Microarrays

Current Technology, Innovations and Applications

Edited by
Zhili He

Caister Academic Press

Chapter from:

Microarrays
Current Technology, Innovations and Applications

Edited by Zhili He

ISBN: 978-1-908230-49-2 (hardback)
ISBN: 978-1-908230-59-1 (ebook)

Microarray of 16S rRNA Gene Probes for Quantifying Population Differences Across Microbiome Samples

5

Alexander J. Probst, Pek Yee Lum, Bettina John, Eric A. Dubinsky, Yvette M. Piceno, Lauren M. Tom, Gary L. Andersen, Zhili He and Todd Z. DeSantis

Abstract

Deciphering microbial communities and their role in Earth's biosphere is crucial for addressing challenges in human health, agriculture, bioremediation and other natural processes. While next-generation sequencing platforms are still under development to improve accuracy, read length and sequencing depth, microarray-based methods have become an attractive alternative for 16S rRNA gene microbial community comparisons. The hybridization method is well-established in the laboratory. Thus, main areas of improvement lie with the development of enhanced bioinformatics and statistical procedures for microarray data, rather than with improvements to the platform itself. In this communication we applied recently-developed bioinformatics tools to re-analyse G3 PhyloChip™ DNA microarray data acquired from deep ocean samples collected during the 2010 Deepwater Horizon oil spill in the Gulf of Mexico. We show that data collected with the G3 PhyloChip assay can be analysed at various stages of resolution, from individual probes to pairs of probes to quartets of probes and finally at the commonly used probe-set level where each probe-set is associated with one operational taxonomic unit (OTU). Our analysis methods comprised topological data analysis (TDA) to facilitate the detection of outlier bio-specimens and the reconstruction of empirical OTUs (eOTUs) in an unsupervised manner, without the need for pre-defined reference OTUs (rOTUs). We observed that the quartet level provided sufficient resolution for identifying a subtle outlier sample with TDA, while the eOTU reconstruction was useful for annotation of the taxa associated with significant population changes in the elevated hydrocarbon waters. The presented methods will improve the deduction of important biological processes from G3 PhyloChip experiments.

Introduction

Microorganisms represent the greatest biomass of Earth's biosphere. They have the largest metabolic variety of all organisms and consequently contribute profoundly to carbon, nitrogen, sulfur and phosphorus cycling on the planet (Falkowski *et al.*, 2008). These cycles are necessary to maintain life across all kingdoms, support environmental homeostasis by bioremediation of pollutants, and process nutrients in the human gut as examples. In their natural habitat, microorganisms have been shown to act as a community rather than as monospecies with independent metabolism, although some exceptional cases have been posited (Chivian *et al.*, 2008). Communities are assemblages of tens to thousands of species, whose individual populations fluctuate based on changes in local stimuli. Thus far, we have incomplete knowledge of community dynamics and the impact on metabolic networks, since microbiologists have mostly studied microbes under artificial laboratory conditions where typically single strains of bacteria or archaea are monitored. The diversity of strains observed in a laboratory is limited by our knowledge of nutrient, temperature, and atmospheric needs of a given organism. Consequently, it is not surprising that the number of microorganisms that can be cultured under artificial laboratory conditions was estimated to be only 1% of all bacteria and archaea that had been discovered via culture-independent molecular techniques (Amann *et al.*, 1995; Colwell, 1997). Recent metagenomic studies, however, increase the discovery of biological dark matter and help illuminate uncultured microorganisms (Castelle *et al.*, 2013; Marcy *et al.*, 2007; Rinke *et al.*, 2013; Wrighton *et al.*, 2012). A recently documented example of uncultured archaea and bacteria interdependence is the metabolic pathway of anaerobic methane oxidation. Here, archaea are able to oxidize methane, which is found in high amounts in the deep ocean, because of the exergonic redox potential produced by bacteria through active sulfate reduction (Moran *et al.*, 2008; Orphan *et al.*, 2001). Neither of the two organisms has been grown in pure culture under laboratory conditions because of their strong interdependence via synthophy (Morris *et al.*, 2013).

Based on the low cultivability of most microbes and their interdependence as a community, examining all microorganisms' population dynamics simultaneously, or tracking the whole microbiome, in the natural environment or within clinical studies is desirable. The urgency of applying whole-microbiome methods is growing as diseases have become linked to the microbiome (Cho and Blaser, 2012). Determination of the microorganisms undergoing population changes in response to a stimulus is the primary step in the elucidation of metabolic interactions within a community. For example, renal disease is known to raise systemic urea and was hypothesized to influence nitrogen metabolic flux in the gut mucosal environment. As a first step in mapping the microbial enzymatic pathways consequently up-regulated, populations of all known bacteria were profiled from faeces of uraemic patients and compared to control patients to determine the significantly increased populations within *Halomonadaceae*, *Moraxellaceae*, *Polyangiaceae* and other families (Vaziri *et al.*, 2013).

The field of microbial ecology aims to understand microorganisms in their natural habitat and how they interact with biotic (e.g. other organisms) and abiotic factors (e.g. temperature). Similar to all branches of ecology, microbial ecology collects data on population shifts in order to propose hypotheses towards further understanding their mechanisms. Some examples to date include the study of healthy and diseased states as investigated by the Human Microbiome Project (Wortman *et al.*, 2010), between wild type and knock-out mice (Frantz *et al.*, 2012; Noval Rivas *et al.*, 2013), between soils in dissimilar biomes as investigated by the Earth Microbiome Project (Gilbert *et al.*, 2010a,b) and between contaminated and pristine waters as investigated by bioremediation scientists (Dubinsky *et al.*, 2012; Lin *et al.*, 2006). The most useful techniques are those that provide reproducible detection of population changes across a diverse range of all known bacteria. Furthermore, those techniques should not be limited to monitoring only the dominant populations, but should be sensitive to shifts in minority populations as well, since some microbes are 10,000-fold less abundant than majority members.

The method employed in thousands of published manuscripts to collect data on population dynamics is the amplification, classification and quantification of 16S rRNA genes from an entire community. The 16S rRNA gene with its nine hyper-variable regions spread over approximately 1.5 kb is ideal for straightforward amplification using primers that flank the hyper-variable spans. Typically, the amplification step creates over 1 μg of amplicons comprising 10 billion to a trillion double-stranded DNA molecules. The base sequence differences allow taxonomic identification (DeSantis *et al.*, 2006). The quantities of each taxa-specific amplicon can be compared across patient groups, biomes, etc. Initial enumerations with cloning and sequencing of the 16S rRNA gene amplicons sampled tens to thousands of molecules per specimen (as an example see Radosevich *et al.*, 2002). Later, massive parallel amplicon sequencing, or next-generation sequencing (NGS) popularized by Roche and Illumina platforms, allowed molecular barcoding of multiple biospecimens followed by inexpensive clone-less sequencing of multiple biospecimens that simultaneously enabled the routine sampling of 10K to 100K short 16S rRNA molecules per specimen (Jumpstart HMP Consortium, 2012). By 2006 general NGS optimism was at its peak due to a series of successful shotgun genome sequencing projects where short overlapping reads were assembled into a consensus genome

(Moore *et al.*, 2006). The optimism was extended to 16S applications with the assumption that sequencing quality, lengths, and depths would continue to improve so that in standard production individual reads could reliably be associated with specific taxa. In fact, it was demonstrated *in silico* that short 250 base subsegments of Sanger-derived 16S rRNA genes could allow identification of many genera (Liu *et al.*, 2008) and it was expected that real NGS reads would perform similarly.

As NGS 16S rRNA gene sequence data sets were published, the limitations of NGS protocols became better understood. As commonly practised, barcoding and multiplexing samples in cost-saving work-flows results in both (1) barcode biases, where perceived community structure could be influenced by the barcode assigned (Alon *et al.*, 2011; Berry *et al.*, 2011) and, (2) non-uniform sampling depth, where some samples within a multiplexed experiment are sampled at only ~1/10th to ~1/100th as thoroughly as other samples (HMP Consortium, 2012b), creating difficulties in comparative metrics of alpha-diversity. Currently, it is not expected that NGS results would be reproducible for the human microbiome or other complex communities (Zhou *et al.*, 2011, 2013), both due to the limitations of sequencing only thousands to millions of reads out of the billions generated (Caporaso *et al.*, 2012b; Haegeman *et al.*, 2013) and due to base-calling errors not explained by quality scores making individual bases or read segments difficult to filter (Engelbrektson *et al.*, 2010; Kunin *et al.*, 2010). Diversity, or specifically richness, is largely inflated by the NGS technique itself (Kunin *et al.*, 2010; Quince *et al.*, 2009; Reeder and Knight, 2009) adding unverifiable ribotypes to compositional data. High numbers of chimeras are created but are difficult to identify (Edgar *et al.*, 2011; Haas *et al.*, 2011) and the filtering stringency affects alpha-diversity. The previously held assumption that 16S rRNA NGS allowed a true representation of the community structure (percentage of community attributed to each taxon population) within a sample has now been called into question since the comparison of the same 21 member 'even' mock community (HMP Consortium, 2012a) processed through both Roche and Illumina platforms gave surprisingly different perceptions of what organisms dominated the community. For example, depending on the protocol, *Staphylococcus aureus* was perceived as comprising from 2.8% to as much as 50.1% of genomic DNA where a known proportion of approximately 5% was prepared (Edgar, 2013). This unsettling recent report will surely generate further mock community experiments and protocol refinement since it was based on gDNA preparations in separate laboratories, which may contribute to the unexpectedly high 17-fold variation. Since NGS readings of 16S rRNA gene amplicons do not produce an unambiguous documentation of the underling community, results should be interpreted only in comparisons to samples processed with the *exact* same protocol.

In 2010, the third generation of broad-spectrum phylogenetic microarrays, the G3 PhyloChip, was introduced to the field of environmental microbiology (Hazen *et al.*, 2010). Prior to this, the second generation, or G2 PhyloChip (DeSantis *et al.*, 2007) had been applied to the study of urban aerosols (Brodie *et al.*, 2007), clean room environments (Probst *et al.*, 2010), CO_2 elevated soil communities (He *et al.*, 2012) and various human clinical samples (Maldonado-Contreras *et al.*, 2011; Saulnier *et al.*, 2011). G2 PhyloChip probes were designed to complement only the sense strand of the contemporary knowledge base of sequences (DeSantis *et al.*, 2006). The G3 chip is preferred over the G2 due to the broadening of 16S rRNA gene sequence diversity uncovered in recent years (McDonald *et al.*, 2012). Compared to the second generation of the PhyloChip, the G3 contains a greater diversity of probes complimenting both strands, sense and anti-sense, of 16S rRNA gene sequences for greater sensitivity to population dynamics across diverse taxa. The array comprises a square grid of 1,016,064 features containing 994,980 different probes complementing all known taxa, cultured and yet-to-be-cultured, within both the Archaea and Bacteria. It contains mis-matching probes that serve as controls for each complementing probe and replicate probes for internal assessment of signal variation (Hazen *et al.*, 2010).

The PhyloChip approach overcomes many of the NGS difficulties by separating each biospecimen on physically separate arrays, eliminating barcode biases and avoiding non-uniform sampling depth otherwise introduced by multiplexing. The most attractive feature of the PhyloChip approach is the ability to expose up to 1 trillion amplicons to the probes. Since the probes are spatially separated, data are collected from each probe in each experiment, allowing minority community components to be measured. It is impossible in the foreseeable future for NGS workflows to take routine measurements on the minority components since typically ~1 million (which is only 1 millionth of 1 trillion) or fewer molecules are represented in the data. The ability of the PhyloChip G3 array to track changes in bacterial and archaeal abundance was evaluated using a latin-square quantitative design with 26 different microbial species mixed in 26 randomized concentration levels from 1× to 0.0005×, replicated three times on different days. Technical variation, or the variation in replicate measurements for the identical probe sequence, was low ($CV = 0.097$) and the correlation between analyte concentration and fluorescence signal was high ($r = 0.941$) (Hazen *et al.*, 2010), indicating that the PhyloChip G3 is reproducibly sensitive to changes in microbial abundance within complex samples across four orders of magnitude. So far, the PhyloChip G3 has been successfully used on samples of a great variety of different biotopes, such as human microbiome (Renwick *et al.*, submitted), sulfidic freshwater springs (Probst *et al.*, 2013), low biomass clean room environments (Cooper *et al.*, 2011; Vaishampayan *et al.*, 2013), animal models of disease (Lam *et al.*, 2012; Miezeiewski *et al.*, submitted), high-altitude bioaerosols (Smith *et al.*, 2013), coral reefs (Kellogg *et al.*, 2012, 2013), citrus plant leaves (Zhang *et al.*, 2013), hot springs (Briggs *et al.*, 2013), wastewater treatment reactors (Wang *et al.*, 2013), biofuel reactors (Wu and He, 2013), and soils (Ding *et al.*, 2013; Hayden *et al.*, 2012; Mendes *et al.*, 2011). The first published G3 study examined water samples from the Gulf of Mexico during the Deepwater Horizon oil spill (Hazen *et al.*, 2010), one of the greatest environmental catastrophes in the history of the production and conversion of fossil fuels. Approximately 4.1 million barrels of oil were released into the Gulf of Mexico between April and July 2010, which comprised a variety of liquid hydrocarbons (saturated, aromatic and polar) and gaseous components (Camilli *et al.*, 2010; Hazen *et al.*, 2010; Kessler *et al.*, 2011; Reddy *et al.*, 2012; Valentine *et al.*, 2010). From a microbiological perspective, the catastrophe has been studied with many different tools including metagenomics and metatranscriptomics to understand its effects on the natural environment of the ocean (Baelum *et al.*, 2012; Dubinsky *et al.*, 2013; Hazen *et al.*, 2010; Mason *et al.*, 2012; Redmond and Valentine, 2012; Rivers *et al.*, 2013; Valentine *et al.*, 2010).

In this book chapter, we re-analysed PhyloChip G3 data collected during the Deepwater Horizon oil spill using a novel bioinformatics technique as well as a topology-based approach called TDA (Lum *et al.*, 2013). Rather than using probe-sets complimenting database reference-based OTU, or 'rOTU', scoring (Hazen *et al.*, 2010), we employed a detailed, non-OTU, probe-by-probe approach as well as an empirical approach to define OTUs, or 'eOTUs', using taxonomically related probes highly correlated in observed fluorescence intensity (FI) across samples. We also present two intermediate pipelines and their effects on outlier determination and biological interpretation. Moreover, the metadata for each sample is utilized to find significant influences of environmental factors on the microbiome structures observed. The analysed dataset was acquired from 17 samples, 10 taken within and 7 outside the oil plume generated from the rupture point.

The analysis methods herein are presented as a resource for microbial community ecologists and bioinformaticists when considering the benefits of low-level versus high-level data analysis. We present the benefits of probe-level analysis for finely separating samples into subgroups and the benefits of probe-set level analysis for taxonomic annotation of population dynamics.

PhyloChip G3 data analysis

PhyloChip data are compact and rich in information

According to current standard protocols, 500 ng of PCR product are amplified from one sample, fragmented and hybridized onto a single array. Assuming an average GC content of 54% (based on the entire Greengenes database release August 2012) and approximately 1465 bp amplicons of the 16S rRNA gene, 3.3 + E11 molecules are assayed with one single array, which is at least one hundred times more molecules compared to even the deepest sequencing technologies such as dedicating an entire run of 16 lanes and two flow cells in an Illumina HiSeq 2000 (http://www.illumina.com/systems/hiseq_comparison.ilmn; Table 5.1). As far as the authors are aware, no sequencing facility is operating at this depth for routine 16S rRNA amplicon analysis. Moreover, the raw data of one single PhyloChip uses only 26 Mb of storage, which converts to approximately 8×10^{-5} bytes per molecule assayed (Table 5.1). In comparison to data from platforms that are based on 16S rRNA gene sequencing, microarray data is compact, easy to move across data networks, and rich in information, making the PhyloChip technology suitable for high-resolution microbial community profiling. In the analysis steps presented in the following sections, the data is examined as (1) single probes, (2) probe-pairs, (3) sense and antisense pairs combined as probe-quartets and (4) as larger sets of probes associated with an eOTU (= taxonomically classified set of probes). We demonstrate that the compact microarray data is well-suited to data mining and elucidates microbial community changes at all four resolutions.

Sinfonietta for empirical OTU discovery

Previously, PhyloChip probe florescence from samples collected from the Deepwater Horizon oil spill (Hazen *et al.*, 2010) was compared to approximately 60,000 bacterial and archaeal reference OTUs spanning the entire Greengenes database (DeSantis *et al.*, 2006). The analysis method presented at that time, termed 'PhyCA', was restricted to these pre-defined OTUs and pre-defined sets of probes for each OTU. In the method presented here, termed Sinfonietta, the probes are placed into probe sets *after* the data is collected, coupling the microarray's sensitivity to shifts in abundance with exploring microbial communities beyond a reference database.

In Sinfonietta, the empirical finding of eOTUs is a multi-stage process. The first stage of pixel summarization of the florescent image, background subtraction, noise estimation and array scaling were conducted as previously described (Hazen *et al.*, 2010). Array fluorescence intensity (FI) of each pixel in an image was collected as integer values ranging from 0 to 65,536 providing 2^{16} distinct FI values. The summary of FI for each single probe feature on the array was calculated by ranking the FI of the central 9 of 64 image pixels and using the value of at 75th percentile. Background was defined as the mean feature FI in the least bright 1% of features in each of 25 equally divided sectors of the image. The background was subtracted separately in each sector. Next, all probes on the array were scaled by multiplication with a single factor so that average FI of the probes perfectly matching the non-16S spike control mix were equal.

The Sinfonietta method provides the options to evaluate the probes, probe-pairs, probe-quartets and/or probe-sets (eOTUs). In the least summarized option, values from redundant probes were averaged, and then all values were $\log_2$-transformed to generate the simple probe-level table representing the responses of 994,980 probes across 17 samples. Pairs of probes are two probes with similar but non-identical sequences

Table 5.1 Comparison of data and information storage of different community profiling platforms

Feature	PhyloChip G3	454 GS FLX+	Illumina HiSeq 2000 (paired-end)	Illumina MiSeq (paired-end)
Number of molecules assayed	3.30×10^{11}	1.00×10^{6}	3.00×10^{9}	1.00×10^{7}
Disc space required of raw data in Gigabytes	0.025 GB	0.7 GB	30 GB	10 GB
Disc space per molecule assayed in Bytes	8.26×10^{-5}	7.34×10^{4}	1.07×10^{1}	1.07×10^{3}

which align along ≥23 bases with ≥1 mismatch or gap as determined by blastn (Altschul *et al.*, 1990) (-word_size 8 -dust no -perc_identity 92 -evalue 0.005 -penalty −1). Although all probes can produce minor fluorescence from non-specific hybridization, if a sequence-specific hybridization has occurred the probe complementing the target will be brighter than its mis-matching mate as has been observed in 70% of controlled experiments using pairs (Furusawa *et al.*, 2009). As a general caution, perfect matching probes (PM) were considered positive if they fulfilled the following criteria in comparison to their corresponding mis-matching probes (MM). (A) PM/MM≥1.5, (B) PM-MM≥50**N* and *r*≥0.95, where *N* is the array specific noise, and *r* indicates the response score (Hazen *et al.*, 2010). In total, the FI values for 32,011 pairs passed these thresholds in three or more samples. To invoke greater stringency, the two strands of the double-stranded PCR products were leveraged to remove pairs if both sense pairs and anti-sense pairs (collectively termed a probe-quartet) did not pass the above criteria. Of the 32,011 pairs, 20,891 passed the quartet filter.

Within-sample ranked FIs of PM probes from the 20,891 probe-quartets were then used for empirical probe-set discovery. Individual probes were clustered into probe-sets by correlating the ranked FI across all samples and by taxonomic relatedness. In cases where multiple solutions were possible, higher correlation coefficients were preferred over lower correlation coefficients, taxonomic relatedness at lower taxonomic level was preferred over higher ranks, and sets with higher number of probes were favoured over sets with lower numbers of probes. Probe sets contained at least five probes with an average pair-wise correlation coefficient of ≥0.85. After the probe-sets were assembled, each probe-set for each sample was binary-scored with a 1 if ≥80% of the probe pairs assigned passed in that sample, otherwise a 0 was assigned. In total, 909 probe-sets scored a 1 in at least one of the 17 samples and the eOTUs of each probe set were annotated against Greengenes with the aid of a Naïve Bayesian algorithm applied to the 9-mers contained in all probes of the set. Bootstrap cut-off was set to 80% for all taxonomic levels. Afterwards, the mean ranked FI of all probes in one eOTU was determined for each sample. These values are referred to as HybScores (hybridization scores) and used as abundance data of eOTUs.

Topological data analysis

Topological data analysis (TDA) functions as a geometric approach to identify small scale and large scale patterns within datasets. By understanding the shape of the data, which ultimately results in signal detection, this method is unsupervised in that it requires no initial hypotheses Three key characteristics of topology are essential for making the identification of shapes successful for assessing very subtle signals in complex data sets. These key characteristics are coordinate freeness, deformation invariance and compressed representation of shapes (Lum *et al.*, 2013). For the Deepwater Horizon dataset we employed TDA for identifying outlier samples. Using TDA, the 17 samples were binned into overlapping buckets (in other words, samples can be placed in one or more buckets) according to the Gaussian density and then data points in the buckets were clustered into nodes according to their degree of cosine similarity. The calculated topological network was displayed using nodes and edges (Fig. 5.1), where nodes can contain multiple samples and samples can appear in multiple nodes. Nodes are connected by edges when they have samples in common. Nodes that do not contain any shared elements are classified as singletons.

Considering the response of 994,980 individual probes simultaneously from 17 different arrays (10 designated plume samples, 7 non-plume samples), TDA identified two major networks, one comprising of plume samples and the other of non-plume samples. Sample OV01106 (designated as non-plume sample) forms a separate node (a singleton) and could be considered neither plume nor non-plume representative. Comparing the plume and the non-plume networks derived from probes, pairs or quartet (Fig 5.1A–C), we notice that the two main networks do not share samples and the non-plume cluster has lower Gaussian density values than the plume cluster. A possible interpretation is the non-plume network represents the high inter-sample dissimilarity within the pristine deep-sea microbial community and the inter-sample diversity is

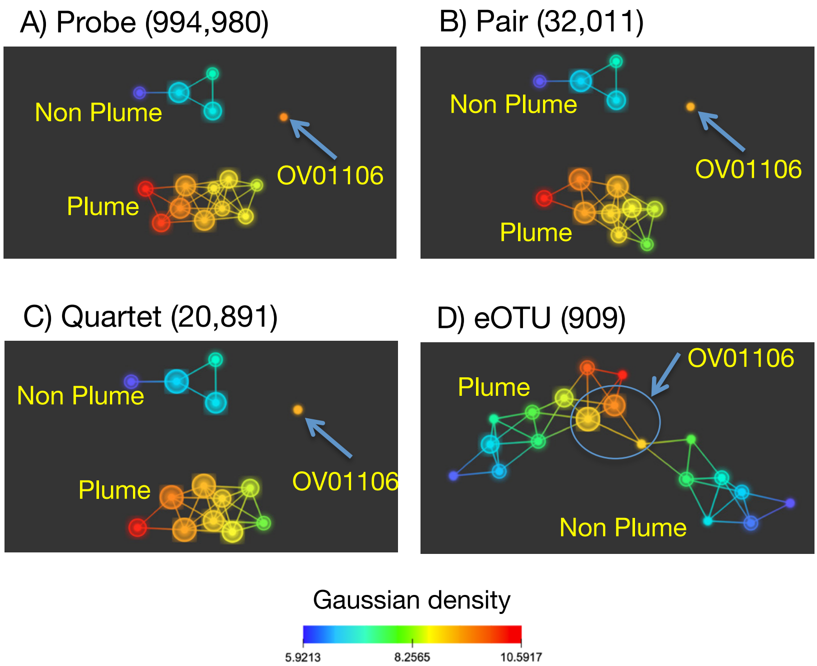

Figure 5.1 Topological network of 17 samples generated using the Iris software (Ayasdi Inc.). Each node (circle) represents a group of samples with high cosine similarity. Node diameter corresponds to the number of samples in each node. An edge (line between nodes) connects any nodes that have samples in common. Node colour indicates Gaussian density of the samples (red is high, blue is low). The four panels represent the topological outcomes using 4 types of input from the same biological samples: (A) all 994,980 unique 25-mer probes, (B) 32,011 responsive probe pairs, (C) 20,981 responsive probe quartets and (D) 909 probe sets (one set for each eOTU). According to these parameters, the relative relatedness of these samples are such that non-plume samples form a strong subcluster and plume samples form another subcluster. There is one sample (arrow, singleton) that does not cluster with either the plume or non-plume samples in panels A, B and C.

attenuated by the plume chemical content, where the relationships between the microbial communities are tighter.

In contrast to probe-, pair- and quartet-level, the analysis at set-level uses the re-construction of empirical OTUs (eOTUs) from clustering probe quartets. Concatenated probe sequences are classified using a Bayesian-algorithm against the Greengenes taxonomy (McDonald *et al.*, 2012), allowing phylogenetic assignments of eOTUs. 909 eOTUs were identified in the entire dataset and their abundances (in units of fluorescence intensity) in each sample were used for TDA. When conducting TDA on probe-level data, pair-level data or quartet-level data, OV01106 was clearly an outlier compared to the other groups, but from the eOTU-level TDA, OV01106 was similar to both plume and non-plume samples. Close examination of the metadata revealed that although sample OV01106 was classified as non-plume based on the fluorescence of the water sample measured immediately after collection, subsequent hydrocarbon analysis revealed that OV01106 contained the highest octadecane concentration of all non-plume samples (Hazen *et al.*, 2010). The slightly increased hydrocarbons may have altered the microbial community, which was detected by the highly sensitive TDA analysis based on probe-, pair- and quartet-level FI. Overall, the observations suggest that the PhyloChip hybridization data contains subtle dissimilarities that can be overlooked when over-summarizing all the probe responses into probe-sets. Regardless, all four levels agreed that only sample OV01106 did not belong unambiguously to one group or the other.

Multivariate statistical analysis of PhyloChip data as revealed by PhyCA-Stats™

Multivariate statistical procedures can be applied to the set-level (eOTUs) based on either their hybridization scores (abundance values) or on binary metrics (presence/absence). These statistics use dissimilarity metrices to calculate the relationship between samples based on the abundance or binary metrics of the entire microbiomes observed. A detailed description of these methods can be found in for example (Kuczynski *et al.*, 2010). In brief, microbial profiles of all samples are inter-compared in a pair-wise fashion to determine a dissimilarity score, resulting in a distance matrix. Two examples used herein are the UniFrac distance metric as published in (Lozupone and Knight, 2005) and the Bray–Curtis Index. While UniFrac utilizes the phylogenetic distance between eOTUs to determine the distance between communities, Bray–Curtis employs a pair-wise normalization by dividing the sum of differences by the sum of all abundances. UniFrac distance measure is either weighted (based on abundance scores) or unweighted (based on binary), whereas the Bray–Curtis Index calculated from a binary data results in the Sørensen Index.

Based on the distance matrix two-dimensional ordination analyses or hierarchal clustering maps in form of dendrograms are popular methods for displaying inter-sample relationships. Two examples of ordination methods used in this book chapter are principal coordinate analysis (PCoA) and non-metric multidimensional scaling (NMDS). PCoA uses the dissimilarity values and NMDS the rank of the dissimilarity values to position samples in relative distance to each other.

Applying these analysis tools to the microarray data set produced during the Deepwater Horizon oil spill (Hazen *et al.*, 2010) provided evidence for microbial community changes of samples inside compared to outside the oil plume. A set of these ordinations is depicted in Figs. 5.2 and 5.3. PCoA and NMDS plots based on abundance dissimilarities clearly demonstrate a separation of the microbiomes of the plume and non-plume samples along NMDS 1 or PCoA 1 axis. These results are in agreement with reference based PhyloChip rOTU analyses published previously (Hazen *et al.*, 2010).

Not all factors influencing changes in the community structure can be grasped from ordination analysis or hierarchal clustering methods. Microbial communities are generally exposed to multiple environmental factors, each of which has a variable degree of influence on community structure. One method to determine if numeric or categorical factors are associated with the observed microbial communities is the Adonis test. Adonis, also called PERMANOVA, utilizes the sample-to-sample distance matrix directly, not a derived ordination or clustering outcome. Low *P*-values across categorical variables would indicate that samples from different categories are generally more dissimilar in their microbiomes than samples from the same category. Low *P*-values for continuous variables, such as age, would indicate the samples from patients that are more similar in age have generally more similar microbiomes. Similar to NMDS and PCoA, the Adonis test can utilize any dissimilarity metric, including UniFrac and Bray–Curtis Index. During the sampling cruise of the *R/V Ocean Veritas* and the *R/V Brooks McCall* after the Deepwater Horizon oil spill, multiple environmental factors were collected. These factors encompassed plume/non-plume, fluorescence, latitude, longitude, sampling depth, acridine orange direct count, phosphate, ammonia nitrogen concentration, dissolved inorganic carbon, total phospholipid fatty acids, phospholipid fatty acids *trans/cis* ratio, phospholipid fatty acids 16:1w5c/16:1w7c ratio, octadecane and docosane concentration. Adonis testing based on weighted UniFrac measures of eOTU data demonstrated a highly significant influence of the oil spill on the microbial community structure (*P*-value 0.001 for plume versus non-plume, Table 5.2); fluorescence, which was used to detect crude oil (Hazen *et al.*, 2010), was also found to be associated with a significant change in the microbial community. Similarly, concentrations of hydrocarbons (octadecane, docosane) in the samples showed a highly significant association with microbial community profiles. Other factors like sampling depth or the total amount of phospholipid-derived fatty acids were also found to be significant but with higher

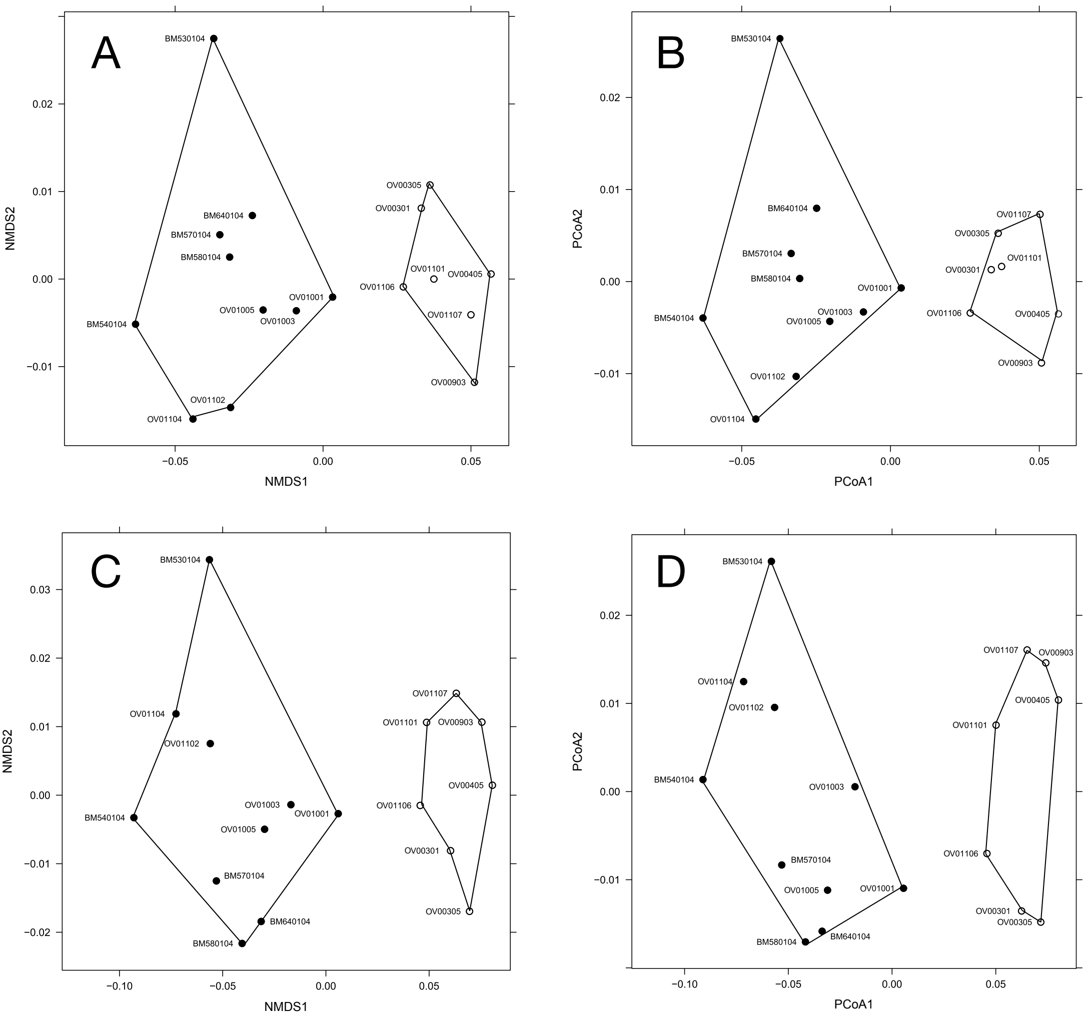

Figure 5.2 Ordination analysis based on abundance scores of eOTUs present in at least one of the samples. All ordination methods show a separation of plume (closed circles) and non-plume (open circles) samples along NMDS1 and PCoA1, respectively. Consequently, different microbial community structures are present in plume and non-plume samples. However, non-plume sample OV01106 places between plume and non-plume groups. (A) NMDS based on Weighted Unifrac distance between samples given abundance of 909 taxa present in at least one sample with stress = 0.0239. (B) PCoA based on Weighted Unifrac distance between samples given abundance of 909 taxa present in at least one sample. Axis 1: 85% of variation explained. Axis 2: 5% of variation explained. (C) NMDS based on Bray–Curtis distance between samples given abundance of 909 taxa present in at least one sample with stress = 0.0321. (D) PCoA based on Bray–Curtis distance between samples given abundance of 909 taxa present in at least one sample. Axis 1: 84% of variation explained. Axis 2: 4% of variation explained.

P-values. Phospholipids are generally a microbial biomarker and are used to estimate microbial community changes via lipidomics (Hazen *et al.*, 2010) or to estimate microbial abundance. An overview of Adonis *P*-values and the multiple factors tested is presented in Table 5.2.

As mentioned above, set-level based analysis also allows also presence/absence calling of eOTUs recorded as the binary variable 1 or 0. Considering binary metrics, the non-plume sample OV01106 was depicted as a possible boundary sample. Due to sample OV01106, incomplete separation of the microbiomes was observed along NMDS1 and PCoA1 axis (Fig.

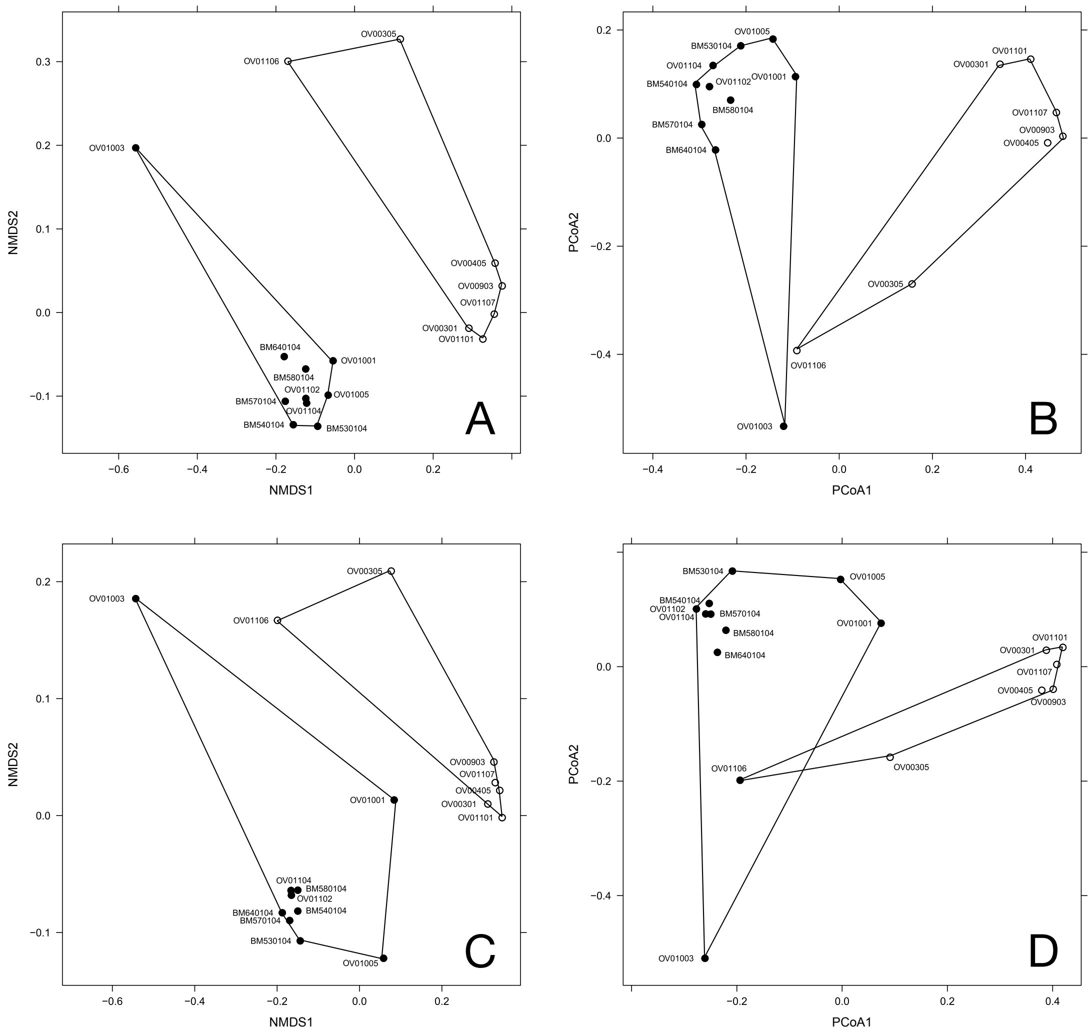

Figure 5.3 Ordination analysis based on binary metrics of eOTUs. In contrast to abundance based ordinations (Fig. 5.2), in three out of four ordination methods plume and non-plume microbiomes are not separated along NMDS1/PCoA1 axis due to the non-plume sample OV01106. This sample is considered an outlier. (A) NMDS based on Unweighted Unifrac distance between samples given presence/absence of 909 taxa present in at least one sample with stress = 0.0381. (B) PCoA based on Unweighted Unifrac distance between samples given presence/absence of 909 taxa present in at least one sample. Axis 1: 43% of variation explained. Axis 2: 14% of variation explained. (C) NMDS based on Bray–Curtis distance between samples given presence/absence of 909 taxa present in at least one sample with stress = 0.0477. (D) PCoA based on Bray–Curtis distance between samples given presence/absence of 909 taxa present in at least one sample. Axis 1: 40% of variation explained. Axis 2: 18% of variation explained.

5.3). This same sample was also categorized as an outlier in TDA (Fig. 5.1) and placed close to plume samples in abundance and binary based ordinations of eOTUs (Figs. 5.2 and 5.3). Presence/absence, or incidence of eOTUs, allows comparison of microbial richness. The number of eOTUs present in plume samples decreased significantly ($P = 0.02$, ANOVA) compared to the number of eOTUs in non-plume samples. This result is in accordance with significant P-values in Adonis testing (Table 5.2) and with the previous rOTU-based analysis (Hazen *et al.*, 2010). In the aforementioned reference, presence/absence of rOTUs and fluorescence microscopy showed a

Table 5.2 Adonis *P*-values based on weighted Unifrac measures of abundance scores of eOTUs. Factors with significant *P*-values (bold) were associated with microbial community changes

Factor	Details	*P*-value
Plume versus non-plume	Sample category determined by fluorescence	**0.001**
Fluorescence	*In situ* fluorescence intensity	**0.001**
Latitude	–	0.065
Longitude	–	0.593
Depth	Sampling depth	**0.037**
AODC	Acridine orange direct count	0.096
Phosphate	Total phosphate	0.419
Ammonia (N)	Total ammonia nitrogen	0.115
d13C_DIC	Dissolved inorganic carbon ^{13}C isotope value	0.222
Total_PLFA	Total phospholipid fatty acids	**0.041**
PLFA_trans_cis	PLFA_trans/cis (ratio)	0.224
X16.1w5c_16.1w7c	PLFA 16:1w5c/16:1w7c (ratio)	**0.009**
Octadecane	Normalized octadecane concentration	**0.002**
Docosane	Normalized docosane concentration	**0.001**

very restricted microbial community profile in samples of high hydrocarbon content.

Identification of taxa enriched in plume samples

With its 1.1 million probes tracking more than 3.0×10^{11} 16S rRNA gene molecules per sample, PhyloChip G3 technology is highly suited for multivariate statistical analysis to display differences between samples. Furthermore, univariate statistics, like a Welch test, can be applied to each individual probe/pair/quartet/set across samples allowing the identification of taxa that significantly increase or decrease in one sample category compared to other. Since TDA identified OV01106 as a boundary sample and both NMDS and PCoA supported that finding, OV01106 was removed from univariate analysis. Table 5.3 provides an overview of the percentage of passing taxa at each of four resolution levels and at each of seven taxonomic levels. Two different corrections for multiple testing were applied and, with the most stringent correction (Benjamini–Hochberg), on average 81.9% of the taxa passed the Welch test. The lowest percentage was retrieved for species level considering probes and the highest percentage for phylum level and eOTUs. In general, with increasing taxonomic summarization (species → genus → family → order → class → phylum) and with increasing probe summarization (probe → pair → quartet → set) the greater the percentage of taxa passed the univariate significance test. Using phylogenetic affiliations, eOTU analysis provides relative abundances for each taxon identified in a sample set. Employing a parametric Welch test, 104 eOTUs were identified to have significantly increased within-plume samples compared to non-plume samples. In contrast, approximately six times as many taxa decreased in abundance within the plume, concordant with the significant richness decrease noted above. Taxa that were significantly lower in abundance in oil-contaminated samples included many Archaea. Archaeal probe sets complementary to *Nitrosopumilus*, known for ammonia-oxidation in the ocean, or to members of the Thermoplasmata marine group II were observed to produce significantly lower FI in plume samples demonstrating that the microbial community of non-contaminated seawater was strongly altered by the spill. A selection of the 25 taxa with the most significant abundance increase and an additional 25 taxa with the most significant decrease is depicted in Fig. 5.4 as a heatmap. Dendrograms were calculated by grouping eOTUs with similar abundance changes across samples and applying

Table 5.3 Proportion of the community with significant changes in abundance observed between non-plume and plume samples. Data was summarized by probe-, pair-, quartet- and set/eOTU-level considering aggregated fluorescence intensities (FI) at various taxonomic levels. For instance, 23,680 responsive pairs belonged to 721 different families. Family sums of FI were compared and approximately 81% of the families exhibited a significant abundance change between plume and non-plume samples depending on the FDR procedure implemented. Outlier sample OV01106 was excluded from all comparisons

Data Resolution	Rank	Count of taxa considered	Per cent of taxa passing Welch test[1]	Per cent of taxa passing Welch test after permutational FDR penalty[2]	Per cent of taxa passing Welch test after BH FDR penalty[3]
Probe-level	Each probe	994,980	NA	NA	NA
	Species	4187	73.4%	73.4%	42.8%
	Genus	2112	82.0%	81.2%	80.8%
	Family	833	87.9%	87.0%	87.3%
	Order	464	91.2%	90.7%	91.2%
	Class	253	96.0%	96.0%	88.1%
	Phylum	93	97.8%	97.8%	97.8%
Pair-level	Each pair	23,680	72.8%	71.0%	69.5%
	Species	2036	76.0%	75.2%	73.8%
	Genus	1493	79.0%	78.2%	76.9%
	Family	721	82.4%	82.1%	80.6%
	Order	410	83.9%	83.2%	82.7%
	Class	232	87.9%	87.1%	87.1%
	Phylum	88	96.6%	96.6%	95.5%
Quartet-level	Each quartet	20,891	53.5%	53.0%	51.2%
	speCies	1453	75.5%	74.3%	73.5%
	Genus	1146	78.9%	77.9%	76.5%
	Family	631	81.1%	80.3%	79.4%
	Order	376	83.2%	82.4%	82.2%
	Class	222	86.5%	85.6%	86.5%
	Phylum	85	92.9%	92.9%	92.9%
Set-level/ eOTU	eOTU HybScore	910	82.1%	76.9%	81.0%
	Species	259	82.6%	80.7%	82.2%
	Genus	244	84.0%	82.4%	83.6%
	Family	207	87.0%	86.0%	86.5%
	Order	140	91.4%	90.7%	90.0%
	Class	102	94.1%	94.1%	94.1%
	Phylum	61	98.4%	98.4%	98.4%

[1] Welch test at $P<0.05$.

[2] Number of taxa passing the Welch test at $P<0.05$ after permutation test for false discovery. Permutation test defined as ≥20 randomizations of taxa-by-sample table followed by Welch test ($P<0.05$). The median number of taxa passing the permutation test is subtracted from the Welch test on the non-permuted data.

[3] Number of taxa passing the Welch test at $P<0.05$ after applying a Benjamini–Hochberg correction at $q<0.05$ for false discovery correction. NA, not calculated.

a pair-wise Euclidean distance measure. A Bray–Curtis Index-based dendrogram separated the plume and non-plume samples into two different clusters. eOTUs that significantly increased in relative abundance in plume samples included taxa classified as Oceanospirillales, an order identified to be responsible for the significant hydrocarbon degradation during the Deepwater Horizon oil

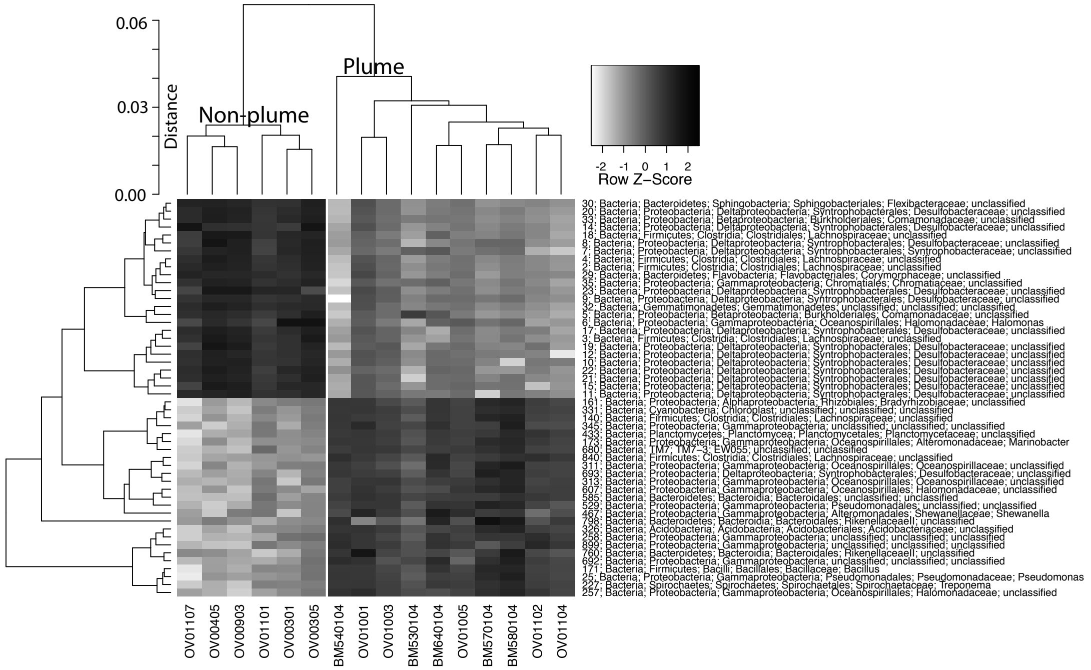

Figure 5.4 Heatmap showing the 25 taxa of each sample category (plume and non-plume) with the most significant changes in abundance scores. Grey intensities are based on normalized abundances (z-scores) across eOTU trajectories of all samples excluding the outlier sample OV01106 (see Fig. 5.1 and 5.3). The dendrogram at left represents hierarchical clustering of the pair-wise Euclidian distances among eOTU trajectories. Sample distances are calculated based on a Bray–Curtis Index of the 50 taxa displayed and presented via hierarchal clustering, which separates non-plume from plume samples.

spill (Hazen *et al.*, 2010; Mason *et al.*, 2012). The G3 PhyloChip identified additional microbial taxa enriched in plume samples (Dubinsky *et al.*, 2013; Hazen *et al.*, 2010) besides the Oceanospirillales. As commonly accepted, a complex interplay of microbial pathways from a diverse set of microorganisms is necessary to degrade crude oil, which may be reflected by these results. Indeed, other taxa were found to be enriched in oil plume samples by multiple NGS methods applied over the course of this bioremediation project including metagenomics, metatranscriptomics and pyrotag-sequencing of 16S rRNA genes (Mason *et al.*, 2012). Nevertheless, all these methods agree that Oceanospirillales microorganisms were enriched in plume samples, in accordance with microscopic analysis and 16S rRNA gene cloning (Hazen *et al.*, 2010). Table 5.4 provides an overview of microbial families reported as plume-enriched by multiple investigations of the Deepwater Horizon spill.

Discussion and future trends

'Everything is everywhere but the environment selects' (Baas-Becking, 1934) is a very common statement in environmental microbiology and is often used to explain microbial community observations (de Wit and Bouvier, 2006). However, a recent study showed that a persistent microbial community with modulation of abundances of certain taxa explains community dynamics observed in a certain environment (Caporaso *et al.*, 2012c). This observation could only be achieved by assaying 100-fold more 16S rRNA gene molecules than previous attempts with typical sequencing depth (Gilbert *et al.*, 2009, 2012). Therefore, past molecular microbial ecology observations based solely on incidence (bacterial taxa 'present' in some samples but 'absent' in others) should be soon re-evaluated. Consequently, the concept 'everything is everywhere but the environment selects' may need to be refined to 'everything is everywhere but the environment

Table 5.4 Families detected by various methods as enriched in oil plume samples compared to samples outside the plume. eOTU and quartet-level analysis are from this book chapter. The table includes the families of all rOTUs (Hazen *et al.*, 2010) described as enriched in plume samples plus all taxa found to have higher proportion in 454-pyrotag sequences in plume samples compared to the reference sample (Mason *et al.*, 2012). The list of families was extended by adding representatives of the 20 most significantly enriched eOTUs applying a Welch test. eOTU and quartet-level analysed excluding outlier OV0110601.

Method	PhyloChip: eOTU	PhyloChip: quartet level	PhyloChip: rOTU	454-pyrotag	Metagenomics	Metatranscriptomics
Taxonomy	Welch test	Welch test	Hazen et al. (2010)	Mason et al. (2012)	Mason et al. (2012)	Mason et al. (2012)
Aeromonadales; Aeromonadaceae	ND	↑	↑	↑	↑	✓
Alteromonadales; Colwelliaceae	≈	↑	↑	↓	ND	ND
Alteromonadales; Pseudoalteromonadaceae	↑	↑	↑	ND	ND	ND
Alteromonadales; Shewanellaceae	↑	↑	↑	ND	ND	ND
Alteromonadales; unclassified	↑	↑	↑	↑	↑	✓
Bacteroidales; Rikenellaceae	↑	↑	NA	NA	NA	NA
Chromatiales; Ectothiorhodospiraceae	ND	↓	NA	↑	ND	ND
Clostridiales; Lachnospiraceae	↑	↑	NA	NA	↓	✓
Desulfobacterales; Desulfobacteraceae	↑	↑	NA	NA	NA	NA
Oceanospirillales; Halomonadaceae	↑	↑	↑	ND	↑	✓
Oceanospirillales; Marinospirillum	ND	ND	↑	ND	ND	ND
Oceanospirillales; unclassified	↑	↑	↑	↑	↑	✓
Pseudomonadales; Moraxellaceae	ND	↑	↑	ND	ND	✓
Pseudomonadales; Pseudomonadaceae	↑	↑	↑	ND	↑	✓
Spirochaetales; Spirochaetaceae	↑	↑	NA	NA	NA	NA

ND, not detected; NA, not available in reference; ↑, increased in plume; ≈, detected but no (significant) increase was given; ☑, detected (for metatranscriptomics as there was no non-plume sample for comparison); ↓, decreased in plume.

selects the abundance'. Moreover, this important finding demonstrates the need to apply molecular methods that can track abundance changes in both the majority and minority populations in order to provide a more complete assessment of microbial population dynamics in the environmental or clinical samples investigated.

The G3 PhyloChip is currently the high-throughput tool that can assay the most number of 16S rRNA gene molecules per sample of all platforms available. In the current study, we explored the resolution of PhyloChip at various stages from annotation-free analysis at probe-level to pair- to quartet- to set-level, where eOTUs are taxonomically classified. Exploring less summarized hybridization FI resulted in higher resolution of microbial community dissimilarities using TDA. With decreasing resolution the ability to capture outliers is diminished (e.g. sample OV01106 in this study). However, as a general caution, relying on a single probe, probe pair or probe quartet has risks since in 30% of paired events in controlled experiments, the mismatch has been shown to out-fluoresce the perfect match probe (Furusawa *et al.*, 2009). Accordingly, relying on multiple pairs, or even better multiple quartets, allows greater confidence in the sequence-specific detection event. Another advantage to the probe sets over individual probes is the increase in classification confidence derived with a larger number of probes aiding in the taxonomic identification of the populations in flux.

In general, annotation of 16S rRNA genes poses a challenge to the scientific community. Associating a short NGS sequence or set of probes to a named taxonomic node is not always reliable. Not every strain or even species has a unique 16S rRNA gene sequence and can hence be accurately identified (e.g. some strains of *Salmonella*, *Shigella* and *Escherichia* have identical 16S rRNA gene sequences). The widely applied method to classify 16S rRNA gene sequences in a high-throughput manner is the application of a Naïve Bayesian algorithm as demonstrated by (Wang *et al.*, 2007). However, the breadth of the reference database affects the perceived confidence of taxonomic classification. Specifically, the less diverse the reference database, the more confident the Bayesian outcomes can appear. By definition, the lower the number of taxonomic bins the greater the chance that a sequence will fall into one bin. More importantly, the sequence length and the region of the 16S rRNA gene (hypervariable versus conserved) are crucial for accurate classification. The shorter the sequence, the less data for matching and species level classification is not normally expected from reads 400 bps and shorter, which are currently produced by next generation sequencing platforms (Wang *et al.*, 2007).

High-throughput sequencing technologies have been developed to analyse microbial communities, especially for understanding the microbial diversity via sequencing of 16S rRNA gene amplicons (Caporaso *et al.*, 2012a; Sogin *et al.*, 2006), and have provided many novel insights (Bartram *et al.*, 2011; Degnan and Ochman, 2012; Deng *et al.*, 2012; Engelbrektson *et al.*, 2010; He *et al.*, 2010; Mason *et al.*, 2012; Yatsunenko *et al.*, 2012; Zhou *et al.*, 2012). However, compared to microarray-based technologies (e.g. PhyloChip), some disadvantages or limitations for sequencing technologies remain. First, there are a variety of sequencing errors and chimeric sequences although they may be difficult to identify (Edgar, 2013; Huse *et al.*, 2007; Kunin *et al.*, 2010; Pinto and Raskin, 2012; Schloss *et al.*, 2011). For instance, based on mock community samples, the average error rate of 16S rRNA by pyrosequencing was 0.6%, and chimera rate was 8% of the total number of reads (Schloss *et al.*, 2011), although those errors could be reduced or minimized with appropriate sequence analysis pipelines which remove a large portion of the data (Edgar, 2013; Schloss *et al.*, 2011). Sequence errors and chimeras may generate numerous spurious OTUs, which can inflate the perceived microbial diversity (Edgar, 2013; Kunin *et al.*, 2010; Roh *et al.*, 2010). Although these spurious OTUs may resemble new species, they have led to an intensive debate regarding how much of the 'rare biosphere' is due to sequencing artefacts (Schloss *et al.*, 2011; Sogin *et al.*, 2006). Second, since very few DNA molecules are actually sequenced, under-sampling occurs resulting in low reproducibility and quantitation (Zhou *et al.*, 2008, 2011, 2013). The effect of random sampling processes on technical reproducibility was explicitly demonstrated by recent mathematical modelling and simulations (Zhou *et al.*, 2013). Also, sequencing approaches tend to sequence dominant species, which may lead to skewed results, especially for rare species due to repeatedly sampling those dominant populations. In addition, 16S rRNA gene sequencing data analysis may be more difficult or still in development (McMurdie and Holmes, 2013), while microarray data analysis approaches have been largely developed and are widely used.

Due to the unique features and advantages and disadvantages provided by both microarray-based and sequencing-based technologies, it is preferred that they are complementarily used for microbial community analysis in order to address fundamental questions in microbial ecology and environmental biology.

Future trends tend to improve not only laboratory-based technologies for producing 16S rRNA gene amplicons but also bioinformatics data mining and the development of novel data analysis tools. Traditionally, 16S rRNA gene amplicons are generated from total DNA extracts from an environmental sample. In contrast to RNA, DNA is more stable and can remain in an environment even after its organism is no longer alive. Thus, DNA-based methods may not accurately reflect the living/active members of a community. One solution for this shortcoming is the extraction of rRNA from the environment and direct use for hybridization or subsequent reverse-transcription to cDNA (DeAngelis *et al.*, 2011). The community profile retrieved by these rRNA-based methods can be significantly different from the community tracked by total DNA-based methods (DeAngelis and Firestone, 2012; DeAngelis *et al.*, 2011). One bottleneck of rRNA-based methods is the amount of rRNA that can be extracted from an environmental sample. This is easily possible for soil microbial communities but hard for low-biomass environments or studies with limited access to biomass. A second alternative to total DNA-based methods is the usage of propidium monoazide (PMA), which has successfully been applied to low-biomass clean room environments, where rRNA extraction fails (Vaishampayan *et al.*, 2013). PMA applied to an environmental sample intercalates into free-DNA molecules (DNA of dead organisms) but cannot enter cells with an intact cell membrane (Nocker *et al.*, 2007). After photoactivation PMA covalently binds to DNA and hampers the binding of the DNA polymerase during PCR amplification of 16S rRNA genes. Similar to the rRNA-based method, PMA treatment significantly alters the microbial community structure of environmental samples, if DNA of dead organisms is present in high amounts (Vaishampayan *et al.*, 2013).

One bioinformatics technique for the analysis of G3 PhyloChip data is the pre-selection of specific probe sequences uniquely complementing 16S rRNA genes of a specific strain of interest. For instance, a strain could be selected which was hypothesized to be enriched in healthy versus diseased samples. After identification of a certain probe for a specific strain, the relative abundance change of the strain can be tracked in any subsequently generated data set. Using this approach, a single probe tracking a specific Archaeon in the environment has been recently leveraged (Probst *et al.*, 2013). We envisage simplistic software applications to read the G3 PhyloChip or future generations of the devise and report on pre-defined probe features useful for dedicated purposes such as waste-water treatment plant monitoring or clinical diagnostics.

One major bottleneck of microarray-based community profiling is the identification of potentially new species, which is theoretically possible with 1.5 kb full-length 16S rRNA gene sequencing technologies but can also lead to the artificial 'rare biosphere' generated by sequencing errors. The new Sinfonietta analysis to generate eOTUs described herein, can be used to track potentially novel species. In a recent publication, a novel subsurface Archaeon could be tracked by the G3 PhyloChip, although its 16S rRNA gene was not included in the original design database due to its novelty (Probst *et al.*, submitted). The authors do not recommend that a novel species should be proposed based solely on PhyloChip or NGS data, however they can be included in the analysis of microbiome dynamics.

In the near future, one major life science research focus will continue to be in comparative analysis of human gut microbiomes in healthy, diseased and treated states. We predict that metatranscriptomics will be essential in discovering the method of action by which microbes can cause or prevent disease. But since taxonomic profiling has been the standard first step in disease profiling, and 16S rRNA gene amplification is relatively simple to apply, we expect that 16S rRNA gene technologies will continue to be employed in the coming years. The past literature can be summarized into four guiding principals for the clinical molecular microbial ecologist: (1) taxon abundance, not incidence, is what differentiates healthy and diseased patient groups in microbiome-associated conditions; (2) tens of billions of DNA fragments are generated from each bio-specimen; (3) precise estimates of per cent composition of each taxa in a community are elusive; and (4) changes in both majority and minority populations may contribute to gut health. Thus, correspondingly, the data collection and analysis methods should have these four properties: (1) ability to gather reproducible measurements of taxon abundance; (2) ability for tens of billions of DNA fragments to affect whatever sensors are used in a routine assay; (3) ability to detect taxon abundance *changes* across health groups; and (4) not inclined to majority populations obscuring minority ones. The G3 PhyloChip fulfils all four.

Conclusions

PhyloChip data analysis can be executed at several stages providing various resolutions to the user. Observations from analysis based on individual probes, pairs of probes, or quartets of probes enabled the detection of outliers via TDA compared with analysis based on probe-sets for empirically derived eOTUs. The advantages of reducing data to empirical OTUs are: (1) the greater resolution on taxonomic annotation of the microorganisms tracked compared to single probes; (2) the non-reliance on pre-defined sets of probes based on reference OTUs from databases; and (3) the potential to track population shifts from microbes previously not included in such databases (Probst *et al.*, 2013).

The variety of data analysis tools presented here agree well with each other concerning the over-arching biological conclusions that can be extracted, despite different mathematical fundamentals. Topological data analysis (TDA) identified two major clusters of samples differentiating the microbiomes of oil-contaminated samples from non-contaminated samples. The same observations were made using principal coordinate analysis and non-metric multidimensional scaling based on eOTUs. Considering the initial results of the Deepwater Horizon oil spill presented in Hazen *et al.* (2010), the eOTU analysis and the TDA herein agree well with reference-based OTU (rOTU) sample grouping and with the lipidomics presented in the aforementioned reference. For these reasons, empirical eOTU discovery is recommended as a starting point for G3 PhyloChip data analysis, and quartet-level processing is recommended for advanced outlier detection and developing novel applications and prototype diagnostics on the PhyloChip platform.

Although next-generation sequencing platforms are continuously improving and may one day inexpensively produce trillions of sequences as long as the entire 16S rRNA gene with high accuracy for the full length of the entire read, array-based methods are mature, reproducible and sensitive to changes in populations when comparing microbiomes. Consequently, any detailed microbiome investigation where population dynamics are hypothesized would benefit from 16S rRNA gene amplification followed by PhyloChip hybridization and analysis.

Web resources

Raw PhyloChip G3 data from Hazen *et al.* (2010). (and rOTU analysis) can be found here: http://greengenes.lbl.gov/Download/Microarray_Data/Hazen_2010_Science.tgz

Further information can be requested from the corresponding author.

PhyloChip G3 assay and data analysis are commercially available at Second Genome Inc.:

(http://www.secondgenome.com/)

Topological data analysis is commercially available at Ayasdi Inc.:

(http://www.ayasdi.com/)

Resources for phylogenetic analysis of 16S rRNA genes and corresponding databases are provided at the Greengenes website hosted by Second Genome Inc.: (http://greengenes.secondgenome.com/)

References

Alon, S., Vigneault, F., Eminaga, S., Christodoulou, D.C., Seidman, J.G., Church, G.M., and Eisenberg, E. (2011). Barcoding bias in high-throughput multiplex sequencing of miRNA. Genome Res. *21*, 1506–1511.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410.

Amann, R.I., Ludwig, W., and Schleifer, K.H. (1995). Phylogenetic identification and *in situ* detection of individual microbial cells without cultivation. Microbiol. Rev. *59*, 143–169.

Baas-Becking, L.G.M. (1934). Geobiologie of inleiding tot de milieukunde (Den Haag, the Netherlands).

Baelum, J., Borglin, S., Chakraborty, R., Fortney, J.L., Lamendella, R., Mason, O.U., Auer, M., Zemla, M., Bill, M., Conrad, M.E., *et al.* (2012). Deep-sea bacteria enriched by oil and dispersant from the Deepwater Horizon spill. Environ. Microbiol. *14*, 2405–2416.

Bartram, A.K., Lynch, M.D.J., Stearns, J.C., Moreno-Hagelsieb, G., and Neufeld, J.D. (2011). Generation of multimillion-sequence 16S rRNA gene libraries from complex microbial communities by assembling paired-end Illumina reads. Appl. Environ. Microbiol. *77*, 3846–3852.

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

Briggs, B.R., Brodie, E.L., Tom, L.M., Dong, H., Jiang, H., Huang, Q., Wang, S., Hou, W., Wu, G., Huang, L., *et al.* (2013). Seasonal patterns in microbial communities inhabiting the hot springs of Tengchong, Yunnan Province, China. Environ. Microbiol. (in press, DOI:10.1111/1462–2920.12311)

Brodie, E.L., DeSantis, T.Z., Parker, J.P., Zubietta, I.X., Piceno, Y.M., and Andersen, G.L. (2007). Urban aerosols harbor diverse and dynamic bacterial populations. Proc. Natl. Acad. Sci. USA *104*, 299–304.

Camilli, R., Reddy, C.M., Yoerger, D.R., Van Mooy, B.A., Jakuba, M.V., Kinsey, J.C., McIntyre, C.P., Sylva, S.P., and Maloney, J.V. (2010). Tracking hydrocarbon plume transport and biodegradation at Deepwater Horizon. Science *330*, 201–204.

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

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

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

Castelle, C.J., Hug, L.A., Wrighton, K.C., Thomas, B.C., Williams, K.H., Wu, D., Tringe, S.G., Singer, S.W., Eisen, J.A., and Banfield, J.F. (2013). Extraordinary phylogenetic diversity and metabolic versatility in aquifer sediment. Nat. Commun. *4*, 2120.

Chivian, D., Brodie, E.L., Alm, E.J., Culley, D.E., Dehal, P.S., DeSantis, T.Z., Gihring, T.M., Lapidus, A., Lin, L.H., Lowry, S.R., *et al.* (2008). Environmental genomics reveals a single-species ecosystem deep within Earth. Science *322*, 275–278.

Cho, I., and Blaser, M.J. (2012). The human microbiome: at the interface of health and disease. Nat. Rev. Genet. *13*, 260–270.

Colwell, R.R. (1997). Microbial diversity: the importance of exploration and conservation. J. Ind. Microbiol. Biotechnol. *18*, 302–307.

HMP Consortium (2012a). A framework for human microbiome research. Nature *486*, 215–221.

HMP Consortium (2012b). Structure, function and diversity of the healthy human microbiome. Nature *486*, 207–214.

Cooper, M., La Duc, M.T., Probst, A., Vaishampayan, P., Stam, C., Benardini, J.N., Piceno, Y.M., Andersen, G.L., and Venkateswaran, K. (2011). Comparison of innovative molecular approaches and standard spore assays for assessment of surface cleanliness. Appl. Environ. Microbiol. *77*, 5438–5444.

DeAngelis, K.M., and Firestone, M.K. (2012). Phylogenetic clustering of soil microbial communities by 16S rRNA but not 16S rRNA genes. Appl. Environ. Microbiol. *78*, 2459–2461.

DeAngelis, K.M., Wu, C.H., Beller, H.R., Brodie, E.L., Chakraborty, R., DeSantis, T.Z., Fortney, J.L., Hazen, T.C., Osman, S.R., Singer, M.E., *et al.* (2011). PCR amplification-independent methods for detection of microbial communities by the high-density microarray PhyloChip. Appl. Environ. Microbiol. *77*, 6313–6322.

Degnan, P.H., and Ochman, H. (2012). Illumina-based analysis of microbial community diversity. ISME J. *6*, 183–194.

Deng, Y., He, Z., Xu, M., Qin, Y., Van Nostrand, J.D., Wu, L., Roe, B.A., Wiley, G., Hobbie, S.E., Reich, P.B., *et al.* (2012). Elevated Carbon Dioxide Alters the Structure of Soil Microbial Communities. Appl. Environ. Microbiol. *78*, 2991–2995.

DeSantis, T.Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E.L., Keller, K., Huber, T., Dalevi, D., Hu, P., and Andersen, G.L. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl. Environ. Microbiol. *72*, 5069–5072.

DeSantis, T.Z., Brodie, E.L., Moberg, J.P., Zubieta, I.X., Piceno, Y.M., and Andersen, G.L. (2007). High-density universal 16S rRNA microarray analysis reveals broader diversity than typical clone library when sampling the environment. Microb. Ecol. *53*, 371–383.

Ding, G.C., Piceno, Y.M., Heuer, H., Weinert, N., Dohrmann, A.B., Carrillo, A., Andersen, G.L., Castellanos, T., Tebbe, C.C., and Smalla, K. (2013). Changes of soil bacterial diversity as a consequence of agricultural land use in a semi-arid ecosystem. PLoS One *8*, e59497.

Dubinsky, E.A., Esmaili, L., Hulls, J.R., Cao, Y., Griffith, J.F., and Andersen, G.L. (2012). Application of phylogenetic microarray analysis to discriminate sources of fecal pollution. Environ. Sci. Technol. *46*, 4340–4347.

Dubinsky, E.A., Conrad, M.E., Chakraborty, R., Bill, M., Borglin, S.E., Hollibaugh, J.T., Mason, O.U., Piceno, Y.M., Reid, F.C., Stringfellow, W.T., *et al.* (2013). Succession of hydrocarbon-degrading bacteria in the aftermath of the deepwater horizon oil spill in the gulf of Mexico. Environ. Sci. Technol. *47*, 10860–10867.

Edgar, R.C. (2013). UPARSE: highly accurate OTU sequences from microbial amplicon reads. Nat. Meth. *10*, 996–998.

Edgar, R.C., Haas, B.J., Clemente, J.C., Quince, C., and Knight, R. (2011). UCHIME improves sensitivity and speed of chimera detection. Bioinformatics *27*, 2194–2200.

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

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

Frantz, A.L., Rogier, E.W., Weber, C.R., Shen, L., Cohen, D.A., Fenton, L.A., Bruno, M.E., and Kaetzel, C.S. (2012). Targeted deletion of MyD88 in intestinal epithelial cells results in compromised antibacterial immunity associated with downregulation of polymeric immunoglobulin receptor, mucin-2, and antibacterial peptides. Mucosal. Immunol. *5*, 501–512.

Furusawa, C., Ono, N., Suzuki, S., Agata, T., Shimizu, H., and Yomo, T. (2009). Model-based analysis of non-specific binding for background correction of high-density oligonucleotide microarrays. Bioinformatics *25*, 36–41.

Gilbert, J.A., Field, D., Swift, P., Newbold, L., Oliver, A., Smyth, T., Somerfield, P.J., Huse, S., and Joint, I. (2009). The seasonal structure of microbial communities in the Western English Channel. Environ. Microbiol. *11*, 3132–3139.

Gilbert, J.A., Meyer, F., Antonopoulos, D., Balaji, P., Brown, C.T., Brown, C.T., Desai, N., Eisen, J.A., Evers, D., Field, D., *et al.* (2010a). Meeting report: the terabase metagenomics workshop and the vision of an Earth microbiome project. Stand. Genomic Sci. *3*, 243–248.

Gilbert, J.A., Meyer, F., Jansson, J., Gordon, J., Pace, N., Tiedje, J., Ley, R., Fierer, N., Field, D., Kyrpides, N., *et al.* (2010b). The Earth Microbiome Project: Meeting report of the '1 EMP meeting on sample selection and acquisition' at Argonne National Laboratory October 6 2010. Stand. Genomic Sci. *3*, 249–253.

Gilbert, J.A., Steele, J.A., Caporaso, J.G., Steinbruck, L., Reeder, J., Temperton, B., Huse, S., McHardy, A.C., Knight, R., Joint, I., *et al.* (2012). Defining seasonal marine microbial community dynamics. ISME J. *6*, 298–308.

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

Haegeman, B., Hamelin, J., Moriarty, J., Neal, P., Dushoff, J., and Weitz, J.S. (2013). Robust estimation of microbial diversity in theory and in practice. ISME J. *7*, 1092–1101.

Hayden, H.L., Mele, P.M., Bougoure, D.S., Allan, C.Y., Norng, S., Piceno, Y.M., Brodie, E.L., DeSantis, T.Z., Andersen, G.L., Williams, A.L., *et al.* (2012). Changes in the microbial community structure of bacteria, archaea and fungi in response to elevated CO(2) and warming in an Australian native grassland soil. Environ. Microbiol. *14*, 3081–3096.

Hazen, T.C., Dubinsky, E.A., DeSantis, T.Z., Andersen, G.L., Piceno, Y.M., Singh, N., Jansson, J.K., Probst, A., Borglin, S.E., Fortney, J.L., *et al.* (2010). Deep-sea oil plume enriches indigenous oil-degrading bacteria. Science *330*, 204–208.

He, Z., Xu, M., Deng, Y., Kang, S., Kellogg, L., Wu, L., Van Nostrand, J.D., Hobbie, S.E., Reich, P.B., and Zhou, J. (2010). Metagenomic analysis reveals a marked divergence in the structure of belowground microbial communities at elevated CO_2. Ecol. Lett. *13*, 564–575.

He, Z., Piceno, Y., Deng, Y., Xu, M., Lu, Z., DeSantis, T., Andersen, G., Hobbie, S.E., Reich, P.B., and Zhou, J. (2012). The phylogenetic composition and structure of soil microbial communities shifts in response to elevated carbon dioxide. ISME J. *6*, 259–272.

Huse, S.M., Huber, J.A., Morrison, H.G., Sogin, M.L., and Welch, D.M. (2007). Accuracy and quality of massively parallel DNA pyrosequencing. Genome Biol. *8*, R143.

Jumpstart HMP Consortium (2012). Evaluation of 16S rDNA-based community profiling for human microbiome research. PLoS One 7, e39315.

Kellogg, C.A., Piceno, Y.M., Tom, L.M., DeSantis, T.Z., Zawada, D.G., and Andersen, G.L. (2012). PhyloChip microarray comparison of sampling methods used for coral microbial ecology. J. Microbiol. Meth. *88*, 103–109.

Kellogg, C.A., Piceno, Y.M., Tom, L.M., DeSantis, T.Z., Gray, M.A., Zawada, D.G., and Andersen, G.L. (2013). comparing bacterial community composition between healthy and white plague-like disease states in *Orbicella annularis* using PhyloChip G3 microarrays. PLoS One *8*, e79801.

Kessler, J.D., Valentine, D.L., Redmond, M.C., Du, M., Chan, E.W., Mendes, S.D., Quiroz, E.W., Villanueva, C.J., Shusta, S.S., Werra, L.M., *et al.* (2011). A persistent oxygen anomaly reveals the fate of spilled methane in the deep Gulf of Mexico. Science *331*, 312–315.

Kuczynski, J., Liu, Z., Lozupone, C., McDonald, D., Fierer, N., and Knight, R. (2010). Microbial community resemblance methods differ in their ability to detect biologically relevant patterns. Nat. Meth. *7*, 813–819.

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.

Lam, V., Moulder, J.E., Salzman, N.H., Dubinsky, E.A., Andersen, G.L., and Baker, J.E. (2012). Intestinal microbiota as novel biomarkers of prior radiation exposure. Radiat. Res. *177*, 573–583.

Lin, L.H., Wang, P.L., Rumble, D., Lippmann-Pipke, J., Boice, E., Pratt, L.M., Sherwood Lollar, B., Brodie, E.L., Hazen, T.C., Andersen, G.L., *et al.* (2006). Long-term sustainability of a high-energy, low-diversity crustal biome. Science *314*, 479–482.

Liu, Z., DeSantis, T.Z., Andersen, G.L., and Knight, R. (2008). Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers. Nucleic Acids Res. *36*, e120.

Lozupone, C., and Knight, R. (2005). UniFrac: a new phylogenetic method for comparing microbial communities. Appl. Environ. Microbiol. *71*, 8228–8235.

Lum, P.Y., Singh, G., Lehman, A., Ishkanov, T., Vejdemo-Johansson, M., Alagappan, M., Carlsson, J., and Carlsson, G. (2013). Extracting insights from the shape of complex data using topology. Sci. Rep. 3, 1236.

McDonald, D., Price, M.N., Goodrich, J., Nawrocki, E.P., DeSantis, T.Z., Probst, A., Andersen, G.L., Knight, R., and Hugenholtz, P. (2012). An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME J. *6*, 610–618.

McMurdie, P.J., and Holmes, S. (2013). Waste Not, Want Not: Why Rarefying Microbiome Data is Inadmissible. arXiv preprint arXiv:13100424.

Maldonado-Contreras, A., Goldfarb, K.C., Godoy-Vitorino, F., Karaoz, U., Contreras, M., Blaser, M.J., Brodie, E.L., and Dominguez-Bello, M.G. (2011). Structure of the human gastric bacterial community in relation to Helicobacter pylori status. ISME J. *5*, 574–579.

Marcy, Y., Ouverney, C., Bik, E.M., Losekann, T., Ivanova, N., Martin, H.G., Szeto, E., Platt, D., Hugenholtz, P., Relman, D.A., *et al.* (2007). Dissecting biological 'dark matter' with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. Proc. Natl. Acad. Sci. U.S.A. *104*, 11889–11894.

Mason, O.U., Hazen, T.C., Borglin, S., Chain, P.S.G., Dubinsky, E.A., Fortney, J.L., Han, J., Holman, H.-Y.N., Hultman, J., Lamendella, R., *et al.* (2012). Metagenome, metatranscriptome and single-cell sequencing reveal microbial response to Deepwater Horizon oil spill. ISME J. *6*, 1715–1727.

Mendes, R., Kruijt, M., de Bruijn, I., Dekkers, E., van der Voort, M., Schneider, J.H., Piceno, Y.M., DeSantis, T.Z., Andersen, G.L., Bakker, P.A., *et al.* (2011). Deciphering the rhizosphere microbiome for disease-suppressive bacteria. Science *332*, 1097–1100.

Miezeiewski, M., Schnaufer, T., Muravsky, M., Wang, S., Caro-Aguilar, I., Secore, S., Thiriot, D.S., Hsu, C., Rogers, I., DeSantis, T.Z., *et al.* An in virto culture model to study the dynamics of colonic microbiota in syrian golden hamster and their susceptibility to infection with Clostridium difficile. ISME J. (submitted).

Moore, M.J., Dhingra, A., Soltis, P.S., Shaw, R., Farmerie, W.G., Folta, K.M., and Soltis, D.E. (2006). Rapid and accurate pyrosequencing of angiosperm plastid genomes. BMC Plant Biol. *6*, 17.

Moran, J.J., Beal, E.J., Vrentas, J.M., Orphan, V.J., Freeman, K.H., and House, C.H. (2008). Methyl sulfides as intermediates in the anaerobic oxidation of methane. Environ. Microbiol. *10*, 162–173.

Morris, B.E., Henneberger, R., Huber, H., and Moissl-Eichinger, C. (2013). Microbial syntrophy: interaction for the common good. FEMS Microbiol. Rev. *37*, 384–406.

Nocker, A., Sossa-Fernandez, P., Burr, M.D., and Camper, A.K. (2007). Use of propidium monoazide for live/dead distinction in microbial ecology. Appl. Environ. Microbiol. *73*, 5111–5117.

Noval Rivas, M., Burton, O.T., Wise, P., Zhang, Y.Q., Hobson, S.A., Garcia Lloret, M., Chehoud, C., Kuczynski, J., DeSantis, T., Warrington, J., *et al.* (2013). A microbiota signature associated with experimental food allergy promotes allergic sensitization and anaphylaxis. J. Allergy Clin. Immunol. *131*, 201–212.

Orphan, V.J., House, C.H., Hinrichs, K.U., McKeegan, K.D., and DeLong, E.F. (2001). Methane-consuming archaea revealed by directly coupled isotopic and phylogenetic analysis. Science *293*, 484–487.

Pinto, A.J., and Raskin, L. (2012). PCR Biases Distort Bacterial and Archaeal Community Structure in Pyrosequencing Datasets. PLoS ONE *7*, e43093.

Probst, A., Vaishampayan, P., Osman, S., Moissl-Eichinger, C., Andersen, G.L., and Venkateswaran, K. (2010). Diversity of anaerobic microbes in spacecraft assembly clean rooms. Appl. Environ. Microbiol. *76*, 2837–2845.

Probst, A.J., Holman, H.Y., DeSantis, T.Z., Andersen, G.L., Birarda, G., Bechtel, H.A., Piceno, Y.M., Sonnleitner, M., Venkateswaran, K., and Moissl-Eichinger, C. (2013). Tackling the minority: sulfate-reducing bacteria in an archaea-dominated subsurface biofilm. ISME J. *7*, 635–651.

Probst, A.J., Birarda, G., Holman, H.-Y.N., DeSantis, T.Z., Wanner, G., Andersen, G.L., Perras, A.K., Meck, S., Völkel, J., Bechtel, H.A., *et al.* (2014). Coupling genetic and chemical microbiome profiling reveals heterogeneity of Archaeome and Bacteriome in subsurface biofilms that are dominated by the same Archaeal species. PLoS ONE, in press.

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. Meth. *6*, 639–641.

Radosevich, J.L., Wilson, W.J., Shinn, J.H., DeSantis, T.Z., and Andersen, G.L. (2002). Development of a high-volume aerosol collection system for the identification of air-borne micro-organisms. Lett. Appl. Microbiol. *34*, 162–167.

Reddy, C.M., Arey, J.S., Seewald, J.S., Sylva, S.P., Lemkau, K.L., Nelson, R.K., Carmichael, C.A., McIntyre, C.P., Fenwick, J., Ventura, G.T., *et al.* (2012). Composition and fate of gas and oil released to the water column during the Deepwater Horizon oil spill. Proc. Natl. Acad. Sci. USA *109*, 20229–20234.

Redmond, M.C., and Valentine, D.L. (2012). Natural gas and temperature structured a microbial community response to the Deepwater Horizon oil spill. Proc. Natl. Acad. Sci. USA *109*, 20292–20297.

Reeder, J., and Knight, R. (2009). The 'rare biosphere': a reality check. Nat. Meth. *6*, 636–637.

Renwick, J., McNally, P., John, B., DeSantis, T., Linnane, B., and Murphy, P. The microbial community of the cystic fibrosis airway is disrupted in early life. PLoS One (submitted).

Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N.N., Anderson, I.J., Cheng, J.F., Darling, A., Malfatti, S., Swan, B.K., Gies, E.A., *et al.* (2013). Insights into the phylogeny and coding potential of microbial dark matter. Nature *499*, 431–437.

Rivers, A.R., Sharma, S., Tringe, S.G., Martin, J., Joye, S.B., and Moran, M.A. (2013). Transcriptional response of bathypelagic marine bacterioplankton to the Deepwater Horizon oil spill. ISME J. *7*, 2315–2329.

Roh, S.W., Abell, G.C.J., Kim, K.-H., Nam, Y.-D., and Bae, J.-W. (2010). Comparing microarrays and next-generation sequencing technologies for microbial ecology research. Trends Biotechnol. *28*, 291–299.

Saulnier, D.M., Riehle, K., Mistretta, T.A., Diaz, M.A., Mandal, D., Raza, S., Weidler, E.M., Qin, X., Coarfa, C., Milosavljevic, A., *et al.* (2011). Gastrointestinal microbiome signatures of pediatric patients with irritable bowel syndrome. Gastroenterology *141*, 1782–1791.

Schloss, P.D., Gevers, D., and Westcott, S.L. (2011). Reducing the Effects of PCR Amplification and Sequencing Artifacts on 16S rRNA-Based Studies. PLoS ONE *6*, e27310.

Smith, D.J., Timonen, H.J., Jaffe, D.A., Griffin, D.W., Birmele, M.N., Perry, K.D., Ward, P.D., and Roberts, M.S. (2013). Intercontinental dispersal of bacteria and archaea by transpacific winds. Appl. Environ. Microbiol. *79*, 1134–1139.

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.

Vaishampayan, P., Probst, A.J., La Duc, M.T., Bargoma, E., Benardini, J.N., Andersen, G.L., and Venkateswaran, K. (2013). New perspectives on viable microbial communities in low-biomass cleanroom environments. ISME J. *7*, 312–324.

Valentine, D.L., Kessler, J.D., Redmond, M.C., Mendes, S.D., Heintz, M.B., Farwell, C., Hu, L., Kinnaman, F.S., Yvon-Lewis, S., Du, M., *et al.* (2010). Propane respiration jump-starts microbial response to a deep oil spill. Science *330*, 208–211.

Vaziri, N.D., Wong, J., Pahl, M., Piceno, Y.M., Yuan, J., DeSantis, T.Z., Ni, Z., Nguyen, T.H., and Andersen, G.L. (2013). Chronic kidney disease alters intestinal microbial flora. Kidney Int. *83*, 308–315.

Wang, Q., Garrity, G.M., Tiedje, J.M., and Cole, J.R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl. Environ. Microbiol. *73*, 5261–5267.

Wang, S., Fan, C., Low, A., and He, J. (2013). Comparison of microbial communities in sequencing batch reactors (SBRs) exposed to trace erythromycin and erythromycin-H_2O. Appl. Microbiol. Biotechnol. (in press, DOI 10.1007/s00253–013–5205–2).

de Wit, R., and Bouvier, T. (2006). 'Everything is everywhere, but, the environment selects'; what did Baas Becking and Beijerinck really say? Environ. Microbiol. *8*, 755–758.

Wortman, J., Giglio, M., Creasy, H., Chen, A., Liolios, K., Chu, K., Davidovics, N., Mazaitis, M., DeSantis, T., Singh, N., *et al.* (2010). A data analysis and coordination center for the human microbiome project. Genome Biol. *11*, 1–1.

Wrighton, K.C., Thomas, B.C., Sharon, I., Miller, C.S., Castelle, C.J., VerBerkmoes, N.C., Wilkins, M.J., Hettich, R.L., Lipton, M.S., Williams, K.H., *et al.* (2012). Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla. Science *337*, 1661–1665.

Wu, Y.R., and He, J. (2013). Characterization of anaerobic consortia coupled lignin depolymerization with biomethane generation. Bioresour. Technol. *139*, 5–12.

Yatsunenko, T., Rey, F.E., Manary, M.J., Trehan, I., Dominguez-Bello, M.G., Contreras, M., Magris, M., Hidalgo, G., Baldassano, R.N., Anokhin, A.P., *et al.* (2012). Human gut microbiome viewed across age and geography. Nature *486*, 222–227.

Zhang, M., Powell, C.A., Benyon, L.S., Zhou, H., and Duan, Y. (2013). Deciphering the Bacterial Microbiome of Citrus Plants in Response to 'Candidatus Liberibacter asiaticus'-Infection and Antibiotic Treatments. PLoS One *8*, e76331.

Zhou, J., Kang, S., Schadt, C.W., and Garten, C.T. (2008). Spatial scaling of functional gene diversity across various microbial taxa. Proc. Natl. Acad. Sci. USA *105*, 7768–7773.

Zhou, J., Wu, L., Deng, Y., Zhi, X., Jiang, Y.-H., Tu, Q., Xie, J., Van Nostrand, J.D., He, Z., and Yang, Y. (2011). Reproducibility and quantitation of amplicon sequencing-based detection. ISME J. *5*, 1303–1313.

Zhou, J., Xue, K., Xie, J., Deng, Y., Wu, L., Cheng, X., Fei, S., Deng, S., He, Z., Van Nostrand, J.D., *et al.* (2012). Microbial mediation of carbon-cycle feedbacks to climate warming. Nat. Clim. Change *2*, 106–110.

Zhou, J., Jiang, Y.-H., Deng, Y., Shi, Z., Zhou, B.Y., Xue, K., Wu, L., He, Z., and Yang, Y. (2013). Random Sampling Process Leads to Overestimation of β-Diversity of Microbial Communities. mBio *4*, e00324–13.