Lawrence Berkeley National Laboratory

Title

GeoChip 3.0 as a high-thoughput tool for analyzing microbial community composition, structure, and functional activity

Permalink

https://escholarship.org/uc/item/78415602

Author

He, Z.

Publication Date

2010-04-15

Peer reviewed

GeoChip 3.0 as a High-Throughput Tool for Analyzing Microbial Community Composition, Structure, and Functional Activity

Zhili He[1,8,9], Ye Deng[1,8,9], Joy D Van Nostrand[1,8], Qichao Tu[1,8], Meiying Xu[1,2], Christopher L Hemme[1,8], Xingyuan Li[3], Liyou Wu[1,8],, Terry J Gentry[4], Yifeng Yin[5], Jost Liebich[6], Terry C Hazen[7,8] and Jizhong Zhou[1,7,8]

[1]Institute for Environmental Genomics and Department of Botany and Microbiology, University of Oklahoma, Norman, OK, USA; [2]Guangdong Provincial Key Lab of Microbial Culture Collection and Application, Guangdong Institute of Microbiology, Guangzhou, China; [3]BioinformaticsPro, Newton, MA, USA; [4]Department of Crop and Soil Sciences, Texas A&M University, College Station, TX, USA; [5]Trinity Biosystems Inc., Menlo Park, CA, USA; [6]Institute of Chemistry and Dynamics of the Geosphere 4: Agrosphere, Forschungszentrum Ju¨lich, Ju¨lich, Germany; [7]Earth Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA and [8]Virtual Institute for Microbial Stress and Survival, vimss.lbl.gov

A new generation of functional gene arrays (FGAs; GeoChip 3.0) has been developed, with ~28,000 probes covering approximately 57,000 gene variants from 292 functional gene families involved in carbon, nitrogen, phosphorus and sulfur cycles, energy metabolism, antibiotic resistance, metal resistance and organic contaminant degradation. GeoChip 3.0 also has several other distinct features, such as a common oligo reference standard (CORS) for data normalization and comparison, a software package for data management and future updating and the gyrB gene for phylogenetic analysis. Computational evaluation of probe specificity indicated that all designed probes would have a high specificity to their corresponding targets. Experimental analysis with synthesized oligonucleotides and genomic DNAs showed that only 0.0036–0.025% false-positive rates were observed, suggesting that the designed probes are highly specific under the experimental conditions examined. In addition, GeoChip 3.0 was applied to analyze soil microbial communities in a multifactor grassland ecosystem in Minnesota, USA, which showed that the structure, composition and potential activity of soil microbial communities significantly changed with the plant species diversity. As expected, GeoChip 3.0 is a high-throughput powerful tool for studying microbial community functional structure, and linking microbial communities to ecosystem processes and functioning.

Introduction

Microorganisms are the foundation of the earth's biosphere, and have integral and unique roles in ecosystem functions, such as biogeochemical cycling of carbon, nitrogen, sulfur, phosphorus and various metals. They inhabit almost all imaginable environments to form communities that are always undergoing dynamic changes in structure, composition and function over space and time. However, the identification, detection, characterization and quantification of microbial communities face several challenges. First, microbial communities are extremely diverse with thousands to ten thousands of microbial species in a single gram of soil (Torsvik et al., 2002; Gans et al., 2005). Characterizing such a vast diversity and understanding the mechanisms shaping it presents numerous obstacles. Also, most of these microorganisms (>99%) have not yet been cultured (Whitman et al., 1998). In addition, it presents even more difficulty for microbiologists to establish mechanistic linkages between microbial diversity and ecosystem functioning (Zhou, 2009). In order to tackle these challenges, microarray-based technology has been developed for the analysis of microbial

communities. Currently, this technology has become a powerful and high-throughput tool for analyzing microbial communities and monitoring environmental processes and ecosystem functions (c.f., Wu *et al.*, 2001, 2006, 2008; Loy *et al.*, 2002; Taroncher-Oldenburg *et al.*, 2003; Bodrossy and Sessitsch, 2004; Rhee *et al.*, 2004; Steward *et al.*, 2004; Tiquia *et al.*, 2004; Dix *et al.*, 2006; Rodriguez-Martinez *et al.*, 2006; He *et al.*, 2007; Leigh *et al.*, 2007; Yergeau *et al.*, 2007; Zhou *et al.*, 2008; Liang *et al.*, 2009; Mason *et al.*, 2009; Tas *et al.*, 2009; Van Nostrand *et al.*, 2009; Waldron *et al.*, 2009; Wang *et al.*, 2009). Especially, one of our previous functional gene arrays (FGAs), GeoChip 2.0, containing more than 24 000 probes and covering more than 10 000 gene sequences from ∼150 gene categories involved in key microbially mediated biogeochemical processes (He *et al.*, 2007) has been widely used to analyze microbial communities from different resources, such as soils (Yergeau *et al.*, 2007; Zhou *et al.*, 2008), waters (He *et al.*, 2007; Leigh *et al.*, 2007; Tas *et al.*, 2009; Van Nostrand *et al.*, 2009; Waldron *et al.*, 2009), oil fields (Liang *et al.*, 2009), marine sediments (Wu *et al.*, 2008), extreme environments (Mason *et al.*, 2009; Wang *et al.*, 2009), bioreactor systems (Rodriguez-Martinez *et al.*, 2006) and other habitats (Kimes *et al.*, 2010), to address a variety of questions related to biogeochemical cycles, bioremediation, global climate changes, microbial biography and ecological theories. All these results indicate that FGA-based microarrays such as GeoChip 2.0 can be used not only to analyze the structure, functional activity and dynamics of microbial communities, but also to link microbial communities with ecosystem processes and functions.

However, several major challenges still remain in the full use of such technologies for studying environmental microbiology and microbial ecology. First, with rapid development of high-throughput sequencing, the number of functional gene sequences of interest has increased exponentially. Thus, the GeoChip must be continuously updated and improved. Also, a retrieval of specific sequences for a particular functional gene is difficult because of the size of databases and the inconsistency or inaccuracy of sequence annotations. New strategies and advanced software tools for automatic sequence retrieval, data pre-processing and further statistical analysis are needed. In addition, because a universal standard such as genomic DNA for a pure organism is not applicable for community samples, a quantitative and accurate comparison of different microbial communities is difficult. To address these needs, in this study we report the design, construction, evaluation and application of a new generation of FGAs, GeoChip 3.0, that has distinct features from GeoChip 2.0 in terms of design, gene coverage, data normalization and comparison and future updates. GeoChip 3.0 was applied to show that plant diversity affected the structure, composition and potentially functional activity of belowground microbial communities.

Materials and methods

The following is the summary of methods used in this study, and more detailed information is provided in the Supplementary Data A.

Retrieval and verification of functional gene sequences
Sequence retrieval was performed by a GeoChip design pipeline (Figure 1). For each functional gene, a query of key words was first submitted to GenBank Protein Database to fetch all candidate amino acid sequences. Then, all candidate sequences were verified by HMMER 2.3.2 (Ashburn, VA, USA) (Eddy, 1998). In addition, all confirmed protein sequences were used to obtain their nucleic acid sequences from GenBank for probe design. For GeoChip 3.0, all sequences were downloaded from the GenBank databases before 5 May 2009.

Oligonucleotide probe design, synthesis and microarray fabrication
A new version of CommOligo (Li *et al.*, 2005) with group-specific probe design features, CommOligo 2.0, was used to design both gene- and group-specific probes, and the best set of oligonucleotide probes were synthesized by Invitrogen (Carlsbad, CA, USA), and arrayed onto Corning UltraGAPS (Corning, New York, NY, USA) slides using a
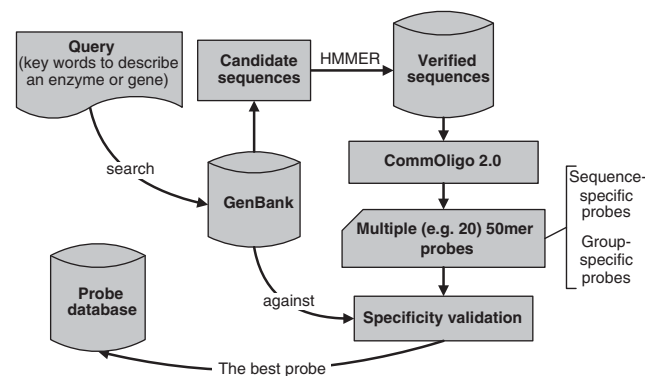

Figure 1 The design pipeline for GeoChip 3.0 construction. The whole pipeline runs on a web-based Common Gateway Interface (CGI) server and the scripts were written in Perl. Sequences for each functional gene were retrieved by key words and confirmed by HMMER (Eddy, 1998) with seed sequences. A new version of CommOligo 2.0 with group-specific probe design features was used to design 50-mer oligonucleotide probes (both gene-specific and group-specific) using the same criteria as described for GeoChip 2.0 (He *et al.*, 2007). After the specificity of all designed probes was computationally checked with currently available databases (GenBank), the best probe for each sequence or each group of sequences was selected to synthesize for GeoChip 3.0 construction. Because all seed sequences and key words are stored in databases, automatic updates can be performed in the future.

Microgrid II Arrayer (Genomic Solutions, Ann Arbor, MI, USA) as described previously (He *et al.*, 2007).

BioCON experimental site, plant species and sampling
The BioCON (Biodiversity, CO_2 and N) experimental site is located at the Cedar Creek Ecosystem Science Reserve, Minnesota, USA (lat. 45° N, long. 93° W), and its main field experiment has a total of 296 plots $(2 \times 2\,m)$ evenly distributed in six 20-m diameter rings with three treatments: CO_2 (ambient, $368\,\mu mol\,mol^{-1}$ vs elevated, $560\,\mu mol\,mol^{-1}$), nitrogen (N; ambient vs $4\,g\,m^{-2}$ per year) and plant diversity. Four levels of plant diversity, 1, 4, 9, or 16 species, were chosen randomly for each plot from 16 perennial species native or naturalized to the Cedar Creek Ecosystem Science Reserve, including (1) four C3 grasses (*Agropyron repens, Bromus inermis, Koeleria cristata, Poa pratensis*), (2) four C4 grasses (*Andropogon gerardii, Bouteloua gracilis, Schizachyrium scoparium, Sorghastrum nutans*), (3) four N-fixing legumes (*Amorpha canescens, Lespedeza capitata, Lupinus perennis, Petalostemum villosum*) and (4) four non N-fixing herbaceous species (*Achillea millefolium, Anemone cylindrica, Asclepias tuberosa, Solidago rigida*) (Reich *et al.*, 2001). Similar to our previous study with soil microbial communities under ambient and elevated CO_2 at the same site (He *et al.*, 2010), this study analyzed 31 soil samples from ring 2 (ambient CO_2 and without N supply) with 11 plots each for 1 and 4 species, 5 plots for 9 species, and 4 plots for 16 species in July 2007.

DNA extraction, purification and quantification
Genomic DNA targets from two *Shewanella* species, MR-4 and W3–18–1, were extracted and purified as previously described (He *et al.*, 2005), and soil DNA from 31 samples (11, 11, 5, and 4 plots from 1, 4, 9 and 16 species, respectively) taken at the BioCON experiment site (Reich *et al.*, 2001) were extracted and purified as described previously (Zhou *et al.*, 1996). DNA quality was assessed by the ratios of A260/A280 and A260/A230, and final DNA concentrations were quantified with a PicoGreen method (Ahn *et al.*, 1996).

GeoChip 3.0 data and statistical analyses
Three types of targets were used for GeoChip 3.0 analysis: (1) 24 synthesized oligonucleotides, (2) genomic DNAs from two *Shewanella* species, MR-4 and W3–18–1 and (3) 31 soil samples from the BioCON study. Genomic DNAs or amplified soil DNAs were labeled, and all hybridizations and image processing were conducted as previously described (He *et al.*, 2007). Scanned images were quantified using the software ImaGene 6.0 (Biodiscovery Inc., El Segundo, CA, USA), and the quantified microarray data were preprocessed as described previously (Wu *et al.*, 2006; He *et al.*,

2007). Pre-processed GeoChip data were further analyzed with different statistical methods, such as cluster analysis (Eisen *et al.*, 1998) and detrended correspondence analysis, as described previously (Zhou *et al.*, 2008).

Results

Selection of functional genes for GeoChip 3.0
For GeoChip 3.0, 292 key enzymes/genes have been selected to target a variety of microbially mediated processes (Table 1).

Carbon cycling. Microorganisms have critical roles in C cycling of the biosphere. A total of 41 enzymes/genes are selected to detect different functional processes of the carbon cycle as described below. (1) CO_2 fixation. Five pathways for autotrophic CO_2 fixation have been identified so far (Berg *et al.*, 2007). Four key enzymes were selected for the following pathways: ribulose-l, 5-bisphosphate carboxylase/oxygenase (Rubisco) for the Calvin cycle (Calvin, 1961), carbon monoxide dehydrogenase (CODH) for the reductive acetyl-CoA pathway (Evans *et al.*, 1966), propionyl-CoA/acetyl-CoA carboxylase (PCC/ACC) for the 3-hydroxypropionate/malyl-CoA cycle (Herter *et al.*, 2002) and ATP citrate lyase (AclB) for the reductive acetyl-CoA pathway (Ragsdale, 1991). (2) Starch. Eight enzymes/genes (alpha-amylase/*amyA*, amylopullulanase/*amyX*/*apu*, cyclomaltodextrin dextrin-hydrolase/*cda*, glucoamylase, isopullulanase, neopullulanase II/*nplT* and pullulanase/*pulA*) are selected to detect starch degradation activity. (3) Cellulose. Four enzymes/genes (cellobiose dehydrogenase, cellobiase/*bgl*, endoglucanase/*egl* and exoglucanase) are selected to detect cellulose degradation. (4) Hemicellulose. Five enzymes/genes (bacterial arabinofuranosidase, fungal arabinofuranosidase, mannanase, xylose isomerase/*xylA* and xylanase) are chosen to monitor hemicellulose degradation. (5) Chitin. Three enzymes (acetylglucosaminidase, endochitinase and exochitinase) are used to detect microbial chitin degradation. (6) Pectin. One enzyme (pectinase) is chosen to monitor pectin degradation. (7) Lignin. Four enzymes/genes (glyoxal oxidase/*glx*, lignin peroxidase or ligninase/*lip*, manganese peroxidase/*mnp* and phenol oxidase/*lcc*) are selected for lignin degradation. (8) Methanogenesis. The key gene *mcrA* encoding methyl coenzyme M reductase is selected for methanogenesis. (9) Methane oxidation. Two enzymes/genes (methane monooxygenase/*mmoX* and particulate methane monooxygenase/*pmoA*) are selected to detect methanotrophic activity. (10) Acetogenesis. Formyltetrahydrofolate synthetase is selected to detect both the Wood-Ljungdahl pathway of autotrophic CO_2 fixation (acetogenesis) and the glycine synthase/reductase pathways of purinolysis (Ljungdahl, 1984, 1986; Mackenzie, 1984). (11) Others. Another eight enzymes/genes (camphor hydroxylase/*camDCAB*,

Table 1 The summary of probe and covered sequence information on GeoChip 3.0

Gene category	No. of genes or enzymes	No. of probes	Sequence-specific probes	Group-specific probes	Covered CDS
Carbon cycling	**41**	**5196**	**1765**	**3431**	**10 573**
Carbon degradation	33	3777	1324	2453	7337
Starch	8	772	263	509	1429
Cellulose	4	305	151	154	484
Hemicellulose	5	505	182	323	940
Lignin	4	330	274	56	419
Chitin	3	574	227	347	1179
Pectin	1	33	31	2	35
Others	8	1258	196	1062	2851
Carbon fixation	4	1043	235	808	2544
Acetogenesis	1	122	35	87	250
Methane metabolism	3	254	171	83	442
Methane production	1	136	84	52	249
Methane oxidation	2	118	87	31	193
Nitrogen (N) cycling	**16**	**3763**	**2148**	**1615**	**7839**
N fixation	1	1224	764	460	2277
Nitrification	2	111	69	42	257
Denitrification	5	1543	1061	482	2995
Ammonification	2	315	87	228	742
Dissimilatory N reduction	2	262	85	177	568
Assimilatory N reduction	3	266	80	186	749
Anaerobic ammonium oxidation	1	42	2	40	251
Phosphorus (P) utilization	**3**	**599**	**183**	**416**	**1220**
Organic P utilization	1	31	26	5	36
Inorganic P biosynthesis	1	182	90	92	318
Inorganic P degradation	1	386	67	319	866
Sulfur (S) cycling	**4**	**1504**	**1083**	**421**	**2042**
S reduction	2	1123	881	242	1438
S oxidation	1	195	66	129	339
Others	1	186	136	50	265
Energy process	**2**	**508**	**410**	**98**	**671**
c-type cytochromes	1	384	370	14	398
Hydrogenases	1	124	40	84	273
Metal resistance	**41**	**4870**	**603**	**4267**	**10 962**
Aluminum	1	81	13	68	128
Arsenic	3	392	59	333	814
Cadmium	2	252	33	219	558
Cadmium and cobalt	3	1002	99	903	2289
Chromium	1	543	65	478	1292
Cobalt	1	49	6	43	96
Cobalt and nickel	3	7	1	6	17
Copper	5	804	71	733	1797
Lead	3	33	3	30	65
Mercury	6	291	77	214	620
Nickel	1	35	3	32	71
Selenium	1	3	2	1	4
Silver	4	213	17	196	419
Tellurium	4	558	50	508	1504
Zinc	2	589	103	486	1246
Miscellaneous	1	18	1	17	42
Organic contaminant degradation	**173**	**8614**	**2165**	**6449**	**17 441**
Aromatics	124	6192	1654	4538	12 416
Aromatic carboxylic acids	38	3302	830	2472	6554
BTEX and related aromatics	18	532	147	385	1105
Chlorinated aromatics	10	460	147	313	869
Heterocyclic aromatics	9	70	35	35	126
Nitroaromatics	9	500	70	430	1115
Polycyclic aromatics	18	241	119	122	416
Other aromatics	22	1087	306	781	2231
Chlorinated solvents	6	355	111	244	747
Herbicide-related compounds	12	816	135	681	1682
Pesticide-related compounds	4	288	37	251	604
Other hydrocarbons	15	454	155	299	894
Others	12	509	73	436	1098
Antibiotic resistance	**11**	**1594**	**265**	**1329**	**3944**
Transporters	5	1181	104	1077	2913
β-lactamases	4	351	147	204	858
Others	2	62	14	48	173
Phylogenetic marker (*gyrB*)	**1**	**1164**	**629**	**535**	**2298**
Total	**292**	**27 812**	**9251**	**18 561**	**56 990**

Abbreviations: BTEX, benzene, toluene, ethylbezene and xylene; CDS, corresponding coding sequence.

limonene epoxide hydrolase/*limEH*, limonene monooxygenase/*lmo*, vanillate monooxygenase/*vanA*, vanillin dehydrogenase/*vdhA*, isocitrate lyase/*aceA*, malate synthase/*aceB* and alkylsuccinate synthase/*assA*) are also chosen for the metabolism of other carbon compounds (for example, aromatics).

Nitrogen (N) cycling. The N cycle is one of the major biogeochemical cycles, and consists of a variety of N pools (for example, N_2, NH_4^+, NO_3^-, NO_2^- and NO) and functional processes (for example, N fixation, nitrification and denitrification). A total of 16 enzymes/genes are selected to target different N cycling processes. (1) *nifH* for N fixation. Microbial N_2 fixation is performed by free-living or symbiotic N-fixing microorganisms through the dinitrogenase reductase gene (*nifH*). (2) Nitrification. Nitrifying microorganisms (bacteria and archaea) are able to oxidize NH_4^+ to NO_3^- in the presence of O_2. Two key enzymes/genes (ammonia monooxygenase/*amoA* and hydroxylamine oxidoreductase/*hao*) are selected. (3) Ammonification. Organic N is converted into inorganic N (NH_4^+) through ammonification (N mineralization). Two key enzymes/genes (glutamate dehydrogenase/*gdh* and urease/*ureC*) are selected. (4) Dissimilatory N reduction to ammonium (DNRA). NO_3^- and NO_2^- can be reduced to NH_4^+ by a variety of microorganisms through a dissimilatory N reduction process, and two key enzymes/genes (nitrate reductase/*napA* and c-type cytochrome nitrite reductase/*nrfA*) are selected for this process. (5) Assimilatory N reduction to ammonium. NO_3^- and NO_2^- can also be reduced to NH_4^+ by an assimilatory N reduction process, and three key genes (nitrate reductase/*nasA*, nitrite reductase/*nirA*/*nirB*) are selected. (6) Denitrification. Denitrifying bacteria are able to convert NO_3^- and NO_2^- to NO, N_2O, and finally to N_2 anaerobically through denitrification, which is one of the key processes in the biogeochemical N cycle and assumed to be one of the major sources of NO and N_2O (greenhouse gases) emissions. Five key enzymes/genes (nitrate reductase/*narG*, nitrite reductase/*nirS&nirK*, nitric oxide reductase/*norB* and nitrous oxide reductase/*nosZ*) are selected for the detection of the denitrification process. (7) Anaerobic ammonium oxidation (anammox). Nitrite and ammonium are converted directly into N_2 by anammox bacteria with a key enzyme/gene hydrazine oxidoreductase/*hzo*.

Sulfur cycling. Four enzymes/genes are used to detect the sulfur cycling of microbial communities with two (sulfite reductase/*dsrA*/*dsrB*) for sulfur reduction, one (sulfite oxidase/*sox*) for sulfur oxidation and one (dissimilatory adenosine-5′-phospho-sulfate reductase/*aprA*) for both the microbial sulfate reduction and sulfur oxidation processes (Meyer and Kuever, 2007).

Phosphorus cycling. Three enzymes were used to detect the phosphorus cycling of microbial communities with two (polyphosphate kinase/*ppk* and phytase) for polyphosphate biosynthesis and one (exopolyphosphatase/*ppx*) for polyphosphate degradation.

Key energy metabolism. We selected only two enzymes, hydrogenase and cytochrome, from some well-known genera (for example, *Geobacter, Anaeromyxobacter, Desulfovibrio, Shewanella, Desulfurobacterium, Desulfobacterium, Rhodobacter* and *Pseudomonas*) for detecting energy metabolism processes of microbial communities.

Metal resistance. Metals and radionuclides are common contaminants in the environment due to anthropogenic activities or natural emissions (Adriano, 2001). As a result of the presence of potentially toxic metals, microorganisms have developed resistance mechanisms to limit exposure to or minimize damage from these metals (Silver, 1996; Silver and Phung, 2005). A total of 41 genes/enzymes comprising resistance mechanisms for 13 metals (Ag, Al, As, Cd, Co, Cr, Cu, Hg, Ni, Pb, Se, Te and Zn) as well as bacterial metallothioneins and metallothionein-like proteins are selected to detect resistance and biotransformation of metals. These metals are common pollutants and microbial mechanisms of resistance to these metals have been well studied (Silver, 1996; Nies, 2003; Silver and Phung, 2005). The majority of these genes confer resistance through transporters, the most common mechanism of prokaryotic metal resistance (Nies, 2003). In addition to transporters, three genes involved in arsenic resistance, including arsenate reductase, and seven for mercury resistance, including mercuric reductase and organomercural lyase, are present.

Degradation of organic contaminants. Contamination by organic chemicals is a worldwide concern and much research has been undertaken to understand the role of microorganisms in the degradation and remediation of organic contaminants. To develop a molecular tool to monitor various degradation pathways, a gene from each step of a contaminant degradation pathway is selected for probe design. As a result, a total of 173 genes/enzymes were selected to detect the potential for degradation of 86 organic contaminants commonly found in the environment. These include 38 enzymes involved in aromatic carboxylic acid (for example, benzoate, phenylpropionate, phthalate) degradation, 18 for BTEX (benzene, toluene, ethylbezene and xylene), 10 for chlorinated aromatics (for example, 2-, 3-, and 4-chlorobenzoate, 2,4,5-trichlorophenoxyacetic acid), 9 for heterocyclic aromatics (for example, carbazole, dibenzothiophene), 9 for nitroaromatics (for example, nitrobenzene, nitropehnol), 18 for polycyclic aromatics (for example, biphenyl, fluorene, naphthalene), 22 for other aromatics (for example, aniline, catechol, phenol), 6 for chlorinated solvents (for example, chloromethane, dichloromethane), 15 for other

hydrocarbons (for example, alkanes, cyclohexane, tetrahydrofuran), 12 for herbicides, 4 for pesticides (for example, atrazine, parathion, lindane), and the remainder (12) are for various chemicals, including acrylonitrile, methanesulfonic acid and tetrahydrofuran. These contaminants are selected from the Priority List of Hazardous Substances (HazDat, 2007) of the Comprehensive Environmental Response, Compensation, and Liability Act (CERCLA; the Superfund program), which ranks contaminants based on their frequency of occurrence and environmental toxicity, and from the Biocatalysis/Biodegradation Database (Ellis *et al.*, 2006) of the University of Minnesota, which lists known degradation pathways for hundreds of chemicals.

Antibiotic resistance. Antibiotic resistance is a growing concern as more and more pathogens develop resistance to common antibiotics. Microorganisms are often exposed to antibiotics in the environment, and antibiotic resistance is often associated with metal resistance (Stepanauskas *et al.*, 2005). In all, 11 genes for antibiotic resistance are included with five for transporters including ATP-binding cassette, multidrug toxic compound extrusion, major facilitator superfamily, Mex, and small multidrug resistance efflux pumps, four β-lactamase genes (Classes A–D) and genes for tetracycline and vancomycin resistance.

Phylogenetic marker. 16S rRNA gene sequences are widely used as a phylogenetic marker for bacterial systematics and ecology, but it is difficult to use them to obtain fine-scale resolutions at species/strain levels (Yamamoto and Harayama, 1995) because the rate of evolution of this molecule is low. An alternative phylogenetic marker molecule, *gyrB*, which encodes DNA gyrase β-subunit gene (*gyrB*), has been widely used for differentiating closely related species/strains (Yamamoto and Harayama, 1998). A phylogenetic tree based on *gyrB* results in a magnitude higher resolution than a tree based on 16S rRNA gene (Yamamoto and Harayama, 1995, 1998; Wang *et al.*, 2007). Thus, *gyrB* gene was selected as a phylogenetic marker on GeoChip 3.0.

Pipeline for GeoChip development
A web-based GeoChip design pipeline has been developed for implementing our GeoChip 3.0 design strategies in four major components: sequence retrieval and verification, oligonucleotide probe design, probe validation and output, and automatic update (Figure 1).

(1) *Sequence retrieval and verification.* Three key steps were implemented to ensure the accuracy of retrieved sequences. First, a query consisting of key words describing the selected functional gene/protein was submitted to GenBank, and all available protein sequences were retrieved.

Generally, the query found all sequences from bacteria, archaea and fungi. However, for cytochromes and hydrogenases, we restricted sequences to be retrieved only in a few common bacterial genera, such as *Geobacter*, *Anaeromyxobacter*, *Desulfovibrio*, *Shewanella*, *Desulfurobacterium*, *Desulfobacterium*, *Rhodobacter* and *Pseudomonas*. Second, seed sequences (normally >5) were identified on the basis of experimental evidence and knowledge of a functional gene from the literature. Third, a consensus model was built with the selected seed sequences, and unrelated sequences were removed by HMMER (Eddy, 1998). Normally, only those sequences with global alignment e-values of <0.1 were considered as highly confident targets and others with local e-values of <1.0 were manually determined to be targets by the user. All verified protein sequences were used to fetch corresponding coding sequences and stored in our local databases (Figure 1). The query, seed sequences and functional gene sequences can also serve as the basis for future automatic updates.

(2) *Oligonucleotide probe design.* Owing to the nature of functional gene sequences, a novel software CommOligo 2.0 based on our previous version of CommOligo (Li *et al.*, 2005) was used to select 50-mer oligonucleotide probes with experimentally established oligonucleotide design criteria (He *et al.*, 2005). To increase the coverage for each functional gene, two types of probes, sequence- and group-specific probes, were designed for GeoChip 3.0 (Figure 1). Sequence-specific probes were preferentially designed, which target a single sequence, whereas group-specific probes target a group of highly homologous sequences exclusively.

(3) *Probe validation.* Multiple (for example, 20) probes for each sequence or each group of sequences were designed, and they were verified by MegaBLAST (Zhang *et al.*, 2000) against all nucleic acid sequences of GenBank. Only those probes with $\leqslant 20$-bp long stretches, $\leqslant 90\%$ sequence identities and $\geqslant -35\,\text{kcal}\,\text{mol}^{-1}$ free energy with their non-targets were kept. The best probe for each single or group of sequences was then stored in the probe database, and commercially synthesized (Figure 1).

(4) *Automatic update.* To reflect the current status of functional genes, this pipeline has a feature to update gene sequences and their probe design information automatically.

Overall description of GeoChip 3.0
GeoChip 3.0 covers 56 990 gene sequences for 292 functional genes with 27 812 probes: 9251 (33.3%) sequence-specific and 18 561 (66.7%) group-specific probes (Table 1). On the basis of functional categories, most of probes target genes involved in
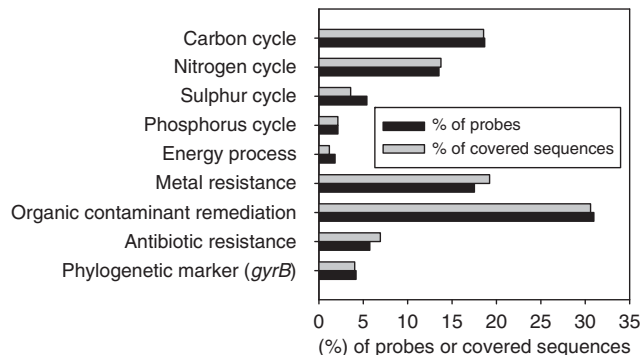
Figure 2 Distribution of all designed probes and covered gene sequences among different categories.

Table 2 Summary of GeoChip 3.0 probes by covered microbial domain and phylum

Domain	Phylum	No. of species	No. of probes	No. of covered CDS
Archaea		140	886	1807
	Crenarchaeota	19	237	560
	Euryarchaeota	107	577	1090
	Thaumarchaeota	1	1	1
	Unclassified archaeon	13	71	156
Bacteria		2744	24 939	52 228
	Acidobacteria	6	107	212
	Actinobacteria	312	2468	5181
	Aquificae	18	29	56
	Bacteroidetes	166	684	1305
	Chlamydiae	9	31	68
	Chlorobi	13	178	361
	Chloroflexi	15	302	642
	Cyanobacteria	115	585	1259
	Deinococcus-Thermus	8	100	215
	Dictyoglomi	2	2	2
	Firmicutes	371	2155	4881
	Fusobacteria	3	12	34
	Lentisphaerae	2	20	40
	Nitrospirae	3	12	16
	Planctomycetes	11	156	277
	Proteobacteria	1428	13 937	30 107
	Spirochaetes	24	46	128
	Synergistetes	1	2	4
	Tenericutes	36	47	88
	Thermodesulfobacteria	4	8	10
	Thermotogae	12	67	135
	Verrucomicrobia	180	3956	7148
	Unclassified bacterium	5	35	59
Fungi		262	1759	2372
	Ascomycota	153	1377	1879
	Basidiomycota	81	279	381
	Glomeromycota	1	2	2
	Microsporidia	1	1	2
	Neocallimastigomycota	4	8	8
	Unclassified fungus	22	92	100
Others[a]		26	228	583
Total		3172	27 812	56 990

Abbreviation: CDS, corresponding coding sequence.
[a]Others included some plasmids and uncultured/unidentified prokaryote organisms.

organic contaminant degradation (31.0%), carbon cycling (18.7%), metal resistance (17.5%) and N cycling (13.5%), with 30.6, 18.6, 19.2 and 13.8% of all covered sequences, respectively (Figure 2 and Table 1). Among 27 812 probes, 24 939 probes (89.7%) target 2744 species of bacteria, 886 (3.2%) target 140 species of archaea, 1759 (6.3%) target 262 species of fungi and 228 (0.8%) target plasmids and uncultured/unidentified prokaryote organisms (Table 2). GeoChip 3.0 also contains eight degenerate probes for the 16S rRNA gene for positive controls (spotted on each sub-grid at least two times), and 672 unique probes designed from hypothetical genes of seven sequenced genomes of hyperthermophiles for negative controls. In addition, a 50-mer common oligonucleotide reference standard (CORS) was mixed with all these probes, including gene probes and controls, and co-spotted on GeoChip 3.0 as a common reference standard for data normalization and comparison (Liang *et al.*, 2010). With such diverse probes and CORS, GeoChip 3.0 could provide more comprehensive analysis of the structure, composition and functional activity of bacterial communities, and also provide useful information for study of soil fungal communities, although those fungal probes mainly target fungal genes involved in limited processes such as carbon degradation.

Specificity evaluation
To examine the specificity of designed probes, the distributions of maximum sequence identity, maximum stretch length and minimal free energy to their closest non-targets were computationally evaluated. On one hand, most of designed sequence-specific probes fell in the ranges of sequence identity, stretch or free energy far away from the thresholds of probe design criteria with their non-targets (Figure 3). Only 4.8% probes had 86–90% identities with their closest non-targets (Figure 3a), 8.0% with continuous stretch lengths of 19–20 bases (Figure 3b) and 6% with free energy of −16 to −35 kcal mol^{-1} (Figure 3c). On the other hand,

group-specific probes covered the most similar regions with all members in the same group (Figure 4). Approximately 96.1% of group-specific probes had 100% sequence identities, and only 3.2% and 0.7% of probes had 98% and 96% identities, respectively, with their targets (Figure 4a). In addition, 98.1% of group-specific probes perfectly matched with their targets, and the rest of them (1.9%) had at least 35 identical bases (Figure 4b). In addition, 86.3% of group-specific probes had maximum free energies from −65 to −80 kcal mol^{-1}, and 12.7% and 1.0% had free energy from −60 to −65, and < −80 kcal mol^{-1}, respectively (Figure 4c). The results suggest that all designed probes have a generally high specificity to their corresponding targets.

The specificity of the designed GeoChip 3.0 was further evaluated experimentally. Both complementary synthesized oligonucleotides and genomic DNA targets were used to hybridize with GeoChip 3.0.
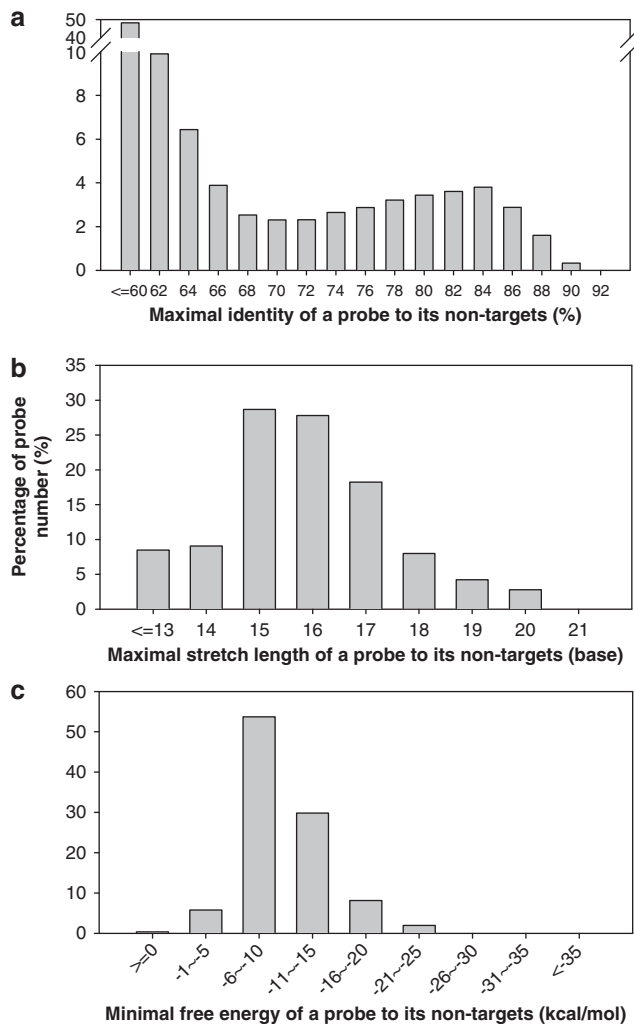
Figure 3 Distribution of 9251 sequence-specific designed probes at their (**a**) maximal sequence identities, (**b**) maximal stretch lengths or (**c**) minimal free energy with their non-targets.

Figure 4 Distribution of 18 561 group-specific probes at their (**a**) minimal sequence identities, (**b**) minimal stretch lengths or (**c**) maximal free energy with their group targets.

First, the specificity was evaluated using an equal mixture (10 pg for each target) of 24 synthesized oligonucleotides (Supplementary Table S1). The hybridizations were performed at 42, 45 and 50 °C in the presence of 50% formamide (1% formamide = 0.6 °C) that was equivalent to 72, 75 and 80 °C, respectively. The hybridization results showed that all 24 probes corresponding to their complementary oligonucleotides showed positive signals at 42 °C and 45 °C, but that one of them was unable to be detected (false negative) at 50 °C. However, seven (0.025%) unexpected probes were detected (false positive) at 42 °C, whereas only one (0.0036%) false positive was detected at 45 or 50 °C (Figure 5). The results suggest that the optimal hybridization temperature for GeoChip 3.0 was 45 °C with 50% formamide because the fewest false-positive and negative spots were detected. This is also consistent with GeoChip 2.0 and our other previous FGAs with fewer probes (Rhee *et al.*, 2004;
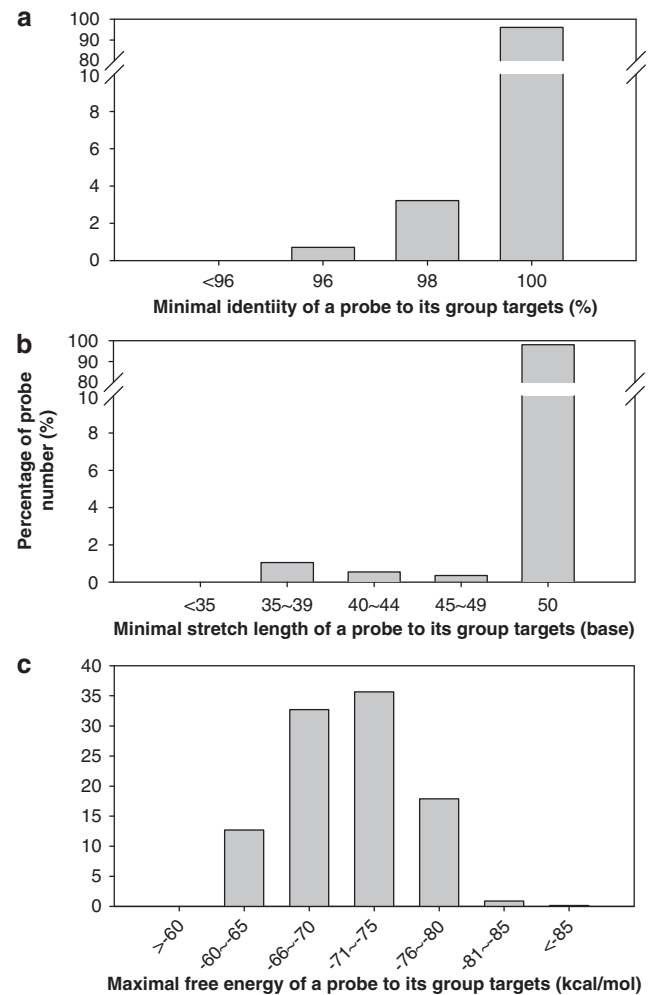
Tiquia *et al.*, 2004; He *et al.*, 2007). Second, genomic DNAs from two *Shewanella* strains, MR-4 and W3–18–1, were mixed with 500 ng for each genome, and used as the target to evaluate GeoChip 3.0 specificity. It is expected that nine probes for *Shewanella* sp. MR-4 and 35 for *Shewanella* sp. W3–18–1 on GeoChip 3.0 would hybridize with the target. The hybridizations conducted at 45 °C and 50% formamide showed that 42 of 44 expected probes had positive signals with an average intensity of 11 428 ± 7223, and an average signal-to-noise ratio = 19.8 ± 9.7 (Table 3). In addition, seven probes showed false-positive signals on GeoChip 3.0, but they had much lower average signal of 3687 ± 2191 than the real positive spots of *Shewanella* sp. MR-4 and W3–18–1. Similar results were observed with synthesized oligonucleotide targets (Table 3). Considering the total number of probes on GeoChip 3.0, the percentage (0.0036–0.025%) of false positives was negligible. Therefore, all the above results
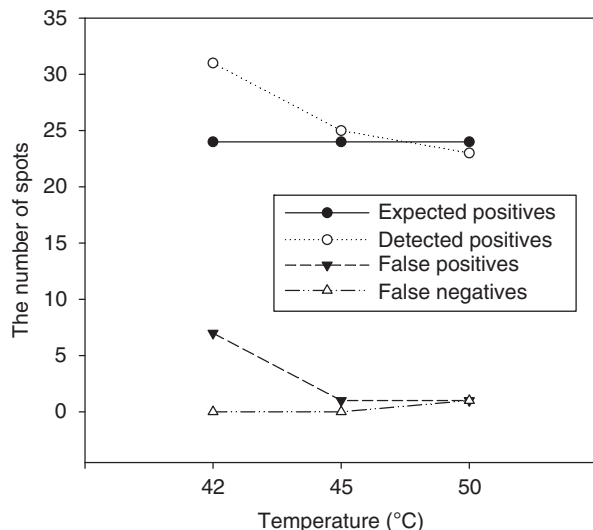
Figure 5 The GeoChip 3.0 was hybridized with a mixture of 24 synthesized oligonucleotide targets at 42 °C, 45 °C and 50 °C in the presence of 50% formamide. The numbers of detected spots, expected spots, false positives and false negatives are shown with five replicates for each condition.

Figure 6 Detrended correspondence analysis (DCA) of soil microbial communities using GeoChip 3.0 data from 11, 11, 5 and 4 (a total of 31) plots planted with 1, 4, 9 and 16 species, respectively. These species were randomly selected from four functional groups, C3, C4, legume and forb, with four species for each group (Reich *et al.*, 2001).

Table 3 Summary of GeoChip 3.0 evaluation with different targets (synthesized oligonucleotides or genomic DNAs) at 45 °C and with 50% formamide

Targets	Oligonucleotide	Genomic DNA
No. of targets	24	2
Expected no. of probes detected	24	44
No. of probes hybridized	25	49
No. of false negatives	0 (0%)	2 (0.0072%)
No. of false positives	1 (0.0036%)	7 (0.025%)
Average signal intensity of targets	6056 ± 4556	$11{,}428 \pm 7223$
Average SNR of targets	13.6 ± 11.9	19.8 ± 9.7
Average signal intensity of false positives	3365 ± 960	$3{,}687 \pm 2191$
Average SNR of false positives	4.3 ± 1.5	6.7 ± 4.3

Abbreviation: SNR, signal-to-noise ratio.

indicate that the probes on GeoChip 3.0 are highly specific.

Applications of GeoChip 3.0 for profiling soil microbial communities

To show the power of the developed GeoChip 3.0, we examined the effects of plant species diversity on soil microbial communities in the BioCON site. Soil samples were taken in July 2007 from 31 plots (11 each for 1 and 4 species, 5 for 9 species and 4 for 16 species) grown at ambient CO_2 without additional N supply (Reich *et al.*, 2001; Reich, 2009). All 31 samples were analyzed by the developed GeoChip 3.0. The signal intensity for all hybridizations was normalized using the CORS strategy. Detrended correspondence analysis of all detected 4012 genes showed that the microbial community and structure
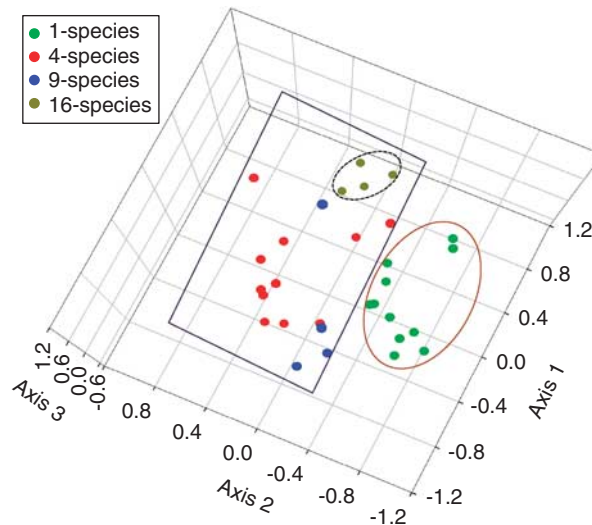
was different between mono-species (1 species) and multispecies, and that 16-species samples appeared to be well separate from 4- or 9-species samples (Figure 6). Consistently, the average number of detected genes was 909.4 ± 103.30 for mono-cultures, which was significantly lower than those for multispecies with 1465 ± 201.18 for 4 species, 1549.6 ± 278.88 for 9 species, and 1540.8 ± 173.08 for 16 species. To understand how plant diversity affects the structure and functional activity of soil microbial communities, we further analyzed the *nifH* gene for N_2 fixation. Two clusters, monoculture and multi-species samples, were observed at a high level, and in the multi-species cluster, four sub-clusters were further formed generally with one for 16-species samples, one enriched for 9-species samples and two enriched for 4-species samples (Figure 7). These results suggested that the structure and functional activity of soil microbial communities were affected by the plant species diversity, and that mono-cultures of plants may lead to a decrease in the belowground microbial community diversity.

Discussion

Microarray-based genomic technology has been widely used for microbial community analysis, and it is expected that microarray-based genomic technologies will revolutionize the analysis of microbial community structure, function and dynamics. The developed GeoChip 3.0 has several distinct features when compared with GeoChip 2.0 (He *et al.*, 2007), and similar to GeoChip 2.0, our
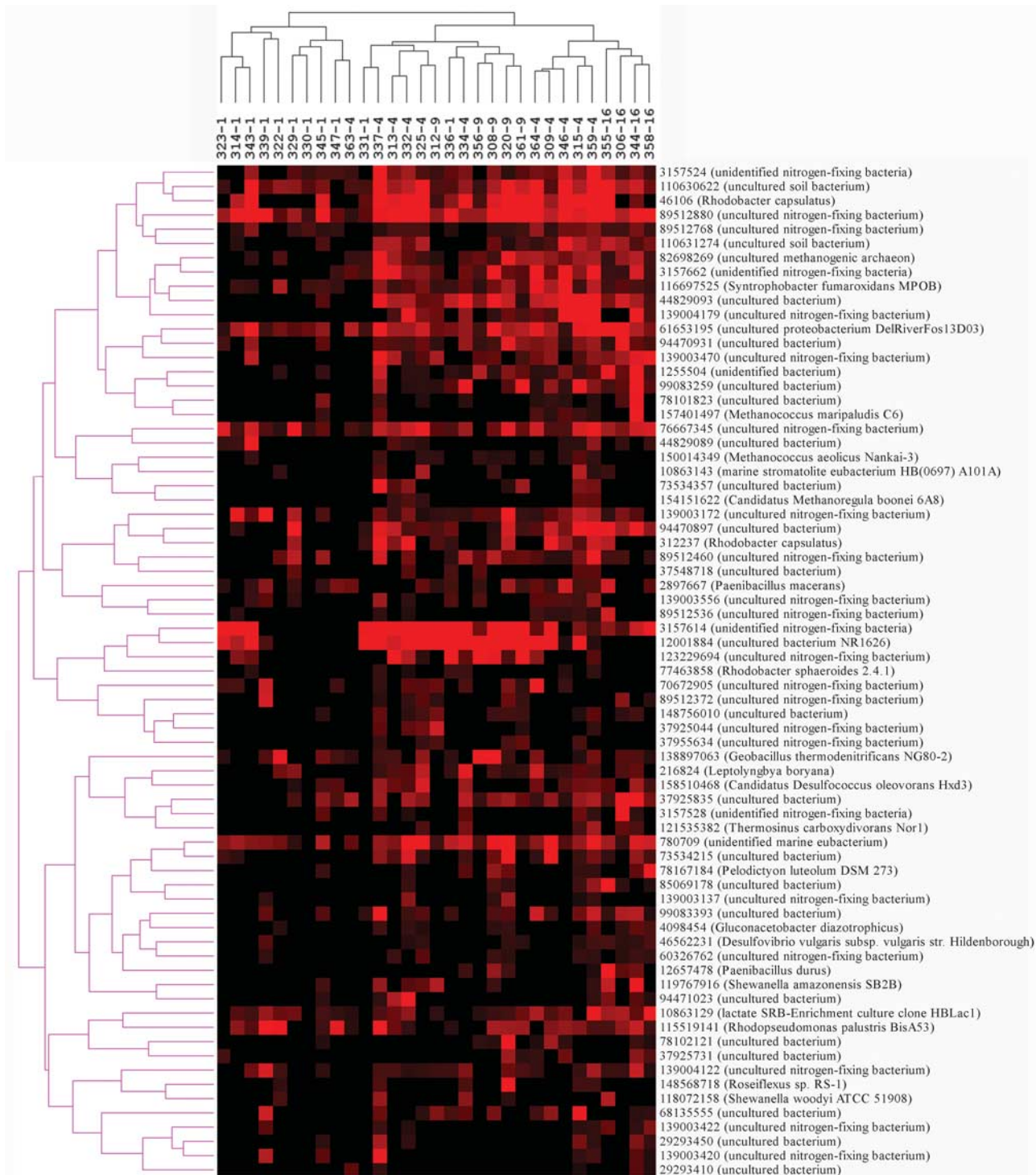
Figure 7 Cluster analysis of *nifH* genes detected by GeoChip 3.0. In total, 114 *nifH* genes were detected, and 71 of them detected in at least 5 out of 31 samples (11, 11, 5 and 4 from 1, 4, 9 and 16 species, respectively) were used for hierarchical clustering analysis. The figure was generated using CLUSTER 3.0 and visualized in TREEVIEW (Berkeley, CA, USA) (Eisen *et al.*, 1998). Black indicates signal intensities below the threshold value and red indicates a positive hybridization signal. The color intensity indicates differences in signal intensity.

computational and experimental evaluation indicates that GeoChip 3.0 is a specific, sensitive and quantitative tool. Using GeoChip 3.0, the BioCON study showed that plant diversity significantly

affected the structure, composition and potential functional activity of soil microbial communities.

GeoChip 3.0 has a few distinct features when compared with GeoChip 2.0 (He *et al.*, 2007). First,

GeoChip 3.0 is more comprehensive, covering approximately 57 000 gene sequences for 292 functional gene families whereas GeoChip 2.0 covers approximately 10 000 gene sequences in 150 gene families, which allows us to obtain more information about functional processes of microbial communities and to analyze more diverse environmental samples. Second, automatically retrieved sequences by key words are verified by HMMER using seed sequences instead of manual verification of sequence homology in GeoChip 2.0, and hence it is much quicker to retrieve the sequences of interest. Third, a CORS has been implemented in GeoChip 3.0 so that data normalization and comparison of different microbial communities can be performed (Liang *et al.*, 2010), and this method can also minimize the effects of cross-hybridization and increase quantitative capability. Fourth, a software package and associated GeoChip design pipeline have been developed for sequence retrieval and verification, probe design and validation, array construction, data analysis, information storage and automatic update, which greatly facilitate the management of such a complicated array, especially for future updates. In addition, GeoChip 3.0 includes the phylogenic marker *gyrB*, which allows us to analyze functional activity and phylogenetic identity of microbial communities together and link microorganisms to their functions. Finally, we have added functional genes involved in antibiotic resistance, which allows us to detect and monitor microbial communities related to public health. All these unique features enable GeoChip 3.0 to be a powerful tool for analysis of the composition, structure, potential function and dynamics of microbial communities, and also link the microbial community structure with ecosystem functioning.

Specificity is a central issue in microarray technology. To ensure GeoChip 3.0 specificity, a few measurements have been taken. First, with experimentally established probe design criteria (He *et al.*, 2005; Liebich *et al.*, 2006), we used a novel software tool, CommOligo 2.0 for oligonucleotide probe design. CommOligo 2.0 considers multiple criteria (for example, sequence identity, continuous stretch length and free energy), and unique algorithms (for example, global alignment and ranking probe candidates) for the optimal probe selection. Second, each designed probe is validated against currently available databases. Third, a computational evaluation showed that a majority (90–95%) of probes on GeoChip 3.0 were far away from the thresholds of probe design criteria, indicating that they should be highly specific to their corresponding targets. Fourth, the GeoChip 3.0 specificity was examined with synthesized oligonucleotides and *Shewanella* genomic DNAs, and only a very small portion of false positives (0.0036–0.025%) was observed. As it was discussed in GeoChip 2.0, those false positives may be because of random error (He *et al.*, 2007).

Group-specific probes are very important for environmental studies because functional genes are generally highly homologous. GeoChip 3.0 has more group-specific probes (66.7%) and covers more than 47 000 sequences in comparison with GeoChip 2.0 with 17.7% group-specific probes covering 3000 sequences. Such an increase in group-specific probes is mainly because of a significant increase in the number of sequences for each functional gene. Our computational analysis of these group-specific probes showed that the majority (95%) of those probes on GeoChip had 100% sequence homology to their corresponding target sequences and had no more than 90% of homology to their non-target sequences, suggesting that these group-specific probes are able to detect groups of homologous sequences. Although we did not evaluate sensitivity and quantitative capability for GeoChip 3.0, it is believed that the same properties remain with GeoChip 3.0 as previously described (Rhee *et al.*, 2004; Tiquia *et al.*, 2004; Wu *et al.*, 2006) because we used the same probe format (50-mer oligonucleotides) and the same array fabrication technologies. Therefore, all these results suggest that GeoChip 3.0 is a specific, sensitive and quantitative tool.

GeoChip is a powerful tool for analyzing the structure, function and dynamics of microbial communities, and linking these properties to ecosystem processes and functions. Currently, high-throughput metagenomic technologies, such as GeoChip (He *et al.*, 2007), PhyloChip (Brodie *et al.*, 2006, 2007) and pyrosequencing (Margulies *et al.*, 2005), provide us opportunities to better characterize microbial communities. In contrast to the 16S rRNA gene-based PhyloChip, GeoChip contains probes from all functionally known geochemical, ecological and environmental processes, and hence it can provide information about the structure and function of microbial communities. Because all GeoChip probes are selected from coding sequences of functional genes, GeoChip can be used not only for measuring the abundance, but also for the expression of functional genes in a microbial community if high quality of mRNAs can be recovered from environmental samples.

Pyrosequencing of targeted functional genes becomes possible, but requires conserved PCR primers and PCR amplifications. Designing PCR primers may be very difficult for most functional genes of interest. Even for genes with PCR primers available, some environmental samples may not be amplified. In addition, quantitative accuracy may be jeopardized because of PCR amplifications. In contrast, GeoChip does not rely on conserved primers or PCR amplifications, although it only detects functionally known genes. Therefore, GeoChip is an ideal tool for analyzing the structure, function and dynamics of microbial communities and for linking microbial diversity with ecosystem functioning. To our knowledge, GeoChip 3.0 is the most comprehensive microarray currently available for studying

microbial communities associated with geobio-chemical cycling, global climate change, bioenergy, agricture, land use, ecosystem management, environmental cleanup and restoration, bioreactor systems and human health.

Acknowledgements

References

Adriano DC. (2001). *Trace Elements in Terrestrial Environments: Biogeochemistry, Bioavailability, and Risks of Metals*, 2nd edn. Springer-Verlag: New York.

Ahn S, Costa J, Emanuel J. (1996). PicoGreen quantitation of DNA: effective evaluation of samples pre- or post-PCR. *Nucleic Acids Res* **24**: 2623–2625.

Berg IA, Kockelkorn D, Buckel W, Fuchs G. (2007). A 3-hydroxypropionate/4-hydroxybutyrate autotrophic carbon dioxide assimilation pathway in Archaea. *Science* **318**: 1782–1786.

Bodrossy L, Sessitsch A. (2004). Oligonucleotide microarrays in microbial diagnostics. *Curr Opin Microbiol* **7**: 245–254.

Brodie EL, Desantis TZ, Joyner DC, Baek SM, Larsen JT, Andersen GL *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.

Brodie EL, DeSantis TZ, Parker JP, Zubietta IX, Piceno YM, Andersen GL. (2007). Urban aerosols harbor diverse and dynamic bacterial populations. *Proc Natl Acad Sci USA* **104**: 299–304.

Calvin M. (1961). Quantum conversion in photosynthesis. *J Theor Biol* **1**: 258–287.

Dix DJ, Gallagher K, Benson WH, Groskinsky BL, McClintock JT, Dearfield KL *et al.* (2006). A framework for the use of genomics data at the EPA. *Nat Biotechnol* **24**: 1108–1111.

Eddy SR. (1998). Profile hidden Markov models. *Bioinformatics* **14**: 755–763.

Eisen MB, Spellman PT, Brown PO, Botstein D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA* **95**: 14863–14868.

Ellis LB, Roe D, Wackett LP. (2006). The University of Minnesota Biocatalysis/Biodegradation Database: the first decade. *Nucleic Acids Res* **34**: D517–D521.

Evans MC, Buchanan BB, Arnon DI. (1966). New cyclic process for carbon assimilation by a photosynthetic bacterium. *Science* **152**: 673.

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

HazDat. (2007). Agency for Toxic Substances and Disease Registry (ATSDR). http://www.atsdr.cdc.gov/cercla/07list.html.

He Z, Gentry TJ, Schadt CW, Wu L, Liebich J, Chong SC *et al.* (2007). GeoChip: a comprehensive microarray for investigating biogeochemical, ecological and environmental processes. *ISMEJ* **1**: 67–77.

He Z, Wu L, Li X, Fields MW, Zhou J. (2005). Empirical establishment of oligonucleotide probe design criteria. *Appl Environ Microbiol* **71**: 3753–3760.

He Z, Xu M, Deng Y, Kang S, Kellogg L, Wu L *et al.* (2010). Metagenomic analysis reveals a marked divergence in the structure of belowground microbial communities at elevated CO_2. *Ecol Lett* **13**: 564–575.

Herter S, Busch A, Fuchs G. (2002). L-Malyl-coenzyme A lyase/beta-methylmalyl-coenzyme A lyase from Chloroflexus aurantiacus, a bifunctional enzyme involved in autotrophic CO_2 fixation. *J Bacteriol* **184**: 5999–6006.

Kimes NE, Van Nostrand JD, Weil E, Zhou J, Morris PJ. (2010). Microbial functional structure of Montastraea faveolata, an important Caribbean reef-building coral, differs between healthy and yellow-band diseased colonies. *Environ Microbiol* **12**: 541–556.

Leigh MB, Pellizari VH, Uhlik O, Sutka R, Rodrigues J, Ostrom NE *et al.* (2007). Biphenyl-utilizing bacteria and their functional genes in a pine root zone contaminated with polychlorinated biphenyls (PCBs). *ISMEJ* **1**: 134–148.

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

Liang Y, He Z, Wu L, Deng Y, Li G, Zhou J. (2010). Development of a common oligo reference standard (CORS) for microarray data normalization and comparison across different microbial communities. *Appl Environ Microbiol* **76**: 1088–1094.

Liang Y, Li G, Van Nostrand JD, He Z, Wu L, Deng Y *et al.* (2009). Microarray-based analysis of microbial functional diversity along an oil contamination gradient in oil field. *FEMS Microbiol Ecol* **70**: 324–333.

Liebich J, Schadt CW, Chong SC, He Z, Rhee SK, Zhou J. (2006). Improvement of oligonucleotide probe design criteria for functional gene microarrays in environmental applications. *Appl Environ Microbiol* **72**: 1688–1691.

Ljungdahl LG. (1984) In: Blakely RL, Benkovic SJ (eds). *Folates and Pterins*. Wiley-Interscience: New York, pp 555–579.

Ljungdahl LG. (1986). The autotrophic pathway of acetate synthesis in acetogenic bacteria. *Annu Rev Microbiol* **40**: 415–450.

Loy A, Lehner A, Lee N, Adamczyk J, Meier H, Ernst J *et al.* (2002). Oligonucleotide microarray for 16S rRNA gene-based detection of all recognized lineages of sulfate-reducing prokaryotes in the environment. *Appl Environ Microbiol* **68**: 5064–5081.

Mackenzie RE. (1984) In: Blakely RL, Benkovic SJ (eds). *Folates and Pterins*. Wiley-Interscience: New York, pp 255–306.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA *et al.* (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**: 376–380.

Mason OU, Di Meo-Savoie CA, Van Nostrand JD, Zhou J, Fisk MR, Giovannoni SJ. (2009). Prokaryotic diversity, distribution, and insights into their role in biogeochemical cycling in marine basalts. *ISMEJ* **3**: 231–242.

Meyer B, Kuever J. (2007). Molecular analysis of the diversity of sulfate-reducing and sulfur-oxidizing prokaryotes in the environment, using aprA as functional marker gene. *Appl Environ Microbiol* **73**: 7664–7679.

Nies DH. (2003). Efflux-mediated heavy metal resistance in prokaryotes. *FEMS Microbiol Rev* **27**: 313–339.

Ragsdale SW. (1991). Enzymology of the acetyl-CoA pathway of CO_2 fixation. *Crit Rev Biochem Mol Biol* **26**: 261–300.

Reich PB. (2009). Elevated CO_2 reduces losses of plant diversity caused by nitrogen deposition. *Science* **326**: 1399–1402.

Reich PB, Knops J, Tilman D, Craine J, Ellsworth D, Tjoelker M *et al.* (2001). Plant diversity enhances ecosystem responses to elevated CO_2 and nitrogen deposition. *Nature* **410**: 809–812.

Rhee SK, Liu X, Wu L, Chong SC, Wan X, 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.

Rodriguez-Martinez EM, Perez EX, Schadt CW, Zhou J, Massol-Deya AA. (2006). Microbial diversity and bioremediation of a hydrocarbon-contaminated aquifer (Vega Baja, Puerto Rico). *Int J Environ Res Public Health* **3**: 292–300.

Silver S. (1996). Bacterial resistances to toxic metal ions–a review. *Gene* **179**: 9–19.

Silver S, Phung LT. (2005). A bacterial view of the periodic table: genes and proteins for toxic inorganic ions. *J Ind Microbiol Biotechnol* **32**: 587–605.

Stepanauskas R, Gless T, Jagoe C, Tuckfield RC, Lindell A, McArthur JV. (2005). Elevated microbial tolerance to metals and antibiotics in metal-contaminated industrial environments. *Environ Sci Technol* **39**: 3671–3678.

Steward GF, Jenkins BD, Ward BB, Zehr JP. (2004). Development and testing of a DNA macroarray to assess nitrogenase (nifH) gene diversity. *Appl Environ Microbiol* **70**: 1455–1465.

Taroncher-Oldenburg G, Griner EM, Francis CA, Ward BB. (2003). Oligonucleotide microarray for the study of functional gene diversity in the nitrogen cycle in the environment. *Appl Environ Microbiol* **69**: 1159–1171.

Tas N, van Eekert MH, Schraa G, Zhou J, de Vos WM, Smidt H. (2009). Tracking functional guilds: 'Dehalococcoides' spp. in European river basins contaminated with hexachlorobenzene. *Appl Environ Microbiol* **75**: 4696–4704.

Tiquia SM, Wu L, Chong SC, Passovets S, Xu D, Xu Y *et al.* (2004). Evaluation of 50-mer oligonucleotide arrays for detecting microbial populations in environmental samples. *Biotechniques* **36**: 664–670.

Torsvik V, Ovreas L, Thingstad TF. (2002). Prokaryotic diversity—magnitude, dynamics, and controlling factors. *Science* **296**: 1064–1066.

Van Nostrand JD, Wu WM, Wu L, Deng Y, Carley J, Carroll S *et al.* (2009). GeoChip-based analysis of functional microbial communities during the reoxidation of a bioreduced uranium-contaminated aquifer. *Environ Microbiol* **11**: 2611–2626.

Waldron PJ, Wu L, Van Nostrand JD, Schadt CW, He Z, Watson DB *et al.* (2009). Functional gene array-based analysis of microbial community structure in groundwaters with a gradient of contaminant levels. *Environ Sci Technol* **43**: 3529–3534.

Wang L, Lee F, Tai C, Kasai H. (2007). Comparison of gyrB gene sequences, 16S rRNA gene sequences and DNA-DNA hybridization in the Bacillus subtilis group. *Intl J Syst Evol Microbiol* **57**: 1846–1850.

Wang F, Zhou H, Meng J, Peng X, Jiang L, Sun P *et al.* (2009). GeoChip-based analysis of metabolic diversity of microbial communities at the Juan de Fuca Ridge hydrothermal vent. *Proc Natl Acad Sci USA* **106**: 4840–4845.

Whitman WB, Coleman DC, Wiebe WJ. (1998). Prokaryotes: the unseen majority. *Proc Natl Acad Sci USA* **95**: 6578–6583.

Wu L, Kellogg L, Devol AH, Tiedje JM, Zhou J. (2008). Microarray-based characterization of microbial community functional structure and heterogeneity in marine sediments from the Gulf of Mexico. *Appl Environ Microbiol* **74**: 4516–4529.

Wu L, Liu X, Schadt CW, 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 DK, Li G, Hurt RA, Tiedje JM, Zhou J. (2001). Development and evaluation of functional gene arrays for detection of selected genes in the environment. *Appl Environ Microbiol* **67**: 5780–5790.

Yamamoto S, Harayama S. (1995). PCR amplification and direct sequencing of gyrB genes with universal primers and their application to the detection and taxonomic analysis of Pseudomonas putida strains. *Appl Environ Microbiol* **61**: 1104–1109.

Yamamoto S, Harayama S. (1998). Phylogenetic analysis of Pseudomonas putida strains deduced from the nucleotide sequences of gyrB, rpoD, and 16S rRNA genes. *Int J Syst Bacteriol* **46**: 813–819.

Yergeau E, Kang S, He Z, Zhou J, Kowalchuk GA. (2007). Functional microarray analysis of nitrogen and carbon cycling genes across an Antarctic latitudinal transect. *ISMEJ* **1**: 163–179.

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

Zhou J. (2009). Predictive microbial ecology. *Microb Biotechnol* **2**: 154–156.

Zhou J, Kang S, Schadt CW, Garten Jr CT. (2008). Spatial scaling of functional gene diversity across various microbial taxa. *Proc Natl Acad Sci USA* **105**: 7768–7773.

Zhou J, Bruns MA, Tiedje JM. (1996). DNA recovery from soils of diverse composition. *Appl Environ Microbiol* **62**: 316–322.