

Development and application of functional gene arrays for microbial community analysis

Z. L. HE, J. D. VAN NOSTRAND, L. Y. WU, J. Z. ZHOU

Institute for Environmental Genomics, Department of Botany and Microbiology,
University of Oklahoma, Norman, OK 73019, USA

Received 20 September 2008; accepted 5 November 2008

Abstract: Functional gene markers can provide important information about functional gene diversity and potential activity of microbial communities. Although microarray technology has been successfully applied to study gene expression for pure cultures, simple, and artificial microbial communities, adapting such a technology to analyze complex microbial communities still presents a lot of challenges in terms of design, sample preparation, and data analysis. This work is focused on the development and application of functional gene arrays (FGAs) to target key functional gene markers for microbial community studies. A few key issues specifically related to FGAs, such as oligonucleotide probe design, nucleic acid extraction and purification, data analysis, specificity, sensitivity, and quantitative capability are discussed in detail. Recent studies have demonstrated that FGAs can provide specific, sensitive, and potentially quantitative information about microbial communities from a variety of natural environments and controlled ecosystems. This technology is expected to revolutionize the analysis of microbial communities, and link microbial structure to ecosystem functioning.

Key words: microarray; application; function gene marker; microbial community

1 Introduction

A great difficulty in studying microbial communities is that a majority (>99%) of microorganisms are as yet uncultivated[1]. As such, the study of microbial composition, structure, function, and dynamics in natural and/or controlled environments requires the use of molecular markers. Most studies use either the 16S rRNA or the *gyrB* genes as phylogenetic markers to identify microorganisms and their phylogenetic relationships in microbial communities of interest. Another option for molecular markers is the use of functional genes, such as *amoA*, which encodes ammonia monooxygenase, a key enzyme for ammonia oxidization[2], and *nifH*, which encodes dinitrogenase reductase, a key enzyme in nitrogen fixation[3–5]. Additional functional genes that are currently being used in environmental studies are described in Ref.[6]. A great benefit in the use of functional gene markers rather

than phylogenetic markers is that the functional gene diversity and functional potential of microbial communities can be evaluated. However, though conventional molecular methods, such as PCR-based cloning and *in situ* hybridization can provide information on microbial community diversity, structure and function, they cannot provide a complete picture of microbial activity and dynamics in a rapid, parallel, and high-through-put manner.

Microarrays provide a rapid mechanism for high-through-put analysis of microbial communities. This technology has been used successfully with pure cultures in the analysis of global gene expression. However, fully adapting this technology for use with environmental samples still presents many challenges in terms of probe design, gene sequence coverage, specificity, sensitivity and quantitative capability. To overcome these obstacles, functional gene arrays(FGAs) have been developed. This type of array targets microbial genes involved in key biogeochemical processes, such as

carbon, nitrogen, phosphorus, and sulfur cycling, organic contaminant degradation, and metal reduction, oxidation and resistance. Since most of processes targeted by the FGA are geochemical processes, these arrays are also called GeoChip[7]. GeoChip is a powerful tool which can be used to address several fundamental questions related to microbial ecology, biogeochemistry, and environmental biology. These include: 1) What functional processes is a microbial community capable of performing? 2) What are the dominant populations within a microbial community? 3) What is the dynamics of functional activity for a given microorganism, gene, or process? 4) How are microbial community structure and composition linked to ecosystem functioning? 5) What are the relationships between functional gene activity and environmental geochemistry? 6) How do environmental stressors and other environmental fluctuations affect microbial communities?

Since the first report of a microarray designed to examine gene expression in *Arabidopsis thaliana*[8], microbial ecologists have tried to answer the central question of whether FGA-based technologies can provide specific, sensitive and quantitative detection of microbial populations and activities within the context of environmental applications. FGA-based technology has evolved over time to include increased coverage of functional gene sequences, improved bioinformatics systems for sequence retrieval, probe selection, data analysis, and information storage. The first prototype FGA was constructed using 89 PCR-amplicon probes from pure cultures and laboratory clones targeting nitrogen cycling genes[9]. However, the use of PCR amplicon probes complicates the design of a comprehensive FGA since a large diversity of environmental clones and bacterial strains would be required. To overcome this difficulty, the use of oligonucleotide probes has become popular in microarrays due to their high specificity, ease of construction, and low cost. Oligonucleotide FGAs are fabricated with synthetic oligonucleotide probes and can be spotted onto glass slides[10–12] or nylon membranes [13]. FGAs have been systematically evaluated to determine the sensitivity, specificity, and quantitative capability. In addition to providing information regarding the functional abilities of microbial communities, FGAs can also be used to examine phylogenetic relationships. For example, LOY et al[14] constructed an array containing both phylogenetic gene (16S rRNA gene) and functional gene (*dsrA/B*) markers.

All these prototype or evaluation versions of FGAs could be designated by GeoChip 1.0. A newer, more comprehensive version, GeoChip 2.0, has been designed

and evaluated[7]. This version was designed to provide a comprehensive tool for the analysis of microbial community structure, function, and dynamics in natural, contaminated, or/and controlled environments. The gene sequence coverage and probe number have been greatly increased from hundreds to tens of thousands. In addition, the sequence retrieval, probe design, data analysis and other related computational techniques have been greatly improved. GeoChip 2.0 contains 24 243 oligonucleotide (50-mer) probes and covers >1 000 genes in >150 functional groups involved in nitrogen, carbon, sulfur and phosphorus cyclings, metal reduction and resistance, and organic contaminant degradation[7]. This array used experimentally established probe design criteria[15–16] and a new computational software tool, CommOligo[17], for oligonucleotide probe selection. Further, a complementary whole-community genome amplification (WCGA) technique has been demonstrated which allows for the analysis of sub-nanogram quantities of microbial community DNA[18]. This approach makes it possible to analyze environmental samples with very low biomass which greatly expands the number and type of samples that can be analyzed with the GeoChip. The next generation FGA, GeoChip 3.0, has been developed and is currently being evaluated. GeoChip 3.0 provides an expanded coverage, approximately 47 000 gene variants (sequences) from more than 290 gene categories. Many new features, including automated sequence retrieval, sequence verification, and sequence updates have been implemented.

2 Development of functional gene arrays

2.1 Challenges

Design and development of FGAs requires overcoming several technical challenges in both microarray-based technologies[19] and bioinformatics software development. Sequences specific to proteins or genes of interest must be retrieved from public databases. Optimally, this is done by using keywords to search the database. However, several obstacles must be surmounted in obtaining appropriate sequences and then designing specific probes. First, genes or proteins may have a more general annotation rather than a specific gene designation, or may be differently annotated in different organisms. Second, functional gene variants can be very similar and most are homologues. Third, if a given sequence is obtained from uncultured microbes or laboratory clones, the complete gene may not be sequenced resulting in shorter sequences to design probes. Fourth, the sheer number of sequences available for each gene can be staggered and the number of

sequences available is rapidly increased, making the selection of variant-specific probes for all of the variant sequences quite difficult and often resulting in a low coverage of most functional genes. Fifth, development of a standardized set of criteria for oligonucleotide probe design remains challenging. Sixth, because FGAs are designed for use with microbial communities, the oligonucleotide probes must be specific enough to detect their target sequences even in complex samples with many unknown sequences. In addition, since FGAs are markedly different from the typical gene expression arrays, novel methods for data normalization and analysis must be developed. Finally, as mentioned above, in the current genomic and meta-genomic era, the number of microbial sequences is increased exponentially on practically daily basis. So periodic updates are necessary to maintain the applicability of FGAs.

2.2 Strategies

Several strategies have been devised to address many of the above challenges. To overcome some of the complexities of sequence searching and selection, a new strategy has been developed to obtain and screen appropriate sequences. First, functional gene sequences are retrieved by using very broad key words to obtain related sequences as many as possible. Next, all of the sequences are screened by selecting seed sequences (that are known to encode the protein or enzyme of interest) and then HMMER is used to remove all unrelated sequences. Then, sequences for each functional gene are aligned by using a multiple sequence alignment(MSA) program and only shared gene regions are used for subsequent probe design. Finally, oligonucleotide probes

are designed using experimentally established design criteria and a novel software package. Use of both gene- and group-specific probes allows for the detection of both divergent and closely related sequences.

The major steps used in the development of FGAs are shown in Fig.1. 1) Individual functional gene sequences are retrieved from public databases (e.g. NCBI GenBank). 2) All retrieved sequences are screened by using HMMER software and unrelated sequences are removed. 3) 50 mer oligonucleotides are designed using experimentally established design criteria[15–16] with a modified CommOligo[17] program, which allows for selection of group-specific probes. The following criteria are used for the gene- and group-specific probe: 1) gene-specific probes must have a sequence identity $\leq 90\%$, stretch ≤ 20 bases, and free energy ≥ -147 kJ/mol[16]; 2) group-specific probes must have a sequence identity $\geq 96\%$, continuous stretch length ≥ 35 bases, and free energy ≤ -252 kJ/mol[15]. Oligonucleotide probes are designed based on all variants (sequences) of each individual gene. To ensure specificity for the entire array, each designed probe is verified by evaluating the sequence identity, stretch and free energy against a larger database, such as NCBI and EMBL. The probes that have passed all of the selected criteria are then commercially synthesized, and spotted onto glass slides (e.g. Corning UltraGAPS), air dried and then UV-cross-linked according to the slide manufacturers suggested methods.

3 Target preparation and hybridization

A key element in obtaining the best microarray data possible is the use of high quality DNA or RNA. The

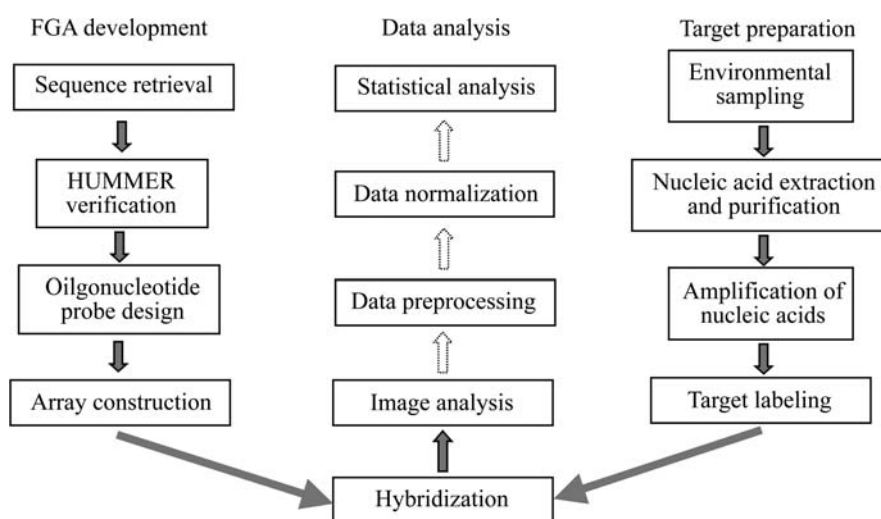


Fig.1 Major steps for development of functional gene arrays, preparations of environmental samples, and data analysis of array hybridization

major steps in target preparation and hybridization are outlined in Fig.1.

3.1 DNA extraction and purification

Microbial community DNA is generally extracted and purified by using a freeze-grind method established by Ref.[20]. Modifications may be required depending on sample type and the particular experimental requirements [21]. There are also many commercial kits available for DNA extraction and purification. Several of these have been evaluated with regards to extraction efficiency using a variety of sample types[22]. Optimally, purified DNA should have $A_{260}/A_{280} > 1.80$, and $A_{260}/A_{230} > 1.70$. If a sample has little (e.g., $< 2.0 \mu\text{g}$) DNA, WCGA can be used to increase the amount of DNA available for hybridization[18].

3.2 RNA preparation

RNA, as an ideal indicator of microbial activity, can also be used with FGAs. The use of mRNA allows for the study of the transcriptional activities of microbial communities and provides insight into microbial activity rather than just ability. However, it is difficult to obtain quality mRNA from environmental samples especially since mRNA tends to be in low abundance. A few methods are currently available for extraction of community RNA from environmental samples. HURT et al[21] describes a simultaneous DNA and RNA extraction and purification method. A new gel electrophoresis method to isolate community RNA has been reported[23]. As with DNA, RNA quality is important as well. Purified RNA should have ratios of $A_{260}/A_{280} > 1.90$ and $A_{260}/A_{230} > 1.70$. A large amount of RNA (10–20 μg) is required for hybridization, but often insufficient quantities of RNA are obtained from community samples. So, an amplification with a novel whole community RNA amplification(WCRA)[24] approach may be required. This method can produce a 1 200- to 1 800-fold amplification from 10 to 100 ng of environmental RNA[24].

3.3 Nucleic acid labeling

The community DNA (5 μg) is fluorescently labeled with Cy5 or Cy3 by random priming using the Klenow fragment of DNA polymerase as described previously [18]. Total community RNA (10–20 μg) is labeled using Cy5 or Cy3 with SuperscriptTM II/III RNase H- reverse transcriptase (Invitrogen Life Technologies, CA, USA) as described by HE et al[25]. The labeled DNA or cDNA targets are purified using a QIAquick PCR purification column and dried in a speedvac centrifuge (Savant Instruments Inc., Holbrook, NY, USA). The labeled targets can then be hybridized or co-hybridized (if two

dyes are used) to FGA chips. Hybridizations are carried out under stringent conditions at 50 °C and 50% formamide as described previously[12, 18]. The hybridization temperature can be lowered (e.g., 45 °C, 42 °C) in order to detect more diverse sequences.

3.4 Array scanning and image processing

A microarray scanner (e.g. ProScan Array, Perkin Elmer, Boston, MA, USA) equipped with lasers at a resolution of 10 μm or higher can be used to process FGA slides. The laser power and photomultiplier tube (PMT) gain are adjusted to avoid saturation of spots. Scanned image displays are saved as 16-bit TIFF files and analyzed by quantifying the pixel density (intensity) of each spot using image quantification software tools (e.g. ImaGene 6.0, Biodiscovery Inc. Los Angeles, CA, USA).

4 Data analysis

The most difficult task involved in the use of FGAs for the study of microbial communities is data analysis and interpretation. Generally, FGA data analysis includes three major steps, preprocessing, normalization, and statistical analysis (Fig.1).

4.1 Array quality validation

The first step is to check the quality of individual spots, regions, and the whole array. The following is a simple method for preprocessing digital array data output from image processing software (e.g. ImaGene). 1) Poor-quality spots, which are flagged by the image processing software using predetermined criteria, are removed. 2) Spots with low signal intensities are removed based on the signal-to-noise ratio(SNR). Normally, an SNR of 2.0 is used as the cutoff[26]. 3) Any outliers are removed. An outlier is defined as any positive spot whose [signal-signal mean of all probe replicates (slides)] is more than three times the standard deviation of the probe replicates. 4) One or more arrays are discarded if the overlap of all detected spots is lower than 80% among replicates.

4.2 Data normalization

The second step is data normalization. The purpose of data normalization is to adjust the microarray data for effects that bring 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, or the use of different reagents, which can be systematically corrected. Normalization can be performed within a chip and/or

among replicate chips. However, a normalization of FGA data needs to consider the following situations. 1) What percentage of spots on each array has positive signals? An FGA (e.g. GeoChip 2.0) normally contains a comprehensive set of probes and some microbial communities may be very simple. 2) Which spots are used as the control for normalization? 3) What is the distribution of signals among positive spots or all spots on the array? Normalization methods may need to be modified based on the answers to the above questions. For example, after data preprocessing, FGA data can be normalized by the total or the mean signal intensity of replicates since it is assumed that the microbial community is not different among replicates.

4.3 Statistical analysis

The third step is statistical analysis of FGA data. The following methods are commonly used. 1) Principal component analysis. This is an exploratory multivariate statistical method for simplifying data sets which reduces the dimensionality of the variables by finding new variables 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 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. 2) 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 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 establishing functionally related groups of genes to gain insights into structure and function of a given microbial community. The most popular clustering method was developed by EISEN et al[27], and other algorithms were described by HEYER et al[28], TAVAZOIE and CHURCH[29], and ZHOU et al[30]. 3) 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[31–33]. Unsupervised neural networks, such as self-organizing maps(SOMs), are more robust and accurate 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[33]. To overcome some of the limitations of SOMs, an unsupervised neural network, termed the self-organizing tree algorithm(SOTA), was proposed[34]. 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[33]. 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[35]. 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 FGA 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[7].

4.4 Data interpretation

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[36]. To compare different samples, some genes are specifically detected in one sample, and some in all samples tested. The numbers of these two type of genes can be calculated as unique and overlap genes as described previously[18]. 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 also be applied to each gene/category with its variants. The software Cluster can be used for cluster analysis and TreeView for visualization[27]. In addition, the network analysis of microarray data has received significant attention, and such a method may be used to present and interpret FGA results.

5 FGA specificity, sensitivity and quantitative capability

5.1 Specificity

Specificity is one of most important parameters to ensure the production of high quality microarray data.

This is especially true for analyzing environmental samples since there may be numerous variants for each gene present in the sample. Microarray specificity can be controlled by probe design and hybridization conditions. Using a 50-mer FGA containing 763 probes for genes involved in nitrogen (e.g. *nirS*, *nirK*, *nifH*) and sulfur cycling (e.g. *dsrA/B*), TIQUIA et al[11] showed that hybridizations at 50 °C could differentiate sequences with <86% identity, while at 55 °C sequences with <90% identity could be distinguished. RHEE et al[12] examined a more comprehensive FGA with 1 662 50-mer probes for genes involved in contaminant degradation and found that hybridization at 50 °C and 50% formamide could differentiate sequences with <88% identity. However, significant cross-hybridization was observed with 70-mer probes which had free energy <-147 kJ/mol[37]. We therefore experimentally established probe design criteria based on sequence identity, continuous sequence stretches and free energy in order to increase specificity[15–16]. These criteria have been implemented in a novel software tool, CommOligo, for microarray probe design[17]. Probes for GeoChip 2.0 were designed by using this newly developed software. Evaluation of GeoChip 2.0 revealed only a very small percentage of false positives (0.002%–0.004%) and no false negatives under experimental conditions[7]. Thus, the use of appropriate probe design criteria allows for a highly specific array.

5.2 Sensitivity

FGA sensitivity is also a major concern, especially for samples from complex environments where many of the sequences will be in low abundance. When PCR-generated probes were used for functional gene arrays(FGA), the detection limit for *nirS* was approximately 1.0 ng of pure gDNA or 25 ng of soil community DNA[9]. The detection sensitivity for *Geobacter chapellei* SSU rRNA gene sequences in soil extracts was approximately 500 ng of total RNA using oligonucleotide capture and detector probes[38]. Other studies of 50-mer FGA probes have shown similar detection limits[11–12]. Coupled to WCGA, as little as 10 pg of community DNA could be detected with 50-mer FGAs[18]. These studies demonstrate that the current microarray technology can provide detection of very small amounts of DNA especially if combined with available amplification techniques, although improved sensitivity is still necessary for environments with extremely low biomass.

5.3 Quantitative capacity

Another important issue is the capability of

microarray-based technologies to provide quantitative information regarding microbial communities. Previous studies using pure culture, mixed culture, and environmental samples have demonstrated linear relationships between target DNA or RNA concentrations and hybridization signal intensity[9, 12, 18]. Similar linear relationships were observed with randomly amplified DNA[18] or RNA[24] using 50 mer FGAs. These studies demonstrate the promise for the use of FGAs as a quantitative tool to analyze environmental samples.

6 Applications of functional gene arrays for microbial community analysis

Since all probes for GeoChip are designed by using functional gene coding sequences, both DNA and RNA can be used as targets for measuring gene abundance and gene expression, respectively. Therefore, GeoChip can be used in a variety of studies, including (but not limited to) 1) detection of functional genes and/or organisms in a particular environment; 2) linkage of microbial community structures to ecosystem functioning; and 3) estimation of gene abundance and activity. GeoChip can be used for analysis of any environmental samples, such as soil, water, sediments, oil fields, deep sea, and animal guts. Here, a few studies are presented to demonstrate the power of FGAs as tools to analyze microbial communities in different ecosystems.

The first study examined microbial communities at the Oak Ridge ERSP (environmental remediation science program) FRC (field research center) site contaminated with nitrate, uranium, and organic compounds using an FGA (GeoChip 1.0) with 2 006 (50-mer) oligonucleotide probes[11–12]. As expected, the highest number of genes was detected in uncontaminated background samples, while the lowest number of genes was detected for the highly contaminated sample, suggesting contaminants strongly affected the microbial communities[18].

GeoChip 2.0 has been used for the analysis of microbial communities in different ecosystems. First, GeoChip 2.0 has been used to track the dynamics of metal-reducing bacteria and associated communities for an in-situ bioremediation study at the Oak Ridge ERSP FRC[7]. This system demonstrated for the first time that uranium can be bioremediated to concentrations below the USA EPA maximum contaminant level(MCL) for drinking water[39]. A significant correlation between U(VI) concentration and the amount of cytochrome genes detected ($r=0.73$, $p<0.05$) was observed, indicating the importance of this group of micro- organisms in U(VI) reduction[7]. Second, RODRÍGUEZ-MARTÍNEZ et

al[40] used GeoChip 2.0 to monitor microbial community shifts in a diesel bioremediation system in Vega Baja, Puerto Rico, and found that an increase in the amount of genes involved in anaerobic degradation of organic contaminants was also observed over time, which is consistent with other evidence that the system shifted to a predominantly anaerobic process over time. Third, YERGEAU et al[41] investigated microbial N and C cyclings in Antarctic sediments using GeoChip 2.0, and found that cellulose degradation and denitrification genes positively correlated with soil temperature. Fourth, GeoChip has been used to investigate microbial community structures in bioremediating a hydrocarbon-contaminated aquifer, and active members were identified in stable isotope experiments fed with labeled biphenyl[42]. Fifth, GeoChip 2.0 has also been used to assess the gene–area relationship of microbial communities in soils, and the results suggested that the forest soil microbial community exhibited a relatively flat gene–area relationship, but the z values varied considerably across different functional and phylogenetic groups[43]. Another study characterized the structure of deep sea hydrothermal vent microbial communities of a mature deep-sea hydrothermal vent chimney and the inner and outer portions of a five-day old chimney (unpublished data). The results showed that microbial functional diversity was much lower in the inner chimney than the outer or mature chimnies, which was confirmed by real-time PCR and clone library approaches.

The latest version of GeoChip (GeoChip 3.0) has been used to study the effects of elevated CO_2 on belowground microbial community, and the results suggest that the concentration of atmospheric CO_2 significantly affects soil microbial community, and that soil microorganisms may influence global carbon and nitrogen cyclings, and probably mitigate global warming effects (unpublished data).

In addition to those comprehensive GeoChips mentioned above, some FGAs have been developed for study of specific ecosystems. For example, an oligonucleotide FGA has been developed for analyze microbial populations in acid mine drainages and bioleaching systems[44], and it is expected that such specific FGAs will be widely applied to particular ecosystems such as bioleaching, bioreactor, and microbial fuel cell(MFC) systems.

Results from these studies demonstrate that FGAs can provide rapid, specific, sensitive, and potentially quantitative analysis for microbial communities in different ecosystems, and that the FGA technology can provide clearer and more comprehensive pictures of

microbial communities in the environment.

7 Summary and future perspectives

FGA technology has received a great deal of attention during its development and subsequent applications. Several recent studies have demonstrated FGAs ability to provide specific, sensitive, and potentially quantitative information regarding microbial communities from a variety of natural environments and controlled ecosystems. This technology, with its high-density and high-through-put capacity, is expected to revolutionize the analysis of the dynamics of microbial community structure and function. FGAs have also been shown to provide direct linkages between microbial genes/populations and ecosystem functions and processes.

However, in order for FGAs to fully realize their potential, a more rigorous and systematic assessment and development will be required. Key issues needing to be addressed include 1) improving experimental design to overcome and minimize inherent variations among hybridizations, 2) developing strategies to improve FGA-based quantitative accuracy, 3) developing approaches to increase hybridization sensitivity so as to detect extremely low biomass, 4) designing novel bioinformatic tools for data analysis and interpretation, and 5) integrating microarray technology into environmental studies to address overarching ecological and environmental questions and hypotheses. Future research and development focus must look not only to improve FGA technology, but also to improve and refine data analysis, interpretation and modeling. Novel strategies for experimental controls are needed in order to fully compare microarray data across samples, experiments, and labs. Improved bioinformatics tools are needed in order to automate sequence retrieval and selection and oligonucleotide probe design and testing for FGA updates. Automated tools are also needed for data analysis and statistical testing. Higher order analysis techniques and mathematical modeling are needed to expand FGA data analysis into network analysis, environmental and cellular modeling, simulation and prediction.

Acknowledgements

This study was supported by the Oklahoma Applied Research Support (OARS), Oklahoma Center for the Advancement of Science and Technology (OCAST), the State of Oklahoma through the Project AR062-034, and the United States Department of Energy under the Genomics: GTL program through the Virtual Institute of

Microbial Stress and Survival (VIMSS; <http://vimss.lbl.gov>), Environmental Remediation Science Program (ERSP), Office of Biological and Environmental Research, Office of Science.

References

- [1] WHITMAN W B, COLEMAN D C, WIEBE W J. Prokaryotes: the unseen majority [J]. *Proc Natl Acad Sci USA*, 1998, 95: 6578–6583.
- [2] ROTTHAUWE J H, WITZEL K P, LIESACK W. The ammonia monooxygenase structural gene *amoA* as a functional marker: Molecular fine-scale analysis of natural ammonia-oxidizing populations [J]. *Appl Environ Microbiol*, 1997, 63: 4704–4712.
- [3] KIRSSTEIN J D, PAERL H W, ZEHR J. Amplification, cloning, and sequencing of a *nifH* segment from aquatic microorganisms and natural communities [J]. *Appl Environ Microbiol*, 1991, 57: 2645–2650.
- [4] ZEHR J P, MELLON M, BRAUN S, LITAKER W, STEPPE T, PAERL H W. Diversity of heterotrophic nitrogen fixation genes in a marine cyanobacterial mat [J]. *Appl Environ Microbiol*, 1995, 61: 2527–2532.
- [5] WIDMER F, SHAFFER B T, PORTEOUS L A, SEIDLER R J. Analysis of *nifH* gene pool complexity in soil and litter at a Douglas fir forest site in the Oregon Cascade mountain range [J]. *Appl Environ Microbiol*, 1999, 65: 374–380.
- [6] SCHENA M, SHALON D, DAVIS R W, BROWN P O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray [J]. *Science*, 1995, 270: 467–470.
- [7] HE Z, GENTRY T J, SCHADT C W, WU L, LIEBICH J, CHONG S C, HUNAG Z, WU W, GU B, JARDINE P, CRIDDLE C, ZHOU J. GeoChip: A comprehensive microarray for investigating biogeochemical, ecological, and environmental processes [J]. *ISME J*, 2007, 1: 67–77.
- [8] SCHADT C W, LIEBICH J, CHONG S C, GENTRY T J, HE Z, PAN H, ZHOU J. Design and use of functional gene microarrays (FGAs) for the characterization of microbial communities [M]// *Methods in Microbiology* Volume. London, UK: Academic Press, 2004: 329–365.
- [9] WU L, THOMPSON D K, LI G S, HURT R, HUANG H, TIEDJE J M, ZHOU J. Development and evaluation of functional gene arrays for detection of selected genes in the environment [J]. *Appl Environ Microbiol*, 2001, 67: 5780–5790.
- [10] TARONCHER-OLDENBURG G, GRINER E M, FRANCIS C A, WARD B B. Oligonucleotide microarray for the study of functional gene diversity in the nitrogen cycle in the environment [J]. *Appl Environ Microbiol*, 2003, 69: 1159–1171.
- [11] TIQUIA S M, WU L, CHONG S C, PASSOVETS S, XU D, XU Y, ZHOU J. Evaluation of 50-mer oligonucleotide arrays for detecting microbial populations in environmental samples [J]. *Biotechniques*, 2004, 36: 664–675.
- [12] RHEE S K, LIU, X, WU L, CHONG S C, WAN X, ZHOU J. Detection of genes involved in biodegradation and biotransformation in microbial communities by using 50-mer oligonucleotide microarrays [J]. *Appl Environ Microbiol*, 2004, 70: 4303–4317.
- [13] STEWARD G F, JENKINS B D, WARD B B, ZEHR J P. Development and testing of a DNA microarray to assess nitrogenase (*nifH*) gene diversity [J]. *Appl Environ Microbiol*, 2004, 70: 1455–1465.
- [14] LOY A, KÜSEL K, LEHNER A, DRAKE H L, WAGNER M. Microarray and functional gene analyses of sulfate-reducing prokaryotes in low-sulfate, acidic fens reveal co-occurrence of recognized Genera and novel lineages [J]. *Appl Environ Microbiol*, 2004, 70: 6998–7009.
- [15] HE Z, WU L, LI X, FIELDS M, ZHOU J. Empirical establishment of oligonucleotide probe design criteria using perfect match and mismatch probes and artificial targets [J]. *Appl Environ Microbiol*, 2005, 71: 3753–3760.
- [16] LIEBICH J, CHONG S C, SCHADT C W, HE Z, ZHOU J. Improvement of oligonucleotide probe design criteria for the development of functional gene microarrays for environmental applications [J]. *Appl Environ Microbiol*, 2006, 72: 1688–1691.
- [17] LI X, HE Z, ZHOU J. Selection of optimal oligonucleotide probes for microarrays using multiple criteria, global alignment and parameter estimation [J]. *Nucleic Acids Res*, 2005, 33: 6114–6123.
- [18] WU L, LIU X, SCHADT C W, ZHOU J. Microarray-based analysis of subnanogram quantities of microbial community DNAs by using whole-community genome amplification [J]. *Appl Environ Microbiol*, 2006, 72: 4931–4941.
- [19] ZHOU J. Microarrays for bacterial detection and microbial community analysis [J]. *Curr Opin Microbiol*, 2003, 6: 288–294.
- [20] ZHOU J, BRUNS M A, TIEDJE J M. DNA recovery from soils of diverse composition [J]. *Appl Environ Microbiol*, 1996, 62: 316–322.
- [21] HURT R A, QIU X, WU L, ROH Y, PALUMBO A V, TIEDJE J M, ZHOU J. Simultaneous recovery of RNA and DNA from soils and sediments [J]. *Appl Environ Microbiol*, 2001, 67: 4495–4503.
- [22] KLERKS M M, VAN BRUGGEN A H C, ZIJLSTRA C, DONNIKOV M. Comparison of methods of extracting salmonella enterica serovar enteritidis DNA from environmental substrates and quantification of organisms by using a general internal procedural control [J]. *Appl Environ Microbiol*, 2006, 72: 3879–3886.
- [23] MCGRATH K C, THOMAS-HALL S R, CHENG C T, LEO L, ALEXA A, SCHMIDT S, SCHENK P M. Isolation and analysis of mRNA from environmental microbial communities [J]. *J Microbiol Meth*, 2008, 75: 172–176.
- [24] GAO H, YANG Z K, GENTRY T J, WU L, SCHADT C W, ZHOU J. Microarray-based analysis of microbial community RNAs by whole-community RNA amplification [J]. *Appl Environ Microbiol*, 2007, 73: 563–571.
- [25] HE Z, WU L, FIELDS M, ZHOU J. Use of microarrays with different probe sizes for monitoring gene expression [J]. *Appl Environ Microbiol*, 2005, 71: 5154–5162.
- [26] HE Z, ZHOU J. Empirical evaluation of a new method for calculating signal to noise ratio (SNR) for microarray data analysis [J]. *Appl Environ Microbiol*, 2008, 74: 2957–2966.
- [27] EISEN M B, SPELLMAN P T, BROWN P O, BOTSTEIN D. Cluster analysis and display of genome-wide expression patterns [J]. *Proc Natl Acad Sci USA*, 1998, 95: 14863–14868.
- [28] HEYER L J, KRUGLYAK S, YOOSEPH S. Exploring expression data: Identification and analysis of coexpressed genes [J]. *Genome Res*, 1999, 9: 1106–1115.
- [29] TAVAZOIE S, CHURCH G M. Quantitative whole-genome analysis of DNA-protein interactions by in vivo methylase protection in *E. coli* [J]. *Nat Biotechnol*, 1998, 16: 566–571.
- [30] ZHOU Y, KALOCSAI P, CHEN J, SHAMS S. Information processing issues and solutions associated with microarray technology [M]// SCHENA M, ed. *Microarray Biochip Technology*. Natick, MA: Eaton Publishing, 2000: 167–200.
- [31] TAMAYO P, SLONIM D, MESIROV J, ZHU Q, KITAREEWAN S, DMITROVSKY E, LANDER E S, GOLUB T R. Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation [J]. *Proc Natl Acad Sci USA*, 1999, 96: 2907–2912.
- [32] TORONEN P, KOLEHMAINEN M, WONG G, CASTREN E. Analysis of gene expression data using self-organizing maps [J]. *FEBS Lett*, 1999, 451: 142–146.
- [33] HERRERO J, VALENCIA A, DOPAZO J. A hierarchical

- unsupervised growing neural network for clustering gene expression patterns [J]. *Bioinformatics*, 2001, 17: 126–136.
- [34] DOPAZO J, CARAZO J M. Phylogenetic reconstruction using a growing neural network that adapts the topology of a phylogenetic tree [J]. *J Mol Evol*, 1997, 44: 226–233.
- [35] TUSHER V G, TIBSHIRANI R, CHU G. Significance analysis of microarrays applied to the ionizing radiation response [J]. *Proc Natl Acad Sci USA*, 2001, 98: 5116–5121.
- [36] BEGON M, HARPER J L, TOWNSEND C R. *Ecology: Individuals, populations, and communities* (3th Edition) [M]. Oxford: Blackwell Science, 1996.
- [37] BOZDECH Z, ZHU J, JOACHIMIAK M P, COHEN F E, PULLIAM B, DERISI J L. Expression profiling of the schizont and trophozoite stages of *Plasmodium falciparum* with a long-oligonucleotide microarray [EB/OL]. <http://genomebiology.com/2003/4/2/R9>.
- [38] SMALL J, CALL D R, BROCKMAN F J, STRAUB T M, CHANDLER D P. Direct detection of 16S rRNA in soil extracts by using oligonucleotide microarrays [J]. *Appl Environ Microbiol*, 2001, 67: 4708–4716.
- [39] WU W, CARLEY J, LUO J, GINDER-VOGEL M A, CARDENAS E, LEIGH M B, HWANG C, KELLY S D, RUAN C, WU L, VAN NOSTRAND J, GENTRY T, LOWE K, MEHLHORN T, CARROLL S, FIELDS M W, GU B, WATSON D, KEMNE K M, MARSH T, TIEDJE J M, ZHOU J, FENDORF S, KITANIDIS P K, JARDINE P M, CRIDDLE C S. *In situ* bioreduction of uranium (VI) to submicromolar levels and reoxidation by dissolved oxygen [J]. *Environ Sci Technol*, 2007, 41: 5716–5723.
- [40] RODRÍGUEZ-MARTÍNEZ E M, PÉREZ E X, SCHADT C W, ZHOU J, MASSOL-DEYÁ A A. Microbial diversity and bioremediation of a hydrocarbon-contaminated aquifer in vega baja, puerto rico [J]. *Int J Environ Res Public Health*, 2006, 3: 292–300.
- [41] YERGEAU E, KANG S, HE Z, ZHOU J, KOWALCHUK G A. Functional microarray analysis of nitrogen and carbon cycling genes across an Antarctic latitudinal transect [J]. *ISME J*, 2007, 1: 163–179.
- [42] LEIGH M B, PELLIZARI V H, UHLIK O, SUTKA R, RODRIGUES J, OSTROM N E, ZHOU J, IEDJE J M. Biphenyl-utilizing bacteria and their functional genes in a pine root zone contaminated with polychlorinated biphenyls (PCBs) [J]. *ISME J*, 2007, 1: 134–148.
- [43] ZHOU J, KANG S, SCHADT C W, CHARLES T, GARTEN C T JR. Spatial scaling of functional gene diversity across various microbial taxa [J]. *Proc Nat Acad Sci USA*, 2008, 105: 7768–7773.
- [44] YIN H, CA, L, QIU G, WANG D, KELLOGG L, ZHOU J, DAI Z, LIU X. Development and evaluation of 50-mer oligonucleotide arrays for detecting microbial populations in Acid Mine Drainages and bioleaching systems [J]. *J Microbiol Meth*, 2007, 70: 165–178.

(Edited by YANG Bing)