The ISME Journal (2013) 7, 1974–1984

www.nature.com/ismej

ORIGINAL ARTICLE

# The PathoChip, a functional gene array for assessing pathogenic properties of diverse microbial communities

Yong-Jin Lee[1], Joy D van Nostrand[1], Qichao Tu[1], Zhenmei Lu[1,2], Lei Cheng[1], Tong Yuan[1], Ye Deng[1], Michelle Q Carter[3], Zhili He[1], Liyou Wu[1], Fang Yang[4], Jian Xu[5] and Jizhong Zhou[1,6,7]

[1]*Institute for Environmental Genomics and Department of Microbiology and Plant Biology, University of Oklahoma, Norman, OK, USA;* [2]*College of Life Sciences, Zhejiang University, Hangzhou, China;* [3]*Produce Safety and Microbiology Unit, Western Regional Research Center, Agricultural Research Service, USDA, Albany, CA, USA;* [4]*Oral Research Center, Qingdao Municipal Hospital, Qingdao, Shandong Province, China;* [5]*Chinese Academy of Sciences, Qingdao Institute of Bioenergy and Bioprocess Technology, Qingdao, Shandong, China;* [6]*Earth Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA and* [7]*State Key Joint Laboratory of Environment Simulation and Pollution Control, School of Environment, Tsinghua University, Beijing, China*

**Pathogens present in the environment pose a serious threat to human, plant and animal health as evidenced by recent outbreaks. As many pathogens can survive and proliferate in the environment, it is important to understand their population dynamics and pathogenic potential in the environment. To assess pathogenic potential in diverse habitats, we developed a functional gene array, the PathoChip, constructed with key virulence genes related to major virulence factors, such as adherence, colonization, motility, invasion, toxin, immune evasion and iron uptake. A total of 3715 best probes were selected from 13 virulence factors, covering 7417 coding sequences from 1397 microbial species (2336 strains). The specificity of the PathoChip was computationally verified, and approximately 98% of the probes provided specificity at or below the species level, proving its excellent capability for the detection of target sequences with high discrimination power. We applied this array to community samples from soil, seawater and human saliva to assess the occurrence of virulence genes in natural environments. Both the abundance and diversity of virulence genes increased in stressed conditions compared with their corresponding controls, indicating a possible increase in abundance of pathogenic bacteria under environmental perturbations such as warming or oil spills. Statistical analyses showed that microbial communities harboring virulence genes were responsive to environmental perturbations, which drove changes in abundance and distribution of virulence genes. The PathoChip provides a useful tool to identify virulence genes in microbial populations, examine the dynamics of virulence genes in response to environmental perturbations and determine the pathogenic potential of microbial communities.**

*The ISME Journal* (2013) **7,** 1974–1984; doi:10.1038/ismej.2013.88; published online 13 June 2013
**Subject Category:** Integrated genomics and post-genomics approaches in microbial ecology
**Keywords:** virulence genes; functional gene array; climate warming; oil-contamination; caries

## Introduction

Presence and persistence of pathogens in natural environments may pose potential threats to public health. For example, many opportunistic pathogens reside most of their life cycle in non-host environments (Friman *et al.*, 2011), and these pathogens can be transmitted to hosts, including humans, and cause outbreaks and epidemics (Yildiz, 2007). Such public concerns become worldwide as climate change and globalization become visible and pervasive (Smith *et al.*, 2007; Jones *et al.*, 2008; Slenning, 2010). Recent disease outbreaks, emergence and re-emergence of pathogens and zoonoses all point to a possible link to the survival and growth of pathogens in the environment. Indeed, it has been reported that there is a positive correlation between bacterial virulence and their survival in environmental reservoirs (Friman *et al.*, 2011). Because natural environments serve as reservoirs and transmission routes for many pathogens (Woolhouse and

Correspondence: J Zhou, Institute for Environmental Genomics and Department of Microbiology and Plant Biology, University of Oklahoma, 101 David L. Boren Boulevard, Norman, OK 73019, USA.
E-mail: jzhou@ou.edu
Received 22 January 2013; revised 10 April 2013; accepted 20 April 2013; published online 13 June 2013

Gaunt, 2007), information on the virulence properties of an environment, such as abundance and diversity of virulence genes and their shifts in response to environmental perturbations, is critical to understand the nature and extent of this potential threat and to identify the source and transmission routes during disease outbreaks or epidemics.

Bacteria, the most dominant group of pathogens, account for approximately 40% of all pathogens followed by fungi, helminths, viruses and prions and protozoa (Taylor *et al.*, 2001; Woolhouse and Gowtage-Sequeria, 2005). Bacterial pathogens have evolved to produce a number of virulence factors that may directly or indirectly have a role in establishing infection and causing disease. However, knowledge on the diversity and distribution of these factors in the environment is scarce. Thus, a more comprehensive and intensive survey of environmental virulence profiles is necessary. Various methods have been developed to detect and identify bacterial pathogens from environmental samples. Among them, microarray-based technology offers a powerful and high-throughput tool and has been developed to detect and identify pathogens (Wu *et al.*, 2003; Bruant *et al.*, 2006; Anjum *et al.*, 2007; Tembe *et al.*, 2007; Jaing *et al.*, 2008; Miller *et al.*, 2008; Geue *et al.*, 2010; Hyman *et al.*, 2010; Peterson *et al.*, 2010; Quinones *et al.*, 2011). However, many of these methods rely on a limited number of genetic markers that were designed for detection of specific pathogens.

A recently developed functional gene array, GeoChip, has shown to be a reliable and comprehensive tool for analyzing the functional diversity, composition and metabolic potential of microbial communities (He *et al.*, 2007; Zhou *et al.*, 2008; Van Nostrand *et al.*, 2009; Waldron *et al.*, 2009; Wang *et al.*, 2009; He *et al.*, 2010b; Lu *et al.*, 2012; Trivedi *et al.*, 2012; Zhou *et al.*, 2012). In this study, we employed a GeoChip-based strategy and developed a comprehensive functional gene array, the PathoChip, targeting the major virulence factors described in bacterial pathogens. Because the pathogenicity of most bacterial pathogens is endowed by numerous virulence factors directly or indirectly, genes encoding such factors are excellent indicators to assess the virulence potential of a given environment. We further applied the developed PathoChip to assess virulence gene composition and structure, changes of the pathogenic properties of microbial communities and their responses to environmental stresses in three distinct environments (soil, seawater and human saliva).

## Materials and methods

*Designing oligonucleotide probes with selected virulence genes*

A virulence gene array was developed based on the major virulence factors present in most bacterial pathogens, including genes involved in adhesion, colonization, motility, invasion, toxin production, immune evasion and iron uptake (Finlay and Falkow, 1997; Wu *et al.*, 2008). The 50-mer probes targeting virulence genes as of May 2010 were designed using the GeoChip pipeline (Figure 1). Briefly, keywords related to virulence factors were used to search the GenBank database to retrieve specific gene sequences. The retrieved sequences were verified by HMMER 2.3.2 (Eddy, 1998) and then used to design specific probes using CommOligo 2.0 (Li *et al.*, 2005). The criteria for designing probes were set in terms of sequence identity, continuous stretch length and free energy. For the sequence-specific probes, maximal sequence identity, maximal stretch length and minimal free energy with non-targets were set at 90%, 20 bases and $-35\,\mathrm{kcal\,mol^{-1}}$, respectively. For the group-specific probes, minimal sequence identity, minimal stretch length and maximal free energy with targets were set at 96%, 35 bases and $-60\,\mathrm{kcal\,mol^{-1}}$, respectively. The best sets of oligonucleotide probes were selected by CommOligo 2.0, and their specificity were further validated through BLAST analysis. A total of 3715 best probes were fabricated on the NimbleGen array platform.

*Details on probes designed from 13 virulence factors*

Both toxins and hemolysins have an important role in toxigenesis by affecting and damaging a host cell directly and aggressively. A total of 70 probes with 5 sequence-specific probes and 65 group-specific probes were designed from 14 toxin-related genes, including cytolethal distending toxin A, B and C,

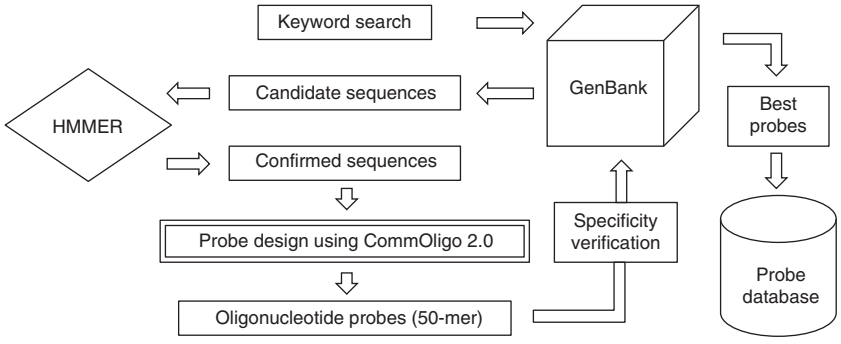


**Figure 1** Scheme of constructing a functional gene array using virulence genes.

toxin B, murine toxin, shiga toxin IA and IIA, exfoliative toxin A, B and D, epsilon toxin, RTX toxin A and diphtheria toxin. The hemolysin-related probes were comprised of 620 probes (26 sequence-specific and 594 group-specific), targeting mostly hemolysin III and others, including thermostable direct hemolysin-related hemolysin, heat-labile hemolysin, thermostable hemolysin, adenylate cyclase hemolysin, TlyC family hemolysin, VHH/TLH hemolysin, hemolysin A and B and hemolysin II.

Adhesins are cell-surface components or appendages of bacteria that facilitate bacterial colonization within their hosts (Kline *et al.*, 2009). Probes in this gene category consisted of 113 (7 sequence-specific and 106 group-specific) probes and were designed from adhesin A, sialic acid-binding adhesin, adhesin Aha1, Dr adhesin, AFA-III adhesin, P fimbrial adhesin PapG, autotransporter adhesin, adhesin protein HpaA, adhesin MafA, MafB2, adhesin B, type V secretory pathway adhesin AidA, fimbrial adhesin FimH, pilus adhesin HifE, collagen adhesin Cna, collagen-binding adhesin Cnm, adhesin P1, adhesin Hia, F17 fimbrial adhesin protein and adhesin 20 K.

Pilin is the major subunit protein of pili in many bacteria (Craig *et al.*, 2003). Besides motility, bacterial pili also have roles in surface attachment and DNA transfer by conjugation (Yang and Bourne, 2009; Carter *et al.*, 2010). A total of 1037 (49 sequence-specific and 988 group-specific) probes related to pilin were designed targeting type IV pilin PilA, type IV pili biogenesis protein PilE, MSHA pilin protein MshABCD, V10 pilin, fimbrial protein EcpC, fimbrial protein pilin FimTU and toxin-coregulated pilin subunit precursor TcpA. Unlike pili, fimbriae are short proteinaceous appendages present in many Gram-negative bacteria and some Gram-positive bacteria and also mediate the surface attachment (Blomfield *et al.*, 1993). The fimbriae-related probes consisted of 16 (5 sequence-specific and 11 group-specific) probes, targeting adhesin F41, F18 fimbrial adhesin FedEF, K88 fimbrial protein FaeG, fimbrial subunit F17A, S fimbrial adhesin major subunit SfaA and AC/I pili.

Bacterial capsules promote virulence by reducing host immune responses (Singh *et al.*, 2011). Probes of capsule-related genes were comprised of 38 sequence-specific and 347 group-specific probes and designed from capsule biosynthesis protein CapABCD, polysialic acid capsule expression protein KpsF, capsule polysaccharide export protein KpsC, capsule polysaccharide export protein KpsS, capsule polysaccharide modification protein LipAB, capsule biosynthesis protein YwsC, capsule polysaccharide modification protein HcsB, capsule polysaccharide biosynthesis/export protein GfcE, capsule biosynthesis protein SiaB and capsule polysaccharide export protein PhyAB.

Colonization factors are surface structures that allow bacteria to bind and colonize on host cells (Tobias and Svennerholm, 2012). A total of 27 (1 sequence-specific and 26 group-specific) probes of colonization factors were designed from accessory colonization factor AcfA, tracheal colonization factor TcfA, colonization factor antigen 1 and antigen b, major fimbriae subunit CsfA and CsaB, CS14 major fimbrial subunit CsuA1 and A2 and CS17 fimbriae major subunit CsbA.

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 of surface proteins, adhesion to host cells and colonization of tissues (Cossart and Jonquières, 2000; Mazmanian *et al.*, 2001). Probes in this gene category comprised 14 sequence-specific and 271 group-specific probes targeting genes *srtABCDF*.

Many Gram-negative bacterial pathogens rely on the type III secretion system, one of the most complex protein secretion systems and particularly prevalent among Gram-negative bacterial pathogens to transport effectors directly into their hosts cells (Galan and Collmer, 1999). A total of 288 (84 sequence-specific and 204 group-specific) probes were designed for type III secretion system. These probes were designed from translocation protein PscU, type III secretion proteins HrcCJQRTUV, EscRV, RhcV, RscU, SsaR, BasJ, BsaQ, FlhA and HrpEJQ, type III secretion component proteins HrpE, PsaN, SctCRU and SpaPS, type III secretion invasion protein InvA, type III secretion system ATP synthases FliI/YscN, HrcN, SsaN and SctN, type III secretion inner membrane protein YscR, type III secretion inner membrane channel proteins YscV and BcrD, type III secretion pathway protein LcrD/SctV, type III secretion system apparatus proteins VcrD2, EpaP and SsaV, type III secretion FHIPEP protein, type III secretion outer membrane protein and type III secretion effector delivery regulator.

A total of 64 and 73 probes were designed for invasins and virulence proteins, respectively. Probes for invasins consisted of 12 sequence-specific and 52 group-specific probes, targeting invasion protein Inv1 and 2, InvAE, IbeA, IagB, CipA, IpaB, HilA, SipAB and YopH, oxygen-regulated invasion protein OrgA, invasion and intracellular persistence protein IipB and invasion-associated protein p60. Probes for virulence proteins included 11 sequence-specific and 62 group-specific probes specific to CrfA, SrfB, EsaA, EssB, pGP2-D and IpgD, surface-exposed virulence protein BigA, iron-regulated outer membrane virulence protein IrgA, adherence and virulence protein A and virulence proteins S and Q.

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 numbers of probes developed for siderophore and aerobactin were 602 (57 sequence-specific) and 125 (7 sequence-specific), respectively. Among the

probes designed from these gene categories are Iro protein, TonB-dependent siderophore receptor, siderophore staphylobactin biosynthesis protein SbnG, rhizobactin siderophore biosynthesis protein EntF3 and RhbD, catecholate siderophore receptor Fiu, vulnibactin siderophore synthesis protein VenB, siderophore non-ribosomal peptide synthetase MbaF, aerobactin siderophore biosynthesis protein IucABCD and ferric aerobactin receptor IutA.

*Microarray analysis of environmental samples*

Details about sample collection and processing, DNA extraction and microarray analysis are available in the Supplementary Information. Briefly, three sets of environmental samples were collected from soil, seawater and human saliva. First, both control (C1–C6) and warming (W1–W6) soils were collected from the unclipped subplots of control and warmed ($+2\,^{\circ}C$) plots, respectively, in Purcell, OK in July 2010 (Zhou *et al.*, 2012). Second, five oil-contaminated samples (BM053, BM054, BM057, BM058 and BM064) and five control samples (OV003, OV004, OV009, OV013 and OV014) were collected from the Gulf of Mexico during two monitoring cruises from 27 May to 2 June in 2010 (Hazen *et al.*, 2010; Lu *et al.*, 2012). Third, human saliva samples were taken from each of 10 caries-active and 10 caries-free individuals after an oral health survey at Sun Yat-sen University in China in 2009 (Yang *et al.*, 2012).

Genomic DNAs isolated from each sample were purified, labeled and then hybridized on the developed array. The hybridized arrays were scanned with a NimbleGen MS 200 Microarray Scanner (Lu *et al.*, 2012). Scanned images were extracted and quantified using NimbleScan software (Roche NimbleGen, Madison, WI, USA), followed by data preprocessing (Wu *et al.*, 2006; He *et al.*, 2007; Van Nostrand *et al.*, 2009; He *et al.*, 2010a). Positive and negative controls were included to check hybridization, gridding and data normalization and comparison, containing (i) 8 degenerate probes targeting 16S rRNA sequences as positive controls, (ii) 563 strain-specific probes targeting 7 hyperthermophile genomes as negative controls, and (iii) common oligonucleotide reference standard for data normalization and comparison (Liang *et al.*, 2010). All hybridization data of functional gene arrays are available at the Institute for Environmental Genomics, University of Oklahoma, OK, USA (http://ieg.ou.edu/4download/).

*Statistical analysis*

Preprocessed microarray data obtained from each environment sample were used for statistical analyses using the vegan package in R 2.9.1 (R Development Core Team, 2006). Virulence gene diversity was calculated using Simpson's reciprocal index ($1/D$), Shannon-Wiener's diversity index ($H'$) and evenness ($E$). Student's *t*-test and Response ratio (Luo *et al.*, 2006) were performed to measure the effects of environmental stimuli on the structure and diversity of microbial communities harboring virulence genes. Detrended correspondence analysis (DCA) was used to determine the overall changes in the occurrence and distribution of virulence genes in each microbial community (Zhou *et al.*, 2008). Three different non-parametric analyses for multivariate data were performed to examine the structure and diversity of microbial communities harboring virulence genes in soil, oil-contaminated seawater and oral samples: (1) analysis of similarities (ANOSIM; Clarke, 1993), (2) non-parametric multivariate analysis of variance using distance matrices (ADONIS; Anderson, 2001), and (3) multi-response permutation procedure (MRPP; Mielke and Berry, 2001; McCune and Grace, 2002). Bray–Curtis similarity index was used to calculate the distance matrix for all the three methods.

## Results

*Overview of the PathoChip*

Pathogenicity is determined by multiple virulence factors, including adherence, colonization, immune evasion, secretion system, invasion, toxin production and iron uptake (Wu *et al.*, 2008). Thus these virulence factors may serve as specific markers for the detection and identification of pathogenic potential in clinical and environmental samples. The new functional gene array, the PathoChip, contained a total of 3715 best probes designed from 16 762 protein-coding sequences belonging to 13 bacterial virulence factors (available at http://ieg.ou.edu/FGPD/summary4.cgi; Table 1). The PathoChip covers 7417 protein-coding sequences

**Table 1** Summary of probes included in the PathoChip

| *Virulence factors* | *No. of HMMER confirmed sequence* | *Total probe* | *Covered CDS* | *Sequence-specific probe* | *No. of covered microbial species (strains)* |
|---|---|---|---|---|---|
| Toxin | 1101 | 70 | 226 | 5 | 31 (75) |
| Hemolysin | 2511 | 620 | 1059 | 26 | 347 (495) |
| Capsule | 1510 | 385 | 613 | 38 | 199 (275) |
| Adhesin | 1214 | 113 | 316 | 7 | 30 (85) |
| Pilin | 2572 | 1037 | 1509 | 49 | 241 (348) |
| Fimbriae | 122 | 16 | 57 | 5 | 3 (5) |
| Colonization factor | 140 | 27 | 137 | 1 | 13 (38) |
| Siderophore | 1977 | 602 | 1708 | 57 | 243 (436) |
| Aerobactin | 490 | 125 | 188 | 7 | 41 (58) |
| Type III secretion | 1899 | 288 | 546 | 84 | 77 (143) |
| Invasin | 1128 | 64 | 131 | 12 | 34 (60) |
| Virulence protein | 454 | 73 | 96 | 11 | 35 (50) |
| Sortase | 1644 | 295 | 831 | 14 | 103 (268) |

Abbreviation: CDS, coding sequence.

from 1397 microbial species (2336 strains). Among the probes in the chip, 316 (8.5%) probes are sequence-specific, while 3399 (91.5%) probes are group-specific. Those probes targeting two or more sequences were classified as group-specific probes; however, 98.2% of the group-specific probes were identified as strain- or species-specific probes.

*Computational validation of probe specificity*

The specificity of the designed probes was examined computationally based on their sequence identity, continuous stretch length and free energy (Rhee *et al*., 2004; He *et al*., 2005, 2010a). For the sequence-specific probes, maximal sequence identity with non-targets was set at 90%, which means sequences that have $\leqslant 4$ mismatches with the probe could be potential targets. However, all sequence-specific probes showed $<90\%$ identity with their closest non-targets (Supplementary Figure S1a), indicating absolute specificity of all the probes. Maximal stretch length with non-targets was set at 20 bases and all sequence-specific probes had continuous stretches at $\leqslant 20$ bases (Supplementary Figure S1b). Minimal free energy with non-targets was set at $-35\,\mathrm{kcal\,mol^{-1}}$ and $>97\%$ of sequence-specific probes showed more than $-35\,\mathrm{kcal\,mol^{-1}}$ free energy (Supplementary Figure S1c). For the group-specific probes, minimal sequence identity, minimal stretch length and maximal free energy with targets were set at 96%, 35 bases, and $-60\,\mathrm{kcal\,mol^{-1}}$, respectively. More than 97% of the group-specific probes shared at least 96% sequence identity with their targets (Supplementary Figure S2a). Minimal stretch lengths of probes with their group targets were calculated and showed that $>97\%$ had sequence stretches with at least 35 bases (Supplementary Figure S2b). More than 97% of group-specific probes had less maximal binding free energy than $-60\,\mathrm{kcal\,mol^{-1}}$ (Supplementary Figure S2c). The measured probe specificity based on a computational evaluation suggested that almost all designed probes were highly specific to their target sequences.

*Applications of the PathoChip to the analysis of environmental samples*

To determine whether the PathoChip is useful in assessing the diversity and responses of microbial communities harboring virulence genes under natural settings, microbial community samples from three distinct environmental habitats (that is, soil, seawater and oral cavities) were analyzed.

*Distribution of virulence genes in soils in response to elevated temperature*

Temperature is one of the major environmental factors affecting the composition and structure of microbial communities (Brock, 1970; Ratkowsky *et al*., 1982; Paerl and Huisman, 2008; Böer *et al*., 2009). To assess how temperature drives changes in the community structure of bacterial pathogens, soil samples collected from an experimental warming site were analyzed. An average of 1587 genes was detected in warming samples while 1372 genes in control samples. The occurrence of each virulence factor was significantly ($P=0.006$) different between warming and control samples. The gene richness index ($S$) showed virulence genes were significantly ($P=0.004$) abundant in warming samples, suggesting that elevated temperature could enrich microbial populations carrying virulence genes (Supplementary Table S1). Based on the Shannon index ($H'$), the diversity of virulence genes was also significantly ($P=0.023$) higher in warming samples than in control samples. Response ratio analysis at 95% confidence interval (CI) showed that elevated temperature led to an increase in the relative abundance of all virulence gene categories except fimbriae and virulence proteins (Figure 2a). In addition, the response ratio analysis of each individual gene at 95% CI showed that most of the individual virulence genes were significantly increased while only 20 genes were significantly decreased (data not shown). A total of 126 virulence genes detected only in warming samples were from a diverse group of bacteria, including 78 genes of known pathogens. Among them, 58 genes were associated with human pathogens (37 species, 48 strains), such as *Acinetobacter baumannii*, *Bacillus cereus*, *Campylobacter coli*, *Clostridium perfringens*, *Clostridium tetani*, *Corynebacterium diphtheria*, *Enterococcus faecalis*, *Haemophilus influenza*, *Helicobacter pylori*, *Mycobacterium ulcerans*, *Neisseria gonorrhoeae*, *Neisseria meningitides*, *Staphylococcus aureus*, *Streptococcus pyogenes* and *Vibrio cholerae*. Some genes were related with soilborne plant pathogens, including *Dickeya dadantii* Ech586, *Pectobacterium atrosepticum* SCRI1043, *Pectobacterium carotovorum* subsp. *carotovorum* WPP14, *Pseudomonas syringae* pv. *phaseolicola* 1448A and *Xanthomonas campestris* pv. *campestris* str. B100. On the contrary, only 10 genes were detected only in control samples, including one soilborne plant pathogen, *Pseudomonas syringae* pv. *tagetis*. To examine further whether elevated temperature made an impact on virulence gene composition and structure in soil microbial communities, DCA was performed and revealed that the overall structure of soil microbial community was altered (Figure 3a). Three complementary non-parametric multivariate statistical tests (MRPP, ANOSIM and ADONIS) also confirmed that the functional structure of the microbial community was significantly ($P<0.01$) different between warming and control samples (Table 2).

*Distribution of virulence genes in deep-sea water in response to oil contamination*

Because many virulence genes have a role in hydrocarbon degradation (Rojo and Martínez,

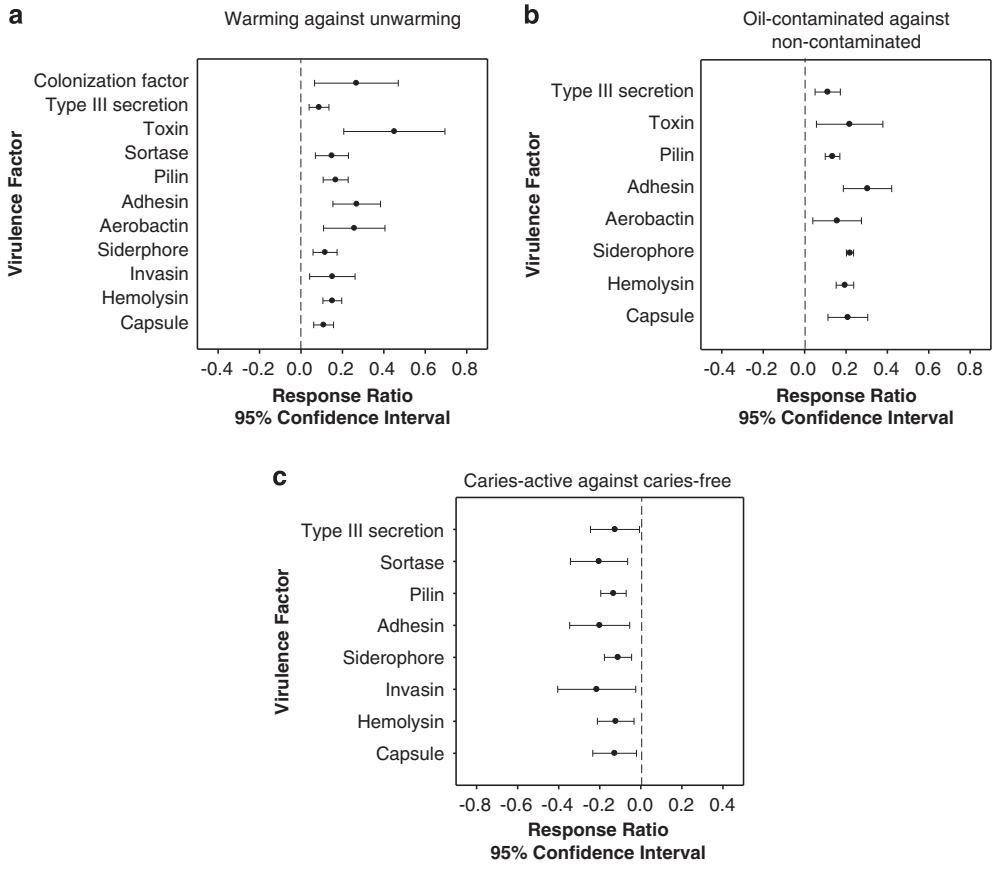


**Figure 2** Response ratio analyses of virulence genes of microbial communities in (**a**) soil, (**b**) seawater and (**c**) oral samples.

2010), we investigated whether there were any considerable changes in the virulence gene diversity in oil-contaminated seawater. An average of 440 genes was detected in oil-contaminated seawater samples while 372 in non-contaminated samples. Richness (*S*) of virulence genes was significantly ($P=0.002$) higher in oil-contaminated samples (Supplementary Table S2), indicating that oil spill stimulated the growth of indigenous microorganisms harboring virulence genes detected. A total of 116 virulence genes were found only in oil-contaminated seawater, including 56 genes of known pathogens. Among them were 38 genes of human pathogens (28 species, 34 strains), such as *Burkholderia cenocepacia*, *Burkholderia multivorans*, *Enterobacter cancerogenus*, *Klebsiella pneumonia*, *Neisseria meningitidis*, *Proteus mirabilis*, *Pseudomonas aeruginosa*, *Ralstonia pickettii*, *Vibrio alginolyticus*, *Vibrio furnissii* and *Vibrio parahaemolyticus*. By contrast, 30 genes detected only in non-contaminated samples showed that 13 genes were from known pathogens, including 12 genes of human pathogens. Response ratio analysis at 95% CI revealed that the relative abundance of genes related to toxin, hemolysin, capsule, adhesin, pilin, type III secretion, aerobactin and siderophore were significantly increased in oil-contaminated samples (Figure 2b). Similarly, microbial community structure was significantly different between oil-contaminated and non-contaminated samples based on the Shannon index ($H'$), Simpson index ($1/D$) and Simpson evenness ($E$) ($P=0.002$, 0.002 and 0.01, respectively) (Supplementary Table S2). In addition, both DCA analysis and three dissimilarity tests showed that the overall virulence gene diversity was significantly altered by the oil contamination (Figure 3b; Table 2).

*Distribution of virulence genes in oral samples in response to caries*

The oral microbial communities respond to different environmental and pathological conditions by modifying their species composition and population size (Avila *et al*., 2009; Polgárová *et al*., 2010). Saliva samples were collected from the caries-active (CA) and caries-free (CF) groups to investigate the impact of caries on the occurrence of virulence genes in human oral microbial community. An average of 923 and 928 probes were detected with the abundance of 16.7–29.5% and 20.6–28.2% in the CA and CF groups, respectively. The evenness ($E$) of virulence genes was significantly ($P=0.017$) different between the CA and CF groups, but there was no significant difference in the richness ($S$) or the diversity between the two groups (Supplementary Table S3). A total of 109 virulence genes were found only in CA, including 47 genes from known pathogens, while 78 genes, including 40 genes from known

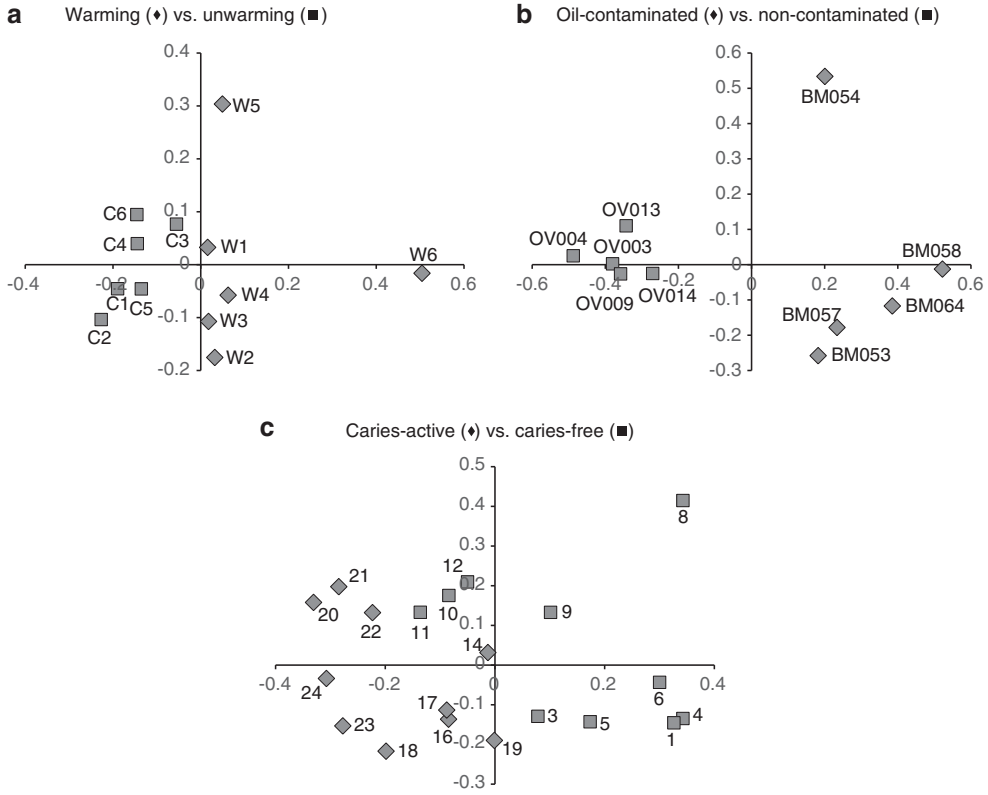


**Figure 3** Detrended correspondence analyses of virulence genes detected in (**a**) soil, (**b**) seawater and (**c**) oral samples.

**Table 2** Dissimilarity tests using three statistical approaches on the overall virulence gene structures of microbial communities in various environments

| *Sample* | *MRPP* | | *ANOSIM* | | *ADONIS* | |
|---|---|---|---|---|---|---|
| | $\delta$ | P-*value* | *R* | P-*value* | *F* | P-*value* |
| Warming vs ambient | 0.149 | 0.000999 | 0.389 | 0.005 | 0.219 | 0.001 |
| Oil-contaminated vs non-contaminated | 0.127 | 0.005 | 0.9 | 0.008 | 0.448 | 0.001 |
| Caries-active vs caries-free | 0.236 | 0.048 | 0.123 | 0.071 | 0.104 | 0.042 |

Abbreviations: ADONIS, analysis of variance using distance matrices; ANOSIM, analysis of similarities; MRPP, multi-response permutation procedure.

pathogens, were found only in CF (data not shown). The response ratio at 95% CI showed that the relative abundance of genes belonging to virulence factors, such as hemolysin, capsule, adhesin, type III secretion, invasin, siderophore, pilin and sortase were significantly decreased in CA (Figure 2c). Both DCA analysis and three dissimilarity tests ($P<0.01$) showed that the structure and composition of microbial communities harboring virulence genes were significantly different between the CA and CF groups (Figure 3c; Table 2).

## Discussion

Environmental pathogens pose a significant threat to human health, thus it is important to obtain an integrated and comprehensive picture of pathogenic potential in various environments. As molecular biological techniques emerged, many molecular methods have been developed for detecting new emergent strains and indicators as well as specific pathogens (Girones *et al.*, 2010). However, most of those methods rely on a limited number of genetic markers that were designed for a particular pathogen or a group. Thus, there is an urgent need for development of a rapid and high-throughput as well as a reliable method for a more comprehensive and intensive survey of environmental pathogen profiles. The recent advance of metagenomic technologies, such as microarrays (Brodie *et al.*, 2006; He *et al.*, 2007, 2010a) and high-throughput sequencing (Iwai *et al.*, 2009; Qin *et al.*, 2010), provides powerful high-throughput tools for analyzing microbial communities (He *et al.*, 2012). Among several types of microarrays, functional gene arrays are

designed primarily to detect specific metabolic groups in microbial communities and provide a platform for analyzing simultaneously multiple functional genes of interest (He *et al.*, 2007; Jaing *et al.*, 2008; He *et al.*, 2010a). Therefore, we developed a comprehensive functional gene array targeting diverse virulence genes to assess pathogenic properties of microbial communities in the environment.

We primarily aimed to develop a functional gene array targeting a broad range of pathogens and thus chose 13 common virulence factors that are present in the vast majority of bacterial pathogens. The PathoChip developed in this study contains 3715 probes, covering 7417 protein-coding sequences from 1397 microbial species (2336 strains). As array specificity is the most critical issue in microarray development, we controlled the specificity of probes in the PathoChip by using experimentally determined probe design and hybridization conditions (He *et al.*, 2005; Liebich *et al.*, 2006). During the probe design, the total number of probes initially designed was 78 870, but the number of the final best probes was extensively decreased after validation using CommOligo and then against GenBank to prevent non-target cross hybridization. For example, aerobactin is a siderophore mainly found in *Escherichia coli*, but all probes designed for this species were discarded because those probes were not specific. During preprocessing, the cutoff value of probe intensity was set at 1000 and the signal-to-noise ratio at 2, which were also used to minimize false positives (He *et al.*, 2010a; Lu *et al.*, 2012).

There are more group-specific probes (91.5%) than sequence-specific probes (8.5%) adopted in the PathoChip. Ideally, more sequence-specific probes provide more specificity to the array by targeting only one sequence, which in turn increases the level of resolution of the array up to the strain level. Previous arrays showed a high percentage of sequence-specific probes ranging from 33.3% to 82.3% (Rhee *et al.*, 2004; He *et al.*, 2007, 2010a). The probes developed in this study showed a small number of sequence-specific probes (316 probes). However, the majority of the group-specific probes showed specificity at least at the species level. A total of 61 probes (1.6%) showed genus-level specificity, of which only six probes (0.2%), designed for pilin (1), siderophores (4) and type III secretion (1) showed specificity at the family level. These results are mainly due to sequence redundancy in the sequence database (that is, GenBank) and/or poor annotation and errors. Therefore, $>98\%$ of the probes adopted in the PathoChip provide specificity at or below the species level, proving an excellent capability of this array for the detection of target sequences with high discrimination power.

The PathoChip in this study was validated experimentally, which also uncovered the pathogenic identity of three environments, illustrated the impacts of environmental perturbations on microbial communities harboring virulence genes and presented a glimpse of how bacterial pathogens thrive in the environment. The application of the PathoChip to a soil environment under climate warming showed that more virulence genes were detected in warming samples, indicating a possible increase in abundance of pathogenic bacteria in response to elevated temperature. Sheik *et al.* (2011) reported that temperature along with drought negatively affected the abundance, diversity and structure of microbial communities and suggested that warming treatment selected for a subset of the total community. However, our study showed that elevated temperature significantly increased the diversity of virulence genes, suggesting that elevated temperature selected for microorganisms carrying the detected virulence genes in the soil microbial community. Another plausible explanation is that the abundance of virulence genes may be increased via the stress response mechanisms of bacteria in response to elevated temperature (Kazmierczak *et al.*, 2003; Allen *et al.*, 2008). All statistical analyses indicated a significant increase of most virulence factors in the warming samples. For example, among the 25 toxin-related genes detected, 11 genes were significantly increased while only two were significantly decreased in warming samples. This significant increase of toxin genes indicates that elevated temperature increases the abundance of microbial populations carrying those toxin genes in soil, which in turn enhances pathogenic potential in soil. This study together with the previous reports support that elevated temperature may increase the level of infectivity and virulence of pathogens as well as pathogen survival in soil (Griffiths, 1991; Dunn *et al.*, 2010; Eastburn *et al.*, 2011; Pritchard, 2011).

Virulence genes are also widespread in many marine bacteria and have important roles in marine bacterial behavior (Persson *et al.*, 2009). Increased abundance of virulence genes is a responsive action of a microbial community to relevant environmental stresses. As many pathogens are capable of degrading hydrocarbons efficiently, the oil spill may stimulate the growth of pathogenic bacteria in oil-contaminated seawater (Rojo and Martínez, 2010). Virulence genes detected in oil-contaminated seawater showed that there was an increase in abundance of virulence genes related to siderophore, aerobactin, pilin and adhesin. Virulence genes of siderophore and pilin that increased or were detected only in oil-contaminated samples were from many hydrocarbon-degrading bacteria, such as *Alcanivorax* sp., *Rhodococcus erythropolis* SK121, *Comamonas testosteroni* KF-1, *Ralstonia eutropha* JMP134, *Delftia acidovorans* SPH-1, *Novosphingobium aromaticivorans* DSM 12444, *Variovorax paradoxus* S110 and *Burkholderia* sp. The increase of virulence genes for both iron uptake and adherence indicates that marine bacteria

facilitate the uptake of iron (a major limiting nutrient) and adherence to hydrocarbons, thereby increasing the potential of *in situ* bioremediation. Other virulence factors that were increased significantly were toxin, hemolysin, type III secretion, capsule and invasin. Virulence genes that increased significantly in oil-contaminated samples were associated with many opportunistic pathogens, including *Staphylococcus pseudintermedius* and *Haemophilus parasuis* SH0165 (Devriese *et al.*, 2005; Xu *et al.*, 2011). These opportunistic pathogens released into seawater may be amplified under certain environmental conditions such as the oil spill and pose potential threat to public health (Yildiz, 2007). Thus the probable success of marine bacteria harboring various virulence genes and/or pathogens in competition with other microorganisms in oil-contaminated ecosystems should be considered during and after the bioremediation processes, especially in terms of microbial risk assessment for offshore coast oil contamination.

The changes in structure and composition of oral microbial communities can either induce or be induced by pathological conditions (Polgárová *et al.*, 2010). Comparison of virulence gene distribution of saliva samples revealed that overall abundance and diversity of virulence genes were not significantly different between the CA and CF groups. However, evenness of virulence genes was significantly different between the two groups, suggesting a selective response of the oral microbial community during the carious infection. In addition, a few subsets were found in each group based on the DCA analysis, indicating that the response of oral microbial communities could be individual host-dependent (heterogeneity of oral microflora). This individual heterogeneity of the microbial community in saliva may be caused by the variability of endogenous and exogenous factors driving the selection and enrichment of oral microflora (Ruby and Barbeau, 2002) or could be due to the artifacts associated with random sampling processes (Zhou *et al.*, 2008, 2011).

A functional gene array with bacterial virulence genes developed in this study provides a powerful tool for detecting a broad range of virulence genes and characterizing virulence gene composition and structure in diverse environments. This study also presented a glimpse of how pathogens thrive in the environment and how pathogenic properties of a microbial community change in response to environmental perturbations, which may provide important information for public health. For more accurate assessment of pathogenic microbial community, it is necessary to discriminate the non-pathogenic strains from the pathogenic strains within the same species and develop a more comprehensive array covering other pathogenic microorganisms, such as viruses, fungi and protozoan parasites. Therefore, further addition of virulence genes should be directed to construct a more specific as well as a comprehensive functional gene array to increase discrimination power as well as to detect and identify a wide variety of pathogens in particular ecosystems.

## Conflict of Interest

The authors declare no conflict of interest.

## Acknowledgements

We thank T. Hazen for providing samples from the Gulf of Mexico. This work was supported, in part, by the US Department of Energy, Biological Systems Research on the Role of Microbial Communities in Carbon Cycling Program (DE-SC0004601).

## References

Allen KJ, Lepp D, McKellar RC, Griffiths MW. (2008). Examination of stress and virulence gene expression in *Escherichia coli* O157:H7 using targeted microarray analysis. *Foodborne Pathog Dis* **5**: 437–447.

Anderson M. (2001). A new method for non-parametric multivariate analysis of variance. *Austral Ecol* **26**: 32–46.

Anjum MF, Mafura M, Slickers P, Ballmer K, Kuhnert P, Woodward MJ *et al.* (2007). Pathotyping *Escherichia coli* by using miniaturized DNA microarrays. *Appl Environ Microbiol* **73**: 5692–5697.

Avila M, Ojcius DM, Yilmaz Ö. (2009). The oral microbiota: living with a permanent guest. *DNA Cell Biol* **28**: 405–411.

Blomfield IC, Calie PJ, Eberhardt KJ, McClain MS, Eisenstein BI. (1993). Lrp stimulates phase variation of type 1 fimbriation in *Escherichia coli* K-12. *J Bacteriol* **175**: 27–36.

Böer SI, Hedtkamp SIC, van Beusekom JJE, Fuhrman JA, Boetius A, Ramette A. (2009). Time- and sediment depth-related variations in bacterial diversity and community structure in subtidal sands. *ISME J* **3**: 780–791.

Bossier P, Hofte M, Vestraete W. (1988). Ecological significance of siderophores in soil. *Adv Microb Ecol* **10**: 385–414.

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.

Bruant G, Maynard C, Bekal S, Gaucher I, Masson L, Brousseau R *et al.* (2006). Development and validation of an oligonucleotide microarray for detection of multiple virulence and antimicrobial resistance genes in *Escherichia coli*. *Appl Environ Microbiol* **72**: 3780–3784.

Brock TD. (1970). High temperature systems. *Annu Rev Ecol Syst* **1**: 191–210.

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

Clarke KR. (1993). Non-parametric multivariate analyses of changes in community structure. *Austral Ecol* **18**: 117–143.

Cossart P, Jonquières R. (2000). Sortase, a universal target for therapeutic agents against Gram-positive bacteria? *Proc Natl Acad Sci USA* **97**: 5013–5015.

Craig L, Taylor RK, Pique ME, Adair BD, Arvai AS, Singh M *et al.* (2003). Type IV pilin structure and assembly: X-ray and EM analyses of *Vibrio cholerae* toxincoregulated pilus and *Pseudomonas aeruginosa* PAK pilin. *Mol Cell* **11**: 1139–1150.

Devriese LA, Vancanneyt M, Baele M, Vaneechoutte M, Graef ED, Snauwaert C *et al.* (2005). *Staphylococcus pseudintermedius* sp. nov., a coagulase-positive species from animals. *Int J Syst Evol Microbiol* **55**: 1569–1573.

Dunn RR, Davies TJ, Harris NC, Gavin MC. (2010). Global drivers of human pathogen richness and prevalence. *Proc R Soc B* **277**: 2587–2595.

Eastburn DM, McElrone AJ, Bilgin DD. (2011). Influence of atmospheric and climatic change on plant–pathogen interactions. *Plant Pathol* **60**: 54–69.

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

Finlay BB, Falkow S. (1997). Common themes in microbial pathogenicity revisited. *Microbiol Mol Biol Rev* **61**: 136–169.

Friman V-P, Hiltunen T, Jalasvuori M, Lindstedt C, Laanto E, Örmälä A-M *et al.* (2011). High temperature and bacteriophages can indirectly select for bacterial pathogenicity in environmental reservoirs. *PLoS ONE* **6**: e17651.

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

Geue L, Schares S, Mintel B, Conraths FJ, Muller E, Ehricht R. (2010). Rapid microarray-based genotyping of enterohemorrhagic *Escherichia coli* (EHEC) serotypes O156:H25/H-/Hnt isolated from cattle and analysis of the clonal relationship. *Appl Environ Microbiol* **76**: 5510–5519.

Girones R, Ferrús MA, Alonso JL, Rodriguez-Manzano J, Calgua B, Corrêa Ade A *et al.* (2010). Molecular detection of pathogens in water–the pros and cons of molecular techniques. *Water Res* **44**: 4325–4339.

Griffiths E. (1991). Environmental regulation of bacterial virulence implications for vaccine design and production. *Trends Biotechnol* **9**: 309–315.

Hazen TC, Dubinsky EA, DeSantis TZ, Andersen GL, Piceno YM, Singh N *et al.* (2010). Deep-sea oil plume enriches indigenous oil-degrading bacteria. *Science* **330**: 204–208.

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

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

He Z, Piceno Y, Deng Y, Xu M, Lu Z, DeSantis T *et al.* (2012). The phylogenetic composition and structure of soil microbial communities shifts in response to elevated carbon dioxide. *ISME J* **6**: 259–272.

He Z, Wu L, Li XY, Fields MW, Zhou JZ. (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.* (2010b). Metagenomic analysis reveals a marked divergence in the structure of belowground microbial communities at elevated $CO_2$. *Ecol Lett* **13**: 564–575.

Hyman RW St, Onge RP, Allen EA, Miranda M, Aparicio AM, Fukushima M *et al.* (2010). Multiplex identification of microbes. *Appl Environ Microbiol* **76**: 3904–3910.

Iwai S, Chai B, Sul WJ, Cole JR, Hashsham SA, Tiedje JM. (2009). Genetargeted-metagenomics reveals extensive diversity of aromatic dioxygenase genes in the environment. *ISME J* **4**: 279–285.

Jaing C, Gardner S, McLoughlin K, Mulakken N, Alegria-Hartman M, Banda P *et al.* (2008). A functional gene array for detection of bacterial virulence elements. *PLoS ONE* **3**: e2163.

Jones KE, Patel NG, Levy MA, Storeygard A, Balk D, Gittleman JL *et al.* (2008). Global trends in emerging infectious diseases. *Nature* **451**: 990–993.

Kazmierczak MJ, Mithoe SC, Boor KJ, Wiedmann M. (2003). Listeria monocytogenes $\sigma^B$ regulates stress response and virulence functions. *J Bacteriol* **185**: 5722–5734.

Kline KA, Falker S, Dahlberg S, Normark S, Henriques-Normark B. (2009). Bacterial adhesins in host–microbe interactions. *Cell Host Microbe* **5**: 580–592.

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.

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.

Lu Z, Deng Y, Van Nostrand JD, He Z, Voordeckers J, Zhou A *et al.* (2012). Microbial gene functions enriched in the Deepwater Horizon deep-sea oil plume. *ISME J* **6**: 451–460.

Luo Y, Hui D, Zhang D. (2006). Elevated $CO_2$ stimulates net accumulations of carbon and nitrogen in land ecosystems: a meta-analysis. *Ecology* **87**: 53–63.

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

McCune B, Grace JB. (2002). *Analysis of Ecological Communities*. MjM Software: Gleneden Beach, OR, USA.

Mielke PW, Berry KJ. (2001). *Permutation Methods: a Distance Function Approach*. Springer: New York, NY, USA.

Miller SM, Tourlousse DM, Stedtfeld RD, Baushke SW, Herzog AB, Wick LM *et al.* (2008). In situ-synthesized virulence and marker gene biochip for detection of bacterial pathogens in water. *Appl Environ Microbiol* **74**: 2200–2209.

Neilands JB. (1995). Siderophores: structure and function of microbial iron transport compounds. *J Biol Chem* **270**: 26723–26726.

Paerl HW, Huisman J. (2008). Blooms like it hot. *Science* **320**: 57–58.

Persson OP, Pinhassi J, Riemann L, Marklund B-I, Rhen M, Normark S *et al.* (2009). High abundance of virulence gene homologues in marine bacteria. *Environ Microbiol* **11**: 1348–1357.

Peterson G, Bai J, Nagaraja TG, Narayanan S. (2010). Diagnostic microarray for human and animal bacterial diseases and their virulence and antimicrobial resistance genes. *J Microbiol Met* **80**: 223–230.

Polgárová K, Behuliak M, Celec P. (2010). Effect of saliva processing on bacterial DNA extraction. *New Microbiol* **33**: 373–379.

Pritchard S. (2011). Soil organisms and global climate change. *Plant Pathol* **60**: 82–99.

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

Quinones B, Swimley MS, Taylor AW, Dawson ED. (2011). Identification of *Escherichia coli* O157 by using a novel colorimetric detection method with DNA microarrays. *Foodborne Pathog Dis* **8**: 705–711.

R Development Core Team (2006). *R: a Language and Environment for Statistical Computing*. R Foundation for Statistical Computing: Vienna, Austria, ISBN 3-900051-07-0. http://www.R-project.org.

Ratkowsky DA, Olley J, McMeekin TA, Ball A. (1982). Relationship between temperature and growth rate of bacterial cultures. *J Bacteriol* **149**: 1–5.

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

Rojo F, Martínez JL. (2010). Oil Degraders as Pathogens. In: Timmis KN (ed) *Handbook of Hydrocarbon and Lipid Microbiology*. Springer: New York, NY, USA, pp 3293–3303.

Ruby J, Barbeau J. (2002). The buccale puzzle: the symbiotic nature of endogenous infections of the oral cavity. *Can J Infect Dis* **13**: 34–41.

Sheik CS, Beasley WH, Elshahed MS, Zhou X, Luo Y, Krumholz LR. (2011). Effect of warming and drought on grassland microbial communities. *ISME J* **5**: 1692–1700.

Singh A, Wyant T, Anaya-Bergman C, Aduse-Opoku J, Brunner J, Laine ML *et al.* (2011). The capsule of *Porphyromonas gingivalis* leads to a reduction in the host inflammatory response, evasion of phagocytosis, and increase in virulence. *Infect Immun* **79**: 4533–4542.

Slenning BD. (2010). Global climate change and implications for disease emergence. *Vet Pathol* **47**: 23–33.

Smith KF, Sax DF, Gaines SD, Guernier V, Guegan J-F. (2007). Globalization of human infectious disease. *Ecology* **88**: 1903–1910.

Taylor LH, Latham SM, Woolhouse ME. (2001). Risk factors for human disease emergence. *Philos Trans R Soc Lond B Biol Sci* **356**: 983–989.

Tembe W, Zavaljevski N, Bode E, Chase C, Geyer J, Wasieloski L *et al.* (2007). Oligonucleotide fingerprint identification for microarray-based pathogen diagnostic assays. *Bioinformatics* **23**: 5–13.

Tobias J, Svennerholm A-M. (2012). Strategies to overexpress enterotoxigenic *Escherichia coli* (ETEC) colonization factors for the construction of oral whole-cell inactivated ETEC vaccine candidates. *Appl Microbiol Biotechnol* **93**: 2291–2300.

Trivedi P, He Z, Van Nostrand JD, Albrigo G, Zhou J, Wang N. (2012). Huanglongbing alters the structure and functional diversity of microbial communities associated with citrus rhizosphere. *ISME J* **6**: 363–383.

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 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.

Woolhouse M, Gaunt E. (2007). Ecological origins of novel human pathogens. *Crit Rev Microbiol* **33**: 231–242.

Woolhouse MEJ, Gowtage-Sequeria S. (2005). Host range and emerging and reemerging pathogens. *Emerg Infect Dis* **11**: 1842–1847.

Wu CF, Valdes JJ, Bentley WE, Sekowski JW. (2003). DNA microarray for discrimination between pathogenic 0157:H7 EDL933 and non-pathogenic *Escherichia coli* strains. *Biosens Bioelectron* **19**: 1–8.

Wu H-J, AH-J Wang, Jennings MP. (2008). Discovery of virulence factors of pathogenic bacteria. *Curr Opin Chem Biol* **12**: 93–101.

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.

Xu Z, Yue M, Zhou R, Fan Y, Bei W, Chen H. (2011). Genomic characterization of *Haemophilus parasuis* SH0165, a highly virulent strain of serovar 5 prevalent in China. *PLoS One* **6**: e19631.

Yang S, Bourne PE. (2009). The evolutionary history of proteins domains viewed by species phylogeny. *PLoS One* **4**: e8378.

Yang F, Zeng X, Ning K, Liu K-L, Lo C-C, Wang W *et al.* (2012). Saliva microbiomes distinguish caries-active from healthy human populations. *ISME J* **6**: 1–10.

Yildiz FH. (2007). Processes controlling the transmission of bacterial pathogens in the environment. *Res Microbiol* **158**: 195–201.

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

Zhou J, Wu L, Deng Y, Zhi X, Jiang Y-H, Tu Q *et al.* (2011). Reproducibility and quantitation of amplicon sequencing-based detection. *ISME J* **5**: 1303–1313.

Zhou J, Xue K, Xie J, Deng Y, Wu L, Cheng X *et al.* (2012). Microbial mediation of carbon-cycle feedbacks to climate warming. *Nat Climate Change* **2**: 106–110.

Supplementary Information accompanies this paper on The ISME Journal website (http://www.nature.com/ismej)