The Use of Microarrays in Microbial Ecology

By Gary L. Andersen, Zhili He, Todd Z. DeSantis, Eoin L. Brodie, and Jizhong Zhou

Abstract

Microarrays have proven to be a useful and high-throughput method to provide targeted DNA sequence information for up to many thousands of specific genetic regions in a single test. A microarray consists of multiple DNA oligonucleotide probes that, under high stringency conditions, hybridize only to specific complementary nucleic acid sequences (targets). A fluorescent signal indicates the presence and, in many cases, the abundance of genetic regions of interest. In this chapter we will look at how microarrays are used in microbial ecology, especially with the recent increase in microbial community DNA sequence data. Of particular interest to microbial ecologists, phylogenetic microarrays are used for the analysis of phylotypes in a community and functional gene arrays are used for the analysis of functional genes, and, by inference, phylotypes in environmental samples. A phylogenetic microarray that has been developed by the Andersen laboratory, the PhyloChip, will be discussed as an example of a microarray that targets the known diversity within the 16S rRNA gene to determine microbial community composition. Using multiple, confirmatory probes to increase the confidence of detection and a mismatch probe for every perfect match probe to minimize the effect of cross-hybridization by non-target regions, the PhyloChip is able to simultaneously identify any of thousands of taxa present in an environmental sample. The PhyloChip is shown to reveal greater diversity within a community than rRNA gene sequencing due to the placement of the entire gene product on the microarray compared with the analysis of up to thousands of individual molecules by traditional sequencing methods. A functional gene array that has been developed by the Zhou laboratory, the GeoChip, will be discussed as an example of a microarray that

dynamically identifies functional activities of multiple members within a community. The recent version of GeoChip contains more than 24,000 50mer oligonucleotide probes and covers more than 10,000 gene sequences in 150 gene categories involved in carbon, nitrogen, sulfur, and phosphorus cycling, metal resistance and reduction, and organic contaminant degradation. GeoChip can be used as a generic tool for microbial community analysis, and also link microbial community structure to ecosystem functioning. Examples of the application of both arrays in different environmental samples will be described in the two subsequent sections.

1. INTRODUCTION

1.1 The use of microarray technologies for microbial ecology studies

Microarrays have become an increasingly popular method in the microbial ecologist’s toolkit for analyzing microbial communities. This trend has also been observed in other disciplines resulting in an increase in PubMed entries mentioning microarrays from 20 in 1998 to over 34,300 by November 2008 (Loring, 2006). Within the scope of this chapter, microarrays can be simply thought of as a high throughput way to detect the presence and concentration of multiple nucleotide sequences within an environmental sample. Sequence complementarity between single stranded nucleic acid molecules, one of which is typically immobilized on the microarray, leads to the hybridization of specific target sequences from a sample. The advent of large genomic sequencing centers and the continuing decrease in sequencing costs have lead to an explosion of DNA sequence data, including data from environmental microorganisms. This has created an even greater need for high throughput methods such as microarrays to make efficient use of the expanding sequence databases.

In this chapter we will discuss two specific examples of microarrays that are being used to answer questions of interest to microbial ecologists. The first example targets the 16S rRNA gene of bacteria and archaea (PhyloChip) to identify specific members within a complex microbial community. We will discuss how this microarray has been used to characterize the microbial diversity of selected environments. The second example targets known functional gene markers to study functional gene diversity and activities of microorganisms in the environment. We will demonstrate how functional gene arrays (GeoChips) have been used to analyze microbial communities, and provide linkages of microbial genes/populations to ecosystem processes and functions.

1.2 DNA microarrays

Based on the target of molecular markers, two types of major microarrays have been used for microbial community analysis. The first are phylogenetic microarrays that target phylogenetic genes, such as the 16S rRNA gene and *gyrB* gene. Phylogenetic microarrays allow us to identify microorganisms and their phylogenetic relationships in a community of interest. The second are functional gene arrays that target key functional gene markers that are indicative of a specific physiological or metabolic process, such as *nirK* and *nirS* encoding nitrite reductases, key enzymes of the denitrification process (Braker et al., 2000), and *amoA* encoding ammonia monooxygenase, a key enzyme for ammonia oxidization (Rotthauwe et al., 1997). Functional gene arrays allow us to study functional gene diversity and activities of microbial communities. In this section, both microarrays are introduced below.

1.2.1 Phylogenetic microarray

1.2.1.1 Analysis of microbial communities with 16S rRNA targeted microarrays

Molecular methods for detecting and monitoring bacteria and archaea routinely rely upon classifying heterogeneous 16S rRNA molecules, either as RNA or as gene fragments encoding RNA that are amplified by universal PCR primers. The general method of sampling sequence types has been to clone and sequence PCR products derived from these biomarkers. However, the number of clones required to adequately catalogue the majority of taxa in a sample is typically unwieldy assuming a log-normal distribution (Curtis and Sloan, 2005). For example, in a typical soil sample with one billion bacterial cells and 10,000 different species, one would need to sample at least one million sequences (Gans et al., 2005). Pyrosequencing has provided a alternative to cloning by producing a greater sample size at a lower cost (Huse et al., 2008). But, as they are currently practiced, neither method is able to cover environmental microbial populations spanning multiple orders of magnitude within a single sample. Sequencing community members is an essential but inefficient process, where the biomarkers representing the most abundant phylotypes or species mask less abundant but potentially significant members.

As an alternative approach to biomarker sampling by cloning, hybridization of target sequences to an array of probes, permits much greater numbers of molecules to be sampled compared with the hundreds or thousands that usually comprise an environmental clone library. Because of the small, oftentimes single nucleotide differences within the probed regions of the biomarker genes for the differentiation of microbial taxa, a high level of sequence specificity is desired. Oligonucleotide DNA microarrays often consist of large numbers of individual short, 15- to 30-nucleotide capture probes to offer the highest level of specificity for the identification of specific target sequences, particularly in a background of closely related sequences. As stated earlier in Chapter 1, the 16S rRNA gene sequence provides a number of advantages in its use as a biomarker for the identification of individual bacterial components in complex environmental samples. Rather than translating its genetic code into protein, rRNA acts directly in the protein assembly machinery as a functional molecule. Due to structural constraints of this molecule, specific regions throughout the 16S rRNA gene have a highly conserved nucleotide sequence while non-structural segments may have a high degree of variability (Woese et al., 1975). Probing the regions of high variability can be used to identify microorganisms at the species level while regions of less variability are used for group-level identification. With only one to a few nucleotides of sequence variability, at best, within any 15- to 30-bp region that may be targeted by a probe for discrimination between related microbial species, it is imperative to maximize the probe-target sequence specificity in the microarray system.

One example of a microarray that has been successfully used to discriminate bacterial species uses a hierarchical set of oligonucleotide probes to target organisms at different levels of taxonomic specificity on a matrix of acrylamide gel pads on a glass slide (Liu et al., 2001). Developed by Liu and Stahl, this method uses the 3-dimensional nature of the gel matrix to allow solution-based probe kinetics with a non-equilibrium dissociation approach for high levels of discrimination between target and non-target 16S rRNA gene sequences. Amplified 16S rRNA gene sequences are placed on the gel-pad microarray and allowed to hybridize under low-stringency conditions. Increasing the hybridization temperature (increased stringency) results in the preferential dissociation of non-homologous probe-target complexes. The simultaneously generated melting curves for the perfect match (PM) and mismatch (MM) duplexes are used to define the temperature at which 50% of the starting duplex remains intact (dissociation temperature, T_d.) It was found that, for the most part, a probe-target duplex with one MM has a greater than two-fold level of discrimination from a PM duplex at T_d, thus allowing greater specificity of detection by differentiating between PM and MM complexes.

By contrast, other 16S rRNA gene-targeted or phylogenetic microarrays use short oligonucleotide probes bound to a two-dimensional surface (e.g., glass) with specific sequences located at defined two dimensional coordinates. Many studies have successfully used these 16S rRNA gene-targeted microarrays to differentiate bacteria in specific groups, such as *Enterococcus* (Lehner et al., 2005), *Cyanobacteria* (Castiglioni et al., 2004), nitrifying bacteria (Kelly et al., 2005) and fish pathogens (Warsen et al., 2004) and for quantitative tuning of sampling protocols to enhance detection of desired taxonomic groups.

1.2.1.2 A comprehensive view of microbial diversity

Instead of targeting specific groups or classes of organisms, another strategy is to perform a comprehensive screen for all known bacterial or archaeal taxa on a single microarray. This relies, initially, on obtaining all known 16S rRNA gene sequences from the major sequence repositories, including Lawrence Berkeley National Laboratory's Greengenes (greengenes.lbl.gov) (DeSantis et al., 2006b), Michigan State University's ribosomal database project (RDP; rdp.cme.msu.edu) (Cole et al., 2005), the Max Planck Institute for Marine Microbiology's Silva database (http://www.arb-silva.de/), and the National Institute of Health's NCBI (www.ncbi.nlm.nih.gov). There are currently over 700,000 individual sequences housed in these repositories but due to the lack of peer-review before inputting the sequences from individual submitters there are a number of quality control issues that may not have been addressed, producing inaccurate data. To reduce this problem, a series of filters on the sequence data may be employed to reduce the possibility that an assay is created for a non-existing target. The Greengenes database of 16S rRNA gene sequences compiles information from other databases to produce a set of sequences that are compatible with a comprehensive phylogenetic microarray design. Among the issues addressed by this database are: (i) standardized taxonomic placement of individual 16S rRNA gene sequences, (ii) removal of chimeric sequences, (iii) removal of poor quality (ambiguous) sequences, and (iv) distribution of data in a consistently aligned sequence format.

Because the discovery rate of 16S sequence records from uncultured organisms now exceeds that from their cultured counterparts, taxonomic placement of sequences lags behind. In fact, over one-third of full-length 16S records in GenBank are presented without taxonomic nomenclature and are simply annotated as "environmental samples" or "unclassified". In contrast, records in the Greengenes dataset are annotated with taxonomy proposed by five independent curators: NCBI, RDP, based on Bergey's Manual?; ((Cole et al., 2005), Wolfgang Ludwig (Ludwig et al., 2004), Phil Hugenholtz (Hugenholtz, 2002) and Norm Pace (Pace, 1997), collectively covering over 95% of the database. Incongruent taxonomic nomenclature exists among curators even at the phylum-level, yet each is tracked to promote user awareness of several estimations of phylogenetic descent allowing a balanced approach to node and operational taxon unit (OTU) nomenclature when labeling probe specificity.

Since 16S rRNA genes from environmental DNA are usually PCR amplified, it has been suspected that many chimeric sequences are present in the public repositories. A PCR-generated chimeric sequence usually comprises two phylogenetically distinct parent sequences and occurs when a prematurely terminated amplicon re-anneals to a foreign DNA strand and is copied to completion in the following PCR cycles. The point at which the chimeric sequence changes from reflecting one parent to the next (break point) can be misconstrued as novel biomarker for a unique organism (Figure **1.**). The trend has been observed in 3% of the sequences from uncultured organisms and 0.2% of sequences annotated as pure cultures (Ashelford et al., 2005; DeSantis et al., 2006b). Very recently, large 16S rRNA clone libraries have been deposited to

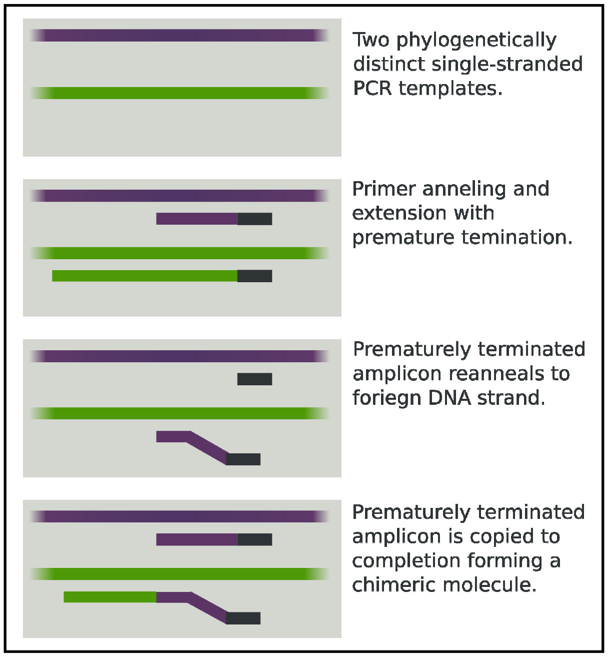
Two phylogenetically distinct single-stranded PCR templates.
Primer anneling and extension with premature temination.
Prematurely terminated amplicon reanneals to foriegn DNA strand.
Prematurely terminated amplicon is copied to completion forming a chimeric molecule.

Figure 1. An example of a chimeric artifact generated during PCR amplification of a mixed population using broad-specificity 16S rRNA gene primers. Partial amplicons may form hybrids with dissimilar templates because conserved regions exist at positions medial to the PCR primer targets. The partial amplicon can be extended using the dissimilar 16S gene as a template.

GenBank with an inferred chimeric content of up to 8% of the sequences (Ashelford et al., 2006). The Bellerophon (Huber et al., 2004) chimera test result is available for each Greengenes 16S record allowing putative chimeras to be avoided. This step helps to ensure that all probes in a microarray design are complementary to sequences from true organisms.

Public DNA data repositories rarely quality-check the primary data (chromatograms) of the sequences they distribute. Unfortunately, the majority of 16S rRNA gene sequences in the public databases are decoded from single coverage sequencing reads typically yielding “non-ambiguous” yet non-reproducible base calls in 2 of 400 chromatogram peaks (0.5%) even under ideal conditions where data is collected in small batches by experienced research groups (Fields et al., 2006). The problem is magnified in high-throughput efforts. Assuming chromatograms or quality scores are not obtainable for the sequence collection, various poor quality records can still be identified. Obvious sequencing reaction failures for the majority of the gene or a just a region can be easily discarded or trimmed based on the distribution of ambiguous base-calls (non-“ACGT” characters). Interpreted chromatograms should also have zero or few long homopolymeric regions (i.e. eight or more sequential adenines reported) since these have rarely been confirmed in isolates and base calls downstream of the homopolymer commonly have increased error rates. Furthermore, an emphasis should be placed on reads covering a large portion of the gene of interest so that multiple probes can be picked from throughout the domains. Greengenes enables de-selection of records with an undesirable percentage of ambiguous calls, one or more homopolymeric runs, and low gene coverage (short reads).

1.2.2 Functional gene arrays

As described in Chapter X, one gram of soil contains more than 5000 microorganisms, and a majority of them (>99%) have not yet been cultivated (Whitman et al., 1998), which presents enormous difficulties for microbiologists to study microbial composition, structure, function, and dynamics in natural and/or contaminated environments. Using functional gene markers to investigate such microbial communities is therefore necessary. Schadt et al. (Schadt et al., 2004) listed some of those functional genes currently used in environmental studies that allow us to study functional gene diversity and activities of microorganisms in the environment. Currently, conventional molecular methods, such as PCR-based cloning, and *in situ* hybridization are very useful in providing snap shots for microbial diversity, structure and function, but they fail to provide a full picture of microbial activities and dynamics in a rapid and high-through-put fashion. Although microarray technology has been used successfully to analyze global gene expression in pure culture studies, adapting microarray technology for use in environmental studies presents numerous challenges in terms of probe design, the coverage of gene sequences, specificity, sensitivity and quantitative capability. To overcome such obstacles for studying microbial communities in natural settings, a particular type of microarrays, called functional gene arrays (FGAs), has been developed and used. This type of microarray contains probes from genes involved in key microbially-mediated biogeochemical processes, such as C, N, P, and S cycling and utilization, organic contaminant degradation, and metal reduction and resistance. FGAs that mainly target geochemical processes are also called GeoChips (He et al., 2007). FGAs are powerful tools to address some fundamental questions in microbial ecology, biogeochemistry, and environmental biology, such as: (i) What functional genes/microorganisms are in a microbial community? (ii) What biological or geochemical processes dominate in a microbial community? (iii) What are the dynamics of activity for a given gene or process? (iv) How does microbial community structure link to its function? (v) What are the relationships between functional gene activities/abundance and geochemical parameters?

For almost one decade, microbiologists, especially microbial ecologists have tried to answer a central question whether microarray-based FGA technologies can provide specific, sensitive and quantitative detection of microbial populations and activities within the context of environmental applications. The technology has evolved several generations in terms of gene/probe coverage and related computational techniques for functional gene sequence retrieval, probe selection, data analysis, and information storage. The first generation of FGAs was prototype microarrays with a focus on proof of the concept. For example, such a prototype microarray was constructed with 89 PCR-amplicon probes targeting functional genes involved in nitrogen cycling using pure cultures and laboratory clones (Tiquia et al., 2004; Wu et al., 2001). It presents a great challenge to obtain PCR amplicon probes from a diversity of environmental clones and bacterial strains from various sources, so constructing a comprehensive FGA is very difficult. Therefore, oligonucleotide-based FGAs have become more popular due to their high specificity, ease of construction, and low cost. Oligonucleotide FGAs are fabricated with synthetic oligonucleotide probes with 50mer {Rhee, 2004 #436) or 70mer (Taroncher-Oldenburg et al., 2003) in length, and they can be spotted on glass slides (Rhee et al., 2004; Taroncher-

Oldenburg et al., 2003; Tiquia et al., 2004; Wu et al., 2001), or nylon membranes (Steward et al., 2004). Different FGAs have been systematically evaluated in terms of their sensitivity, specificity, and quantitation, and the results demonstrate that FGA technology holds promise for the analysis of microbial communities. In addition, an FGA may target both functional and phylogenetic markers. For example, Loy et al. (Loy et al., 2004) constructed an array containing both phylogenetic (16S rRNA gene) and functional (*dsrA/B*) markers and those example FGAs were classified as GeoChip 1.0. The current version, GeoChip 2.0 has been designed and used for a comprehensive analysis of microbial community structure, function, and dynamics in a natural or contaminated environment. The coverage of gene sequences/probes has greatly increased from hundreds to tens of thousands, and related computational techniques, such as sequence retrieval, probe design, and data analysis have been greatly improved. GeoChip 2.0 contains 24,243 oligonucleotide (50mer) probes and covers > 10,000 genes in >150 functional groups involved in nitrogen, carbon, sulfur and phosphorus cycling, metal reduction and resistance, and organic contaminant degradation **(**Table 1; (He et al., 2007)). This array uses experimentally established probe design criteria (He et al., 2005b; Liebich et al., 2006) and a new computational software tool, CommOligo (Li et al., 2005) for oligonucleotide probe selection. In addition, an analysis of sub-nanogram quantities of microbial community DNA has been achieved by whole-community genome amplification (Wu et al., 2006). This approach made it possible to analyze microbial communities with low biomass. The next generation, GeoChip 3.0, is currently being developed. It is expected to be more comprehensive, covering >45,000 gene variants (sequences) in > 290 gene categories. Many new features will be implemented.

1.3 General Experimental Procedures

1.3.1 Sample preparation

For more detailed description on how microarrays are used to identify and quantify specific genetic sequences for biological research in general please refer to the excellent recent reviews by R. B. Stoughton and J. W. Edwards (Allison et al., 2007; Stoughton, 2005). One of the main factors distinguishing the use of microarrays within microbial ecology from other areas of study is that the genetic sequences are derived from multiple organisms in what is usually a complex microbial community. Because the constituents of a microbial community typically have very different types of cell walls it is important to find a target isolation and purification method that is suitable for a wide range of conditions. Equally important for microbial ecology applications is to remove environmental inhibitors of nucleic acid amplification. An example of a robust nucleic acid extraction and purification protocol is the method established by Zhou et al. (Zhou et al., 1996) or using other approaches described in Chapter X. The method used normally requires modification depending on the experimental goals and environmental sample type, such as soils, sediments, and groundwater (Hurt et al., 2001). Many DNA extraction and purification kits are commercially available, and these were recently compared in terms of DNA extraction efficiency from different types of samples (Klerks et al., 2006). Purified DNA samples should have $A_{260}/A_{280} > 1.80$, and $A_{260}/A_{230} > 1.70$. Since a typical hybridization for GeoChip analysis requires 2-5 µg purified DNA, samples with lower than 2 µg require amplification using the currently developed WCGA (whole community genome amplification) method (Wu et al., 2006). Obtaining purified mRNA from environmental samples is an even greater challenge than DNA, especially for low-abundance mRNAs. This is still often desirable since mRNA is an ideal indictor of microbial activity. Total RNA can be isolated and purified using the approach described by Hurt et al. (Hurt et al., 2001). This method can isolate DNA and RNA simultaneously with the same sample. Recently, a new gel electrophoresis method to isolate community RNA was developed (McGrath et al., 2008). Normally, the ratios of A_{260}/A_{280} and A_{260}/A_{230} for a purified RNA are expected to be >1.90, and >1.70, respectively. For analysis of a microbial community, a normal hybridization requires 10-20 µg purified RNA, samples with lower than 5 µg will again require amplification, as may be performed with the novel whole community RNA amplification approach (WCRA) (Gao et al., 2007) to obtain cDNA. With such a method, 1,200- to 1,800-fold amplification can be obtained with 10 to 100 ng of RNA as templates (Gao et al., 2007).

1.3.2 Microarray fabrication

There are several different styles of microarrays used for microbial community analysis. The Affymetrix (Santa Clara, CA) platform DNA arrays that are used on phylogenetic arrays such as the PhyloChip have the short oligonucleotide probes (~25-mer) synthesized directly on the glass surface by a photolithography method at an approximate density of 10,000 molecules per µm2 (Chee et al., 1996). Spotted DNA arrays use oligonucleotides that are synthesized individually at a predefined concentration and are applied to a chemically activated glass surface. Oligonucleotide length can range from a few nucleotides to hundreds of bases in length but are typically in the 50-mer range for functional arrays such as the GeoChip.

1.3.3 Target labeling

The nucleic acid targets are labeled so that a laser scanner tuned to a specific wavelength can measure the number of fluorescent molecules that hybridized to a specific DNA probe. For photolithography arrays such as Affymetrix, the nucleic acid targets are fragmented to between 50 and 100-bp size and a biotinylated nucleotide is added to the end of the fragment by terminal DNA transferase. At a later stage, the biotinylated fragments that hybridize to the oligonucleotide probes are used as a substrate for the addition of multiple phycoerythrin fluorophores by a sandwich antibody (Streptavidin) method.

For spotted arrays such as the GeoChip, the purified community DNA can be fluorescently labeled by random priming using the Klenow fragment of DNA polymerase as described previously (Wu et al., 2006) and more than one fluorescent moiety can be used (e.g. controls could be labeled with Cy3, and experimental samples labeled with Cy5 for direct comparison by hybridization to a single microarray). Total community RNA (e.g. 5-10 µg) can be labeled using Cy5 or Cy3 with Superscript^{TM} II/III RNase H^{-} reverse transcriptase (Invitrogen Life Technologies, CA) as described by He et al. (He et al., 2005a). The labeled cDNA target is then purified and concentrated.

1.3.4 Hybridization

Microarray hybridizations are then carried out under stringent conditions described previously (Rhee et al., 2004; Wu et al., 2006). The temperature can be lowered to reduce stringency and allow the detection of more divergent sequences. Robotic hybridization and stringency wash stations can be used to give more consistent results. The photolithography arrays use an automated Affymetrix hybridization and fluidics station for all washes as well as fluorescent staining with antibody and phycoerythrin. Spotted arrays can use the TECAN HS4800 (Tecan U.S., Inc., Durham, NC) to replace manual hybridization, which allows 48 hybridizations to be completed in 6 hours.

1.3.5 Signal quantification and analysis

After hybridization the arrays are scanned using a microarray scanner (e.g. GeneChip Scanner 3000, Affymetrix, Santa Clara, CA for PhyloChip, or ProScan Array, Perkin Elmer, Boston, MA for GeoChip) equipped with lasers at a resolution of 10 µm or finer. The scanned image displays are saved and analyzed by quantifying the pixel density (intensity) of each spot using image quantification software (e.g. GeneChip Microarray Analysis Suite, version 5.1 Affymetrix, Santa Clara, CA for PhyloChip, or ImaGene 6.0, Biodiscovery Inc. Los Angeles, CA for GeoChip).

2. PhyloChip, A PHYLOGENETIC MICROARRAY

2.1 PhyloChip Design.

The key considerations that must be taken into account in designing a 16S rRNA gene-based microarray to identify individual organisms in a complex environmental mixture are 1.) natural sequence diversity and 2.) potential cross-hybridization. Sequence diversity is an issue as we sample new and distinctive environments such as bioaerosols. There may be many undocumented organisms with 16S rRNA gene sequences that are similar, but not identical to the sequences that were used for array design. Microarrays based upon single sequence-specific hybridizations (single probes) per OTU may be ineffective in detecting such environmental sequences with one or several polymorphisms. To overcome this obstacle, an Affymetrix-style photolithography chip was designed with a minimum of 11 different, short oligonucleotide probes for each taxonomic grouping, allowing for the failure of one or more probes. Also important is non-specific cross hybridization, especially when an abundant 16S rRNA gene shares sufficient sequence similarity to non-targeted probes, such that a weak but detectable signal is obtained. It has been found that the perfect match-mismatch (PM-MM) probe pair approach effectively minimizes the influence of cross-hybridization. Widely used on expression arrays as a control for non-specific binding (Chee et al., 1996), the central nucleotide is replaced with any of the three non-matching bases so that the increased hybridization intensity signal of the PM over the paired MM indicates a sequence-specific, positive hybridization. By requiring multiple PM-MM probe-pairs to have a positive interaction, we substantially increase the chance that the hybridization signal is due to a predicted target sequence.

Once a reference set of valid genes is established it is important to create a multiple sequence alignment (MSA). The MSA allows confident comparisons between sequences when selecting probes. For instance, when a candidate probe does not complement a sequence it is practical to determine if sequence data is available at the expected probe position. Filtered, aligned 16S rRNA sequence records can be exported directly from Greengenes. Since the Greengenes database consistently spreads the gene into an alignment of 7682 characters in width, then private, in-house sequences can be formatted into the same MSA using the NAST (Nearest Alignment Space Termination) web tool (DeSantis et al., 2006a). Other strategies exist for compiling MSAs from various sequence sources and the choice is governed by the size of the project. In general, alignment can be done with clustalw (Thompson et al., 1994) for small MSAs (<500 sequences), MUSCLE (Edgar, 2004) for mid-size MSAs, (500-10,000 sequences) or NAST (DeSantis et al., 2006a) for large MSAs (>10,000 genes).

2.2 Simultaneous clustering of genes and probes.

The objective of the probe selection strategy for comprehensive prokaryotic identification is to obtain an effective set of probes capable of correctly categorizing mixed amplicons into their proper operational taxonomic unit (OTU) designations. Each OTU is formed from sequences, which have common oligomer targets, considering only those targets that meet the G+C, melting temperature, and secondary structure constraints of the design. Whereas clustering implemented to infer phylogeny may utilize similarities along the entire gene, creating OTUs relies on finding shared attributes that can be assayed. A supervised clustering procedure is used which consists first of generating numerous candidate OTUs by unsupervised hierarchical clustering using pair-wise gene distances measured by megaBLAST (Zhang et al., 2000) or counts of unique targets. Then, each candidate OTU is evaluated to determine the count of targets which are simultaneously prevalent across the genes of the candidate OTU and also incapable of hybridization to genes outside the OTU.

In designing the G2 PhyloChip, probes presumed to have the capacity to correctly hybridize were those unique 25-mers that also contain a central 17-mer not matching any sequences outside the OTU (Urakawa et al., 2002). Thus, probes that were unique to an OTU solely due to a distinctive base in one of the outer four bases were avoided. For each OTU harvested from the hierarchical trees, a set of 11 or more specific 25-mers (probes) was sought. Three classes of probe sets resulted from the clustering procedure. The resulting 8,741 OTUs, each containing an average of 3% sequence divergence, represented all 121 demarcated bacterial and archaeal orders. In most cases, as expected, the OTUs contained sequences that were previously identified as related using the phylogenic tree approaches. For a majority of the OTUs represented on the PhyloChip (5,737; 65%), probes were designed from regions of gene sequences that have been identified only within a given taxon. For 1,198 taxa (14%), no probe-level sequence could be identified that was not shared with other groups of 16S rRNA gene sequences, although the gene sequence as a whole was distinctive. For these taxonomic groupings, an average of 24 probes (but no less than 11 probes) was designed to a combination of regions on the 16S rRNA gene that taken together as a whole did not exist in any other taxa. For the remaining 1,806 taxa (21%), a set of probes were selected to minimize the number of putative cross-reactive taxa. Although more than half of the probes in this group have a hybridization potential to one outside sequence, this sequence was typically from a phylogenetically similar taxon. For all three probe set classes, the advantage of the hybridization approach used was that multiple OTUs could be identified simultaneously by targeting unique regions or combinations or regions from genes.

2.3 Analyzing PhyloChip data

The G2 PhyloChip consists of 506,944 probe features, arranged as a grid of 712 rows and columns. Of these features, 297,851 are oligonucleotide PM or MM probes with exact or inexact complementarity, respectively, to 16S rRNA genes. The remaining probes are used for image orientation, normalization controls, or for pathogen-specific signature amplicon detection using additional targeted regions of the chromosome (Wilson et al., 2002). Each high-density 16S rRNA gene microarray is designed with additional probes that: 1) target amplicons of prokaryotic metabolic genes spiked into the 16S rRNA gene amplicon mix in defined quantities just prior to fragmentation and 2) are complimentary to pre-labeled oligonucleotides added into the hybridization mix. The first control collectively tests the target fragmentation, labeling by biotinylation, array hybridization, and staining/scanning efficiency. It also allows the overall fluorescent intensity to be normalized across all the arrays in an experiment. The second control directly assays the hybridization, staining and scanning.

Complementary targets to the probe sequences hybridize to the array and fluorescent signals are captured as pixel images using standard Affymetrix software (GeneChip Microarray Analysis Suite, version 5.1) that reduces the data to an individual signal value for each probe and is typically exported as a human readable "CEL" file. Background probes are identified from the CEL file as those producing intensities in the lowest 2% of all intensities. The average intensity of the background probes is subtracted from the fluorescence intensity of all probes. The noise value (N) is the variation in pixel intensity signals observed by the scanner as it reads the array surface. The standard deviation of the pixel intensities within each of the identified background probe intensities is divided by the square root of the number of pixels comprising that feature. The average of the resulting quotients is used for N in the calculations described below.

Probe pairs scored as positive are those that meet two criteria: (i) the fluorescence intensity from the perfectly matched probe (PM) is at least 1.3 times greater than the intensity from the mismatched control (MM), and (ii) the difference in intensity, PM minus MM, is at least 130 times greater than the squared noise value ($>130\ N^2$). The positive fraction (PosFrac) is calculated for each probe set as the number of positive probe pairs divided by the total number of probe pairs in a probe set. An OTU is considered "present" when its PosFrac for the corresponding probe set is >0.92 (based on empirical data from clone library analyses).

Replicate arrays can be used collectively in determining the presence of each OTU by requiring each to exceed a PosFrac threshold. Present calls are propagated upwards through the taxonomic hierarchy by considering any node (sub-family, family, order, etc.) as "present" if at least one of its subordinate OTUs was present.

Hybridization intensity is the measure of OTU abundance and is calculated in arbitrary units for each probe set as the trimmed average (maximum and minimum values removed before averaging) of the PM minus MM intensity differences across the probe pairs in a given probe set. All intensities <1 are shifted to 1 to avoid errors in subsequent logarithmic transformations.

2.4 Comparing PhyloChip and Clone Libraries

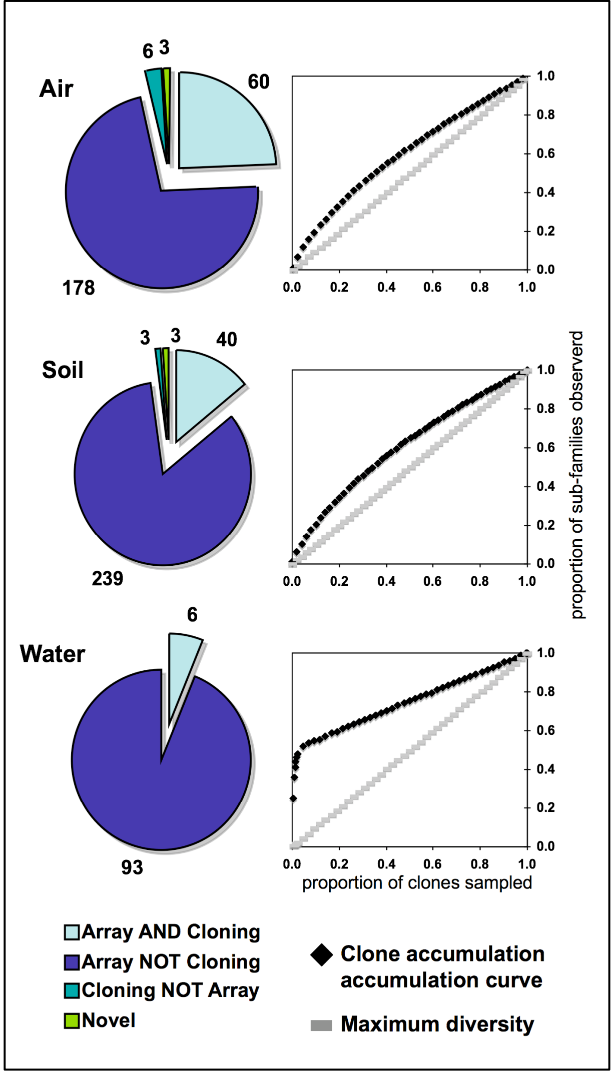

Figure 2. Comparison of cloning and PhyloChip (Array) analysis of air, soil and water samples. Left panel shows numbers of bacterial sub-families detected by array and cloning, array only, cloning only and novel sequences not detectable by the array. Right panel shows sub-family accumulation curves obtained by serial clone observations. The maximum diversity scenario where every clone sampled represents a new sub-family is indicated by grey line with constant slope = 1.

The breadth and accuracy of the PhyloChip in detecting diverse 16S rRNA gene sequence types compared to cloning-and-sequencing can be directly compared (DeSantis et al., 2007). Using three environmental samples: urban aerosol, subsurface soil and subsurface water, 16S PCR products were classified using both methods. Approximately 8% of the clones could not be placed into a known sub-family and were considered novel. The microarray results confirmed the majority of clone-detected sub-families and additionally demonstrated greater amplicon diversity (Figure 2, left) extending into phyla not observed by the cloning method. Sequences within the phyla *Nitrospira*, *Planctomycetes*, and TM7, which were uniquely detected by the array, were verified with specific primers and subsequent amplicon sequencing. Communities dominated by few sub-families are most likely to errantly appear well sampled (See accumulation curve for water, Fig 2, Right) by the clone library. The PhyloChip enables observation of the rarer bacterial populations since the entire mass of PCR products (~10^{12} molecules) are sampled as opposed to a mere hundreds to thousands when cloning. In general, although the microarray is unreliable in identifying novel taxa, it reveals greater diversity in environmental samples than sequencing a typically sized (300-700) clone library. Deeper sequencing of up to several hundred thousand targets is possible with pyrosequencing methods. This has the potential to reduce the difference in diversity detected by both methods but at present such efforts are typically confined to fewer samples and not to fully replicated ecological studies.

2.5 Direct analysis of microbial community activity by microarray analysis of ribosomal RNA (rRNA).

While analysis of DNA by high-density phylogenetic microarrays provides information regarding relative changes in biomass within a microbial community, cell division in natural ecosystems may occur on the order of hours to decades or even millennia (Phelps et al., 1994). Therefore the immediate environmental impacts on microorganisms may not be apparent through the analysis of DNA turnover rates. As in studies using stable-isotope labeled substrates such as ^{13}C (see Chapter 4), cellular RNA due to its higher synthesis rates, responds rapidly to environmental stimuli. For this reason, assaying the more labile RNA molecules permits immediate impacts to be monitored independent of replication.

Many standard approaches to sample preparation for RNA analysis require multiple enzymatic steps to replicate, amplify and label nucleic acid targets (e.g. reverse transcription, in vitro transcription or PCR). An advantage of these approaches is the ability to detect lower-abundance species but the quantitative representation of the original population may be altered during enzymatic amplification steps, such as PCR, or by the labeling procedure itself. Direct labeling of RNA on the other hand does not typically require amplification and also necessitates fewer manipulations, leading to a reduction in bias.

Direct labeling of RNA can be accomplished by various means such as base-incorporation with labeled nucleotides, or chemical modification. We have chosen to use a recently developed direct-labeling method using T4 RNA ligase to attach a biotinylated nucleotide donor molecule to the 3′ end of RNA targets according to the approach of Cole et al (Cole et al., 2004). This approach results in uniform end-labeling of RNA fragments, avoiding sequence bias inherent in methods which require incorporation of labeled nucleotides during synthesis (e.g. biotin–dUTP). End-labeled RNA produced through this approach is predicted to have higher target–probe affinity because hybridization is unimpaired by label molecules, whereas methods that either label the nucleic acid strand internally, such as biotin–ULS (Vanbelkum et al., 1994), or chemically modify a nucleotide (Kelly et al., 2002) can result in decreased hybridization efficiency (Cook et al., 1988). As the probes on the Affymetrix platform are tethered at the 3’ end , the 3′ position of the label will result in increased exposure of the biotin for more efficient binding by the streptavidin-fluorophore conjugate. Furthermore the donor molecule used to label targets with biotin contains biotin moieties tethered to the 3'-hydroxyl, rather than attached to the nucleobase. This permits multiple biotin molecules (3 in this case) to be tethered sequentially to the donor nucleotide without any significant reduction in ligation efficiency and has the added potential of enhancing overall signal intensity to improve the detection of lower-abundance/activity taxa.

Other groups have also been successful with direct microarray-based rRNA analysis. Small et al. (Small et al., 2001) hybridized total RNA from a soil extract to a simple microarray and demonstrated that rRNA could be analyzed without amplification, however they used biotin labeled detector probes rather than direct RNA target labeling. An important finding in this study was that fragmentation of the RNA produced greater hybridization specificity and sensitivity presumably through reduction of the complex secondary structure inherent in rRNA molecules. With this in mind, the choice of the most appropriate method for fragmentation must consider compatibility with the labeling procedure. Again, we have chosen to use the approach of Cole et al (Cole et al., 2004) by using an enzymatic RNA fragmentation procedure. This procedure uses RNaseIII (a double strand-specific ribonuclease) and shrimp alkaline phosphatase to simultaneously perform controlled fragmentation and dephosphorylation of the target RNA, yielding fragments in the range of 20-200 nucleotides. The dephosphorylation step is necessary because of the requirement of T4 RNA ligase for a 3'-hydroxyl on the RNA target.

2.6 Example application of direct RNA labeling and PhyloChip analysis - Monitoring bioremediation of uranium contaminated sediments.

A promising strategy for the containment of uranium within contaminated sites is in situ bioprecipitation. The valency state of uranium in contaminated groundwater and sediment, U(VI), is soluble and thus capable of transport by advective and diffusive processes. The tetravalent form, U(IV) is highly insoluble and therefore has reduced mobility. Metal-reducing bacteria such as those that typically respire iron have been proposed as efficient catalysts for reductive precipitation of uranium and several laboratory and field studies have demonstrated the feasibility of this approach (e.g. (Anderson et al., 2003; Brodie et al., 2006; Chang et al., 2005;

Peacock et al., 2004; Reardon et al., 2004; Wan et al., 2005)). One concern however, is the long-term stability of the bioreduced U(IV) and recent studies have demonstrated re-oxidation under anaerobic conditions (Ginder-Vogel et al., 2006; Wan et al., 2005).

In our studies we noted a carbonate-associated re-oxidation of U(IV) and further experiments were carried out to determine the influence of organic carbon (OC) electron donor concentration on the extent of re-oxidation (Wan et al., 2005). In one set of columns, concentrations (32mM OC) were retained at the concentration where re-oxidation was first noted after a period of successful reduction. In another set of columns, initially at 32mM OC, we increased the concentration to 100mM OC while eliminating OC from a third set of columns (0mM OC). As hypothesized, uranium re-oxidation increased with increasing OC concentration and was correlated with increased carbonate concentrations. Uranium re-oxidation ceased in the columns without OC additions. To examine the impact of changing carbon concentrations on the microbial community structure we first employed PhyloChip analysis using DNA extracted from the column sediments. Following amplification of the 16S rRNA gene, array analysis between the three treatments (0, 32 and 100 mM OC) indicated no significant differences in hybridization intensity among the 271 subfamilies detected (of the 842 subfamilies represented on the array). Figure 3A shows a comparison by DNA analysis between 0 and 100 mM OC treatments. At the point of sampling, columns had been running for almost two years with continuous OC addition, and most available electron acceptors (oxygen, nitrate, iron, manganese, sulfate) had been depleted or were of low bioavailability, therefore limitation on microbial replication may have been due to the limited electron acceptor and not the electron donor. Despite depletion of electron donors and the re-oxidation of uranium, several genera of metal-reducing bacteria including *Geobacter* and *Desulfovibrio* were still detected at similar intensities to those detected during the reducing phase. We hypothesized that cell-turnover was electron-acceptor limited but cellular activity was being maintained through interspecies electron-transfer. To assess this, we extracted RNA from the sediment samples, gel-purified the 16S rRNA band, fragmented, labeled and hybridized the material directly to the PhyloChip without any primers or amplification steps.

[insert fig 3 here]

[Figure 3. Comparison of microbial population dynamics under contrasting carbon amendment conditions using PhyloChip analysis of (A) DNA (PCR amplicons) and (B) RNA hybridization. X and Y axes display probe set intensities in arbitrary units under two carbon conditions (100 and 0 mM.]

Figure 3B shows a comparative analysis of hybridization intensities between the 0 mM and 100 mM organic carbon (OC) treatments. Immediately apparent was the detection of *Geobacter* 16S rRNA, indicating continued activity of this proven uranium reducing bacterial genus, despite observed re-oxidation. We also detected 16S rRNA of the methanogenic genus *Methanosarcina*. Various methanogens and sulfate reducers (also detected) are known to enter into syntrophic associations with the methanogens acting as biological electron acceptors in the absence of sulfate (Bryant et al., 1977). Other active organisms detected such as the acetogens, *Syntrophobacter* and *Syntrophus,* are specialized in exocellular electron transfer and may also serve as a sink for electrons in the absence of sulfate or possibly iron. Intriguingly, *Geobacter* species may also enter into syntrophic electron transfer associations although to date their association with methanogens is unclear.

3. GeoChip, A FUNCTIONAL GENE ARRAY

3.1 Design and construction

Construction of FGAs faces many challenges in both microarray-based technologies (Zhou, 2003) and computational software development. First, a retrieval of sequences specific to a particular functional gene by key words alone is difficult because gene/protein names are not always specific, or they can be differently annotated in different organisms. On the other hand, using functionally characterized and known sequences to search databases usually leads to a large number of hits, and it really depends on the threshold used. Additional difficulties are that variants of a functional gene are often highly similar, and most of them are homologues; sequences from uncultured microbes or laboratory clones may not be complete; and the number of deposited sequences for each gene continuously increases. Thus, it is very difficult to select specific probes for all variants (sequences) of a functional gene, which leads to a low coverage for most functional genes. In addition, standardization of oligonucleotide probe design criteria and software development is still challenging. FGA oligonucleotide probes should be capable of specifically detecting their targets in a complex sample whose components are not known. In addition, for data normalization and analysis, FGAs are very different from the most commonly used gene expression arrays, and novel methods are needed. Finally, in the era of genomics and meta-genomics, microbial sequences are daily produced in a gigabyte or terabyte fashion. A comprehensive FGA must be periodically updated to reflect the latest information.

To address the above challenges, some strategies have been developed or are in development. First, functional gene sequences are first retrieved using key words, and then unrelated sequences are removed by the HMMER program (http://hmmer.wustl.edu/). Second, functional gene sequences are aligned using a multiple sequence alignment (MSA) program after verification by HMMER. Only the shared regions of the functional genes are used for probe design. Third, experimentally established oligonucleotide design criteria and a novel software tool specifically targeting highly similar sequences are used to select oligonucleotide probes. Fourth, to detect both divergent and closely related sequences, both gene- and group-specific probes can be designed.

Details for GeoChip 2.0 design and construction were described by He et al. (He et al., 2007). Briefly, the major steps include: (1) Sequences of individual functional genes were retrieved from publicly available databases (e.g. NCBI GenBank). (2) 50mer oligonucleotides were designed after the removal of unrelated sequences using the oligouncleotide probe design software CommOligo (Li et al., 2005) with a new feature for group-specific probe selection, and experimentally established oligonucleotide design criteria (He et al., 2005b; Liebich et al., 2006). Both gene-specific and group-specific probes were designed with the following criteria: (i) gene-specific probes with sequence identity ≤90%, stretch ≤20 bases, and free energy ≥-35 kcal/mol (Liebich et al., 2006); and (ii) group-specific probes with sequence identity ≥96%, continuous stretch length ≥ 35 bases, and free energy ≤-60 kcal/mol (He et al., 2005b). (3) Oligonucleotide probes were designed on the basis of all variants (sequences) of the same gene. To ensure the whole array specificity, all designed probes of different genes were verified using the same design criteria against larger databases, such as NCBI and EMBL. (4) All verified probes were commercially synthesized, and spotted on glass slides (e.g. Corning UltraGAPS). (5) After printing, the slides were dried and then UV-cross-linked according to manufacturer’s instructions. It should be kept in mind that since the number of sequences in databases increases rapidly, it is important to periodically update the FGA databases and chips to reflect the current status.

3.2 Specificity, sensitivity and quantitative capability

Specificity is one of most important parameters to ensure that high quality microarray data can be obtained, and it is even more critical for FGA analysis of environmental samples. Microarray specificity can be controlled by probe design and hybridization conditions. Tiquia et al. (Tiquia et al., 2004) showed that sequences could be differentiated with < 86% identity when hybridizations were at 50^{o}C, and sequences with < 90% identity at 55^{o}C using a 50mer FGA with 763 probes for nitrogen cycling (e.g. *nirS*, *nirK*, nifH) and sulfate reduction (e.g. *dsrA/B)* genes. With a 50mer FGA containing 1662 probes for genes involved in contaminant degradation, Rhee et al. (Rhee et al., 2004) showed that at hybridization conditions of 50^{o}C and 50% formamide, the 50mer microarray was able to differentiate sequences with < 88% identities. In addition, Bozdech *et al*. (Bozdech et al., 2003) showed that a significant cross-hybridization could occur if a 70mer probe had free energy < -35 kcal/mol. Therefore, based on sequence identity, continuous sequence stretches and free energy, we have experimentally established probe design criteria (He et al., 2005b; Liebich et al., 2006). Furthermore, those criteria have been implemented in a novel software tool, *CommOligo*, for microarrays probe design (Li et al., 2005). Our GeoChip was designed using the newly developed software, and the evaluation results showed only a very small portion of false positives (0.002-0.004%) and no positive negatives (He et al., 2007). Therefore, a well designed FGA can achieve its desirable specificity.

Sensitivity is a critical parameter, particularly for environmental studies when biomass is low or for low-level molecules in individual cells. When cDNA-based (PCR-generated probes) functional gene arrays (FGA) were used, the detection limit for *nirS* (nitrite reductase) genes was approximately 1 ng of pure gDNA and 25 ng of soil community DNA (Wu et al., 2001). When oligonucleotide arrays with capture and detector probes were used, a previous study demonstrated that the detection sensitivity for *Geobacter chapellei* SSU rRNA gene sequences in soil extracts was approximately 500 ng of total RNA (Small et al., 2001). Recently, studies with a 50mer FGA showed that the detection limit for some functional genes could be 5 to 10 ng of pure gDNA and 50 to 100 ng in a mixture of gDNA from different organisms (Rhee et al, 2004; Tiquia et al, 2004). By taking advantage of the WCGA approach, the 50mer FGA can detect subnanogram quantities of microbial community DNAs as low as 10 pg (Wu et al., 2006). Therefore, the currently available technologies can generally obtain enough nucleic acids for FGA analysis, although challenges may still remain for those genes with low abundances in an environmental sample.

The quantitative capability of microarray-based technology is another central issue for environmental applications. Several previous studies showed that very good linear relationships were obtained between hybridization signal intensity and target DNA or RNA concentration from pure cultures, mixed cultures, and environmental samples (Rhee et al., 2004; Wu et al., 2006; Wu et al., 2001). Recently we showed that reliable quantification could be obtained using a 50mer FGA with randomly amplified DNAs (Wu et al., 2006), or randomly applied RNAs (Gao et al., 2007), as targets. Thus, it is expected that FGAs can serve as a quantitative tool to analyze environmental samples.

3.3 Data normalization and analysis

Microarray data normalization is necessary to adjust microarray data for effects that arise from variations in the microarray technology rather than biological differences between samples, or probes on an array. Microarray technology variations may be due to dye bias, labeling efficiency, different scanning properties and settings, and use of different reagents, which can be systematically corrected. Normalization can be performed within a chip and among replicate chips. However, a normalization of microarray data needs to consider the following aspects: (i) what percentage of spots have positive signals on the array for each slide since a microarray normally contains a comprehensive set of probes and some microbial communities may be very simple, or complex; (ii) what spots are used as the control for normalization; (iii) the distribution of signals among positive spots or all spots on the array; (iv) normalization methods may be different based on different situations mentioned above. The following is a simple method for analyzing digital array data output from image processing software (e.g. ImaGene, BioDiscovery Inc.). This method includes the following key steps: (i) poor-quality spots are removed, (ii) the signal intensity of each spot is normalized by calculating the mean intensity, (iii) spots with low signal intensities are removed based on the signal-to-noise ratio (SNR) (Wu et al., 2006) and normally, an SNR of 2.0 is generally used (Verdick et al., 2002), and (iv) for outlier removal, if any of replicates (slides) has (signal – mean) more than three times the standard deviation, this replicate is removed. This process continues until no such replicates are identified.

Data analysis is the most difficult task for microarray analysis of microbial communities. First, a massive amount of data is generated by microarray hybridization. Second, although many methods and software have been developed for microarray data analysis, most of them are focused on the analysis of gene expression data, especially two-dye microarray hybridization, and they may not be suitable for microarray data analysis. Third, microarray data generally have large variations, and rigorous statistical analysis of such data is needed (He and Zhou, 2008). Statistical analysis of microarray data is complicated. Finally, our ultimate goal is to extract biological insights from such large data sets to understand microbial structure and function, so the results should be of biological relevance and statistical significance.

The following statistical approaches are commonly used to analyze microarray data: (1) *Scatter plot.* This is the simplest way to visualize microarray data. Scatter plots can display signal intensities for a single chip with two-dye hybridization, or for two chips with one-dye hybridization. (2) *Principal component analysis (PCA).* This is an exploratory multivariate statistical method for simplifying data sets that reduces the dimensionality of the variables by finding new variables, which are independent of each other. A few of the new variables, typically 2-3, are selected to explain the majority of variance in the original data. For microarray data analysis, genes or experiments can be considered as variables. The main advantage of PCA is that it identifies outliers in the data or genes that behave differently than most of the genes across a set of experiments. (3) *Cluster analysis.* One of the most commonly used methods is cluster analysis. Cluster analysis is used to identify groups of genes, or clusters that have similar profiles. Clusters and the genes within them can be subsequently examined for commonalities in functions and sequences for better understanding of how and why they behave similarly. Cluster analysis can help establish functionally related groups of genes to gain insights into structure and function of a given microbial community. A popular clustering method was developed by Eisen et al. (Eisen et al., 1998), and other algorithms were described by Heyer et al. (Heyer et al., 1999), Travazoie and Church (Tavazoie and Church, 1998), and Zhou *et al*. (Zhou et al., 2000). (4) *Neural network analysis.* Since clustering methods have some serious drawbacks in dealing with data with a significant amount of noise, a fundamentally different network-based approach has been proposed for microarray data analysis (Herrero et al., 2001; Tamayo et al., 1999; Toronen et al., 1999). Unsupervised neural networks, such as self-organizing maps (SOMs), are a more robust and accurate method for grouping large data sets. The main advantage of SOMs is that they are robust to noise, and SOMs are also reasonably fast and can be easily scaled up to large data sets. One disadvantage of SOMs is that they require pre-determined choices about geometry. In addition, it is very difficult to detect higher-order relationships between clusters of profiles due to the lack of a tree structure (Herrero et al., 2001). To overcome some of the limitations of SOMs, an unsupervised neural network, termed the self-organizing tree algorithm (SOTA), was proposed (Dopazo and Carazo, 1997). This new algorithm combines the advantages of hierarchical clustering (tree topology) and neural network (accuracy and robustness) and was used to analyze gene expression data (Herrero et al., 2001).

There are many commercial and free software tools available for general microarray data analysis. Such tools are as simple as Excel (Microsoft), or as complicated as Matlab (The MathWorks), and those include R, GeneSpring (Agilent Technologies), Genesight (BisDiscovery), S-Plus (Insightful Corporation), SPSS (SPSS Inc.), SAS (SAS Institute Inc.), and SAM (Tusher et al., 2001). It is noted that most currently available tools are focused on the analysis of gene expression data. Therefore, we need to carefully choose suitable tools for microarray data analysis. For example, to examine the correlations between the differences of uranium concentrations and those of various functional gene abundances, we used R to implement the Mantel test (He et al., 2007).

Finally, such large data sets need to be simply presented and biologically interpreted. Richness of different gene categories in the community as a whole in the studied samples can be determined from the number of probes that detect their target(s). With probes by category (e.g. *nifH*) as indicators of individual taxa, Simpson's diversity index, Shannon diversity index, and Evenness based on Simpson's index can be calculated as described previously (Begon et al., 1996). To compare different samples, some genes are specifically detected in one sample, and some in all samples tested. The numbers of those two type of genes can be calculated as unique and overlap genes as described previously (Wu et al., 2006). Clustering is one of the most popular methods to analyze and visualize microarray data. Using a hierarchical clustering algorithm, the relationship between different samples taken at different times/sites and different clusters among those samples can be identified. Such analysis can be also applied to each gene/category with its variants. The software *Cluster* can be used for cluster analysis and *TreeView* for visualization (Eisen et al., 1998). In addition, the network analysis of microarray data has received significant attention, and such a method may be used to present and interpret microarray results.

3. 4 Example applications of functional gene arrays for microbial community analysis

FGAs have been extensively evaluated with artificial microbial communities (He et al., 2007; Loy et al., 2004; Rhee et al., 2004; Steward et al., 2004; Taroncher-Oldenburg et al., 2003; Tiquia et al., 2004; Wu et al., 2006; Wu et al., 2001) and applied to investigate microbial communities in natural and contaminated environments (He et al., 2007). Since all probes were designed using functional gene coding sequences, both DNA and RNA can be used as targets for measuring gene abundance and gene expression, respectively. Therefore, FGAs can be used in a variety of studies, including (but not limiting): (1) detection of functional genes and/or organisms in a particular environment; (2) linking microbial structures to function; and (3) estimation of gene abundance and activity. The GeoChip can be used for analysis of any environmental sample, including soil, water, sediments, oil fields, deep sea, animal guts, etc.

Here, an example study is presented for using an FGA (with 2006 50mer oligonucleotide probes) to analyze the structure and activity of microbial communities in groundwater (Wu et al., 2006). The Environmental Remediation Science Program (ERSP) field research center (FRC) sites are contaminated with nitrate, uranium, and technetium, as well as some residual organic compounds. Samples from five wells with different degrees of contamination were collected at such a site in Oak Ridge, TN. Microarray analyses showed more than 400 genes had positive hybridization signals (Table 2). As expected, the highest number of genes was detected for uncontaminated background samples (well FW300), while the lowest number of genes was detected for the highly contaminated sample (well FW010) (Table 2). Simpson's diversity indices showed that the diversities in the uncontaminated background well (FW300) and less contaminated well (FW003) were much higher than those in more heavily contaminated wells (FW021, FW024, and FW010), suggesting contaminants strongly affected the microbial communities (Table 2). The proportion of overlapping genes in different samples was consistent with the contaminant level and geochemistry (Table 2).

Some important genes involved in denitrification (e.g., *nosZ* and *nirS*), degradation of organic contaminants (e.g., dienelactone hydrolase genes), and metal resistance (e.g., mercuric reductase gene) were observed in all samples (Fig. 5), suggesting that the microbial populations containing these genes are widespread. Dissimilatory sulfate-reducing bacteria are important in the reduction of uranium from soluble U(VI) to insoluble U(IV). In contrast to the results described above, while some dissimilatory sulfite-reducing organisms (*dsrAB*) were found in all of the samples (group D), the abundance and presence of most types (groups A, B, and C) seemed to vary with the origin of the sample (Fig. 5). The above results are consistent with 16S gene phylogenetic analysis, and recent microbial community sequencing data. Therefore, as a specific, sensitive and high-throughput tool, the FGA technology is capable of revealing biological processes of microbial communities in natural/contaminated environmental systems.

The GeoChip 2.0 has also been successfully used in our laboratory for tracking the dynamics of metal-reducing bacteria and associated communities for an *in situ* bioremediation study at the ERSP Oak Ridge FRC (Field Research Center) site, which is the first time to demonstrate that uranium can be bioremediated to the concentrations below the USA EPA maximum contaminant level (MCL) for drinking water (He et al., 2007). In addition, a FGA has been recently used to characterize pure isolates, analyze soil microbial nitrogen and carbon cycles along a south polar latitudinal gradient (Yergeau et al., 2007), to investigate microbial community structure during bioremediation of a hydrocarbon-contaminated aquifer, and to identify active members in a stable isotope experiment fed with labeled biphenyl. The GeoChip 2.0 has also been used to examine the gene-area relationship of microbial communities in soils, and the results suggest that a forest soil microbial community exhibited a relatively flat gene–area relationship, but the z (z is a measurement of the rate of species turnover across space in a power-law relationship: $S = cA^z$, where S is the number of species, A is the area, and c is the intercept in log-log space) values varied considerably across different functional and phylogenetic groups (Zhou et al., 2008). All resulting data from these studies indicate that the functional gene array technology is a powerful tool for studying microbial communities in the environment.

4. CONCLUDING REMARKS

We are entering a new era in environmental microbiology where we are beginning to understand some of the complex and interdependent relationships among prokaryotes in natural communities. Starting with an acid mine drainage (AMD) community (Tyson et al., 2004) and working up in complexity to the Sargasso Sea (Venter et al., 2004) and a Minnesota farm soil

(Tringe et al., 2005), shotgun microbial community sequencing (metagenomic sequencing) has implicated taxa and specific genes that contribute to the overall cellular respiration and other important functions of a microbial community (See Chapter x). Phylogenetic arrays such as the PhyloChip that target the known diversity within bacteria and archaea are important for determining the composition of microbial communities in a number of different environments and conditions. Functional gene arrays such as the GeoChip are important for identifying physiological activities involved in certain biogeochemical processes.

The development and application of microarray technology for environmental studies has received a great deal of attention. Because of its high-density and high-throughput capacity, it is expected that such a technology will revolutionize the analyses of microbial community structure, function and dynamics. Current studies showed that microarray technology is able to provide specific, sensitive, and potentially quantitative analysis of microbial communities from a variety of natural environments. It is also useful for providing direct linkages of microbial genes/populations to ecosystem processes and functions. Unlike more costly and time consuming methods such as metagenomic, or 16S clone library sequencing, microarrays allow high throughput sampling so that multiple replications and multiple treatments can be examined. Identifying the population dynamics of specific taxonomic groups in response to defined conditions is the first step towards understanding how they may contribute to the overall microbial community structure.

However, more rigorous and systematic assessment and development are needed to realize the full potential of microarrays for microbial ecology studies. Several key issues need to be addressed, including novel experimental designs and strategies for minimizing inherent high hybridization variations to improve FGA-based quantitative accuracy, novel approaches for increasing hybridization sensitivity to detect extremely low biomass in natural environments, novel computational tools for microarray data interpretation, and broad integration and application of microarray technologies with environmental studies to address ecological and environmental questions and hypotheses. Therefore, the future focuses will be not only on microarray technology and applications, but also on microarray-related data analysis, interpretation and modeling. First, novel strategies and approaches for experimental controls and design are needed to ensure that microarray hybridization data from different samples are comparable, interpretable and biologically significant because of the inherent variability in microarray hybridization signals. Second, automatic bioinformatic and computational tools are necessary to retrieve gene sequences, design oligonucleotide probes, construct databases, and update sequence, probe and array information. In addition, more advanced automatic mathematical and statistical approaches, such as multivariate analysis, time-series analysis, neural network, artificial intelligence, and differential equation-based modeling, should be extremely useful for rapid pattern recognition, visualization, data mining, cellular modeling, simulation and prediction.

Acknowledgment

We would like to acknowledge funding support from the Director, Office of Science, Office of Biological and Environmental Research, Environmental Remediation Sciences Program, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

Table 1. List of major functional markers on the GeoChip 2.0

Gene category	***Example of key enzyme (gene)***	***Total_probe#***
Nitrogen cycling		**5310**
Nitrogen fixation	Nitrogenase (*nifH*)	1225
Denitrification	Nitrate reductase (*narG*, *napA*, *nasA*), nitrite reductase (*nirS*, *nirK*), nitric oxide reductase (*norB*), nitrous oxide reductase (*nosZ*)	2306
Nitrification	Ammonium monooxygenase (*amoA*), hydroxylamine oxidoreductase (*hao*)	347
Nitrogen mineralization	Urease (*ureC*), glutamate dehydrogenase (*gdh*)	1432
Carbon cycling		**4599**
Carbon fixation	Rubisco (*cbbL*, *rbcL*), Acl (*aclB*), CODH, FTHFS	1018
Cellulose degradation	Cellulase, endoglucanase	1285
Lignin degradation	Laccase, mannanase	513
Chitin degradation	Endochitinase (*chiA*), exochitinase	744
Methane production	Methyl coenzyme M reductase (*mcrA*)	437
Methane oxidation	Methane monooxygenase (*pmoA*)	336
Others	Lignin peroxidase (*lip*), pectinase, cellobiase	266
Sulfate reduction		**1615**
	Sulfite reductase (*dsrA/B*), APS (*apsA*)	1615
Phosphorus utilization		**145**
	Exopolyphosphatase (*ppx*), phytase	145
Metal reduction and resistance		**4546**
Arsenic resistance	Arsenate reductase (arsC, *arsB*, *arsC*)	877
Cadmium resistance	Cadmium transporter (*cadA*, *cadB*, *cadC*)	282
Chromium resistance	Chromium/chromate transporter (*chrA*)	319
Mercury resistance/reduction	Mercuric ion reductase/transporter (*mer*, *merA*, *merB*)	548
Nickel resistance	Nickel transporter (*nccA*), permease (*nreB*)	140
Zinc resistance	Zinc resistance protein (*zntA*)	128
Other metal resistance/reduction	cobalt resistance proteins, selenium reductase, etc.	2252
Contaminant degradation		**8028**
Benzene, toluene, ethylbenzene, and xylene (BTEX) & related aromatics	Benzene 1,2-dioxygenase (*ben*), ethylbenzene dehydrogenase (*ebd*), benzylsuccinate synthase (*bss*), xylene monooxygenase (*xyl*), benzoyl-CoA reductase (*bad*), and catechol 1,2-dioxygenase (*cat*, *tfd*).	4176
Chlorinated aromatics	Chlorophenol reductive dehalogenase (*cpr*)	90
Nitroaromatics	Nitrobenzene nitroreductase (*nbz*), 4-nitrobenzaldehyde dehydrogenase (*ntn*), p-nitrobenzoate reductase (*pnb*)	152
Polycyclic aromatic hydrocarbons (PAHs)	Naphthalene dioxygenase (*nah*), PAH ring-hydroxylating dioxygenase (*pdo*)	741
Polychlorinated biphenyls (PCBs)	Biphenyl dioxygenase (*bph*)	388
Chlorinated solvents (e.g. PCE)	PCE/TCE reductive dehalogenase (*rdh*, *pceA*, *tecA*)	232
Other organic compounds/by-products	Alkane hydroxylase (*alk*), homogentisate 1,2-dioxygenase (*hmg*), vanillate O-demethylase oxygenase (*van*), etc.	2249
Total	***19959***	***24243***

References

Allison, D. B., Page, G. P., Beasley, T. M., and Edwards, J. W. (2007). A Review of: "DNA Microarrays and Related Genomics Techniques: Design, Analysis, and Interpretation of Experiments". Journal of Biopharmaceutical Statistics *17*, 187-190.

Anderson, R. T., Vrionis, H. A., Ortiz-Bernad, I., Resch, C. T., Long, P. E., Dayvault, R., Karp, K., Marutzky, S., Metzler, D. R., Peacock, A., *et al.* (2003). Stimulating the in situ activity of Geobacter species to remove uranium from the groundwater of a uranium-contaminated aquifer. Appl Environ Microbiol *69*, 5884-5891.

Ashelford, K. E., Chuzhanova, N. A., Fry, J. C., Jones, A. J., and Weightman, A. J. (2005). At least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies. Appl Environ Microbiol *71*, 7724-7736.

Ashelford, K. E., Chuzhanova, N. A., Fry, J. C., Jones, A. J., and Weightman, A. J. (2006). New Screening Software Shows that Most Recent Large 16S rRNA Gene Clone Libraries Contain Chimeras. Appl Environ Microbiol *72*, 5734-5741.

Begon, M., Harper, J. L., and Townsend, C. R. (1996). Ecology: individuals, populations, and communities, Third edn: Blackwell Science).

Bozdech, Z., Zhu, J., Joachimiak, M. P., Cohen, F. E., Pulliam, B., and DeRisi, J. L. (2003). Expression profiling of the schizont and trophozoite stages of Plasmodium falciparum with a long-oligonucleotide microarray. Genome Biol *4*, R9.

Braker, G., Zhou, J., Wu, L., Devol, A. H., and Tiedje, J. M. (2000). Nitrite reductase genes (nirK and nirS) as functional markers to investigate diversity of denitrifying bacteria in pacific northwest marine sediment communities. Appl Environ Microbiol *66*, 2096-2104.

Brodie, E. L., DeSantis, T. Z., Joyner, D. C., Baek, S. M., Larsen, J. T., Andersen, G. L., Hazen, T. C., Richardson, P. M., Herman, D. J., Tokunaga, T. K., *et al.* (2006). Application of a High-Density Oligonucleotide Microarray Approach To Study Bacterial Population Dynamics during Uranium Reduction and Reoxidation. Appl Environ Microbiol *72*, 6288-6298.

Bryant, M. P., Campbell, L. L., Reddy, C. A., and Crabill, M. R. (1977). Growth of Desulfovibrio in Lactate or Ethanol Media Low in Sulfate in Association with H2-Utilizing Methanogenic Bacteria. Appl Environ Microbiol *33*, 1162-1169.

Castiglioni, B., Rizzi, E., Frosini, A., Sivonen, K., Rajaniemi, P., Rantala, A., Mugnai, M. A., Ventura, S., Wilmotte, A., Boutte, C., *et al.* (2004). Development of a universal microarray based on the ligation detection reaction and 16S rrna gene polymorphism to target diversity of cyanobacteria. Appl Environ Microbiol *70*, 7161-7172.

Chang, Y. J., Long, P. E., Geyer, R., Peacock, A. D., Resch, C. T., Sublette, K., Pfiffner, S. M., Smithgall, A., Anderson, R. T., Vrionis, H. A., *et al.* (2005). Microbial incorporation of C-13-

labeled acetate at the field scale: Detection of microbes responsible for reduction of U(VI). Environmental Science & Technology *39*, 9039-9048.

Chee, M., Yang, R., Hubbell, E., Berno, A., Huang, X. C., Stern, D., Winkler, J., Lockhart, D. J., Morris, M. S., and Fodor, S. P. (1996). Accessing genetic information with high-density DNA arrays. Science *274*, 610-614.

Cole, J. R., Chai, B., Farris, R. J., Wang, Q., Kulam, S. A., McGarrell, D. M., Garrity, G. M., and Tiedje, J. M. (2005). The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis. Nucleic Acids Res *33*, D294-296.

Cole, K., Truong, V., Barone, D., and McGall, G. (2004). Direct labeling of RNA with multiple biotins allows sensitive expression profiling of acute leukemia class predictor genes. Nucleic Acids Research *32*, -.

Cook, A. F., Vuocolo, E., and Brakel, C. L. (1988). Synthesis and hybridization of a series of biotinylated oligonucleotides. Nucleic Acids Res *16*, 4077-4095.

Curtis, T. P., and Sloan, W. T. (2005). Microbiology. Exploring microbial diversity--a vast below. Science *309*, 1331-1333.

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.

DeSantis, T. Z., Hugenholtz, P., Keller, K., Brodie, E. L., Larsen, N., Piceno, Y. M., Phan, R., and Andersen, G. L. (2006a). NAST: a multiple sequence alignment server for comparative analysis of 16S rRNA genes. Nucleic Acids Res *34*, W394-399.

DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., Huber, T., Dalevi, D., Hu, P., and Andersen, G. L. (2006b). Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB. Appl Environ Microbiol *72*, 5069-5072.

Dopazo, J., and Carazo, J. M. (1997). Phylogenetic reconstruction using an unsupervised growing neural network that adopts the topology of a phylogenetic tree. J Mol Evol *44*, 226-233.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res *32*, 1792-1797.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A *95*, 14863-14868.

Fields, M. W., Schryver, J. C., Brandt, C. C., Yan, T., Zhou, J. Z., and Palumbo, A. V. (2006). Confidence intervals of similarity values determined for cloned SSU rRNA genes from environmental samples. J Microbiol Methods *65*, 144-152.

Gans, J., Wolinsky, M., and Dunbar, J. (2005). Computational improvements reveal great bacterial diversity and high metal toxicity in soil. Science *309*, 1387-1390.

Gao, H., Yang, Z. K., Gentry, T. J., Wu, L., Schadt, C. W., and Zhou, J. (2007). Microarray-based analysis of microbial community RNAs by whole-community RNA amplification. Appl Environ Microbiol *73*, 563-571.

Ginder-Vogel, M., Criddle, C. S., and Fendorf, S. (2006). Thermodynamic constraints on the oxidation of biogenic UO2 by Fe(III) (hydr) oxides. Environmental Science & Technology *40*, 3544-3550.

He, Z., Gentry, T. J., Schadt, C. W., Wu, L., Liebich, J., Chong, S. C., Huang, Z., Wu, W., Gu, B., Jardine, P., *et al.* (2007). GeoChip: A comprehensive microarray for investigating biogeochemical, ecological, and environmental processes. ISME J *1*, 67-77.

He, Z., L. Wu, L., Fields, M. W., and J., Z. (2005a). Comparison of microarray performance with different probe sizes for monitoring gene expression. Appl Environ Microbiol *71*, 5154-5162.

He, Z., Wu, L., Li, X., Fields, M. W., and Zhou, J. (2005b). Empirical establishment of oligonucleotide probe design criteria. Appl Environ Microbiol *71*, 3753-3760.

He, Z., and Zhou, J. (2008). Empirical evaluation of a new method for calculating signal-to-noise ratio for microarray data analysis. Appl Environ Microbiol *74*, 2957-2966.

Herrero, J., Valencia, A., and Dopazo, J. (2001). A hierarchical unsupervised growing neural network for clustering gene expression patterns. Bioinformatics *17*, 126-136.

Heyer, L. J., Kruglyak, S., and Yooseph, S. (1999). Exploring expression data: identification and analysis of coexpressed genes. Genome Res *9*, 1106-1115.

Huber, T., Faulkner, G., and Hugenholtz, P. (2004). Bellerophon: a program to detect chimeric sequences in multiple sequence alignments. Bioinformatics *20*, 2317-2319.

Hugenholtz, P. (2002). Exploring prokaryotic diversity in the genomic era. Genome Biol *3*, 1-8.

Hurt, R. A., Qiu, X., Wu, L., Roh, Y., Palumbo, A. V., Tiedje, J. M., and Zhou, J. (2001). Simultaneous recovery of RNA and DNA from soils and sediments. Appl Environ Microbiol *67*, 4495-4503.

Huse, S. M., Dethlefsen, L., Huber, J. A., Welch, D. M., Relman, D. A., and Sogin, M. L. (2008). Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing. PLoS Genet *4*, e1000255.

Kelly, J. J., Chernov, B. K., Tovstanovsky, I., Mirzabekov, A. D., and Bavykin, S. G. (2002). Radical-generating coordination complexes as tools for rapid and effective fragmentation and fluorescent labeling of nucleic acids for microchip hybridization. Anal Biochem *311*, 103-118.

Kelly, J. J., Siripong, S., McCormack, J., Janus, L. R., Urakawa, H., El Fantroussi, S., Noble, P. A., Sappelsa, L., Rittmann, B. E., and Stahl, D. A. (2005). DNA microarray detection of nitrifying bacterial 16S rRNA in wastewater treatment plant samples. Water Res *39*, 3229-3238.

Klerks, M. M., van Bruggen, A. H., Zijlstra, C., and Donnikov, M. (2006). Comparison of methods of extracting Salmonella enterica serovar Enteritidis DNA from environmental substrates and quantification of organisms by using a general internal procedural control. Appl Environ Microbiol *72*, 3879-3886.

Lehner, A., Loy, A., Behr, T., Gaenge, H., Ludwig, W., Wagner, M., and Schleifer, K. H. (2005). Oligonucleotide microarray for identification of Enterococcus species. FEMS Microbiol Lett *246*, 133-142.

Li, X., He, Z., and Zhou, J. (2005). Selection of optimal oligonucleotide probes for microarrays using multiple criteria, global alignment and parameter estimation. Nucleic Acids Res *33*, 6114-6123.

Liebich, J., Schadt, C. W., Chong, S. C., He, Z., Rhee, S. K., and Zhou, J. (2006). Improvement of oligonucleotide probe design criteria for functional gene microarrays in environmental applications. Appl Environ Microbiol *72*, 1688-1691.

Liu, W. T., Mirzabekov, A. D., and Stahl, D. A. (2001). Optimization of an oligonucleotide microchip for microbial identification studies: a non-equilibrium dissociation approach. Environ Microbiol *3*, 619-629.

Loring, J. F. (2006). Evolution of microarray analysis. Neurobiol Aging *27*, 1084-1086.

Loy, A., Kusel, K., Lehner, A., Drake, H. L., and Wagner, M. (2004). Microarray and functional gene analyses of sulfate-reducing prokaryotes in low-sulfate, acidic fens reveal cooccurrence of recognized genera and novel lineages. Appl Environ Microbiol *70*, 6998-7009.

Ludwig, W., Strunk, O., Westram, R., Richter, L., Meier, H., Yadhukumar, Buchner, A., Lai, T., Steppi, S., Jobb, G., *et al.* (2004). ARB: a software environment for sequence data. Nucleic Acids Res *32*, 1363-1371.

McGrath, K. C., Thomas-Hall, S. R., Cheng, C. T., Leo, L., Alexa, A., Schmidt, S., and Schenk, P. M. (2008). Isolation and analysis of mRNA from environmental microbial communities. J Microbiol Methods *75*, 172-176.

Pace, N. R. (1997). A molecular view of microbial diversity and the biosphere. Science *276*, 734-740.

Peacock, A. D., Chang, Y. J., Istok, J. D., Krumholz, L., Geyer, R., Kinsall, B., Watson, D., Sublette, K. L., and White, D. C. (2004). Utilization of microbial biofilms as monitors of bioremediation. Microb Ecol *47*, 284-292.

Phelps, T. J., Murphy, E. M., Pfiffner, S. M., and White, D. C. (1994). Comparison between geochemical and biological estimates of subsurface microbial activities. Microbial Ecology *V28*, 335-349.

Reardon, C. L., Cummings, D. E., Petzke, L. M., Kinsall, B. L., Watson, D. B., Peyton, B. M., and Geesey, G. G. (2004). Composition and diversity of microbial communities recovered from surrogate minerals incubated in an acidic uranium-contaminated aquifer. Appl Environ Microbiol *70*, 6037-6046.

Rhee, S. K., Liu, X., Wu, L., Chong, S. C., Wan, X., and Zhou, J. (2004). Detection of genes involved in biodegradation and biotransformation in microbial communities by using 50-mer oligonucleotide microarrays. Appl Environ Microbiol *70*, 4303-4317.

Rotthauwe, J. H., Witzel, K. P., and Liesack, W. (1997). The ammonia monooxygenase structural gene amoA as a functional marker: molecular fine-scale analysis of natural ammonia-oxidizing populations. Appl Environ Microbiol *63*, 4704-4712.

Schadt, C. W., Liebich, J., Chong, S. C., Gentry, T. J., He, Z., Pan, H., and Zhou, J. (2004). Design and use of functional gene microarrays (FGAs) for the characterization of microbial communities, In Methods in Microbiology, T. Savidge, and H. Pothulakis, eds. (London: Academic Press), pp. 329-365.

Small, J., Call, D. R., Brockman, F. J., Straub, T. M., and Chandler, D. P. (2001). Direct detection of 16S rRNA in soil extracts by using oligonucleotide microarrays. Appl Environ Microbiol *67*, 4708-4716.

Steward, G. F., Jenkins, B. D., Ward, B. B., and Zehr, J. P. (2004). Development and testing of a DNA macroarray to assess nitrogenase (nifH) gene diversity. Appl Environ Microbiol *70*, 1455-1465.

Stoughton, R. B. (2005). Applications of DNA microarrays in biology. Annu Rev Biochem *74*, 53-82.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E. S., and Golub, T. R. (1999). Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proc Natl Acad Sci U S A *96*, 2907-2912.

Taroncher-Oldenburg, G., Griner, E. M., Francis, C. A., and Ward, B. B. (2003). Oligonucleotide microarray for the study of functional gene diversity in the nitrogen cycle in the environment. Appl Environ Microbiol *69*, 1159-1171.

Tavazoie, S., and Church, G. M. (1998). Quantitative whole-genome analysis of DNA-protein interactions by in vivo methylase protection in E. coli. Nat Biotechnol *16*, 566-571.

Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res *22*, 4673-4680.

Tiquia, S. M., Wu, L., Chong, S. C., Passovets, S., Xu, D., Xu, Y., and Zhou, J. (2004). Evaluation of 50-mer oligonucleotide arrays for detecting microbial populations in environmental samples. Biotechniques *36*, 664-670, 672, 674-665.

Toronen, P., Kolehmainen, M., Wong, G., and Castren, E. (1999). Analysis of gene expression data using self-organizing maps. FEBS Lett *451*, 142-146.

Tringe, S. G., von Mering, C., Kobayashi, A., Salamov, A. A., Chen, K., Chang, H. W., Podar, M., Short, J. M., Mathur, E. J., Detter, J. C., *et al.* (2005). Comparative metagenomics of microbial communities. Science *308*, 554-557.

Tusher, V. G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A *98*, 5116-5121.

Tyson, G. W., Chapman, J., Hugenholtz, P., Allen, E. E., Ram, R. J., Richardson, P. M., Solovyev, V. V., Rubin, E. M., Rokhsar, D. S., and Banfield, J. F. (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature *428*, 37-43.

Urakawa, H., Noble, P. A., El Fantroussi, S., Kelly, J. J., and Stahl, D. A. (2002). Single-base-pair discrimination of terminal mismatches by using oligonucleotide microarrays and neural network analyses. Appl Environ Microbiol *68*, 235-244.

Vanbelkum, A., Linkels, E., Jelsma, T., Vandenberg, F. M., and Quint, W. (1994). Nonisotopic Labeling of DNA by Newly Developed Hapten-Containing Platinum Compounds. Biotechniques *16*, 148-&.

Venter, J. C., Remington, K., Heidelberg, J. F., Halpern, A. L., Rusch, D., Eisen, J. A., Wu, D., Paulsen, I., Nelson, K. E., Nelson, W., *et al.* (2004). Environmental genome shotgun sequencing of the Sargasso Sea. Science *304*, 66-74.

Verdick, D., Handran, S., and Pickett, S. (2002). Key considerations for accurate microarray scanning and image analysis, In DNA Array Image Analysis: Nuts and Bolts, G. Kamberova, ed. (Salem, MA: DNA Press), pp. 83-98.

Wan, J., Tokunaga, T. K., Brodie, E., Wang, Z., Zheng, Z., Herman, D., Hazen, T. C., Firestone, M. K., and Sutton, S. R. (2005). Reoxidation of bioreduced uranium under reducing conditions. Environ Sci Technol *39*, 6162-6169.

Warsen, A. E., Krug, M. J., LaFrentz, S., Stanek, D. R., Loge, F. J., and Call, D. R. (2004). Simultaneous discrimination between 15 fish pathogens by using 16S ribosomal DNA PCR and DNA microarrays. Appl Environ Microbiol *70*, 4216-4221.

Whitman, W. B., Coleman, D. C., and Wiebe, W. J. (1998). Prokaryotes: the unseen majority. Proc Natl Acad Sci U S A *95*, 6578-6583.

Wilson, W. J., Strout, C. L., DeSantis, T. Z., Stilwell, J. L., Carrano, A. V., and Andersen, G. L. (2002). Sequence-specific identification of 18 pathogenic microorganisms using microarray technology. Mol Cell Probes *16*, 119-127.

Woese, C. R., Fox, G. E., Zablen, L., Uchida, T., Bonen, L., Pechman, K., Lewis, B. J., and Stahl, D. (1975). Conservation of primary structure in 16S ribosomal RNA. Nature *254*, 83-86.

Wu, L., Liu, X., Schadt, C. W., and Zhou, J. (2006). Microarray-based analysis of subnanogram quantities of microbial community DNAs by using whole-community genome amplification. Appl Environ Microbiol *72*, 4931-4941.

Wu, L., Thompson, D. K., Li, G., Hurt, R. A., Tiedje, J. M., and Zhou, J. (2001). Development and evaluation of functional gene arrays for detection of selected genes in the environment. Appl Environ Microbiol *67*, 5780-5790.

Yergeau, E., Kang, S., He, Z., Zhou, J., and Kowalchuk, G. A. (2007). Functional microarray analysis of nitrogen and carbon cycling genes across an Antarctic latitudinal transect. (in preparation).

Zhang, Z., Schwartz, S., Wagner, L., and Miller, W. (2000). A greedy algorithm for aligning DNA sequences. J Comput Biol *7*, 203-214.

Zhou, J. (2003). Microarrays for bacterial detection and microbial community analysis. Curr Opin Microbiol *6*, 288-294.

Zhou, J., Bruns, M. A., and Tiedje, J. M. (1996). DNA recovery from soils of diverse composition. Appl Environ Microbiol *62*, 316-322.

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

Zhou, Y., Kalocsai, P., Chen, J., and Shams, S. (2000). Information processing issues and solutions associated with microarray technology, In Microarray Biochip Technology, M. Schena, ed. (Natick, MA: Eaton Publishing), pp. 167–200.