the 16S and the 23S rRNA regions (Fisher & Triplett 1999). Notably, ARISA uses only the length heterogeneity of the ITS and gives hundreds ITS phylotypes of 400-1,200 bp per sample (Fisher & Triplett 1999).

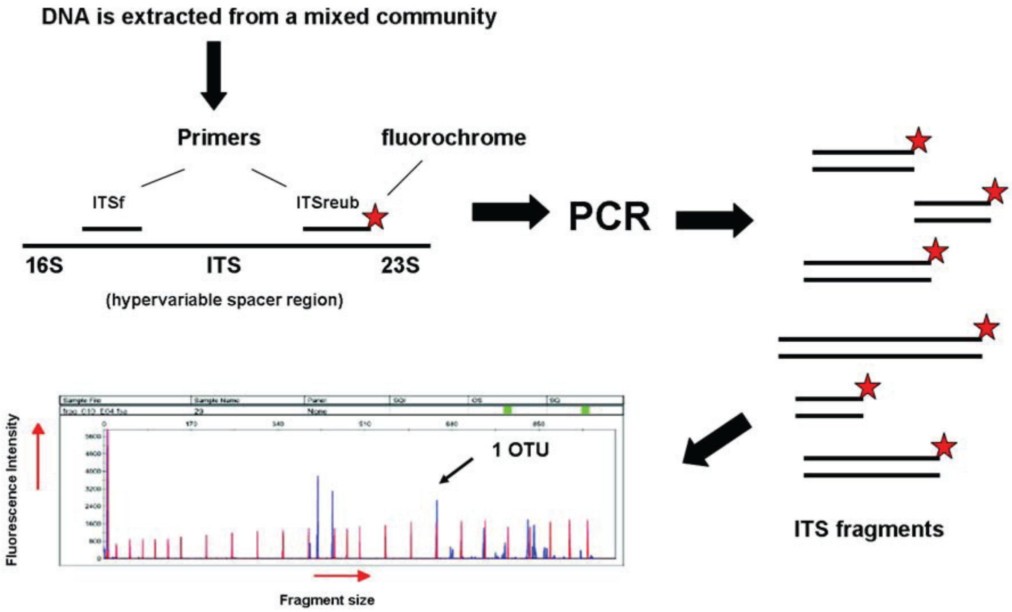


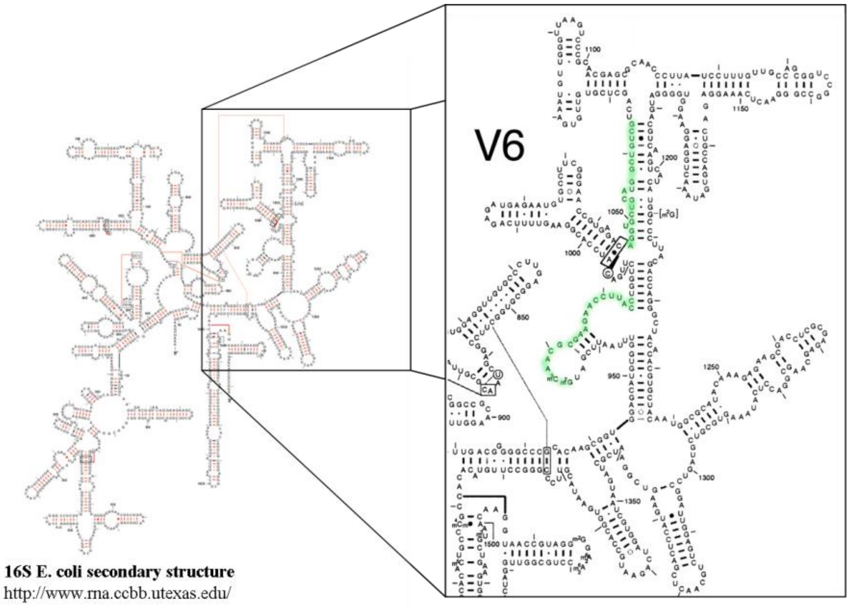
Figure. 1.5. Steps of the automated rRNA intergenic spacer analysis, (Böer 2008).

In a few words, the genomic DNA is amplified in triplicates by polymerase chain reaction (PCR) with an universal primer set, flanking the ITS region of the rRNA amplicon, one of the primers being fluorescently tagged. The amplified DNA is then cleaned and analyzed via capillary electrophoresis for detection of fluorescent DNA fragments. Fragment sizes are discriminated by comparison with a base standard and each peak of the electropherogram corresponds to one or several phylotypes of equal length [(Crosby & Criddle 2003, Yannarell & Triplett 2005), (**Fig. 1.5.**]. ARISA profiles are then analyzed and the obtained data are then formatted and binned before further ecological interpretation (Cardinale et al. 2004, Hewson & Fuhrman 2006, Böer et al. 2009). For more details on the technique and data processing, see (Böer 2008, Böer et al. 2009) and Chapter 1.

ARISA was applied during a previous PhD work (Böer 2008) and the data output was also used for the first chapter of this thesis.

1.4.2.2 454 Massively Parallel Tag Sequencing (MPTS)

454 massively parallel tag sequencing is a high-throughput technique allowing the sequencing of a significantly great amount of bases in a short time and for low cost. In 2005, Margulies and colleagues announced the sequencing of 25 million bases, with a high accuracy, in a four-hour run. The output resulted in about 100 fold increase over state-of-the-art Sanger sequencing (Margulies et al. 2005).



16S E. coli secondary structure
http://www.rna.ccbb.utexas.edu/

Figure 1.6. V6 variable region of the 16S rRNA gene. The V6 variable region of the 16S rRNA gene is situated between the highlighted parts on the figure (modified from www.rna.ccb.utexas.edu)

Creation of a V6 amplicon library. The same genomic DNA used as for ARISA was amplified to create an amplicon library based on the V6 hypervariable region of the 16S rRNA gene (**Fig. 1.6.**). Bacterial-specific primers that flank the 16S rRNA-V6 region were ligated with the 454 life sciences adapters A and B (**Fig. 1.7.A**).

454 MPTS steps. The latter primer-adapter complexes were then used to amplify the V6 rRNA region. Single stranded assemblies were annealed onto a bead presenting an immobilized primer which is complimentary to either the A or B adapter. Beads are then emulsified in droplets of a water-in-oil solution containing PCR reagents. The PCR occurs within each droplet and generates about ten million copies of the initial captured DNA template onto the bead (**Fig. 1.7.B**). After breaking the emulsion, DNA strands are denatured, and beads carrying single-stranded DNA templates are deposited into wells of a fiber-optic PicoTiter plate (**Fig. 1.7.C**). Smaller beads carrying immobilized enzymes required for a solid phase pyrophosphate sequencing reaction are also included into each well (**Fig. 1.7.D**). The pyrosequencing consists in each base (A, T, G, C) flowed at a time, and the incorporation of a base to the sequence induces the release of light (**Fig. 1.7.E**). For more details, see (Margulies et al. 2005).

The sequencing machine. The 454 sequencing instrument consists of three main parts (**Fig. 1.7.E**): a fluidic assembly (**1**), a flow cell that includes the well-containing PicoTiter plate (**2**), a CCD camera-based imaging assembly with its own fiber-optic bundle used to image the PicoTiter plate (**3**), and a computer that provides the necessary user interface and instrument control [**3**, (Rothberg & Leamon 2008)].

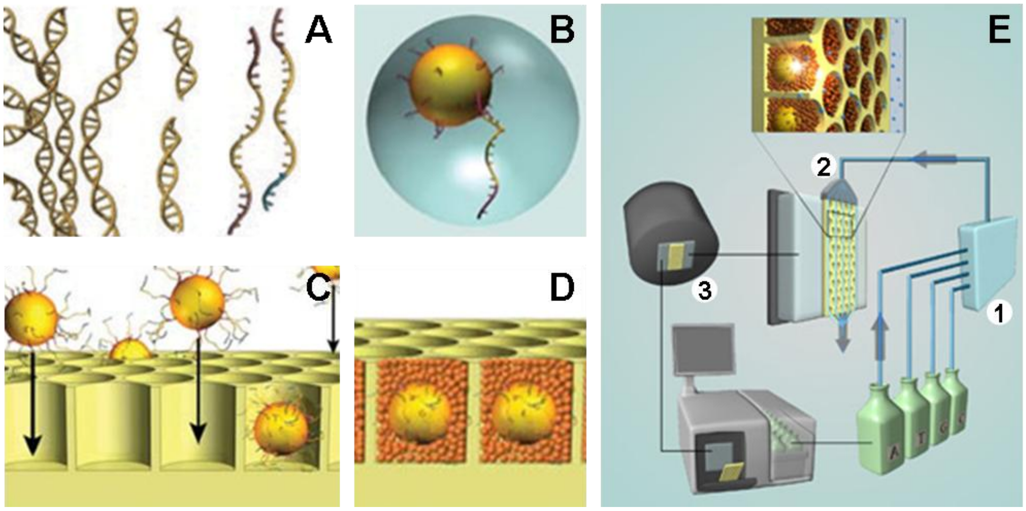


Figure 1.7. Overview of the 454 MPTS sequencing steps. (**A**) Creation of a V6 (variable region of the 16S rRNA gene) amplicon library by ligation of adapters to the V6 region, (**B**) the V6-adapters assembly is attached to a bead which is incorporated in a water-in-oil emulsion with PCR reagents to be amplified on the bead, (**C**) the bead with the amplified V6-adapters assembly is deposited into a well of a PicoTiter plate, (**D**) sequencing enzymes are deposited into each well, (**E**) each base (A, T, G, C) is flowed at a time, and the incorporation of a base to the sequence induces the release of light. Figure modified from www.roche-applied-science.com.

The genomic DNA was sent to the Josephine Bay Paul Center, at the Marine Biological Laboratory at Woods Hole, MA, USA, and was sequenced in Mitchell L. Sogin's laboratory facilities. For details about primers, or data processing, see the Visualization and Analysis of Microbial Population Structures website (http://vamps.mbl.edu/).

The data output obtained from 454 MPTS was used for the analyses of all three chapters of the thesis.

1.5 Thesis Aims and Content Overview

1.5.1 Aims

Only we are starting to unveil microbial community ecology by using high-throughput molecular techniques coupled with community ecology theories. The use of these techniques still needs some adjustments regarding the processing of large amount of data and the interpretation of such output. This PhD work seeks to improve the field of microbial community ecology by (i) comparing ARISA with 454 MPTS for an in-depth comprehension of microbial ecological patterns, (ii) making available new user-friendly statistical pipeline to analyze complex community data sets (*e.g.* 454 MPTS data sets), (iii) seeking for consistent ecological patterns at successive taxonomic levels.

These technologies offer promising advances into a better comprehension of microbial community ecology and should help us answering questions regarding (iv) the processes responsible for variations in the microbial community in space and time (*e.g.* selection, speciation), (v) fluctuations of dominant and rare microbes and their impact on the structuring of microbial communities, (vi) the characterization microbial community ecology in temperate coastal sands.

1.5.2 Thesis Content

As high-throughput molecular techniques are emerging, there is a need (i) to evaluate their accuracy and potential bias and (ii) to improve the way to analyze the colossal data output. The first part of this PhD thesis is rather technical, with a **first chapter** comparing ARISA, a traditional fingerprinting techniques in parallel with 454 MPTS, a recent next-generation sequencing techniques to study microbial communities in temperate coastal sands. Despite significant differences in community turnover (*i.e.* 50% with ARISA and 70-80% with 454 MPTS), variations in microbial community structure described by both techniques indicated similar patterns. The same combination of environmental parameters could also explain a similar amount of biological variation for ARISA data, 454 MPTS data at the phylum level and the data set containing resident OTU, present at all sampling times. Also, similar combinations of biogeochemical parameters could explain the biological variation from the latter data sets. This study validates the robustness of applying ARISA together with 454 MPTS for a high resolution in describing microbial ecology. The **second chapter** proposes a pipeline to analyze consequent data output from high-throughput sequencing techniques. It consists in a systematic truncation of proportions of rare or dominant bacterial types from large bacterial community data sets to test for the effect of the truncation on the resulting ecological interpretation. About 40% of the rare bacterial types could be removed from the original data set and a similar ecological signal was still obtained.

The second part of the thesis describes the ecology of dominant and rare types of the microbial community in temperate coastal sands. The **second chapter** already gives insights about the impact of removing either abundant or rare bacterial types on microbial ecological patterns. The **third chapter** gives a more thorough interpretation of each fraction of the bacterial community. Dominant and resident (*i.e.* present at all times) types presented similar ecological patterns as that of the overall community while rare types' patterns were different. Actually, rare types' fluctuations were driving the really high turnover of the microbial community through depth and time. Notably, rare microbial types were not randomly fluctuating as biogeochemical parameters could explain their variation.

1.5.3 Outline

Chapter I: Comparison of the consistency of molecular fingerprinting techniques for the description of microbial ecological patterns

Authors: Angélique Gobet, Antje Boetius and Alban Ramette.

In preparation for Applied and Environmental Microbiology.

Personal contribution: designed study together with Alban Ramette and Antje Boetius, conducted all analyses on available data sets, and wrote the first draft of the manuscript.

Brief overview: Comparison of two high-throughput molecular techniques, ARISA and 454 MPTS to interpret ecological patterns in bacterial communities and to give recommendation as for the suitability of each method for specific applications.

Chapter II: Multivariate Cutoff Level Analysis (MultiCoLA) of Large Community Datasets

Authors: Angélique Gobet, Christopher Quince and Alban Ramette.

Published in Nucleic Acids Research in June 2010.

Personal contribution: designed study together with Alban Ramette, wrote the R programs and conducted all data analyses except a subset by Christopher Quince, wrote the first draft of the manuscript, and finalized the submitted version with Alban Ramette and Christopher Quince.

Brief overview: Exploration of the effects of removing successive proportions of rare bacterial types from large bacterial community data sets on the resulting ecological interpretation.

Chapter III: Diversity and dynamics of the rare and resident bacterial biosphere in coastal sands

Authors: Angélique Gobet, Simone I. Böer, Susan M. Huse, Justus E.E. van Beusekom, Christopher Quince, Mitchell L. Sogin, Antje Boetius and Alban Ramette.

Submission to Proceedings of the National Academy of Science planned by December, 5th 2010.

Personal contribution: designed study with Alban Ramette, Simone I. Böer, Justus E.E. van Beusekom and Antje Boetius, conducted most of the data analyses except two smaller subsets provided by Alban Ramette and Christopher Quince, and wrote the core manuscript with Antje Boetius and Alban Ramette. The 454 MPTS data were provided by Susan M. Huse, Mitchell L. Sogin, all authors commented on the final version of the manuscript.

Brief overview: High resolution description of the bacterial diversity patterns in coastal sandy habitats through time and sediment depth by applying 454 MPTS. This study highlighted the very high turnover of the bacterial community and then deciphered the main drivers of such high community dynamics.

Fundings

The thesis was supported by the Marie Curie Early Stage Training fellowship Site in Marine Microbiology [MarMic EST contract MEST-CT-2004-007776], the International Max Planck Research School of Marine Microbiology, and by the Max Planck Society. The sequencing was financed by a Keck Foundation grant to the International Census of Marine Microbes (ICoMM), project AB_SAND and FIS to AR and AB.

2 Chapter I.

Comparison of the consistency of molecular fingerprinting techniques for the description of microbial ecological patterns

Angélique Gobet, Antje Boetius and Alban Ramette

In preparation for Applied and Environmental Microbiology.

Keywords: Bacterial diversity, ARISA, fingerprinting, 454 pyrosequencing, multivariate analyses, sand, coastal environment

Abstract. As fast method to describe microbial community structures, cost-effective fingerprinting techniques such as automated rRNA intergenic spacer analysis (ARISA), are nowadays often preferred to labor-intensive molecular techniques such as 16S rRNA library-based approaches. Since high-throughput sequencing is becoming more available and enables a high resolution description of the taxonomic composition of microbial communities, information based on traditional fingerprinting techniques may now appear limited. Yet, it may be more convenient to process data output from fingerprinting techniques as compared with the complex data flood derived from high-throughput techniques. Indeed, each approach may be well suited for a specific type of question. We thus compared bacterial community structure in coastal sands as obtained by ARISA and 454 massively parallel tag sequencing (MPTS), at several levels of taxonomic resolution or data set truncation, so as to account for the effects of the number of taxa considered by the two approaches. Despite revealing different microbial community turnover between sediment depth layers or sampling times (*i.e.* 50% with ARISA and 70-80% with 454 MPTS), variations in community structure were similar with both approaches. Also, similar combinations of biogeochemical parameters could explain the biological variation from ARISA and broad taxonomic levels from 454 MPTS, and this probably reflected the ecology of main microbial players in the ecosystem. This study first confirms ARISA as a valid technique to describe microbial community patterns and further suggests combining community fingerprinting and high-throughput sequencing techniques to obtain both a broad and quick overview of diversity patterns in many samples and a detailed description of community composition and dynamics for specific samples.

2.1 Introduction

Microbes represent the largest pool of global biomass and contribute to ecosystem processes by recycling carbon and nutrients (Whitman et al. 1998, Azam & Worden 2004). Deciphering the complexity of the microbial world is thus essential to understand structure and function of different habitats and ecosystems, despite the large range of stochastic and deterministic environmental conditions that may shape microbial communities (Curtis & Sloan 2004). Whereas traditional sequence-library based approaches are time consuming to process the many samples required for a good description of microbial communities, molecular fingerprinting techniques [*e.g.* denaturing gradient gel electrophoresis [DGGE, (Muyzer et al. 1993)], terminal restriction fragment length polymorphism [T-RFLP, (Avaniss-Aghajani et al. 1994)], automated rRNA intergenic spacer analysis [ARISA, (Fisher & Triplett 1999)]] allow a rapid processing of many samples with consistent reproducible patterns (Fisher & Triplett 1999). ARISA targets the Intergenic Transcribed Spacer (ITS) between the 16S and the 23S rRNA regions (Fisher & Triplett 1999), highly variable in nucleotide sequence and length [*e.g.* from 60 bp to 1529 bp (Gürtler & Stanisich 1996)], and gives hundreds ITS phylotypes of 400-1,200 bp per sample (Fisher & Triplett 1999). ARISA renders fluorescence profiles where each peak corresponds to one or several phylotypes of equal length (Crosby & Criddle 2003, Yannarell & Triplett 2005). Consequently, although suitable to study community changes, this approach does not allow the assessment of the number of species in a given sample (Bent & Forney 2008), as well as their taxonomy (Fisher & Triplett 1999, Brown et al. 2005).

Although traditional sequence-library based techniques have already described a large fraction of microbial diversity, the major part of it escapes our sampling efforts and even large 16S rRNA clone libraries highly underestimate microbial diversity (Curtis & Sloan 2005, Quince et al. 2008). For instance, Sanger sequencing on coastal waters retrieved 516 unique OTU while the estimated richness reached 1,633 OTU (Acinas et al. 2004). The advent of high-throughput sequencing techniques has revolutionized the microbial ecology field by giving a high resolution description of microbial diversity. Indeed, 454 massively parallel tag sequencing (MPTS) gives thousands to tens of thousands short variable regions of the 16S rRNA [~60 bp, (Sogin et al. 2006)] per sample, which can be further taxonomically annotated. Using this method 4,000 OTU were obtained from a deep sea sample with a total estimated diversity of 11,296 OTU (Sogin et al. 2006). Despite offering a deeper coverage of microbial diversity, 454 MPTS data output has to be analyzed with care due to the presence of PCR and sequencing artifacts such as chimera and homopolymer formation, which inflate microbial diversity estimates (Kunin et al. 2010). Consequently, several studies have provided various ways to trim and correct sequences (Quince et al. 2009, Gobet et al. 2010, Kunin et al. 2010).

High-throughput sequencing techniques are increasingly used to complement methods describing microbial communities with lower resolution such as Sanger sequencing or fingerprinting techniques (*e.g.* DGGE, T-RFLP, ARISA). Most of these studies have shown the efficiency of 454 MPTS for giving a higher resolution of the microbial community description than the latter techniques. For example, 454 MPTS sequences were included in a Sanger sequencing-based phylogenetic tree and revealed greater depth coverage to potentially describe new taxonomic groups (Gillevet et al. 2009, Roh et al. 2010). Other studies indicated an estimated richness from pyrosequenced OTU much higher than that from T-RFLP and ARISA-ITS fragments (Roesch et al. 2009, Koopman et al. 2010). Also, a study showed that T-RFLP data did not follow similar cyclic or seasonal microbial community patterns as with 454 MPTS (Gilbert et al. 2009). Overall, these studies provide evidence of the high resolution of 454 MPTS for describing microbial diversity, whereas the consistency to describe microbial ecological patterns as with traditional molecular approaches remains unknown.

This study resorts to 454 MPTS and ARISA data from bacteria communities in coastal North Sea sands. ARISA analyses indicated that variations in bacterial community composition were strongly related to vertical changes in biogeochemical gradients. Within a period of two years, the turnover of the microbial community based on ARISA data was about ~50% (Böer et al. 2009) while a selection of pyrosequenced DNA templates from the latter study revealed even higher turnover (~70-80% new OTU) of the bacterial community (Gobet et al. *Submitted*). Additional multivariate analyses indicated the importance of vertical and temporal variations in biogeochemical gradients on the structuring of the bacterial community. These two approaches already gave deep insights into the description of microbial ecology in temperate coastal sands. However, these two molecular techniques target different parts of the rRNA gene and offer a different level of resolution of the microbial community description. The ecological interpretations from both ARISA and 454 MTPS applied on temperate coastal sands were thus compared to assess whether similar conclusions could be obtained or whether each technique may be better suited to address specific microbial ecological questions. The application of multivariate analyses on the ARISA data set, and of successive taxonomic levels and corrections of the 454 MPTS data set [*i.e.* using PyroNoise (Quince et al. 2009) and MultiCoLA (Gobet et al. 2010)] allows for the comparison of diversity patterns and of the effects of the environment on the structuring of bacterial communities from different angles of the data set.

2.2 Materials and Methods

Sampling procedures and contextual parameters.

In February, April, July and November 2005, beginning and end of March 2006, sediment push cores were collected at low tide on the shallow subtidal sandy area of the island Sylt (55° 00’47.7’’N, 8° 25’59.3’’E, North Sea, Germany). Cores were sectioned every 5 cm down to 15 cm and the sections were directly processed (*e.g.* measurements of extracellular enzymatic activities) or stored at -4°C and -20°C until further analyses [DNA extraction or measurements of nutrients, pigments, carbohydrates, bacterial cell counts; (Böer et al. 2009, Gobet et al. *Submitted*)]. Other environmental parameters (wind speed and water column data: chlorophyll a, pH and water temperature) were added to the data set as described earlier (Böer et al. 2009, Gobet et al. *Submitted*).

Community structure analysis.

DNA extraction. DNA from the 16 sandy samples was extracted and purified using an UltraClean Soil DNA Isolation Kit (MoBio Laboratories Inc. Carlsbad, CA) following the manufacturer’s protocol, as described earlier (Böer et al. 2009, Gobet et al. *Submitted*). The same DNA templates were used to analyze the bacterial community structure samples by automated rRNA intergenic spacer analysis (Fisher & Triplett 1999) and 454 massively parallel tag sequencing (MPTS).

Automated rRNA intergenic spacer analysis. Extracted DNA was amplified in triplicates using bacteria-specific primers and normalized DNA quantities of 25 ng per reaction. The resulting amplified fragments were purified with Sephadex G-50 Superfine (Sigma Aldrich, Munich, Germany) and identified by capillary electrophoresis on an ABI PRISM 3130*xl* Genetic Analyzer (Applied Biosystems). For details of the ARISA protocol, see Böer et al. (Böer et al. 2009). The obtained ARISA profiles were analyzed using the GeneMapper Software v 3.7 (Applied Biosystems, Carlsbad, CA, USA) and further data formatting and binning were done as described elsewhere (Cardinale et al. 2004, Hewson & Fuhrman 2006, Böer et al. 2009).

454 massively parallel tag sequencing. The V6 region of the bacterial 16S rRNA gene was amplified by using a mixture of five forward (967F) and four reverse (1046R) primer sets including 454 Life Science's A or B sequencing adapter (Huber et al. 2007). The V6 fragments were pyrosequenced on a Genome Sequencer 20 system (Roche, Basel, Switzerland) at 454 Life Sciences (Branford, CT) by primer extension (Margulies et al. 2005). Sequences were first trimmed and corrected (Huse et al. 2007) and then, annotated by an automatic annotation pipeline using several known databases (Entrez Genome, RDP, SILVA), following the approach of Sogin et al. (Sogin et al. 2006). All 454 MPTS sequences are publicly available at the Visualization and Analysis of Microbial Populations Structure (VAMPS) website (http://vamps.mbl.edu/).

Data analyses.

Data sets. In this study, analyses were performed by defining OTU (Operational Taxonomic Units) either as ITS phylotype [hereafter, OTU_{ARISA} correspond to binned ARISA peaks (Böer et al. 2009)], or as unique 454 MPTS sequences (hereafter, two sequences are considered as two different OTU_{unique} when they differ by at least one base pair). For the 454 MPTS data sets, the following subsets were considered: 1) all un-annotated sequences that we referred to as OTU_{all}, 2) the fully annotated sequences (*i.e.* from phylum to genus levels and the corresponding $OTU_{annotated}$ level, each data set representing 20% of the original OTU_{all} data set), 3) the PyroNoise-corrected data clustered at different percentages of sequence dissimilarity (0%, 3%, 5% and 10% sequence dissimilarity) 4) the MultiCoLA-truncated data sets, consisting in the original OTU_{all} data set without successive proportions of low occurring OTU_{unique}; *i.e.* OTU_{unique} with lowest number of sequences than a given threshold are removed (Gobet et al. 2010).

Variations in bacterial community structure and ecological patterns. OTU numbers were compared by pairwise Student t-tests. The amount of shared OTU between either two sampling dates or two depth layers was calculated for all community matrices [ARISA data, OTU_{all}, the fully annotated sequences, the PyroNoise-corrected and the truncated data sets; (Gobet et al. *Submitted*)].

Pairwise distance matrices were calculated from the relative abundance data (ARISA and 454 MPTS data sets) using the Bray-Curtis dissimilarity index (Bray & Curtis 1957). The resulting dissimilarity matrices were compared to one another using Mantel's test with Pearson's product moment correlation coefficient (Pearson 1901). For the comparison of dissimilarity matrices from MultiCoLA-truncated data sets, simple Pearson's correlations were calculated without Mantel's test. In this case, the significance of each pairwise comparison (*i.e.* each Pearson's correlation) cannot be calculated as the matrices depend on one another and testing correlations would only make sense in case of data set independence (Legendre & Legendre 1998, Legendre et al. 2005).

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

Relationships between the structuring of the microbial community and the environment. In a previous related study (Gobet et al. *Submitted*), multivariate regression approaches (Legendre & Legendre 1998) were applied to test the relationships between the variation of time (with sampling dates set as ranks), depth and measured environmental parameters (pH, water temperature, wind speed, salinity, pigments, nutrients, extra-cellular enzymatic activities and cell properties). As time and depth were found to significantly covary with most biogeochemical factors, they were discarded from the environmental data set for further analyses (Gobet et al. *Submitted*). Some explanatory variables of the remaining data set (pigments, nutrients, extra-cellular enzymatic activities and cell properties) were $\log_{10}$-transformed before describing the microbial community distribution in the relative abundance matrices (ARISA data, OTU_{all}, the fully annotated sequences, the PyroNoise-corrected data and the MultiCoLA-trimmed data). As the inclusion of all environmental parameters for explaining the community variation may distort the ecological interpretation (multicollinearity of the environmental variables), a forward selection [based on 999 Monte Carlo permutation tests and Akaike Information Criterion (AIC)] of the environmental parameters was applied. Consequently, we obtained the best-fitting models that could significantly explain the variation in the Hellinger-transformed (Legendre & Legendre 1998, Legendre & Gallagher 2001) community tables. The effect of pure environmental variables (pigments, nutrients, extra-cellular enzymatic activities, cell abundance) selected previously and their covariation on microbial community structure was then tested by canonical variation partitioning.

The forward selection was performed with the software package CANOCO for Windows 4.5 (ter Braak & Šmilauer 2002). Other statistical analyses were carried out using the R statistical environment [R version 2.10.0 (R_Development_Core_Team 2009)], using the *vegan* package (Oksanen et al. 2009) and custom R scripts.

2.3 Results & Discussion

2.3.1 Alpha-diversity of the bacterial community in temperate coastal sands as described by the data output from ARISA and 454 MPTS.

A total of 16 coastal sandy samples were analyzed in parallel by ARISA (Böer et al. 2009) and 454 MPTS (Gobet et al. *Submitted*). From the application of ARISA, 306 different $\mathrm{OTU}_{\mathrm{ARISA}}$ were detected in the whole data set, with 100-202 $\mathrm{OTU}_{\mathrm{ARISA}}$ per sample (in 5 cm sediment layer at a given sampling time). We obtained 88 $\mathrm{OTU}_{\mathrm{ARISA}}$ (29% of the total number of $\mathrm{OTU}_{\mathrm{ARISA}}$ in the whole data set) that were resident, *i.e.* present at all sampling times, while 15% $\mathrm{OTU}_{\mathrm{ARISA}}$ were present only once in the data set (**Table S2.1.**). The application of 454 MPTS on the same extracted DNA generated 197,684 sequences in total, which correspond to 27,630 $\mathrm{OTU}_{\mathrm{unique}}$ in the $\mathrm{OTU}_{\mathrm{all}}$ data set, with a range of 1,042 to 5,577 $\mathrm{OTU}_{\mathrm{unique}}$ identified per sample (Gobet et al. *Submitted*). About 5% of the total number of $\mathrm{OTU}_{\mathrm{unique}}$ were resident while 74.7% (20,640 $\mathrm{OTU}_{\mathrm{unique}}$) were present only once in the $\mathrm{OTU}_{\mathrm{all}}$ data set (Gobet et al. *Submitted*). These observations first indicate the large difference in the amount of data output obtained by the two molecular methods.

As high-throughput techniques potentially induce biases due to pyrosequencing errors, some solutions were proposed such as correcting by pre-clustering electrophoregram output (Quince et al. 2009) or by clustering the tag sequences (Kunin et al. 2010). We first studied the distribution of OTU numbers over time and through depth for each approach and after correction of the 454 MPTS data set (PyroNoise at 3% sequence dissimilarity, [$\mathrm{OTU}_{3\%}$], **Fig. S2.1.**). It seemed that the number of OTU increased with sediment depth in all cases (**Fig. S2.1. A-C**), as previously observed when more samples were considered in the previous ARISA study (Böer et al. 2009) and in other coastal sediments by using T-RFLP (Urakawa et al. 2000). This was confirmed by significant differences in OTU number between depths (Student t-tests, $P < 0.05$). The top 5 cm layer was clearly different from the deeper layer with both molecular techniques. Notably, the mid 5-10 cm layer behaved differently according to each technique (**Figs. S2.1. A-C**) and these differences might be due to the different levels of resolution of each technique. For instance, some of the $\text{OTU}_{\text{ARISA}}$ may represent several bacterial types (Crosby & Criddle 2003) and ARISA data output may not reflect rare bacterial types (Bent & Forney 2008), especially depending on the way the data output was processed. Here, the ARISA data output considered only fragments above a threshold of 50 fluorescence units and between 100–1000 bp length (Böer et al. 2009). Also, both techniques induced no significant changes in OTU number through time (**Fig. S2.1. D-F**), as shown in the previous study where OTU were temporally stable between April 2005 and March 2006 (Böer et al. 2009).

2.3.2 Comparison of beta-diversity patterns obtained from ARISA and from different resolution levels of the 454 MPTS output.

Microbial community turnover. Microbial communities can be described according to various levels of resolution; either by using different techniques (here, ARISA or 454 MPTS) or by transforming the output (*i.e.* taxonomic levels, PyroNoise correction, rare $\text{OTU}_{\text{unique}}$ removal). The resulting data sets were compared with each other to see whether similar ecological information could be obtained from each of them. Depth-related and temporal microbial community turnovers from ARISA and 454 MPTS were mostly similar. Indeed, the previous ARISA study indicated about 50% shared $\text{OTU}_{\text{ARISA}}$ in the sand over two years (Böer et al. 2009) while the previous 454 MPTS study showed 20-30% shared $\text{OTU}_{\text{unique}}$ in the bacterial community between two depth layers or any two sampling dates [(Gobet et al. *Submitted*) and **Figs. 2.1., S2.2.**]. When analyzing the turnover of the bacterial community on 16 ARISA-samples from the previous study (Böer et al. 2009), the turnover of the bacterial community was much lower, with 66-78% shared $\text{OTU}_{\text{ARISA}}$ between two depth layers and 70-91% shared $\text{OTU}_{\text{ARISA}}$ between sampling times.

As $\text{OTU}_{\text{unique}}$ were annotated by using the GAST taxonomic pipeline (Sogin et al. 2006), it was interesting to explore the amount of shared $\text{OTU}_{\text{unique}}$ at successive taxonomic levels. Indeed, when performing the analyses from the genus to the phylum levels, there were 63% to 97% shared $\text{OTU}_{\text{unique}}$ over sediment depth, respectively (**Figs. 2.1., S2.3.**). Interestingly, a relatively constant proportion of turnover with time was also observed at all taxonomic levels investigated. The ratio of shared taxa increased from OTU_{all} to phylum levels from 18 to 100% between two sampling times, respectively (**Figs. 2.1., S2.2.-S2.3.**). The $\text{OTU}_{\text{annotated}}$ data set, consisting only of sequences with a full taxonomic annotation from genus to phylum, showed similar patterns as the OTU_{all} data set (**Fig. S2.3.**), suggesting that ecologically meaningful patterns can still be obtained from the taxonomically identified fraction of the community.

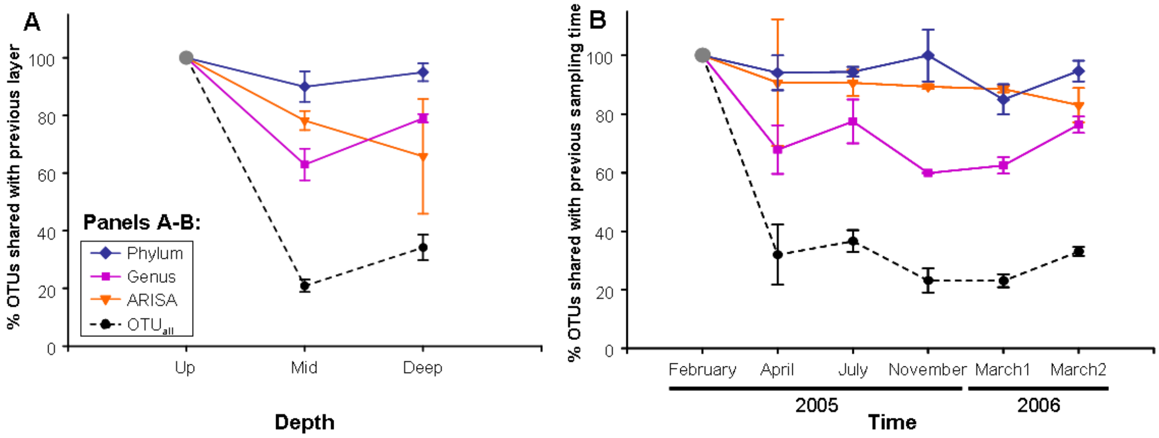


Figure 2.1. Turnover of the bacterial community between two consecutive (A) depth layers or (B) sampling times. The percentage of OTU shared between two successive sampling depth layers (or sampling dates) was calculated. The turnover of the community was compared between different 454 MPTS data sets (at the Phylum, Genus and OTU_{all} levels) and the ARISA data set. Bars correspond to standard deviation calculated (**A**) over 4-6 sampling dates and (**B**) over three depth layers, except for July and November 2005 where 2 depth layers were considered. The first depth layer and sampling date (February 2005) are indicated by the grey point as 100% of common OTU. OTU_{all} represents the original data set with all OTU, used here as a reference to study the effects of the taxonomic classification of OTU on the interpretation of the dynamics of the bacterial community.

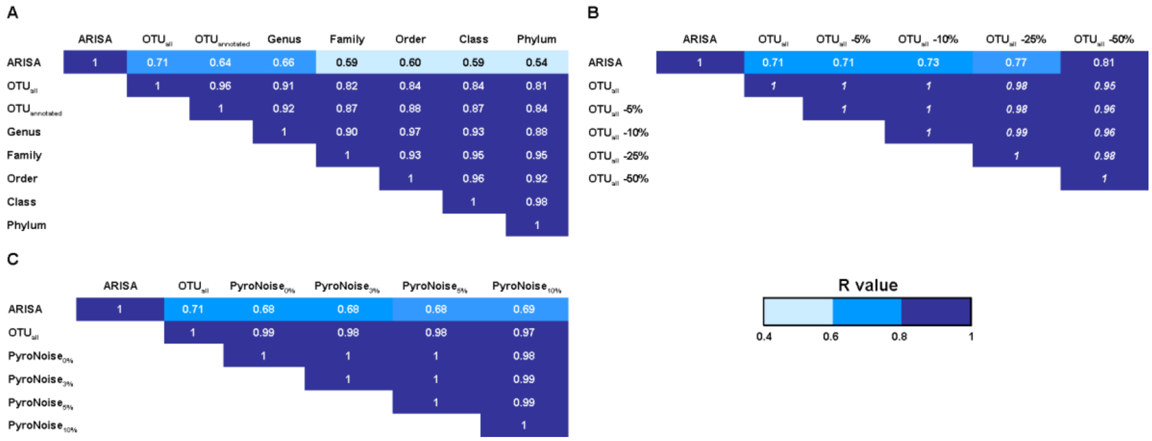
However, the observed high turnover at the OTU_{all} level might overestimate real dynamics of the bacterial community as it might result from pyrosequencing artifacts (Quinlan et al. 2008). Even though we applied the PyroNoise algorithm (Quince et al. 2009) and different levels of clustering, only about 20-40% of shared OTU in the bacterial community could still be observed over depth or time (**Fig. S2.2.**). If we consider that 3% sequence similarity thresholds for OTU correspond roughly to cutoff levels defining bacterial species level (Schloss & Handelsman 2005), patterns observed with PyroNoise-corrected data seem to present a continuum with the taxonomic resolution up to the genus level (**Fig. S2.2.-2.3.**). Hence, the observed large community turnover may not be due to technical biases and may reflect real dynamics of the bacterial community in marine sandy sediments.

As a large part of pyrosequencing data sets represent low abundant OTU_{unique} (Gobet et al. 2010), we also studied the effect of removing successive proportions of rare OTU_{unique} on the turnover of the microbial community [**Fig. S2.4.**, (Gobet et al. *Submitted*)]. As observed previously, removal of rare OTU_{unique} leads to a decrease in community turnover, and would be supported by the low fraction of abundant types in the community (Gobet et al. *Submitted*). Interestingly, when comparing the turnover at successive taxonomic levels or percentages of rare OTU_{unique} removed, some patterns were found to be similar (**Figs. S2.3.-S2.4.**). For instance, it seemed that the turnover after removal of 15% rare OTU_{unique} corresponds to the genus level or that the removal of 30% rare OTU_{unique} would lead to a similar turnover as the class or phylum levels (**Figs. S2.3.-S2.4.**). This may be explained by the loss of community resolution at broader taxonomic levels where the patterns of many different types are lumped together and by the fact that most rare OTU_{unique} were not identified from the genus to the phylum level. Also, this indicates the high consistency of the ecological information obtained at various taxonomic levels, as also observed in global benthic and pelagic marine realms (Zinger et al. *In preparation*).

Interestingly, the amount of shared OTU_{ARISA} over depth or through time seemed to reveal similar information for the description of the microbial community turnover as obtained at the family level or when removing 20-25% of low abundant OTU_{unique} in the OTU_{all} data set (**Figs. 2.1., S2.2.-S2.4.**). These observations confirm the consistency of the community turnover observed with ARISA and at successive taxonomic and corrected levels of the 454 MPTS data set. This also highlights the importance of the degree of resolution to describe bacterial community patterns.

Structure of the data sets and resulting microbial ecological interpretation. We thus evaluated changes in data structure after creating sample-by-sample dissimilarity matrices from the relative abundances (*i.e.* by using the Bray-Curtis dissimilarity index to calculate the dissimilarity between samples) and those matrices were then compared with each other by using Pearson's correlations (**Fig. 2.2.**).

Overall, the comparison of the community structure between the relative abundance-data sets showed little variation at different taxonomic levels, after correcting for pyrosequencing noise, or truncating the data sets (**Fig. 2.2.**). This indicated how the main community patterns stayed consistent regardless of the chosen level of resolution. The ARISA data set structure was slightly more different than most of the other data sets, but was quite similar to data sets truncated from 25-50% of their rarer OTU_{unique} (**Fig. 2.2.**).



A

	ARISA	OTU_{all}	$OTU_{annotated}$	Genus	Family	Order	Class	Phylum
ARISA	1	0.71	0.64	0.66	0.59	0.60	0.59	0.54
OTU_{all}		1	0.96	0.91	0.82	0.84	0.84	0.81
$OTU_{annotated}$			1	0.92	0.87	0.88	0.87	0.84
Genus				1	0.90	0.97	0.93	0.88
Family					1	0.93	0.95	0.95
Order						1	0.96	0.92
Class							1	0.98
Phylum								1

B

	ARISA	OTU_{all}	OTU_{all} -5%	OTU_{all} -10%	OTU_{all} -25%	OTU_{all} -50%
ARISA	1	0.71	0.71	0.73	0.77	0.81
OTU_{all}		*1*	*1*	*1*	*0.98*	*0.95*
OTU_{all} -5%			*1*	*1*	*0.98*	*0.96*
OTU_{all} -10%				*1*	*0.99*	*0.96*
OTU_{all} -25%					*1*	*0.98*
OTU_{all} -50%						*1*

C

	ARISA	OTU_{all}	$PyroNoise_{0\%}$	$PyroNoise_{3\%}$	$PyroNoise_{5\%}$	$PyroNoise_{10\%}$
ARISA	1	0.71	0.68	0.68	0.68	0.69
OTU_{all}		1	0.99	0.98	0.98	0.97
$PyroNoise_{0\%}$			1	1	1	0.98
$PyroNoise_{3\%}$				1	1	0.99
$PyroNoise_{5\%}$					1	0.99
$PyroNoise_{10\%}$						1

Figure 2.2. Comparison of the structure of modified data sets. Pearson's correlation coefficient was used to compare the ARISA and the OTU_{all} data sets with (**A**) various levels of taxonomic annotation, (**B**) successive removal of the rare OTU and (**C**) successive clustering of the PyroNoise-corrected data to define OTU. The correlation coefficient was calculated from the distance matrices resulting from the relative sequence abundances. Significances of the correlation were tested using Mantel tests. For (**B**) values *in italic* indicate simple Pearson correlations of the truncated matrices, but without a test of significance as the truncated matrices are not statistically independent from each other (see Materials and Methods). Only significant values (after Bonferroni correction) are shown.

When the amount of extracted ecological variation was analyzed by non-metric multidimensional scaling (NMDS), similar depth-related patterns of the microbial community were obtained, regardless of the chosen taxonomic or correction levels (**Fig. 2.3.**). These observations were confirmed after testing for differences between sampling depth layers by analysis of similarities (**Fig. 2.3.**). When comparing the obtained NMDS ordinations by Procrustes rotation (*i.e.* a measure of the correlation between two ordination solutions), a similar picture emerged. For instance, the comparison of NMDS axes from ARISA with those from OTU_{all} data set or PyroNoise-corrected data set after

3% clustering were highly similar, with a R value reaching 0.88 and 0.74, respectively (**Fig. S2.5.**). Together with the turnover results, all these observations indicate that our ARISA data may represent the most dominant types in the community (*e.g.* without 25% of rare types, **Fig. S2.5.**). Indeed, a Mantel test between ARISA data and resident OTU_{unique} (OTU_{unique} present at all times) reached 0.43 and a Procrustes rotation between the resulting NMDS axes reached 0.79, indicating significant similar patterns, even after Bonferroni correction. ARISA could thus mimic the effect of the resident types which may shape the microbial community structure, and play a major role in the main functions of the ecosystem (Gobet et al. *Submitted*). These similar microbial ecological patterns obtained confirm the reproducibility and the consistency of both molecular techniques.

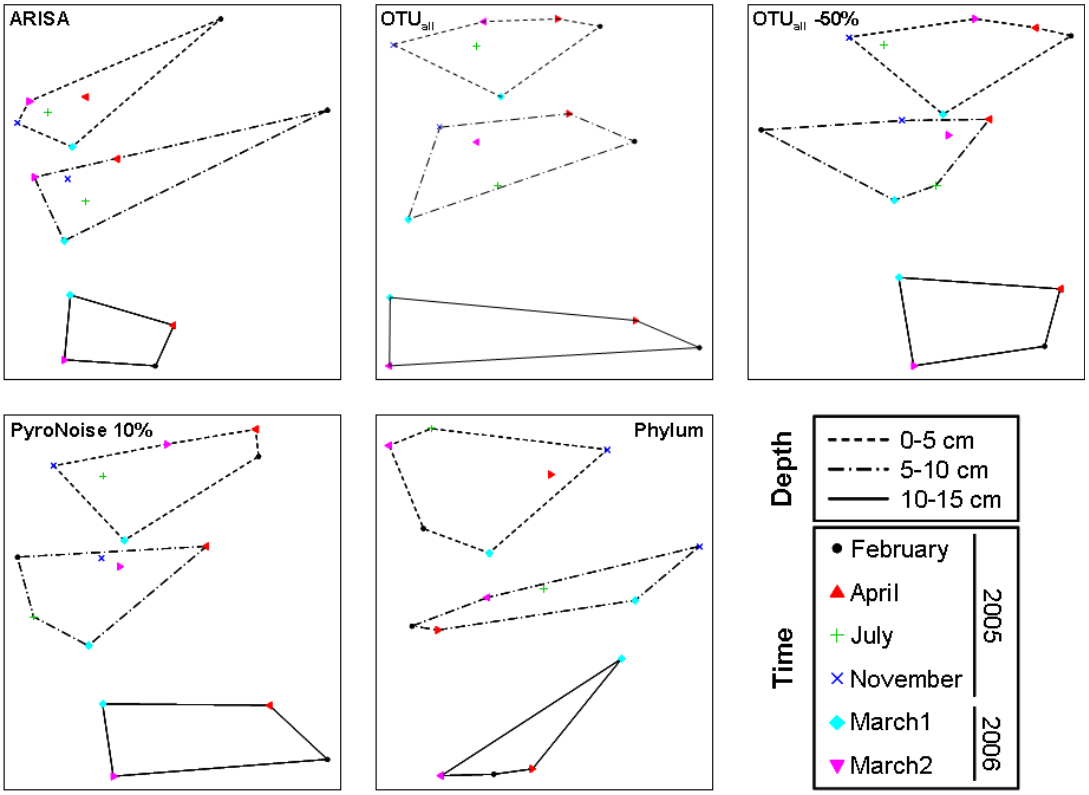


Figure 2.3. Examples and comparison of extracted variation from the ARISA and 454 MPTS data sets. Non-metric multidimensional scaling (NMDS) ordination (Bray-Curtis distance matrix) of the relative abundance data sets from ARISA (stress = 6.2%), the original OTU_{all} data set (stress = 8.1%), the OTU_{all} data set with 50% of rare OTU removed (stress = 9.5%), the PyroNoise-corrected data at 10% sequence dissimilarities (stress = 8.3%) and, at the Phylum level (stress = 4.7%). Analyses of similarities (ANOSIM) indicated significant differences between samples grouped per sediment depth ($R > 0.3$, P value ≤ 0.01).

2.3.3 Ecological modeling of beta-diversity patterns.

Microbial ecology in temperate coastal sediments from the North Sea island Sylt has already been studied using several types of molecular methods. 16S rRNA-based libraries and fluorescence *in situ* hybridization allowed the description of the main bacterial groups present in the sand (Musat et al. 2006). The application of ARISA allowed further details about depth-related and ecological patterns (Böer et al. 2009) and 454 MPTS permitted a high resolution description of the fluctuations of rare and resident OTU in the sand (Gobet et al. *Submitted*). By studying relationships of the bacterial community and the surrounding environment, these studies allowed a deep comprehension of bacterial ecology in temperate coastal sands.

We applied a multivariate variation partitioning approach to analyze the impact of time, sediment depth, cell abundance and biogeochemical gradients (*i.e.* pigments, nutrients and extra-cellular enzymes), and of their combined effects on the structure of the bacterial community (Legendre & Gallagher 2001, Ramette & Tiedje 2007b). In order to avoid collinearity in the analyses, the factors time and depth, which were significantly covarying with most environmental factors, were removed from the data set (Gobet et al. *Submitted*). The remaining factors, cell abundance and biogeochemical gradients, and their covariation were then linked to the variations in community structure in the ARISA and the 454 MPTS data sets. Interestingly, the same combinations of significant biogeochemical variables could explain data sets that had a similar degree of resolution. For instance, a model containing salinity, pigments, the same nutrients and extra-cellular enzymes as well as cell abundance, could explain 51-75% of the biological variation from the genus to the phylum level in sandy sediments (**Fig. 2.4.**, **Table S2.2.**). Also, almost the same environmental model as for the taxonomic annotated data set could explain the biological variation in the ARISA data set (**Table S2.2.**).

For more complex data sets (*i.e.* more levels of variation are present), a similar environmental model (*i.e.* chlorophyll *a*, extra-cellular phosphatase activity, cell abundance, **Table S2.2.**) could explain 14-20% of biological variation in the $\mathrm{OTU}_{\mathrm{annotated}}$, the raw $\mathrm{OTU}_{\mathrm{all}}$ and after PyroNoise-correction and clustering at 0% and 3% sequence dissimilarity (**Fig. 2.4.**). Notably, the same environmental factors could explain a similar amount of the microbial community variation for the OTU$_{\text{all}}$ and the OTU$_{\text{annotated}}$ data sets. This observation follows our previous assumption that the annotated subset of the data set consists of an unbiased subset of the whole data set, and that the taxonomically identified OTU may be used to describe overall patterns in microbial communities in temperate coastal sandy sediments. Also, when 1-5% rare OTU$_{\text{unique}}$ were removed from the OTU$_{\text{all}}$ data set, similar models were obtained as with the original OTU$_{\text{all}}$ data set (18-20% explained variation, **Fig. S6**, **Table S2.3.**). Interestingly, some truncated data sets followed similar environmental models as that of data sets defined at specific taxonomic levels. The same combination of environmental parameters could explain the OTU$_{\text{all}}$ data set without 30% rare OTU$_{\text{unique}}$ as the data set at the genus level, while the OTU$_{\text{all}}$ data set without 35-50% rare OTU$_{\text{unique}}$ followed similar ecological patterns as that of the family to phylum levels (**Figs. 2.4.**, **S2.6.**, and **Tables S2.2.-2.3.**).

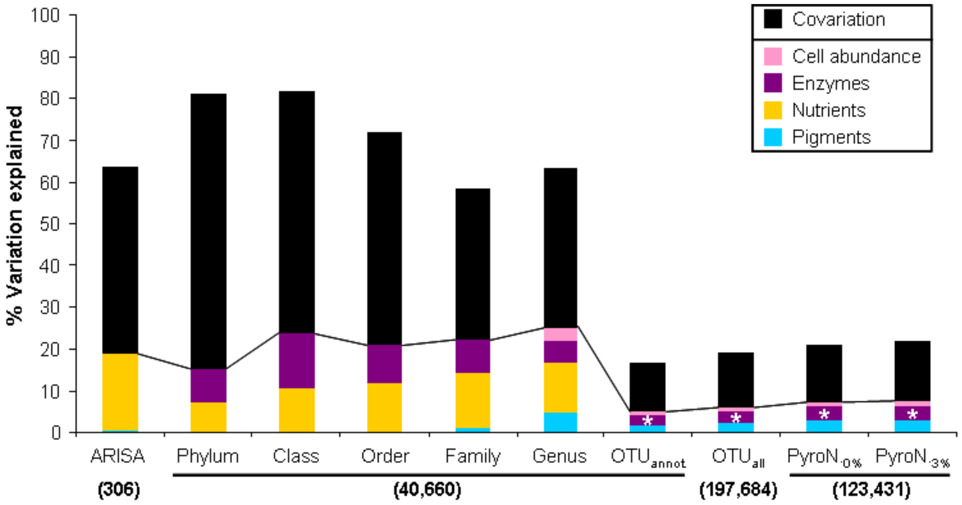


Figure 2.4. Partitioning of the biological variation in the bacterial community structure based on the ARISA and 454 MPTS. Environmental parameters accounted for include pigments (chlorophyll *a* and pheophytin), nutrients (silicate, phosphate, nitrite, nitrate, ammonium), extra-cellular enzyme activities (chitinase, α-glucosidase, β-glucosidase, lipase, aminopeptidase, phosphatase), cell abundance and their combined effects. The black line in each panel separates the pure factor effects from their covariations. Covariation of any of the 4 environmental factors is represented under the category "covariation". Negative values, unexplained variation and non-significant (NS) multivariate models are not shown. Here, the OTU$_{\text{annot.}}$ level represents sequences with a complete annotation from the phylum to the genus level (*i.e.* 20% of the total number of sequences in the original data set), while the OTU$_{\text{all}}$ level includes also sequences without complete annotation. PyroN.$_{0\%}$, PyroN.$_{3\%}$ represent the PyroNoise-corrected OTU$_{\text{all}}$ data set, clustered at 0% and 3% of sequence dissimilarity, respectively. Numbers in parentheses represent the total number of sequences in each data set. White stars indicate pure factors that significantly explain the biological variation (P value $\leq$ 0.05) after 1000 Monte Carlo permutations.

As previously observed (Gobet et al. *Submitted*), our results confirm that a greater amount of biological variation in the microbial community can be explained as data sets become less complex. Also, consistent ecological patterns were obtained with data sets with lower resolution such as that from ARISA or from the 454 MPTS taxonomic annotation or truncation. These patterns may be driven by the main bacterial types, *e.g.* the resident bacterial types, present at all times that probably maintain the general functions of the ecosystem. This follows a recent assumption regarding the ecological coherence of high bacterial taxonomic ranks (*i.e.* phylum to the genus level) which suggests that members of a same taxon would share the same main functions in the ecosystem (Philippot et al. 2010). Whereas ARISA would permit to understand the ecology of the main bacterial types in the ecosystem, the 454 MPTS data sets, with higher resolution of the microbial community, would give more details on the ecology of the community, also influenced by the large fraction of rare types in the microbial community. Finally, the pre-clustering and clustering corrections of the 454 MPTS data set confirmed the robustness of the ecological trends, as the conclusions were the same as without correction of the original OTU_{all} data set.

In summary, the above analyses showed how each molecular approach can not only answer specific questions, but also lead to the same ecological conclusions. ARISA may be better suited for a general overview of the bacterial community structure, and the data output may be easier to process than large data sets from high-throughput sequencing techniques. Highly similar ecological patterns obtained from ARISA and different taxonomic levels, PyroNoise-corrected or truncated 454 MPTS data sets indicated how both molecular approaches in fact produce a fingerprint of the bacterial community. Our study thus confirms ARISA as a valuable technique for a rapid and consistent evaluation of bacterial ecological patterns. In addition, the knowledge gained from previous studies based on ARISA or classical community fingerprinting techniques is not obsolete and may be fruitfully extended by using new molecular techniques.

Acknowledgements

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

2.4 Supplementary Information

2.4.1 Supplementary Figures

Fig. S2.1. Total OTU numbers from all depths over time for (A) ARISA, (B) $\mathrm{OTU}_{\mathrm{all}}$, and (C) $\mathrm{PyroNoise}_{3\%}$.

Fig. S2.2. Turnover of the bacterial community between sediment depth layers or sampling dates after correction and OTU clustering of the 454 MPTS data set and of the ARISA data set.

Fig. S2.3. Turnover of the bacterial community between sediment depth layers or sampling dates at successive taxonomic levels.

Fig. S2.4. Turnover of the bacterial community between sediment depth layers or sampling dates after applying MultiCoLA.

Fig. S2.5. Comparison of most important axes of extracted variation from the different categories of data sets.

Fig. S2.6. Partitioning of the biological variation in the bacterial community structure.

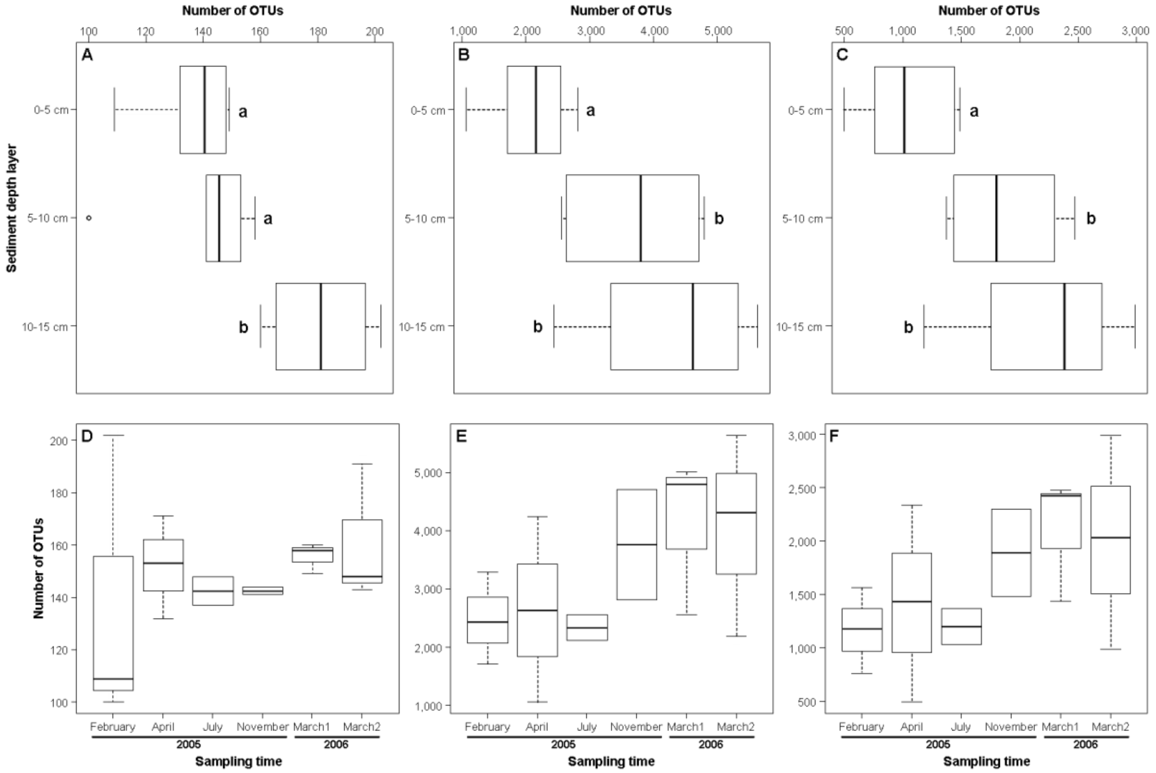
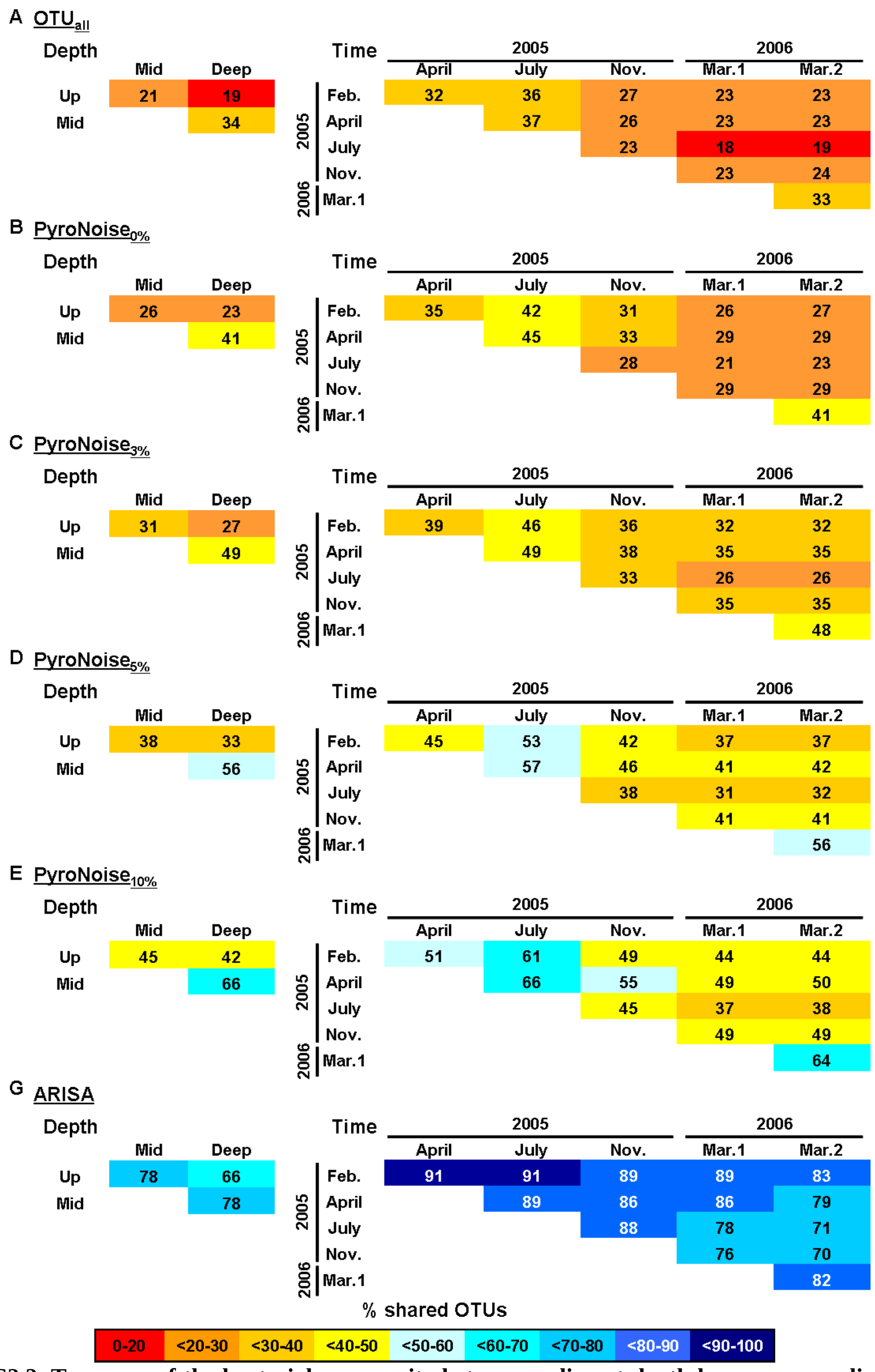


Figure S2.1. Total OTU numbers along sediment depth or over time for (A, D) ARISA, (B, E) OTU$_{all}$, and (C, F) PyroNoise$_{3\%}$ data sets. Horizontal lines outside of the box represent the smallest and the largest observations in the data set. The first and third quartiles are indicated by the lowest and highest limits of the box, respectively. The median is indicated by the thick bar in the middle of each box. (**A**, **B**, **C**) boxplots were calculated at all sampling times along the three depth layers. A different letter (**a**, **b**) indicates a significant mean difference in OTU number between the selected depth layer and the other(s) (Student t-test, *P value* ≤ 0.05). There were 6 sampling dates considered for the upper 10 cm layers and 4 sampling times for the 10-15 cm layer. In (**A**), the circle represents an outlier: a data point inferior or superior to the first or third quartile, respectively, by 1.5 times the interquartile range. (**D**, **E**, **F**) boxplots were calculated at all depths over time. The 3 depth layers were considered except for July and November 2005 where 2 upper depth layers were available. There was no significant mean difference in OTU number between sampling times after Student t-testing.



A OTU$_{all}$

Depth	Mid	Deep
Up	21	19
Mid		34

Time		2005			2006	
		April	July	Nov.	Mar.1	Mar.2
2005	Feb.	32	36	27	23	23
	April		37	26	23	23
	July			23	18	19
	Nov.				23	24
2006	Mar.1					33

B PyroNoise$_{0\%}$

Depth	Mid	Deep
Up	26	23
Mid		41

Time		2005			2006	
		April	July	Nov.	Mar.1	Mar.2
2005	Feb.	35	42	31	26	27
	April		45	33	29	29
	July			28	21	23
	Nov.				29	29
2006	Mar.1					41

C PyroNoise$_{3\%}$

Depth	Mid	Deep
Up	31	27
Mid		49

Time		2005			2006	
		April	July	Nov.	Mar.1	Mar.2
2005	Feb.	39	46	36	32	32
	April		49	38	35	35
	July			33	26	26
	Nov.				35	35
2006	Mar.1					48

D PyroNoise$_{5\%}$

Depth	Mid	Deep
Up	38	33
Mid		56

Time		2005			2006	
		April	July	Nov.	Mar.1	Mar.2
2005	Feb.	45	53	42	37	37
	April		57	46	41	42
	July			38	31	32
	Nov.				41	41
2006	Mar.1					56

E PyroNoise$_{10\%}$

Depth	Mid	Deep
Up	45	42
Mid		66

Time		2005			2006	
		April	July	Nov.	Mar.1	Mar.2
2005	Feb.	51	61	49	44	44
	April		66	55	49	50
	July			45	37	38
	Nov.				49	49
2006	Mar.1					64

G ARISA

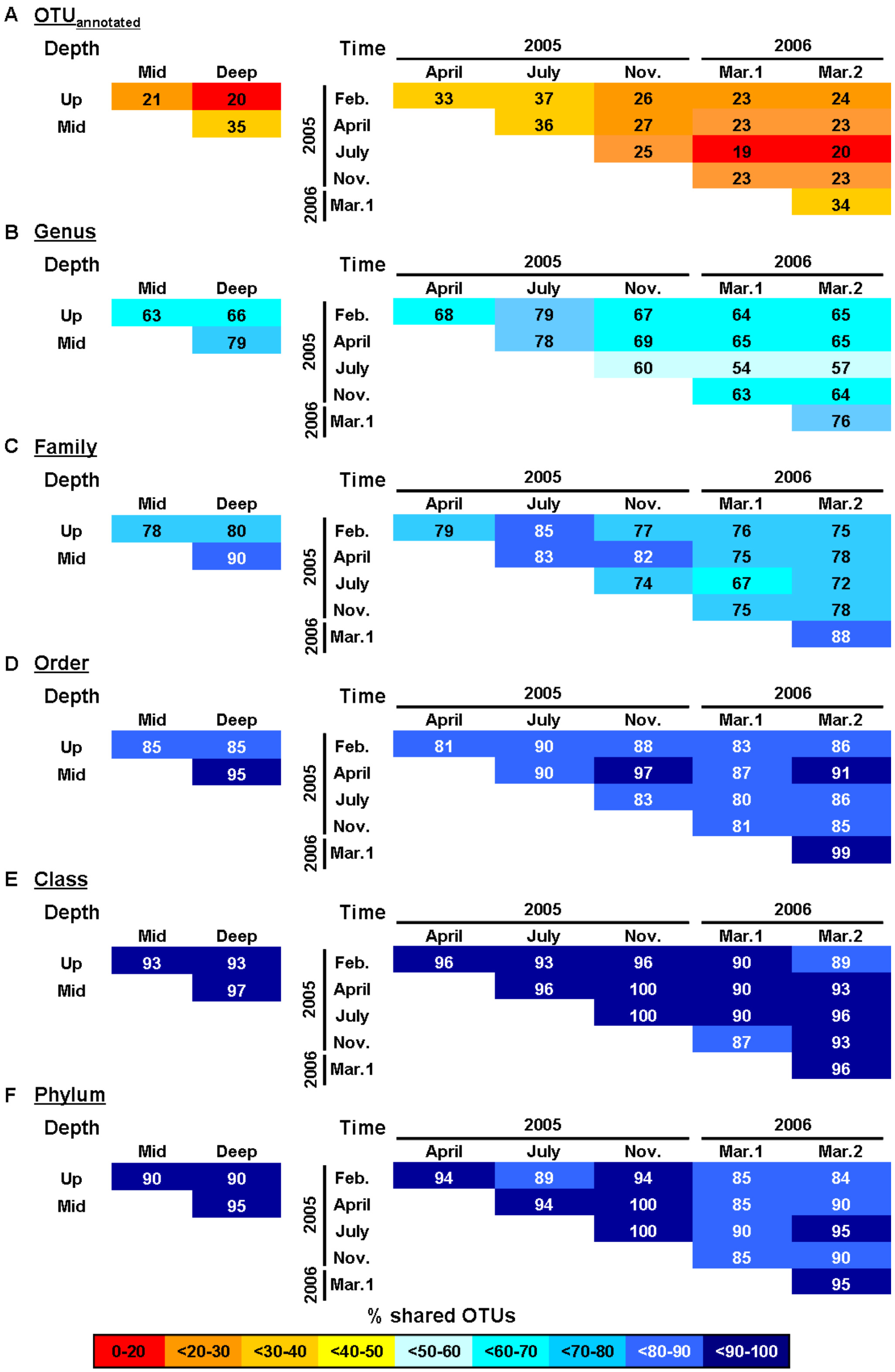
Depth	Mid	Deep
Up	78	66
Mid		78

Time		2005			2006	
		April	July	Nov.	Mar.1	Mar.2
2005	Feb.	91	91	89	89	83
	April		89	86	86	79
	July			88	78	71
	Nov.				76	70
2006	Mar.1					82

% shared OTUs

0-20	<20-30	<30-40	<40-50	<50-60	<60-70	<70-80	<80-90	<90-100

Figure S2.2. Turnover of the bacterial community between sediment depth layers or sampling dates after correction and OTU clustering of the 454 MPTS data set and of the ARISA data set. (**A**) OTU$_{all}$, (**B**) PyroNoise unique, (**C**) PyroNoise 3%, (**D**) PyroNoise 5%, (**E**) PyroNoise 10%, and (**F**) ARISA. The percentage of OTU shared between a sampling depth (or date) and the previous one was calculated and values were represented by heatmap matrices. OTU$_{all}$ represents the original data set with all OTU$_{unique}$. Figure modified from (Gobet et al. *Submitted*).



A OTU$_{annotated}$

Depth	Mid	Deep
Up	21	20
Mid		35

Time	2005			2006	
	April	July	Nov.	Mar.1	Mar.2
2005 Feb.	33	37	26	23	24
April		36	27	23	23
July			25	19	20
Nov.				23	23
2006 Mar.1					34

B Genus

Depth	Mid	Deep
Up	63	66
Mid		79

Time	2005			2006	
	April	July	Nov.	Mar.1	Mar.2
2005 Feb.	68	79	67	64	65
April		78	69	65	65
July			60	54	57
Nov.				63	64
2006 Mar.1					76

C Family

Depth	Mid	Deep
Up	78	80
Mid		90

Time	2005			2006	
	April	July	Nov.	Mar.1	Mar.2
2005 Feb.	79	85	77	76	75
April		83	82	75	78
July			74	67	72
Nov.				75	78
2006 Mar.1					88

D Order

Depth	Mid	Deep
Up	85	85
Mid		95

Time	2005			2006	
	April	July	Nov.	Mar.1	Mar.2
2005 Feb.	81	90	88	83	86
April		90	97	87	91
July			83	80	86
Nov.				81	85
2006 Mar.1					99

E Class

Depth	Mid	Deep
Up	93	93
Mid		97

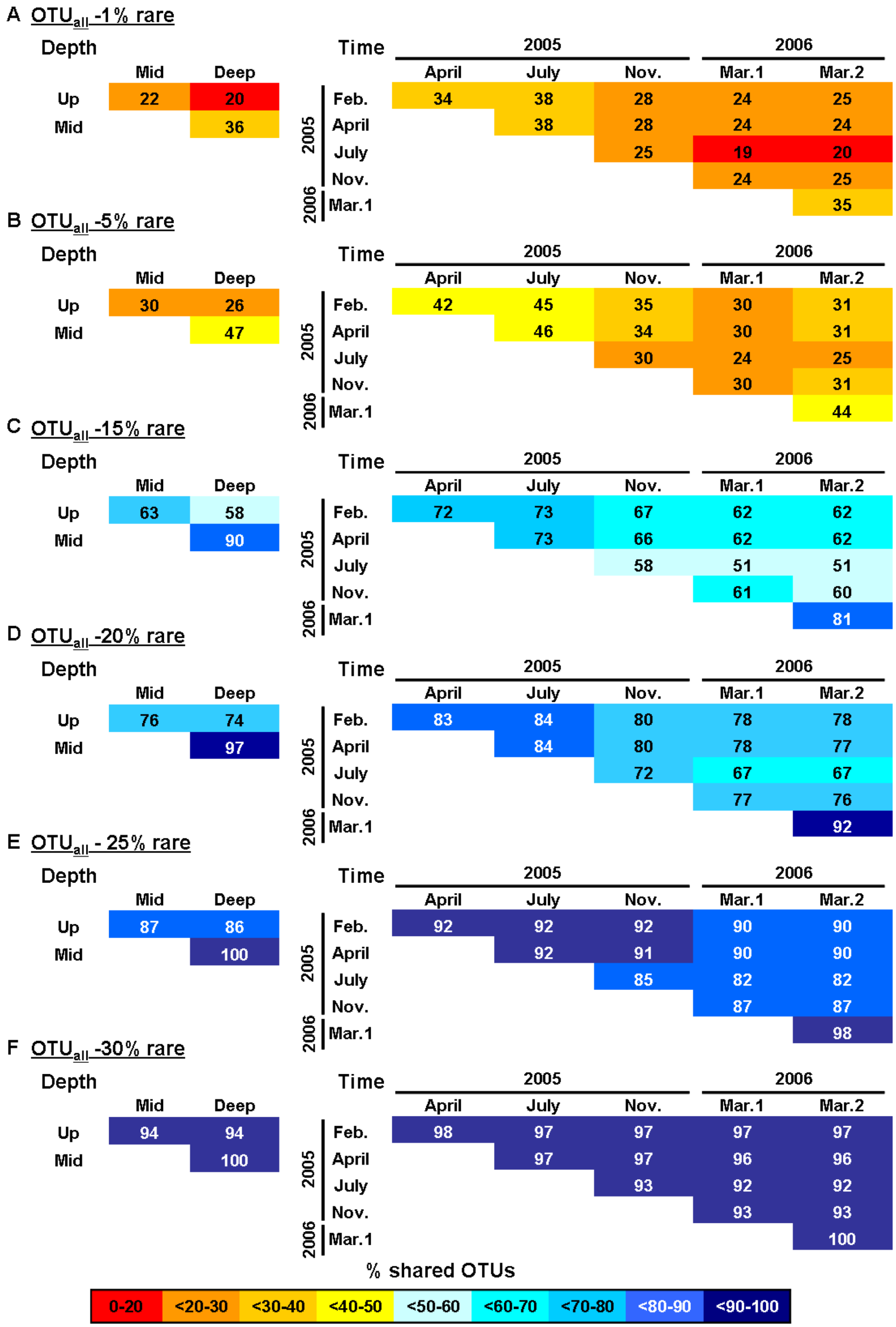
Time	2005			2006	
	April	July	Nov.	Mar.1	Mar.2
2005 Feb.	96	93	96	90	89
April		96	100	90	93
July			100	90	96
Nov.				87	93
2006 Mar.1					96

F Phylum

Depth	Mid	Deep
Up	90	90
Mid		95

Time	2005			2006	
	April	July	Nov.	Mar.1	Mar.2
2005 Feb.	94	89	94	85	84
April		94	100	85	90
July			100	90	95
Nov.				85	90
2006 Mar.1					95

Figure S2.3. Turnover of the bacterial community between sediment depth layers or sampling dates at successive taxonomic levels. (**A**) OTU$_{annotated}$, (**B**) Genus, (**C**) Family, (**D**) Order, (**E**) Class, and (**F**) Phylum levels. OTU$_{annotated}$ represents OTU$_{unique}$ that are completely annotated (phylum to genus levels). See Fig. S2 for further details.



A OTU_{all}-1% rare

Depth	Mid	Deep
Up	22	20
Mid		36

Time		2005			2006	
		April	July	Nov.	Mar.1	Mar.2
2005	Feb.	34	38	28	24	25
	April		38	28	24	24
	July			25	19	20
	Nov.				24	25
2006	Mar.1					35

B OTU_{all}-5% rare

Depth	Mid	Deep
Up	30	26
Mid		47

Time		2005			2006	
		April	July	Nov.	Mar.1	Mar.2
2005	Feb.	42	45	35	30	31
	April		46	34	30	31
	July			30	24	25
	Nov.				30	31
2006	Mar.1					44

C OTU_{all}-15% rare

Depth	Mid	Deep
Up	63	58
Mid		90

Time		2005			2006	
		April	July	Nov.	Mar.1	Mar.2
2005	Feb.	72	73	67	62	62
	April		73	66	62	62
	July			58	51	51
	Nov.				61	60
2006	Mar.1					81

D OTU_{all}-20% rare

Depth	Mid	Deep
Up	76	74
Mid		97

Time		2005			2006	
		April	July	Nov.	Mar.1	Mar.2
2005	Feb.	83	84	80	78	78
	April		84	80	78	77
	July			72	67	67
	Nov.				77	76
2006	Mar.1					92

E OTU_{all}- 25% rare

Depth	Mid	Deep
Up	87	86
Mid		100

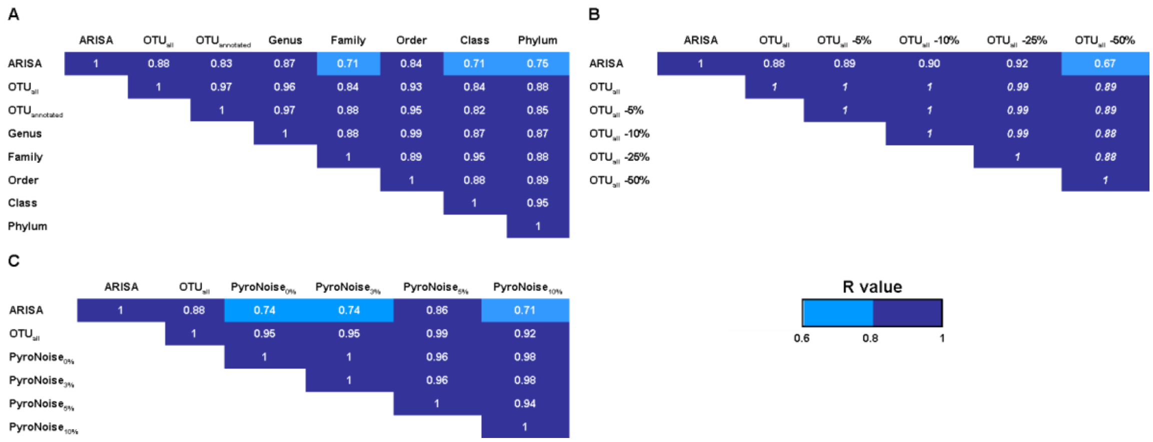
Time		2005			2006	
		April	July	Nov.	Mar.1	Mar.2
2005	Feb.	92	92	92	90	90
	April		92	91	90	90
	July			85	82	82
	Nov.				87	87
2006	Mar.1					98

F OTU_{all}-30% rare

Depth	Mid	Deep
Up	94	94
Mid		100

Time		2005			2006	
		April	July	Nov.	Mar.1	Mar.2
2005	Feb.	98	97	97	97	97
	April		97	97	96	96
	July			93	92	92
	Nov.				93	93
2006	Mar.1					100

Figure S2.4. Turnover of the bacterial community between sediment depth layers or sampling dates after applying MultiCoLA. Successive percentages of rare OTU were removed from the OTU_{all} data set: (**A**) 1%, (**B**) 5%, (**C**) 15%, (**D**) 20%, (**E**) 25%, and (**F**) 30%. See Fig. S2 for further details. Figure modified from (Gobet et al. *Submitted*).



A

	ARISA	OTU_{all}	$OTU_{annotated}$	Genus	Family	Order	Class	Phylum
ARISA	1	0.88	0.83	0.87	0.71	0.84	0.71	0.75
OTU_{all}		1	0.97	0.96	0.84	0.93	0.84	0.88
$OTU_{annotated}$			1	0.97	0.88	0.95	0.82	0.85
Genus				1	0.88	0.99	0.87	0.87
Family					1	0.89	0.95	0.88
Order						1	0.88	0.89
Class							1	0.95
Phylum								1

B

	ARISA	OTU_{all}	OTU_{all} -5%	OTU_{all} -10%	OTU_{all} -25%	OTU_{all} -50%
ARISA	1	0.88	0.89	0.90	0.92	0.67
OTU_{all}		*1*	*1*	*1*	*0.99*	*0.89*
OTU_{all} -5%			*1*	*1*	*0.99*	*0.89*
OTU_{all} -10%				*1*	*0.99*	*0.88*
OTU_{all} -25%					*1*	*0.88*
OTU_{all} -50%						*1*

C

	ARISA	OTU_{all}	$PyroNoise_{0\%}$	$PyroNoise_{3\%}$	$PyroNoise_{5\%}$	$PyroNoise_{10\%}$
ARISA	1	0.88	0.74	0.74	0.86	0.71
OTU_{all}		1	0.95	0.95	0.99	0.92
$PyroNoise_{0\%}$			1	1	0.96	0.98
$PyroNoise_{3\%}$				1	0.96	0.98
$PyroNoise_{5\%}$					1	0.94
$PyroNoise_{10\%}$						1

Figure S2.5. Comparison of extracted variation from the different categories of data sets. The Procrustes' correlation coefficient was used to compare the ARISA and the OTU_{all} extracted variation with (**A**) various levels of taxonomic annotation, (**B**) successive removal of the rare OTU and (**C**) successive clustering of the PyroNoise-corrected data to define OTU. The Procrustes' correlation coefficient was calculated from the variation in the main axes of extracted variation via non-metric multidimensional scaling [NMDS], based on the distance matrices resulting from the relative sequence abundances. For (**B**) values *in italic* indicate no test of significance as the truncated matrices are not statistically independent from each other (see Materials and Methods). Only significant values (after Bonferroni correction) are shown.

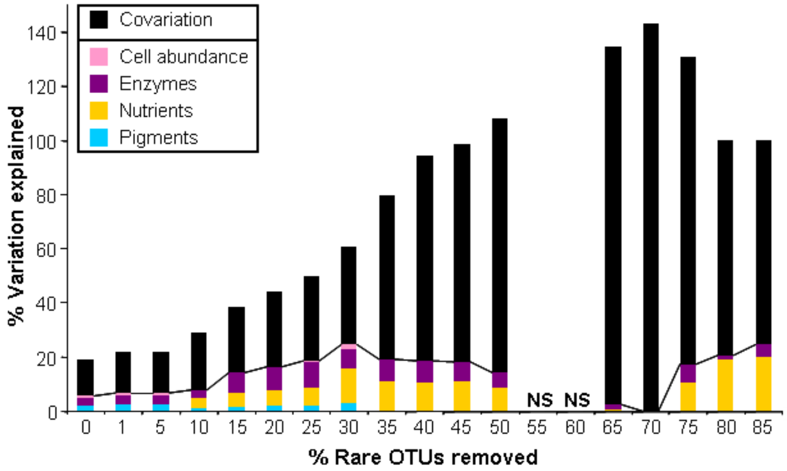


Figure S2.6. Partitioning of the biological variation in the bacterial community structure based on the removal of successive percentages (here 0 to 85%) of rare OTU from the OTU_{all} data set by using MultiCoLA. Environmental parameters accounted for include pigments (chlorophyll *a* and pheophytin), nutrients (silicate, phosphate, nitrite, nitrate, ammonium), extra-cellular enzyme activities (chitinase, α-glucosidase, β-glucosidase, lipase, aminopeptidase, phosphatase), cell abundance and their combined effects. The black line in each panel separates the pure factor effects from their covariations. Covariation of any of the 4 environmental factors is represented under the category “covariation”. Negative values, unexplained variation and non significant (NS) multivariate models are not shown.

N.B: A non significant (NS) model appears when trying to explain the biological variation from data sets without 55-60% of rare OTU. This is probably not an artifact. Indeed, as an increasing amount of rare OTU is removed, the original structure of the data set may be disorganized. Hence, as rare OTU are removed from the data set, the amount of explained variation likely varies. Additionally, when OTU are removed from the data set, the data set structure may vary. The interesting point here is that similar patterns can be kept after a consequent amount of OTU removed.

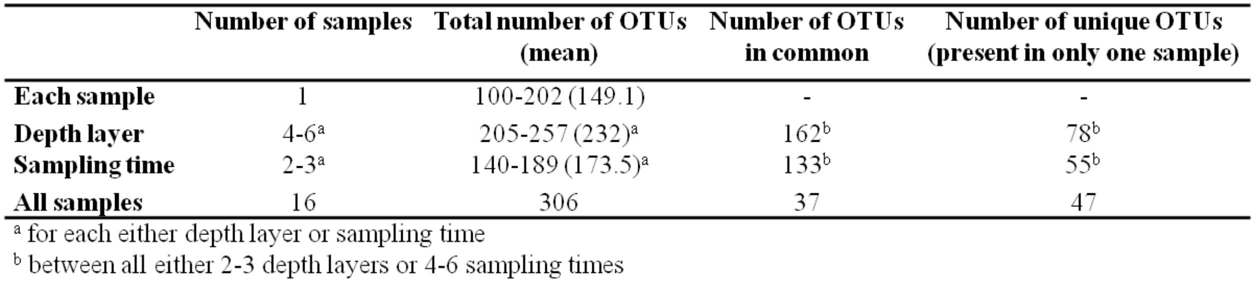
2.4.2 Supplementary Tables

Table S2.1. Summary of OTU abundances in the ARISA data set for all samples at three different sediment depths at different sampling times

Table S2.2. Contribution of environmental parameters to the variation in the data sets at successive taxonomic levels.

Table S2.3. Contribution of environmental parameters to the variation in truncated data sets [by applying MultiCoLA, (Gobet et al. 2010)], at the OTU level for all sequences available.

Table S2.1. Summary of OTUs abundances in the ARISA data set for all samples at different sediment depths and at different sampling times



	Number of samples	Total number of OTUs (mean)	Number of OTUs in common	Number of unique OTUs (present in only one sample)
Each sample	1	100-202 (149.1)	-	-
Depth layer	4-6[a]	205-257 (232)[a]	162[b]	78[b]
Sampling time	2-3[a]	140-189 (173.5)[a]	133[b]	55[b]
All samples	16	306	37	47

[a] for each either depth layer or sampling time

[b] between all either 2-3 depth layers or 4-6 sampling times

Table S2.2. Contribution of environmental parameters to the variation in the data sets at successive taxonomic levels.

Cutoff levels[a]	Total number of sequences	$R^{2,b}$	Individual factor contribution[c]											
					Nutrients					Enzymes				
			Salinity	Pigments	SiO_2	PO_4	NO_2	NO_3	NH_4	Chit	α-glu	Lip	Phos	Cell abundance
$PyroNoise_{3\%}$	123,431	20%***		-1									-0.8	-0.9
$PyroNoise_{0\%}$	123,431	19%***		-1									-0.8	-0.9
OTU_{all}	197,684	17%***		-1									-0.8	-0.9
$OTU_{annotated}$	40,660	14%***		-1									-0.8	-0.9
Genus	40,660	51%***	-0.1	-1	0.4	-0.3	0.3	-0.5	0.6		-0.8	-0.1	-0.8	-0.9
Family	40,660	57%***	-0.1	-1	0.4	-0.3	0.3	-0.5	0.6	-0.6	-0.8	-0.2	-0.8	-0.9
Order	40,660	64%**	-0.1	-1	0.4	-0.3	0.4	-0.5	0.5	-0.6	-0.8	-0.2	-0.8	-0.9
Class	40,660	75%***	-0.2	-1	0.5	-0.2	0.4	-0.5	0.6	-0.6	-0.8	-0.2	-0.8	-0.9
Phylum	40,660	75%**	0.2	1	-0.6	0.2	-0.4	0.4	-0.6	0.6	0.8	0.2	0.8	0.9
ARISA[d]	-	51%***		-0.8	0.2	-0.4	0.2	-0.6	0.4				-0.6	-0.8

[a] Cutoff levels were defined based on the whole data set strategy (see Supplementary Fig. 1). Cutoff levels were applied until samples were lost due to lack of sequences.

[b] Adjusted R^2 indicates the amount of variation explained by environmental parameters (salinity, pigments, nutrients, enzymes and cell abundance), their significance is indicated as NS (non significant), * ($P \leq 0.05$), ** ($P \leq 0.01$), and *** ($P \leq 0.001$). Values were rounded to one decimal after the comma.

[c] Only significant, standardized correlation coefficients to the first redundancy analysis (RDA) axis are indicated for each parameter.

Chl a, chlorophyll a; SiO_2, silicate; PO_4, phosphate; NO_2, nitrite; NO_3, nitrate; NH_4, ammonium; Chit, chitinase; α-glu, α-glucosidase; Lip, lipase; Phos, phosphatase.

[d] No total number of sequences were indicated for the ARISA data set as it is peaks areas.

Table S2.3. Contribution of environmental parameters to the variation in truncated data sets (by applying MultiCoLA, (Gobet et al. 2010)), at the OTU level for all sequences available.

Cutoff levels[a]	Total number of sequences	$R^{2,b}$	Individual factor contribution[c]														
			Salinity	Pigments		Nutrients					Enzymes						Cell abundanc e
				Chl a	Pheo	SiO_2	PO_4	NO_2	NO_3	NH_4	Chit	α-glu	β-glu	Lip	Amin	Phos	
0%	197,684	17%***		-1												-0.8	-0.9
1%	195,707	18%***		-1												-0.8	-0.9
5%	187,799	20%***		-1												-0.8	-0.9
10%	177,915	27%***		-1		0.4	-0.3		-0.5							-0.8	-0.9
15%	168,029	36%***		-1		0.4	-0.3		-0.5			-0.7		-0.1		-0.8	-0.9
20%	158,143	41%***		-1		0.4	-0.3		-0.5			-0.7		-0.1		-0.8	-0.9
25%	148,258	46%***		-1		0.4	-0.3		-0.5			-0.7		-0.1		-0.8	-0.9
30%	138,377	55%***	-0.1	-1		0.4	-0.3	0.3	-0.5	0.5		-0.7		-0.1		-0.8	-0.9
35%	128,459	60%*	-0.1	-1		0.4	-0.3	0.3	-0.5	0.5	-0.6	-0.7	-0.6	-0.1	-0.8	-0.8	-0.9
40%	118,590	63%*	-0.1	-1		0.4	-0.3	0.3	-0.5	0.5	-0.6	-0.7	-0.6	-0.1	-0.8	-0.8	-0.9
45%	108,709	62%***	-0.1	-1		0.4	-0.3	0.3	-0.5	0.5	-0.6	-0.7	-0.6	-0.1	-0.8	-0.8	-0.9
50%	98,762	63%*	-0.1	-1		0.4	-0.3	0.3	-0.5	0.5	-0.6	-0.7	-0.6	-0.2	-0.8	-0.8	-0.9
55%	88,871	NS															
60%	78,951	NS															
65%	68,789	60%*		-0.9	-0.9	0.3	-0.3	0.3	-0.5	0.4	-0.6	-0.7		-0.2		-0.7	-0.9
70%	58,637	55%*		0.9	0.9	-0.2	0.4	-0.2	0.5	-0.4	0.6	0.7		0.1		0.7	0.9
75%	49,110	52%*	0.3	0.8		-0.4	0.1	-0.4	0.3	-0.3	0.7	0.6		0.2		0.8	0.7
80%	38,299	39%***							0.5	-0.5	-0.7	-0.1					
85%	25,961	47%***				-0.4				-0.7	0.6				-0.1		
90%	16,852																
95%	6,550																
99%	0																

[a]Cutoff levels were defined based on the whole data set strategy (see Supplementary Fig. 1). Cutoff levels were applied until samples were lost due to lack of sequences.

[b]Adjusted R^2 indicates the amount of variation explained by environmental parameters (salinity, pigments, nutrients, enzymes and cell abundance), their significance is indicated as NS (non significant), * ($P \leq 0.05$), ** ($P \leq 0.01$), and *** ($P \leq 0.001$). Values were rounded to one decimal after the comma.

[c]Only significant, standardized correlation coefficients to the first redundancy analysis (RDA) axis are indicated for each parameter.

Chl a, chlorophyll a; Pheo, Pheophytin; SiO_2, silicate; PO_4, phosphate; NO_2, nitrite; NO_3, nitrate; NH_4, ammonium; Chit, chitinase; α-glu, α-glucosidase; β-glu, β-glucosidase; Lip, lipase; Amin, aminopeptidase; Phos, phosphatase.

3 Chapter II.

Multivariate Cutoff Level Analysis (MultiCoLA) of Large Community Datasets

Angélique Gobet, Christopher Quince and Alban Ramette

Published in Nucleic Acids Research in June 2010.

Keywords: Bacterial diversity, 454 massively parallel tag sequencing, multivariate analysis, rare biosphere, microbial ecology

3.1 Published article

Nucleic Acids Research Advance Access published June 14, 2010

Nucleic Acids Research, 2010, ***1–9***
doi:10.1093/nar/gkq545

Multivariate Cutoff Level Analysis (MultiCoLA) of large community data sets

Angélique Gobet[1,2], Christopher Quince[3] and Alban Ramette[1,*]

[1]Microbial Habitat Group, Max Planck Institute for Marine Microbiology, Bremen, [2]Jacobs University Bremen GmbH, Bremen, Germany and [3]Department of Civil Engineering, University of Glasgow, Rankine building, Glasgow, UK

Received March 1, 2010; Revised May 26, 2010; Accepted May 29, 2010

ABSTRACT

High-throughput sequencing techniques are becoming attractive to molecular biologists and ecologists as they provide a time- and cost-effective way to explore diversity patterns in environmental samples at an unprecedented resolution. An issue common to many studies is the definition of what fractions of a data set should be considered as rare or dominant. Yet this question has neither been satisfactorily addressed, nor is the impact of such definition on data set structure and interpretation been fully evaluated. Here we propose a strategy, MultiCoLA (Multivariate Cutoff Level Analysis), to systematically assess the impact of various abundance or rarity cutoff levels on the resulting data set structure and on the consistency of the further ecological interpretation. We applied MultiCoLA to a 454 massively parallel tag sequencing data set of V6 ribosomal sequences from marine microbes in temperate coastal sands. Consistent ecological patterns were maintained after removing up to 35–40% rare sequences and similar patterns of beta diversity were observed after denoising the data set by using a preclustering algorithm of 454 flowgrams. This example validates the importance of exploring the impact of the definition of rarity in large community data sets. Future applications can be foreseen for data sets from different types of habitats, e.g. other marine environments, soil and human microbiota.

INTRODUCTION

Community ecologists traditionally deal with data sets consisting of large tables of samples by 'species' (hereafter referred to as 'types'). The scientific community has yet not reached a general agreement on the optimal way to deal with rare types (1): for some, rare types are noise in data sets which may originate from sampling artifacts and thus do not represent the whole community. Rare types are often removed so as to decrease the large amount of zeros stored in data sets, and to reduce the challenging task of their taxonomic identification (1). For others, rare types are valuable as they may provide critical insights into the functioning of ecosystems such as resistance against invasive species or into the likely existence of multiple niches (1). It is thus left at the discretion of the authors to define their own concept of rarity: rare plants and animals may be defined according to their restricted geographical distribution (2) or to their low proportions in data sets (3).

In microbial ecology, the current revolution in high-throughput DNA sequencing technology has revealed the existence of a 'rare biosphere', consisting of the many microbial types displaying long distribution tails in rank-abundance curves (4,5). Because sequencing artifacts may produce chimeric types (6), several studies have put into doubt the true existence of rare types in the high-throughput sequencing data sets and have provided various ways to trim and correct sequences: for instance, clustering threshold at 97% sequence identity (7) on 454 massively parallel tag sequencing (MPTS) data or a flowgram-based preclustering algorithm (8) may be applied. When rare types are not considered as artifacts, they can be defined by applying arbitrary abundance cutoffs to the original data set (9). However, the effects of the definition of rare organisms on the stability of the data structure and ecological conclusions that derive from the resulting, truncated data sets have not been examined so far.

We propose a new approach, Multivariate Cutoff Level Analysis (MultiCoLA), to systematically explore how large community data sets are affected by different definitions of rarity. First, MultiCoLA truncates the original data set by discarding rare types according to successive increasing

[*]To whom correspondence should be addressed. Tel: +49 421 2028 863; Fax: +49 421 2028 690; Email: aramette@mpi-bremen.de

abundance cutoffs. The effects of removing rare types are then measured at the levels of (i) variation of data set structure, (ii) amounts of extracted variation between the original and the truncated data sets and (iii) the ecological interpretation of the original and each truncated data sets when environmental parameters are available.

MATERIALS AND METHODS

Data set

In this study, the analyses were performed on a data set consisting of hyper-variable V6 sequences of the 16S rRNA gene, which were obtained from the application of 454 MPTS on temperate subtidal sandy samples at three sediment depth layers (0–15 cm depth, with a 5-cm interval) taken over 2 years (2005–2006). Detailed sample processing and DNA extraction has been described earlier (10) and the 454 MPTS of the extracted DNA was processed as described previously (5). The output from 454 MPTS was retrieved from the publicly available Visualization and Analysis of Microbial Populations Structure (VAMPS) web site (http://vamps.mbl.edu/). An automatic annotation pipeline [Global Alignment for Sequence Taxonomy (GAST) (5)] using several known databases (Entrez Genome, RDP and SILVA) allowed the taxonomic assignment of the sequences. Despite the limitations of current databases, only 6% of sequences from this data set were not taxonomically identified at all. However, about 20% of sequences were annotated from the phylum to the genus level. In this study, the analyses were performed by defining OTUs (operational taxonomic units) as unique sequences (i.e. sequences differing by at least one base were considered as different OTUs. Note, however, that MultiCoLA could also have been applied to sequence subsets based on another OTU definition) and the following subsets were considered: (i) all, unannotated sequences that we referred to as ‘OTU whole data set (DS)’, (ii) on the 20% fully annotated sequences (i.e. from phylum to genus levels and the corresponding OTU level) and (iii) on PyroNoise-corrected data defined at different percentages of sequence similarity.

Data analyses

Truncated tables. Data sets were analyzed by applying two types of cutoff abundance levels (Figure 1): (i) Whole-data set-based cutoffs*:* truncated matrices were obtained by removing chosen proportions (0, 1, 5–95 and 99%) of rare OTUs from the total sum of sequences in the data set (Figure 1A). The original data set was first sorted according to the decreasing number of sequences per OTU. Then low-abundance OTUs were removed according to the given cutoff levels. (ii) Sample-based cutoffs: a total of 15 cutoffs were selected from 1 to 208 total number of sequences per OTU per sample (because certain samples did not contain any more OTUs for cutoff levels higher than 208 sequences, i.e. 208 was the lowest number of the maximum OTU occurrences per sample), in order to select OTUs with more sequences than the applied cutoff (Figure 1B). This number is obviously

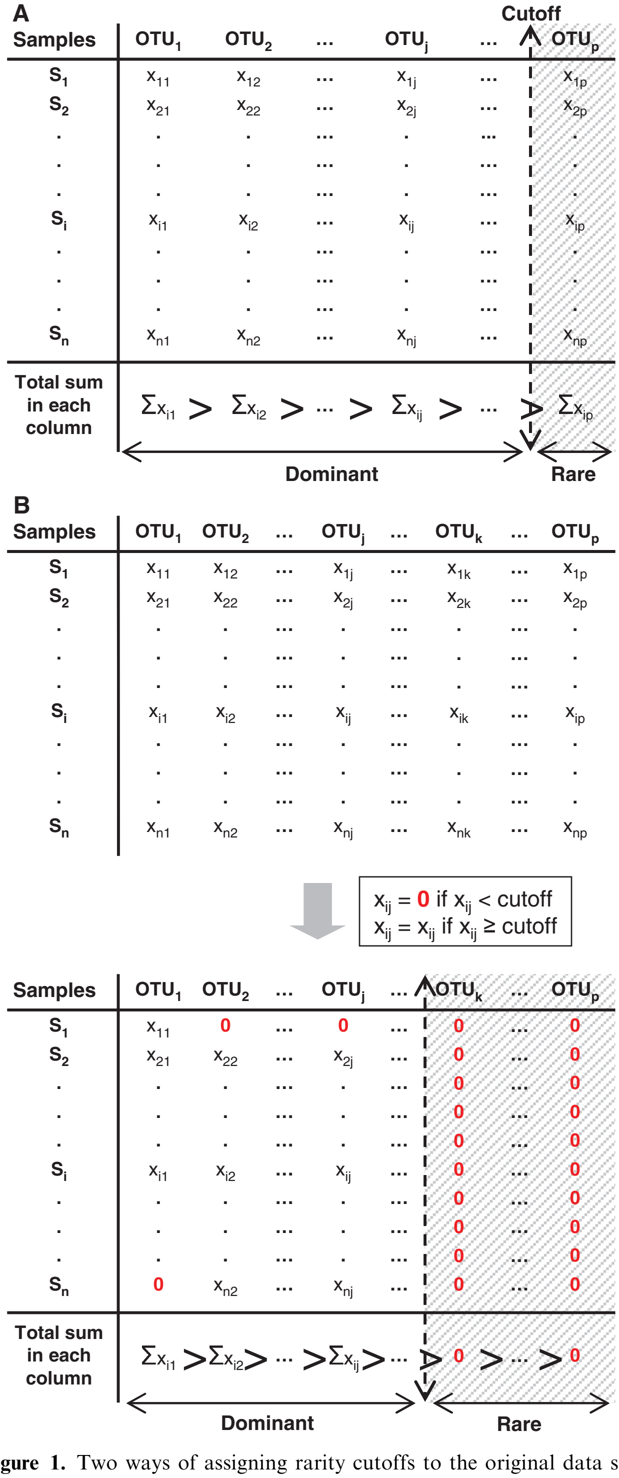
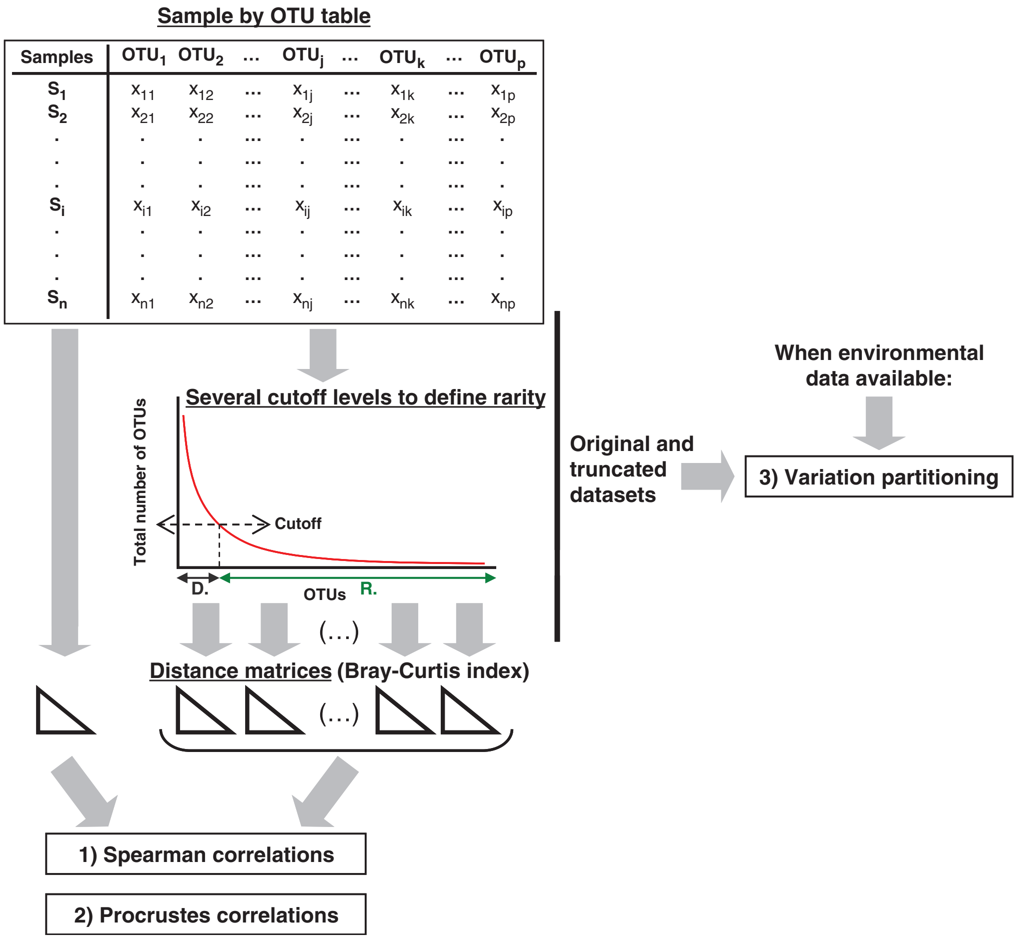


Figure 1. Two ways of assigning rarity cutoffs to the original data set. **(A)** In the data set-based approach, cutoff levels are assigned to the original data set according to several percentages (0, 1, 5–95 and 99%) of the total number of sequences in the data set. The data set was sorted according to the decreasing total sum of OTU sequences (columns, here) before selecting out rare OTUs. For instance, a cutoff assignment of 1% removes 1% of the low-abundant OTUs. **(B)** In the sample-based approach, cutoff levels are assigned to the original data set according to the occurrence (1–208 sequences) of each OTU in each sample. The maximum cutoff (here, 208) was chosen according to the lowest number of the maximum OTU occurrences in all samples; this is the limit when some samples did not contain any more OTUs. For example, the assignment of a cutoff level of 3 removes OTUs occurring less than three times in each sample.


Sample by OTU table
Several cutoff levels to define rarity
Total number of OTUs
Cutoff
D.
OTUs
R.
(...)
Distance matrices (Bray-Curtis index)
(...)
Original and truncated datasets
When environmental data available:
3) Variation partitioning
1) Spearman correlations
2) Procrustes correlations

Samples	OTU_1	OTU_2	...	OTU_j	...	OTU_k	...	OTU_p
S_1	x_{11}	x_{12}	...	x_{1j}	...	x_{1k}	...	x_{1p}
S_2	x_{21}	x_{22}	...	x_{2j}	...	x_{2k}	...	x_{2p}
.	.	.	...	.	...	.	...	.
.	.	.	...	.	...	.	...	.
.	.	.	...	.	...	.	...	.
S_i	x_{i1}	x_{i2}	...	x_{ij}	...	x_{ik}	...	x_{ip}
.	.	.	...	.	...	.	...	.
.	.	.	...	.	...	.	...	.
.	.	.	...	.	...	.	...	.
S_n	x_{n1}	x_{n2}	...	x_{nj}	...	x_{nk}	...	x_{np}

Figure 2. MultiCoLA steps. After truncating the original table according to various abundance cutoff levels, the effects of specific rarity definitions are tested by applying three types of analyses: (1) Variations in data set structure are established based on non-parametric correlations of pairwise distance matrices (e.g. calculated with the Bray–Curtis coefficient). (2) The amounts of extracted community variation (using NMDS) from the original data and the truncated data sets are compared by Procrustes correlations. (3) When additional parameters are available, the biological variation that can be explained by environmental parameters in the original and in the truncated data sets are then systematically compared. D, dominant OTUs; R, rare OTUs.

specific to each data set and should be taken into consideration if one wants to consider the same number of samples in all comparative analyses.

Analyses of changes in bacterial community structure and in main patterns of community variation. Pairwise distance matrices were calculated from the data (original and truncated matrices) using the Bray–Curtis dissimilarity index (11). The resulting dissimilarity matrices were compared with one another using the non-parametric Spearman rho correlation coefficient (12), which ranges from 0 to 1 (a score closer to 1 indicates higher correlations between dissimilarity matrices).

Variations in the main axes of extracted variation in community structure were explored via non-metric multidimensional scaling [NMDS (13)], a method commonly used to identify diversity patterns from molecular fingerprinting results (14). The Procrustes method (15) was then used to compare the NMDS ordination results from the original distance matrix with those from the truncated distance matrices. Procrustes rotation produces an R value that ranges from 0 to 1 [a score closer to 1 indicates highest similarities between the NMDS results (16)]. In other words, this approach enables to quantify the agreement between the most important axes of extracted variation from the original versus truncated data sets. This is particularly relevant because multivariate analyses that are typically applied to such data sets generally focus on the first few axes of main biological variation in the data.

In both profiles of data structure and extracted variation, a limitation is that one cannot calculate either the confidence interval or the significance of each pairwise comparison (i.e. for each single point). This is because the truncated matrices depend on the original matrix and testing correlations would only make sense in the case of data set independence (17,18). Yet, those limitations are not critical to our approach because we are more interested in overall changes in profiles rather than single-point variation or estimation. Indeed, the emphasis here is to measure (such as an index would do) the deviation from the signals in the original data set under the various hypothetical scenarios, i.e. when applying various cutoff levels.

Relationships between community structure and environment. For illustration purposes, four major

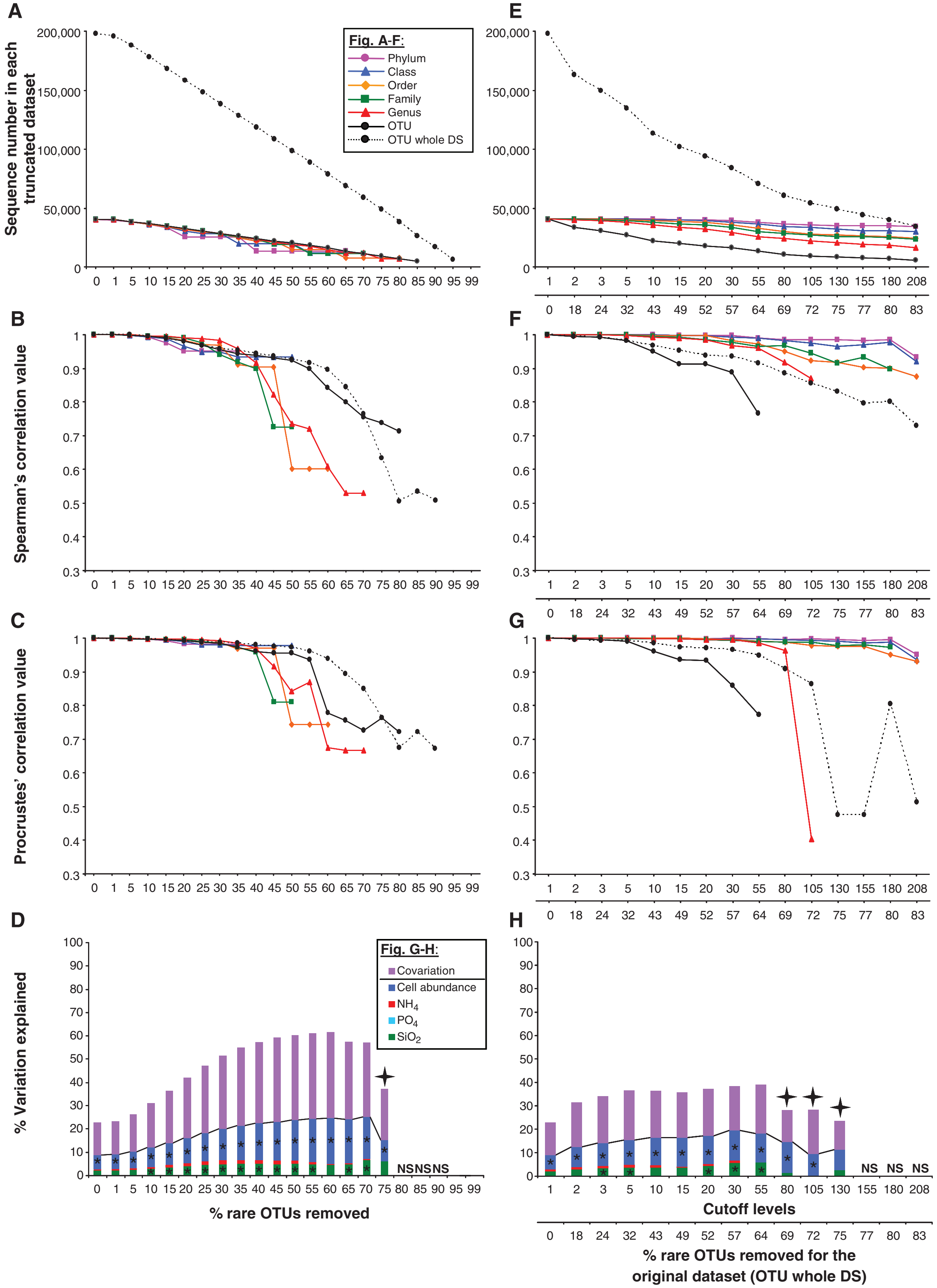


Figure 3. MultiCoLA profiles for data set structure, most important axes of extracted variation and interpretation of biological variation based on the data set-based (**A–D**) and sample-based (**E–H**) approaches. (A, E) Abundance of dominant OTUs in each truncated data set at the phylum, class, order, family, genus and OTU levels. A black solid line indicates comparisons at the OTU level for the data set with a complete annotation and a black dashed line indicates the OTU level with the whole data set (OTU whole DS). (B, F) Non-parametric Spearman correlations comparing the

contextual parameters [silicate, phosphate, ammonium and cell abundance from Böer *et al.* (10), which were $\log_{10}$-transformed prior to analyses] were used to investigate the relationships between the bacterial community structure (at successive assigned cutoffs and taxonomic levels) and environmental parameters. Each response community data set was Hellinger-transformed as recommended when dealing with data sets to be analyzed via linear multivariate models (19). Canonical variation partitioning (19,20) was then applied to the community data to test for the effects of each environmental variable (silicate, phosphate, ammonium and cell abundance) and their covariation on microbial community structure (21). Significances of the global and partial regression models were determined by using 1000 data permutations.

Creation of the MultiCoLA scripts. All statistical analyses were carried out using the R statistical environment (22), and specific routines in the *vegan* (23) and *MASS* (24) packages. The resulting MultiCoLA scripts are available at http://www.ecology-research.com. Some MultiCoLA scripts require some time and a certain computing power (10 min of calculations for an example matrix with 1000 OTUs on an Intel Pentium 4), but this may vary as a function of data set size and complexity, and choice of the analyses (i.e. Spearman correlations, Procrustes correlation or variation partitioning at multiple cutoff levels).

RESULTS AND DISCUSSION

Two approaches may be applied to truncate the original data set when removing an increasing proportion of rare types: either the whole data set is considered or each sample is considered individually (Figure 1). Because there is no reason to *a priori* choose a given threshold value, various cutoffs need to be systematically applied to explore their effects. The resulting, truncated data sets are then evaluated at three levels: first, the data sets are converted to sample-by-sample dissimilarity matrices (e.g. here we used the Bray–Curtis coefficient to calculate the dissimilarity between samples but other dissimilarity coefficients may be used) and those matrices are compared with the matrix produced by the whole dataset using non-parametric Spearman correlations (Figure 2), so as to assess changes in data structure. Second, the amounts of extracted ecological variation, obtained by the application of the NMDS ordination, in the truncated and original data sets are compared by Procrustes rotation (i.e. a measure of the correlation between two ordination solutions). Third, when contextual parameters (e.g. space, time or environment) are available, it is possible to systematically compare the ecological interpretation of each truncated data set with that of the original data set. This is achieved by partitioning the biological variation from the different truncated data sets as a function of explanatory variables (Materials and Methods section).

We applied MultiCoLA to a large 454 MPTS data set representing a case of high microbial diversity retrieved from temperate coastal sediments (10), which included a considerable amount of singletons (68% unique OTUs with a single sequence and 10% unique sequences in the whole data set) and low-abundant types. Another level of interest came from the fact that many sequences could also be taxonomically classified by applying the GAST taxonomic pipeline (5). It was thus possible to systematically explore the effects of rarity definition on the structure and interpretation of a data set at different taxonomic levels.

The systematic truncation of the whole data set produced a quasi linear decrease in sequence number as a function of increasing cutoff levels, and a similar trend was observed for the taxonomically annotated OTUs (Figure 3A). When the structure of community tables were compared between the truncated and the original matrices (Figure 3B), little variation in data structure was observed up to a removal threshold of 40% of the rare parts of the data set, indicating robustness in the signal far beyond the usual removal of singletons. Beyond the 40% threshold, the correlation coefficients greatly varied in a non-linear and non-predictive fashion, with higher taxonomic levels mostly associated with higher correlation values. When the most important patterns of extracted variation were compared between the various truncated and the original data sets (Figure 3C), a similar picture emerged with 40% representing a cutoff level up to which very little change in extracted variation could be observed. Beyond this threshold, Procrustes coefficients also greatly varied in a non-predictable and non-linear way, again regardless of the taxonomic level of the analysis.

When the truncated data sets were further analyzed as a function of environmental parameters, a surprising picture emerged (Figure 3D): nutrients (phosphate, silicate and ammonium) and total cell abundance seemed to consistently affect community variation at different cutoff levels. Not surprisingly, more explained variation was obtained overall when data complexity was reduced via the application of increasing cutoff levels or at higher taxonomic levels (Supplementary Figure S1). Noticeably, different multivariate models could be retained at each cutoff level or at each taxonomic level of the analyses, indicating that each truncated data set may be explained by slightly different combinations or covariations of environmental factors (Supplementary Tables S1–S7). It seemed overall

deviation in complete data structure between the original matrix and truncated matrices. (C, G) Comparison of most important axes of extracted variation between the original and truncated data sets. (D, H) Partitioning of the biological variation at the OTU level (all OTUs) into the respective effects of environmental factors (nutrients and cell abundance). Negative values, unexplained variation and non-significant models are not shown. SiO_2, silicate; PO_4, phosphate; NH_4, ammonium; covariation of any of the four environmental factors is represented under the same category. Asterisk indicates a significant effect of the pure factors ($P < 5\%$), whereas 'NS' indicates non-significant models. A cross indicates non-significant Bonferroni corrected models. Lacking points or bars are due to sample loss by applying a given cutoff to the original data set. In (E–H), the upper x-axis corresponds to cutoff levels defined as a function of the sample-based approach, and the lower x-axis represents the corresponding proportion of removed sequences in the OTU data set (all OTUs). This enables the comparison of the data set-based approach with the sample-based approach. Note that (D and H) have a different legend than (A–C) and (E–G).

that the rather broad taxonomic classification of the sequences was sufficient to describe general ecological patterns and that the interpretation of the effects of the structuring factors was robust and would not be affected by the removal of a large fraction of the rare types.

When applying the sample-based approach to the data to reveal changes in data structure and extracted variation (Figure 3F and G, respectively), changes in data structure varied in a narrower range (Spearman correlation coefficient from 0.8 to 1), while changes in extracted ecological variation varied over a larger range (Procrustes correlations from 0.5 to 1) and less predictably, as compared with their counterparts from the whole-data set approach (Figure 3B and C, respectively). A similar critical threshold of 35–40% for which profiles became more dissimilar from each other was also observed. For instance, by removing sequences occurring less than five times in the data set (i.e. removing 32% of all sequences), only a small drop in Spearman correlation coefficient to 0.98 would be observed, as compared with the original data set matrix, regardless of the taxonomic affiliation of the sequences (Figure 3F). Yet, the explained variations in

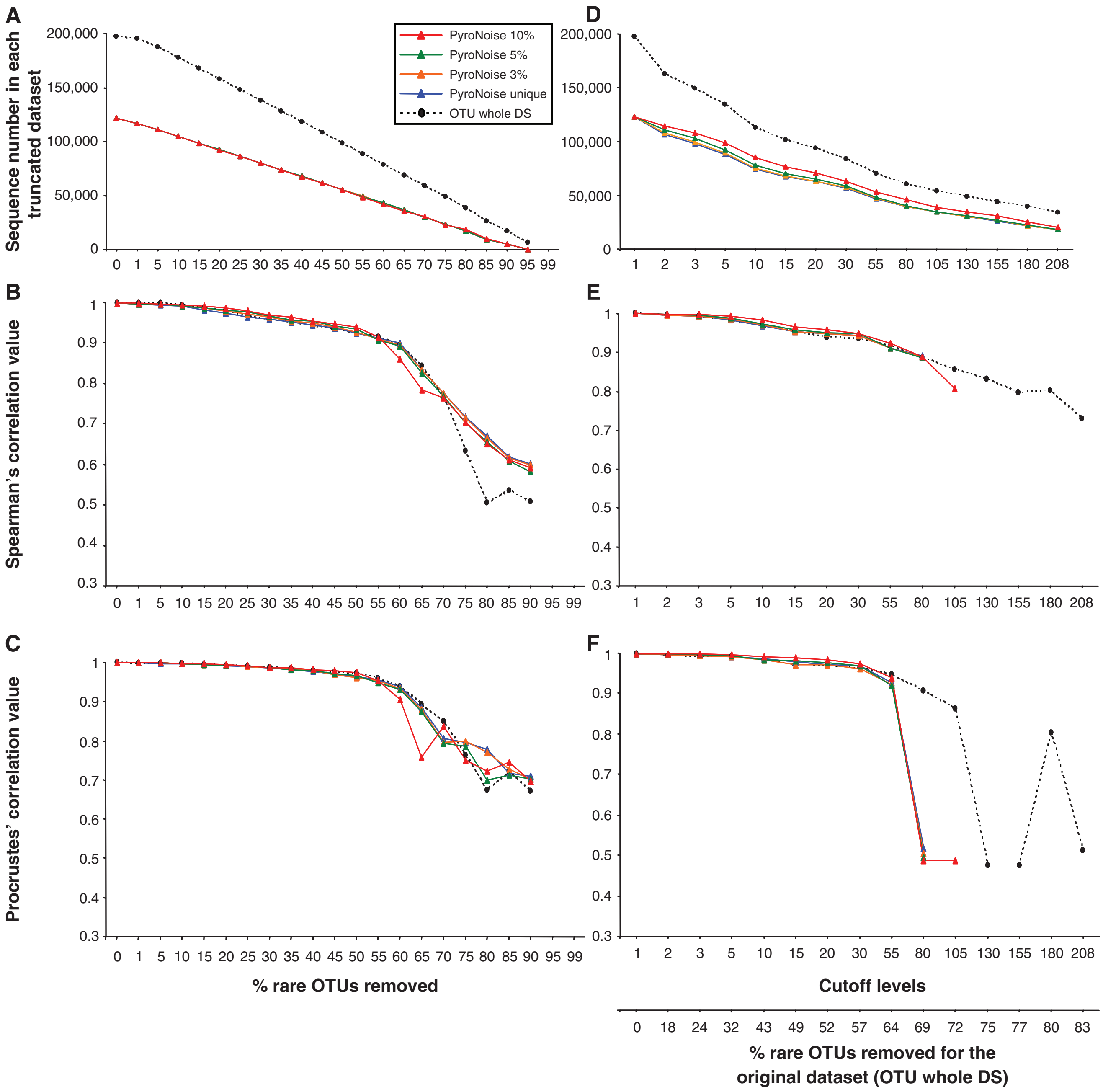


Figure 4. MultiCoLA profiles for data set structure and most important axes of extracted variation based on the data set (**A–C**) and sample (**D–F**) cutoff approaches for PyroNoise-corrected 454 MPTS data and the original 454 MPTS data set at the OTU level. Different colored lines indicate PyroNoise-corrected data sets whose sequences were further clustered at various sequence dissimilarity values. See Figure 3 for further details.

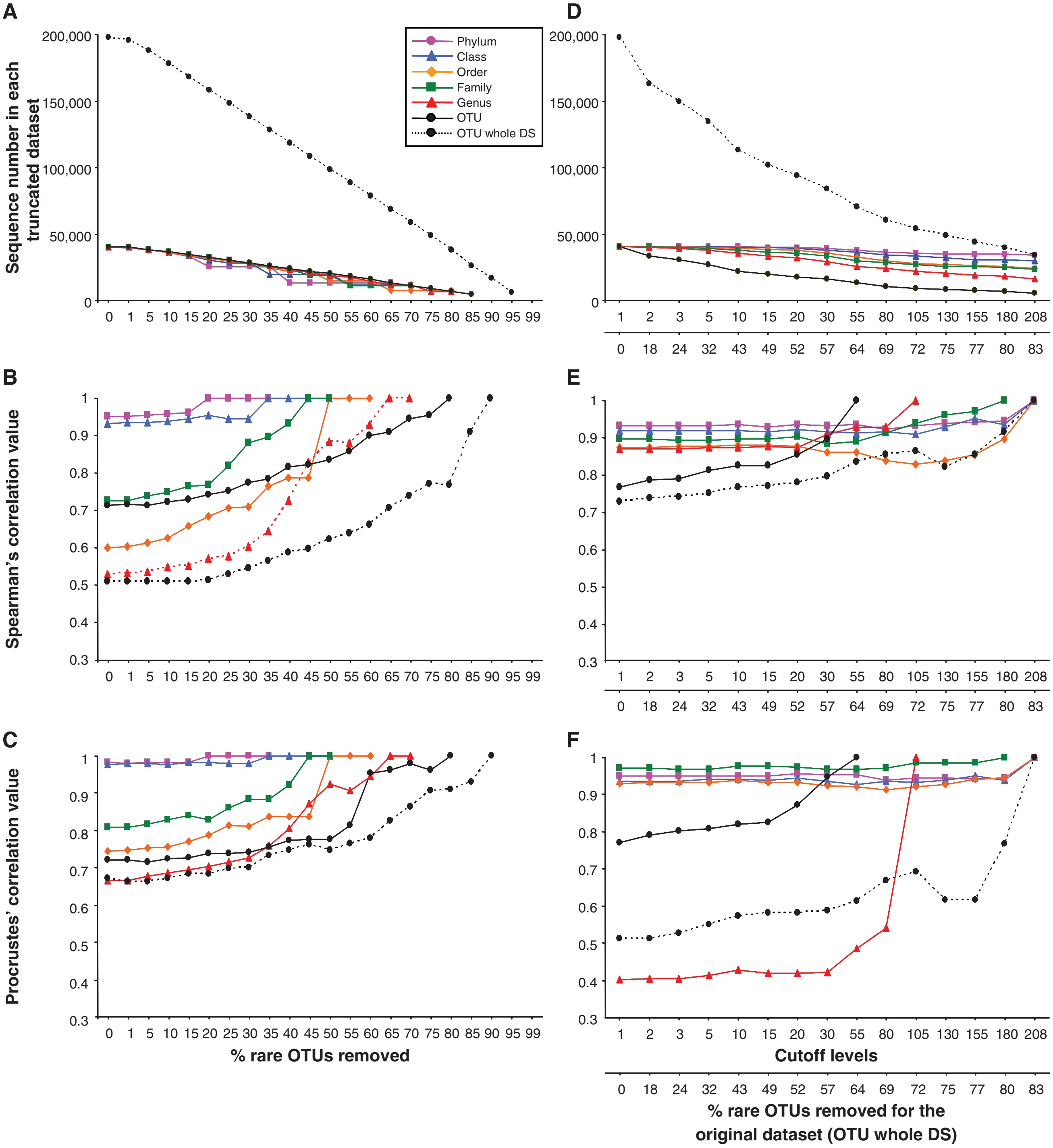


Figure 5. MultiCoLA profiles using the matrix with the most abundant OTUs as a reference for the comparison with the truncated matrices. (**A–C**) are based on the data set-based approach and (**D–F**) on the sample-based approach. See Figure 3 for further descriptions of each panel.

community structure as explained by nutrients and cell abundance (Figure 3D and H) were qualitatively similar to those based on the data set approach. More variation was again explained at higher taxonomic levels (Supplementary Figure S2 and Tables S8–S14). Therefore, choosing the sample- or data set- based approach would lead to the same ecological conclusions, despite their contrasting effects on data structure and amount of extracted ecological variation.

Because sequencing and PCR noise may generate spurious, low-abundance types, especially in high-throughput sequencing data sets (6), two strategies have been proposed to correct for sequence artifacts: a clustering threshold at 97% sequence identity (7) or a flowgram-based preclustering algorithm (8). A central question is therefore whether the afore-described variation observed in MultiCoLA profiles could be due to the presence of sequence artifacts. When MultiCoLA was applied to PyroNoise-corrected data (Supplementary Table S15), both the data set-based (Figure 4A–C) and sample-based (Figure 4D–F) approaches produced very similar profiles as those obtained with uncorrected data. The main differences consisted of generally less fluctuations in the profiles and of higher cutoff levels of 55–60% (i.e. 30–55 individual sequence abundance in the data set) that should be reached to drastically deviate from the signal in the original data set. Explanation of the community variation by additional environmental parameters yielded the same conclusions as with uncorrected data (Supplementary Figure S3). Therefore, we can conclude that the observed variations in profiles at different cutoff and taxonomic levels were mostly due to non-technical fluctuations in the data, i.e. to real structural and ecological characteristics of the studied data sets.

In this study, the original data set was used as reference for the MultiCoLA profiles, because usually one wants to remove only a small fraction of the data. Yet, it is also possible to choose the table of the most abundant types as reference for comparisons, so as to assess the effects of an increasing amount of rare types in the data set. By doing so (Figure 5), different profiles and fluctuation patterns could be observed, indicating a significant impact of the addition of rare types on data structure and ecological interpretation. Another possibility of analysis is to systematically remove the abundant fraction from each truncated data set and thus only retain the rare types (Supplementary Figure S4). This approach mimics the addition of an increasing amount of dominant types in the data set, and would enable a characterization of the data structure and ecological patterns, or lack of, present within the rare fraction of any data set. The resulting profiles and patterns (Supplementary Figure S4) were different from those obtained by systematically keeping the dominant fractions (Figure 3), suggesting that the rare fraction has a different structure and ecological signal than the more dominant fraction of the community. This observation opens the door to many new questions, but their exploration would go beyond the scope of the current study. In any case, these observations exemplify the usefulness of MultiCoLA to generate new knowledge about the nature of rarity in data sets.

In conclusion, MultiCoLA enables a systematic and data-driven exploration of the impact of rarity or dominance of specific fractions of large community data sets and on their further ecological interpretations. This would be especially useful for data sets containing a large fraction of singletons, as found in previous high-throughput Sanger sequencing data sets [e.g. from clone libraries (25) or shotgun sequencing libraries (26)], and in ongoing, high-throughput 16S rRNA-based pyrosequencing projects [e.g. the International Census of Marine Microbes (ICoMM) (5,9), http://icomm.mbl.edu], and high-throughput metagenomic projects [e.g. the International Soil Metagenome Sequencing Consortium (Terragenome) (27), http://www.terragenome.org/; or the International Human Microbiome Consortium (IHMC) (28), http://www.human-microbiome.org/] where the rare sequence issue is generally addressed arbitrarily [e.g. a threshold of two reads was chosen to identify a gene in a human microbiome metagenomic data set (28)]. This analytical approach will also help scientists to move beyond the debate of sequence accuracy and in the future, it would be particularly interesting to determine how the threshold range of profile stability varies as a function of sequencing strategy, data set sizes, samples or habitat types.

The MultiCoLA software with its respective manual and examples are available at: http://www.ecology-research.com.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We acknowledge Antje Boetius for helpful suggestions and discussions, Susan M. Huse and Philip R. Neal for curating and annotating the pyrotag sequences; Pierre Legendre for helpful statistical discussions.

FUNDING

This work was supported by the Marie Curie Early Stage Training fellowship Site in Marine Microbiology (MarMic EST contract MEST-CT-2004-007776 to A.G.); the International Max Planck Research School of Marine Microbiology (to A.G.); the Max Planck Society (to A.R.). The sequencing was financed by a W.M. Keck Foundation award to the Marine Biological Laboratory at Woods Hole, MA. This is a contribution to the International Census of Marine Microbes (ICoMM). Funding for open access charge: Max Planck Society.

Conflict of interest statement. None declared.

REFERENCES

1. Gauch,H.G. (1982) *Multivariate Analyses in Community Ecology*. Cambridge University Press, Cambridge.
2. Prendergast,J.R., Quinn,R.M., Lawton,J.H., Eversham,B.C. and Gibbons,D.W. (1993) Rare species, the coincidence of diversity hotspots and conservation strategies. *Nature*, **365**, 335–337.
3. Magurran,A.E. and Henderson,P.A. (2003) Explaining the excess of rare species in natural species abundance distributions. *Nature*, **422**, 714–716.
4. Pedrós-Alió,C. (2006) Marine microbial diversity: can it be determined? *Trends Microbiol.*, **14**, 257–263.
5. Sogin,M.L., Morrison,H.G., Huber,J.A., Welch,D.M., Huse,S.M., Neal,P.R., Arrieta,J.M. and Herndl,G.J. (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proc. Natl Acad. Sci. USA*, **103**, 12115–12120.
6. Quinlan,A.R., Stewart,D.A., Stromberg,M.P. and Marth,G.T. (2008) Pyrobayes: an improved base caller for SNP discovery in pyrosequences. *Nat. Methods*, **5**, 179–181.
7. Kunin,V., Engelbrektson,A., Ochman,H. and Hugenholtz,P. (2010) Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environ. Microbiol.*, **12**, 118–123.
8. Quince,C., Lanzen,A., Curtis,T.P., Davenport,R.J., Hall,N., Head,I.M., Read,L.F. and Sloan,W.T. (2009) Accurate determination of microbial diversity from 454 pyrosequencing data. *Nat. Methods*, **6**, 639–641.

9. Galand,P.E., Casamayor,E.O., Kirchman,D.L. and Lovejoy,C. (2009) Ecology of the rare microbial biosphere of the Arctic Ocean. *Proc. Natl Acad. Sci. USA*, **106**, 22427–22432.
10. Böer,S.I., Hedtkamp,S.I.C., van Beusekom,J.E.E., Fuhrman,J.A., Boetius,A. and Ramette,A. (2009) Time- and sediment depth-related variations in bacterial diversity and community structure in subtidal sands. *ISME J.*, **3**, 780–791.
11. Bray,J.R. and Curtis,J.T. (1957) An ordination of the upland forest communities of Southern Wisconsin. *Ecol. Monogr.*, **27**, 326–349.
12. Kendall,M.G. (1949) Rank and product-moment correlation. *Biometrika*, **36**, 177–193.
13. Shepard,R.N. (1966) Metric structures in ordinal data. *J. Math. Psychol.*, **3**, 287–315.
14. Ramette,A. (2007) Multivariate analyses in microbial ecology. *FEMS Microbiol. Ecol.*, **62**, 142–160.
15. Gower,J.C. (1966) Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, **53**, 325–338.
16. Peres-Neto,P.R. and Jackson,D.A. (2001) How well do multivariate data sets match? The advantages of a Procrustean superimposition approach over the Mantel test. *Oecologia*, **129**, 169–178.
17. Legendre,L. and Legendre,P. (1998) *Numerical Ecology*, Elsevier Science BV, Amsterdam, The Netherlands.
18. Legendre,P., Borcard,D. and Peres-Neto,P.R. (2005) Analyzing beta diversity: partitioning the spatial variation of community composition data. *Ecol. Monogr.*, **75**, 435–450.
19. Legendre,P. and Gallagher,E.D. (2001) Ecologically meaningful transformations for ordination of species data. *Oecologia*, **129**, 271–280.
20. Ramette,A. and Tiedje,J.M. (2007) Multiscale responses of microbial life to spatial distance and environmental heterogeneity in a patchy ecosystem. *Proc. Natl Acad. Sci. USA*, **104**, 2761–2766.
21. Borcard,D., Legendre,P. and Drapeau,P. (1992) Partialling out the spatial component of ecological variation. *Ecology*, **73**, 1045–1055.
22. R Development Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.
23. Oksanen,J., Kindt,R., Legendre,P., O'Hara,B., Simpson,G.L., Solymos,P., Stevens,M.H.H. and Wagner,H. (2009) vegan: Community Ecology Package. R package version 1.15-2. http://CRAN.R-project.org/package = vegan.
24. Venables,W.N. and Ripley,B.D. (2002) Modern applied statistics with S. Fourth Edition. Springer, New York, ISBN 0-387-95457-0.
25. Ley,R.E., Backhed,F., Turnbaugh,P., Lozupone,C.A., Knight,R.D. and Gordon,J.I. (2005) Obesity alters gut microbial ecology. *Proc. Natl Acad. Sci. USA*, **102**, 11070–11075.
26. Venter,J.C., Remington,K., Heidelberg,J.F., Halpern,A.L., Rusch,D., Eisen,J.A., Wu,D.Y., Paulsen,I., Nelson,K.E., Nelson,W. *et al.* (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, **304**, 66–74.
27. Vogel,T.M., Simonet,P., Jansson,J.K., Hirsch,P.R., Tiedje,J.M., van Elsas,J.D., Bailey,M.J., Nalin,R. and Philippot,L. (2009) TerraGenome: a consortium for the sequencing of a soil metagenome. *Nat. Rev. Microbiol.*, **7**, 252.
28. Qin,J.J., Li,R.Q., Raes,J., Arumugam,M., Burgdorf,K.S., Manichanh,C., Nielsen,T., Pons,N., Levenez,F., Yamada,T. *et al.* (2010) A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, **464**, 59–65.

3.2 Supplementary Information

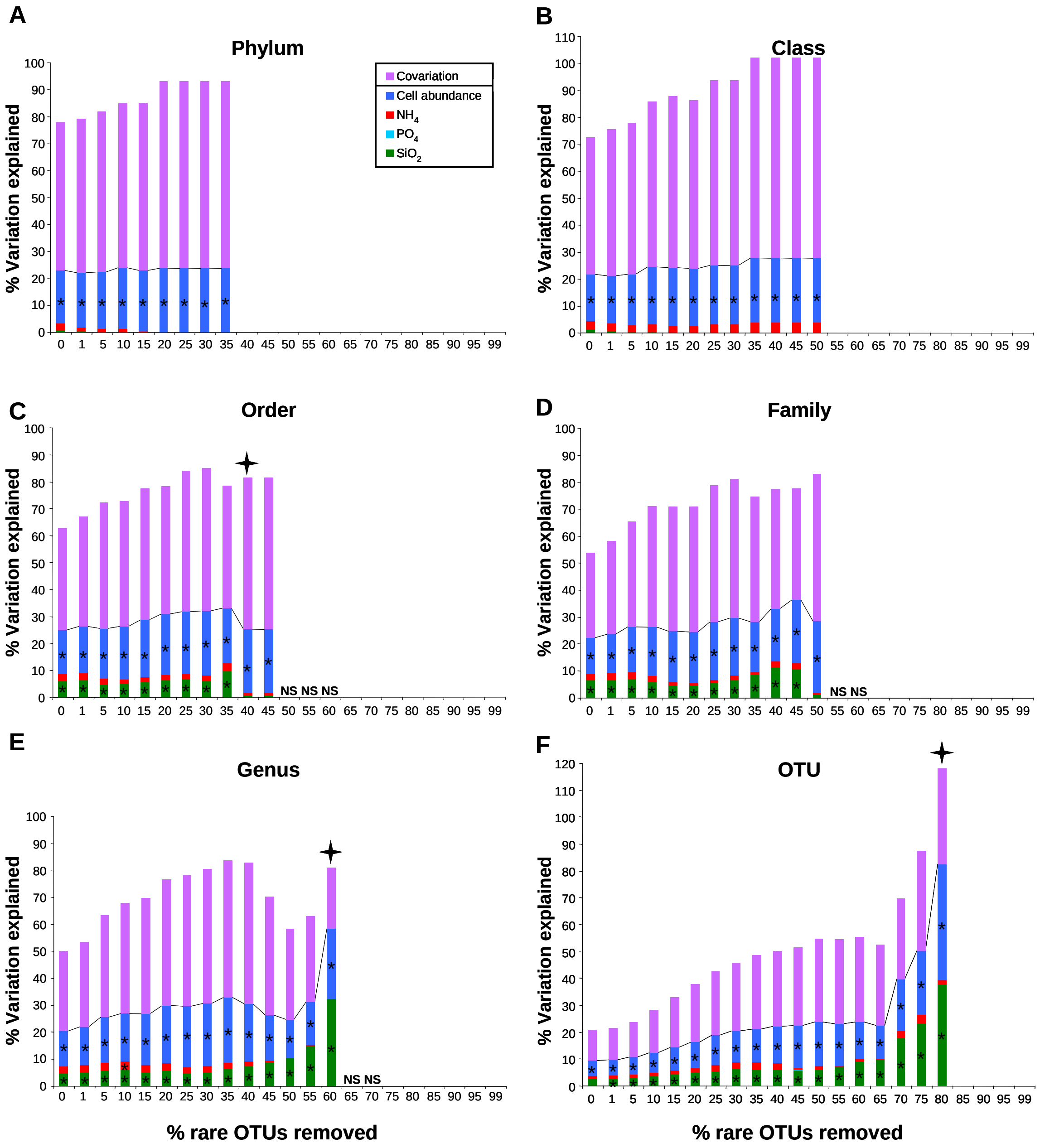
3.2.1 Supplementary Figures

Supplementary Figure 1. MultiCoLA profiles of biological variation with the dataset-based cutoff approach.

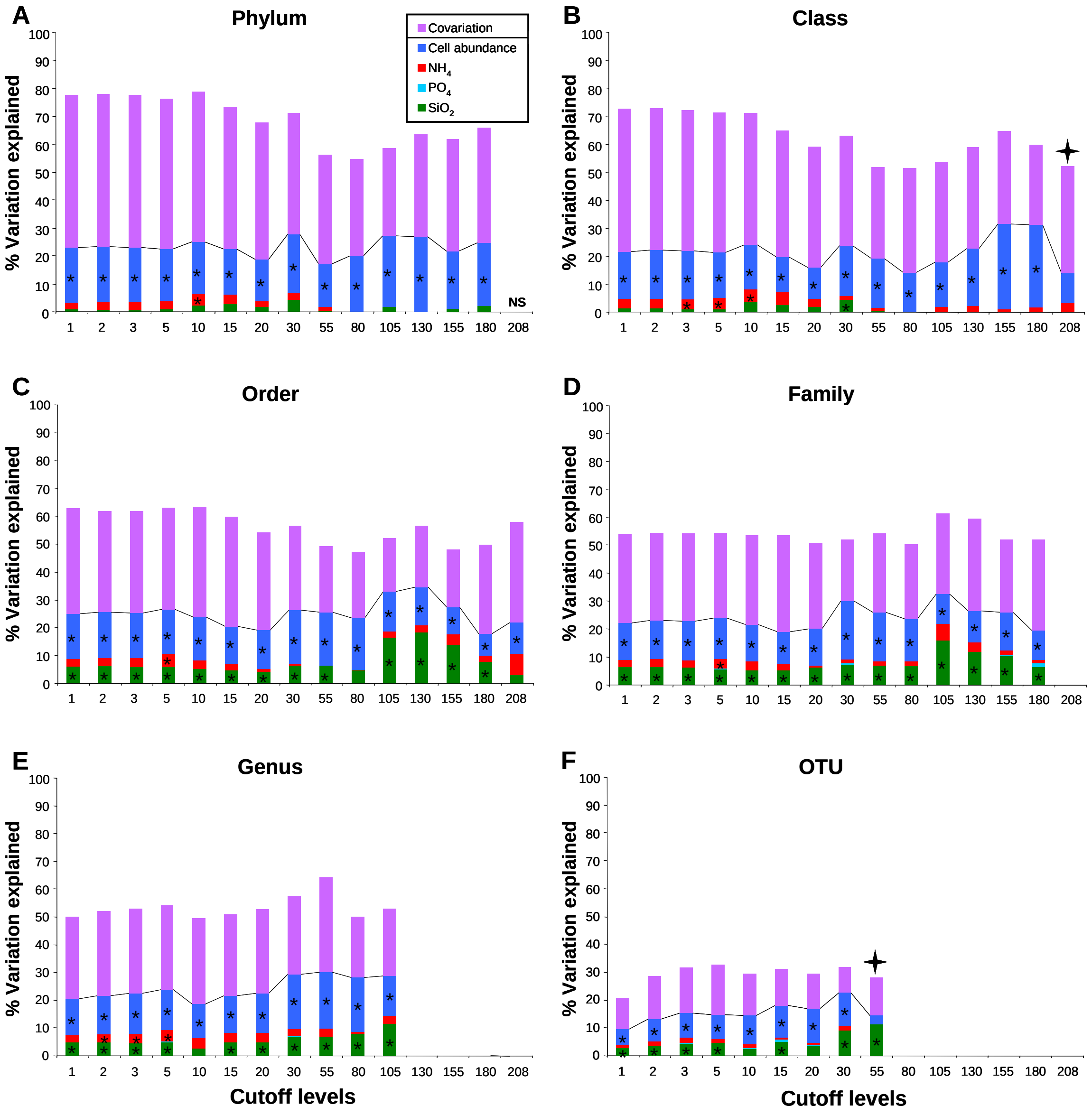
Supplementary Figure 2. MultiCoLA profiles of biological variation with the sample-based cutoff approach.

Supplementary Figure 3. MultiCoLA profiles of biological variation with the dataset-based and sample-based approaches on PyroNoise corrected data.

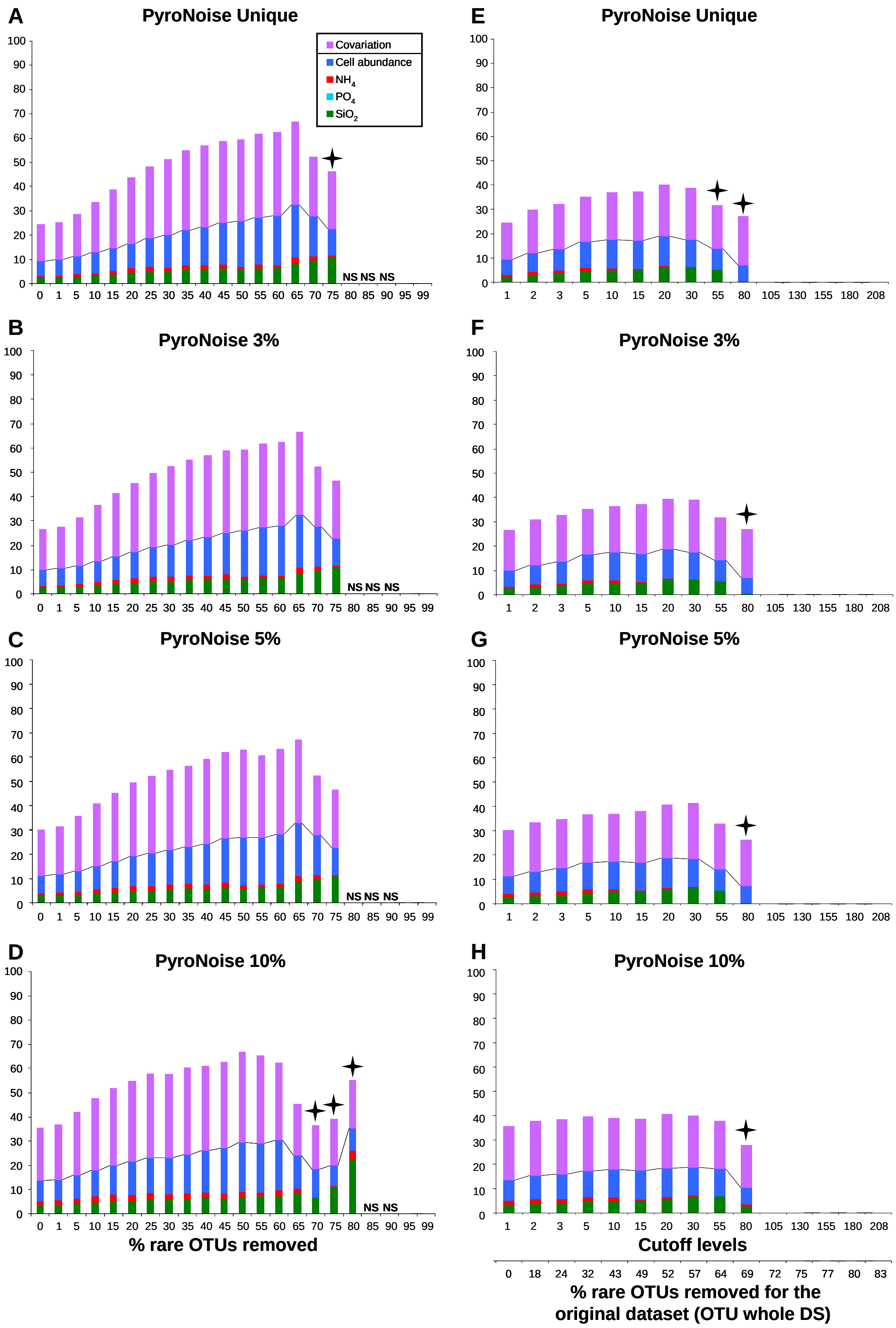
Supplementary Figure 4. MultiCoLA profiles based on the dataset- (A, B, C) and sample-based (D, E, F) cutoff approaches only retaining the rare OTU in each truncated dataset. (A, D) Abundance of rare OTU in each truncated dataset at the phylum, class, order, family, genus and OTU levels. A black solid line indicates comparisons at the OTU level for the dataset with a complete annotation and a black dashed line indicates the OTU level for the whole dataset (OTU whole DS). (B, E) Non-parametric Spearman correlations comparing the deviation in complete data structure between the original matrix and truncated matrices. (C, F) Comparison of most important axes of extracted variation between the original and truncated datasets. Lacking points are due to sample loss by applying a given cutoff to the original dataset. In the panels D, E, F, the upper x-axis corresponds to cutoff levels defined as a function of the sample-based approach, and the lower x-axis represents the corresponding proportion of removed sequences in the OTU dataset (all OTU). This enables the comparison of the dataset-based approach with the sample-based approach. ODS, original dataset.



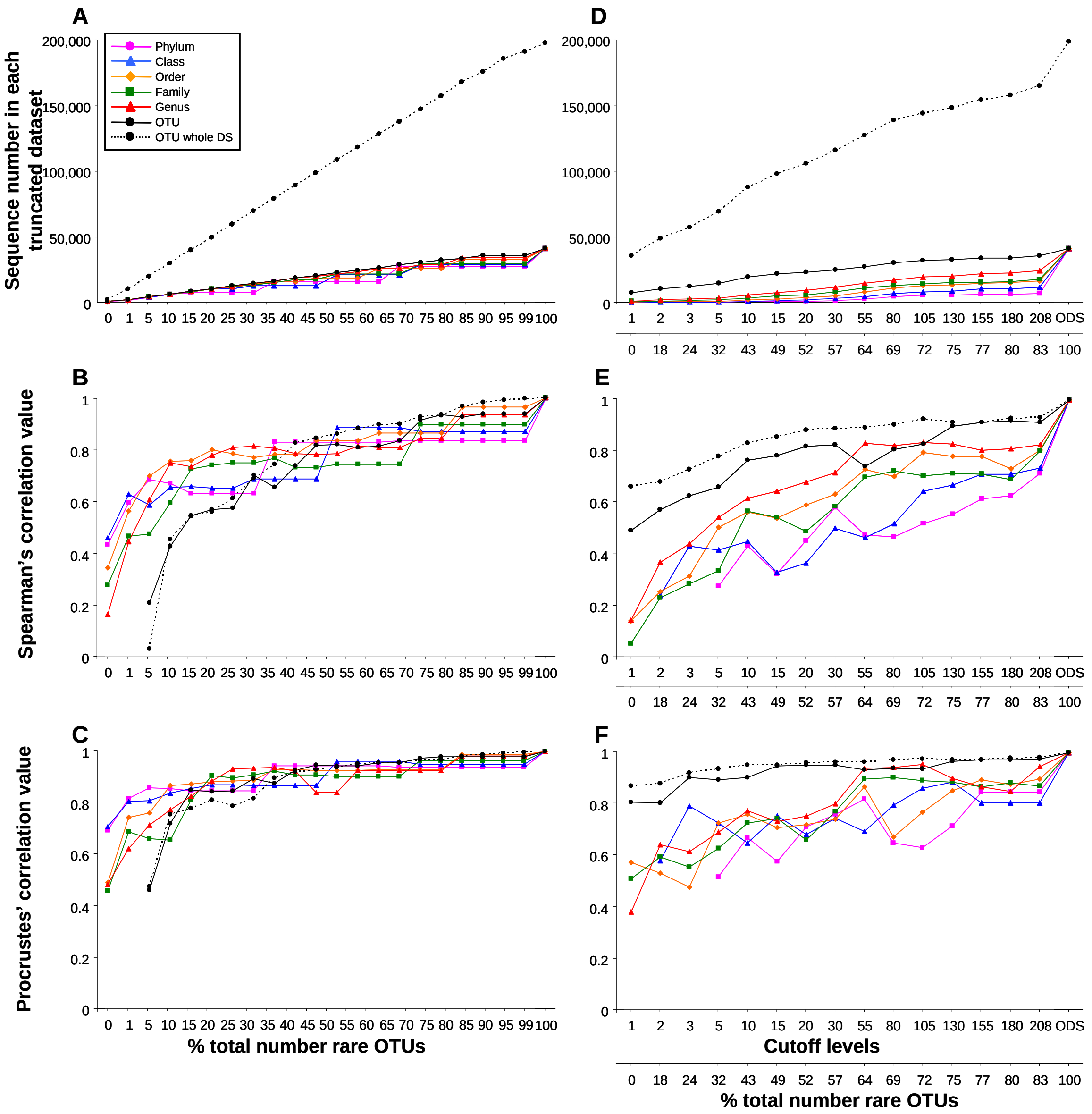
Supplementary Figure 1. MultiCoLA profiles of biological variation with the dataset-based cutoff approach.. Partitioning of the biological variation at the (A) phylum, (B) class, (C) order, (D) family, (E) genus and (F) OTU levels for the dataset with a complete annotation, into the respective effects of environmental factors (nutrients and cell abundance). Negative values, unexplained variation and non-significant models are not shown. SiO2, silicate; PO4, phosphate; NH4, ammonium; Covariation of any of the 4 environmental factors is represented under the same category. A star indicates a significant effect of the pure factors (P<5%), whereas “NS” indicates non-significant models. A cross indicates non-significant Bonferroni corrected models. Absence of data (lacking bar) is due to sample loss by applying a given cutoff to the original dataset.



Supplementary Figure 2. MultiCoLA profiles of biological variation with the sample-based cutoff approach. See Supplementary Figure 1 for details.


A
PyroNoise Unique
Covariation
Cell abundance
NH_4
PO_4
SiO_2
100
90
80
70
60
50
40
30
20
10
0
0 1 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 99
NS NS NS
B
PyroNoise 3%
NS NS NS
C
PyroNoise 5%
NS NS NS
D
PyroNoise 10%
NS NS
% rare OTUs removed
E
PyroNoise Unique
1 2 3 5 10 15 20 30 55 80 105 130 155 180 208
F
PyroNoise 3%
G
PyroNoise 5%
H
PyroNoise 10%
Cutoff levels
0 18 24 32 43 49 52 57 64 69 72 75 77 80 83
% rare OTUs removed for the original dataset (OTU whole DS)

Supplementary Figure 3. MultiCoLA profiles of biological variation with the dataset-based (A, B, C, D) and sample-based (E, F, G, H) approaches for PyroNoise corrected data. For the sample-based approach panel H, the upper x-axis corresponds to cutoff levels defined as a function of the sample-based approach (as for panels E, F, G), and the lower x-axis represents the corresponding proportion of removed sequences in the OTU dataset (all OTU). This enables the comparison of the sample-based with dataset-based approach. Each panel consists of PyroNoise-corrected datasets whose sequences were clustered at various sequence dissimilarity levels (0-10%). See Supplementary Figure 1 for further details.



Supplementary Figure 4. MultiCoLA profiles based on the dataset- (A, B, C) and sample-based (D, E, F) cutoff approaches only retaining the rare OTU in each truncated dataset. (**A**, **D**) Abundance of rare OTU in each truncated dataset at the phylum, class, order, family, genus and OTU levels. A black solid line indicates comparisons at the OTU level for the dataset with a complete annotation and a black dashed line indicates the OTU level for the whole dataset (OTU whole DS). (**B**, **E**) Non-parametric Spearman correlations comparing the deviation in complete data structure between the original matrix and truncated matrices. (**C**, **F**) Comparison of most important axes of extracted variation between the original and truncated datasets. Lacking points are due to sample loss by applying a given cutoff to the original dataset. In the panels **D**, **E**, **F**, the upper x-axis corresponds to cutoff levels defined as a function of the sample-based approach, and the lower x-axis represents the corresponding proportion of removed sequences in the OTU dataset (all OTU). This enables the comparison of the dataset-based approach with the sample-based approach. ODS, original dataset.

3.2.2 Supplementary Tables

Supplementary Table 1. Contribution of environmental parameters to the variation in truncated datasets (dataset-based approach), at the OTU level for all sequences available.

Supplementary Table 2. Contribution of environmental parameters to the variation in truncated datasets (dataset-based approach), at the OTU level for the dataset with a complete annotation.

Supplementary Table 3. Contribution of environmental parameters to the variation in truncated datasets (dataset-based approach), at the Genus level.

Supplementary Table 4. Contribution of environmental parameters to the variation in truncated datasets (dataset-based approach), at the Family level.

Supplementary Table 5. Contribution of environmental parameters to the variation in truncated datasets (dataset-based approach), at the Order level.

Supplementary Table 6. Contribution of environmental parameters to the variation in truncated datasets (dataset-based approach), at the Class level.

Supplementary Table 7. Contribution of environmental parameters to the variation in truncated datasets (dataset-based approach), at the Phylum level.

Supplementary Table 8. Contribution of environmental parameters to the variation in truncated datasets (sample-based approach), at the OTU level for all sequences available.

Supplementary Table 9. Contribution of environmental parameters to the variation in truncated datasets (sample-based approach), at the OTU level for the dataset with a complete annotation.

Supplementary Table 10. Contribution of environmental parameters to the variation in truncated datasets (sample-based approach), at the Genus level.

Supplementary Table 11. Contribution of environmental parameters to the variation in truncated datasets (sample-based approach), at the Family level.

Supplementary Table 12. Contribution of environmental parameters to the variation in truncated datasets (sample-based approach), at the Order level.

Supplementary Table 13. Contribution of environmental parameters to the variation in truncated datasets (sample-based approach), at the Class level.

Supplementary Table 14. Contribution of environmental parameters to the variation in truncated datasets (sample-based approach), at the Phylum level.

Supplementary Table 15. Summary of OTU numbers after PyroNoise correction of the 454 MPTS dataset.

Supplementary Table 1. Contribution of environmental parameters to the variation in truncated datasets (dataset-based approach), **at the OTU level for all sequences available.**

Cutoff levels[a]	Total number of sequences	$R^{2,b}$	Individual factor contribution[c]		
			Cell	Nutrients	
			abundance	SiO_2	NH_4
0%	197,684	17%***	-0.97		
1%	195,707	17%***	-0.97		
5%	187,799	19%***	-0.97		
10%	177,915	23%***	-0.97	0.42	
15%	168,029	27%***	-0.96	0.42	
20%	158,143	31%***	-0.96	0.42	
25%	148,258	34%***	-0.96	0.43	
30%	138,377	38%***	-0.96	0.43	
35%	128,459	40%***	-0.97	0.43	
40%	118,590	42%***	-0.97	0.43	
45%	108,709	44%***	-0.97	0.43	
50%	98,762	44%***	-0.97	0.41	
55%	88,871	45%***	-0.97	0.38	
60%	78,951	45%***	-0.98		
65%	68,789	41%***	-0.97	0.35	
70%	58,637	39%***	0.95	-0.26	
75%	49,110	22%**	0.96		
80%	38,299	NS			
85%	25,961	NS			
90%	16,852	NS			
95%	6,550				
99%	0				

[a] Cutoff levels were defined based on the whole dataset strategy (see Supplementary Fig. 1). Cutoff levels were applied until samples were lost due to lack of sequences.

[b] Adjusted R^2 indicate the amount of variation explained by cell abundance and nutrients (SiO2, silicate; NH4, ammonium), their significance is indicated as NS (non significant), * ($P \leq 0.05$), ** ($P \leq 0.01$), and *** ($P \leq 0.001$).

[c] Only significant, standardized correlation coefficients to the first redundancy analysis (RDA) axis are indicated for each parameter.

Supplementary Table 2. Contribution of environmental parameters to the variation in truncated datasets (dataset-based approach), **at the OTU level for the dataset with a complete annotation.**

Cutoff levels[a]	Total number of sequences	$R^{2, b}$	Individual factor contribution[c]		
			Cell abundance	Nutrients	
				SiO_2	NH_4
0%	40,660	16%***	-0.96		
1%	40,253	16%***	-0.96	0.41	
5%	38,627	18%***	-0.96	0.41	
10%	36,593	21%***	-0.96	0.41	
15%	34,560	25%***	-0.96	0.4	
20%	32,524	29%***	-0.96	0.4	
25%	30,489	32%***	-0.96	0.41	
30%	28,453	35%***	-0.96	0.44	
35%	26,391	38%***	-0.96	0.46	
40%	24,371	38%***	-0.97	0.43	
45%	22,315	40%***	-0.97	0.44	
50%	20,267	41%***	-0.97	0.39	
55%	18,234	39%***	-0.96	0.31	
60%	16,106	38%***	-0.89	0.1	
65%	13,731	31%**	-0.6	-0.28	
70%	11,332	41%***	-0.29	-0.58	
75%	8,884	54%***	-0.35	-0.55	
80%	7,181	58%**	-0.54	-0.39	
85%	5,165				
90%	0				
95%	0				
99%	0				

For detailed explanations, see Supplementary Table 1.

Supplementary Table 3. Contribution of environmental parameters to the variation in truncated datasets (dataset-based approach), **at the Genus level.**

Cutoff levels[a]	**Total number of sequences**	**R^{2,b}**	**Individual factor contribution**[c]		
			Cell abundance	**Nutrients**	
				SiO$_2$	**NH$_4$**
0%	40,660	38%***	-0.96	0.43	
1%	40,251	41%***	-0.96	0.43	
5%	38,600	48%***	-0.96	0.44	
10%	36,569	52%***	-0.96	0.44	0.58
15%	34,549	53%***	-0.97	0.42	
20%	32,374	58%***	-0.97	0.41	
25%	30,336	59%***	-0.97	0.42	
30%	28,367	61%***	-0.97	0.42	
35%	25,952	62%***	-0.97	0.4	
40%	23,612	60%***	-0.97	0.39	
45%	21,013	49%***	-0.97	0.34	
50%	19,464	39%***	-0.99	0.41	
55%	17,702	43%***	-0.95	0.65	
60%	14,982	36%*	-0.42	-0.51	
65%	12,129	NS			
70%	12,129	NS			
75%	7,029				
80%	7,029				
85%	0				
90%	0				
95%	0				
99%	0				

For detailed explanations, see Supplementary Table 1.

Supplementary Table 4. Contribution of environmental parameters to the variation in truncated datasets (dataset-based approach), **at the Family level.**

			Individual factor contribution[c]		
Cutoff levels[a]	**Total number of sequences**	**R^{2,b}**	**Cell abundance**	**Nutrients**	
				SiO$_2$	**NH$_4$**
0%	40,660	41%***	-0.97	0.47	
1%	40,246	44%***	-0.97	0.47	
5%	38,555	50%***	-0.97	0.48	
10%	36,534	53%***	-0.97	0.48	
15%	34,494	53%***	-0.98	0.5	
20%	32,441	53%***	-0.98	0.49	
25%	30,201	59%***	-0.98	0.47	
30%	28,455	61%***	0.97	-0.46	
35%	26,140	56%***	0.97	-0.59	
40%	24,378	58%***	0.93	-0.65	
45%	22,143	59%***	0.96	-0.55	
50%	19,290	62%***	0.96		
55%	15,089	NS			
60%	15,089	NS			
65%	8,059				
70%	8,059				
75%	8,059				
80%	8,059				
85%	0				
90%	0				
95%	0				
99%	0				

For detailed explanations, see Supplementary Table 1.

Supplementary Table 5. Contribution of environmental parameters to the variation in truncated datasets (dataset-based approach), **at the Order level.**

Cutoff levels[a]	**Total number of sequences**	**R^{2,b}**	**Individual factor contribution**[c]		
			Cell abundance	**Nutrients**	
				SiO_2	**NH_4**
0%	40,660	47%***	-0.97	0.46	
1%	40,217	50%***	-0.97	0.47	
5%	38,446	54%***	-0.98	0.5	
10%	36,412	54%***	-0.98	0.49	
15%	34,218	58%***	-0.98	0.49	
20%	32,241	59%***	0.98	-0.47	
25%	29,586	63%***	0.97	-0.48	
30%	28,379	64%***	-0.97	0.51	
35%	24,963	60%***	-0.91	0.66	
40%	22,110	63%**	0.94		
45%	22,110	63%**	0.94		
50%	15,089	NS			
55%	15,089	NS			
60%	15,089	NS			
65%	8,059				
70%	8,059				
75%	8,059				
80%	8,059				
85%	0				
90%	0				
95%	0				
99%	0				

For detailed explanations, see Supplementary Table 1.

Supplementary Table 6. Contribution of environmental parameters to the variation in truncated datasets (dataset-based approach), **at the Class level.**

Cutoff levels[a]	**Total number of sequences**	**$R^{2, b}$**	**Individual factor contribution**[c]		
			Cell	**Nutrients**	
			abundance	**SiO_2**	**NH_4**
0%	40,660	55%***	-0.97		
1%	40,127	56%***	-0.97		
5%	38,297	58%***	-0.97		
10%	36,176	64%***	-0.97		
15%	34,013	66%***	-0.97		
20%	30,704	65%***	-0.96		
25%	28,065	71%***	-0.94		
30%	28,065	71%***	-0.94		
35%	20,006	78%***	-0.94		
40%	20,006	78%***	-0.94		
45%	20,006	78%***	-0.94		
50%	20,006	78%***	-0.94		
55%	11,937				
60%	11,937				
65%	11,937				
70%	11,937				
75%	0				
80%	0				
85%	0				
90%	0				
95%	0				
99%	0				

For detailed explanations, see Supplementary Table 1.

Supplementary Table 7. Contribution of environmental parameters to the variation in truncated datasets (dataset-based approach), **at the Phylum level.**

Cutoff levels[a]	**Total number of sequences**	**$R^{2,b}$**	**Individual factor contribution**[c]		
			Cell abundance	**Nutrients**	
				SiO_2	**NH_4**
0%	40,660	60%***	-0.97		
1%	40,045	60%***	-0.97		
5%	38,600	62%***	-0.97		
10%	36,242	65%***	-0.97		
15%	33,575	67%***	-0.93		
20%	25,516	73%***	-0.93		
25%	25,516	73%***	-0.93		
30%	25,516	73%***	-0.93		
35%	25,516	73%***	-0.93		
40%	13,579				
45%	13,579				
50%	13,579				
55%	13,579				
60%	13,579				
65%	13,579				
70%	0				
75%	0				
80%	0				
85%	0				
90%	0				
95%	0				
99%	0				

For detailed explanations, see Supplementary Table 1.

Supplementary Table 8. Contribution of environmental parameters to the variation in truncated datasets (sample-based approach), **at the OTU level for all sequences available.**

Cutoff levels[a]	**Total number of sequences**	**$R^{2,b}$**	**Individual factor contribution**[c]		
			Cell abundance	**Nutrients**	
				SiO_2	**NH_4**
1	197,684	17%***	-0.97		
2	162,639	23%***	-0.97		
3	149,499	25%***	-0.97	0.43	
5	134,258	27%***	-0.97	0.43	
10	113,172	27%***	-0.98		
15	101,666	27%***	-0.98		
20	94,126	27%***	-0.97	0.38	
30	84,180	28%***	-0.96	0.30	
55	70,118	28%***	-0.97	0.36	
80	60,612	20%**	-0.97		
105	54,317	16%*	-0.98		
130	49,187	13%*			
155	44,384	NS			
180	39,872	NS			
208	33,931	NS			

[a] Cutoff levels were defined based on the sample-based strategy (see Supplementary Fig. 1). Cutoff levels were applied until samples were lost due to lack of sequences.

[b] Adjusted R^2 indicate the amount of variation explained by cell abundance and nutrients (SiO2, silicate; NH4, ammonium), their significance is indicated as NS (non significant), * ($P \leq 0.05$), ** ($P \leq 0.01$), and *** ($P \leq 0.001$).

[c] Only significant, standardized correlation coefficients to the first redundancy analysis (RDA) axis are indicated for each parameter.

Supplementary Table 9. Contribution of environmental parameters to the variation in truncated datasets (sample-based approach), **at the OTU level for the dataset with a complete annotation.**

Cutoff levels[a]	**Total number of sequences**	**R^{2,b}**	**Individual factor contribution**[c]		
			Cell	**Nutrients**	
			abundance	**SiO_2**	**NH_4**
1	40,660	16%***	-0.96	0.41	
2	33,533	22%***	-0.96	0.42	
3	30,655	25%***	-0.97	0.44	
5	27,352	25%***	-0.97	0.48	
10	22,164	23%***	-0.98		
15	19,591	25%***	-0.98	0.39	
20	18,010	24%***	-0.97		
30	16,113	23%***	-0.76	-0.13	
55	13,314	17%**		-0.80	
80	10,761				
105	9,256				
130	8,331				
155	7,491				
180	6,982				
208	5,436				

For detailed explanations, see Supplementary Table 8.

Supplementary Table 10. Contribution of environmental parameters to the variation in truncated datasets (sample-based approach), **at the Genus level.**

Cutoff levels[a]	Total number of sequences	$R^{2,b}$	Individual factor contribution[c]		
			Cell	Nutrients	
			abundance	SiO_2	NH_4
1	40,660	38%***	-0.96	0.43	
2	39,800	40%***	-0.96	0.42	0.59
3	39,114	41%***	-0.95	0.43	0.61
5	37,979	42%***	-0.95	0.43	0.63
10	35,585	36%***	-0.94		
15	33,526	38%***	-0.95	0.41	
20	31,993	40%***	-0.96	0.42	
30	29,405	44%***	-0.95	0.36	
55	25,952	46%***	0.95	-0.31	
80	24,252	38%***	0.98	-0.37	
105	21,847	43%***	-0.99	0.54	
130	20,689				
155	19,254				
180	18,239				
208	16,501				

For detailed explanations, see Supplementary Table 8.

Supplementary Table 11. Contribution of environmental parameters to the variation in truncated datasets (sample-based approach), **at the Family level.**

Cutoff levels[a]	**Total number of sequences**	**$R^{2,b}$**	**Individual factor contribution**[c]		
			Cell abundance	**Nutrients**	
				SiO_2	**NH_4**
1	40,660	41%***	-0.97	0.47	
2	40,233	42%***	-0.96	0.45	
3	39,841	42%***	-0.96	0.45	
5	39,087	42%***	-0.96	0.44	0.60
10	37,588	39%***	-0.94	0.42	
15	36,328	40%***	-0.96	0.50	
20	35,245	38%***	-0.97	0.44	
30	33,077	41%***	-0.97	0.42	
55	29,427	39%***	-0.97	0.37	
80	27,830	37%***	-0.97	0.45	
105	26,303	49%***	-0.93	0.73	
130	25,053	45%***	-0.94	0.63	
155	24,346	40%***	-0.96	0.57	
180	23,357	38%***	-0.96	0.44	
208	21,827				

For detailed explanations, see Supplementary Table 8.

Supplementary Table 12. Contribution of environmental parameters to the variation in truncated datasets (sample-based approach), **at the Order level.**

Cutoff levels[a]	**Total number of sequences**	**R**[2, b]	**Individual factor contribution**[c]		
			Cell abundance	**Nutrients**	
				SiO_2	**NH_4**
1	40,660	47%***	-0.97	0.46	
2	40,526	47%***	-0.97	0.45	
3	40,342	47%***	-0.96	0.44	
5	39,992	48%***	-0.96	0.42	0.60
10	39,237	45%***	-0.95	0.39	
15	38,496	43%***	-0.96	0.44	
20	37,580	38%***	-0.96	0.40	
30	35,878	41%***	-0.96	0.39	
55	32,563	37%***	-0.99	0.44	
80	29,987	35%***	-0.98		
105	27,960	41%***	-0.99	0.62	
130	27,358	44%***	-0.98	0.68	
155	26,382	34%***	-0.91	0.58	
180	25,541	35%**	-0.94	0.54	
208	24,372	42%***	-0.94		

For detailed explanations, see Supplementary Table 8.

Supplementary Table 13. Contribution of environmental parameters to the variation in truncated datasets (sample-based approach), **at the Class level.**

Cutoff levels[a]	**Total number of sequences**	**$R^{2, b}$**	**Individual factor contribution**[c]		
			Cell abundance	**Nutrients**	
				SiO_2	**NH_4**
1	40,660	55%***	-0.97		
2	40,636	55%***	-0.97		
3	40,594	54%***	-0.96		0.62
5	40,495	54%***	-0.96		0.61
10	40,144	53%***	-0.96		0.61
15	39,753	48%***	-0.95		
20	39,172	42%***	-0.95		
30	37,932	48%***	-0.96	0.48	
55	36,156	39%***	-0.97		
80	34,338	31%**	-0.99		
105	33,168	36%***	-0.99		
130	32,351	39%**	-0.99		
155	30,658	50%***	-0.98		
180	30,503	46%***	-0.97		
208	29,564	34%*			

For detailed explanations, see Supplementary Table 8.

Supplementary Table 14. Contribution of environmental parameters to the variation in truncated datasets (sample-based approach), **at the Phylum level.**

Cutoff levels[a]	**Total number of sequences**	**$R^{2,b}$**	**Individual factor contribution**[c]		
			Cell	**Nutrients**	
			abundance	**SiO_2**	**NH_4**
1	40,660	60%***	-0.97		
2	40,644	60%***	-0.97		
3	40,614	59%***	-0.97		
5	40,544	58%***	-0.97		
10	40,358	59%***	-0.96		0.62
15	40,167	55%***	-0.96		
20	39,901	49%***	-0.96		
30	39,285	53%***	-0.96		
55	37,868	40%***	-0.96		
80	36,601	39%***	-0.99		
105	35,579	44%***	-0.99		
130	34,876	50%***	-0.99		
155	34,745	45%***	-0.99		
180	34,580	48%***	-0.99		
208	33,993	NS			

For detailed explanations, see Supplementary Table 8.

Supplementary Table 15. Summary of OTUs numbers after PyroNoise correction of the 454 MPTS dataset.

Sampling date		Depth layer (cm)	Read number	Filtered number	Unique sequences	3% OTUs	5% OTUs	10% OTUs
2005	February	0-5	9,526	4,553	784	763	692	591
		5-10	18,409	8,732	1,639	1,568	1,385	1,102
		10-15	7,971	3,364	1,207	1,179	1,061	876
	April	0-5	5,146	2,162	505	496	470	407
		5-10	10,733	5,740	1,487	1,436	1,314	1,113
		10-15	15,259	8,862	2,476	2,341	2,083	1,650
	July	0-5	10,690	5,970	1,068	1,036	927	763
		5-10	9,632	4,940	1,402	1,373	1,265	1,063
	November	0-5	14,526	8,950	1,539	1,487	1,326	1,083
		5-10	21,996	12,981	2,469	2,301	2,072	1,618
2006	March1	0-5	10,439	5,839	1,492	1,442	1,324	1,107
		5-10	20,182	11,463	2,643	2,473	2,181	1,737
		10-15	19,564	9,981	2,586	2,426	2,154	1,703
	March2	0-5	11,339	5,956	1,035	994	904	772
		5-10	20,866	11,941	2,133	2,036	1,808	1,467
		10-15	21,470	11,997	3,210	2,993	2,586	2,008

4 Chapter III.

Diversity and Dynamics of the Rare and Resident Bacterial Biosphere in Coastal Sands

Angélique Gobet, Simone I. Böer, Susan M. Huse, Justus E.E. van Beusekom, Christopher Quince, Mitchell L. Sogin, Antje Boetius and Alban Ramette

Submission to Proceedings of the National Academy of Science planned by December, 5th 2010.

Keywords: Bacterial diversity, 454 pyrosequencing, multivariate analysis, coastal seas, rare biosphere

Abstract. The use of pyrosequencing as a fingerprinting technique for environmental microbial communities has expanded our knowledge of the enormous diversity of Bacteria in the ocean, especially regarding the dominance of rare types (Sogin et al. 2006), and revealed non-random patterns (Galand et al. 2009a). Using 454 massively parallel tag sequencing, we have investigated fluctuations of both rare and resident bacterial types in temperate coastal sands, which represent a highly dynamic marine environment characterized by strong physical mixing and seasonal variation. About 60-70% of the bacterial types consisted of tag sequences occurring only once over a period of 1 year. Most members of the rare biosphere did not become abundant at any time or at any sediment depth, but varied significantly with environmental parameters associated with nutritional stress. Only 3-5% of all bacterial types of a given depth zone were present at all times, but 50-80% of them belonged to the most abundant types in the data set. Despite the large proportion and turnover of rare organisms, overall community patterns were driven by deterministic relationships associated with seasonal fluctuations in key biogeochemical parameters related to primary productivity. The maintenance of major biogeochemical functions throughout the observation period suggests that the small proportion of resident bacterial types in sands perform the key biogeochemical processes, while the majority of rare taxa are transient.

4.1 Introduction

Marine coastal areas represent highly dynamic ecosystems where the atmosphere, continents and the ocean interact. Permeable sands constitute the dominant sediment type on continental shelves (Emery 1968, Boudreau et al. 2001) and play a central role for global carbon and nutrient cycles. They act as biocatalytic filters for various types of materials advected by currents and winds, including dissolved and particulate organic matter derived from living and dead biomass of terrestrial or marine origin (de Beer et al. 2005). Sandy sediments are also constantly subjected to biotic (*e.g.* bioturbation) and abiotic disturbances [*e.g.* mixing by currents, seasonal and tidal temperature fluctuations, anoxia; (Boudreau et al. 2001)]. They may also host human pathogens, depending on human impact by *e.g.* recreational use of beaches, and temperature anomalies (Ruppert et al. 2004, Dinsdale et al. 2008).

Microorganisms produce extracellular polymeric substances (EPS) to attach to surfaces such as those of sand grains, forming a biofilm together with other organisms. In the physically dynamic environment of sands, a biofilm provides an ideal environment for microorganisms to thrive; *e.g.* it enables cell adhesion, cell protection from dehydration and gives an external digestive system containing extracellular enzymes and particles from various origins (Flemming & Wingender 2010). The pore volume between the sand grains represents a place of constant particle exchange due to turbulent flow of currents and pore water advection which may induce significant fluxes of particulate organic matter (Stoodley et al. 2005). It remains yet unknown whether sand grain-associated biofilms trap microbes from the water column flushing through the sand, or whether cells are carried away from the biofilms with the pore water. Indeed, it is known that bacterial cells need a strong ultrasonic treatment to be dislodged from sand grains (Epstein et al. 1997) and that less than 0.2% of total bacterial cells in the sand can be found in the pore water (Rusch et al. 2003).

Consequently, the high diversity of niches provided in sands could support a rich community of bacteria fluctuating with environmental variations. The dynamics of environmental parameters in coastal sands could still be a challenge to many microbial populations, selecting for few, but tolerant and well-adapted resident types. Nevertheless, the rapid transport of materials through those niches may support a high bacterial richness of transient rare types, characterized by high temporal fluctuation. One of the principal aims of this study was to test these different hypotheses.

So far, few studies have provided insights into the structuring and the ecology of microbial communities in coastal sediments, and these were mostly based on traditional culture-independent community fingerprinting approaches (Urakawa et al. 2000, Bertics & Ziebis 2009, Böer et al. 2009). While those molecular techniques have permitted a first description of the diversity of abundant populations, they generally fail to describe low abundance, rare populations that might represent considerable diversity (Acinas et al. 2004). High-throughput sequencing strategies such as shotgun sequencing (Venter et al. 2004) or pyrosequencing (Sogin et al. 2006) allow a higher resolution of the description of microbial diversity. The latter technique has already permitted the exploration of the rare microbial biosphere in several types of ecosystems such as surface and deep-sea water (Sogin et al. 2006), extreme environments (Huber et al. 2007, Brazelton et al. 2010), or the human hand surface (Fierer et al. 2008). Patterns of the rare biosphere in the Arctic Ocean have been related to differences in water masses and attributed to dispersal limitation (Galand et al. 2009a). Investigating patterns in the rare microbial biosphere is, however, difficult, because of technical noise in pyrosequencing data (Quince et al. 2009, Kunin et al. 2010) that may lead to the description of spurious patterns. Beyond correcting the raw data (Quince et al. 2009), a systematic analysis of the effects of rarity on community structure and ecological interpretation was recently proposed to test whether non-stochastic ecological signals were present for different definitions of the rare biosphere in a given data set (Gobet et al. 2010).

Here, a high-resolution description of the bacterial diversity of resident and rare types in temperate, marine sandy sediments was obtained over six sampling dates within a year (2005-2006) by applying 454 massively parallel tag sequencing (MPTS) targeting the V6 region of the 16S rRNA gene. Patterns of bacterial diversity were correlated with contextual environmental parameters to quantify the structuring effects of time, sediment depth and biogeochemical conditions on microbial community composition and turnover for the whole community as well as for fractions of different tag abundance and taxonomic resolution.

4.2 Materials and Methods

Study site, sampling procedures, contextual parameters.

Detailed sample processing and environmental measurements have been published elsewhere (Böer et al. 2009). Briefly, sediment samples were collected on a shallow subtidal sand flat off the "Hausstrand" beach on the North Sea island Sylt (55° 00'47.7''N, 8° 25'59.3''E) in February, April, July and November 2005, beginning and end of March 2006. Sandy sediment cores were sectioned right after collection in three 5 cm layers down to 15 cm. Subsamples were then stored at -20°C for DNA extraction or used for diverse environmental measurements (extracellular enzymatic activities, nutrients, pigments, carbohydrates, bacterial cell counts; **Tables S3** and **S4**). Other contextual data were added to the environmental data set: data from the water column consisting of chlorophyll *a*, pH and water temperature that were obtained from the Sylt time series (van Beusekom et al. 2009) and data on wind speed that were obtained from the German Weather Service (Deutsche Wetterdienst) and measured at the weather station of List on Sylt.

DNA extraction and 454 MPTS.

The same DNA extract from sixteen of the samples previously used in Böer et al. 2009 (6) was used here for 454 pyrosequencing. The DNA was extracted using an UltraClean Soil DNA Isolation Kit (MoBio Laboratories Inc. Carlsbad, CA) and further stored in a final volume of 50 to 100 µL of Tris-EDTA buffer. At each step of the molecular protocols, the DNA quantities were spectrophotometrically adjusted using a NanoDrop ND-1000 Spectrophotometer (NanoDrop Technologies, Inc. Wilmington, DE). Extracted DNA from sandy samples was amplified using primers targeting the V6 region of the bacterial 16S rRNA gene and including 454 Life Science's A or B sequencing adapter according to (Sogin et al. 2006). Fragments were sequenced by pyrosequencing on a Genome Sequencer 20 system (Roche, Basel, Switzerland) at 454 Life Sciences (Branford, CT) by primer extension (Margulies et al. 2005).

Data analyses.

Output from the application of the 454 MPTS on sandy samples. Data from the 454 MPTS were retrieved from the publicly available “Visualization and Analysis of Microbial Populations Structure (VAMPS)” website (http://vamps.mbl.edu/). The taxonomic assignment of the sequences was performed by an automatic annotation pipeline (Sogin et al. 2006) using several known databases (Entrez Genome, RDP, SILVA). Although current databases are still sequence-limited, only 6% of sequences from the whole data set were not annotated at all. In our study, analyses were based on a definition of Operational Taxonomic Units (OTU) as unique (*i.e.* two sequences are considered as to belong to two different OTU_{unique} when they differ by at least one base) so as to keep a consistent definition throughout. Additional subsets were also considered and are indicated when necessary using a subscript notation: all, un-annotated sequences that we referred to as [OTU_{all}], the PyroNoise-corrected data sets (Quince et al. 2009), [$PyroNoise_{0\%}$, $PyroNoise_{3\%}$] defined at different percentages of sequence similarity, and the truncated OTU_{all} data set, without successive percentages (1-50%) of rare OTU_{unique} as proposed in a previous study (Gobet et al. 2010).

Taxa-environment relationships. As time and depth effects on biological variation may be confounded by the covariation with other measured parameters (pH, water temperature, wind speed, salinity, pigments, nutrients, extra-cellular enzymatic activities and cell properties), we first tested by using multiple regression how much the response variable time or depth could be explained by explanatory environmental variables. Thus, this analysis helps decide whether time and depth should be included or removed from further analyses to reduce collinearity in the ecological models.

To investigate taxa-environment relationships, most of the parameters (except pH, water temperature, wind speed and salinity) were log_{10}-transformed while the community matrices [OTU_{all} data set, the resident and SSO_{rel} data sets or the (potential) pathogen abundance matrix (including the genera *Parachlamydia*, *Arcobacter*, *Francisella*, *Acinetobacter*, *Rickettsiella*, *Pseudomonas*, and *Ralstonia*)] were Hellinger transformed (Legendre & Gallagher 2001). A forward selection (based on a canonical redundancy analysis (RDA) algorithm and 999 Monte Carlo permutation tests) of the environmental factors was done to find the set of parameters that could significantly explain the variation in the community table. The best-fitting models were chosen using the Akaike Information Criterion (AIC). Canonical variation partitioning (Borcard et al. 1992, Ramette & Tiedje 2007b) was then applied on the community data to test for the effects of pure environmental variables (pigments, nutrients, extra-cellular enzymatic activities, cell abundance) selected previously and their covariation on microbial community structure. Statistical analyses were carried out using the R statistical environment [R version 2.10.0, R Development Core Team 2009], using the *vegan* (Oksanen et al. 2009) and *gplots* (Bolker et al. 2009) packages and custom R scripts.

4.3 Results and Discussion

4.3.1 Dominant bacterial phyla of coastal marine sands

The three compartments sand, pore water, and the overlying water column may be expected to share a large proportion of the same microbial assemblages because the pore space of permeable sands is constantly flushed by the overlying water, trapping detritus and living cells from the water column (Boudreau et al. 2001). Microscopic observations of sand grains (**Fig. 4.1.A**), show that they are covered by a biofilm consisting of cells embedded in extracellular polymeric substances, with potentially little exchange with cells flowing through the sediment (Rusch et al. 2001). At any time, the porewater contains less than 0.2% of the cell abundance associated with the sand grains (Rusch et al. 2003). In addition, the exchange of bacterial populations between the sand and the water column is rather low, at the level of both community composition and evenness (**Figs. 4.1.B-C**, **S4.1.**). Here, only 2-3% of all $\mathrm{OTU}_{\mathrm{unique}}$ (sequences from the original $\mathrm{OTU}_{\mathrm{all}}$ data set that have at least one nucleotide difference are considered as $\mathrm{OTU}_{\mathrm{unique}}$ here) were shared between the three compartments, confirming previous findings (Llobet-Brossa et al. 1998).

The Sylt (North Sea) water column was mainly dominated by *Bacteroidetes* and by the *Alpha*- and *Gamma*- subdivisions of the *Proteobacteria*, as described previously by fluorescence *in situ* hybridization (FISH) and 16S rRNA gene-based clone libraries (Glöckner et al. 1999, Eilers et al. 2000, Zubkov et al. 2002). Sequences of the phyla *Verrucomicrobia* and *Actinobacteria* were also abundant (**Fig. 4.1.B**). The phyla dominating the top 5 cm of Sylt sand were *Bacteroidetes*, *Gammaproteobacteria*, *Deltaproteobacteria*, and *Planctomycetes*, as previously shown with FISH and 16S rRNA libraries (Llobet-Brossa et al. 1998, Musat et al. 2006). *Acidobacteria* sequences were also abundant in the sand microbial community (**Fig. 4.1.B, S4.2.**). At the phylum level, the pore water microbial composition resembled the sand community with a dominance of *Bacteroidetes*, *Gammaproteobacteria*, *Deltaproteobacteria*, *Acidobacteria* (**Fig. 4.1.B**). However, only a small number of OTU were shared with the sand-associated bacterial community or the water column, indicating a distinct ecological community in the pore water (**Fig. S4.3.**). For a complete analysis of temporal and spatial changes in bacterial communities of coastal sands, the sand-associated biofilm and pore water were not further distinguished.

Six sampling dates of coastal sands at the island of Sylt (North Sea) generated a total of 197,684 sequences, corresponding to 27,630 unique Operational Taxonomic Units ($\mathrm{OTU}_{\mathrm{unique}}$). Per sample, sequences ranged from about 5,000-19,000, corresponding to 496-2,993 $\mathrm{OTU}_{3\%}$ ($\mathrm{OTU}_{3\%}$ represent V6 sequences that were corrected by PyroNoise and clustered at 3% sequence dissimilarity, **Table S1**). Bacterial richness estimates reached high values as previously reported for soils, sediments and crusts which were also analyzed using 454 MPTS (Sogin et al. 2006, Huber et al. 2007, Roesch et al. 2007, Brazelton et al. 2010).

Gammaproteobacteria and *Deltaproteobacteria* were the dominant groups, representing 25-30 and 16-23%, respectively, of the total sequences from 0-15 cm sediment depth. Previous studies of coastal sediments from the North Sea also recognized *Bacteroidetes*, *Planctomycetes*, *Betaproteobacteria* and *Deltaproteobacteria* as the most abundant groups associated with sand grains, comprising more than 17% of the total cell number in the sands (Llobet-Brossa et al. 1998, Musat et al. 2006). Interestingly, total sequence abundance and OTU numbers systematically increased with depth from 0-15 cm at most of the sampling dates (**Table S1**). A positive relation between sediment depth and bacterial diversity was previously observed using different fingerprinting techniques (Urakawa et al. 2000, Böer et al. 2009), and explained by increasing physico-chemical stability of the habitat. The *Delta*- and *Beta*-subdivisions of the *Proteobacteria,* the *Deferribacteres, Spirochaetes* and *Nitrospira* showed an increasing OTU richness with depth. *Cyanobacteria* and *Bacteroidetes* were among the most represented phyla in the upper layers and decreased with depth (**Fig. S4.2.**).

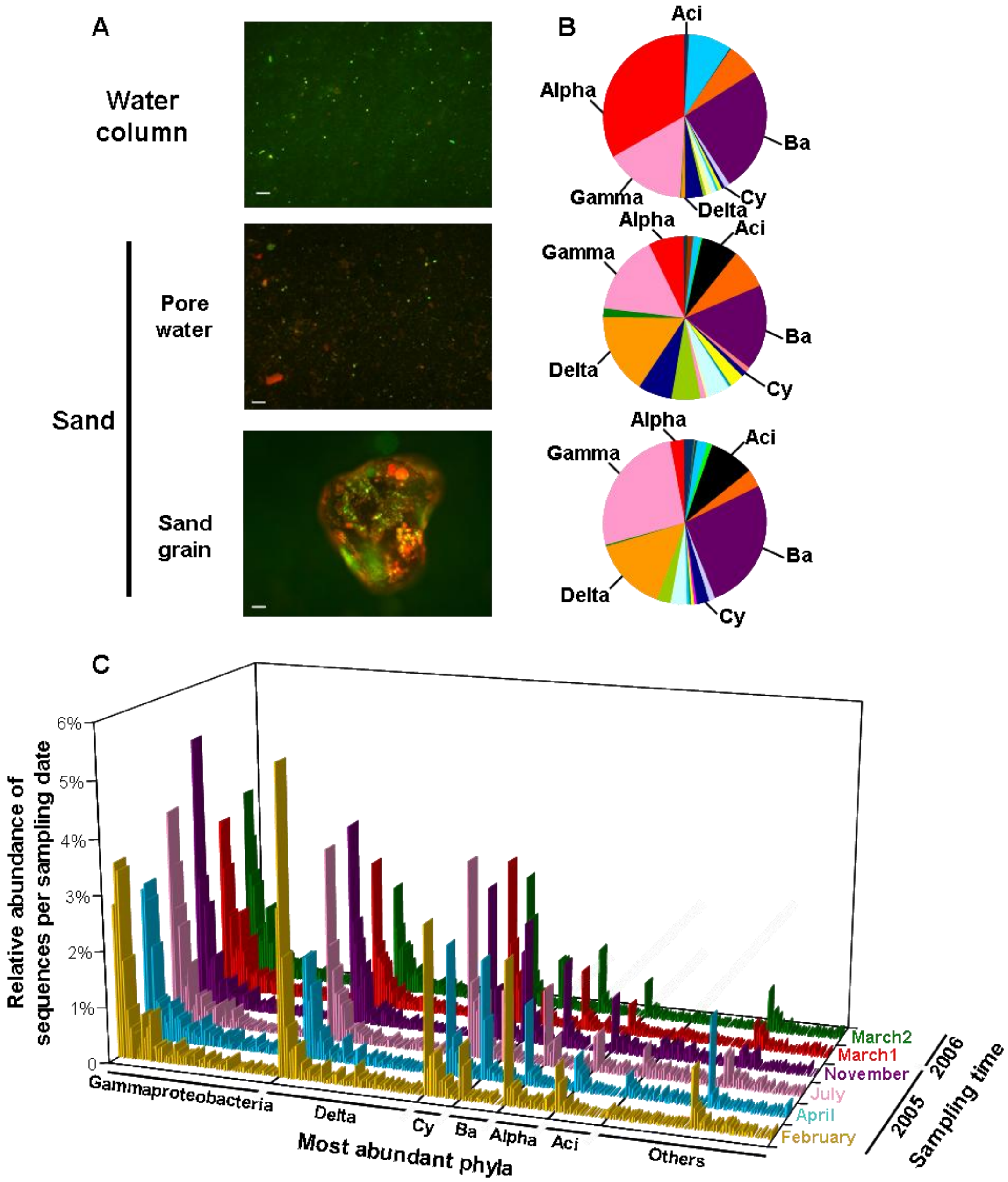


Figure 4.1. Microbial community distribution in the sand and the water column. (**A**) From top to bottom: acridine orange staining of bacteria in the water column, the porewater and on the surface of a sand grain (scale bar = 50 µm). (**B**) Relative number of sequences in different compartments for the top 5-cm sand layer and in the overlying water column in April 2008. Here the phylum level was chosen for illustrative purposes. (**C**) Sequence distribution in the sand over time where each bar represents an $\text{OTU}_{\text{unique}}$ (only $\text{OTU}_{\text{unique}}$ occurring more than 100 times in the whole data set are shown). The *Proteobacteria* phylum was further split into its corresponding classes *e.g.* Alpha, Gamma, Delta; Cy, *Cyanobacteria*; Ba, *Bacteroidetes*; Aci, *Acidobacteria*; Others: *Actinobacteria*, NA (not annotated)-*Proteobacteria*, *Planctomycetes*, *Chloroflexi*, *Verrucomicrobia*, *WS3*, *Firmicutes*, *Lentisphaerae*, *Deferribacteres*, *Gemmatimonadetes*. See SI for Materials and Methods for Fig. 4.1. A-B.

4.3.2 High turnover of bacterial diversity with sediment depth and time

From the phylum to the class level of taxonomic resolution, which are commonly used to explore the diversity of microbial communities based on whole cell fluorescence *in situ* hybridization or 16S-based clone libraries (Llobet-Brossa et al. 1998, Musat et al. 2006), microbial community patterns were mostly unchanged over time and sediment depth (**Fig. S4.2.**). However, drastic changes in bacterial community structure and composition were observed when a higher taxonomic resolution was used: Overall, only 152 OTU_{unique} (0.55% of the total number of 27,630 OTU_{unique}) were present in all sample depths at all times, and their sequence abundance ranged from 54 to 6,550. Also, only 3-5% of all OTU_{unique} within a sampling depth layer were present at all times (**Table 4.1.**). The majority, some 77%, of these OTU_{unique}, which we define as resident OTU, consisted of the phyla *Acidobacteria*, *Actinobacteria*, *Bacteroidetes*, *Cyanobacteria*, *Alpha-*, *Delta-* and *Gammaproteobacteria* and *Verrucomicrobia*. For the OTU_{all} data set (*i.e.* data set with all OTU_{unique}), a low percentage of OTU_{unique} were shared between sediment layers or any two sampling dates. Only about 20% of sequences were shared between the deeper layers and the upper layer (**Fig. S4.4.A**), the communities of the two deeper layers were slightly more similar (**Fig. 4.2.A**). Interestingly, resident OTU seemed to be characterized by a high and relatively stable number of sequences (**Figs. 4.3.B**, **S4.5.**). About 70% of the resident OTU had abundances of more than 10 sequences per sample; (**Table 4.1.**, **Fig. 4.3.B**), suggesting that they represented the common sand microbial communities.

When time was considered alone, only 18 to 37% of the OTU_{unique} were found to be shared between any two sampling times (**Fig. 4.2.B**, **Fig. S4.4.**). This indicates that a very large fraction of the community may be constantly replaced. Yet the fact that the turnover rate did not increase with sampling time suggests that some populations vanished and reappeared during the investigated time period (**Fig. S4.4.**). Some of the most sequence-abundant groups were found to be positively correlated with time (**Fig. 4.4.**); *e.g. Gammaproteobacteria* and *Planctomycetes*, following the seasonal fluctuations of cell abundances observed in Sylt sediment as previously described by FISH (Musat et al. 2006). Most interestingly, the fluctuation within a month (March 1 and 2 2006) was almost as high as within a year. This is explained by the large environmental variations occurring within March 2006, from the end of the winter with cold temperatures and high nutrient levels (March 1), to the spring bloom and stormy conditions (March 2).

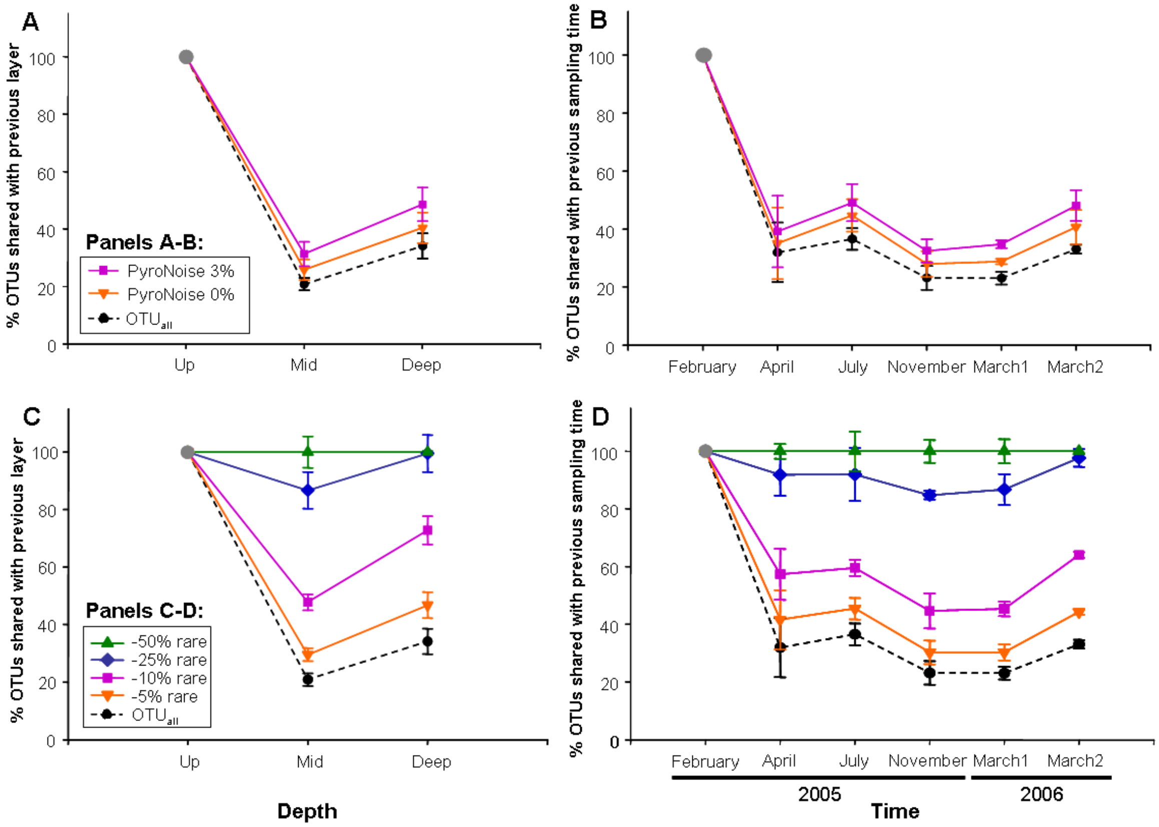


Figure 4.2. Turnover of the bacterial community between two consecutive (A, C) depth layers or (B, D) sampling times. The percentage of OTU shared between a sampling depth (or date) and the previous one was calculated (**A**, **B**) after PyroNoise correction and OTU clustering of the 454 MPTS data set at several levels of sequence dissimilarity and, (**C**, **D**) after removing successive percentages (here 5%, 10%, 25% and 50%) of rare OTU_{unique} from the OTU_{all} data set. Bars correspond to standard deviation calculated over 4-6 sampling dates (**A**, **C**) or three depth layers (**B**, **D**) except for July and November 2005 where 2 depth layers were considered. The first depth layer and sampling date (February 2005) are indicated by the grey point as 100% of OTU. OTU_{all} represents the original data set with all OTU_{unique}, used here as a reference to study the effects of the various attempts to correct the data set on the dynamics of the bacterial community. Note that the panels have different legends.

Observing a high turnover for a substantial fraction of the bacterial community in marine sands raises several questions about the ecological significance of these dynamics. An interesting question is as to the abundance and role of the resident, potentially sand-biofilm associated populations vs. rare populations getting advected through the sands. Indeed, in our data set, more than 50% of the OTU_{unique} present at all times in the three depth layers were members of the *Gammaproteobacteria* and *Deltaproteobacteria*, two classes that were also previously found to be dominating Sylt sands, representing respectively up to 23% and up to 10% of the total cell counts, as determined by FISH (Musat et al. 2006). But also 40% of all $\text{OTU}_{\text{unique}}$ appearing only once in a given sample or in the whole OTU_{all} data set, in the investigated time period belonged to these classes. The rare members of dominant classes never became abundant at any sampling time, confirming the conclusion of a previous study in Arctic waters (14). Among the rare bacteria in coastal sands were also potential pathogens (including *Parachlamydia*, *Arcobacter*, *Francisella*, *Acinetobacter*, *Rickettsiella*, *Pseudomonas*, *Ralstonia*) with a total of 16-88 sequences corresponding to 2-54 $\text{OTU}_{\text{unique}}$ in the OTU_{all} data set (0.22% of all sequences in OTU_{all}).

Large turnover in community composition could be due to various phenomena, which are not mutually exclusive, such as: 1) Migration of non-resident bacterial populations into and out of an ecosystem due to the rapid physical transport processes prevailing in sands (Sloan et al. 2006). 2) High adaptability of microbial communities to ever-changing and complex environments, as previously observed in the laboratory (Rosenzweig et al. 1994). 3) Emergence of latent “rare” prokaryotic stages (Finlay 2002, Pedrós-Alió 2006) that may become dominant when appropriate conditions are met. This fluctuation from rare to dominant types will be supported by the “seed bank” hypothesis (Finlay 2002). 4) Finally, the hypothesis that free-floating DNA from seawater was also retrieved by our approach cannot be excluded, since sands are known to act as natural filters that concentrate particles and DNA in suspension (Naviaux et al. 2005). However, the dissimilarity between water column and sand $\text{OTU}_{\text{unique}}$ described above does not support Hypothesis 4. To test hypotheses 1-3, we further investigated the contribution of the rare biosphere to the overall community turnover.

Table 4.1. Temporal patterns of OTU variation[a]

								Sequence similarity thresholds for PyroNoise-corrected data		
							OTU_{all}	$\text{PyroN.}_{0\%}$	$\text{PyroN.}_{3\%}$	$\text{PyroN.}_{5\%}$
Number of OTU (%)[b]										
Total							19,785 (100)	9,036 (100)	7,668 (100)	5,927 (100)
SSO_{abs}							13,676 (69)	5,272 (58)	3,970 (52)	2,624 (44)
SSO_{rel}							3,735 (19)	2,136 (24)	2,268 (30)	2,087 (35)
DSO_{abs}							1,108 (6)	697 (8)	536 (7)	386 (6)
TSO_{abs}							236 (1)	187 (2)	170 (2)	133 (2)
With $\geq$ 5 sequences minimum in any sample							203 (1)	124 (1)	125 (1)	135 (2)
With $\geq$ 10 sequences minimum in any sample							106 (0.5)	67 (0.7)	67 (0.9)	69 (1)
Temporal patterns of occurrence (%)[c]										
	Feb	Apr	Jul	Nov	Mar1	Mar2				
	■	■	■	■	■	■	654 (3)	394 (4)	412 (5)	443 (7)
	■	■	■				29 (0.2)	11 (0.1)	12 (0.2)	10 (0.2)
				■	■	■	157 (0.8)	110 (1)	116 (1)	106 (2)
	■		■		■		34 (0.2)	21 (0.2)	23 (0.3)	23 (0.4)
		■		■		■	32 (0.2)	18 (0.2)	18 (0.2)	19 (0.3)
	■						1,998 (10/9) [d]	685 (8/6)	515 (7/5)	304 (5/4)
		■					1,371 (7/6)	622 (7/6)	492 (6/5)	345 (6/5)
			■				1,807 (9/8)	704 (8/7)	554 (7/6)	349 (6/5)
				■			3,657 (18/17)	1,571 (17/15)	1,171 (15/13)	821 (14/11)
					■		3,539 (18/16)	1,706 (19/15)	1,301 (17/13)	915 (15/12)
						■	2,833 (14/13)	1,018 (11/9)	767 (10/8)	501 (8/7)
Type of temporal relationships (%)[e]										
Linear (total)							490 (2)	287 (3)	284 (4)	261 (4)
Linear increase							375 (2)	231 (2)	230 (3)	215 (4)
Linear decrease							115 (0.6)	56 (0.6)	54 (0.7)	46 (0.8)
Quadratic (total)							227 (1)	113 (1)	114 (1)	103 (2)
Quadratic positive							208 (1)	100 (1)	97 (1)	90 (1)
Quadratic negative							19 (0.1)	13 (0.1)	17 (0.2)	13 (0.2)

[a] Sequences from the first top 10-cm sediment layers pooled were processed with the VAMPS pipeline and with the PyroNoise algorithm to remove pyrosequencing and PCR amplification artifacts. Thresholds from 0 to 5% sequence similarity were used to define OTU. [b] Numbers of sequences are given for Single Sequence OTU occurring in the whole dataset (SSO_{abs}) or at least in one sample (SSO_{rel}), Double Sequence OTU (DSO_{abs}), or Triple Sequence OTU (TSO_{abs}) in the whole dataset. Figures in parentheses correspond to the percentage of the total number of sequences in the first top 10-cm sediment layers. [c] Occurrence patterns were defined as the presence (black square) or absence of OTU at specific sampling dates from February 05 (Feb), April 05 (Apr), July 05 (Jul), November 05 (Nov), beginning of March 06 (Mar1), to end of March 06 (Mar2). [d] The second value in parentheses indicates the respective percentage of SSO_{abs}. [e] Different patterns of OTU abundance were examined by specifying linear or quadratic models of change with time. A positive quadratic relationship with time implies a U-shape relationship, *i.e.* associated with high abundance at the beginning and at the end of the study, while low abundances observed at the 3rd and 4th sampling times. A negative quadratic relationship would be conversely described by an inverted U-shape abundance function over time.

4.3.3 Impact of the rare biosphere on community turnover

The rare biosphere has been postulated to consist of low-abundance microbial organisms that would not be subjected to predation or viral lysis, and would likely represent a huge proportion of microbial communities, as generally indicated by long distribution tails in rank-abundance curves (Pedrós-Alió 2007). In order to understand which fraction of the community may be associated with the large diversity turnover observed in the sands, we gradually removed increasing fractions of the rare sequences in the data set starting from the rarest ones (Gobet et al. 2010): interestingly, when up to 50% of the rare sequences were removed, the large turnover previously described for the complete data set over both sediment depth and time (**Figs. 4.2.C**, **4.2.D**, **S4.6.**) was no longer observed, indicating that most of the community turnover was due to changes in the rare tail of the data set. The rare biosphere could be identified as OTU appearing only once in a given sample (*i.e.* Single Sequence OTU relative [SSO_{rel}], representing about 20% of all PyroNoise-corrected $OTU_{0\%}$), or as OTU appearing only once in the whole data set (*i.e.* Single Sequence OTU absolute [SSO_{abs}], representing about 58% of the $OTU_{0\%}$), as compared to the 3-5% of the $OTU_{0\%}$ that were resident (**Table 4.1.**). Moreover, when only the SSO_{rel} fraction of the data set was retained, very similar fractions of total explained biological variation were identified as for the total data set (**Fig. S4.7.**). This study is the first to report that such a large fraction of the community consists of rare types, which undergo substantial replacement over few months of time or few centimeters of sediment depth. In conclusion, both the presence of this large proportion of singleton $OTU_{0\%}$ in sandy sediments, and the large turnover in community composition could be explained by the dispersal of OTU from other sand locations by advective transport and physical mixing (Boudreau et al. 2001). Further research would be needed to examine the biogeography of rare and resident bacterial types of coastal sands.

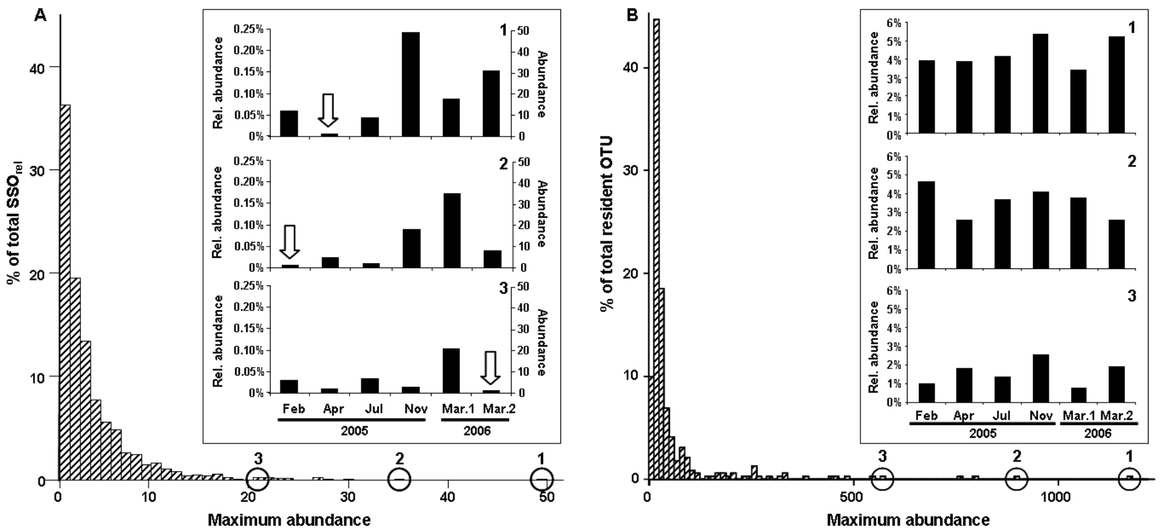


Figure 4.3. Distribution of the maximum abundance of (A) SSO_{rel} (*i.e.* $OTU_{0\%}$ that, at least in one sample, consisted of only one sequence) and (B) resident $OTU_{0\%}$ (*i.e.* $OTU_{0\%}$ present at all times) in the top 10 cm layer. In (**A**), panels **1**, **2**, **3** are examples of cases where particular high fluctuations from the single sequence case (white arrow) to higher sequence abundances were observed. In (**B**), panels **1**, **2**, **3** are examples of cases where particular high fluctuations of relative abundance were observed. No absolute abundances were calculated in this case. All data were initially processed to remove pyrosequencing noise. Rel. abundance, relative abundance to the total of sequences per sampling time: February 2005 (13,285), April 2005 (7,902), July 2005 (10,910), November 2005 (21,931), March1 2006 (17,302), March2 2006 (17,897).

4.3.4 Ecological interpretation of overall microbial diversity patterns

The next question was to determine whether observed temporal patterns may be attributed to either deterministic (niche-based) or stochastic processes, or both (Ramette & Tiedje 2007a). We used the previously published contextual data (**Tables S4.4.** and **S4.5**, Böer et al. 2009) for the investigated samples, to test the effects of time, sediment depth, cell abundance and biogeochemical gradients (*i.e.* pigments, nutrients and extra-cellular enzymes,**Fig. 4.4.**), and their combined effects on bacterial community structure. A multiple regression analysis indicated that the factors time and depth significantly covaried with most biogeochemical factors. Temporal variation significantly explained 70% of nutrients' variation while depth significantly explained the variation of pigments, nutrients, extra-cellular enzymes as well as cell abundances [93%, 74%, 80% and 79% of the variation explained, respectively, using multiple regression analyses (data not shown)], which overall confirms previous observations (Böer et al. 2009). Thus, in the next step, time and depth were removed from the pool of independent factors to investigate the role of other environmental factors.

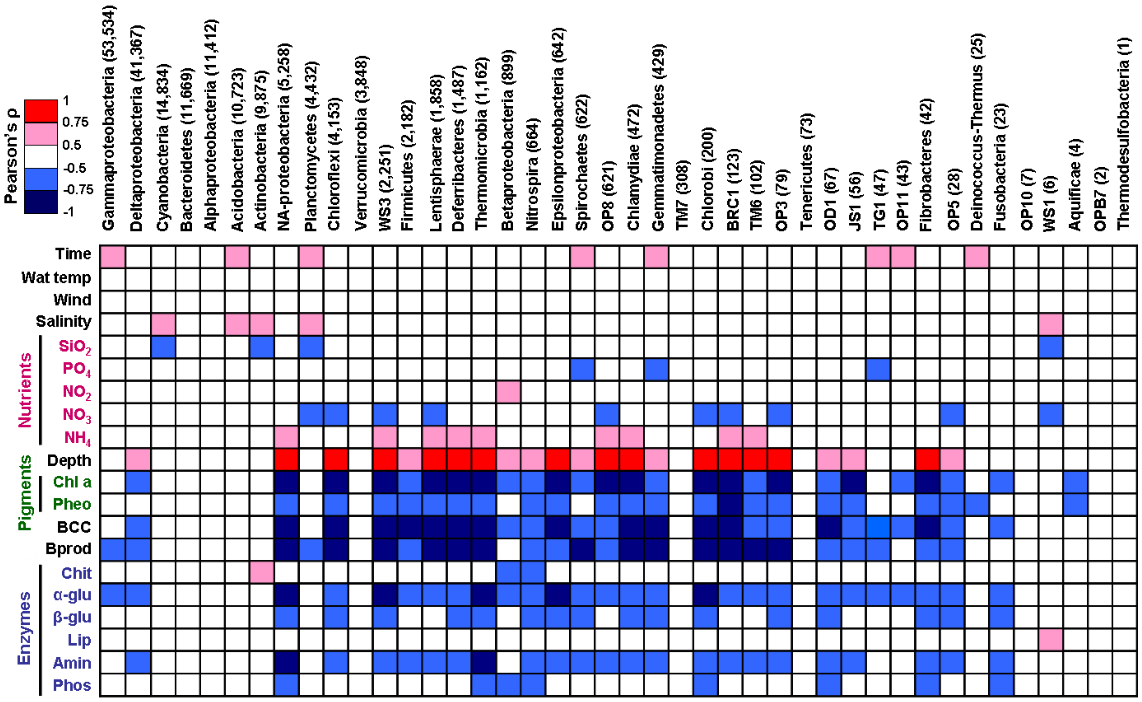


Figure 4.4. Impact of the environment on the community structure in the sand for all phyla. Pearson's ρ indicates correlations between phyla distribution and several environmental parameters. For example, a red square between sediment depth and *Chloroflexi* indicates a higher number of sequences with increasing depth. The *Proteobacteria* phylum level was separated into its corresponding classes for higher resolution. NA-*Proteobacteria* are *Proteobacteria* with missing class annotation. The total number of sequences in each phylum is indicated in brackets. SiO_2, silicate; PO_4, phosphate; NO_2, nitrite; NO_3, nitrate; NH_4, ammonium; Chl *a*, chlorophyll *a*; Pheo, pheophytin; BCC, Bacterial abundance; Bprod, Bacterial Carbon production; Chit, chitinase; α-glu, α-glucosidase; β-glu, β-glucosidase; Lip, lipase; Amin, aminopeptidase; Phos, phosphatase.

A multivariate variation partitioning approach (Borcard et al. 1992, Ramette & Tiedje 2007b) showed that biogeochemical (pigments, nutrients and extra-cellular enzymes) gradients, cell abundance and their covariation were directly related to the major changes in community structure (**Fig. S4.7.**). Yet, the significant biogeochemical variables included in the most parsimonious multivariate models were qualitatively almost the same for the resident OTU, the phylum, OTU_{all} and SSO_{rel} levels (**Fig. S4.7.**, **Table S4.2.**). Also, a greater amount of biological variation (R^2) could be explained when a lower taxonomic resolution was used. When data sets become less complex, it is easier to explain the biological variation present in the community (Gobet et al. 2010).

Depending on the research question, this can be achieved by removal of the rare members of the biosphere from the data set to be analyzed, or by decreasing taxonomic resolution. Furthermore, that environmental selection acts directly on the phylum level can be best explained in the case of taxonomic groups with distinct environmental functions or niches, such as the *Cyanobacteria* or the *Deltaproteobacteria* (see discussion below).

The main parameters significantly influencing the variation for several OTU definition levels were chlorophyll *a*, activity of the extracellular enzyme phosphatase and cell abundance (**Table S4.2.**), with all three parameters highly correlated with sediment depth. Cell abundance was also identified as an important factor influencing the variation in the different OTU data sets (**Table S4.2.**). Measuring chlorophyll *a* from coastal sands allows the biomass of the microphytobenthos to be quantified (Lorenzen 1967). Microphytobenthos contains mostly cyanobacteria and diatoms and produces labile exudates enhancing bacterial growth. It secretes extracellular polymeric substances forming biofilms around sand grains to avoid desiccation and to capture particles and organic matter (Stal 2003). The phosphatase enzyme activity also correlates positively with bacterial cell growth (Böer et al. 2009), and may be relevant in times of nutrient limitation by cleaving inorganic phosphate from organophosphate complexes (Kloeke et al. 1999). It is important to notice that vertical gradients of microbial functions (*e.g.* biomass, benthic oxygen consumption and extra-cellular enzymatic activities) in Sylt sandy sediments varied to different extents, but usually much less than community diversity and composition. This suggests that the few abundant, resident microbes perform the main microbial functions in sandy ecosystems, and that the rare types being replaced at high rates may have little effects on bulk functions. An alternative hypothesis is that a substantial level of functional redundancy exists in microbial communities, regardless of them being resident or transient organisms. However, a large amount (58%) of $OTU_{0\%}$ in Sylt sands occurred as a single sequence, this hypothesis seems unlikely to explain the observed functional stability.

It is interesting that the temporal dynamics of bacterial communities in sandy sediments were clearly distinct from the seasonally reoccurring, cyclic bacterial patterns that were observed in water column samples offshore Southern California (Fuhrman et al. 2006), in the English channel (Gilbert et al. 2009) and in the Baltic Sea (Andersson et al.

2010). In contrast, a study of the bacterioplankton in the Arctic has highlighted the stability of the community through seasons (Kirchman et al. 2010), indicating that patterns of temporal variation may differ between different microbial habitats and oceanic regions.

Deeper ecological insights may also be gained by the detailed analysis of the correlations between sequence abundances of distinct bacterial groups and specific environmental parameters (**Fig. 4.4.**). Although the environmental interpretation of whole-community patterns was overall straightforward, more contrasting patterns were obtained when phyla were examined individually, with some of them being strongly, but not necessarily identically influenced by depth and related environmental parameters (such as pigments, bacterial properties and some enzymes; **Fig. 4.4.**). For example, some of the most abundant phyla showed different response patterns with changes in environmental conditions: *Deltaproteobacteria* sequence abundance was positively correlated with depth and negatively correlated with chlorophyll *a*, bacterial abundance, alpha-glucosidase, bacterial productivity and aminopeptidase. About 67% of the *Deltaproteobacteria* was composed of *Desulfobacterales*, an order representing sulfate-reducing bacteria that thrive in anoxic North Sea sand (de Beer et al. 2005). The cyanobacterial group was found to be positively correlated with salinity and negatively correlated with silicate, a factor significantly correlating with depth. The order *Oscillatoriales* accounted for 47% of the *Cyanobacteria* and the main genera represented here was *Oscillatoria*, which can grow in harsh environmental conditions with high seasonal fluctuations of salinity and temperature, and which may play a role in degradation of different hydrocarbon compounds in intertidal oil-polluted sediment (Al-Thukair et al. 2007). In contrast, variations in sequence abundance for other prominent phyla, such as *Bacteroidetes* or *Alphaproteobacteria*, were not significantly correlated with the environmental parameters analyzed here (**Fig. 4.4.**). Certainly, other environmental parameters such as biological interaction with other bacteria or other organisms, grazing, organic matter composition, *etc.* are also important factors in structuring the diversity of bacterial communities in sand.

4.3.5 The rare biosphere responds to environmental drivers

The biogeography (Galand et al. 2009a) and the dynamics over a thousand-year period (Brazelton et al. 2010) of rare microbial biospheres of different marine habitats have previously been described, but an ecological explanation of the rare microbial biosphere in its environmental context has not been undertaken yet. In order to better understand whether patterns in the rare fraction were random or environmentally-driven, the proportion of SSO_{rel} (**Table S3**) in each sample was correlated with environmental parameters. Interestingly, fluctuations in SSO_{rel} were not random (as determined by model selection using Monte Carlo permutation test and by selecting the lowest Akaike-Information-Criterion model values; data not shown), and seemed to be negatively correlated with pigments (chlorophyll *a*, phaeopigments) and bacterial carbon production in the sediments (**Fig. S4.8.**). For $OTU_{0\%}$, the corresponding Pearson correlation coefficients between the proportions of SSO_{rel} and chlorophyll *a*, phaeopigments and bacterial carbon production were -0.730, -0.695, and -0.626, respectively, and were all highly significant. However, neither sediment depth nor sampling time could significantly explain the variation in the proportion of rare types. Because pigment concentration and bacterial carbon production may indicate the food availability status in sand ecosystems (Rusch et al. 2003), an increased proportion of rare types that is concomitant with a decrease of those parameters would indicate that rarity becomes more prevalent when the environmental conditions become harsher for microbial life. This is when primary and secondary production in Sylt sands reaches a minimum in late winter, at temperatures close to the freezing point, and high wind forces.

In summary, the high turnover of bacterial community composition observed was explained by high fluctuations of rare bacterial types, which made up 60% of all $OTU_{0\%}$. Although less than 5% of all $OTU_{0\%}$ were present at all times and sampling depths, they mostly comprised the abundant types represented by high biomass in sands. Accordingly, the main environmental functions were maintained in the sandy coastal ecosystem despite the turnover of a high proportion of all sequences. Fluctuations in the bacterial community were related to those of biogeochemical gradients at all levels of taxonomic resolution, including the phylum level. The rare biosphere presented contrasting ecological patterns associated with low productivity phases in winter. Future functional analyses are needed to examine whether the large proportion of single OTU are commonplace also in other sandy habitats, including those of tropical seas characterized by lesser environmental fluctuations, and hence typical for this microbial realm dominated by physical transport.

Acknowledgements and Funding

We acknowledge Martina Alisch, Marianne Jacob, Susanne Menger and Shalin Seebah for great assistance with Sylt sampling and help with sample processing. We thank Rafael Stiens for cell microscopy. This work was supported by the Marie Curie Early Stage Training fellowship in Marine Microbiology [MarMic EST contract MEST-CT-2004-007776 to A.G.] and by the International Max Planck Research School of Marine Microbiology. [A.G.]; as well as by the Max Planck Society [A.R., A.B.], Helmholtz Association [A.B., J.B.] and Leibniz Program of the DFG [A.B]. CQ is supported by an Engineering and Physical Sciences Research Council Career Acceleration Fellowship (EP/H003851/1). The sequencing was financed by a Keck Foundation grant to the International Census of Marine Microbes led by Mitchell L. Sogin and Linda A. Amaral-Zettler. This is a contribution to the International Census of Marine Microbes (ICoMM).

4.4 Supplementary Information

4.4.1 Supplementary Text

4.4.1.1 Materials and Methods for Fig. 4.1.A-B

Study site and sampling procedures. Sandy sediment cores (55° 2' 28'' N, 8° 24' 26''E) and seawater (55° 1'41''N, 8° 26'10''E) were collected at the "Königshafen" intertidal in April 2008. 1l seawater was successively filtered through 10 µm and 0.2 µm (filters type GTTP; diameter, 47 mm) before further DNA extraction. Pore water of the upper 5 cm of the sediment was separated from the sand grains (with bacterial biofilms) through GF-C filters (47 mm) by low speed centrifugation (10 min at 1801 rcf and 4°C). We fabricated centrifuge tubes according to the following experimental set-up: 1) the upper part of a 50 ml Falcon tube was cut up to the 20 ml indication and the conic bottom of the tube was pierced with 16 holes in a circular way, in staggered rows, 2) a GF-C filters was put inside the latter Falcon tube so as to retain the sediment during the centrifugation, 3) the whole assembly was then taped on top of another Falcon tube and finally, 4) 10 g of sand was put in the assembly before centrifugation (**Fig. S4.9.**). The resulting 8 g of "dry" sediment were put aside and the resulting 2 ml pore water was filtered through 0.2 µm (filters type GTTP; diameter, 47 mm) before further DNA extraction.

DNA extraction, 454 MPTS and taxonomic annotation. DNA was extracted from 1) the 8 g sediment and, 2) the 0.2 µm filters (pore water and water column) cut into pieces with a sterilized cutter, using an UltraClean Soil DNA Isolation Kit (MoBio Laboratories Inc. Carlsbad, CA) and further stored in a final volume of 50 to 100 µL of Tris-EDTA buffer. 454 MPTS and taxonomic annotation of the sequences were performed as for the other sixteen samples.

Acridine Orange staining. The water column and pore water samples were fixed with 1:10 seawater-diluted sterile-filtered (0.2 µm disposable syringe filter) 37% formaldehyde (methanol stabilized) solution. 1 ml of sand was fixed in 9 ml of a particle-free 2% formaldehyde/seawater solution. Pictures of the bacteria in the different samples were done by epifluorescence microscopy after staining with Acridine Orange. The samples of porewater and water column were non-diluted while some sand grains were directly applied onto 0.2 µm black polycarbonate filters, stained with Acridine Orange solution for 3 min and rinsed with 1 ml citrate buffer.

4.4.1.2 Results and Discussion

Turnover of the overall microbial community with sediment depth and time: testing the data reliability (§ 4.3.2)

Several studies have questioned the accuracy of high-throughput sequencing data due to the likely existence of chimeric sequences originating from sequencing or PCR amplification artifacts [*e.g.* (Quinlan et al. 2008)]. This issue was of particular importance for our study because it may have erroneously inflated the observed community turnover rates. Yet, even after applying the PyroNoise algorithm (Quince et al. 2009) to remove pyrosequencing and amplification noise from the data and reclustering of the sequences at different levels of sequence similarity, a very large turnover (*i.e.* 40-70% of sequence replacement) of the bacterial community over depth or time could still be observed at various sequence dissimilarity levels used to cluster the data (**Figs. 4.2.A-B**, **S4.4.**). If we consider the PyroNoised-corrected data clustered at 3% sequence dissimilarity, turnover patterns were still really high, as observed at the OTU_{all} level (**Figs. 4.2.A-B**, **S4.4.**). It may therefore be concluded that the large community turnover present in marine sandy sediments is not due to technical artifacts and is consistently observed at different taxonomic levels, yet to different extent depending on taxonomic resolution.

Impact of the rare biosphere on community turnover (§ 4.3.3)

We tested several scenarios for temporal fluctuations of the rare biosphere investigating $OTU_{0\%}$ and SSO_{rel} occurrence for the top 10 cm sediment layer over the six sampling times (**Table 4.1.**). For example, less than 1% of the $OTU_{0\%}$ appeared at either the first or last three sampling dates only, or had patterns that would skip one sampling date each time. Noticeably, 6-17% of the $OTU_{0\%}$ per sample appeared only at one sampling date and mostly consisted of SSO_{abs}, *i.e.* OTU that occurred only in one sample with a sequence abundance of one. The fluctuation in sequence abundance of noise-corrected SSO_{rel} was also explored (**Fig. 4.3.**): altogether, 94% of SSO_{rel} had a maximum abundance below 10 sequences when all samples were considered, indicating that when an $OTU_{0\%}$ was rare it remained rare and was not likely to become abundant within a year, as already observed through a seasonal survey in the Arctic water column (Kirchman et al. 2010). Few SSO_{rel} displayed very high fluctuation in abundance (**Fig. 4.3.A-C**), which further supports the idea that blindly removing SSO_{rel} or SSO_{abs} from the data set would also remove meaningful patterns of OTU variation (Huse et al. 2010).

The rare biosphere responds to environmental drivers (§ 4.3.5)

The presence of pathogenic bacteria in the environment is of great interest for public health and for fisheries, and further research concerning regulative factors of pathogen distribution may be of prime importance especially for coastal ecosystems (Stewart et al. 2008). Both the survival and proliferation of pathogens can be influenced by the variation of environmental factors. Potential pathogenic genera in our data set were selected according to the literature. A combination of environmental parameters could explain 55% of their overall biological variation (**Table S4.6.**, **Fig. S4.10.**). Noticeably, such pathogens are not usually observed in marine sediments and occurred at low sequence abundance in our samples, thus they could be dispersed by various ways of human interaction with coastal habitats, and occur in a latent stage in sands, further supporting the idea of a “seed bank” of rare organisms as previously proposed by several authors (Finlay 2002, Pedrós-Alió 2006).

4.4.2 Supplementary Figures

Fig. S4.1. Bacterial phylum composition and distribution in different sand-associated compartments and in the water column.

Fig. S4.2. Relative sequence abundance of phyla in the sand through depth and time.

Fig. S4.3. Total number of shared OTU between the three compartments: Sand grain-associated biofilm, sand porewater, and overlying water column.

Fig. S4.4. Turnover of the bacterial community between sediment depth layers or sampling dates after PyroNoise correction and successive OTU clustering of the 454 MPTS data.

Fig. S4.5. Variation in minimum sequence abundance among resident $OTU_{0\%}$ (*i.e.* $OTU_{0\%}$ present at all times) in the top 10 cm sediment layer.

Fig. S4.6. Turnover of the bacterial community between sediment layers or sampling dates after applying MultiCoLA.

Fig. S4.7. Partitioning of the biological variation in the bacterial community structure as a function of explanatory variables.

Fig. S4.8. Effects of environmental conditions on the proportion of SSO_{rel} per sample.

Fig. S4.9. Experimental set-up to separate the pore water from sand grains.

Fig. S4.10. Partitioning of the biological variation in the bacterial community structure of potential pathogens.

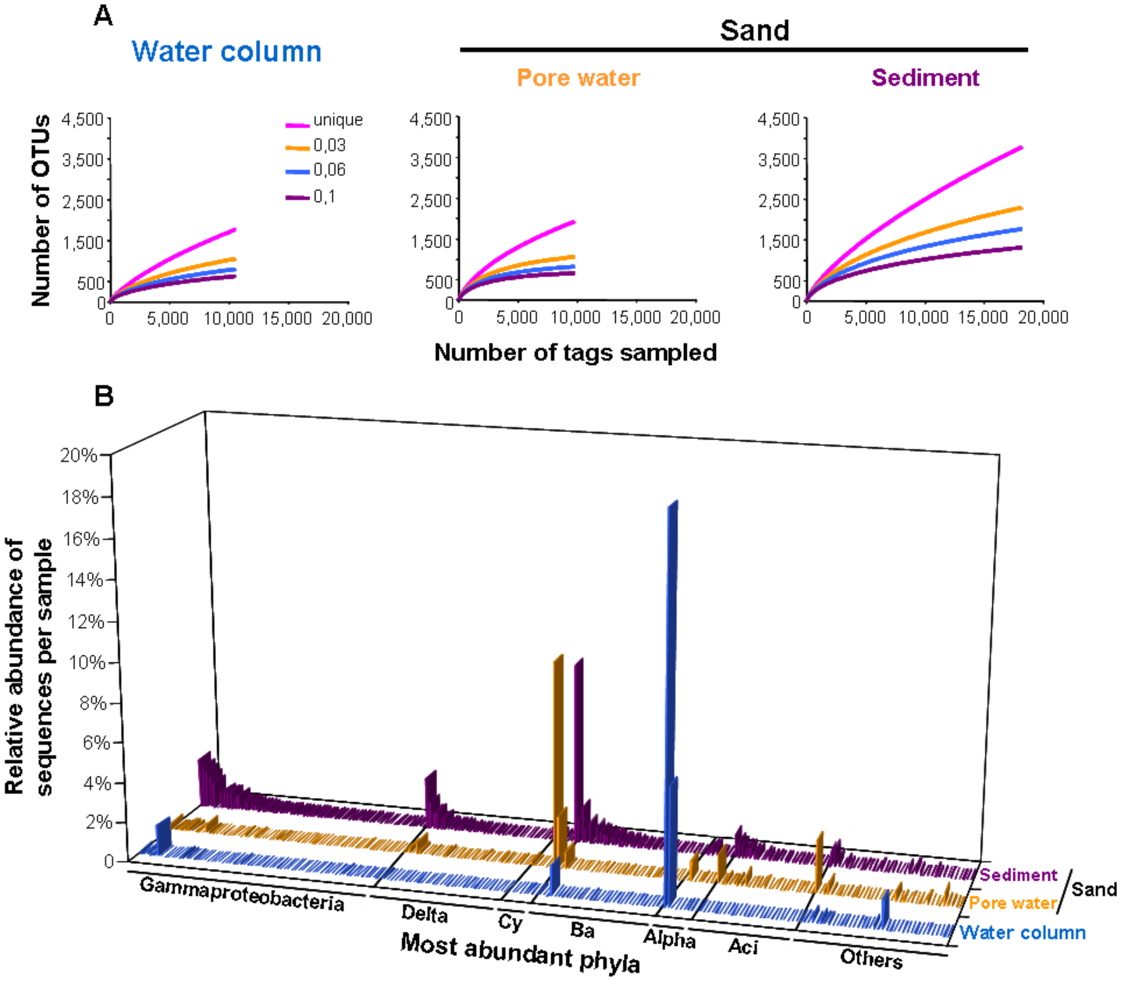


Fig. S4.1. Bacterial phylum composition and distribution in different sand-associated compartments and in the water column. (**A**) Rarefaction curves in different compartments of the sand and in the water column at the unique, 3%, 6% and 10% dissimilarities to define OTU. (**B**) Sequence distribution in different compartments of the sand and in the water column in April 2008. Each bar represents an $\text{OTU}_{\text{unique}}$ (only $\text{OTU}_{\text{unique}}$ occurring more than 10 times in the OTU_{all} data set are shown on the skyline plot). The *Proteobacteria* phylum was separated into its corresponding classes for higher resolution. Gamma, *Gammaproteobacteria*; Delta, *Deltaproteobacteria*; Cy, *Cyanobacteria*; Ba, *Bacteroidetes*; Alpha, *Alphaproteobacteria*; Aci, *Acidobacteria*; Others: *Actinobacteria*, NA-*Proteobacteria* (*Proteobacteria* with class annotation missing), *Planctomycetes*, *Chloroflexi*, *Verrucomicrobia*, WS3, *Firmicutes*, *Lentisphaerae*, *Deferribacteres*, *Epsilonproteobacteria*, *Gemmatimonadetes*.

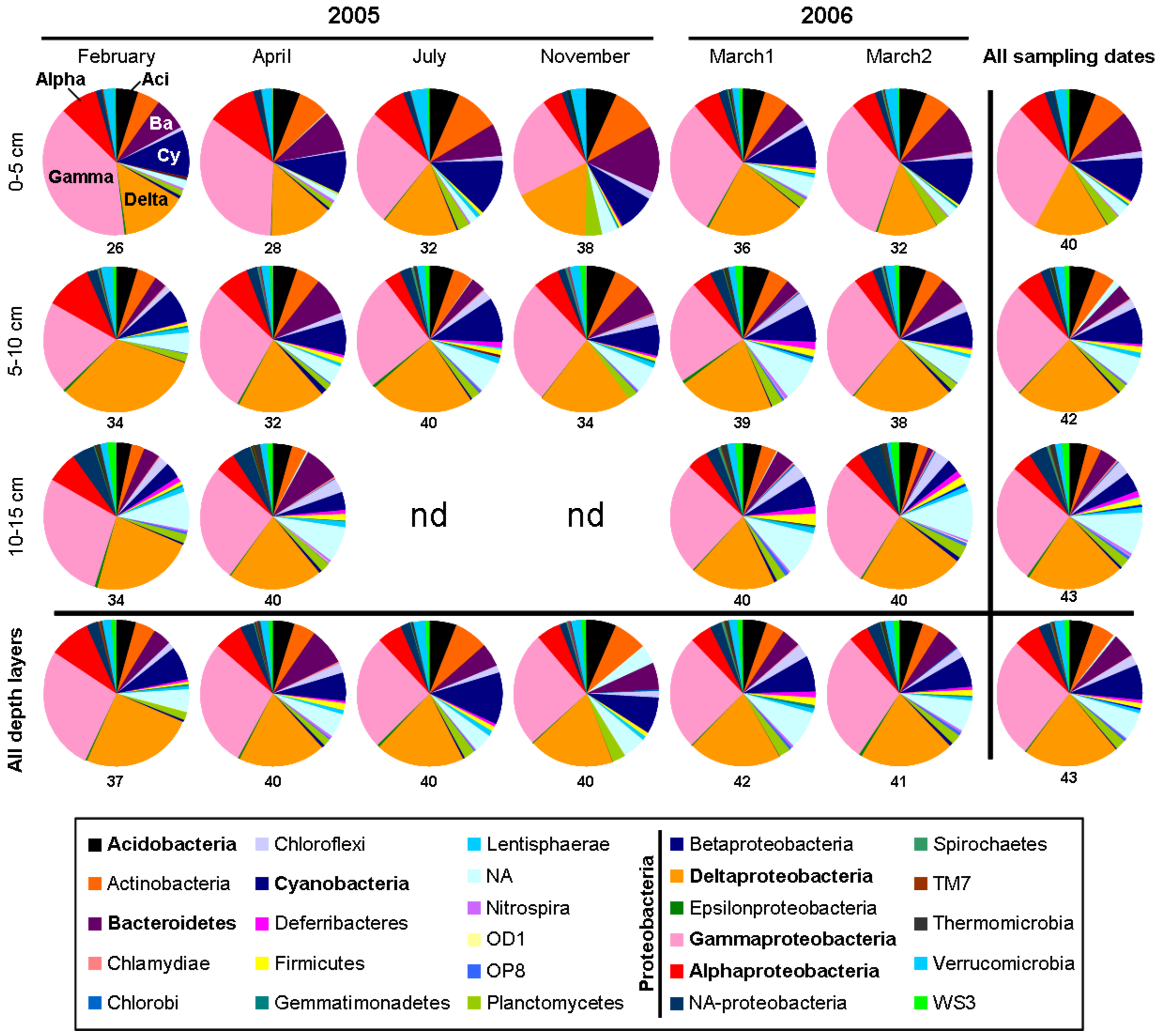


Fig. S4.2. Relative sequence abundance of phyla in the sand through depth and time. Some of the most abundant phyla were indicated on the first pie chart and in bold in the legend: Gamma, *Gammaproteobacteria*; Delta, *Deltaproteobacteria*; Cy, *Cyanobacteria*; Ba, *Bacteroidetes*; Alpha, *Alphaproteobacteria*; Aci, *Acidobacteria*. The *Proteobacteria* phylum was separated into its corresponding classes for higher resolution. Unassigned $\text{OTU}_{\text{unique}}$ were grouped as one phylum in this figure (NA). nd, missing samples. The total number of phyla (including the *Proteobacteria* divided as classes) is indicated under each pie chart. The legend indicates the phyla color code of the pie charts. Phyla with too few sequences to be visible on the pie charts were removed from the legend: *Aquificae*, *Deinococcus-Thermus*, *Fibrobacteres*, *Fusobacteria*, *Tenericutes*, *Thermodesulfobacteria* and, the candidate divisions: BRC1, JS1, OP10, OP11, OP3, OP5, OPB7, TG1, TM6 and, WS1.

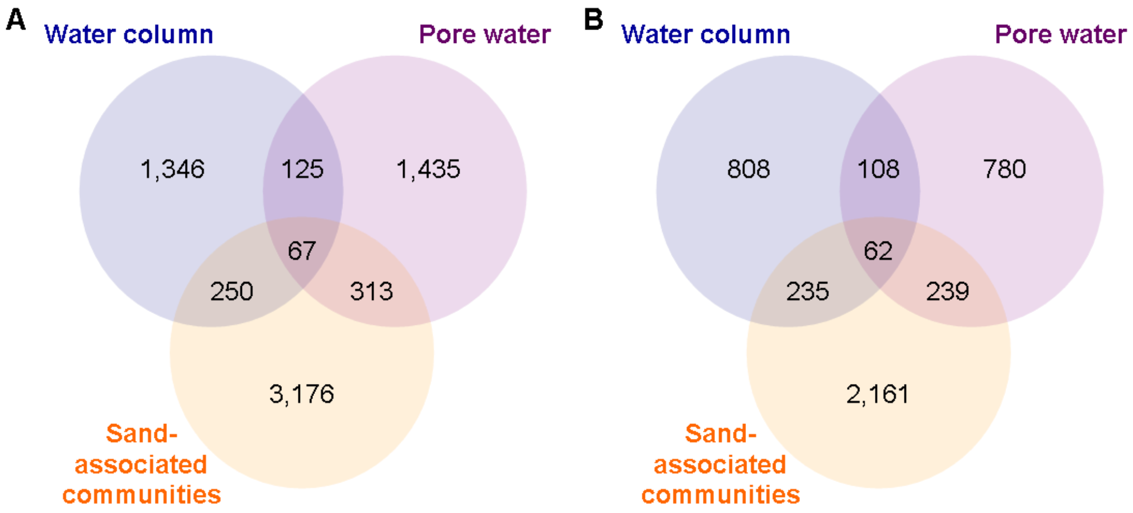
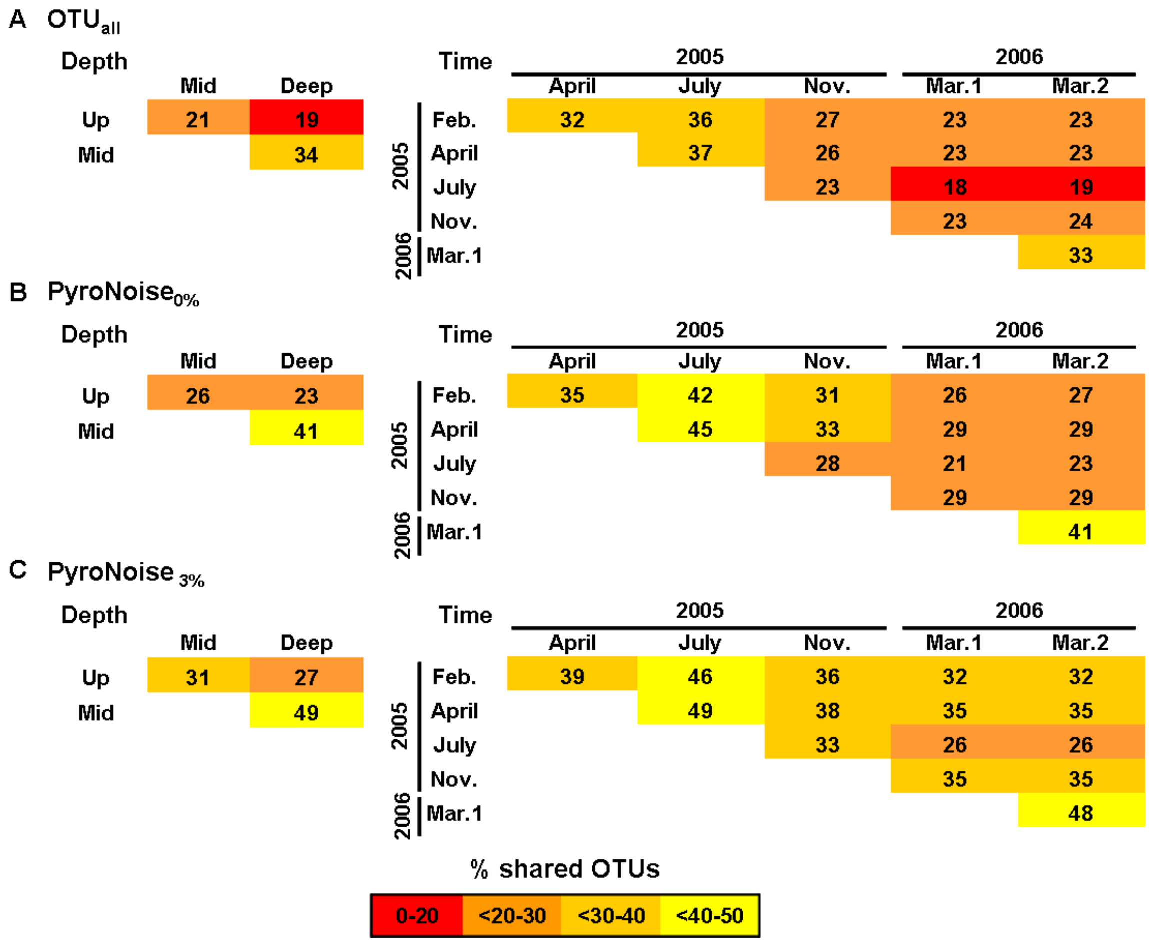


Fig. S4.3 Total number of shared OTU between the three compartments: Sand grain-associated biofilm, sand porewater, and overlying water column. (**A**) $\text{OTU}_{\text{unique}}$, (**B**) PyroNoise-corrected $\text{OTU}_{3\%}$. Each entire circle represents the total number of OTU in a given compartment (here, one sample).



A OTU$_{all}$

Depth	Mid	Deep
Up	21	19
Mid		34

Time		2005			2006	
		April	July	Nov.	Mar.1	Mar.2
2005	Feb.	32	36	27	23	23
	April		37	26	23	23
	July			23	18	19
	Nov.				23	24
2006	Mar.1					33

B PyroNoise$_{0\%}$

Depth	Mid	Deep
Up	26	23
Mid		41

Time		2005			2006	
		April	July	Nov.	Mar.1	Mar.2
2005	Feb.	35	42	31	26	27
	April		45	33	29	29
	July			28	21	23
	Nov.				29	29
2006	Mar.1					41

C PyroNoise$_{3\%}$

Depth	Mid	Deep
Up	31	27
Mid		49

Time		2005			2006	
		April	July	Nov.	Mar.1	Mar.2
2005	Feb.	39	46	36	32	32
	April		49	38	35	35
	July			33	26	26
	Nov.				35	35
2006	Mar.1					48

Fig. S4.4. Turnover of the bacterial community between sediment depth layers or sampling dates after PyroNoise correction and successive OTU clustering of the 454 MPTS data. (**A**) OTU$_{all}$, (**B**) PyroNoise$_{0\%}$, (**C**) PyroNoise$_{3\%}$. The percentage of OTU shared between a sampling depth (or date) and the previous one was calculated and values were represented according to heatmap matrices. OTU$_{all}$ represents the original data set with all OTU$_{unique}$, used here as a reference to test for the effects of correction and clustering on the resolution of bacterial community dynamics.

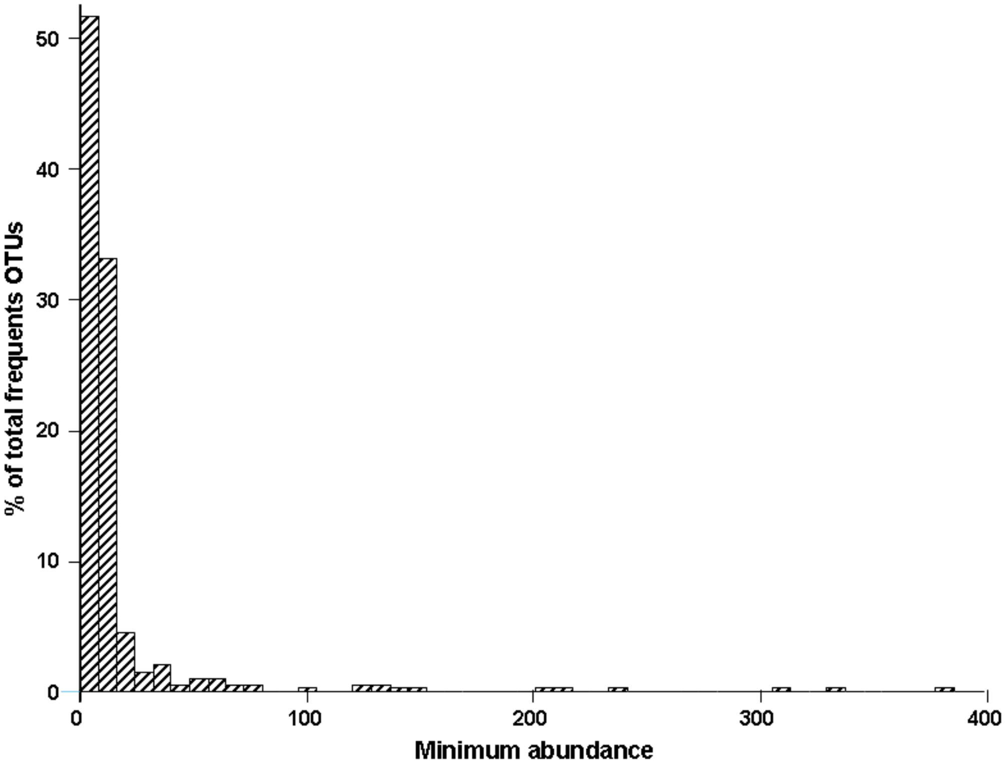
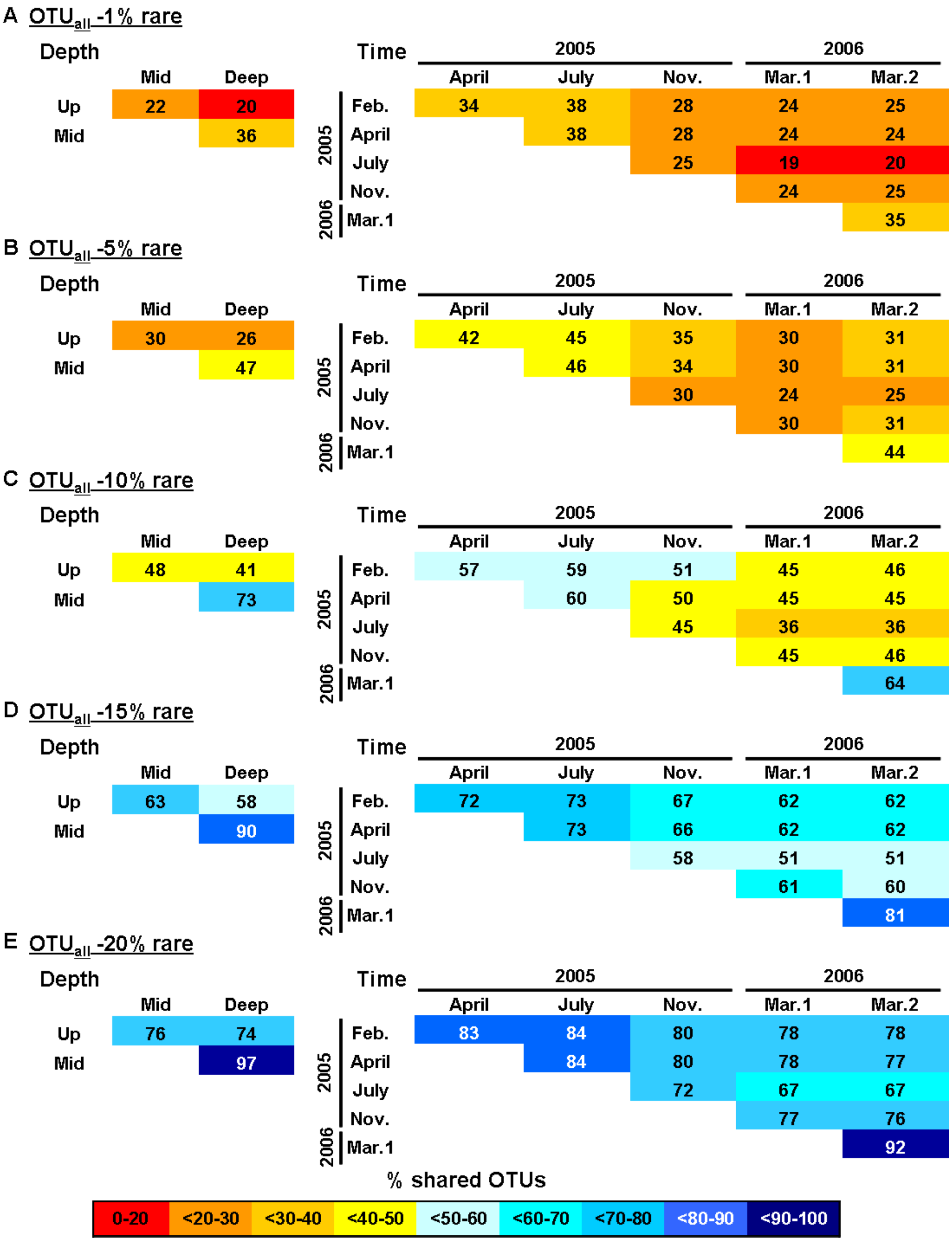


Fig. S4.5. Variation in minimum sequence abundance among resident $OTU_{0\%}$ (*i.e.* $OTU_{0\%}$ present at all times) in the top 10 cm sediment layer. All data were initially processed to remove pyrosequencing noise



A OTU$_{all}$-1% rare

Depth	Mid	Deep
Up	22	20
Mid		36

Time		2005			2006	
		April	July	Nov.	Mar.1	Mar.2
2005	Feb.	34	38	28	24	25
	April		38	28	24	24
	July			25	19	20
	Nov.				24	25
2006	Mar.1					35

B OTU$_{all}$-5% rare

Depth	Mid	Deep
Up	30	26
Mid		47

Time		2005			2006	
		April	July	Nov.	Mar.1	Mar.2
2005	Feb.	42	45	35	30	31
	April		46	34	30	31
	July			30	24	25
	Nov.				30	31
2006	Mar.1					44

C OTU$_{all}$-10% rare

Depth	Mid	Deep
Up	48	41
Mid		73

Time		2005			2006	
		April	July	Nov.	Mar.1	Mar.2
2005	Feb.	57	59	51	45	46
	April		60	50	45	45
	July			45	36	36
	Nov.				45	46
2006	Mar.1					64

D OTU$_{all}$-15% rare

Depth	Mid	Deep
Up	63	58
Mid		90

Time		2005			2006	
		April	July	Nov.	Mar.1	Mar.2
2005	Feb.	72	73	67	62	62
	April		73	66	62	62
	July			58	51	51
	Nov.				61	60
2006	Mar.1					81

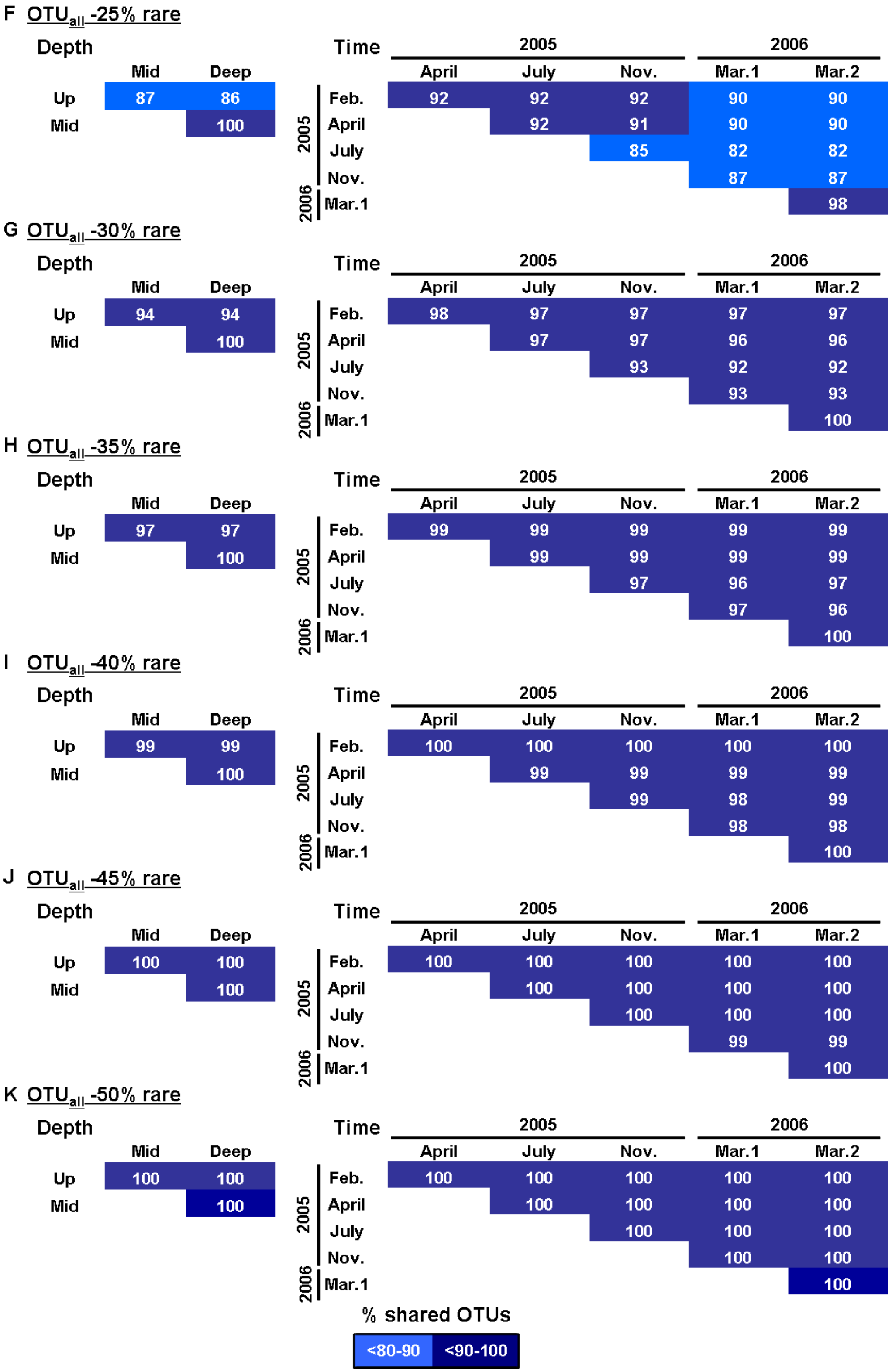
E OTU$_{all}$-20% rare

Depth	Mid	Deep
Up	76	74
Mid		97

Time		2005			2006	
		April	July	Nov.	Mar.1	Mar.2
2005	Feb.	83	84	80	78	78
	April		84	80	78	77
	July			72	67	67
	Nov.				77	76
2006	Mar.1					92

% shared OTUs

0-20	<20-30	<30-40	<40-50	<50-60	<60-70	<70-80	<80-90	<90-100



F OTU_{all}-25% rare

Depth	Mid	Deep
Up	87	86
Mid		100

Time		2005			2006	
		April	July	Nov.	Mar.1	Mar.2
2005	Feb.	92	92	92	90	90
	April		92	91	90	90
	July			85	82	82
	Nov.				87	87
2006	Mar.1					98

G OTU_{all}-30% rare

Depth	Mid	Deep
Up	94	94
Mid		100

Time		2005			2006	
		April	July	Nov.	Mar.1	Mar.2
2005	Feb.	98	97	97	97	97
	April		97	97	96	96
	July			93	92	92
	Nov.				93	93
2006	Mar.1					100

H OTU_{all}-35% rare

Depth	Mid	Deep
Up	97	97
Mid		100

Time		2005			2006	
		April	July	Nov.	Mar.1	Mar.2
2005	Feb.	99	99	99	99	99
	April		99	99	99	99
	July			97	96	97
	Nov.				97	96
2006	Mar.1					100

I OTU_{all}-40% rare

Depth	Mid	Deep
Up	99	99
Mid		100

Time		2005			2006	
		April	July	Nov.	Mar.1	Mar.2
2005	Feb.	100	100	100	100	100
	April		99	99	99	99
	July			99	98	99
	Nov.				98	98
2006	Mar.1					100

J OTU_{all}-45% rare

Depth	Mid	Deep
Up	100	100
Mid		100

Time		2005			2006	
		April	July	Nov.	Mar.1	Mar.2
2005	Feb.	100	100	100	100	100
	April		100	100	100	100
	July			100	100	100
	Nov.				99	99
2006	Mar.1					100

K OTU_{all}-50% rare

Depth	Mid	Deep
Up	100	100
Mid		100

Time		2005			2006	
		April	July	Nov.	Mar.1	Mar.2
2005	Feb.	100	100	100	100	100
	April		100	100	100	100
	July			100	100	100
	Nov.				100	100
2006	Mar.1					100

Fig. S4.6. Turnover of the bacterial community between sediment depth layers or sampling dates after applying MultiCoLA. Successive percentages of rare OTU_{unique} were removed from the OTU_{all} data set: (**A**) 1%, (**B**) 5%, (**C**) 10%, (**D**) 15%, (**E**) 20%, (**F**) 25%, (**G**) 30%, (**H**) 35%, (**I**) 40%, (**J**) 45% and (**K**) 50%. See Fig. S4 for further details.

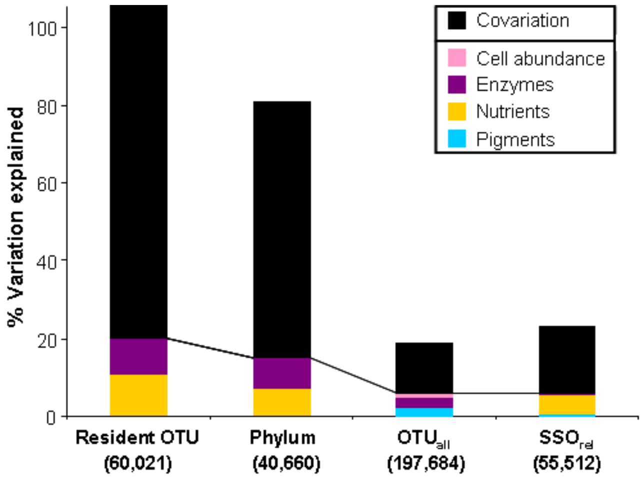


Fig. S4.7. Partitioning of the biological variation in the bacterial community structure as a function of explanatory parameters based on the Phylum level, OTU_{all}, resident OTU and SSO_{rel} data sets. Environmental parameters accounted for included pigments (chlorophyll *a* and pheophytin), nutrients (silicate, phosphate, nitrite, nitrate, ammonium), extra-cellular enzyme activities (chitinase, α-glucosidase, β-glucosidase, lipase, aminopeptidase, phosphatase), cell abundance and their combined effects. The black line in each panel separates the pure factor effects from their covariations. Covariation of any of the 4 environmental factors is represented under one category "Covariation". Here, the OTU_{all} level includes also sequences without complete annotation. The total number of sequences in each data set is indicated in parentheses.

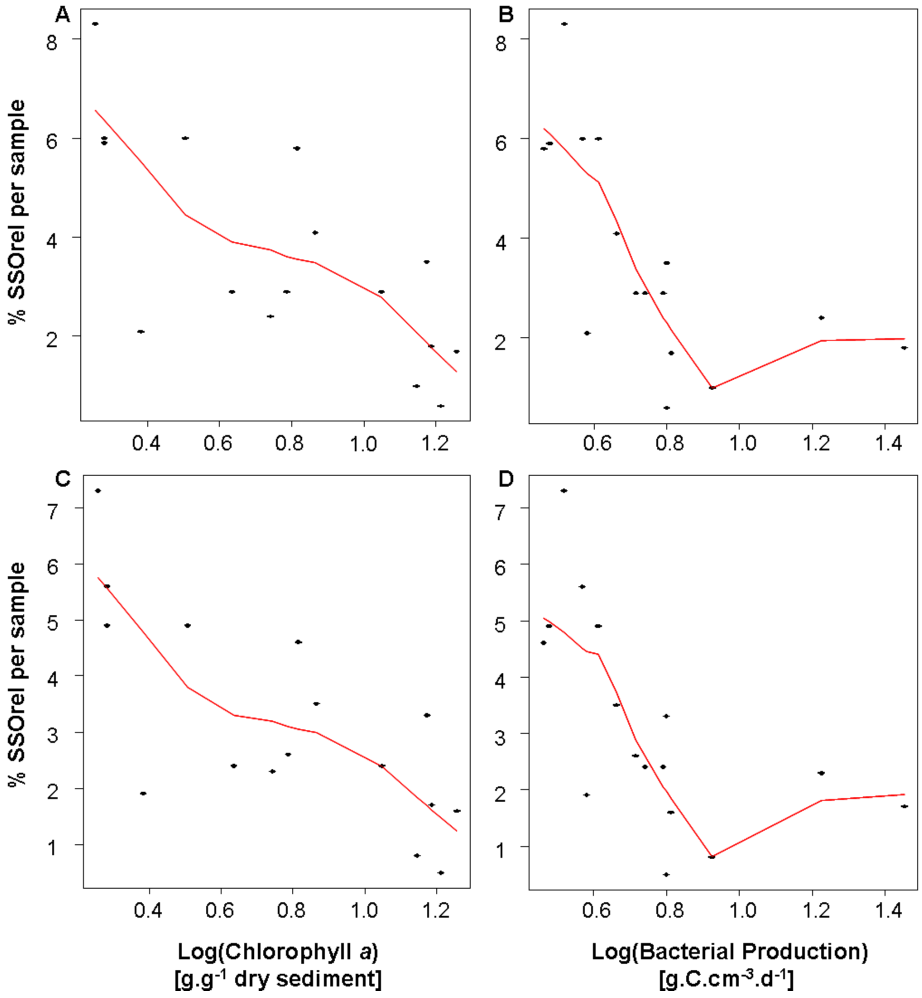


Fig. S4.8. Effects of environmental conditions on the proportion of SSO_{rel} per sample. PyroNoise-corrected data were clustered to define OTU at (**A**, **B**) 0% and (**C**, **D**) 3% sequence dissimilarity levels. The red line in each plot represents the best local fitting regression line between the variables.

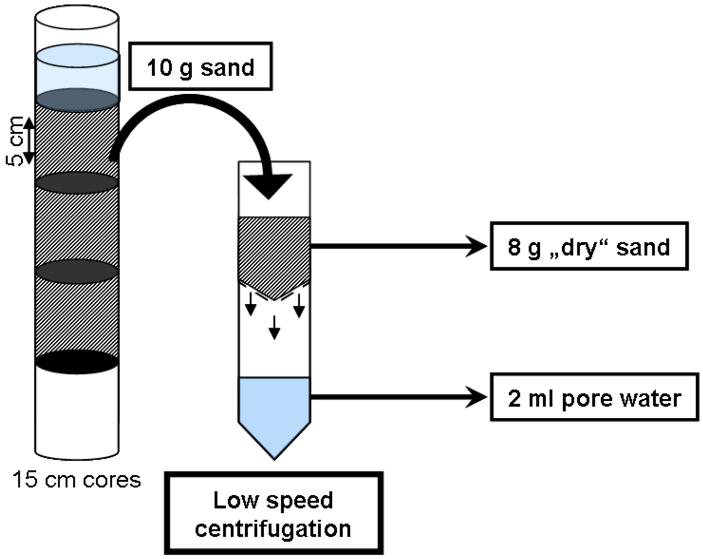


Fig. S4.9. Experimental set-up to separate the pore water from sand grains. 10 g of the upper 5 cm of the sediment were put in a Falcon assembly before low speed centrifugation. 8 g of sand grains were separated from 2 ml of pore water.

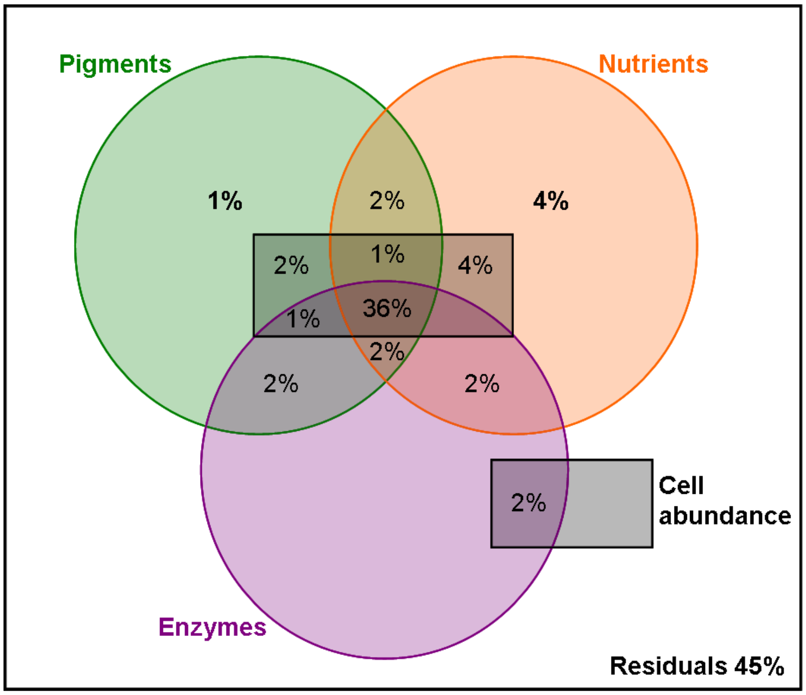


Fig. S4.10. Partitioning of the biological variation in the bacterial community structure of potential pathogens (including *Parachlamydia*, *Arcobacter*, *Francisella*, *Acinetobacter*, *Rickettsiella*, *Pseudomonas*, *Ralstonia*). Environmental parameters accounted for include pigments (chlorophyll *a*), nutrients (silicate, phosphate, nitrate, ammonium), extra-cellular enzyme activities (α-glucosidase, lipase, phosphatase), cell abundance and their combined effects. Unexplained variation is not shown.

4.4.3 Supplementary Tables

Table S4.1. Summary of diversity estimators of richness for all samples at three different sediment depths and in the water column at different sampling times.

Table S4.2. Contribution of environmental parameters to the variation in the Phylum level, OTU_{all}, resident OTU and SSOrel data sets.

Table S4.3. Percentages of SSO_{rel} in each sample at different levels of OTU definition (PyroNoise-corrected data).

Table S4.4. Average values of environmental parameters for 5-cm sediment intervals in Sylt water column and sandy sediment.

Table S4.5. Cell-specific extracellular enzymatic activities, bacterial abundances and bacterial carbon production rates in Sylt sandy sediment.

Table S4.6. Contribution of environmental parameters to the variation in sequences potentially affiliated with pathogens.

Table S4.1. Total OTU number for all samples at three different sediment depths and in the water column at different sampling times, for the raw OTU$_{all}$ and the PyroNoise-corrected data sets at different percentages of sequence clustering (OTU$_{0\%}$ and OTU$_{3\%}$).

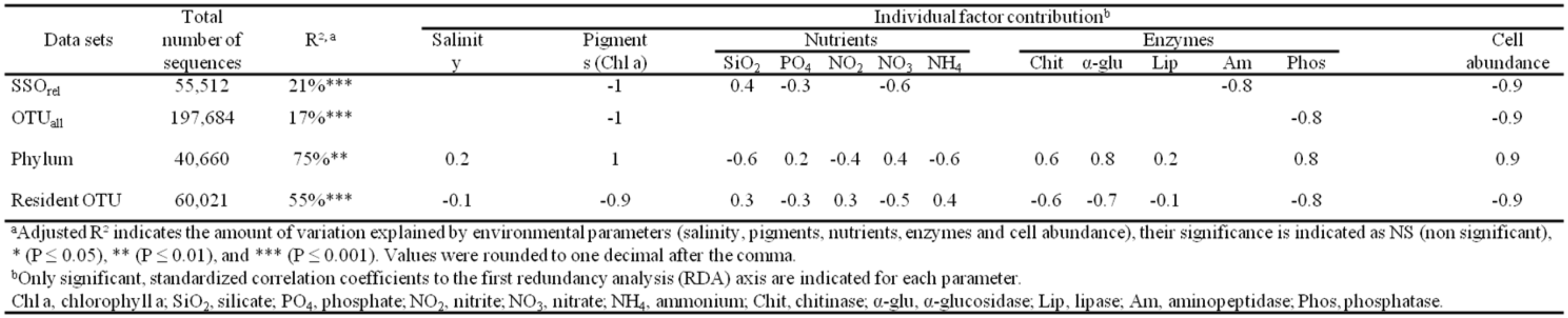
0-5 cm	Average over time	February 2005	April 2005	July 2005	November 2005	March1 2006	March2 2006
Total number of V6 sequences	8,827 ±2,398	8,527	4,722	9,035	12,153	8,856	9,667
Total OTU$_{unique}$	2,029 ±613	1,660	1,042	2,081	2,747	2,518	2,126
Total OTU$_{0\%}$	1,071±400	784	505	1,068	1,539	1,492	1,035
Total OTU$_{3\%}$	1,036±384	763	496	1,036	1,487	1,442	994

5-10 cm							
Total number of V6 sequences	14,736 ±4,717	16,197	9,648	8,044	18,948	16,770	18,806
Total OTU$_{unique}$	3,648 ±988	3,228	2,606	2,521	4,605	4,694	4,231
Total OTU$_{0\%}$	1,962±528	1,639	1,487	1,402	2,469	2,643	2,133
Total OTU$_{3\%}$	1,865±470	1,568	1,436	1,373	2,301	2,473	2,036

10-15 cm							
Total number of V6 sequences	14,078 ±5,118	7,132	13,593	nd	nd	16,672	18,914
Total OTU$_{unique}$	4,279 ±1,369	2,408	4,194	nd	nd	4,935	5,577
Total OTU$_{0\%}$	2,370±840	1,207	2,476	nd	nd	2,586	3,210
Total OTU$_{3\%}$	2,235±761	1,179	2,341	nd	nd	2,426	2,993

	Sand (0-5 cm)		
April 2008	Sediment	Pore water	Water column
Total number of V6 sequences	18,157	9,726	10,557
Total OTU$_{unique}$	3,806	1,940	1,788
Total OTU$_{0\%}$	2,830	1,376	1,336
Total OTU$_{3\%}$	2,697	1,189	1,213

Table S4.2. Contribution of environmental parameters to the variation in the Phylum level, OTU_{all}, resident OTU and SSOrel data sets.



	Total		Individual factor contribution[b]												
Data sets	number of	$R^{2,a}$	Salinit	Pigment	Nutrients					Enzymes					Cell
	sequences		y	s (Chl a)	SiO_2	PO_4	NO_2	NO_3	NH_4	Chit	α-glu	Lip	Am	Phos	abundance
SSO_{rel}	55,512	21%***		-1	0.4	-0.3		-0.6					-0.8		-0.9
OTU_{all}	197,684	17%***		-1										-0.8	-0.9
Phylum	40,660	75%**	0.2	1	-0.6	0.2	-0.4	0.4	-0.6	0.6	0.8	0.2		0.8	0.9
Resident OTU	60,021	55%***	-0.1	-0.9	0.3	-0.3	0.3	-0.5	0.4	-0.6	-0.7	-0.1		-0.8	-0.9

[a] Adjusted R^2 indicates the amount of variation explained by environmental parameters (salinity, pigments, nutrients, enzymes and cell abundance), their significance is indicated as NS (non significant), * ($P \leq 0.05$), ** ($P \leq 0.01$), and *** ($P \leq 0.001$). Values were rounded to one decimal after the comma.

[b] Only significant, standardized correlation coefficients to the first redundancy analysis (RDA) axis are indicated for each parameter.

Chl a, chlorophyll a; SiO_2, silicate; PO_4, phosphate; NO_2, nitrite; NO_3, nitrate; NH_4, ammonium; Chit, chitinase; α-glu, α-glucosidase; Lip, lipase; Am, aminopeptidase; Phos, phosphatase.

Table S4.4. Percentages of SSO_{rel} in each sample at different levels of OTU definition (PyroNoise-corrected data).

Sampling date	Sediment layer [cm]	0%	3%	5%	10%
February 2005	0-5	1	0.8	0.6	0.5
	5-10	2.9	2.4	1.7	1.2
	10-15	2.1	1.9	1.6	1.2
April 2005	0-5	0.6	0.5	0.4	0.3
	5-10	2.9	2.6	2.4	1.8
	10-15	6	5.6	4.9	3.9
July 2005	0-5	1.8	1.7	1.1	0.8
	5-10	2.4	2.3	1.9	1.7
November 2005	0-5	3.5	3.3	2.7	2.4
	5-10	5.8	4.6	4.3	3.1
March1 2006	0-5	2.9	2.4	2.2	1.9
	5-10	6	4.9	4.1	3.4
	10-15	5.9	4.9	4.4	3.3
March2 2006	0-5	1.7	1.6	1.2	1
	5-10	4.1	3.5	2.8	2
	10-15	8.3	7.3	6.3	5.2

Table S4.4. Average values of environmental parameters for 5-cm sediment intervals in Sylt water column and sandy sediment (Böer et al. 2009).

	Water column			Sediment								
Sampling date	Temp. [°C]	pH	Wind speed [m.s^{-1}]	Sediment layer [cm]	Salinity [°/°°]	Pigments [µg.g^{-1}]		Nutrients [µM]				
						Chl a	Pheo	SiO_2	PO_4	NO_2	NO_3	NH_4
February 2005	1.5	8.00	6.67	0-5	25.4	13.0	1.6	67.4	6.7	1.6	68.3	23.4
				5-10	27.1	3.3	1.2	60.0	9.5	2.6	10.0	55.5
				10-15	26.1	1.4	0.8	71.4	4.6	1.8	5.7	48.0
April 2005	6.9	8.04	11.02	0-5	23.2	15.3	1.2	66.4	5.8	2.2	56.6	22.2
				5-10	20.3	5.1	0.9	143.1	6.9	3.5	13.5	59.6
				10-15	19.8	0.9	0.6	170.7	7.3	2.6	5.6	61.8
July 2005	18.0	8.03	4.73	0-5	30.0	14.4	0.9	15.0	10.9	0.9	3.3	37.6
				5-10	30.0	4.5	0.7	21.2	7.3	0.7	2.0	55.0
November 2005	10.8	8.08	9.50	0-5	29.3	13.9	1.1	9.7	6.0	0.4	4.2	31.7
				5-10	30.0	5.5	1.1	17.9	3.1	0.8	4.0	67.8
March1 2006	1.7	7.90	5.65	0-5	27.8	10.2	1.3	36.1	2.7	1.6	28.5	21.6
				5-10	28.2	2.2	0.9	48.2	3.3	1.6	8.6	33.6
				10-15	28.0	0.9	0.4	55.4	2.6	1.7	9.6	41.8
March2 2006	4.1	8.13	8.82	0-5	26.4	17.0	1.2	8.2	3.0	1.9	42.7	10.7
				5-10	28.0	6.3	1.1	25.9	5.2	2.1	4.1	46.7
				10-15	27.8	0.8	0.3	34.1	4.1	2.6	3.0	38.8

The different parameters were Chl a, chlorophyll a; Pheo, pheophytin; SiO_2, silicate; PO_4, phosphate; NO_2, nitrite; NO_3, nitrate; NH_4, ammonium.

Table S4.5. Cell-specific extracellular enzymatic activities, bacterial abundances and bacterial carbon production rates in Sylt sandy sediment (see Böer et al. 2009).

Sampling date	Sediment layer [cm]	Enzymes [amol.h^{-1}.cell]						Bacterial properties	
		Chit	α-glu	β-glu	Lip	Amin	Phos	BCC [cells.10^{9}.cm^{-3}]	Bprod [mg C.L^{-1}.d^{-1}]
February 2005	0-5	0.8	0.6	1.0	0.0	14	7	2.4	7.4
	5-10	2.2	0.4	0.8	0.1	7	6	1.3	4.5
	10-15	0.3	0.2	0.3	0.0	3	2	0.9	2.8
April 2005	0-5	0.9	0.8	1.6	0.0	20	7	2.0	5.3
	5-10	0.9	0.5	1.0	0.0	14	5	1.4	4.2
	10-15	0.5	0.3	0.6	0.1	8	3	0.6	2.7
July 2005	0-5	3.4	1.6	2.7	0.3	35	14	2.9	27.4
	5-10	1.2	0.8	2.4	0.3	26	9	2.0	15.8
November 2005	0-5	1.5	0.9	1.3	0.1	20	10	2.3	5.3
	5-10	1.2	0.5	1.3	0.1	13	8	1.7	1.9
March1 2006	0-5	1.3	0.4	0.6	0.1	8	6	1.7	5.2
	5-10	0.9	0.3	0.5	0.1	4	7	0.8	3.1
	10-15	0.4	0.2	0.5	0.0	3	4	0.6	2.0
March2 2006	0-5	0.5	0.5	0.8	0.0	17	7	2.1	5.5
	5-10	0.9	0.2	0.4	0.0	8	4	1.4	3.6
	10-15	0.4	0.1	0.4	0.0	5	3	0.5	2.3

The different parameters were Chit, chitinase; α-glu, α-glucosidase; β-glu , β-glucosidase; Lip, lipase; Amin, aminopeptidase; Phos, phosphatase; BCC, bacterial abundance; Bprod, bacterial carbon production.

Table S4.6. Contribution of environmental parameters to the variation in sequences potentially affiliated with pathogens.

	Total number of sequences	$R^{2,a}$	Individual factor contribution[b]														
			Salinity	Pigments		Nutrients					Enzymes						Cell abundance
				Chl a	Pheo	SiO_2	PO_4	NO_2	NO_3	NH_4	Chit	α-glu	β-glu	Lip	Amin	Phos	
Pathogens	426	55%**	0.1	1		-0.4	0.2		0.6	-0.6		0.7		0.2		0.8	0.9

[a]Adjusted R^2 indicates the amount of variation explained by environmental parameters (salinity, pigments, nutrients, enzymes and cell abundance), their significance is indicated as ** ($P \leq 0.01$). Values were rounded to one decimal after the comma.

[b]Only significant, standardized correlation coefficients to the first redundancy analysis (RDA) axis are indicated for each parameter.

Chl a, chlorophyll a; Pheo, Pheophytin; SiO_2, silicate; PO_4, phosphate; NO_2, nitrite; NO_3, nitrate; NH_4, ammonium; Chit, chitinase; α-glu, α-glucosidase; β-glu, β-glucosidase; Lip, lipase; Amin, aminopeptidase; Phos, phosphatase.

5 General Discussion

As introduced earlier, community ecology concepts started with plant and animals studies, without considering microorganisms. Microbial community ecology is thus a recent field of study and the processes shaping the microbial community structure are still not well understood. Studying microbial community ecology requires the comprehension of interactions between microorganisms, as well as the impact of time, space and the environment on it. This PhD thesis focused on explaining the processes shaping microbial community structure by studying temperate subtidal coastal sands, a highly dynamic ecosystem constantly influenced by the overlying water column where environmental conditions are well characterized, as a case study.

5.1 Towards the Complementary Use of Classical and Next-Generation Molecular Ecology Tools

5.1.1 Consistency of Molecular Tools

One important aim of this thesis work was to allow a better use and comprehension of the possibilities made available through emergent molecular tools. The parallel application of two time- and cost-effective techniques: the traditional fingerprinting technique ARISA and the high-throughput 454 MPTS, led to similar ecological patterns (**Chapter 1**) with still some differences in targeting certain fractions of the microbial community. ARISA may describe patterns of resident bacterial types, which are present in the community at all times (**Chapter 1**), while 454 massively parallel tag sequencing (MPTS) may allow higher resolution into identifying dominant, resident and rare microbes (**Chapters 1 and 2**).

5.1.2 Towards a More Thorough Interpretation of High-Throughput Data Sets

High-throughput molecular tools are being developed at an unprecedented rate, and the data output is accumulating without any clear pipeline available to extract the deeper meaning[2] from it. A systematic way to analyze the impressive 454 MPTS data output was made available to microbial ecologists through the implementation of a software user-friendly (MultiCoLA, **Chapter 2**).

5.1.3 Consistency of the Available Taxonomy

When comparing ARISA data with 454 MPTS taxonomy or successive truncated data sets, similar ecological patterns could be observed for all taxonomic levels (phylum to the

[2] « *En extraire la substantifique moelle* », Rabelais.

genus level), and ARISA data (**Chapter 1**). This indicated first the consistency of ecological patterns obtained with ARISA and, second, it validated the ecological meaning resulting from the available taxonomic annotation. In addition, the taxonomic annotation available may represent the dominant part of the bacterial community (**Chapters 1 and 2**). This also follows a recent statement regarding the ecological coherence of high bacterial taxonomic ranks (*i.e.* phylum to the genus level), suggesting that members of a same taxon share the same main functions in the ecosystem (Philippot et al. 2010).

5.2 On the Commonness and Rarity of Microbes: Identification, Distribution and Ecological Patterns

5.2.1 Defining Rare and Dominant Types in the Microbial Community

Whereas dominant types are usually distinguished from rare types in most community ecology studies, there is no clear definition of each of these fractions of the microbial community. In a recent 454 MPTS study (Galand et al. 2009a), rare (<0.01% within a sample) or abundant types (>1% within a sample) in the community were defined by applying the arbitrary cutoff proposed by (Pedrós-Alió 2006). Another 454 MPTS study does not even give any definitions, rare bacterial types are discriminated from the abundant ones only depending on the large variations in sequence proportion within a sample (Brazelton et al. 2010). The systematic truncation of defined proportions of 454 MPTS data sets permitted to distinguish rare from dominant bacterial types (**Chapter 2**). The microbial community was composed of few dominant, few resident (about 5% microbial types present at all times) and many rare types (*e.g.* about 30% of OTU appearing only once in a given sample, SSO_{rel} and 50% of OTU appearing only once in the whole data set, SSO_{abs}, **Chapter 3**). Interestingly, most rare bacterial types stayed rare and most abundant stayed abundant over the 2-year period (2005-2006), as observed over 6 years in the Arctic ocean, where 99% of the rare bacterial OTU were always rare (Galand et al. 2009a, Kirchman et al. 2010).

5.2.2 Diversity Patterns of Rare and Dominant Types in the Microbial Community

Comparisons between the original 454 MPTS data set and the truncated ones allowed the assessment of its impact on the resulting ecological interpretation in the microbial community (**Chapter 2**). Consistent ecological patterns could be kept until a truncation of up to 40% of rare types in the data set and this indicated that dominant bacterial types may maintain the function in the community (**Chapter 2**). Accordingly, despite representing less than 5% of bacterial types in the data set, resident types seemed to maintain the community structure as similar combinations of biogeochemical parameters could explain their distribution and that of the whole community (**Chapter 3**). Despite presenting different ecological patterns than that of resident types, the distribution of rare bacterial types was not random as it could be explained by biogeochemical parameters (**Chapter 3**). Indeed, rare types had a major role in the really high turnover of the microbial community across sediment depth and through sampling time (**Chapter 3**). The ecological meaning of such fluctuations has yet to be determined.

5.2.3 Potential Processes Influencing the Structuring of the Rare Biosphere

If we refer to community ecology concepts to explain the structuring of a community, not all of the four major processes (Vellend 2010) may apply in the current study.

In **chapter 3**, we could observe that rare bacterial types always stayed rare and abundant types always stayed abundant through a two-year period. However, a recent 454 MPTS study in hydrothermal vents indicated that rare may become abundant over a longer period (*i.e.* a thousand years), due to changing conditions of the chimneys (Brazelton et al. 2010). The hypothesis of the seed-bank in dormancy waiting for appropriate conditions to develop may thus apply in this study.

Also, we observed that the distribution of abundant, resident and rare bacterial types were not random as specific combinations of biogeochemical parameters could explain it (**Chapters 2 and 3**). This agrees with the Brazelton's hypothesis stating that environmental conditions influence the distribution of microbes on the chimneys (Brazelton et al. 2010). Consequently, this does not follow the hypothesis of an ecological drift (stochastic changes in species abundance) but rather events of **selection** (*e.g.* bottom-up or top-down influences). Notably, a high turnover of bacterial types was observed with time and across depth, and this may be due to high migration events due to **dispersal**. However, as temperate coastal sands represent highly dynamic ecosystems with lots of mixing, the concept of speciation may not apply here.

To conclude, it seems that deterministic processes (*e.g.* competition) could also explain the structuring of microbial communities. Also, as proposed by Sloan and colleagues (2006), Hubbel's neutral community model may apply here to explain microbial community structuring (Sloan et al. 2006). Indeed, microbial community structure may be influenced by stochastic immigration and birth-death processes. **Chance** and **immigration** may thus be important factors to shape microbial communities.

5.3 Spatial Study of Microbial Communities in Temperate Coastal Sediments

Microbial community ecology in temperate coastal sands of the North Sea island Sylt is still not well understood. Two molecular studies have shown the diversity and abundance of main bacterial groups (Musat et al. 2006), and the fluctuations of the community structure through time and depth (Böer et al. 2009).

This whole PhD thesis is based on a sampling design consisting of a single area, where sandy cores were pooled together for a better chance of recovering the diversity, sampled over a two-year period. To complement this thesis work and further study the biogeography of microbial communities in temperate coastal sediments, an additional sampling was carried out at several locations of Sylt's coastal environment. Samples were taken by following gradients depending on:

(i) the location on the shore: samples may be taken on the intertidal, perpendicular and parallel to the tide, so as to study the effect of, for instance, the waves or desiccation on the microbial communities,

(ii) grain size, which may be an important factor in structuring microbial communities as each sediment type may offer a specific microbial habitat, from biofilm-dominated porous sands, to diffusion-limited organic rich muds, and

(iii) sampling depth, as oxic and anoxic conditions may vary depending on the distance to the sea as well as grain size.

DNA was extracted from three compartments defined from the coastal environment: sediment grains' biofilms, pore water and the overlying water column. ARISA and 454 MPTS were conducted on these samples and several contextual environmental parameters were measured (*e.g.* salinity, pigments, nutrients, extra-cellular enzyme activities, porosity, permeability, grain size). As observed in the **third chapter**, preliminary ARISA results indicated gradients in microbial communities from smaller to bigger grain sizes (*i.e.* mud, mix of mud and sand, sand), which may also reflect gradients or distances to the shore (**Fig. 5.1.A**). Notably, NMDS analysis indicated that sediment containing a mix of mud and sand had a similar microbial community structure as in the sand while the mud's microbial community was significantly different. Also, there were clear differences between the sediment (dry sediment, without pore water and wet sediment, representing dry sediment and pore water), the pore water and the water column (**Fig. 5.1.B**). The pore water's microbial community was significantly different from that of the sediment but shared 32% similarity with water column, as tested with ANOSIM.

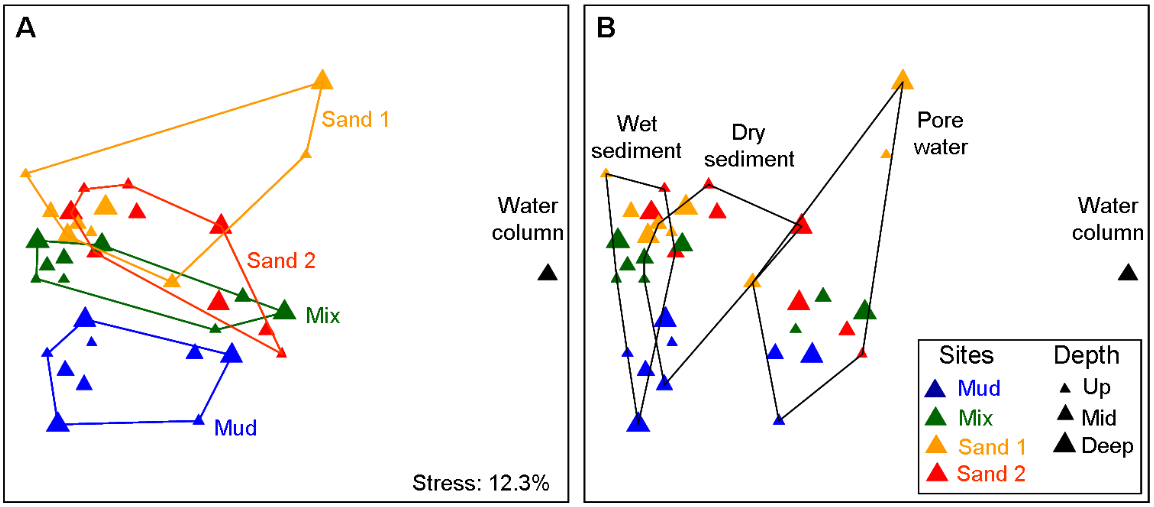


Figure 5.1. Comparison of sediment samples similarities. Samples were grouped according to (**A**) their grain type, or (**B**) their compartment. Analyses of similarities (ANOSIM) tested for differences between samples grouped per (**A**) grain type, or (**B**) compartment (P < 0.05).

When studying the effect of environmental parameters on the structuring of the microbial community, the factors space and depth indicated the importance of spatial location in structuring the community and grain sizes indicated the significant effect of habitat structure (**Fig. 5.2.**). As observed earlier (**Chapter 3**), the pure effect of compartments [*e.g.* the separation of pore water from dry sediment (containing microbial-associated biofilms)] was explaining most of the biological variation (**Fig. 5.2.**). Clearly, variations in the microbial community structure are mainly influenced by sediment type together with spatial coordinates – which may correspond to hydrodynamic influences – and may represent specific microbial habitats (**Fig. 5.2.**).

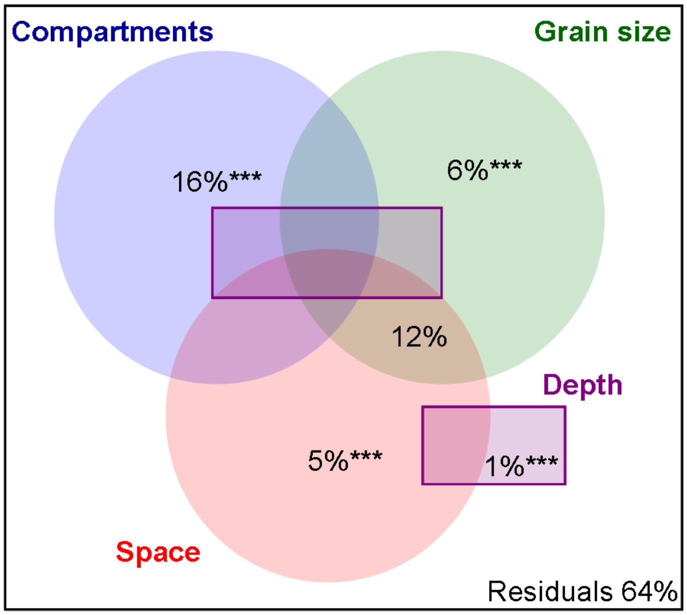


Figure 5.2. Partitioning of the biological variation in the bacterial community structure. Environmental parameters accounted for include compartments (dry and wet sediment, pore water, water column), grain size (mud, mix of mud and sand, sand), space and depth (0-5, 5-10, 10-15 cm). Unexplained variation is not shown (stars indicate P < 0.001).

5.4 Conclusion

As high-throughput molecular techniques allow the accumulation of data, it became clear that improvements were needed to extract the most information out of such data. A deeper understanding of the ecology of dominant and rare fractions of microbial communities in temperate coastal sands was made possible by:

(i) validating classical molecular tools to describe microbial community structure and comprehend microbial ecological patterns (**Chapter 1**),

(ii) implementing new user-friendly statistical tools to analyze complex community data sets (MultiCoLA, www.ecology-research.com), such as high-throughput 454 MPTS data sets (**Chapter 2**),

(iii) testing the reliability of the known taxonomy on the resulting ecological interpretation (**Chapters 1 and 2**),

(iv) determining different fractions (rare, dominant and resident members) of the microbial community and seeking their effect on the overall microbial community structure (**Chapters 2 and 3**),

(v) studying the impact of the environment in shaping the microbial community structure and its different fractions (**Chapters 2 and 3**), and

(vi) improving the characterization of temperate coastal sands and its microbial community ecology (**Chapter 3**).

6 Perspectives

This thesis work shed light on some of the processes that may shape microbial communities but other processes should also be of interest. To complete this study of microbial community ecology, some remaining concepts to investigate are (i) the biogeography of microbial communities and of its different fractions (dominant, resident or rare bacteria), and (ii) the functional redundancy of microbial communities.

6.1 Towards a Better Characterization of Microbial Communities in Temperate Coastal Sediments

All work of this PhD thesis is based on one location and one sediment type, sampled at six irregular sampling times. This study may be improved by higher temporal resolution of the sampling scheme, especially as we observed a high turnover of the community within one month, when stormy conditions were observed (March 2006, **Chapter 3**). It may thus be of interest to sample at higher frequencies, for instance, before, during and after stormy conditions. This may allow an estimation of the impact of rapid changes in environmental conditions on the microbial ecology of sands.

As described in **§ 5.3.**, coastal samples of several sediment types were taken at several locations of the North Sea island Sylt. ARISA results indicated interesting patterns linked to sediment type, compartment type and space. These preliminary analyses coupled with further analyses on the 454 MPTS data set should give promising results to better understand the biogeography of microbial communities in temperate coastal sediments.

6.2 Function of Microbial Communities

The work of this PhD thesis allowed more insights into microbial community ecology, by explaining structuring patterns of microbial communities. However, the understanding of microbial community ecology could be completed if the function of the ecosystem would be further studied (Konopka 2009). To analyze microbial community ecology and ecosystem functioning, determining functional diversity is of great importance. Several possibilities to study functional diversity may be possible. First, one can relate ecological functions to taxonomic groups. However, phylogeny may not be always appropriate as some groups may reflect several biological functions and a single function may also be common to different taxa (Konopka 2009). Fingerprinting techniques targeting a specific functional gene may also be used (Santillano et al. 2010) to assess functional diversity. The limit of such rapid and cost-effective techniques being that the targeted gene should be conserved enough to be representative, and that there are no way to confirm whether the gene of interest is active. Another option would be to make experiments, inducing perturbations to an ecosystem and check for the fluctuations of the gene in correlation with the perturbations (Konopka 2009). This hypothesis may also be biased as, as said earlier, several microbes may present the same functions. A last alternative would be to apply metatranscriptomics or metaproteomics, describing all active genes or proteins from the microbial community, and allowing to relate functions to organisms.

In the case of Sylt's temperate coastal sediments, some interesting functions to study would be sulfate reduction or degradation of complex macromolecules. Sulfate-reduction may be interesting to study as sulfate reduction ranges were found to vary widely (de Beer et al. 2005) and fluorescence *in situ* hybridization [FISH, (Musat et al. 2006)], indicated that sulfate-reducing bacteria represent a dominant group of the community.

7 Annexes

7.1 MultiCoLA Manual (Chapter II)

MultiCoLA

v1.2

Angélique Gobet & Alban Ramette, July 2010

MultiCoLA Manual

Angélique Gobet, Alban Ramette, July 2010
Version 1.2

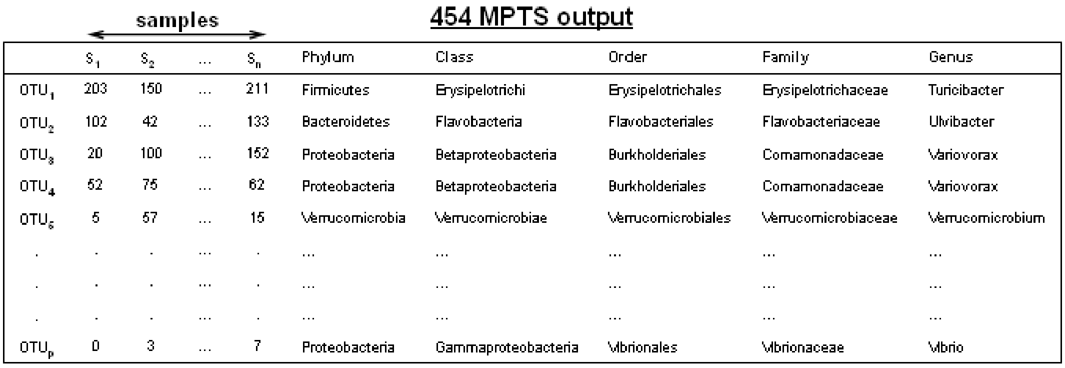
Table of contents

1. **Prepare the input file**

Abundance table with the according taxonomy (e.g. output from the application of 454 massively parallel pyro-tag sequencing (MPTS)): Sample by [OTUs and taxonomy] (abundance matrix) to save as a .txt file, e.g. “input.txt”. In case there is no taxonomic annotation available, the input file can also be an abundance table (e.g. sample by OTUs).

You will find as an example “input.txt” in the .zip file which consists of a simplified 454 MPTS dataset with OTUs abundances and the according taxonomy.

<u>*Example:*</u>



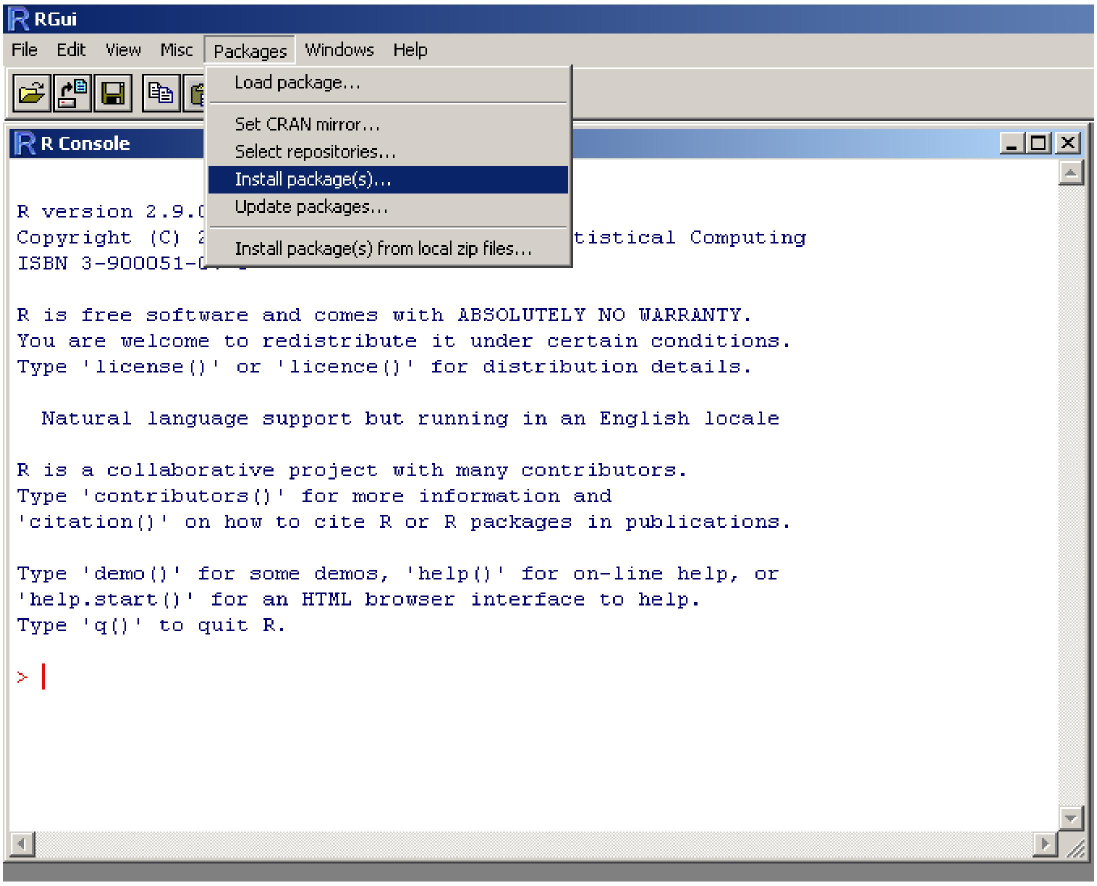
samples

454 MPTS output

	S_1	S_2	...	S_n	Phylum	Class	Order	Family	Genus
OTU_1	203	150	...	211	Firmicutes	Erysipelotrichi	Erysipelotrichales	Erysipelotrichaceae	Turicibacter
OTU_2	102	42	...	133	Bacteroidetes	Flavobacteria	Flavobacteriales	Flavobacteriaceae	Ulvibacter
OTU_3	20	100	...	152	Proteobacteria	Betaproteobacteria	Burkholderiales	Comamonadaceae	Variovorax
OTU_4	52	75	...	62	Proteobacteria	Betaproteobacteria	Burkholderiales	Comamonadaceae	Variovorax
OTU_5	5	57	...	15	Verrucomicrobia	Verrucomicrobiae	Verrucomicrobiales	Verrucomicrobiaceae	Verrucomicrobium
.	.	.	...	.	...	...	...	...	...
.	.	.	...	.	...	...	...	...	...
.	.	.	...	.	...	...	...	...	...
OTU_p	0	3	...	7	Proteobacteria	Gammaproteobacteria	Vibrionales	Vibrionaceae	Vibrio

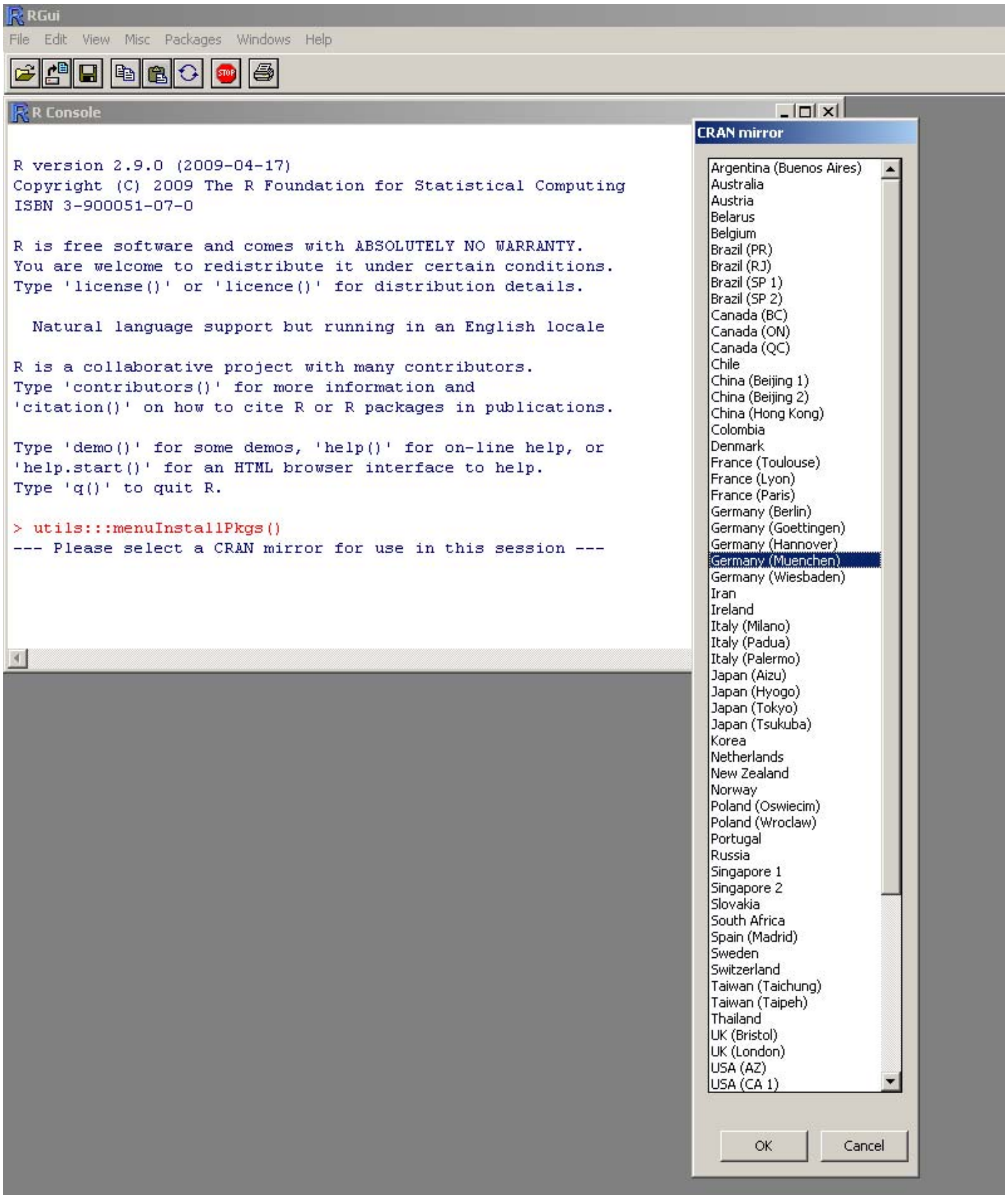
2. **Start the R interface (freely available at: http://www.r-project.org/) and install *vegan* and *MASS* packages**

- Go to “Packages/Install packages”:



```
R version 2.9.0
Copyright (C) 2                   tistical Computing
ISBN 3-900051-0

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> |
```

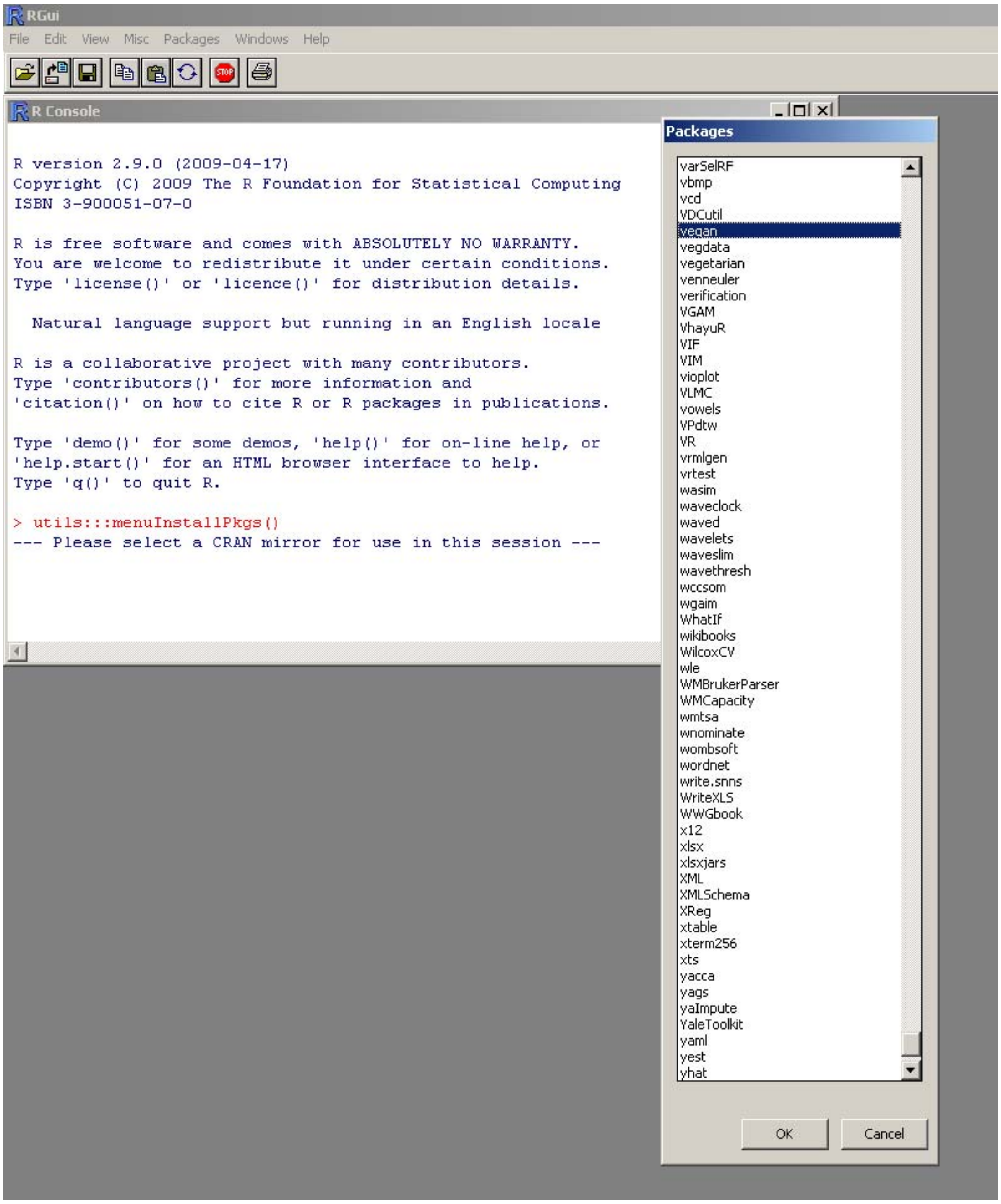
- Select a CRAN mirror closer to the place where you work:



```
R version 2.9.0 (2009-04-17)
Copyright (C) 2009 The R Foundation for Statistical Computing
ISBN 3-900051-07-0

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> utils:::menuInstallPkgs()
--- Please select a CRAN mirror for use in this session ---
```

Argentina (Buenos Aires)
Australia
Austria
Belarus
Belgium
Brazil (PR)
Brazil (RJ)
Brazil (SP 1)
Brazil (SP 2)
Canada (BC)
Canada (ON)
Canada (QC)
Chile
China (Beijing 1)
China (Beijing 2)
China (Hong Kong)
Colombia
Denmark
France (Toulouse)
France (Lyon)
France (Paris)
Germany (Berlin)
Germany (Goettingen)
Germany (Hannover)
Germany (Muenchen)
Germany (Wiesbaden)
Iran
Ireland
Italy (Milano)
Italy (Padua)
Italy (Palermo)
Japan (Aizu)
Japan (Hyogo)
Japan (Tokyo)
Japan (Tsukuba)
Korea
Netherlands
New Zealand
Norway
Poland (Oswiecim)
Poland (Wroclaw)
Portugal
Russia
Singapore 1
Singapore 2
Slovakia
South Africa
Spain (Madrid)
Sweden
Switzerland
Taiwan (Taichung)
Taiwan (Taipeh)
Thailand
UK (Bristol)
UK (London)
USA (AZ)
USA (CA 1)

- Select the package you would like to install (e.g. *vegan*):



```
R version 2.9.0 (2009-04-17)
Copyright (C) 2009 The R Foundation for Statistical Computing
ISBN 3-900051-07-0

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> utils:::menuInstallPkgs()
--- Please select a CRAN mirror for use in this session ---
```

3. Load the data into the R workspace

First, set the directory where you want to work, i.e. where your input file and the series of scripts should be, and where you will find the several outputs from these scripts. The directory should be created beforehand and the name should not contain spaces to be readable by the software R (e.g. use underscore to separate words). The path to the working directory (e.g. "454_MPTS") may be indicated as followed:

```
setwd("C:\\R\\454_MPTS")
```

The data can then be stored into the object M in the workspace:

```
M<-read.table("input.txt",header=TRUE,row.names=1)
```

The scripts can now be used on the sample by OTUs (or taxonomy) matrix M according to the different steps:

4. Load and run the scripts: community structure

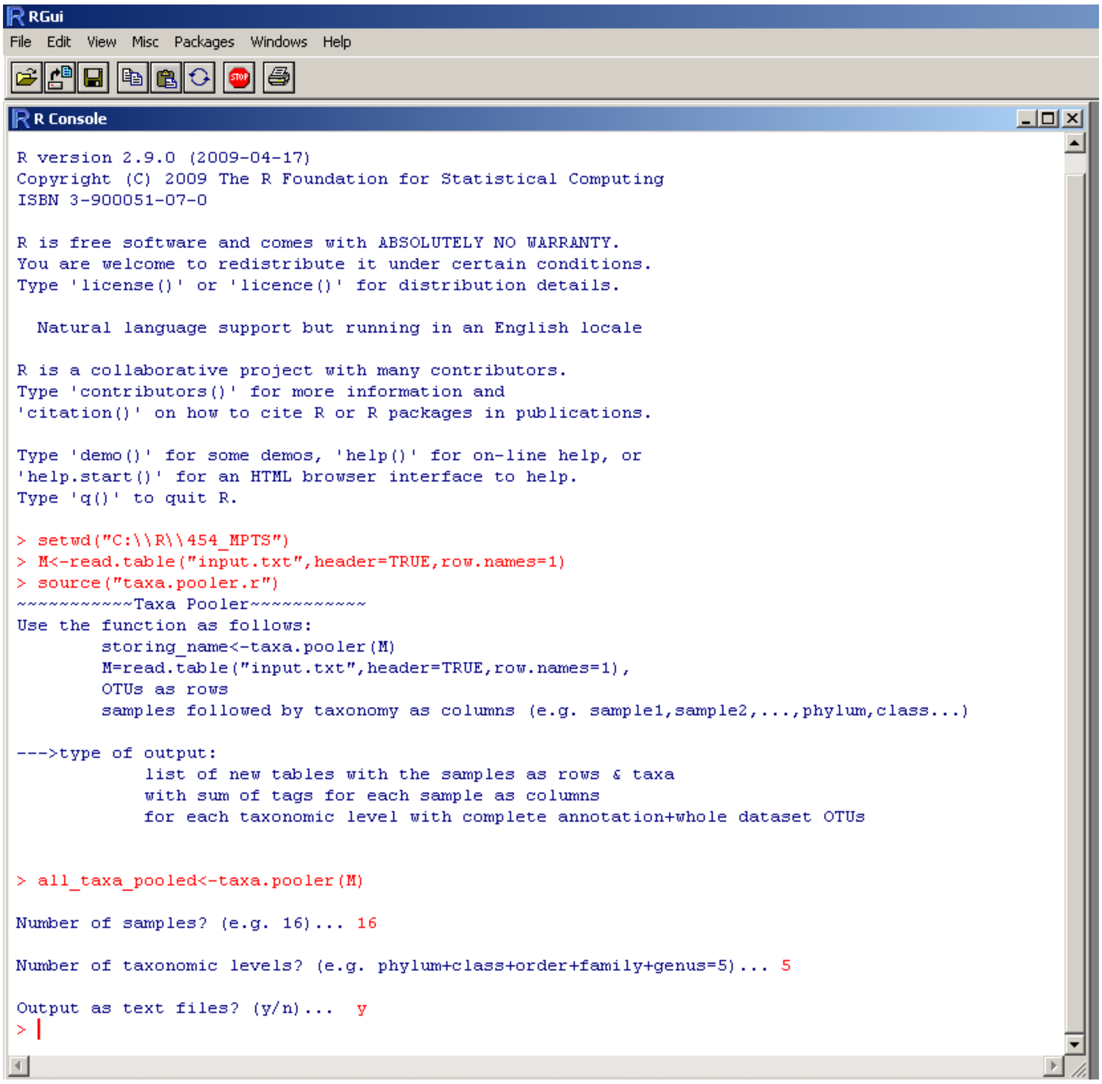
4.1. To obtain a matrix for each taxonomic level (when the taxonomic annotation is available only):

```
source("taxa.pooler.1.2.r")
```

Some explanations about the function are then appearing. To execute the script, the output can be stored in the R workspace under a name of your choice, for instance:

```
all_taxa_pooled<-taxa.pooler(M)
```

Some questions will then appear:



```
R version 2.9.0 (2009-04-17)
Copyright (C) 2009 The R Foundation for Statistical Computing
ISBN 3-900051-07-0

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> setwd("C:\\R\\454_MPTS")
> M<-read.table("input.txt",header=TRUE,row.names=1)
> source("taxa.pooler.r")
~~~~~~~~~~~Taxa Pooler~~~~~~~~~~~
Use the function as follows:
        storing_name<-taxa.pooler(M)
        M=read.table("input.txt",header=TRUE,row.names=1),
        OTUs as rows
        samples followed by taxonomy as columns (e.g. sample1,sample2,...,phylum,class...)

--->type of output:
            list of new tables with the samples as rows & taxa
            with sum of tags for each sample as columns
            for each taxonomic level with complete annotation+whole dataset OTUs


> all_taxa_pooled<-taxa.pooler(M)

Number of samples? (e.g. 16)... 16

Number of taxonomic levels? (e.g. phylum+class+order+family+genus=5)... 5

Output as text files? (y/n)...  y
> 
```

The output is a list of matrices for each taxonomic level and two other matrices describing the occurrence of each OTU: one for only OTUs with a complete annotation and another one with all the OTUs.

4.2. Application of successive cutoffs on each original matrices

```
source("COtables.1.2.r")
```

The truncated datasets can be stored as follows:

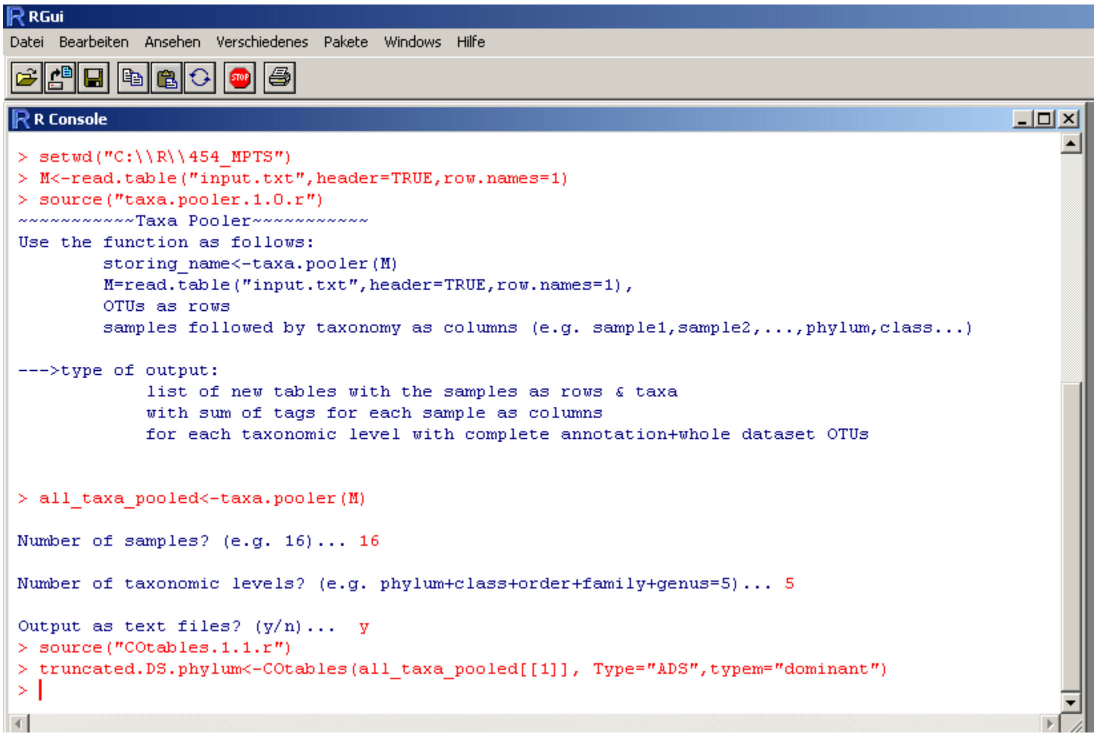
```
truncated.DS.i<-COtables(all_taxa_pooled[[i]], Type="ADS",typem="dominant")
```

With:

- The input: “all_taxa_pooled[[i]]”, representing one of the matrix obtained from the taxa.pooler(), with i from 1 (phylum level here) to the total number of taxonomic levels (here, 7), for example:

```
truncated.DS.phylum<-COtables(all_taxa_pooled[[1]], Type="ADS",typem="dominant")
truncated.DS.class<-COtables(all_taxa_pooled[[2]], Type="ADS",typem="dominant")
                                        .
                                        .
                                        .
truncated.DS.OTUwholeDS<-COtables(all_taxa_pooled[[7]], Type="ADS",typem="dominant")
```

- Type = Type of cutoff: all dataset-,”ADS”, or sample-,”SAM”, based;
- typem = choice of the fraction of the matrix to work on: “dominant” types or “rare” types.



```
> setwd("C:\\R\\454_MPTS")
> M<-read.table("input.txt",header=TRUE,row.names=1)
> source("taxa.pooler.1.0.r")
~~~~~~~~~~~Taxa Pooler~~~~~~~~~~~
Use the function as follows:
        storing_name<-taxa.pooler(M)
        M=read.table("input.txt",header=TRUE,row.names=1),
        OTUs as rows
        samples followed by taxonomy as columns (e.g. sample1,sample2,...,phylum,class...)

--->type of output:
            list of new tables with the samples as rows & taxa
            with sum of tags for each sample as columns
            for each taxonomic level with complete annotation+whole dataset OTUs


> all_taxa_pooled<-taxa.pooler(M)

Number of samples? (e.g. 16)... 16

Number of taxonomic levels? (e.g. phylum+class+order+family+genus=5)... 5

Output as text files? (y/n)...  y
> source("COtables.1.1.r")
> truncated.DS.phylum<-COtables(all_taxa_pooled[[1]], Type="ADS",typem="dominant")
> |
```

4.3. Calculation of Spearman (or Pearson, Kendall) correlations and Procrustes correlations between the original dataset and the truncated ones

In this script, the truncated datasets are automatically calculated.

```
source("cutoff.impact.1.2.r")
```

Some explanations about the function are then appearing. Store the output in the R workspace under a name of your choice, for instance:

```
corr.all<-cutoff.impact(all_taxa_pooled,Type="ADS",corcoef="spearman",typem="dominant")
```

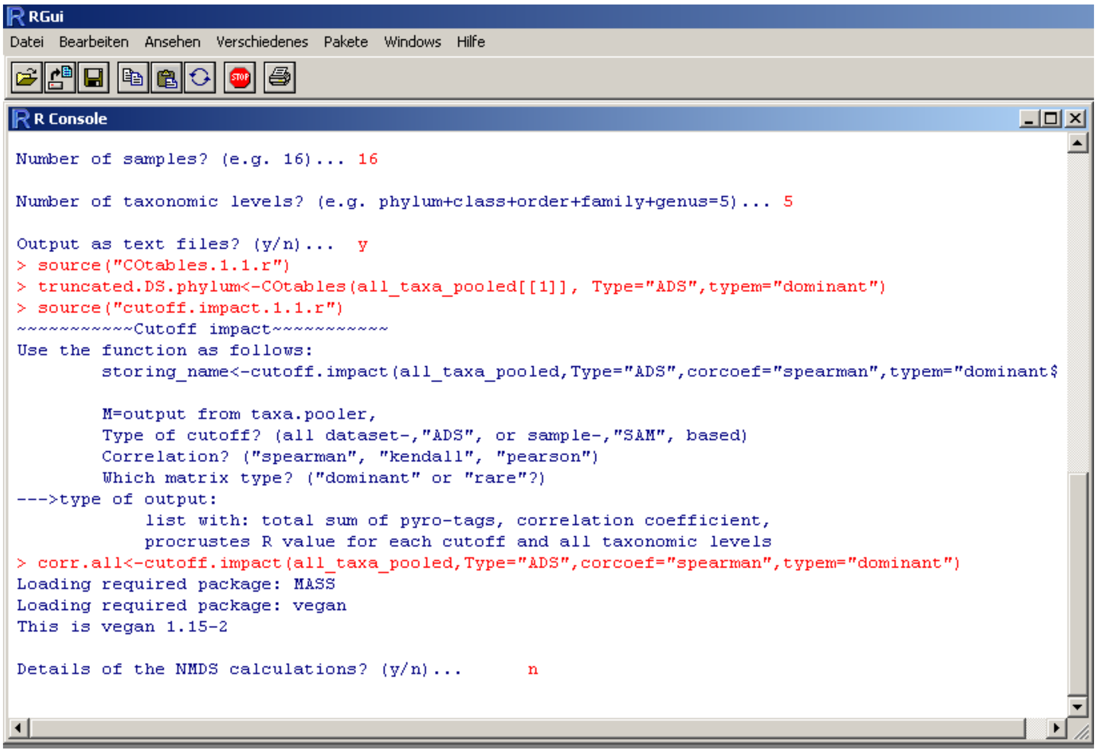
With:

- The input, “all_taxa_pooled” here, should be a list (e.g. the output from the taxa.pooler);

- Type = Type of cutoff: all dataset-,"ADS", or sample-,"SAM", based;
- corcoef = the chosen non-parametric correlation coefficient: "spearman" ("pearson" for a linear coefficient).
- typem = choice of the fraction of the matrix to work on: "dominant" types or "rare" types.

Also, if one does not need to see the details of the NMDS calculations, some computing time might be saved by answering no ("n") to the following question:

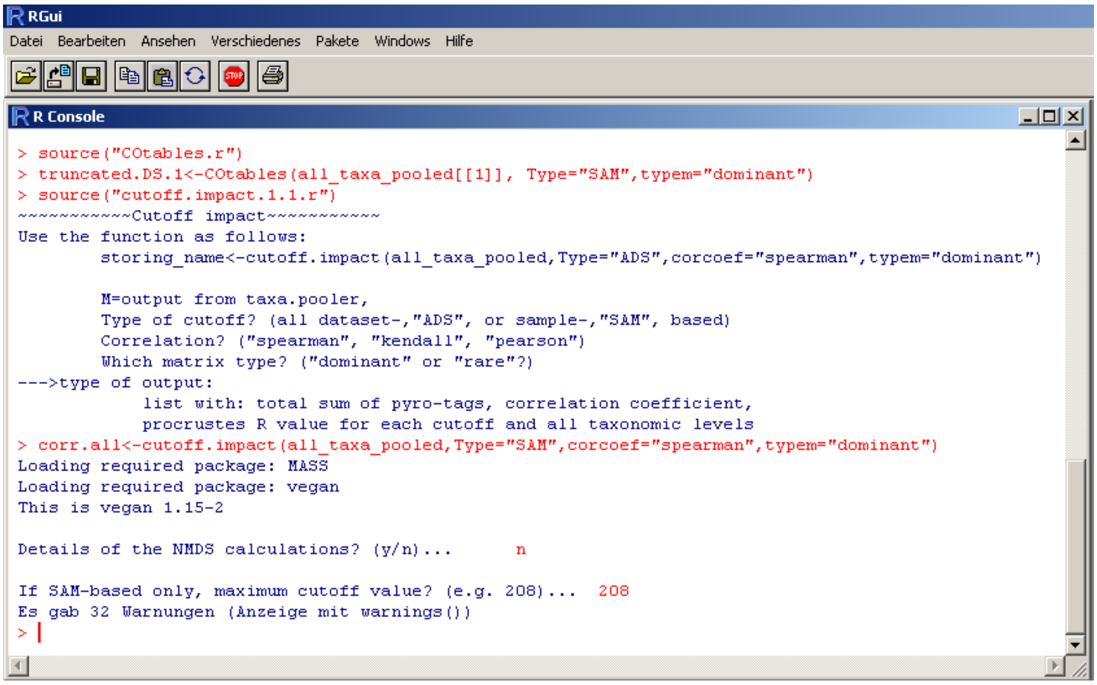
```
Details of the NMDS calculations? (y/n)...
```



```
Number of samples? (e.g. 16)... 16

Number of taxonomic levels? (e.g. phylum+class+order+family+genus=5)... 5

Output as text files? (y/n)...  y
> source("COtables.1.1.r")
> truncated.DS.phylum<-COtables(all_taxa_pooled[[1]], Type="ADS",typem="dominant")
> source("cutoff.impact.1.1.r")
~~~~~~~~~~~Cutoff impact~~~~~~~~~~~
Use the function as follows:
        storing_name<-cutoff.impact(all_taxa_pooled,Type="ADS",corcoef="spearman",typem="dominant$

        M=output from taxa.pooler,
        Type of cutoff? (all dataset-,"ADS", or sample-,"SAM", based)
        Correlation? ("spearman", "kendall", "pearson")
        Which matrix type? ("dominant" or "rare"?)
--->type of output:
            list with: total sum of pyro-tags, correlation coefficient,
            procrustes R value for each cutoff and all taxonomic levels
> corr.all<-cutoff.impact(all_taxa_pooled,Type="ADS",corcoef="spearman",typem="dominant")
Loading required package: MASS
Loading required package: vegan
This is vegan 1.15-2

Details of the NMDS calculations? (y/n)...      n
```

If sample-based cutoff chosen, the following question will appear:

```
If SAM-based only, maximum cutoff value? (e.g. 208)...
```



```
> source("COtables.r")
> truncated.DS.1<-COtables(all_taxa_pooled[[1]], Type="SAM",typem="dominant")
> source("cutoff.impact.1.1.r")
~~~~~~~~~~~Cutoff impact~~~~~~~~~~~
Use the function as follows:
        storing_name<-cutoff.impact(all_taxa_pooled,Type="ADS",corcoef="spearman",typem="dominant")

        M=output from taxa.pooler,
        Type of cutoff? (all dataset-,"ADS", or sample-,"SAM", based)
        Correlation? ("spearman", "kendall", "pearson")
        Which matrix type? ("dominant" or "rare"?)
--->type of output:
            list with: total sum of pyro-tags, correlation coefficient,
            procrustes R value for each cutoff and all taxonomic levels
> corr.all<-cutoff.impact(all_taxa_pooled,Type="SAM",corcoef="spearman",typem="dominant")
Loading required package: MASS
Loading required package: vegan
This is vegan 1.15-2

Details of the NMDS calculations? (y/n)...      n

If SAM-based only, maximum cutoff value? (e.g. 208)...  208
Es gab 32 Warnungen (Anzeige mit warnings())
> |
```

The output is a list of tables with the different assigned cutoffs (all dataset- or sample-based) by the sum of each truncated table, the correlation value between the original table and the truncated table, and the Procrustes value between the non-metric multidimensional scaling (NMDS) from the original table and the truncated table for all taxonomic levels.

!!! This script requires some time and a certain computing power (10 min of calculations for the example matrix with 1,000 OTUs on an Intel Pentium 4)

In order to obtain similar figures as Fig. 3 in the article, another script is needed:

```
source("cutoff.impact.fig.1.2.r")
output.all<-cutoff.impact.fig(corr.all)
```

With the input, “corr.all” here, as a list (e.g. the output from the cutoff.impact) and you can choose to have the output as a text file:

```
Output as text files? (y/n)...
```

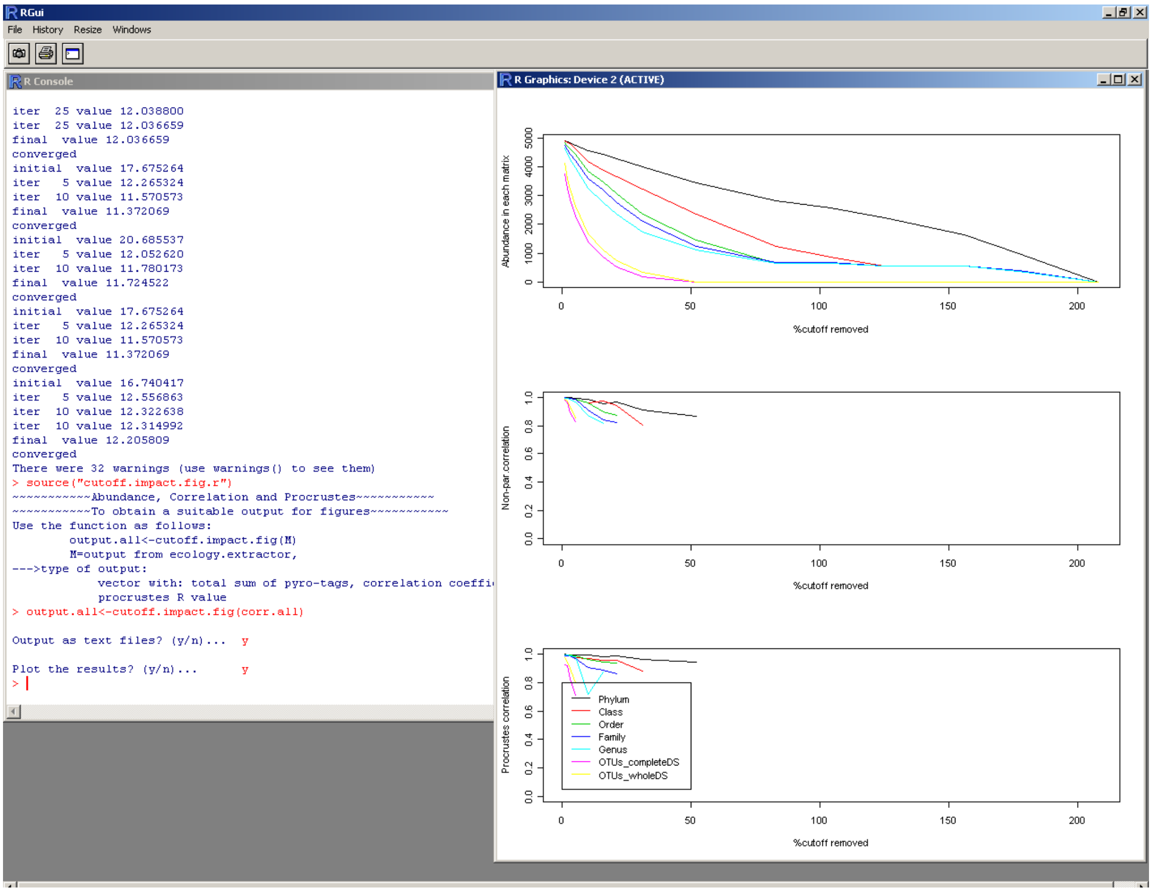
Then three files will appear in the directory:

- "abundance.txt"
- "non-par.correlation.txt"
- "procrustes.txt"

And they can be further used to produce figures with Microsoft Excel for example.

Or you can also choose if you want to directly plot the data:

```
Plot the results? (y/n)...
```



```
iter  25 value 12.038800
iter  25 value 12.036659
final  value 12.036659
converged
initial  value 17.675264
iter   5 value 12.265324
iter  10 value 11.570573
final  value 11.372069
converged
initial  value 20.685537
iter   5 value 12.052620
iter  10 value 11.780173
final  value 11.724522
converged
initial  value 17.675264
iter   5 value 12.265324
iter  10 value 11.570573
final  value 11.372069
converged
initial  value 16.740417
iter   5 value 12.556863
iter  10 value 12.322638
iter  10 value 12.314992
final  value 12.205809
converged
There were 32 warnings (use warnings() to see them)
> source("cutoff.impact.fig.r")
~~~~~~~~~~~Abundance, Correlation and Procrustes~~~~~~~~~~~
~~~~~~~~~~~To obtain a suitable output for figures~~~~~~~~~~~
Use the function as follows:
        output.all<-cutoff.impact.fig(M)
        M=output from ecology.extractor,
--->type of output:
            vector with: total sum of pyro-tags, correlation coeffi
            procrustes R value
> output.all<-cutoff.impact.fig(corr.all)

Output as text files? (y/n)...  y

Plot the results? (y/n)...      y
> 
```

5. Load and run the scripts: ecological patterns

5.1. Variation partitioning at several cutoff levels for all taxonomic levels

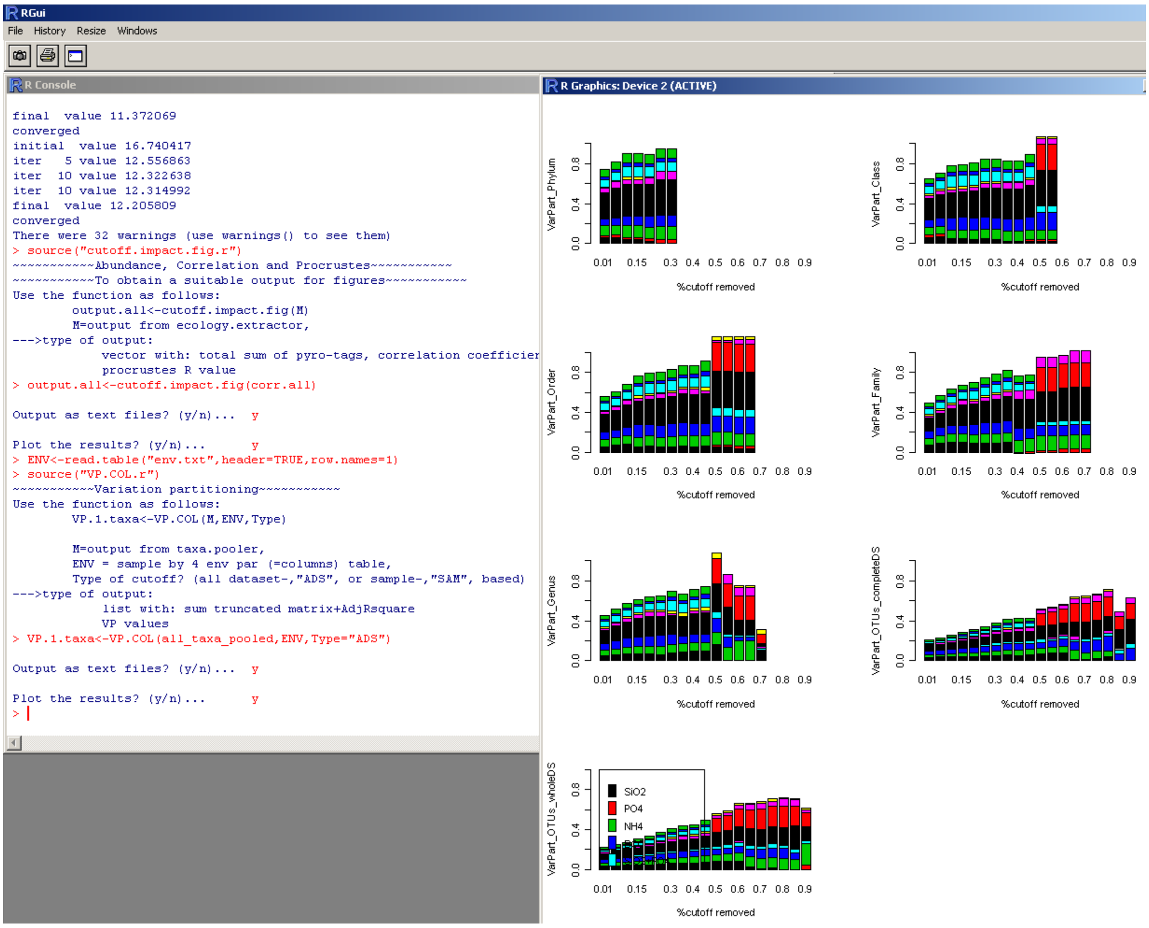
Load the environmental table with samples as rows and environmental parameters as columns (here the script is written for an environmental table with 4 columns) and the script:

```
ENV<-read.table("env.txt",header=TRUE,row.names=1)
source("VP.COL.1.2.r")
```

Some explanations about the function are then appearing. Store the output in the R workspace under a name of your choice, for instance:

```
VP.1.taxa<-VP.COL(all_taxa_pooled,ENV,Type="ADS")
```

With:
- The input, “all_taxa_pooled” here, is the output from the taxa.pooler;
- ENV = the environmental table;
- Type = Type of cutoff: all dataset-,”ADS”, or sample-,”SAM”, based.



```
final  value 11.372069
converged
initial  value 16.740417
iter   5 value 12.556863
iter  10 value 12.322638
iter  10 value 12.314992
final  value 12.205809
converged
There were 32 warnings (use warnings() to see them)
> source("cutoff.impact.fig.r")
~~~~~~~~~~~Abundance, Correlation and Procrustes~~~~~~~~~~~
~~~~~~~~~~~To obtain a suitable output for figures~~~~~~~~~~~
Use the function as follows:
        output.all<-cutoff.impact.fig(M)
        M=output from ecology.extractor,
--->type of output:
            vector with: total sum of pyro-tags, correlation coefficier
            procrustes R value
> output.all<-cutoff.impact.fig(corr.all)

Output as text files? (y/n)...  y

Plot the results? (y/n)...      y
> ENV<-read.table("env.txt",header=TRUE,row.names=1)
> source("VP.COL.r")
~~~~~~~~~~~Variation partitioning~~~~~~~~~~~
Use the function as follows:
        VP.1.taxa<-VP.COL(M,ENV,Type)

        M=output from taxa.pooler,
        ENV = sample by 4 env par (=columns) table,
        Type of cutoff? (all dataset-,"ADS", or sample-,"SAM", based)
--->type of output:
            list with: sum truncated matrix+AdjRsquare
            VP values
> VP.1.taxa<-VP.COL(all_taxa_pooled,ENV,Type="ADS")

Output as text files? (y/n)...  y

Plot the results? (y/n)...      y
> 
```

The output is a list of two tables, for each taxonomic level:
- one with the partition of the variation by the different assigned cutoffs (all dataset- or sample-based);
- one with the different assigned cutoffs by the sum of each truncated table, and the adjusted R square.

You can choose if you want the output as a text file:

```
Output as text files? (y/n)...
```

Then two files x the number of taxonomic level will appear in the directory:
- "taxonomiclevel.VarPart.txt"
- "taxonomiclevel.sum.adjRsq.txt"

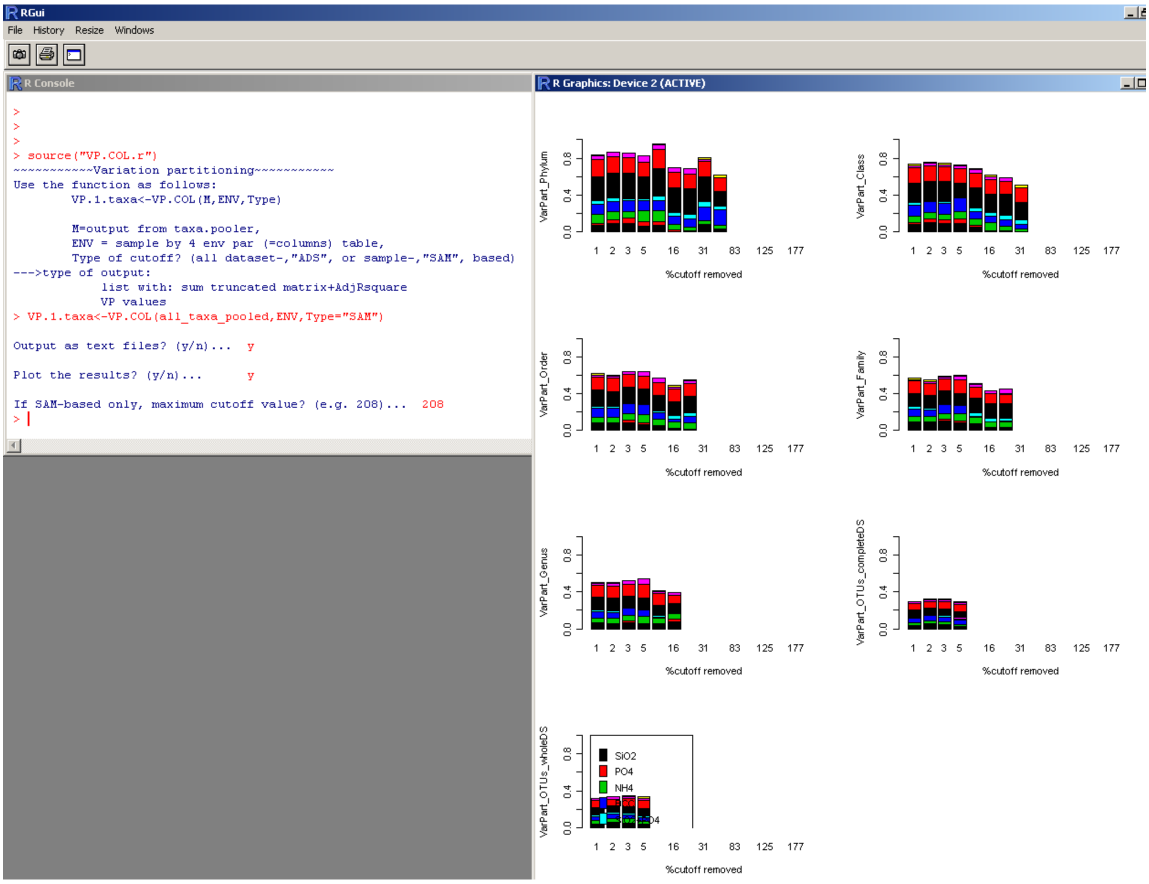
And then can be further used to produce figures with Microsoft Excel for example.

Or you can also choose if you want to plot the data:

```
Plot the results? (y/n)...
```

If sample-based cutoff chosen, the following question will appear:

```
If SAM-based only, maximum cutoff value? (e.g. 208)...
```



```
>
>
>
> source("VP.COL.r")
~~~~~~~~~~~Variation partitioning~~~~~~~~~~~
Use the function as follows:
        VP.1.taxa<-VP.COL(M,ENV,Type)

        M=output from taxa.pooler,
        ENV = sample by 4 env par (=columns) table,
        Type of cutoff? (all dataset-,"ADS", or sample-,"SAM", based)
--->type of output:
            list with: sum truncated matrix+AdjRsquare
            VP values
> VP.1.taxa<-VP.COL(all_taxa_pooled,ENV,Type="SAM")

Output as text files? (y/n)...   y

Plot the results? (y/n)...       y

If SAM-based only, maximum cutoff value? (e.g. 208)...  208
> |
```

5.2. Calculation of correlation coefficients for the environmental parameters (for the first RDA axis)

Load the following script:

```
source("corrcoeff.ENV.1.2.r")
```

However, a whole “automatic” script could not be realized as the R software can present some scoping problems. Instead, you may copy and paste the following lines (here an example for the original table at the OTU level with the whole dataset; we work here on the 7th element of the VP.1.taxa output):

- for all dataset-based cutoffs:

```
#create a matrix to store corrcoeff output at all 21 cutoffs
corrcoeff.table.ADS<-matrix(NA,21,5)
row.names(corrcoeff.table.ADS)<-c(paste("CO_",c(0.01,seq(0.05,0.95,by=0.05),0.99),sep=""))
colnames(corrcoeff.table.ADS)<-c("Sum",paste("RDA1.",colnames(ENV),sep=""))

#store the original matrix
#7: whole dataset at the OTU level
#3: where the cutoff matrices are
OTU.ADS<-VP.1.taxa[[c(7,3)]]

#application of corrcoeff at all cutoffs
SPE<-OTU.ADS[[1]];corrcoeff.table.ADS[1,]<-corrcoeff(SPE,ENV);rm(SPE)
SPE<-OTU.ADS[[2]];corrcoeff.table.ADS[2,]<-corrcoeff(SPE,ENV);rm(SPE)
                                                  .
                                                  .
                                                  .
SPE<-OTU.ADS[[21]];corrcoeff.table.ADS[21,]<-corrcoeff(SPE,ENV);rm(SPE)

#application of corrcoeff on the original table with no cutoff
SPE<-all_taxa_pooled[[7]]
corrcoeff.table.ADS.orig<-corrcoeff(SPE,ENV)
row.names(corrcoeff.table.ADS.orig)<-c("CO_1")
corrcoeff.table.ADS<-rbind(corrcoeff.table.ADS,corrcoeff.table.ADS.orig)

#output as a text file
write.table(corrcoeff.table.ADS,"corrcoeff.table.ADS.txt",quote=FALSE)
```

- for sample-based cutoffs:

```
#create a matrix to store corrcoeff output at all 15 cutoffs
corrcoeff.table.SAM<-matrix(NA,15,5)
row.names(corrcoeff.table.SAM)<-
c(paste("CO_",c(1,2,3,5,10,15,20,30,55,80,105,130,155,180,208),sep=""))
colnames(corrcoeff.table.SAM)<-c("Sum",paste("RDA1.",colnames(ENV),sep=""))

#store the original matrix
#7: whole dataset at the OTU level
#3: where the cutoff matrices are
OTU.SAM<-VP.1.taxa[[c(7,3)]]

#application of corrcoeff at all cutoffs
SPE<-OTU.SAM[[1]];corrcoeff.table.SAM[1,]<-corrcoeff(SPE,ENV);rm(SPE)
SPE<-OTU.SAM[[2]];corrcoeff.table.SAM[2,]<-corrcoeff(SPE,ENV);rm(SPE)
                                                  .
                                                  .
                                                  .
SPE<-OTU.SAM[[15]];corrcoeff.table.SAM[15,]<-corrcoeff(SPE,ENV);rm(SPE)
```

```
#output as a text file
write.table(corrcoeff.table.SAM,"corrcoeff.table.SAM.txt",quote=FALSE)
```

5.3. Calculation of the significance of the whole variation partitioning model and the impact of the pure environmental parameters

Load the following script:

```
source("signif.1.2.r")
```

However, a whole “automatic” script could not be realized as the R software can present some scoping problems. Instead, you may copy and paste the following lines (here an example for the original table at the OTU level with the whole dataset; we work here on the 7th element of the VP.1.taxa output):

- for all dataset-based cutoffs:

```
#create a matrix to store signif output at all 21 cutoffs
signif.table.ADS<-matrix(NA,21,5)
row.names(signif.table.ADS)<-c(paste("CO_",c(0.01,seq(0.05,0.95,by=0.05),0.99),sep=""))
colnames(signif.table.ADS)<- c("whole.sig","ENV1.sig","ENV2.sig","ENV3.sig","ENV4.sig")

#store the original matrix
#7: whole dataset at the OTU level
#3: where the cutoff matrices are
OTU.ADS<-VP.1.taxa[[c(7,3)]]

#application of signif at all cutoffs
SPE<-OTU.ADS[[1]];signif.table.ADS[1,]<-signif(SPE,ENV);rm(SPE)
SPE<-OTU.ADS[[2]];signif.table.ADS[2,]<-signif(SPE,ENV);rm(SPE)
                                             .
                                             .
                                             .
SPE<-OTU.ADS[[21]];signif.table.ADS[21,]<-signif(SPE,ENV);rm(SPE)

#application of signif on the original table with no cutoff
SPE<-all_taxa_pooled[[7]]
signif.table.ADS.orig<-signif(SPE,ENV)
row.names(signif.table.ADS.orig)<-c("CO_1")
signif.table.ADS<-rbind(signif.table.ADS, signif.table.ADS.orig)

#output as a text file
write.table(signif.table.ADS,"signif.table.ADS.txt",quote=FALSE)
```

- for sample-based cutoffs:

```
#create a matrix to store signif output at all 15 cutoffs
signif.table.SAM<-matrix(NA,15,5)
row.names(signif.table.SAM)<-
c(paste("CO_",c(1,2,3,5,10,15,20,30,55,80,105,130,155,180,208),sep=""))
colnames(signif.table.SAM)<- c("whole.sig","ENV1.sig","ENV2.sig","ENV3.sig","ENV4.sig")

#store the original matrix
#7: whole dataset at the OTU level
#3: where the cutoff matrices are
OTU.SAM<-VP.1.taxa[[c(7,3)]]

#application of signif at all cutoffs
SPE<-OTU.SAM[[1]];signif.table.SAM[1,]<-signif(SPE,ENV);rm(SPE)
SPE<-OTU.SAM[[2]];signif.table.SAM[2,]<-signif(SPE,ENV);rm(SPE)
```

```
                                                   .
                                                   .
                                                   .
SPE<-OTU.SAM[[15]];signif.table.SAM[15,]<-signif(SPE,ENV);rm(SPE)

#output as a text file
write.table(signif.table.SAM,"signif.table.SAM.txt",quote=FALSE)
```

6. Save your R workspace

```
save.image("MultiCoLA.RData")
```

All variables will then be saved and then available to work on them without running all the scripts again.

How to cite the script?
Gobet, A., Quince, C., and Ramette, A. 2010. **Multivariate Cutoff Level Analysis (MultiCoLA) of Large Community Datasets.** *Nucl. Acids Res.*

Comments and corrections are always welcome. Please address email correspondence to:
Angélique Gobet: agobet@mpi-bremen.de
or
Alban Ramette: aramette@mpi-bremen.de

7.2 MultiCoLA R scripts (Chapter II)

```
# Copyright (C) 2010 Angelique Gobet & Alban Ramette
# This program is free software; you can redistribute it and/or modify
# it under the terms of the GNU General Public License as published by
# the Free Software Foundation; either version 2 of the License, or
# (at your option) any later version.
#
# This program is distributed in the hope that it will be useful,
# but WITHOUT ANY WARRANTY; without even the implied warranty of
# MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.  See the
# GNU General Public License for more details.

cat("~~~~~~~~~~~~Taxa Pooler~~~~~~~~~~~~\n")
cat('Use the function as follows:\n')
cat('       storing_name<-taxa.pooler(M)\n')
cat('       M=read.table("input.txt",header=TRUE,row.names=1),\n')
cat('       OTUs as rows\n')
cat('       samples followed by taxonomy as columns (e.g. sample1,sample2,...,phylum,class...)\n\n')
cat('--->type of output:\n')
cat('         list of new tables with the samples as rows & taxa\n')
cat('         with sum of tags for each sample as columns\n')
cat('         for each taxonomic level with complete annotation+whole dataset OTUs\n\n\n')

taxa.pooler<-function(M1){
sa=as.numeric(readline("\nNumber of samples? (e.g. 16)...\t"))
ta=as.numeric(readline("\nNumber of taxonomic levels? (e.g. phylum+class+order+family+genus=5)...\t"))
OUTP=readline("\nOutput as text files? (y/n)...\t")

M<-M1[-which(apply(M1,1,function(x)any(is.na(x)))),]

pool.1.level<-function(M,j){
        N<-matrix(NA,length(unique(M[,sa+j])),sa)
        row.names(N)<-sort(unique.default(M[,sa+j]))  #name of unique taxa
                colnames(N)<-colnames(M[,1:sa])  #name of samples
        for (i in 1:sa){
  N[,i]<-by(M[,i],factor(M[,sa+j]),sum)
 }      #end for i
        return(t(N))
}       #end pool.1.level()


taxa_res<-vector("list",ta+2)
names(taxa_res)<-c(colnames(M[,(sa+1):(sa+ta)]),"OTUs_completeDS","OTUs_wholeDS")

#loop to apply the function pool.1.level at all taxonomic levels
for (k in 1:ta){
        taxa_res[[k]]<-pool.1.level(M,k)
        }       #end for k

#table at the OTU level with only OTUs with a complete annotation
taxa_res[[ta+1]]<-t(M[,1:sa])

#table at the OTU level with all the OTUs
taxa_res[[ta+2]]<-t(M1[,1:sa])

 if(OUTP=="y"){
         for(j in 1:(ta+2)){
```

```
    write.table(taxa_res[[j]],paste(names(taxa_res[j]),".matrix.txt",sep=""),quote=FALSE)
    }
  } #end if

return(taxa_res)

}            #end taxa.pooler
```

```
# Copyright (C) 2010 Angelique Gobet & Alban Ramette
# This program is free software; you can redistribute it and/or modify
# it under the terms of the GNU General Public License as published by
# the Free Software Foundation; either version 2 of the License, or
# (at your option) any later version.
#
# This program is distributed in the hope that it will be useful,
# but WITHOUT ANY WARRANTY; without even the implied warranty of
# MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.  See the
# GNU General Public License for more details.

COtables<-function(ODS,Type="ADS",typem="dominant"){
       COP<-function(ODS,z,Type,typem){##################
    #to remove all the lines in the matrix for which the sum of the line is 0
    CLrow<-function(m) {
     #to create a column of 0 in the last column
     m=cbind(m,matrix(0,nrow(m),1))
     m[,ncol(m)]=apply(m,1,sum)
     #to keep only the lines without 0
     mclean=subset(m,m[,ncol(m)]!=0)
     mclean=mclean[,-ncol(mclean)]
     return (mclean)
     }
     #######################
    CLcol<-function(m) {
     #transpose the matrix and apply the same function as before
     #then transpose back
     m=t(m)
     #to create a column of 0 in the last column
     m=cbind(m,matrix(0,nrow(m),1))
     m[,ncol(m)]=apply(m,1,sum)
     #to keep only the lines without 0
     mclean=subset(m,m[,ncol(m)]!=0)
     mclean=mclean[,-ncol(mclean)]
     mclean=t(mclean)
     return (mclean)
     }

                     #Application of a percentage cut-off to the original dataset to obtain abundant
dataset
                     ##all dataset-based cutoff
                     if(Type=="ADS"){
                            M<-rbind(ODS,apply(ODS,2,sum))#add the column sum as a last row
of the matrix M
                            if(typem=="dominant"){N<-
M[,order(M[nrow(M),],decreasing=TRUE)]}         #order the columns of M by their decreasing sum
              if(typem=="rare"){N<-M[,order(M[nrow(M),])]}  #order the columns of M by their
increasing sum
                            Q<-N[1:(nrow(N)-1),]      #remove the last row (with the sum of the
columns)
                            L<-ncol(Q)
                            K<-nrow(Q)
                            M1<-t(matrix(NA,L))       #create a vector to store sum of successive
matrices
                            Q1<-matrix(NA,K,L)        #create a matrix to store new data
                            perc<-z*sum(ODS)
```

```
                                  for (i in 1:L){     ###for #1
                                         M1[,i]<-sum(Q[,1:i])
                                         if (M1[,i]<=perc) {Q1[,1:i]=Q[,1:i]}
                                         row.names(Q1)=row.names(Q)
                                         colnames(Q1)=colnames(Q)
                                         if (M1[,i]>perc) {Q1[,1:i]==0}
                                 }#end for #1
                                 Q3<-Q1[,-which(apply(Q1,2,function(x)all(is.na(x))))]
                                 Q3<-CLcol(CLrow(as.data.frame(Q3)))       #remove rows and
columns whose sum=0
                         } #end"ADS"

                         ##sample-based cutoff
                         if(Type=="SAM"){
                                 Q1<-ODS
                                 if(typem=="dominant"){Q1[Q1<z]<-0}        # all species presents less
than j times =0
                                 if(typem=="rare"){Q1[Q1>z]<-0}   # all species presents more than j
times =0
                                 Q3<-CLcol(CLrow(Q1))   #remove rows and columns whose sum=0
                         } #end "SAM"
return(Q3)
} #end COP

        if(Type=="ADS"){
                  #create a matrix to store VPvalues for each CO
                  LISTRES<-vector("list",21)
   names(LISTRES)<-c(0.01,seq(0.05,0.95,by=0.05),0.99)
                         for(i in 1:21){
    LISTRES[[i]]=COP(ODS,z=as.numeric(names(LISTRES)[i]),Type,typem)
                  }
        } #end if "ADS"

        if(Type=="SAM"){
   #create a matrix to store VPvalues for each CO
                  LISTRES<-vector("list",15)
   names(LISTRES)<-c(1,2,3,5,10,15,20,30,55,80,105,130,155,180,208)
                         for(i in 1:15){
    LISTRES[[i]]=COP(ODS,z=as.numeric(names(LISTRES)[i]),Type,typem)
                  }
        } #end if "SAM"
        return(LISTRES)
} #end of VP.COL
###############################################################################
```

```
# Copyright (C) 2010 Angelique Gobet & Alban Ramette
# This program is free software; you can redistribute it and/or modify
# it under the terms of the GNU General Public License as published by
# the Free Software Foundation; either version 2 of the License, or
# (at your option) any later version.
#
# This program is distributed in the hope that it will be useful,
# but WITHOUT ANY WARRANTY; without even the implied warranty of
# MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.  See the
# GNU General Public License for more details.

cat("~~~~~~~~~~~~Cutoff impact~~~~~~~~~~~~\n")
cat('Use the function as follows:\n')
cat('       storing_name<-
cutoff.impact(all_taxa_pooled,Type="ADS",corcoef="spearman",typem="dominant")\n\n')
cat('       M=output from taxa.pooler,\n')
cat('       Type of cutoff? (all dataset-,"ADS", or sample-,"SAM", based)\n')
cat('       Correlation? ("spearman", "kendall", "pearson")\n')
cat('       Which matrix type? ("dominant" or "rare"?)\n')
cat('--->type of output:\n')
cat('         list with: total sum of pyro-tags, correlation coefficient,\n')
cat('         procrustes R value for each cutoff and all taxonomic levels\n')

cutoff.impact<-function(MM,Type="ADS",corcoef="spearman",typem="dominant"){
require(MASS)
require(vegan)
details=readline("\nDetails of the NMDS calculations? (y/n)...\t")
#calculation of the cut-off matrices, correlation coefficient, procrustes R value
CoCalc<-function(ODS,z,Type,corcoef){
            res1<-matrix(NA,1,3)       #create a matrix to store mantel and procrustes data
            colnames(res1)<-c("Sum","corrcoeff","R Procrustes")

        #to remove all the lines in the matrix for which the sum of the line is 0
        CLrow<-function(m) {
          #to create a column of 0 in the last column
          m=cbind(m,matrix(0,nrow(m),1))
          m[,ncol(m)]=apply(m,1,sum)
          #to keep only the lines without 0
          mclean=subset(m,m[,ncol(m)]!=0)
          mclean=mclean[,-ncol(mclean)]
          return (mclean)
         }
         #######################
        CLcol<-function(m) {
          #transpose the matrix and apply the same function as before
          #then transpose back
          m=t(m)
          #to create a column of 0 in the last column
          m=cbind(m,matrix(0,nrow(m),1))
          m[,ncol(m)]=apply(m,1,sum)
          #to keep only the lines without 0
          mclean=subset(m,m[,ncol(m)]!=0)
          mclean=mclean[,-ncol(mclean)]
          mclean=t(mclean)
          return (mclean)
         }
```

```
                        #Application of a percentage cut-off to the original dataset to obtain abundant
dataset
                        ##all dataset-based cutoff
                        if(Type=="ADS"){
                                M<-rbind(ODS,apply(ODS,2,sum))#add the column sum as a last row
of the matrix M
                                if(typem=="dominant"){N<-
M[,order(M[nrow(M),],decreasing=TRUE)]}            #order the columns of M by their decreasing sum
                if(typem=="rare"){N<-M[,order(M[nrow(M),])]}  #order the columns of M by their
increasing sum
                                Q<-N[1:(nrow(N)-1),]       #remove the last row (with the sum of the
columns)
                                L<-ncol(Q)
                                K<-nrow(Q)
                                M1<-t(matrix(NA,L))        #create a vector to store sum of successive
matrices
                                Q1<-matrix(NA,K,L)         #create a matrix to store new data
                                perc<-z*sum(ODS)
                                  for (i in 1:L){     ###for #1
                                        M1[,i]<-sum(Q[,1:i])
                                        if (M1[,i]<=perc) {Q1[,1:i]=Q[,1:i]}
                                        row.names(Q1)=row.names(Q)
                                        colnames(Q1)=colnames(Q)
                                        if (M1[,i]>perc) {Q1[,1:i]==0}
                                }#end for #1
                                Q2<-Q1[,-which(apply(Q1,2,function(x)all(is.na(x))))]
                                Q2<-CLcol(CLrow(as.data.frame(Q2)))       #remove rows and
columns whose sum=0
                                res1[1,1]<-sum(Q2)
    } #end "ADS"

                        ##sample-based cutoff
                        if(Type=="SAM"){
                                Q1<-ODS
                                if(typem=="dominant"){Q1[Q1<z]<-0}        # all species presents less
than j times =0
                                if(typem=="rare"){Q1[Q1>z]<-0}   # all species presents more than j
times =0
                                Q2<-CLcol(CLrow(Q1))   #remove rows and columns whose sum=0
                                res1[1,1]<-sum(Q2)
                 } #end "SAM"

        if (res1[1,1]==0){res1[1,2:3]<-cbind(NA,NA)} # to avoid conflicts when comparing original
dataset to NA
        else {     if (length(Q2)<=nrow(ODS)){res1[1,2:3]<-cbind(NA,NA)} ###else #1
                else { if (nrow(Q2)<nrow(ODS)) {res1[1,2:3]<-cbind(NA,NA)}
                        else { ###else #2
#####################################
        #Correlation and Procrustes calculations
                ODSdist<-vegdist(ODS,method="bray")       #distance matrix of the original dataset
                Q2dist<-vegdist(Q2,distance="bray")          #distance matrix of the truncated dataset
                ODSQ2cor<-cor.test(ODSdist,Q2dist,method=corcoef) #correlation between matrices
                ODSdist2<-ODSdist
                ODSdist2[ODSdist2==0]<-10e-20  #replace 0 by 10e-20 for original dataset
                Q2dist2<-Q2dist
```

```
              Q2dist2[Q2dist2==0]<-10e-20        #replace 0 by 10e-20 for truncated dataset
              if(details=="y"){
               ODSNMDS<-isoMDS(ODSdist2) #NMDS for original dataset
               Q2NMDS<-isoMDS(Q2dist2)                  #NMDS for truncated dataset
               }
               else{if(details=="n"){     ###else #3
               ODSNMDS<-isoMDS(ODSdist2,trace=0)  #NMDS for original dataset
               Q2NMDS<-isoMDS(Q2dist2,trace=0)                  #NMDS for truncated dataset
               }}
              ODSQ2procrustes<-protest(ODSNMDS,Q2NMDS) #procrustes
              res1[1,2]<-cbind(ODSQ2cor$estimate)
              res1[1,3]<-cbind(ODSQ2procrustes$t0)
              } #end else #2
       } #end else #1
 }
####################################
return(res1)
}#end function CoCalc

###############################################################
#application of the function COP at different all dataset-based cutoffs
       if(Type=="ADS"){
  ecol.ext.all<-function(MM,Type,corcoef){
               allcorr<-function(ODS,Type,corcoef){
    ADS_perc<-c(0.01,seq(0.05,0.95,by=0.05),0.99)
    table_taxa<-matrix(NA,length(ADS_perc),3)
    row.names(table_taxa)<-ADS_perc
    colnames(table_taxa)<-c("Sum","corrcoeff","R Procrustes")
    for(i in 1:length(ADS_perc)){
                             table_taxa[((length(ADS_perc)+1)-i),]<-
CoCalc(ODS,z=as.numeric(ADS_perc[i]),Type,corcoef)
                }
                #ROW<-row.names(table_taxa)
    #table_taxa<-table_taxa[order(as.numeric(row.names(table_taxa)),decreasing=TRUE),]
    #row.names(table_taxa)<-ROW
                return(table_taxa)
               } #end allcorr

 list.ecol<-vector("list",length(MM))
 names(list.ecol)<-names(MM)

 for(j in 1:length(MM)){
  list.ecol[[j]]<-allcorr(MM[[j]],Type,corcoef)
 }
 return(list.ecol)
 } #end ecol.ext.all
 }#end if "ADS"

#application of the function COP at different sample-based cutoffs
       if(Type=="SAM"){
  ecol.ext.all<-function(MM,Type,corcoef){
   limSAMco=as.numeric(readline("\nIf SAM-based only, maximum cutoff value? (e.g. 208)...\t"))
              allcorr<-function(ODS,Type,corcoef){
    SAM_perc<-limSAMco*c(0.005,0.01,0.015,0.025,0.05,0.075,0.1,0.15,0.25,0.4,0.5,0.6,0.75,0.85,1)
    table_taxa<-matrix(NA,length(SAM_perc),3)
    row.names(table_taxa)<-round(SAM_perc,0)
```

```
        colnames(table_taxa)<-c("Sum","corrcoeff","R Procrustes")
        for(i in 1:length(SAM_perc)){
                                                           table_taxa[i,]<-
CoCalc(ODS,z=as.numeric(SAM_perc[i]),Type,corcoef)
        }
                                return(table_taxa)
                              } #end allcorr

  list.ecol<-vector("list",length(MM))
  names(list.ecol)<-names(MM)

  for(j in 1:length(MM)){
    list.ecol[[j]]<-allcorr(MM[[j]],Type,corcoef)
  }
  return(list.ecol)
  } #end ecol.ext.all
 }#end if "SAM"

result.allcorr<-ecol.ext.all(MM,Type,corcoef)
return(result.allcorr)

} #end cutoff.impact
```

```
# Copyright (C) 2010 Angelique Gobet & Alban Ramette
# This program is free software; you can redistribute it and/or modify
# it under the terms of the GNU General Public License as published by
# the Free Software Foundation; either version 2 of the License, or
# (at your option) any later version.
#
# This program is distributed in the hope that it will be useful,
# but WITHOUT ANY WARRANTY; without even the implied warranty of
# MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.  See the
# GNU General Public License for more details.

cat("~~~~~~~~~~~Abundance, Correlation and Procrustes~~~~~~~~~~~\n")
cat("~~~~~~~~~~~To obtain a suitable output for figures~~~~~~~~~~~\n")
cat('Use the function as follows:\n')
cat('       output.all<-cutoff.impact.fig(M)\n')
cat('       M=output from ecology.extractor,\n')
cat('--->type of output:\n')
cat('           vector with: total sum of pyro-tags, correlation coefficient,\n')
cat('           procrustes R value\n')

cutoff.impact.fig<-function(M){
  OUTP=readline("\nOutput as text files? (y/n)...\t")
  PLOT=readline("\nPlot the results? (y/n)...\t")
  list.all<-vector("list",3)
  names(list.all)<-c("Abundance","Non-par.correlation","Procrustes")
  list.all[[1]]<-matrix(NA,nrow(M[[1]]),length(M))
  list.all[[2]]<-matrix(NA,nrow(M[[1]]),length(M))
  list.all[[3]]<-matrix(NA,nrow(M[[1]]),length(M))
  colnames(list.all[[1]])<-names(M)
  colnames(list.all[[2]])<-names(M)
  colnames(list.all[[3]])<-names(M)
  row.names(list.all[[1]])<-row.names(M[[1]])
  row.names(list.all[[2]])<-row.names(M[[1]])
  row.names(list.all[[3]])<-row.names(M[[1]])

  for(i in 1:length(M)){
    list.all[[1]][,i]<-cbind(M[[i]][,1])
    list.all[[2]][,i]<-cbind(M[[i]][,2])
    list.all[[3]][,i]<-cbind(M[[i]][,3])
  } #end for

  if(OUTP=="y"){
  write.table(list.all[[1]],"abundance.txt",quote=FALSE)
  write.table(list.all[[2]],"non-par.correlation.txt",quote=FALSE)
  write.table(list.all[[3]],"procrustes.txt",quote=FALSE)
  } #end if

  if(PLOT=="y"){
    par(mfrow=c(3,1))
    plot(row.names(list.all[[1]]),list.all[[1]][,1],type="l",xlab=c("%cutoff removed"),ylab=c("Abundance in
each matrix"))
    for(i in 1:length(M)){
      lines(row.names(list.all[[1]]),list.all[[1]][,i],col=i)
    }
    plot(row.names(list.all[[2]]),list.all[[2]][,1],type="l",ylim=c(0,1),xlab=c("%cutoff
removed"),ylab=c("Non-par.correlation"))
```

```
    for(i in 1:length(M)){
      lines(row.names(list.all[[2]]),list.all[[2]][,i],col=i)
    }
    plot(row.names(list.all[[3]]),list.all[[3]][,1],type="l",ylim=c(0,1),xlab=c("%cutoff
removed"),ylab=c("Procrustes correlation"))
    for(i in 1:length(M)){
      lines(row.names(list.all[[3]]),list.all[[3]][,i],col=i)
      legend(0,0.8,colnames(as.data.frame(list.all[[1]])),col=seq(1:length(M)),lty=1,y.intersp=0.7)
    }
  } #end if

return(list.all)
} #end cutoff.impact.fig
```

```
# Copyright (C) 2010 Angelique Gobet & Alban Ramette
# This program is free software; you can redistribute it and/or modify
# it under the terms of the GNU General Public License as published by
# the Free Software Foundation; either version 2 of the License, or
# (at your option) any later version.
#
# This program is distributed in the hope that it will be useful,
# but WITHOUT ANY WARRANTY; without even the implied warranty of
# MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.  See the
# GNU General Public License for more details.


cat("~~~~~~~~~~~Variation partitioning~~~~~~~~~~~\n")
cat('Use the function as follows:\n')
cat('       VP.1.taxa<-VP.COL(M,ENV,Type)\n\n')
cat('       M=output from taxa.pooler,\n')
cat('       ENV = sample by 4 env par (=columns) table,\n')
cat('       Type of cutoff? (all dataset-,"ADS", or sample-,"SAM", based)\n')
cat('--->type of output:\n')
cat('          list with: sum truncated matrix+AdjRsquare\n')
cat('          VP values\n')


VP.COL<-function(MM,ENV,Type){
require(vegan)
  OUTP=readline("\nOutput as text files? (y/n)...\t")
  PLOT=readline("\nPlot the results? (y/n)...\t")

        COP<-function(ODS,z,ENV,Type){##################
                        #vector to store sum,adjRsq
                        res1<-matrix(NA,1,2)
                        #Application of a percentage cut-off to the original dataset to obtain abundant
dataset
                        ##all dataset-based cutoff
                        if(Type=="ADS"){
                                M<-rbind(ODS,apply(ODS,2,sum))#add the column sum as a last row
of the matrix M
                                N<-M[,order(M[nrow(M),],decreasing=TRUE)]       #order the
columns of M by their decreasing sum
                                Q<-N[1:(nrow(N)-1),]    #remove the last row (with the sum of the
columns)
                                L<-ncol(Q)
                                K<-nrow(Q)
                                M1<-t(matrix(NA,L))     #create a vector to store sum of successive
matrices
                                Q1<-matrix(NA,K,L)      #create a matrix to store new data
                                perc<-z*sum(ODS)
                                for (i in 1:L){     ###for #1
                                        M1[,i]<-sum(Q[,1:i])
                                        if (M1[,i]<=perc) {Q1[,1:i]=Q[,1:i]}
                                        row.names(Q1)=row.names(Q)
                                        colnames(Q1)=colnames(Q)
                                        if (M1[,i]>perc) {Q1[,1:i]==0}
                                }#end for #1
                                Q3<-Q1[,-which(apply(Q1,2,function(x)all(is.na(x))))]
                                res1[,1]<-sum(Q3)
```

```
                        } #end"ADS"

                        ##sample-based cutoff
                        if(Type=="SAM"){
                             #to remove all the lines in the matrix for which the sum of the line is 0
    CLrow= function(m) {
     #to create a column of 0 in the last column
     m=cbind(m,matrix(0,nrow(m),1))
     m[,ncol(m)]=apply(m,1,sum)
     #to keep only the lines without 0
     mclean=subset(m,m[,ncol(m)]!=0)
     mclean=mclean[,-ncol(mclean)]
     return (mclean)
    } #end CLrow
    ######################
    CLcol= function(m) {
     #transpose the matrix and apply the same function as before
     #then transpose back
     m=t(m)
     #to create a column of 0 in the last column
     m=cbind(m,matrix(0,nrow(m),1))
     m[,ncol(m)]=apply(m,1,sum)
     #to keep only the lines without 0
     mclean=subset(m,m[,ncol(m)]!=0)
     mclean=mclean[,-ncol(mclean)]
     mclean=t(mclean)
     return (mclean)
    } #end CLcol
                                 Q1<-ODS
                                 Q1[Q1<z]<-0       # all species presents less than j times =0
                                 Q3<-CLcol(CLrow(Q1))   #remove rows and columns whose sum=0
                                 res1[1,1]<-sum(Q3)
                        } #end "SAM"

##########################
                        if (res1[,1]==0){
                                 res1[,2]<-NA
                                 VP_Rsq<-matrix(NA,nrow(ODS),1)
                                 List=list(res1,VP_Rsq,Q3)
                                 names(List)=c("res1","VP_Rsq","cutoff.table")
                        } #end if
   # to avoid conflicts comparing original dataset/NA
                        else {   ###else #1
    if (length(Q3)<=nrow(ODS)){
                                         res1[,2]<-NA
                                         VP_Rsq<-matrix(NA,nrow(ODS),1)
                                         List=list(res1,VP_Rsq,Q3)
                                         names(List)=c("res1","VP_Rsq","cutoff.table")
                         } #end if
                        else {    ###else #2
    if (nrow(Q3)<nrow(ENV)) {
                                         res1[,2]<-NA
                                         VP_Rsq<-matrix(NA,nrow(ODS),1)
                                         List=list(res1,VP_Rsq,Q3)
                                         names(List)=c("res1","VP_Rsq","cutoff.table")
                         }
```

```
                                else { ###else #3
                                        Q2<-decostand(Q3,"hel")[1:nrow(Q3),1:ncol(Q3)]
                                        ####################################
                                        #Variation partitioning
                                        ###Transform input as model matrices
                                        ENV1<-model.matrix(~.,as.data.frame(ENV[,1]))[,-1]
                                        ENV2<-model.matrix(~.,as.data.frame(ENV[,2]))[,-1]
                                        ENV3<-model.matrix(~.,as.data.frame(ENV[,3]))[,-1]
                                        ENV4<-model.matrix(~.,as.data.frame(ENV[,4]))[,-1]
                                        Q2mod<-model.matrix(~.,as.data.frame(Q2))[,-1]

                                        ###Variation partitioning with 4 variables
                                        VP_Q2mod<-varpart(Q2mod,ENV1,ENV2,ENV3,ENV4)
                                        VP_Rsq<-
as.data.frame(VP_Q2mod$part$indfract$Adj.R.square)
                                        res1[,2]<-VP_Q2mod$part$fract[c("[abcdefghijklmno] =
All"),c("Adj.R.square")]
                                        List<-list(res1,VP_Rsq,Q3)
                                        names(List)<-c("res1","VP_Rsq","cutoff.table")
                                } #end else #3
                        } #end else #2
   } #end else #1
return(List)
} #end COP


        if(Type=="ADS"){
        VP.taxa<-function(MM,ENV,Type){
                VPcutoff<-function(ODS,ENV,Type){
                 #create a matrix to store VPvalues for each CO
                 result1<-matrix(NA,nrow(ODS),21)
                 a<-colnames(ENV)[1]
                 b<-colnames(ENV)[2]
                 d<-colnames(ENV)[3]
                 e<-colnames(ENV)[4]
                 row.names(result1)<-
c(a,b,d,e,paste(a,b,sep="+"),paste(a,d,sep="+"),paste(a,e,sep="+"),paste(b,d,sep="+"),paste(b,e,sep="+"),pa
ste(d,e,sep="+"),paste(a,b,d,sep="+"),paste(a,b,e,sep="+"),paste(a,d,e,sep="+"),paste(b,d,e,sep="+"),"All","
Unexplained")
                 colnames(result1)<-c(0.01,seq(0.05,0.95,by=0.05),0.99)
                 #create a matrix to store sum,adjRsq,envAIC of all cutoffs
                 result2<-matrix(NA,21,2)
                 colnames(result2)<-c("Sum","Adj.R.square")
                 row.names(result2)<-c(0.01,seq(0.05,0.95,by=0.05),0.99)
                 LISTRES<-vector("list",21)
                 names(LISTRES)=c(0.01,seq(0.05,0.95,by=0.05),0.99)
                result3<-vector("list",21)
                        names(result3)=c(0.01,seq(0.05,0.95,by=0.05),0.99)
                        for(i in 1:21){
    LISTRES[[i]]=COP(ODS,z=as.numeric(names(LISTRES)[i]),ENV,Type)
                         result1[,(22-i)]<-LISTRES[[c(i,2)]][1:nrow(ODS),]
                         result1[,i][result1[,i]<0]<-0
                         result2[(22-i),]<-LISTRES[[c(i,1)]]
                         result3[[i]]<-LISTRES[[c(i,3)]]
                 }
         LIST2<-list(result1,result2,result3)
        names(LIST2)<-c("VP_Rsq","res1","cutoff.tables")
              return(LIST2)
              } #end VPcutoff

 list.ecol<-vector("list",length(MM))
 names(list.ecol)<-names(MM)
 for(j in 1:length(MM)){
  list.ecol[[j]]<-VPcutoff(MM[[j]],ENV,Type)
 } #end for
 return(list.ecol)
       } #end VP.taxa
} #end if "ADS"

       if(Type=="SAM"){
       VP.taxa<-function(MM,ENV,Type){
  limSAMco=as.numeric(readline("\nIf SAM-based only, maximum cutoff value? (e.g. 208)...\t"))
              VPcutoff<-function(ODS,ENV,Type){
   SAM_perc<-limSAMco*c(0.005,0.01,0.015,0.025,0.05,0.075,0.1,0.15,0.25,0.4,0.5,0.6,0.75,0.85,1)
               #create a matrix to store VPvalues for each CO
               result1<-matrix(NA,nrow(ODS),length(SAM_perc))
               a<-colnames(ENV)[1]
               b<-colnames(ENV)[2]
               d<-colnames(ENV)[3]
               e<-colnames(ENV)[4]
               row.names(result1)<-c(a,b,d,e,paste(a,b,sep="+"),paste(a,d,sep="+"),paste(a,e,sep="+"),paste(b,d,sep="+"),paste(b,e,sep="+"),paste(d,e,sep="+"),paste(a,b,d,sep="+"),paste(a,b,e,sep="+"),paste(a,d,e,sep="+"),paste(b,d,e,sep="+"),"All","Unexplained")
   colnames(result1)<-round(SAM_perc,0)
               #create a matrix to store sum,adjRsq,envAIC of all cutoffs
               result2<-matrix(NA,length(SAM_perc),2)
               colnames(result2)<-c("Sum","Adj.R.square")
   row.names(result2)<-round(SAM_perc,0)
               LISTRES<-vector("list",length(SAM_perc))
               names(LISTRES)<-SAM_perc
              result3<-vector("list",15)
                     names(result3)<-round(SAM_perc,0)
                     for(i in 1:length(SAM_perc)){
    LISTRES[[i]]=COP(ODS,z=as.numeric(names(LISTRES)[i]),ENV,Type)
                      result1[,i]<-LISTRES[[c(i,2)]][1:nrow(ODS),]
                      result1[,i][result1[,i]<0]<-0
                      result2[i,]<-LISTRES[[c(i,1)]]
                     result3[[i]]<-LISTRES[[c(i,3)]]
               } #end for

        LIST2<-list(result1,result2,result3)
        names(LIST2)<-c("VP_Rsq","res1","cutoff.tables")
              return(LIST2)
              } #end VPcutoff

 list.ecol<-vector("list",length(MM))
 names(list.ecol)<-names(MM)
 for(j in 1:length(MM)){
  list.ecol[[j]]<-VPcutoff(MM[[j]],ENV,Type)
 } #end for
 return(list.ecol)
```

```
              } #end VP.taxa
} #end if "SAM"

result.VP<-VP.taxa(MM,ENV,Type)

  if(OUTP=="y"){
      for(i in 1:length(MM)){
        write.table(result.VP[[c(i,2)]],paste(names(MM[i]),".sum.adjRsq.txt",sep=""),quote=FALSE)
        write.table(result.VP[[c(i,1)]],paste(names(MM[i]),".VarPart.txt",sep=""),quote=FALSE)
      }
  } #end if

  if(PLOT=="y"){
    par(mfrow=c(round(length(MM)/2,0),2))
    for(i in 1:length(MM)){
      barplot(result.VP[[c(i,1)]][1:15,],ylim=c(0,1),col=seq(1:15),xlab=c("%cutoff
removed"),ylab=paste("VarPart_",names(MM[i]),sep=""))
    }
  legend(0.1,1,row.names(result.VP[[c(1,1)]][1:15,]),fill=seq(1:15),y.intersp=0.7)
  } #end if

return(result.VP)

} #end of VP.COL
```

```
# Copyright (C) 2010 Angelique Gobet & Alban Ramette
# This program is free software; you can redistribute it and/or modify
# it under the terms of the GNU General Public License as published by
# the Free Software Foundation; either version 2 of the License, or
# (at your option) any later version.
#
# This program is distributed in the hope that it will be useful,
# but WITHOUT ANY WARRANTY; without even the implied warranty of
# MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.  See the
# GNU General Public License for more details.


regcoeff<-function(SPE,ENV){
require(vegan)
res2<-matrix(NA,1,5)
colnames(res2)<-c("Sum",paste("RDA1.",colnames(ENV),sep=""))
        res2[,1]<-sum(SPE)
        if (nrow(SPE)<nrow(SPE)){
                        res2[,2:5]<-c(NA,NA,NA,NA)
                } # to avoid conflicts comparing original dataset/NA
                else {
                ENV1<-model.matrix(~.,as.data.frame(ENV[,1]))[,-1]
                ENV2<-model.matrix(~.,as.data.frame(ENV[,2]))[,-1]
                ENV3<-model.matrix(~.,as.data.frame(ENV[,3]))[,-1]
                ENV4<-model.matrix(~.,as.data.frame(ENV[,4]))[,-1]
                Q2<-decostand(SPE,"hel")[1:nrow(SPE),1:ncol(SPE)]
                #RDA1 axis values for each env par
                R1=rda(Q2~ENV1+ENV2+ENV3+ENV4)
                res2[,2:5]=summary(R1)$biplot[,"RDA1"]
                }
return(res2)
}
```

```
# Copyright (C) 2010 Angelique Gobet & Alban Ramette
# This program is free software; you can redistribute it and/or modify
# it under the terms of the GNU General Public License as published by
# the Free Software Foundation; either version 2 of the License, or
# (at your option) any later version.
#
# This program is distributed in the hope that it will be useful,
# but WITHOUT ANY WARRANTY; without even the implied warranty of
# MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.  See the
# GNU General Public License for more details.

signif<-function(SPE,ENV){
require(vegan)
result3<-matrix(NA,1,5)
colnames(result3)<-c("Pw","Pe1","Pe2","Pe3","Pe4")
#colnames(result3)<-c("whole.sig","ENV1.sig","ENV2.sig","ENV3.sig","ENV4.sig")

        ###Transform input as model matrices
        if (length(SPE)<=nrow(SPE)){
                result3<-c("NA","NA","NA","NA")
        } # to avoid conflicts comparing original dataset/NA
        else {
                ENV1<-model.matrix(~.,as.data.frame(ENV[,1]))[,-1]
                ENV2<-model.matrix(~.,as.data.frame(ENV[,2]))[,-1]
                ENV3<-model.matrix(~.,as.data.frame(ENV[,3]))[,-1]
                ENV4<-model.matrix(~.,as.data.frame(ENV[,4]))[,-1]
                Q2<-decostand(SPE,"hel")[1:nrow(SPE),1:ncol(SPE)]
                #significance
                ###significance of whole model
                whole<-permutest.cca(rda(Q2~ENV1+ENV2+ENV3+ENV4),permutations=1000)
                Sw<-sort(whole$F.perm)
                if(length(Sw[Sw>whole$F.0])==0){
                        result3[,1]<-c("<0.001")
                }
                else{
                        result3[,1]<-length(Sw[Sw>whole$F.0])/length(whole$F.perm)
                }
                ###significance of each environmental parameter
                ENV1.sig<-
permutest.cca(rda(Q2~ENV1+Condition(ENV2)+Condition(ENV3)+Condition(ENV4)),permutation=1000
,model="full")
                Se1<-sort(ENV1.sig$F.perm)
                if(length(Se1[Se1>ENV1.sig$F.0])==0){
                        result3[,2]<-c("<0.001")
                }
                else{
                        result3[,2]<-length(Se1[Se1>ENV1.sig$F.0])/length(ENV1.sig$F.perm)

                }
                ENV2.sig<-
permutest.cca(rda(Q2~ENV2+Condition(ENV1)+Condition(ENV3)+Condition(ENV4)),permutation=1000
,model="full")
                Se2<-sort(ENV2.sig$F.perm)
                if(length(Se2[Se2>ENV2.sig$F.0])==0){
                        result3[,3]<-c("<0.001")
                }
                else{
                        result3[,3]<-length(Se2[Se2>ENV2.sig$F.0])/length(ENV2.sig$F.perm)

                }
                ENV3.sig<-
permutest.cca(rda(Q2~ENV3+Condition(ENV2)+Condition(ENV1)+Condition(ENV4)),permutation=1000
,model="full")
                Se3<-sort(ENV3.sig$F.perm)
                if(length(Se3[Se3>ENV3.sig$F.0])==0){
                        result3[,4]<-c("<0.001")
                }
                else{
                        result3[,4]<-length(Se3[Se3>ENV3.sig$F.0])/length(ENV3.sig$F.perm)

                }
                ENV4.sig<-
permutest.cca(rda(Q2~ENV4+Condition(ENV2)+Condition(ENV3)+Condition(ENV1)),permutation=1000
,model="full")
                Se4<-sort(ENV4.sig$F.perm)
                if(length(Se4[Se4>ENV4.sig$F.0])==0){
                        result3[,5]<-c("<0.001")
                }
                else{
                        result3[,5]<-length(Se4[Se4>ENV4.sig$F.0])/length(ENV4.sig$F.perm)

                }
                #result3<-list(Pw,Pe1,Pe2,Pe3,Pe4)
                #names(result3)<-c("whole.sig","ENV1.sig","ENV2.sig","ENV3.sig","ENV4.sig")
        }

return(result3)
}

######################################################################
```

Literature

Achtman M, Wagner M (2008) Microbial diversity and the genetic nature of microbial species. Nat Rev Microbiol 6:431-440

Acinas SG, Klepac-Ceraj V, Hunt DE, Pharino C, Ceraj I, Distel DL, Polz MF (2004) Fine-scale phylogenetic architecture of a complex bacterial community. Nature 430:551-554

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. Appl Environ Microb 71:8966-8969

Al-Thukair AA, Abed RMM, Mohamed L (2007) Microbial community of cyanobacteria mats in the intertidal zone of oil-polluted coast of Saudi Arabia. Marine Pollution Bulletin 54:173-179

Andersson AF, Riemann L, Bertilsson S (2010) Pyrosequencing reveals contrasting seasonal dynamics of taxa within Baltic Sea bacterioplankton communities. The ISME Journal 4:171-181

Avaniss-Aghajani E, Jones K, Chapman D, Brunk C (1994) A Molecular Technique for Identification of Bacteria Using Small-Subunit Ribosomal-Rna Sequences. Biotechniques 17:144-&

Azam F, Worden AZ (2004) Microbes, molecules, and marine ecosystems. Science 303:1622-1624

Baas-Becking LGM (1934) Geobiologie of Inleiding Tot de Milieukunde, Vol p15, The Hague, The Netherlands

Begon M, Townsend CR, Harper JL (2006) Ecology: from individuals to ecosystems., Vol. Blackwell Publishing, Oxford, UK

Bent SJ, Forney LJ (2008) The tragedy of the uncommon: understanding limitations in the analysis of microbial diversity. The ISME Journal 2:689-695

Bertics VJ, Ziebis W (2009) Biodiversity of benthic microbial communities in bioturbated coastal sediments is controlled by geochemical microniches. The ISME Journal 3:1269-1285

Böer SI (2008) Investigation of the distribution and activity of benthic microorganisms in coastal habitats. University of Bremen.

Böer SI, Arnosti C, van Beusekom JEE, Boetius A (2008) Temporal variations in microbial activities and carbon turnover in subtidal sandy sediments. Biogeosciences Discuss 5:4271-4313

Böer SI, Hedtkamp SIC, van Beusekom JEE, Fuhrman JA, Boetius A, Ramette A (2009) Time- and sediment depth-related variations in bacterial diversity and community structure in subtidal sands. The ISME Journal 3:780-791

Bolker B, Bonebakker L, Gentleman R, Liaw WHA, Lumley T, Maechler M, Magnusson A, Moeller S, M. S, Venables B (2009) gplots: Various R programming tools for plotting data. R package version 2.7.4.

Borcard D, Legendre P, Drapeau P (1992) Partialling out the Spatial Component of Ecological Variation. Ecology 73:1045-1055

Boudreau BP, Huettel M, Forster S, Jahnke RA, McLachlan A, Middelburg JJ, Nielsen P, Sansone F, Taghon G, van Raaphorst W (2001) Permeable marine sediments: overturning an old paradigm. EOS 82:133-136

Bray JR, Curtis JT (1957) An Ordination of the Upland Forest Communities of Southern Wisconsin. Ecol Monogr 27:326-349

Brazelton WJ, Ludwig KA, Sogin ML, Andreishcheva EN, Kelley DS, Shen CC, Edwards RL, Baross JA (2010) Archaea and bacteria with surprising microdiversity show shifts in dominance over 1,000-year time scales in hydrothermal chimneys. P Natl Acad Sci USA 107:1612-1617

Brown MV, 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 7:1466-1479

Cardinale M, Brusetti L, Quatrini P, Borin S, Puglia AM, Rizzi A, Zanardini E, Sorlini C, Corselli C, Daffonchio D (2004) Comparison of different primer sets for use in automated ribosomal intergenic spacer analysis of complex bacterial communities. Appl Environ Microb 70:6147-6156

Casamayor EO, Schafer H, Baneras L, Pedros-Alio C, Muyzer G (2000) Identification of and spatio-temporal differences between microbial assemblages from two neighboring sulfurous lakes: Comparison by microscopy and denaturing gradient gel electrophoresis. Appl Environ Microb 66:499-508

Chao A (1984) Nonparametric-Estimation of the Number of Classes in a Population. Scand J Stat 11:265-270

Chao A, Bunge J (2002) Estimating the number of species in a Stochastic abundance model. Biometrics 58:531-539

Chao A, Lee SM (1992) Estimating the Number of Classes Via Sample Coverage. J Am Stat Assoc 87:210-217

Clarke KR (1993) Non-parametric Multivariate Analyses of Changes in Community Structure. Australian Journal of Ecology 18:117-143

Cottrell MT, Kirchman DL (2003) Contribution of major bacterial groups to bacterial biomass production (thymidine and leucine incorporation) in the Delaware estuary. Limnol Oceanogr 48:168-178

Crosby LD, Criddle CS (2003) Understanding bias in microbial community analysis techniques due to rrn operon copy number heterogeneity. Biotechniques 34:790-+

Curtis TP, Sloan WT (2004) Prokaryotic diversity and its limits: microbial community structure in nature and implications for microbial ecology. Curr Opin Microbiol 7:221-226

Curtis TP, Sloan WT (2005) Exploring microbial diversity - A vast below. Science 309:1331-1333

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

de Beer D, Wenzhofer F, Ferdelman TG, Boehme SE, Huettel M, van Beusekom JEE, Bottcher ME, Musat N, Dubilier N (2005) Transport and mineralization rates in North Sea sandy intertidal sediments, Sylt-Romo Basin, Wadden Sea. Limnol Oceanogr 50:113-127

de Wit R, Bouvier T (2006) 'Everything is everywhere, but, the environment selects'; what did Baas Becking and Beijerinck really say? Environmental Microbiology 8:755-758

Dinsdale EA, Pantos O, Smriga S, Edwards RA, Angly F, Wegley L, Hatay M, Hall D, Brown E, Haynes M, Krause L, Sala E, Sandin SA, Thurber RV, Willis BL, Azam

F, Knowlton N, Rohwer F (2008) Microbial Ecology of Four Coral Atolls in the Northern Line Islands. Plos One 3:-

Eilers H, Pernthaler J, Glockner FO, Amann R (2000) Culturability and in situ abundance of pelagic bacteria from the North Sea. Appl Environ Microb 66:3044-3051

Emery K (1968) Relict sediments on continental shelves of the world. Am Assoc Pet Geol Bull 52:445-464

Epstein SS, Alexander D, Cosman K, Dompe A, Gallagher S, Jarsobski J, Laning E, Martinez R, Panasik G, Peluso C, Runde R, Timmer E (1997) Enumeration of sandy sediment bacteria: Are the counts quantitative or relative? Mar Ecol-Prog Ser 151:11-16

Falkowski PG, Fenchel T, Delong EF (2008) The microbial engines that drive Earth's biogeochemical cycles. Science 320:1034-1039

Fenchel T, Finlay BJ (2004) The ubiquity of small species: Patterns of local and global diversity. Bioscience 54:777-784

Fierer N, Hamady M, Lauber CL, Knight R (2008) The influence of sex, handedness, and washing on the diversity of hand surface bacteria. P Natl Acad Sci USA 105:17994-17999

Finlay BJ (2002) Global dispersal of free-living microbial eukaryote species. Science 296:1061-1063

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

Flemming H-C, Wingender J (2010) The biofilm matrix. Nat Rev Micro 8:623-633

Fuhrman JA, Hewson I, Schwalbach MS, Steele JA, Brown MV, Naeem S (2006) Annually reoccurring bacterial communities are predictable from ocean conditions. P Natl Acad Sci USA 103:13104-13109

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

Galand PE, Casamayor EO, Kirchman DL, Potvin M, Lovejoy C (2009b) Unique archaeal assemblages in the Arctic Ocean unveiled by massively parallel tag sequencing. The ISME Journal 3:860-869

Galand PE, Potvin M, Casamayor EO, Lovejoy C (2010) Hydrography shapes bacterial biogeography of the deep Arctic Ocean. The ISME Journal 4:564-576

Gerbilskii NL, Petrunkevitch A (1955) Intraspecific Biological Groups of Acipenserines and Their Reproduction in the Lower Regions of Rivers with Regulated Flow. Systematic Zoology 4:86-92

Gilbert JA, Field D, Swift P, Newbold L, Oliver A, Smyth T, Somerfield PJ, Huse S, Joint I (2009) The seasonal structure of microbial communities in the Western English Channel. Environmental Microbiology 11:3132-3139

Gillevet PM, Sikaroodi M, Torzilli AP (2009) Analyzing salt-marsh fungal diversity: comparing ARISA fingerprinting with clone sequencing and pyrosequencing. Fungal Ecology 2:160-167

Glöckner FO, Fuchs BM, Amann R (1999) Bacterioplankton compositions of lakes and oceans: a first comparison based on fluorescence in situ hybridization. Appl Environ Microb 65:3721-3726

Gobet A, Böer SI, Huse S, van Beusekom J, Quince C, Sogin ML, Boetius A, Ramette A (*Submitted*) Diversity and dynamics of the rare and resident bacterial biosphere in coastal sands. Proc Natl Acad Sci USA

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

Gotelli NJ, Colwell RK (2001) Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. Ecology Letters 4:379-391

Gower JC (1966) Some Distance Properties of Latent Root and Vector Methods Used in Multivariate Analysis. Biometrika 53:325-338

Green J, Bohannan BJM (2006) Spatial scaling of microbial biodiversity. Trends Ecol Evol 21:501-507

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

Harper JL, Hawksworth DL (1995) Preface. In: Biodiversity: measurement and estimation. Chapman & Hall, London, UK., p 5-12

Hawksworth DL, Kalin-Arroyo MT (1995) In: Heywood VH (ed) Global biodiversity assessment. Cambridge University Press, Cambridge, MA, USA, p 107-191

Hewson I, Fuhrman JA (2006) Improved strategy for comparing microbial assemblage fingerprints. Microb Ecol 51:147-153

Hong SH, Bunge J, Jeon SO, Epstein SS (2006) Predicting microbial species richness. P Natl Acad Sci USA 103:117-122

Horner-Devine MC, Lage M, Hughes JB, Bohannan BJM (2004) A taxa-area relationship for bacteria. Nature 432:750-753

Hubbell SP (2001) The Unified Neutral Theory of Biodiversity and Biogeography., Vol. Princeton University Press, Princeton, NJ.

Huber JA, Mark Welch D, Morrison HG, Huse SM, Neal PR, Butterfield DA, Sogin ML (2007) Microbial population structures in the deep marine biosphere. Science 318:97-100

Huber JA, Morrison HG, Huse SM, Neal PR, Sogin ML, Welch DBM (2009) Effect of PCR amplicon size on assessments of clone library microbial diversity and community structure. Environmental Microbiology 11:1292-1302

Hughes JB, Hellmann JJ, Ricketts TH, Bohannan BJM (2001) Counting the uncountable: Statistical approaches to estimating microbial diversity. Appl Environ Microb 67:4399-4406

Huse SM, Huber JA, Morrison HG, Sogin ML, Mark Welch D (2007) Accuracy and quality of massively parallel DNA pyrosequencing. Genome Biol 8:R143

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

Jaccard P (1901) Etude comparative de la distribution florale dans une portion des Alpes et du Jura. Bulletin de la société vaudoise des sciences naturelles 37:241-272

Kemp PF, Aller JY (2004) Bacterial diversity in aquatic and other environments: what 16S rDNA libraries can tell us. Fems Microbiology Ecology 47:161-177

Kirchman DL, Cottrell MT, Lovejoy C (2010) The structure of bacterial communities in the western Arctic Ocean as revealed by pyrosequencing of 16S rRNA genes. Environmental Microbiology 12:1132-1143

Kjelleberg S, Hermansson M, Marden P (1987) The Transient Phase between Growth and Nongrowth of Heterotrophic Bacteria, with Emphasis on the Marine-Environment. Annual Review of Microbiology 41:25-49

Kloeke FV, Baty AM, Eastburn CC, Diwu Z, Geesey GG (1999) Novel method for screening bacterial colonies for phosphatase activity. Journal of Microbiological Methods 38:25-31

Konopka A (2009) What is microbial community ecology? Isme J 3:1223-1230

Koopman MM, Fuselier DM, Hird S, Carstens BC (2010) The Carnivorous Pale Pitcher Plant Harbors Diverse, Distinct, and Time-Dependent Bacterial Communities. Appl Environ Microb 76:1851-1860

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

Lauber CL, Hamady M, Knight R, Fierer N (2009) Pyrosequencing-Based Assessment of Soil pH as a Predictor of Soil Bacterial Community Structure at the Continental Scale. Appl Environ Microb 75:5111-5120

Legendre L, Legendre P (1998) Numerical Ecology, Vol. Elsevier Science BV, Amsterdam, The Netherlands

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

Legendre P, Gallagher ED (2001) Ecologically meaningful transformations for ordination of species data. Oecologia 129:271-280

Llobet-Brossa E, Rossello-Mora R, Amann R (1998) Microbial community composition of Wadden Sea sediments as revealed by fluorescence in situ hybridization. Appl Environ Microb 64:2691-2696

Lorenzen CJ (1967) Determination of Chlorophyll and Pheo-Pigments - Spectrophotometric Equations. Limnol Oceanogr 12:343-&

Mace GM (1995) Classification of threatened species and its role in conservation planning. In: Extinction rates. Oxford University Press., Oxford, UK., p 197-213

Magurran AE (2004) Measuring Biological Diversity., Vol. Blackwell Publishing, Oxford, UK

Magurran AE, Henderson PA (2003) Explaining the excess of rare species in natural species abundance distributions. Nature 422:714-716

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen ZT, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer MLI, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu PG, Begley RF, Rothberg JM (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376-380

Martiny JBH, Bohannan BJM, Brown JH, Colwell RK, Fuhrman JA, Green JL, Horner-Devine MC, Kane M, Krumins JA, Kuske CR, Morin PJ, Naeem S, Ovreas L, Reysenbach AL, Smith VH, Staley JT (2006) Microbial biogeography: putting microorganisms on the map. Nat Rev Microbiol 4:102-112

Mayr E (1942) Systematics and the origin of species., Vol. Columbia University Press, New York, NY, USA

Musat N, Werner U, Knittel K, Kolb S, Dodenhof T, van Beusekom JEE, de Beer D, Dubilier N, Amann R (2006) Microbial community structure of sandy intertidal sediments in the North Sea, Sylt-Romo Basin, Wadden Sea. Syst Appl Microbiol 29:333-348

Muyzer G, Dewaal EC, Uitterlinden AG (1993) Profiling of Complex Microbial-Populations by Denaturing Gradient Gel-Electrophoresis Analysis of Polymerase Chain Reaction-Amplified Genes-Coding for 16s Ribosomal-Rna. Appl Environ Microb 59:695-700

Naviaux RK, Good B, McPherson JD, Steffen DL, Markusic D, Ransom B, Corbeil J (2005) Sand DNA - a genetic library of life at the water's edge. Mar Ecol-Prog Ser 301:9-22

Norse EA (1986) Conserving biological diversity in our national forests., Vol. Wilderness Society, Washington, D.C., USA

Ochman H, Lawrence JG, Groisman EA (2000) Lateral gene transfer and the nature of bacterial innovation. Nature 405:299-304

Oksanen J, Kindt R, Legendre P, O'Hara B, Simpson GL, Solymos P, Stevens MHH, Wagner H (2009) vegan: Community Ecology Package. R package version 1.15-2

Olsen GJ, Lane DJ, Giovannoni SJ, Pace NR, Stahl DA (1986) Microbial Ecology and Evolution - a Ribosomal-Rna Approach. Annual Review of Microbiology 40:337-365

Pace NR (1997) A molecular view of microbial diversity and the biosphere. Science 276:734-740

Pachepsky E, Crawford JW, Bown JL, Squire G (2001) Towards a general theory of biodiversity. Nature 410:923-926

Parker BO, Marinus MG (1992) Repair of DNA Heteroduplexes Containing Small Heterologous Sequences in Escherichia-Coli. P Natl Acad Sci USA 89:1730-1734

Pearman PB, Guisan A, Broennimann O, Randin CF (2008) Niche dynamics in space and time. Trends Ecol Evol 23:149-158

Pearson K (1901) Mathematical contributions to the theory of evolution - VII On the correlation of characters of not quantitatively measurable. Philosophical Transactions of the Royal Society of London Series a-Containing Papers of a Mathematical or Physical Character 195:1-47

Pedrós-Alió C (2006) Marine microbial diversity: can it be determined? Trends Microbiol 14:257-263

Pedrós-Alió C (2007) Dipping into the rare biosphere. Science 315:192-193

Pernthaler J (2005) Predation on prokaryotes in the water column and its ecological implications. Nat Rev Microbiol 3:537-546

Philippot L, Andersson SGE, Battin TJ, Prosser JI, Schimel JP, Whitman WB, Hallin S (2010) The ecological coherence of high bacterial taxonomic ranks. Nat Rev Microbiol 8:523-529

Pommier T, Canback B, Riemann L, Bostrom KH, Simu K, Lundberg P, Tunlid A, Hagstrom A (2007) Global patterns of diversity and community structure in marine bacterioplankton. Mol Ecol 16:867-880

Prosser JI (2010) Replicate or lie. Environmental Microbiology 12:1806-1810

Purvis A, Hector A (2000) Getting the measure of biodiversity. Nature 405:212-219

Putman RJ (1994) Community Ecology., Vol. Chapman and Hall, London, UK.

Qiu XY, Wu LY, Huang HS, McDonel PE, Palumbo AV, Tiedje JM, Zhou JZ (2001) Evaluation of PCR-generated chimeras: Mutations, and heteroduplexes with 16S rRNA gene-based cloning. Appl Environ Microb 67:880-887

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

Quince C, Lanzen A, Curtis TP, Davenport RJ, Hall N, Head IM, Read LF, Sloan WT (2009) Accurate determination of microbial diversity from 454 pyrosequencing data. Nature Methods 6:639-641

Quinlan AR, Stewart DA, Stromberg MP, Marth GT (2008) Pyrobayes: an improved base caller for SNP discovery in pyrosequences. Nature Methods 5:179-181

R Development Core Team (2009) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria

Ramette A (2007) Multivariate analyses in microbial ecology. Fems Microbiology Ecology 62:142-160

Ramette A, Tiedje JM (2007a) Biogeography: An emerging cornerstone for understanding prokaryotic diversity, ecology, and evolution. Microb Ecol 53:197-207

Ramette A, Tiedje JM (2007b) Multiscale responses of microbial life to spatial distance and environmental heterogeneity in a patchy ecosystem. P Natl Acad Sci USA 104:2761-2766

Reeder J, Knight R (2009) The 'rare biosphere': a reality check. Nature Methods 6:636-637

Reeder J, Knight R (2010) Rapidly denoising pyrosequencing amplicon reads by exploiting rank-abundance distributions. Nature Methods 7:668-669

Reise K, Gatje C (1997) The List tidal basin: a reference area for scientific research in the northern Wadden Sea. Helgolander Meeresuntersuchungen 51:249-251

Roesch LF, Fulthorpe RR, Riva A, Casella G, Hadwin AKM, Kent AD, Daroub SH, Camargo FAO, Farmerie WG, Triplett EW (2007) Pyrosequencing enumerates and contrasts soil microbial diversity. The ISME Journal 1:283-290

Roesch LFW, Lorca GL, Casella G, Giongo A, Naranjo A, Pionzio AM, Li N, Mai V, Wasserfall CH, Schatz D, Atkinson MA, Neu J, Triplett EW (2009) Culture-independent identification of gut bacteria correlated with the onset of diabetes in a rat model. The ISME Journal 3:536-548

Roh SW, Kim KH, Nam YD, Chang HW, Park EJ, Bae JW (2010) Investigation of archaeal and bacterial diversity in fermented seafood using barcoded pyrosequencing. The ISME Journal 4:1-16

Rosenzweig RF, Sharp RR, Treves DS, Adams J (1994) Microbial Evolution in a Simple Unstructured Environment - Genetic Differentiation in Escherichia-Coli. Genetics 137:903-917

Rossello-Mora R, Amann R (2001) The species concept for prokaryotes. Fems Microbiol Rev 25:39-67

Rothberg JM, Leamon JH (2008) The development and impact of 454 sequencing. Nat Biotechnol 26:1117-1124

Ruppert J, Panzig B, Guertler L, Hinz P, Schwesinger G, Felix SB, Friesecke S (2004) Two cases of severe sepsis due to Vibrio vulnificus wound infection acquired in the Baltic Sea. European Journal of Clinical Microbiology & Infectious Diseases 23:912-915

Rusch A, Forster S, Huettel M (2001) Bacteria, diatoms and detritus in an intertidal sandflat subject to advective transport across the water-sediment interface. Biogeochemistry 55:1-27

Rusch A, Huettel M, Reimers CE, Taghon GL, Fuller CM (2003) Activity and distribution of bacterial populations in Middle Atlantic Bight shelf sands. Fems Microbiology Ecology 44:89-100

Sala OE, Chapin FS, Armesto JJ, Berlow E, Bloomfield J, Dirzo R, Huber-Sanwald E, Huenneke LF, Jackson RB, Kinzig A, Leemans R, Lodge DM, Mooney HA, Oesterheld M, Poff NL, Sykes MT, Walker BH, Walker M, Wall DH (2000) Biodiversity - Global biodiversity scenarios for the year 2100. Science 287:1770-1774

Sanger F, Nicklen S, Coulson AR (1977) DNA Sequencing with Chain-Terminating Inhibitors. P Natl Acad Sci USA 74:5463-5467

Santillano D, Boetius A, Ramette A (2010) Improved dsrA-Based Terminal Restriction Fragment Length Polymorphism Analysis of Sulfate-Reducing Bacteria. Appl Environ Microb 76:5308-5311

Schloss PD, Handelsman J (2005) Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. Appl Environ Microb 71:1501-1506

Schwarzenbach K, Enkerli J, Widmer F (2007) Objective criteria to assess representativity of soil fungal community profiles. Journal of Microbiological Methods 68:358-366

Shepard RN (1966) Metric Structures in Ordinal Data. Journal of Mathematical Psychology 3:287-315

Sloan WT, Lunn M, Woodcock S, Head IM, Nee S, Curtis TP (2006) Quantifying the roles of immigration and chance in shaping prokaryote community structure. Environmental Microbiology 8:732-740

Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, Neal PR, Arrieta JM, Herndl GJ (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere". P Natl Acad Sci USA 103:12115-12120

Stal LJ (2003) Microphytobenthos, their extracellular polymeric substances, and the morphogenesis of intertidal sediments. Geomicrobiol J 20:463-478

Stauffer RC (1957) Haeckel, Darwin, and Ecology. Quarterly Review of Biology 32:138-144

Stewart JR, Gast RJ, Fujioka RS, Solo-Gabriele HM, Meschke JS, Amaral-Zettler LA, del Castillo E, Polz MF, Collier TK, Strom MS, Sinigalliano CD, Moeller PDR, Holland AF (2008) The coastal environment and human health: microbial indicators, pathogens, sentinels and reservoirs. Environmental Health 7:S3

Stoodley P, Dodds I, De Beer D, Scott HL, Boyle JD (2005) Flowing biofilms as a transport mechanism for biomass through porous media under laminar and turbulent conditions in a laboratory reactor system. Biofouling 21:161-168

ter Braak CJF, Šmilauer P (2002) CANOCO Reference Manual and CanoDraw for Windows User's Guide: Software for Canonical Community Ordination, Version 4.5.

Thingstad TF (2000) Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic systems. Limnol Oceanogr 45:1320-1328

Torsvik V, Ovreas L, Thingstad TF (2002) Prokaryotic diversity - Magnitude, dynamics, and controlling factors. Science 296:1064-1066

Urakawa H, Yoshida T, Nishimura M, Ohwada K (2000) Characterization of depth-related population variation in microbial communities of a coastal marine sediment using 16S rDNA-based approaches and quinone profiling. Environmental Microbiology 2:542-554

van Beusekom JEE (2005) A historic perspective on Wadden Sea eutrophication. Helgoland Mar Res 59:45-54

van Beusekom JEE, de Jonge VN (2002) Long-term changes in Wadden Sea nutrient cycles: importance of organic matter import from the North Sea. Hydrobiologia 475-476:185-194

van Beusekom JEE, Loebl M, Martens P (2009) Distant riverine nutrient supply and local temperature drive the long-term phytoplankton development in a temperate coastal basin. J Sea Res 61:26-33

Vellend M (2010) Conceptual Synthesis in Community Ecology. Quarterly Review of Biology 85:183-206

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu DY, Paulsen I, Nelson KE, Nelson W, Fouts DE, Levy S, Knap AH, Lomas MW, Nealson K, White O, Peterson J, Hoffman J, Parsons R, Baden-Tillson H, Pfannkoch C, Rogers YH, Smith HO (2004) Environmental genome shotgun sequencing of the Sargasso Sea. Science 304:66-74

Whitman WB, Coleman DC, Wiebe WJ (1998) Prokaryotes: The unseen majority. P Natl Acad Sci USA 95:6578-6583

Whittaker RH (1960) Vegetation of the Siskiyou Mountains, Oregon and California. Ecol Monogr 30:280-338

Whittaker RH (1972) Evolution and Measurement of Species Diversity. Taxon 21:213-251

Whittaker RH (1975) Communities and Ecosystems, Vol. Macmillan, New York, NY, USA

Whittaker RH, Levin SA, Root RB (1973) Niche, Habitat, and Ecotope. Am Nat 107:321-338

Wilson EO (1992) The diversity of life, Vol. Harvard University Press, Harvard, MA, USA

Wilson EO (2000) On the future of conservation biology. Conservation Biology 14:1-3

Woese CR (1987) Bacterial Evolution. Microbiological Reviews 51:221-271

Yannarell AC, Triplett EW (2005) Geographic and environmental sources of variation in lake bacterial community composition. Appl Environ Microb 71:227-239

Zhang HS, DiBaise JK, Zuccolo A, Kudrna D, Braidotti M, Yu YS, Parameswaran P, Crowell MD, Wing R, Rittmann BE, Krajmalnik-Brown R (2009) Human gut microbiota in obesity and after gastric bypass. P Natl Acad Sci USA 106:2365-2370

Zinger L, Amaral-Zettler LA, Fuhrman JA, Horner-Devine MC, Huse SM, Mark Welch DB, Martiny JBH, Neal PR, Sogin ML, Boetius A, Ramette A (*In preparation*) Global patterns of marine bacterial communities from the pelagic and benthic realms.

Zubkov MV, Fuchs BM, Tarran GA, Burkill PH, Amann R (2002) Mesoscale distribution of dominant bacterioplankton groups in the northern North Sea in early summer. Aquat Microb Ecol 29:135-144

Acknowledgements

First, I would like to thank the jury members, for taking the time of evaluating the PhD thesis and PhD defense: Antje Boetius, Matthias Ullrich, and Alban Ramette. The other thesis committee members are also acknowledged for their helpful and constructive advices through the PhD: Justus van Beusekom and Frank Oliver Glöckner.

This PhD work would have never been possible without my two supervisors: Antje Boetius and Alban Ramette. I would like to gratefully thank them for their trust, for giving me the opportunity to learn so much during this PhD work, and to "grow up" as a scientist. Thank you also for your support and understanding when the private life may influence the work…and especially when I had all my biking “issues” ☺

I would like to warmly thank Lucie, Melissa, Petra, Gunter and Ilaria that offered their time to comment and proofread this PhD thesis.

Many thanks to the really friendly Microbial habitat group, especially Susanne, Martina, Erika, and Rafael that were of great help during sampling and/or laboratory work.
Thank you, Shalin and Marianne, for your help and the nice time during the lab rotation.
Thanks a lot to the participants of the habitat lunch, and the ComE meetings.
I also would like to thank my former and current office mates: Gunter, Daniel, Felix, Jan, Simone and especially my friend Sandra with whom I had great conversations and support during nice and tough moments of the PhD and…Salsa Moments.

I would like to thank the MarMic colleagues and faculty and especially Christiane Glöckner for her warm welcoming in Bremen and her help through the process of becoming "Bremerin".

Many, many thanks to meine Bremer Familie, that made me feel home in Bremen by cooking, dancing, partying, travelling: Gunter, Marie, Luciana, Paola, Melissa, Angela, Pablo, Daniel, Maya, Ana, Francesca, Alexandra, Chia-I, Daphne, Lucie, Jean, Yann,

James, Tobias and many other friends. I also would like to thank across borders Belgium supporters: Yannick & Pedro.

I would like to deeply thank Ilaria, for giving the best advices and being always supportive, until the very end of the thesis. Grazie mille per tu grande aiuto amica mia! Sei sempre la piú figa!

Enfin, je remercie ma famille: Christelle, ma mère et JB, pour toujours me soutenir et être toujours présents.

I am really grateful to all of you that helped me making this PhD possible and made my stay in Bremen so nice.