MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2014) doi: 10.1111/1755-0998.12239

GeoChip 4: a functional gene-array-based high-throughput environmental technology for microbial community analysis

QICHAO TU,*[1] HAO YU,*†‡[1] ZHILI HE,* YE DENG,* LIYOU WU,* JOY D. VAN NOSTRAND,* AIFEN ZHOU,* JAMES VOORDECKERS,* YONG-JIN LEE,* YUJIA QIN,* CHRISTOPHER L. HEMME,* ZHOU SHI,* KAI XUE,* TONG YUAN,* AIJIE WANG§ and JIZHONG ZHOU*¶**

*Department of Microbiology and Plant Biology, Institute for Environmental Genomics (IEG), University of Oklahoma, Norman, OK 73019, USA, †State Key Laboratory of Urban Water Resource and Environment, Harbin Institute of Technology, Harbin 150090, China, ‡School of Environmental Science and Engineering, Liaoning Technical University, Fuxin, Liaoning 123000, China, §Research Center for Eco-Environmental Sciences, Chinese Academy of Sciences, Beijing 100085, China, ¶State Key Joint Laboratory of Environment Simulation and Pollution Control, School of Environment, Tsinghua University, Beijing 100084, China, **Earth Science Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

Abstract

Micro-organisms play critical roles in many important biogeochemical processes in the Earth's biosphere. However, understanding and characterizing the functional capacity of microbial communities are still difficult due to the extremely diverse and often uncultivable nature of most micro-organisms. In this study, we developed a new functional gene array, GeoChip 4, for analysing the functional diversity, composition, structure, metabolic potential/activity and dynamics of microbial communities. GeoChip 4 contained approximately 82 000 probes covering 141 995 coding sequences from 410 functional gene families related to microbial carbon (C), nitrogen (N), sulphur (S), and phosphorus (P) cycling, energy metabolism, antibiotic resistance, metal resistance/reduction, organic remediation, stress responses, bacteriophage and virulence. A total of 173 archaeal, 4138 bacterial, 404 eukaryotic and 252 viral strains were targeted, providing the ability to analyse targeted functional gene families of micro-organisms included in all four domains. Experimental assessment using different amounts of DNA suggested that as little as 500 ng environmental DNA was required for good hybridization, and the signal intensities detected were well correlated with the DNA amount used. GeoChip 4 was then applied to study the effect of long-term warming on soil microbial communities at a Central Oklahoma site, with results indicating that microbial communities respond to long-term warming by enriching carbon degradation, nutrient cycling (nitrogen and phosphorous) and stress response gene families. To the best of our knowledge, GeoChip 4 is the most comprehensive functional gene array for microbial community analysis.

Keywords: environmental technology, functional gene array, GeoChip 4, microbial community analysis

Received 10 November 2013; revision received 2 February 2014; accepted 5 February 2014

Introduction

Micro-organisms, which encompass a diverse number of life forms in the Earth's biosphere, play integral and unique roles in ecosystems, such as bio-geochemical cycling of carbon (C) (Bardgett *et al.* 2008), nitrogen (N) (Gruber & Galloway 2008), sulphur (S), phosphorous (P) and various metals. Understanding the diversity, composition, structure, function and interactions of microbial communities over time and space is crucial in microbial ecology. However, it is still very difficult to detect, identify, characterize and quantify the functional capacity of microbial communities due to the extremely diverse and uncultivated nature of most micro-organisms. It is estimated that a total of 1.2×10^{29} bacterial cells are present in the open ocean (Whitman *et al.* 1998), 2.9×10^{29} in the subseafloor sediment (Kallmeyer *et al.* 2012) and 2.6×10^{29} in soil (Whitman *et al.* 1998). Different approaches suggest that the number of bacterial species in a gram of soil varies between 2000 and 8.3 million (Gans *et al.* 2005; Schloss & Handelsman 2006; Roesch *et al.* 2007), and the majority of these (>99%) are as-yet-uncultivated (Rappe & Giovannoni 2003). Characterizing such vast diversity and understanding the mechanisms of community assembly are very difficult. Therefore, it is highly desirable to develop

Correspondence: Jizhong Zhou, Fax: 405 325 7552; E-mail: jzhou@ou.edu

[1]These two authors contributed equally to this work.

high-throughput metagenomic technologies for microbial community analysis.

Indeed, several high-throughput technologies have recently been developed and applied to microbial community analysis, including next-generation sequencing (Sogin *et al.* 2006; Mardis 2008; Shendure & Ji 2008; Ansorge 2009; MacLean *et al.* 2009; Metzker 2010), single cell genomics (Lasken 2007; Walker & Parkhill 2008; Woyke *et al.* 2010; Huang & Zhou 2012), microbial ecological microarrays such as PhyloChip (Brodie *et al.* 2006, 2007; Schatz *et al.* 2010) and functional gene arrays (FGAs), for example, GeoChip (He *et al.* 2007, 2010a, 2012a). Next-generation sequencing technology using the Roche 454 and Illumina platforms has been applied to capture sequences for both targeted genes with available primers (e.g., 16S rDNA, *amoA* and *nifH*) and metagenomes (Sogin *et al.* 2006; He *et al.* 2010b; Qin *et al.* 2010, 2012; Hess *et al.* 2011; Mackelprang *et al.* 2011; Brisson *et al.* 2012; Deng *et al.* 2012; Díez *et al.* 2012; Sintes *et al.* 2013; Yatsunenko *et al.* 2012; Zhou *et al.* 2012). These data have provided many novel insights into the phylogenetic/taxonomic, genetic, and functional diversity, structure and composition of microbial communities from different environments. Microarrays such as PhyloChip and GeoChip have also been applied to profile the phylogenetic and functional structure and composition of known microbial populations (Brodie *et al.* 2007; Zhou *et al.* 2008, 2012; Wang *et al.* 2009; Hazen *et al.* 2010; He *et al.* 2010b), giving novel insights into how environmental factors affect microbial communities in various habitats.

GeoChip is a comprehensive functional gene array targeting hundreds to thousands of different gene families that play important roles in various biogeochemical processes, enabling researchers to comprehensively analyse the functional diversity, composition and structure of microbial communities in various environments. During the past decade, several versions of GeoChip-like FGAs have been developed and were proven to be effective high-throughput tools for microbial community analysis (Wu *et al.* 2001; Rhee *et al.* 2004; He *et al.* 2007, 2010a). GeoChip 2.0 contained more than 24 000 probes and covered more than 10 000 gene sequences from ~150 gene families in key microbial-mediated biogeochemical processes (He *et al.* 2007), while GeoChip 3.0 contained about 28 000 probes and covered approximately 57 000 gene sequences from 292 gene families in various functional processes (He *et al.* 2010a). These two previous versions of GeoChip have been used to analyse microbial communities associated with various environments (He *et al.* 2012a,b), including soil habitats (Zhou *et al.* 2008, 2012; He *et al.* 2010b; Trivedi *et al.* 2012; Yergeau *et al.* 2012), aquatic ecosystems (Taş *et al.* 2009; Kimes *et al.* 2010), extreme environments (Wang *et al.* 2009; Mason *et al.* 2010), contaminated sites (Leigh *et al.* 2007; Liang *et al.* 2009, 2011; Liebich *et al.* 2009; Van Nostrand *et al.* 2009; Xiong *et al.* 2010; Xu *et al.* 2010) and bioreactors (Liu *et al.* 2010, 2012; Zhou *et al.* 2013). All of these studies have indicated that GeoChip is a powerful FGA-based technology to survey the functional diversity, composition, structure, metabolic potential/activity and dynamics of microbial communities, and link them with ecosystem processes and functions.

The objective of this study was to develop a more comprehensive functional gene array, GeoChip 4, for broader applications in analysing biogeochemical processes and microbial responses to environmental perturbations. In addition to updating conventional gene families involved in carbon, nitrogen, sulphur, phosphorous cycles, organic remediation, metal reduction and antibiotics, we also added many previously untargeted gene families, such as those involved in various environmental stress responses, bacteriophages and virulence processes. After the establishment of experimental and data analysis procedures, the developed GeoChip 4 was used to analyse the responses of soil microbial communities to long-term warming at a Central Oklahoma site.

Materials and methods

Sequence retrieval and probe designing

GeoChip 4 employed a similar pipeline (Fig. S1, Supporting information) that was used for GeoChip 3.0 (He *et al.* 2010a). Sequences from gene families covered in previous GeoChip versions were updated using existing keyword queries. For gene families involved in microbial stress responses, bacteriophages and virulence, keywords describing the gene families were created and searched against the NCBI nr database. Candidate protein sequences were fetched and searched against prebuilt HMM models by HMMER (Eddy 1998) for verification. Corresponding nucleotide sequences were retrieved and used for probe design by CommOligo 2.0 (Li *et al.* 2005). Candidate probes were searched against NCBI nt/env_nt databases for specificity.

Microarray construction

GeoChip 4 uses the Roche NimbleGen (Madison, WI, USA) 12 × 135 K (12 arrays on one slide with 135 K probes for each array) platform because of its high density and capacity. The microarray template was obtained from NimbleGen. Probes were placed in every other position across the array so that each probe was adjacent to four void spots (Fig. S2, Supporting information). All GeoChip 4 probes were randomly placed at available

positions except for the 16S probes, which were placed in specific positions throughout the array as positive hybridization controls. Three copies of 563 negative control probes targeting seven thermophile strains were randomly placed on the array. Six thousand 50-mer common oligonucleotide reference standard (CORS) features (with the same probe sequence) were also randomly positioned for data normalization and comparison (Liang *et al.* 2010). Microarrays were synthesized and manufactured by Roche NimbleGen.

DNA extraction, purification and quantification

Soil DNA taken from the BioCON experimental site (Reich *et al.* 2001) and a long-term OK warming site (Luo *et al.* 2001) was extracted by freeze-grinding mechanical lysis (Zhou *et al.* 1996) and purified using a low-melting agarose gel followed by phenol extraction. Community DNA extracted from BioCON soil sample was used for experimental evaluation of GeoChip 4, and DNA from the OK warming site was used in the application study. DNA quality was assessed by the ratios of A260/A280 and A260/A230 using a NanoDrop ND-1000 spectrophotometer (NanoDrop Technologies Inc., Wilmington, DE), and final DNA concentrations were quantified with PicoGreen (Ahn *et al.* 1996) using a FLUOstar Optima microplate reader (BMG Labtech, Jena, Germany).

Target labelling and hybridization

The purified DNA was labelled with Cy-3 using random primers and the Klenow fragment of DNA polymerase I (Wu *et al.* 2006). Labelled DNA was purified using the QIA quick purification kit (Qiagen, Valencia, CA, USA) according to the manufacturer's instructions, measured on a NanoDrop ND-1000 spectrophotometer and then dried down in a SpeedVac (ThermoSavant, Milford, MA, USA) at 45 °C for 45 min. Dried DNA was rehydrated with 2.68 μL sample tracking control (NimbleGen) to confirm sample identity. The samples were incubated at 50 °C for 5 min, vortexed for 30 s and then centrifuged to collect all liquid at the bottom of the tube. Hybridization buffer (7.32 μL), containing 40% formamide, 25% SSC, 1% SDS, 2.38% Cy3-labelled alignment oligo (NimbleGen) and 2.8% Cy5-labelled CORS target, was added. The samples were then mixed by vortexing, spun down, incubated at 95 °C for 5 min and maintained at 42 °C until hybridization. An HX12 mixer (NimbleGen) was placed onto the array using NimbleGen's precision mixer alignment tool, and then, the array was preheated to 42 °C on a hybridization station (MAUI, BioMicro Systems, Salt Lake City, UT, USA) for at least 5 min. Samples (6.8 μL) were then loaded onto the array surface and hybridized approximately 16 h with mixing.

Imaging, data preprocessing and analysis

After hybridization, arrays were scanned at full laser power and 100% photomultiplier tubes gain with a NimbleGen MS 200 Microarray Scanner (Roche NimbleGen). Scanned images were gridded by NimbleScan software using the gridding file containing GeoChip 4 probes and NimbleGen control probes to obtain the signal intensity for each probe. Probe spots with coefficient of variance (CV) >0.8 were removed.

In general, a local background that represents the actual background signal for each spot is preferred for signal-to-noise ratio (SNR) calculations and false-positive filtering instead of the global background generated by NimbleScan; so a different method for background calculation was introduced here. To obtain a local background signal for each probe, a customized void gridding file targeting positions without probes was generated. The scanned images were gridded using the void gridding files. The local background signal for each probe was calculated as the mean signal intensity of the four neighbouring void spots. When all of the four neighbouring void spots were not valid, the set of eight void spots surrounding its closest neighbours was used (Fig. S2, Supporting information). A probe was discarded if all twelve void spots failed to meet the CV criteria (CV <0.8). Signal standard deviation for the background was calculated as:

$$\text{MeanVoidStdev} = \sqrt{\frac{\text{stdev}_1^2 + \text{stdev}_2^2 + \ldots + \text{stdev}_n^2}{n}}$$

where stdev_i is the standard deviation for each void spot, and n ($\leq$8) is the number of void spots. SNR was calculated as previously described (He & Zhou 2008). Signal intensities for each probe were normalized by the mean signals from all spiked CORS probes.

Statistical analysis

Various statistical methods were used for further analysis. Three different nonparametric multivariate analysis methods, adonis (permutational multivariate analysis of variance using distance matrices), anosim (analysis of similarities) and MRPP (multiresponse permutation procedure), as well as detrended correspondence analysis (DCA), were used to measure the overall differences of community functional gene structure between treatment and control samples (Zhou *et al.* 2012). The significance of differences in relative abundance of functional genes between control and treatment samples was evaluated using Student's t-test.

Results

Gene families included in GeoChip 4

In addition to updating those gene families already covered by previous GeoChips, GeoChip 4 gene coverage was expanded by adding 98 new gene families to target more microbially mediated functional processes involved in microbial stress response (45 gene families) and virulence (13) as well as genes for environmentally related bacteriophages (40). Gene families included in previous versions of GeoChip were also included on GeoChip 4, and coverage of these gene families was manually checked and improved by modifying keyword queries for some genes and increasing gene sequence coverage from the latest public database at the time of development (continuous update since ~June 2010). Specifically, 312 gene families were included to target functional processes such as antibiotic resistance (11), carbon cycling (41), energy processing (4), metal resistance (44), nitrogen cycling (17), organic remediation (184), phosphorus utilization (3), sulphur (6), *bchY* and *gyrB* genes. A detailed description of these gene families and functional processes is in (He *et al.* 2010a). Newly targeted gene families in GeoChip 4 are described below:

Stress responses. Micro-organisms are sensitive to environmental conditions, and many abiotic and/or biotic environmental fluctuations, such as temperature, pH, oxygen, and nutrition, could affect their growth, functions, dynamics and evolution. To survive in such environmental variability, micro-organisms have developed different mechanisms for temporary and/or long-term adaptation. To study microbial responses to various environmental stressors, 45 gene families involved in stress responses were selected. (i) Sigma factors, general stress responses and stringent responses. Different genes are activated under different environmental conditions to regulate microbial responses. Four sigma factor genes (σ^{70}, σ^{38}, σ^{32} and σ^{24}), one general stress response gene (*katE*) and one stringent response gene (*obgE*) were selected. (ii) Temperature. Functional genes targeting heat shock (*dnaK*, *grpE*, *groES*, and *groEL*) and cold shock (*cspA* and *cspB*) were selected. Regulatory genes, including *hrcA* for heat shock and two-component genes *desK-desR* for cold shock, were also selected. (iii) Osmolarity. Changes in extracellular osmotic pressure may elicit rapid water fluxes along the osmotic gradient to maintain the proper turgor for normal cellular physiology. Cells respond to osmotic shock by adjusting the cellular concentration of osmolytes or compatible solutes (Kempf & Bremer 1998). Here, four functional genes involved in microbial osmotic stress response were selected, including *opuE*, *proX*, *proV* and *proW*. (iv) Oxidative stress and oxygen limitation. Three functional genes (*ahpC*, *ahpF* and *katA*) and two regulatory genes (*perR* and *oxyR*) were selected for oxidative stress response. The expression of *ahpCF* and *kat* are controlled by transcriptional factor *oxyR* or *perR* (Pomposiello & Demple 2001; Fuangthong *et al.* 2002). For oxygen limitation, two cytochrome genes (*cydA* and *cydB*), three nitrate reductase genes (*narH*, *narI* and *narJ*), one regulatory gene (*fnr*) and a two-component system gene (*arcA-arcB*) were selected. (v) Nutrient limitation. Limited nutrients such as glucose, phosphate and nitrogen are common stresses for micro-organisms in the natural environment. Glucose is the preferred carbon and energy source for most micro-organisms, and two genes (*bglP* and *bglH*) were selected for glucose limitation. Phosphate is the essential nutrient for microbial metabolism for its role in many functional structures such as nucleic acid, ribosomes and membranes, and phosphate-specific transport system genes (*pstS*, *pstA*, *pstB* and *pstC*), alkaline phosphate gene *phoA* and the two-component system gene *phoB* were selected for phosphate limitation. Nitrogen is also a key nutrient for micro-organisms, and glutamine synthase gene *glnA* and regulatory genes *tnrA* and *glnR* for nitrogen limitation were selected. (vi) Protein stress. Overexpression of recombinant proteins stimulates protein stress (Goff & Goldberg 1985; Dong *et al.* 1995), and two genes (*ctsR* and *clpC*) were selected.

Virulence. Bacteria are the most dominant group of pathogens and account for approximately 40% of all pathogens followed by fungi, helminthes, viruses and prions and protozoa (Taylor *et al.* 2001; Woolhouse & Gowtage-Sequeria 2005). For pathogenicity, a pathogen must enter and cause damage to the host while evading the host defence mechanisms. Such pathogenicity is determined by multiple virulence factors such as adherence, colonization, invasion, secretion system, immune evasion, toxin production and iron uptake (Finlay & Falkow 1997; Wu *et al.* 2008). Thus, these virulence factors could serve as specific markers for the detection and monitoring of a variety of pathogens in clinical and environmental samples. Here, 13 bacterial virulence factors, including adhesion, aerobactin, capsule, colonization factor, fimbriae, haemolysin, invasion proteins, siderophore, pilin, type III secretion proteins, sortase, toxin and virulence proteins, were selected. Adhesins are cell surface components or appendages of bacteria that facilitate bacterial colonization within their host (Kline *et al.* 2009). Siderophores, including aerobactin, are small, high-affinity iron-chelating compounds generally produced under iron-limiting conditions to scavenge iron (Bossier *et al.* 1988; Neilands 1995). The bacterial capsule promotes virulence by reducing host immune responses (Singh *et al.* 2011).

Colonization factors are surface structures that allow bacteria to bind and colonize onto host cells (Tobias & Svennerholm 2012). Toxins and haemolysins play an important role in toxigenesis by affecting and damaging a host cell directly and aggressively. Pilin is the major subunit protein of pili in many bacteria (Craig *et al.* 2003) and plays roles in surface attachment and DNA transfer by conjugation (Yang & Bourne 2009; Carter *et al.* 2010). Sortases are a family of enzymes found in Gram-positive bacteria and act as both proteases and transpeptidases, which are required for cell wall anchoring protein production, adhesion to epithelial cells and colonization of the mouse intestine (Cossart & Jonquières 2000; Mazmanian *et al.* 2001). The type III secretion system is particularly prevalent among Gram-negative bacterial pathogens to transport effectors directly into their hosts cells (Galan & Collmer 1999).

Bacteriophages. Bacteriophages are potentially the most numerous form of life in Earth's biosphere (Grath & Sinderen 2007), surpassing the number of bacteria by an order of magnitude (Bergh *et al.* 1989; Wommack & Colwell 2000). In sea water, up to 9×10^8 virions per millilitre have been found in surface microbial mats (Wommack & Colwell 2000), and more than 70% of marine bacteria might be infected by phages (Prescott 1993). It is expected that the number of virus in each gram of soil is ~1.5×10^8, accounting for about 4% of the total bacterial population (Ashelford *et al.* 2003). Bacteriophages play several important roles in the environment, such as turnover of nutrients by lysis of their prokaryotic hosts (Weinbauer 2004), supplying released nutrients to other organisms (Suttle 1994, 2007) and driving evolution by transferring genetic information between multiple hosts (Chen & Novick 2009; Gomez & Buckling 2011).

To monitor the composition and abundance/activity of bacteriophages in the environment over time and space, we designed probes for 40 genes that are related to structure, host recognition, lysis and replication processes in phages. (i) Host recognition. The first critical step in any virus' life cycle is the recognition of its host organism through cell surface components such as receptors or lipopolysaccharides. Two long-tail fibre proteins, p38 and p37, responsible for bacterial host receptor recognition in T7 type phages, T2 (Riede *et al.* 1987) and T4 (Thomassen *et al.* 2003), respectively, were selected. (ii) Replication. Bacteriophages have a number of diverse replication mechanisms that have been reviewed extensively elsewhere (Weigel & Seitz 2006). Some phages encode for all of the genes necessary for the replication of their genomes while others use components from their host's replication mechanism. Here, 25 genes encoded by bacteriophages for replication were selected, among which eight are encoded by T4 bacteriophage and the rest are common to multiple bacteriophages. (iii) Structure. Proteins representing major structural components of the virion were chosen with priority given to those that have previously been used as environmental markers to assess viral populations. Six genes encoding proteins that function in forming unique bacteriophage structures were selected, including contractile tail tube protein, contractile sheath protein, noncontractile tail protein, capsid protein, scaffolding protein and tape measure protein. (iv) Lysis. With the exception of the ssDNA Inoviridae (Hoffmann-Berling & Mazé 1964), the end of the phage life cycle in bacteria involves the lysis and death of the host cell. Several different types of proteins and enzymes can be involved in this process and have been reviewed previously (Young *et al.* 2000). Seven genes involved in phage lysis were selected, including two endolysins (glycosidase and transglycosilase), three holins (class 1, class 2, and class 3), one lysozyme and one lysis protein B.

Overall description of GeoChip 4 features

Instead of the spotted array platform used for GeoChip 3.0 (round spots; 100–150 μm in diameter), an *in situ* synthesized microarray platform is employed for GeoChip 4, resulting in much smaller square spots (13 μm × 13 μm, about 1/46 to 1/105 of the spot size in spotted arrays). GeoChip 4 is a more compact and dense array (8.9 mm × 6.5 mm) with ~135 K probes per array and 12 arrays on each slide, making GeoChip 4 a high-density functional gene array able to hybridize multiple samples on a single slide under nearly identical conditions. In total, 82 074 probes targeting 410 functional gene families were included in GeoChip 4, covering 141 995 coding sequences (CDS) (Table 1). Among these, 18 098 (22.1%) are sequence-specific probes and 63 976 (78.0%) are group-specific probes (Table 2). Specifically, 21 541 probes (26.3%) targeted 45 microbial stress response genes, 17 056 probes (20.8%) targeted 184 organic remediation genes, and 11 034 probes (13.4%) targeted 41 genes involved in carbon cycling processes. At the taxonomic level, 73 106 probes (89.1%) targeted 4332 bacterial strains, 2555 (3.1%) for 188 archaeal strains, 4965 (6.1%) for 420 eukaryotic strains, 1071 (1.3%) for 273 bacteriophage, and the remaining for uncultured/unidentified/environmental organisms (Table 3). In addition, GeoChip 4 also contains 640 (80 replicates × 8 degenerate probes) probes targeting 16S rRNA sequences as positive controls, 1689 (3 replicates × 563 probes) probes targeting seven sequenced hyperthermophile genomes as negative controls. Moreover, 6000 identical probes were included as a common oligonucleotide reference standard (CORS) for data normalization and comparison (Liang *et al.* 2010).

Table 1 Comparisons of major differences between GeoChip 3.0 and GeoChip 4

	GeoChip 3.0	GeoChip 4
Fabrication method	Spotted array	In situ synthesized array
Feature shape and size	Circle, 100–150 μm in diameter	Square, 13 μm × 13 μm
Array size	25 mm × 76 mm	8.9 mm × 6.5 mm
Capacity per slide	1 array × 30 000 probes	12 arrays × 135 000 probes
Number of gene families	292	410
Number of probes	27 812	82 074
Number of covered CDS	56 990	141 995
Number of covered strains	3172	5247
Minimum required community DNA	2 μg	0.5 μg

Specificity for positive and negative control probes as well as CORS was also verified by searching against NCBI nt database.

Computational evaluation of designed probes

Specificity for GeoChip 4 probes was computationally evaluated against the NCBI database based on sequence identity, continuous stretch length and free energy. For sequence-specific probes, the maximum identity, maximum stretch length and minimal free energy to their closest nontarget sequences were calculated. About 66.5% of probes showed maximum sequence identities of 60% or less to their nontargets. Only 4.2% of probes showed 86–90% sequence identity (Fig. 1a), 6.4% had 19–20 base continuous stretch (Fig. 1b), and 6.8% had −35 to −25 kcal/mol free energy to their nontargets (Fig. 1c). For group-specific probes, the minimum identity, minimum stretch length and maximum free energy to its group members were calculated. Approximately 92% of group-specific probes were identical to their group members (Fig. 1d,e), and more than 79% showed −85 to −65 kcal/mol free energy to their group members (Fig. 1f). All these results were consistent with the probe design criteria (He *et al.* 2010a), showing the designed probes were highly specific to their targets.

Determining the minimal DNA amount for array hybridization

DNA extraction of many environmental samples is not always an easy process, thus determining the appropriate amount of community DNA is important for successful microarray-based analyses. An insufficient amount of DNA during microarray hybridization will lead to inadequate observation of hybridized probes/genes, while excessive amounts of DNA will lead to signal intensity saturation of dominant functional genes, resulting in invalid observations for comparative analysis. To address the above questions, ten DNA concentrations (each with three replicates) ranging from 0.001 to 4 μg were hybridized with GeoChip 4 at 42 °C with 40% formamide. Probes with SNR ≥2.0 in at least two of three replicates were treated as positives. The number of positive probes at each DNA concentration and overlap between any two DNA concentrations were examined. More than 10 000 and 17 500 positive hybridization spots were obtained with 50 ng and 100 ng DNA concentrations, respectively. Consistent hybridizations were observed with DNA amounts ≥0.5 μg, and more than 20 000 positive probes were detected for all DNA concentrations ≥0.5 μg (Fig. 2a). Thus, for a good hybridization, a minimum of 0.5 μg total community DNA is required for environmental samples, and 1 μg or more DNA is recommended. Also high overlap percentages of positive probes (~70–95%) were observed between the different DNA concentrations used, suggesting stable hybridizations of GeoChip 4 across replicates.

In addition, the relationship between signal intensity and DNA amount was also examined. Overlapped positive probes among all samples with different DNA concentrations were used. The signal intensity of each probe was averaged across the three technical replicates. Log-transformed average signal intensities of all probes were compared with log-transformed DNA concentrations. A significant linear relationship ($r = 0.925$) was observed between the average signal intensity and DNA concentrations (Fig. 2b), indicating that GeoChip 4 hybridization is quantitative for environmental samples.

Application of GeoChip 4 to analyse soil microbial communities under warming

The response of soil microbial communities to long-term warming was studied using GeoChip 4. Ten soil samples (five treatments and five controls) were collected in April 2008 from an experimental warming site located in Central Oklahoma (34°59′N, 97°31′W), which has been subjected to a continuous warming treatment at 2 °C higher than atmospheric temperature since 1999 (Luo *et al.* 2001). In total, 30 632 probes were detected in at least 2 of 10 samples, with 23 630–26 515 probes per sample. Both DCA and nonparametric multivariate statistical tests showed significantly different microbial community structures and compositions between warming and control samples (Fig. 3) with *P*-values varying from 0.028 to 0.059 (ANOSIM: $R = 0.348$, $P = 0.059$; adonis: $F = 0.234$,

Table 2 Summary of probe and covered coding sequence information of GeoChip 4 based on gene categories

Gene category	No. genes or enzymes	No. probes	me No. sequence-specific probes	No. group-specific probes	No. covered CDS
Carbon cycling	**41**	**11 034**	**3204**	**7830**	**18 071**
Acetogenesis	1	122	35	87	250
Methane cycling	3	498	244	254	1577
Carbon fixation	4	1620	288	1332	3148
Carbon degradation	33	8794	2637	6157	12 996
Cellulose	4	702	268	434	1246
Chitin	3	1473	519	954	2295
Hemicellulose	5	1380	363	1017	2026
Lignin	4	934	569	365	1028
Pectin	1	72	50	22	63
Starch	8	2432	613	1819	3439
Others	8	1801	255	1546	2899
Nitrogen cycling	**17**	**7386**	**3090**	**4296**	**10 744**
Ammonification	2	969	210	759	1554
Anammox	1	47	2	45	282
Assimilatory N reduction	4	469	106	363	692
Dissimilatory N reduction	2	646	193	453	1341
Nitrification	2	1419	497	922	118
Denitrification	5	2612	1318	1294	4534
Nitrogen fixation	1	1224	764	460	2223
Phosphorus utilization	**3**	**1341**	**351**	**990**	**2261**
Sulphur	**6**	**3113**	**1784**	**1329**	**4049**
Adenylsulphate reductase	3	563	301	262	549
Sulphite reductase	2	2084	1336	748	2831
Sulphur oxidation	1	466	147	319	669
Energy process	**4**	**853**	**436**	**417**	**1131**
Cytochrome	1	619	390	229	728
Hydrogenase	2	197	40	157	348
P450	1	37	6	31	55
Metal resistance	**44**	**9272**	**1295**	**7977**	**17 198**
Aluminium	1	162	30	132	270
Arsenic	5	905	208	697	1551
Cadmium	2	876	99	777	1284
Cadmium, cobalt, zinc	3	1333	145	1188	2637
Chromium	1	1067	129	938	1949
Cobalt	1	62	8	54	118
Cobalt, nickel	3	16	5	11	29
Copper	5	1729	199	1530	3063
Lead	3	69	13	56	119
Mercury	7	733	169	564	1237
Nickel	1	36	4	32	71
Selenium	1	4	2	2	6
Silver	4	402	45	357	934
Tellurium	4	1290	136	1154	2542
Zinc	2	555	99	456	1321
Miscellaneous	1	33	4	29	67
Organic remediation	**184**	**17 056**	**4692**	**12 364**	**28 716**
Aromatics	132	12 831	3732	9099	21 056
Aromatic carboxylic acid	38	6068	1900	4168	9637
BTEX and related aromatics	21	1131	225	906	2044
Chorinated aromatics	11	717	205	512	1227
Heterocyclic aromatics	9	154	68	86	234
Nitoaromatics	11	981	187	794	1782
Polycyclic aromatics	19	1138	375	763	1547
Other aromatics	23	2642	772	1870	4585

Table 2 (Continued)

Gene category	No. genes or enzymes	No. probes	me No. sequence-specific probes	No. group-specific probes	No. covered CDS
Chlorinated solvents	6	600	161	439	1060
Herbicides related compounds	13	1408	283	1125	2373
Pesticides related compounds	5	492	96	396	928
Other hydrocarbons	14	746	241	505	1368
Others	14	979	179	800	1931
Antibiotic resistance	**11**	**3,334**	**589**	**2,745**	**5533**
Transporters	5	2316	294	2022	3979
β-lactamases	4	665	220	445	964
Others	2	353	75	278	590
Stress	**45**	**21 541**	**1313**	**20 228**	**40 635**
Cold shock	4	70	0	70	134
Heat shock	5	1655	215	1440	2616
Glucose limitation	2	87	3	84	134
Nitrogen limitation	3	1461	11	1450	3939
Osmotic stress	4	457	41	416	997
Oxygen limitation	7	994	121	873	2020
Oxygen stress	7	4765	255	4510	10 375
Phosphate limitation	6	5484	185	5299	9083
Protein stress	2	587	14	573	1316
Radiation stress	1	1264	54	1210	2695
Sigma factors	4	4717	414	4303	7326
Bacteria phage	**40**	**1071**	**195**	**876**	**1987**
Replication	25	666	113	553	1188
Lysis	7	133	22	111	289
Structural	6	230	42	188	452
Host recognition/structural	2	42	18	24	58
Virulence	**13**	**3726**	**315**	**3411**	**7444**
Other (gyrB, bchY)	**2**	**2347**	**834**	**1513**	**4226**
Total	**410**	**82 074**	**18 098**	**63 976**	**141 995**

$P = 0.036$; MRPP: $\delta = 0.136$, $P = 0.028$). Further analysis of genes involved in carbon degradation, nitrogen cycling and phosphorus cycling showed significantly increased abundances for most of these genes ($P < 0.05$, Fig. S3–S5, Supporting information). In total, 13 of 27 detected genes involved in carbon degradation (five for starch degradation and three for hemi-cellulose/cellulose degradation) increased significantly (Fig. S3, Supporting information), 10 of 16 detected nitrogen cycling genes increased significantly (Fig. S4, Supporting information), and 2 of 3 phosphorous cycling genes increased significantly (Fig. S5, Supporting information). These results suggested that warming could significantly stimulate microbially mediated nutrient cycling. Such observations were consistent with our previous results based on GeoChip 3.0 with soil samples collected in 2007 from the same site (Zhou *et al.* 2012).

As new features in GeoChip 4, gene families related to microbial stress responses (Fig. 4) and bacteriophages (Fig. S6, Supporting information) were also analysed. Specifically, the abundances of all four sigma factors, including σ70 (primary sigma factor), σ24 (extracytoplasmic/extreme heat stress), σ32 (heat shock) and σ38 (starvation/stationary phase), increased significantly (P-value $\leq$0.05) in response to warming. The abundances of three genes involved in heat shock responses (*dnaK*, *grpE* and *hrcA*) also increased significantly, while genes related to cold shock responses remained unchanged. In accordance with the stimulated phosphorus cycling, genes related to phosphate limitation also increased significantly. Also, one gene (*proV*) involved in osmotic stress significantly increased in response to warming. Of 40 genes from bacteriophage, the majority showed no significant difference in response to warming. Only three genes, encoding endolysin transglycosylase, noncontractile major tail protein and scaffold protein, significantly increased in abundance under warming, and one gene for helicase P4 alpha type significantly decreased (Fig. S6, Supporting information), showing a minor effect of warming on bacteriophage in grassland soil. These results showed long-term warming significantly affected the microbial community functional structure and composition, by stimulating genes related to nutrient cycling and stress responses.

Table 3 Summary of GeoChip 4 probes based on covered microbial domains and phylum information. Viruses were classified at the order/family level

Domain	Phylum	No. genes	No. strains	No. probes	No. covered cds
Archaea		118	173	2555	4390
	Euryarchaeota	102	137	1518	2226
	Crenarchaeota	63	31	732	1893
	Korarchaeota	9	1	16	18
	Thaumarchaeota	6	3	11	15
	Nanoarchaeota	1	1	1	2
	Unclassified*	14		277	776
Bacteria		367	4138	73 106	130 347
	Proteobacteria	346	2152	39 141	71 166
	Firmicutes	210	822	10 084	20 585
	Actinobacteria	251	490	6925	11 785
	Bacteroidetes	128	212	2768	4194
	Cyanobacteria	120	142	2495	3671
	Chloroflexi	89	26	995	1489
	Chlorobi	52	20	652	844
	Verrucomicrobia	82	11	623	683
	Spirochaetes	60	42	241	592
	Planctomycetes	88	18	510	563
	Deinococcus-Thermus	77	14	344	540
	Thermotogae	41	18	290	516
	Tenericutes	33	74	207	457
	Acidobacteria	71	5	300	393
	Chlamydiae	26	21	128	341
	Aquificae	44	22	198	290
	Fusobacteria	36	21	138	228
	Lentisphaerae	26	2	77	98
	Nitrospirae	38	7	71	93
	Synergistetes	20	4	77	84
	Dictyoglomi	23	3	45	79
	Gemmatimonadetes	22	1	57	64
	Elusimicrobia	14	2	34	43
	Fibrobacteres	17	1	39	41
	Deferribacteres	15	1	30	34
	Thermodesulphobacteria	5	6	19	15
	Candidatus Poribacteria	1	1	4	3
	Unclassified*	128		6614	11 456
Eukaryota		127	404	4965	4824
	Ascomycota	112	227	3753	3517
	Basidiomycota	48	125	841	839
	Streptophyta	14	10	49	49
	Chordata	5	8	18	16
	Echinodermata	2	1	16	16
	Apicomplexa	2	8	12	15
	Arthropoda	4	7	20	13
	Neocallimastigomycota	7	5	13	11
	Bacillariophyta	3	2	11	10
	Microsporidia	4	3	5	6
	Nematoda	4	3	8	5
	Platyhelminthes	1	1	4	5
	Chlorophyta	1	2	5	3
	Glomeromycota	3	1	4	3
	Phaeophyceae	1	1	2	1
	Unclassified*	13		204	315
Viruses	(Order/Family)	40	252	1071	1997

Table 3 (Continued)

Domain	Phylum	No. genes	No. strains	No. probes	No. covered cds
	Caudovirales	39	216	954	1687
	Leviviridae	2	16	55	179
	Microviridae	2	14	19	63
	Tectiviridae	3	5	7	18
	Corticoviridae	1	1	1	1
	Unclassified*	12		35	49
Others†		29		377	833
Total		410	4967	82 074	141 995

*Unclassified refer to sequences that can only be identified at domain level, but not at phylum level or lower. Investigation of these sequences showed they are mostly from environmental samples.
†Others refer to sequences from some plasmids and uncultured/unidentified prokaryote organisms.

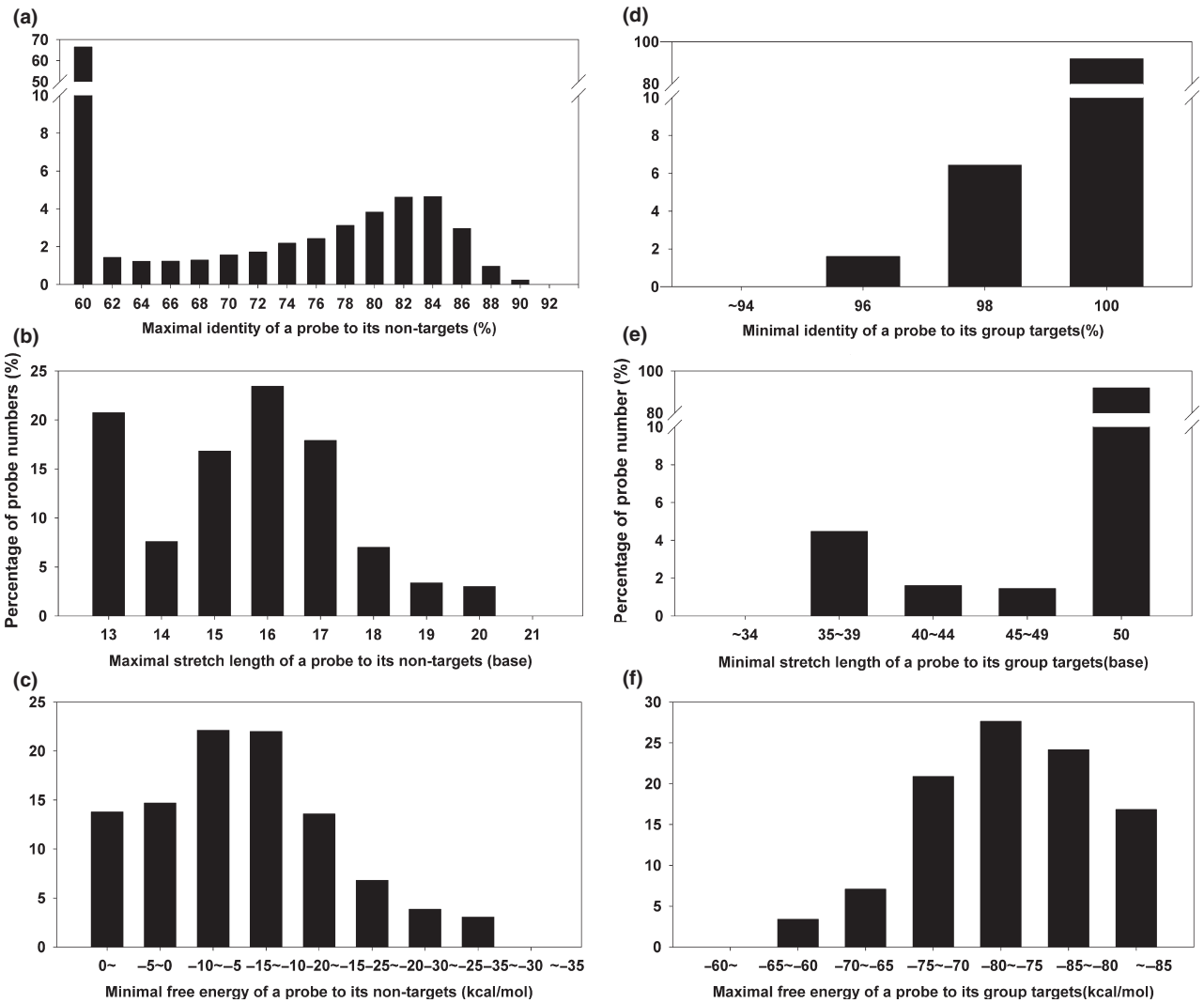

Fig. 1 Computational evaluation of sequence-specific probes at (a) maximal sequence identities, (b) maximal stretch length and (c) minimal free energy with their closest nontarget sequences; and group-specific probes at (d) minimal sequence identities, (e) minimal stretch length and (f) maximal free energy with their group targets.

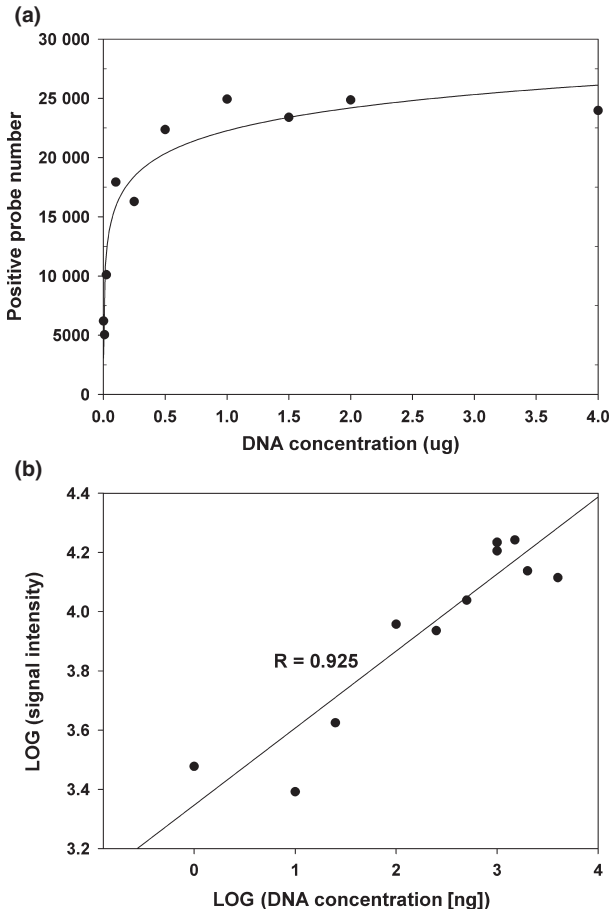

Fig. 2 Experimental assessment using environmental DNA extracted from one soil samples collected from the BioCON experimental sites with DNA amounts varying from 0.001 to 4 μg. (a) Number of positive probes detected when different amounts of DNA were used. Stable hybridization could be observed with DNA amount of more than 0.5 μg. (b) Correlation between Log (signal intensity) vs. log (DNA concentration). A high correlation coefficient value of 0.925 was observed. Three replicates were included for each DNA concentration. Probes showing up in at least two of three replicates were regarded as positive probes.

Discussion

Microarray-based metagenomic technology has been widely used for microbial community analysis, revolutionizing our understanding of microbial community structure, function and dynamics. Several different FGAs with differing purposes have been developed for microbial ecology studies (He *et al.* 2012a). These include GeoChip 2.0 and 3.0 for comprehensive microbial ecology studies (He *et al.* 2007, 2010a), the Hydrogenase Chip for characterizing hydrogen-producing and hydrogen-consuming microbes (Marshall *et al.* 2012) and a *pmoA* functional gene array for methanotrophs (Bodrossy *et al.* 2003).

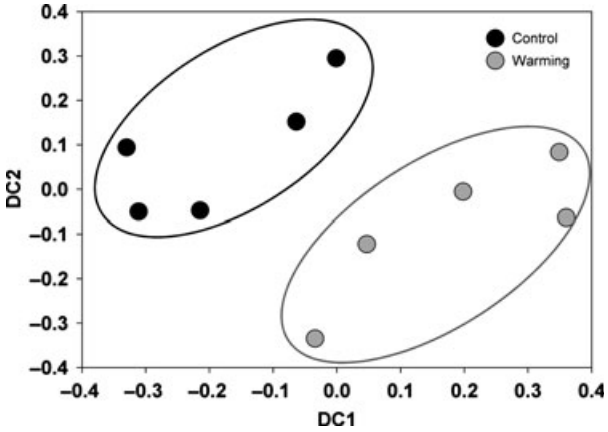

Fig. 3 Detrended correspondence analysis (DCA) of soil microbial communities using GeoChip 4 data of samples collected from long-term warming experimental site located in the Central Oklahoma. A total of 30 632 probes detected in the whole community were analysed.

The developed GeoChip 4 here is much more comprehensive than any other FGAs currently available and has several new features. First, GeoChip has been continuously updated to reflect our most current knowledge of the gene families important to biogeochemistry, ecology and environmental science. Compared with previous versions, GeoChip 4 is more comprehensive, targeting 410 gene families with ~82 000 probes covering ~142 000 CDS from more than 5200 microbial strains, including bacteria, archaea, fungi and viruses, enabling researchers to study more functional genes and microbial lineages within a single hybridization. Second, conventional gene families covered by previous GeoChips were manually checked and updated with twice as many CDS, making GeoChip 4 more comprehensive. Third, several important functional gene categories have been carefully selected and included, such as bacteriophage, microbial stress responses and virulence. In addition, GeoChip 4 has taken advantage of the *in situ* synthesized microarray format where probes are synthesized directly onto the surface of the glass slides, resulting in much smaller, more sensitive probe features. This also resulted in lower amounts of DNA, as low as 500 ng environmental DNA, being required for successful hybridizations, compared with previous versions of GeoChip that required a minimum of 2 μg DNA for microarray hybridization. All these distinct features make GeoChip 4 a more sensitive, powerful and comprehensive metagenomic tool for analysing the functional diversity, composition, structure and dynamics of microbial communities, and linking their structure to environmental factors and ecosystem functioning.

Specificity is one of the most important issues in microarray technology, especially for environmental

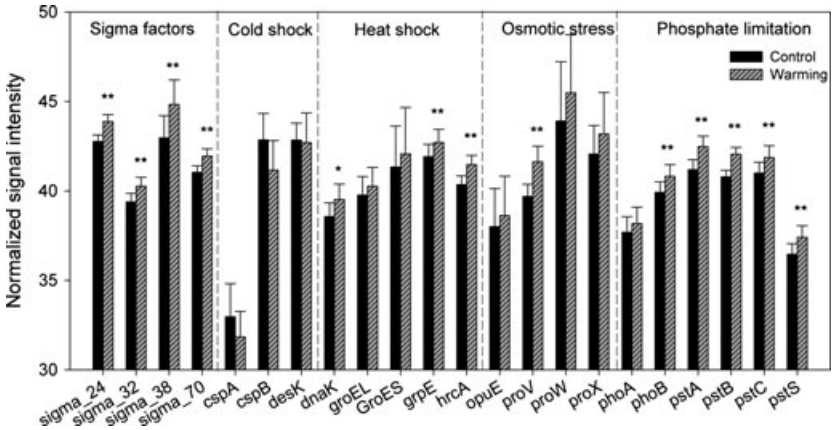

Fig. 4 The normalized average signal intensity of stress response genes related with cold/heat shock, osmotic stress, phosphate limitation and sigma factors under warming and the control. Signal intensities were the average abundances of detected genes under warming or control plots, normalized by the probe number of each gene. Error bars represent standard error. The differences between warming and control samples were tested by two-tailed paired *t*-tests. *: $P < 0.10$; **: $P < 0.05$.

samples with complex microbial communities. The specificity of GeoChip probes was guaranteed in multiple ways. First, the parameters used for designing the 50-mer probes, both sequence specific and group specific, were experimentally evaluated and established (He *et al.* 2005; Liebich *et al.* 2006). Probes designed with those parameters were highly specific to their target sequences. Second, multiple criteria were considered simultaneously for high-quality probe selection by CommOligo 2.0, ensuring all designed probes were specific to all CDS in the input file and with similar thermodynamic properties (Li *et al.* 2005). Third, all designed probes were checked against the most recent NCBI nt and env_nt databases for specificity, and nonspecific probes were discarded. Fourth, computational evaluation showed that only a very small portion of the designed probes (5%) was very close to the criterion thresholds (He *et al.* 2005). Finally, extensive evaluations for functional gene probes designed with the same criteria were carried out in the laboratory during the past 15 years using pure culture DNA, mock community DNA and environmental samples, suggesting high specificity and sensitivity for these probes (Wu *et al.* 2001, 2006; Rhee *et al.* 2004; Tiquia *et al.* 2004; He *et al.* 2007, 2010a).

Application of GeoChip 4 to soil microbial communities sampled in 2008 from a long-term warming site showed consistent results with a previous study of samples collected in 2007 from the same site using GeoChip 3.0 (Zhou *et al.* 2012). Similar abundance patterns of functional genes involved in nutrient cycling processes such as carbon, nitrogen and phosphorous cycling processes were found in both studies, indicating GeoChip results were generally reproducible for similar samples collected from the same experimental site using different versions of the microarray, but with more information. The results indicated several significant mechanisms by which microbial communities responded to climate warming. First, increased abundances of functional genes involved in carbon degradation suggested soil microbial communities play an important role of carbon cycling in response to long-term warming. Second, by increasing the abundance of genes involved in N and P cycling processes, nutrient cycling processes are enhanced and possibly promote plant nutrient use efficiency and plant growth. Third, long-term warming could be a stressor for soil microbial communities, which may adapt to such a stress by increasing the abundance of corresponding stress genes (e.g., sigma factor, heat shock, phosphate limitation genes). All these results indicate that microbial communities may play an important role in response to long-term warming and that the contribution of soil microbial communities should be considered in studying and predicting ecosystem feedbacks to climate warming.

In conclusion, a more comprehensive version of GeoChip 4 was developed in this study. Computational and experimental evaluations indicated GeoChip 4 to be a specific, sensitive and quantitative tool for microbial ecology studies. Application of GeoChip 4 to analyse soil microbial communities under long-term warming indicated warming significantly affected the functional composition and structure of soil microbial communities. To the best of our knowledge, GeoChip 4 is the most comprehensive functional gene array to study the functional diversity, composition, structure and dynamics of microbial communities, and to link their structure to environmental factors and ecosystem functioning. Further development of GeoChip will include more gene families for currently covered and uncovered biogeochemical processes, and new sequence variants from metagenome sequencing projects.

Acknowledgements

This work was conducted by ENIGMA-Ecosystems and Networks Integrated with Genes and Molecular Assemblies (http://enigma.lbl.gov), a Scientific Focus Area Program at Lawrence Berkeley National Laboratory, was supported by the Office of Science, Office of Biological and Environmental Research (OBER), of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. This study was also supported by the OBER Biological Systems Research on the Role of Microbial Communities in Carbon Cycling Program (DE-SC0004601), by the U.S. National Science Foundation MacroSystems Biology programme under the contract (NSF EF-1065844) and by Oklahoma Applied Research Support (OARS), Oklahoma Center for the Advancement of Science and Technology (OCAST) through the Projects AR062-034 and AR11-035, the State of Oklahoma.

Conflict of interest

None declared.

References

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

Ansorge WJ (2009) Next-generation DNA sequencing techniques. *New Biotechnology*, **25**, 195–203.

Ashelford KE, Day MJ, Fry JC (2003) Elevated Abundance of Bacteriophage Infecting Bacteria in Soil. *Applied and Environmental Microbiology*, **69**, 285–289.

Bardgett RD, Freeman C, Ostle NJ (2008) Microbial contributions to climate change through carbon cycle feedbacks. *ISME Journal*, **2**, 805–814.

Bergh O, Borsheim KY, Bratbak G, Heldal M (1989) High abundance of viruses found in aquatic environments. *Nature*, **340**, 467–468.

Bodrossy L, Stralis-Pavese N, Murrell JC *et al.* (2003) Development and validation of a diagnostic microbial microarray for methanotrophs. *Environmental Microbiology*, **5**, 566–582.

Bossier P, Hofte M, Verstraete W (1988) Ecological significance of siderophores in soil. *Advances in microbial ecology*, **10**, 385–414.

Brisson VL, West KA, Lee PK *et al.* (2012) Metagenomic analysis of a stable trichloroethene-degrading microbial community. *ISME Journal*, **6**, 1702–1714.

Brodie EL, DeSantis TZ, Joyner DC *et al.* (2006) Application of a High-Density Oligonucleotide Microarray Approach To Study Bacterial Population Dynamics during Uranium Reduction and Reoxidation. *Applied and Environmental Microbiology*, **72**, 6288–6298.

Brodie EL, DeSantis TZ, Parker JPM *et al.* (2007) Urban aerosols harbor diverse and dynamic bacterial populations. *Proceedings of the National Academy of Sciences, USA*, **104**, 299–304.

Carter MQ, Chen J, Lory S (2010) The Pseudomonas aeruginosa pathogenicity island PAPI-1 is transferred via a novel type IV pilus. *Journal of Bacteriology*, **192**, 3249–3258.

Chen J, Novick RP (2009) Phage-mediated intergeneric transfer of toxin genes. *Science*, **323**, 139–141.

Cossart P, Jonquières R (2000) Sortase, a universal target for therapeutic agents against Gram-positive bacteria? *Proceedings of the National Academy of Sciences, USA*, **97**, 5013–5015.

Craig L, Taylor RK, Pique ME *et al.* (2003) Type IV pilin structure and assembly: X-ray and EM analyses of Vibrio cholerae toxin-coregulated pilus and Pseudomonas aeruginosa PAK pilin. *Molecular Cell*, **11**, 1139–1150.

Deng Y, He Z, Xu M *et al.* (2012) Elevated carbon dioxide alters the structure of soil microbial communities. *Applied and Environmental Microbiology*, **78**, 2991–2995.

Díez B, Bergman B, Pedrós-Alió C, Antó M, Snoeijs P (2012) High cyanobacterial nifH gene diversity in Arctic seawater and sea ice brine. *Environmental Microbiology Reports*, **4**, 360–366.

Dong H, Nilsson L, Kurland CG (1995) Gratuitous overexpression of genes in Escherichia coli leads to growth inhibition and ribosome destruction. *Journal of Bacteriology*, **177**, 1497–1504.

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

Finlay BB, Falkow S (1997) Common themes in microbial pathogenicity revisited. *Microbiology and Molecular Biology Reviews*, **61**, 136–169.

Fuangthong M, Herbig AF, Bsat N, Helmann JD (2002) Regulation of the Bacillus subtilis fur and perR genes by PerR: not all members of the PerR regulon are peroxide inducible. *Journal of Bacteriology*, **184**, 3276–3286.

Galan JE, Collmer A (1999) Type III secretion machines: bacterial devices for protein delivery into host cells. *Science*, **284**, 1322–1328.

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

Goff SA, Goldberg AL (1985) Production of abnormal proteins in E. coli stimulates transcription of ion and other heat shock genes. *Cell*, **41**, 587–595.

Gomez P, Buckling A (2011) Bacteria-phage antagonistic coevolution in soil. *Science*, **332**, 106–109.

Grath SM, Sinderen DV (2007) *Bacteriophage: Genetics and Molecular Biology*. Caister Academic Press, Norfolk, UK.

Gruber N, Galloway JN (2008) An Earth-system perspective of the global nitrogen cycle. *Nature*, **451**, 293–296.

Hazen TC, Dubinsky EA, DeSantis TZ *et al.* (2010) Deep-Sea Oil Plume Enriches Indigenous Oil-Degrading Bacteria. *Science*, **330**, 204–208.

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

He Z, Wu L, Li X, Fields MW, Zhou J (2005) Empirical Establishment of Oligonucleotide Probe Design Criteria. *Applied and Environmental Microbiology*, **71**, 3753–3760.

He Z, Gentry TJ, Schadt CW *et al.* (2007) GeoChip: a comprehensive microarray for investigating biogeochemical, ecological and environmental processes. *ISME Journal*, **1**, 67–77.

He Z, Deng Y, Van Nostrand JD *et al.* (2010a) GeoChip 3.0 as a high-throughput tool for analyzing microbial community composition, structure and functional activity. *ISME Journal*, **4**, 1167–1179.

He Z, Xu M, Deng Y *et al.* (2010b) Metagenomic analysis reveals a marked divergence in the structure of belowground microbial communities at elevated CO2. *Ecology Letters*, **13**, 564–575.

He Z, Deng Y, Zhou J (2012a) Development of functional gene microarrays for microbial community analysis. *Current Opinion in Biotechnology*, **23**, 49–55.

He Z, Van Nostrand JD, Zhou J (2012b) Applications of functional gene microarrays for profiling microbial communities. *Current Opinion in Biotechnology*, **23**, 460–466.

Hess M, Sczyrba A, Egan R *et al.* (2011) Metagenomic Discovery of Biomass-Degrading Genes and Genomes from Cow Rumen. *Science*, **331**, 463–467.

Hoffmann-Berling H, Mazé R (1964) Release of male-specific bacteriophages from surviving host bacteria. *Virology*, **22**, 305–313.

Huang WE, Zhou J (2012) When single cell technology meets omics, the new toolbox of analytical biotechnology is emerging. *Current Opinion in Biotechnology*, **23**, 1.

Kallmeyer J, Pockalny R, Adhikari RR, Smith DC, D'Hondt S (2012) Global distribution of microbial abundance and biomass in subseafloor sediment. *Proceedings of the National Academy of Sciences, USA*, **109**, 16213–16216.

Kempf B, Bremer E (1998) Uptake and synthesis of compatible solutes as microbial stress responses to high-osmolality environments. *Archives of Microbiology*, **170**, 319–330.

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. *Environmental Microbiology*, **12**, 541–556.

Kline KA, Fälker S, Dahlberg S, Normark S, Henriques-Normark B (2009) Bacterial Adhesins in Host-Microbe Interactions. *Cell Host & Microbe*, **5**, 580–592.

Lasken RS (2007) Single-cell genomic sequencing using Multiple ''Displacement Amplification. *Current Opinion in Microbiology*, **10**, 510–516.

Leigh MB, Pellizari VH, Uhlik O *et al.* (2007) Biphenyl-utilizing bacteria and their functional genes in a pine root zone contaminated with polychlorinated biphenyls (PCBs). *ISME Journal*, **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 Research*, **33**, 6114–6123.

Liang Y, Li G, Van Nostrand JD *et al.* (2009) Microarray-based analysis of microbial functional diversity along an oil contamination gradient in oil field. *FEMS Microbiology Ecology*, **70**, 324–333.

Liang Y, He Z, Wu L *et al.* (2010) Development of a Common Oligonucleotide Reference Standard for Microarray Data Normalization and Comparison across Different Microbial Communities. *Applied and Environmental Microbiology*, **76**, 1088–1094.

Liang Y, Van Nostrand JD, Deng Y *et al.* (2011) Functional gene diversity of soil microbial communities from five oil-contaminated fields in China. *ISME Journal*, **5**, 403–413.

Liebich J, Schadt CW, Chong SC *et al.* (2006) Improvement of Oligonucleotide Probe Design Criteria for Functional Gene Microarrays in Environmental Applications. *Applied and Environmental Microbiology*, **72**, 1688–1691.

Liebich J, Wachtmeister T, Zhou J, Burauel P (2009) Degradation of Diffuse Pesticide Contaminants: Screening for Microbial Potential Using a Functional Gene Microarray. *Vadose Zone Journal*, **8**, 703–710.

Liu W, Wang A, Cheng S *et al.* (2010) Geochip-Based Functional Gene Analysis of Anodophilic Communities in Microbial Electrolysis Cells under Different Operational Modes. *Environmental Science & Technology*, **44**, 7729–7735.

Liu W, Wang A, Sun D *et al.* (2012) Characterization of microbial communities during anode biofilm reformation in a two-chambered microbial electrolysis cell (MEC). *Journal of Biotechnology*, **157**, 628–632.

Luo Y, Wan S, Hui D, Wallace LL (2001) Acclimatization of soil respiration to warming in a tall grass prairie. *Nature*, **413**, 622–625.

Mackelprang R, Waldrop MP, DeAngelis KM *et al.* (2011) Metagenomic analysis of a permafrost microbial community reveals a rapid response to thaw. *Nature*, **480**, 368–371.

MacLean D, Jones JDG, Studholme DJ (2009) Application of 'next-generation' sequencing technologies to microbial genetics. *Nature Reviews Microbiology*, **7**, 287–296.

Mardis ER (2008) The impact of next-generation sequencing technology on genetics. *Trends in genetics*, **24**, 133.

Marshall IP, Berggren DR, Azizian MF *et al.* (2012) The Hydrogenase Chip: a tiling oligonucleotide DNA microarray technique for characterizing hydrogen-producing and -consuming microbes in microbial communities. *ISME Journal*, **6**, 814–826.

Mason OU, Nakagawa T, Rosner M *et al.* (2010) First Investigation of the Microbiology of the Deepest Layer of Ocean Crust. *PLoS ONE*, **5**, e15399.

Mazmanian SK, Ton-That H, Schneewind O (2001) Sortase-catalysed anchoring of surface proteins to the cell wall of Staphylococcus aureus. *Molecular Microbiology*, **40**, 1049–1057.

Metzker ML (2010) Sequencing technologies [mdash] the next generation. *Nature Reviews Genetics*, **11**, 31–46.

Neilands JB (1995) Siderophores: Structure and Function of Microbial Iron Transport Compounds. *Journal of Biological Chemistry*, **270**, 26723–26726.

Pomposiello PJ, Demple B (2001) Redox-operated genetic switches: the SoxR and OxyR transcription factors. *Trends in Biotechnology*, **19**, 109–114.

Prescott LM (1993) *Microbiology*. William C Brown Pub, Boston, MA.

Qin J, Li R, Raes J *et al.* (2010) A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, **464**, 59–65.

Qin J, Li Y, Cai Z *et al.* (2012) A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*, **490**, 55–60.

Rappe MS, Giovannoni SJ (2003) The uncultured microbial majority. *Annual Review of Microbiology*, **57**, 369–394.

Reich PB, Knops J, Tilman D *et al.* (2001) Plant diversity enhances ecosystem responses to elevated CO2 and nitrogen deposition. *Nature*, **410**, 809–810.

Rhee SK, Liu X, Wu L *et al.* (2004) Detection of genes involved in biodegradation and biotransformation in microbial communities by using 50-mer oligonucleotide microarrays. *Applied and Environment Microbiology*, **70**, 4303–4317.

Riede I, Drexler K, Eschbach M-L, Henning U (1987) DNA sequence of genes 38 encoding a receptor-recognizing protein of bacteriophages T2, K3 and of K3 host range mutants. *Journal of Molecular Biology*, **194**, 31–39.

Roesch LFW, Fulthorpe RR, Riva A *et al.* (2007) Pyrosequencing enumerates and contrasts soil microbial diversity. *ISME Journal*, **1**, 283–290.

Schatz MC, Phillippy AM, Gajer P *et al.* (2010) Integrated microbial survey analysis of prokaryotic communities for the PhyloChip microarray. *Applied and Environmental Microbiology*, **76**, 5636–5638.

Schloss PD, Handelsman J (2006) Toward a census of bacteria in soil. *PLoS Computational Biology*, **2**, 5.

Shendure J, Ji H (2008) Next-generation DNA sequencing. *Nature Biotechnology*, **26**, 1135–1145.

Singh A, Wyant T, Anaya-Bergman C *et al.* (2011) The capsule of Porphyromonas gingivalis leads to a reduction in the host inflammatory response, evasion of phagocytosis, and increase in virulence. *Infection and Immunity*, **79**, 4533–4542.

Sintes E, Bergauer K, De CD, Yokokawa T, Herndl GJ (2013) Archaeal amoA gene diversity points to distinct biogeography of ammonia-oxidizing Crenarchaeota in the ocean. *Environmental Microbiology*, **15**, 1647–1658.

Sogin ML, Morrison HG, Huber JA *et al.* (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proceedings of the National Academy of Sciences, USA*, **103**, 12115–12120.

Suttle CA (1994) The Significance of Viruses to Mortality in Aquatic Microbial Communities. *Microbial Ecology*, **28**, 237–243.

Suttle CA (2007) Marine viruses–major players in the global ecosystem. *Nature Reviews Microbiology*, **5**, 801–812.

Taş N, van Eekert MHA, Schraa G *et al.* (2009) Tracking Functional Guilds: "Dehalococcoides" spp. in European River Basins Contaminated with Hexachlorobenzene. *Applied and Environmental Microbiology*, **75**, 4696–4704.

Taylor LH, Latham SM, Woolhouse ME (2001) Risk factors for human disease emergence. *Philosophical Transactions of the Royal Society of London. Series B, Biological sciences*, **356**, 983–989.

Thomassen E, Gielen G, Schutz M *et al.* (2003) The structure of the receptor-binding domain of the bacteriophage T4 short tail fibre reveals a knitted trimeric metal-binding fold. *Journal of Molecular Biology*, **331**, 361–373.

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

Tobias J, Svennerholm AM (2012) Strategies to overexpress enterotoxigenic Escherichia coli (ETEC) colonization factors for the construction of oral whole-cell inactivated ETEC vaccine candidates. *Applied Microbiology and Biotechnology*, **93**, 2291–2300.

Trivedi P, He Z, Van Nostrand JD *et al.* (2012) Huanglongbing alters the structure and functional diversity of microbial communities associated with citrus rhizosphere. *ISME Journal*, **6**, 363–383.

Van Nostrand JD, Wu W-M, Wu L *et al.* (2009) GeoChip-based analysis of functional microbial communities during the reoxidation of a

bioreduced uranium-contaminated aquifer. *Environmental Microbiology*, **11**, 2611–2626.
Walker A, Parkhill J (2008) Single-cell genomics. *Nature Reviews Microbiology*, **6**, 176–177.
Wang F, Zhou H, Meng J *et al.* (2009) GeoChip-based analysis of metabolic diversity of microbial communities at the Juan de Fuca Ridge hydrothermal vent. *Proceedings of the National Academy of Sciences, USA*, **106**, 4840–4845.
Weigel C, Seitz H (2006) Bacteriophage replication modules. *FEMS Microbiology Reviews*, **30**, 321–381.
Weinbauer MG (2004) Ecology of prokaryotic viruses. *FEMS Microbiology Reviews*, **28**, 127–181.
Whitman WB, Coleman DC, Wiebe WJ (1998) Prokaryotes: The unseen majority. *Proceedings of the National Academy of Sciences, USA*, **95**, 6578–6583.
Wommack KE, Colwell RR (2000) Virioplankton: viruses in aquatic ecosystems. *Microbiology and Molecular Biology Reviews*, **64**, 69–114.
Woolhouse ME, Gowtage-Sequeria S (2005) Host range and emerging and reemerging pathogens. *Emerging Infectious Diseases*, **11**, 1842–1847.
Woyke T, Tighe D, Mavromatis K *et al.* (2010) One bacterial cell, one complete genome. *PLoS ONE*, **5**, e10314.
Wu L, Thompson DK, Li G *et al.* (2001) Development and Evaluation of Functional Gene Arrays for Detection of Selected Genes in the Environment. *Applied and Environmental Microbiology*, **67**, 5780–5790.
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. *Applied and Environmental Microbiology*, **72**, 4931–4941.
Wu HJ, Wang AH, Jennings MP (2008) Discovery of virulence factors of pathogenic bacteria. *Current Opinion in Chemical Biology*, **12**, 93–101.
Xiong J, Wu L, Tu S *et al.* (2010) Microbial Communities and Functional Genes Associated with Soil Arsenic Contamination and the Rhizosphere of the Arsenic-Hyperaccumulating Plant Pteris vittata L. *Applied and Environmental Microbiology*, **76**, 7277–7284.
Xu M, Wu W-M, Wu L *et al.* (2010) Responses of microbial community functional structures to pilot-scale uranium in situ bioremediation. *ISME Journal*, **4**, 1060–1070.
Yang S, Bourne PE (2009) The evolutionary history of protein domains viewed by species phylogeny. *PLoS ONE*, **4**, 0008378.
Yatsunenko T, Rey FE, Manary MJ *et al.* (2012) Human gut microbiome viewed across age and geography. *Nature*, **486**, 222–227.
Yergeau E, Bokhorst S, Kang S *et al.* (2012) Shifts in soil microorganisms in response to warming are consistent across a range of Antarctic environments. *ISME Journal*, **6**, 692–702.
Young I, Wang I, Roof WD (2000) Phages will out: strategies of host cell lysis. *Trends in Microbiology*, **8**, 120–128.
Zhou J, Bruns MA, Tiedje JM (1996) DNA recovery from soils of diverse composition. *Applied and Environmental Microbiology*, **62**, 316–322.
Zhou J, Kang S, Schadt CW, Garten CT (2008) Spatial scaling of functional gene diversity across various microbial taxa. *Proceedings of the National Academy of Sciences, USA*, **105**, 7768–7773.
Zhou J, Xue K, Xie J *et al.* (2012) Microbial mediation of carbon-cycle feedbacks to climate warming. *Nature Climate Change*, **2**, 106–110.
Zhou J, Liu W, Deng Y *et al.* (2013) Stochastic Assembly Leads to Alternative Communities with Distinct Functions in a Bioreactor Microbial Community. *mBio*, **4**, e00584–e00612.

Q.T., Z.H., J.V.N. and J.Z. wrote the manuscript. H.Y., K.X., L.W., and T.Y. designed and performed microarray experiments. Q.T., H.Y. and K.X. processed and analysed the microarray data. Q.T., Y.D., Y.Q. and Z.S. built the GeoChip probe design pipeline. A.Z., J.V., Y.L. and CH contributed to keywords generation for newly added gene families. A.W. and J.Z. oversaw the whole study. All authors read and approved the final manuscript.

Data Accessibility

Normalized GeoChip 4 data for the long-term warming effects on soil microbial communities study at a Central Oklahoma site are available at http://ieg.ou.edu/4download. Details for the all covered gene families in GeoChip 4 are also available at the above website.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Figure S1 Flowchart of GeoChip 4.0 design pipeline.

Figure S2 Illustration of the layout of GeoChip 4.0 probes.

Figure S3 The normalized average signal intensity of genes involved in carbon degradation process under warming and the control conditions.

Figure S4 The normalized average signal intensity of genes involved in nitrogen cycling process under warming and the control.

Figure S5 The normalized average signal intensity of genes involved in phosphorus cycling process under warming and the control.

Figure S6 The normalized average signal intensity of genes related with bacteriophage under warming and the control.