Supporting Information

Zhou *et al.* 10.1073/pnas.0709016105

SI Results and Discussion

Artifacts and the Advantages of Microarray-Based Approaches. Substantial variations of the *z* values for microbial communities are observed, ranging from 0.02 to 0.47 (1). Although some of the observed *z* values are lower than those in plants, some are much higher. Such variations could be due to the true differences of microbial spatial distributions, but could also be compounded by various sampling artifacts such as under-sampling, unequal sampling, random sampling, and taxonomic lumping (1). In general, the array-based approach has several advantages over the conventional molecular approaches to minimize such artifacts.

Under-sampling. One of the sampling artifacts is undersampling. Because microbial diversity is immense, all studies have greatly undersampled microbial diversity (1). For example, only very small portions (<1%) of a microbial community are randomly sampled for determining TARs, using conventional PCR-based molecular methods (e.g., DGGE, T-RFLP, cloning and sequencing) (2–4). Due to the low detection sensitivity, many rare taxa could not be sampled for analysis (5). Undersampling could result in flatter taxa-area relationships (1).

The GeoChip-based approach has advantages to ameliorate the undersampling problem, although it could not eliminate it. First, the arrays contain tens of thousands genes, and hence a single hybridization can simultaneously survey all of these genes. If the arrays represent well the diversity of the community examined, the majority of microbial populations of interest, if not all, can be surveyed for analysis. Second, the GeoChip contain probes for many different types of functional genes, and hence more broad functional groups will be sampled from a community. Sampling multiple genes of diverse functions could reduce the potential biases from a single gene such as 16S rRNA gene. The information on microbial diversity based on many functional genes/groups will be more representative of the overall picture of microbial diversity at the whole-community level than those based on single gene (e.g., 16S rRNA gene). Third, compared with the conventional approaches, higher detection sensitivity (as low as 2 bacterial cells) can be obtained with the whole-community genome amplification (WCGA)-assisted microarray detection approach (6). Therefore, some less abundant taxa could be possibly detected for analysis with the array-based hybridization. In addition, undersampling problem will be much less severe with functional genes than rRNA genes. Not all microbial populations in a community have certain functional genes (e.g., *nir*S, *nir*K) of interest, and consequently the microbial population diversity based on functional genes will be much lower than that based on 16S rRNA genes. Thus, less sampling efforts are required to survey the functional gene diversity of the microbial community of interest. For instance, generally speaking, up to 5% (0.1–5%) of cultured heterotrophs in a typical soil community are denitrifiers (7). Assuming that 1 g of soil contains 2,000–18,000 genomes (8–10), a total of 100–900 microbial populations would have been expected to be denitrifiers. The GeoChips used in this study (GeoChip2.0) have probes to target 136 *nir*S- or 127 *nir*K-containing populations. Theoretically GeoChip 2.0 could detect at least >30% denitrifiers in a soil community if the numbers used above are valid and applicable to various soils. We are developing next generation of GeoChips (GeoChip 3.0), which is able to detect 2–3 times more populations involved in denitrification. For another example, nitrification is more restrictive to narrow phylogenetic groups (e.g., *β-proteobacteria*) and some Archaea (e.g., *Euryarchaeota*) and *amo*A genes are less divergent among different microbial populations. GeoChip 2.0 contain probes to target 32 *amo*A gene sequences, whereas GeoChip 3.0 have probes to target ≈500 *amo*A sequences from both Bacteria and Archaea. Therefore, it is expected that GeoChips would have been able to detect substantial proportions, if not all, of nitrifiers in a soil community. Finally, distance-decay approach (2, 11) uses relative comparisons of microbial community composition rather than richness to examine TARs. It also has advantages in ameliorating the undersampling problem because it is not at all necessary to fully characterize or sample a community for making relative comparisons, which could still give robust measures of spatial scaling parameters (2). In a word, the undersampling problem can be greatly ameliorated with the GeoChip-based detection methods and appropriate experimental designs and data analysis approaches.

Unequal sampling. The second artifact is unequal sampling (1). To minimize the artifact associated with under-sampling, more sampling efforts are required to describe diversity in larger areas than smaller area (1), which leads to another sampling artifact, unequal sampling, that is, the sampling efforts are not equal between larger and smaller areas. Because some species could be missed during sampling processes, sampling progressively larger area and devoting more time to data collection will increase the number of counting the missed species. As a result, such unequal sampling could lead to increases in species richness estimation in larger area, even if the richness in the real community remains unchanged (12–14).

Different from other conventional sampling approaches, such as PCR amplification-based cloning, terminal fragment length polymorphisms, and mass spectrometry-based protein analysis, which are an open-format detection approach, microarray-based hybridization, is a close-format detection approach, because it provides information only for the genes fabricated on the microarrays. No information will be obtained for the genes that have no probes on the arrays. Because the same arrays were used to analyze the samples from areas of different size, theoretically, no additional genes will be detected with microarray hybridization even if more sampling efforts are used for surveying larger area. Consequently, the unequal sampling problem can also be minimized or eliminated by the array-based close-format detection approach.

Random sampling. The third sampling artifact is associated with random sampling strategy. DNA reassociation kinetic studies estimate that one gram of soil could contain 2,000–18,000 genomes (8–10), suggesting that prokaryotic diversity in soil is extremely high. Generally, only small portions (< 1% of species per population) from a microbial community are randomly sampled for determining TARs, using conventional PCR-based molecular methods. If a community of 2,000–18,000 genomes is randomly divided into 100 portions, the species/genome composition among different portions will be quite different. If the sampling process is completely random, theoretically, the probability of sampling the same portion of a community in various sampling events for determining TARs could be very small (as low as <0.01%). Although certain populations (e.g., dominant populations) will have higher probabilities of being resampled by these methods if they are abundant and/or evenly distributed among these different portions, it is still not possible to ensure that the same portions of a microbial community are measured across the different sampling events. Because different portions of a community are randomly sampled for analysis in different

sampling events, the estimated species richness would increase as more sampling efforts are used, even for the same community without any changes in diversity. Thus, theoretically, the randomly sampling approach could also lead to overestimation of TARs.

With the array-based approach, all of the samples are compared against the same sets of probes on the array, ensuring that the same portion of species/populations (i.e., those species/populations with probes on the arrays) from a community are sampled for comparison across all samples in a study. As a result, the sampling artifact due to the nature of random sampling process can be minimized if not eliminated.

It should be noted that, although the artifacts associated with the nature of random sampling process could occur with the conventional PCR-based fingerprinting and cloning approaches, they still have values in evaluating TARs. Some previous results showed that conventional fingerprinting and cloning-based approaches are able to consistently ($> 90\%$) sample dominant populations (11, 15), and hence the sampling processes are not totally random. Thus, the z values determined by using such approaches should still be valuable (2, 11). However, great cautions should be taken. The relative impacts due to the randomly sampling nature of these conventional approaches on estimating TARs are community-dependent. If a community is very complex and without dominant populations as previously demonstrated in various soils (16), substantial overestimation of z values would be expected with the conventional molecular methods.

Taxonomic lumping. The fourth sampling artifact is associated with taxonomic lumping. The spatial biodiversity patterns depend on the defined taxonomic resolution. The z value increases with increased taxonomic resolution (1, 17). Due to the low rates of molecular evolution, it is difficult to obtain fine-scale resolution at the desired species/strain level with rRNA genes (18, 19). The taxonomic resolutions based on morphortypes, ribosomal RNA genes-based phylotypes and fingerprints probably include a much greater diversity of ecological types than contained within a plant or animal species. Such coarse level taxonomic resolutions could also lead to the flat TARs relative to those of plants and animals (1, 2, 11). Because microbial spatial patterns are very sensitive to taxonomic resolutions used, the observations of microbial cosmopolitanism might be due to taxonomic "lumping" of microorganisms (11).

Microarray hybridization was carried out overnight at 50°C in 50% of formamide. Our previous results showed that, under such conditions, by simultaneously considering probe sequence similarity, sequence stretches and free energy, specific probes of 50mer can be designed with the sequence similarity of $<90\%$ to the target sequences (20, 21). Because many functional genes have $<90\%$ DNA sequence identity at the species level, e.g., *amo*A ($0.75 \pm 0.11\%$), *nir*S ($0.70 \pm 0.18\%$), and *dsr*AB ($0.73 \pm 0.13\%$) (18), the level of $<90\%$ sequence similarity will provide species/strain level of resolution (22). The sequence similarity criterion can be further relaxed if the other two criteria are applied. For instance, specific hybridization was also observed for some probes with $<98\%$ similarity to the target sequences. These results suggest that finer strain level of resolution could be possibly achieved with some 50mer probes under the hybridization conditions examined (21, 22).

In summary, the array-based approach has great advantages in minimizing or eliminating various artifacts associated with evaluation of microbial spatial diversity patterns, and thus it can provide more realistic estimation of z values in microbial communities.

SI Materials and Methods

Experimental Site, Sampling Strategy and Procedure. Generally speaking, the taxa-area relationships (TARs) are determined by both contiguous and noncontiguous sampling approaches (1). The contiguous TARs estimate the increase in taxa richness for nested areas within a single region, whereas the noncontiguous TARs estimate the increase of taxa richness from local to global scales. Several well known contiguous TARs estimations are measured at meters scales (see refs. 1 and 17 for review; see also refs. 2 [0.03 to 300 m], 11 [1–750 m within different land systems], and 35 [1–8 m]). To ensure our results are comparative with previous studies, in this study, we used contiguous sampling approach and examined TARs at spatial scale of meter level (ranging from 0.14 to 1000 m) (Fig. S3).

Two, perpendicular, 1-km transects crossing at the center of each other were located in a deciduous forest near the Clinch River on the Oak Ridge Reservation (Oak Ridge, TN). A forest cover map based on satellite imageries by the USGS and the USDA Forest Service classifies the experimental site as predominantly "oak-hickory" forest (The National Atlas, www.nationalatlas.gov/mld/foresti.html). The mean annual temperature is 14.3°C and annual precipitation is 113 cm based on records from 1971 to 2000 (National Weather Service). The forest soil type was primarily loam and the position of the center point of the two crossing transects was 35°92′10″N and 84°26′40″W. The entire sampling area was $\approx$0.5 km^2 and the elevation difference was $\approx$80 m (244 to 323 m). Sampling locations were placed at the center point and one near each corner of each grid (Fig. S3).

Microbial functional processes and distribution could be horizontally heterogeneous. To make sure that the sampling is representative of the diversity at the meter scale of interest, pooling samples around a sampling point is recommended and commonly used approach (23). Thus, to minimize the possible effects of spatial heterogeneity at the spatial level below meter scale on experimental measurements, after removing loose organic material (O_i horizon), four closely adjacent soil cores (15 cm deep) were randomly collected around each sampling point, using a bucket auger. The collected soils were then mixed, sieved through a 4-mm sieve, and combined. Also, by pooling 15-cm core, we were able to control vertical heterogeneity while able to measure (horizontal) spatial heterogeneity at meter scale**.** Thus, the experimental data obtained in this study represent integrated microbial community structure and environmental heterogeneity within the top 15 cm of the soil. A subset of the composited soil was preserved immediately in the field by freezing in liquid nitrogen, transported to the laboratory on dry ice, and stored at −80°C until DNA extraction. In addition, at each sampling point, the predominant overstory tree species within 10 m of the center point were recorded.

DNA Extraction and Microarray Hybridization. Community DNA was directly extracted and purified from 5 g of each preserved soil sample, using a protocol modified from Zhou *et al.* (24). Approximately 100 ng of purified community DNA was amplified in triplicate, using a Templiphi 500 amplification kit (Amersham Biosciences), using the whole-community genome amplification approach as described by Wu *et al.* (25). Target template labeling and microarray hybridization were carried out as described in ref. 25. The GeoChip used in this study contained 24,243 oligonucleotide probes targeting >150 functional groups of $>10{,}000$ genes essential to the biogeochemical cycles of carbon, nitrogen, phosphorus, and sulfur along with metal resistance, metal reduction and organic contaminant degradation (20). To control experimental error from random amplification and fluorescent labeling, the triplicate samples were pooled before hybridization. Microarray hybridization was carried out overnight at 50°C plus 50% of formamide (25), in which specific hybridization can be achieved for probes with sequence identity of $<90\%$ to the target sequences (20, 21), even up to 98% similar to the target sequences for some probes (10). Many functional genes had DNA sequence identity $<90\%$ at the

species level, e.g., *amo*A (0.75 ± 0.11%), *nir*S (0.70 ± 0.18%), and *dsr*AB (0.73 ± 0.13%) (11). Thus, the GeoChip hybridization could provide species-strain level of resolution (23). Microarray scanning and image processing were accomplished as described in ref. 25.

Data Processing and Analysis. ***Microarray data preprocessing.*** Microarray data preprocessing was carried out as described (20, 25) to remove poor spots and outliers. Hybridization spots with a signal-noise ratio (SNR) < 2.5 were removed from further analysis. When three technical replicates were combined, only those spots that showed >2 positive hybridizations were used. The normalization among slides was done based on summed signal intensity of positive spots excluding control spots. Outliers were removed among three technical replicates with $\alpha = 0.01$ by Grubbs's test (26). The normalized hybridization data for individual functional gene sequences was then reorganized based on functional genes, such as *nif*H and *nir*S, and functional groups, such as nitrification, denitrification for estimating z values.

Gene–area relationships. The hybridization intensity data were transformed into binary data as presence or absence, which were then used to estimate richness in each sample. The power-law exponent z was estimated directly with linear regression approach, using the equation $\log S = \log c + z \log A$, where S is the observed gene richness, and A is the area in the nested design (2, 50, 200, 2,500, 125,000 and 500,000 m^2). This equation was fitted separately based on the richness of all individual functional gene sequences and the richness of functional genes to obtain the z values at the level of individual gene sequences and the level of functional genes (Table 1). This equation was also fitted based on individual gene sequences within each functional group (e.g., carbon degradation, nitrification, nitrogen fixation, N reduction, metal resistance, and organic contaminant degradation) or phylogenetic group (e.g., fungi, archaea, bacteria, Gram-positive bacteria, Gram-negative bacteria, α-, β-, γ-, and δ-*proteobacteria*) to examine how the z values change among different functional or phylogenetic groups (Table 1).

Due to the nested sampling design, the data points in individual areas were not independent. Thus, the significance of regression coefficient (z value) was tested by using bootstrapping with replacement. First, area and richness were randomly paired from the original dataset and the regression coefficients (z values) were calculated 10,000 times in each of the situations described above. Then, one-sample t test was used to determine whether the observed slope was significantly different from the mean of the randomly generated slopes as described in ref. 3. As widely used in phylogenetics and other areas, bootstrapping is a technique for obtaining standard errors and confidence limits of various statistics by randomly resampling the original populations.

One interesting question is that whether all observed z values were significantly different among various functional or phylogenetic groups. However, no standard deviations were available for the estimated z values for individual groups, standard t test or ANOVA could not be implemented. Thus, bootstrapping (1,000 times) was used to estimate the variances of z values, followed by pairwise t test with Bonferroni correction to determine whether the observed z values for individual groups were significantly different from each other.

Variation partitioning analysis. Various similarity matrices were calculated for microbial community, plant composition, environmental chemistry and geographic distance. First, the binary hybridization data were used to calculate community similarity matrices with Sørensen index. The hybridization intensity data were also used to calculate community similarity matrices with Bray-Curtis index, which is a quantitative version of Sørensen index (27). When two types of similarity matrices were used, the results were very similar to each other ($r_M > 0.95$), thus the Sørensen index was used for analysis with BIO-ENV and Mantel test. Also, Euclidean distances were used to construct similarity matrices for geographic distance and soil properties. In addition, presence-absence data of overstory tree species were converted into distance matrix by Sørensen index as well.

Partial Mantel tests (10,000 permutations) and partial CCA were performed as described in refs. 28–30 to determine the effects of environmental heterogeneity or tree diversity on microbial community composition by treating geographic distance constant and vice versa based on individual functional genes and each functional or phylogenetic group. Then function "mantel" in the packages of vegan (v. 1.7–67) and ecodist (v. 1.01) of *R* (v.2.2.0, www.r-project.org) were used to perform the tests. The partial CCA used for variation partition was carried out as described in ref. 30, using the function "cca" in vegan (version 1.7–67) and the package CANOCO for verification purpose (31).

BIO-ENV and CCA were also used to identify the abiotic factors most important to microbial community composition, and they were used to construct soil property matrix for the Mantel test and CCA. BIO-ENV identified the best subset of geochemical variable set [ammonium concentration, C/N ratio and fraction of C in particulate organic matter (fPOM-C)] based on Spearman rank correlation. The selected variables were confirmed by significance of the CCA model based on 1,000 times of permutation, using eigenvalues of constrained and unconstrained correspondence axes. Euclidean distances were used to construct similarity matrices based on these identified abiotic factors, which were then used to test the effects of these factors on microbial community composition in conjunction with geographic distance.

1. Green J, Bohannan BJM (2006) Spatial scaling of microbial biodiversity. *Trends Ecol Evol* 21:501–507.
2. Horner-Devine MC, Lage M, Hughes JB, Bohannan BJM (2004) A taxa-area relationship for bacteria. *Nature* 432:750–753.
3. Noguez AM, Arita HT, Escalante AE, Forney LJ, Garcia-Oliva F, Souza V (2005) Microbial macroecology: Highly structured prokaryotic soil assemblages in a tropical deciduous forest. *Glob Ecol Biogeogr* 14:241–248.
4. van der Gast CJ, Lilley AK, Ager D, Thompson IP (2005) Island size and bacterial diversity in an archipelago of engineering machines. *Environ Microbiol* 7:1220–1226.
5. Woodcock S, Curtis TP, Head IM, Lunn M, Sloan WT (2006) Taxa-area relationships for microbes: The unsampled and the unseen. *Ecol Lett* 9:805–812.
6. Wu L, Liu X, Schadt CW, Zhou J (2006) Microarray-based analysis of sub-nanogram quantities of microbial community DNAs using Whole Community Genome Amplification (WCGA). *Appl Environ Microbiol* 72:4931–4941.
7. Tiedje JM (1994) in *Methods of Soil Analysis Part 2 Microbiological and Biochemical Properties*, eds Weaver RW, Angle JS, Bottomley PS (SSSA, Madison, WI), pp 245–267.
8. Torsvik V, Daae FL, Sandaa RA, Øverås L (1998) Novel techniques for analysing microbial diversity in natural and perturbed environments. *J Biotechnol* 64:53–62.
9. Torsvik V, Øverås L (2002) Microbial diversity and function in soil: From genes to ecosystems. *Curr Opin Microbiol* 5:240–245.
10. Torsvik V, Sorhein R, Goksoyr J (1996) Total bacterial diversity in soil and sediment communities-a review. *J Ind Microbiol* 17:1970–1178.
11. Green JL, *et al.* (2004) Spatial scaling of microbial eukaryote diversity. *Nature* 432:747–750.
12. Cam E, *et al.* (2002) Disentangling sampling and ecological explanations underlying species-area relationships. *Ecology* 83:1118–1130.
13. Rosenzweig ML (1995) *Species diversity in space and time* (Cambridge Univ Press, Cambridge, UK).
14. Turner WR, Tjørve E (2005) Scale-dependence in species-area relationships. *Ecography* 28:721–730.
15. Fernández A, *et al.* (1999) How stable is stable? Function versus community composition. *Appl Environ Microbiol* 65:3697–3704.
16. Zhou J, Xia B, Treves DS, Wu, L.-Y., Marsh TL, O'Neill RV, Palumbo AV, Tiedje JM (2002) Spatial and resource factors influencing high microbial diversity in soil. *Appl Environ Microbiol* 68:326–334.
17. Martiny JBH, *et al.* (2006) Microbial biogeography: Putting microorganisms on the map. *Nat Rev Microbiol* 4:102–112.
18. Stackebrandt E, Goebel BM (1994) Taxonomic note: A place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. *Int J Syst Bacteriol* 44:846–849.

19. Yamamoto S, Harayama S (1995) PCR amplification and direct sequencing of *gyr*B genes with universal primers and their application to the detection and taxonomic analysis of *Pseudomonas putida* strains. *Appl Environ Microbiol* 61:1104–1109.
20. He Z, *et al.* (2007) GeoChip: A comprehensive microarray for investigating biogeochemical, ecological, and environmental processes. *ISME J* 1:67–77.
21. Liebich J, *et al.* (2006) Improvement of oligonucleotide probe design criteria for functional gene microarrays in environmental applications. *Appl Environ Microbiol* 72:1688–1691.
22. Tiquia SM, *et al.* (2004) Evaluation of 50-mer oligonucleotide arrays for detecting microbial populations in environmental samples. *BioTechniques* 36:664–675.
23. van Elsas JD, Smalla K, Lilley AK, Bailey MJ (2002) in *Manual of Environmental Microbiology*, eds Hurst CJ, Crawford RL, Knudsen GR, McInerney MJ, Stetzenbach LD (ASM, Washington, DC), pp 505–515.
24. Zhou J, Bruns MA, Tiedje JM (1996) DNA recovery from soils of diverse composition. *Appl Environ Microbiol* 62:316–322.
25. Wu L, Liu X, Schadt CW, Zhou J (2006) Microarray-based analysis of subnanogram quantities of microbial community DNAs using Whole Community Genome Amplification (WCGA). *Appl Environ Microbiol* 72:4931–4941.
26. Grubbs FE (1969) Procedures for detecting outlying observations in samples. *Technometrics* 11:1–21.
27. Colwell RK (2005) EstimateS 7.5: Statistical estimation of species richness and shared species from samples (Univ of Connecticut, Storrs, CT).
28. Smouse PE, Long JC, Sokal RR (1986) Multiple regression and correlation extensions of the Mantel test of matrix correspondence. *Syst Zool* 35:627–632.
29. Borcard D, Legendre P (1994) Environmental control and spatial structure in ecological communities: An example using oribatid mites (Acari, Oribatei). *Environ Ecol Stat* 1:37–61.
30. Borcard D, Legendre P, Drapeau P (1992) Partialling out the spatial component of ecological variation. *Ecology* 73:1045–1055.
31. ter Braak CJF, Smilauer P (2002) *CNOCO Reference Manual and CanoDraw for Windows User's Guide: Software for Canonical Community Ordination (version 4.5)* (Microcomputer Power, Ithaca, NY).

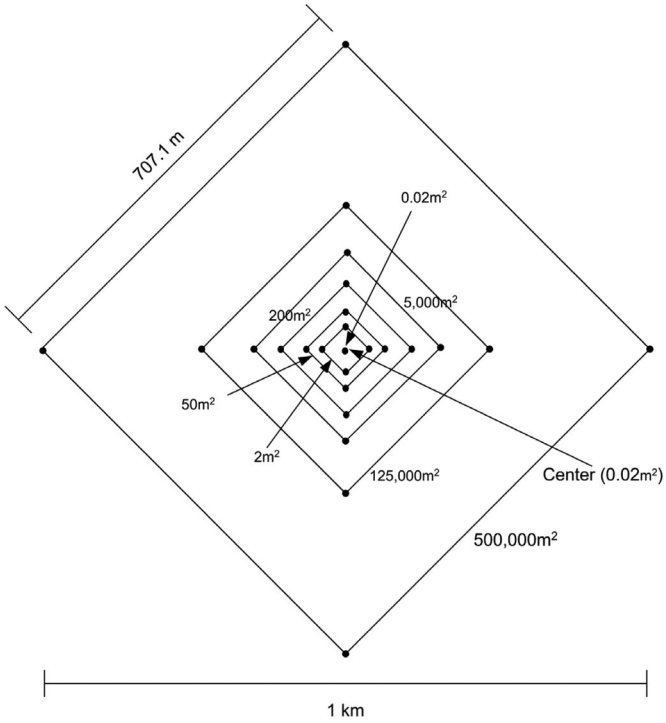

Fig. S1. A schematic map of sampling locations.

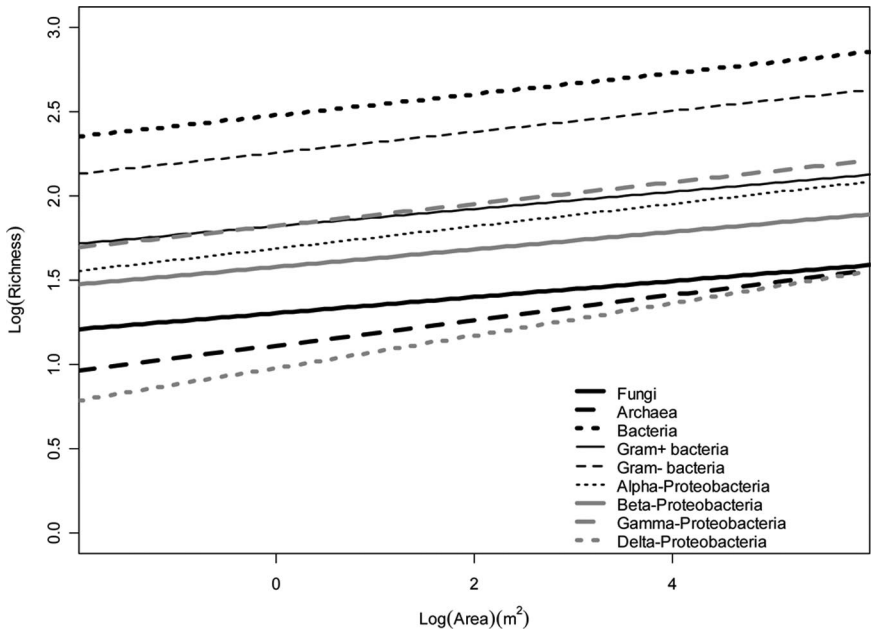

Fig. S2. The gene-area relationships of individual phylogenetics groups based on measurement from the GeoChip hybridization.

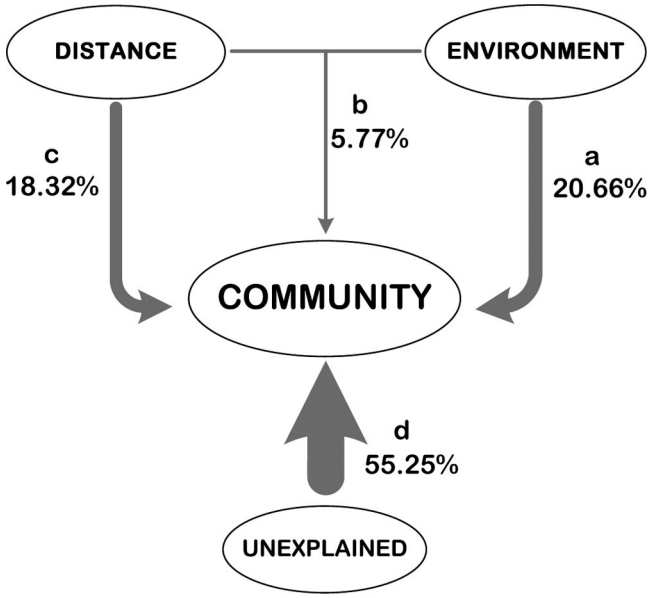

Fig. S3. Variation partitioning based on canonical correspondence analysis (CCA) for all functional gene sequences (FGSs).

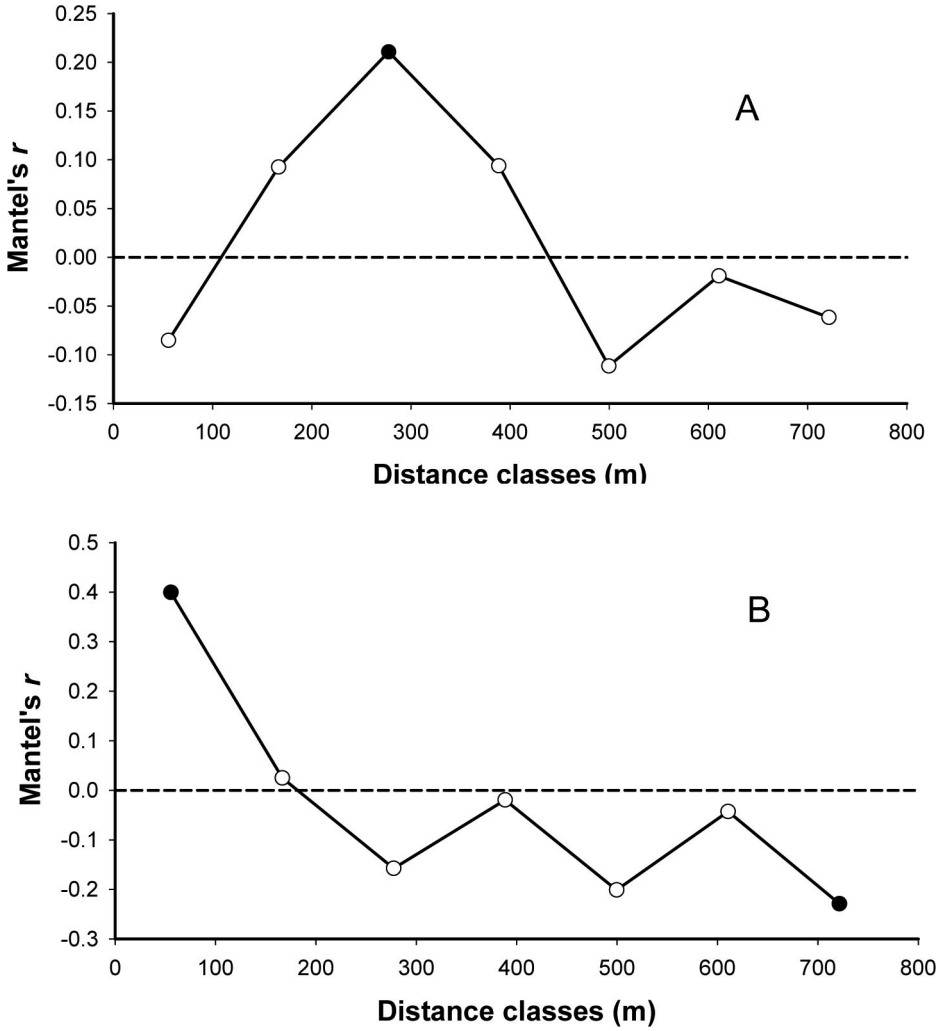

Fig. S4. Multivariate spatial autocorrelogram based on Mantel's *r*. Significant spatial autocorrelation (by permutation and progressive Bonferroni correction) at each distance class is represented by filled circle.

Table S1. Calculated *t* statistics between pairs of functional groups

Functional group	Functional group				
	C degradation	Nitrification	N fixation	N reduction	Organic contaminant degradation
Nitrification	21.67	—	—	—	—
N fixation	47.09	36.80	—	—	—
N reduction	32.72	62.78	75.77	—	—
Organic contaminant degradation	16.06	48.26	65.08	21.09	—
Metal resistance	5.70	39.23	59.11	34.84	14.46

$\alpha' = 0.00333$ (0.05/15). A two-sample *t* test was done between all possible pairs of functional and phylogenetics groups based on variances estimated by bootstrapping (1,000 times without replacement). Bootstrapping was done in *R*, and the *t* test was done in Excel manually. Significance levels were adjusted by Bonferroni correction ($\alpha' = \alpha$/k), and all pairs were significantly different from each other ($P < 0.001$) except between Gram-positive bacteria and and *β-proteobacterium* subdivision ($P = 0.008$) (degree of freedom = 1,998). Critical value of *t* statistics at $\alpha = 0.001$ with df = 1,998 is 3.291.

Table S2. Calculated *t* statistics between pairs of phylogenetics groups

	Phylogenetic group					
Phylogenetic group	Fungi	Archaea	Gram-positive	α	β	γ
Archaea	50.88					
Bacteria	54.79	22.47				
Gram-positive bacteria	13.19	44.25				
Gram-negative bacteria	61.79	24.23	46.36			
α-proteobacteria	66.81	15.77	53.79			
β-proteobacteria	12.99	39.65	2.64	40.89		
γ-proteobacteria	69.86	19.74	54.70	6.90	38.70	
δ-proteobacteria	160.77	37.92	147.92	93.88	119.12	111.19

$\alpha' = 0.00192$ (0.05/26). For further details, see Table S1.

Table S3. The effects of soil geochemistry and geographic distance on microbial community composition

	In association with								
	Soil geochemistry			Geographic distance			Vegetation		
	Controlling for								
	Geographic distance			Soil geochemistry			Soil geochemistry		
All individual gene sequences	r_M 0.002	P_1 0.489	P_2 0.818	r_M 0.064	P_1 0.301	P_2 0.310	r_M −0.004	P_1 0.510	P_2 0.861
Functional genes	0.293	0.043*	0.764	−0.120	0.762	0.231	0.055	0.330	0.825
Functional groups									
C degradation	−0.006	0.481	0.962	0.099	0.234	0.151	0.010	0.449	0.780
Nitrification	−0.038	0.645	0.841	0.004	0.478	0.278	−0.062	0.749	0.911
N fixation	−0.138	0.810	0.920	0.037	0.387	0.333	−0.098	0.728	0.854
N reduction	0.113	0.119	0.901	0.169	0.040*	0.146	0.142	0.057	0.502
Organic cont. degr.	0.029	0.406	0.848	0.029	0.410	0.331	−0.018	0.555	0.884
Metal resistance	−0.007	0.506	0.947	0.028	0.396	0.310	0.028	0.392	0.815
Phylogenetic groups									
Fungi	−0.053	0.628	0.741	0.014	0.391	0.402	−0.048	0.620	0.771
Archaea	0.029	0.348	0.935	0.085	0.232	0.293	0.062	0.291	0.969
Bacteria	−0.009	0.529	0.873	0.065	0.283	0.183	0.003	0.483	0.882
Gram-positive bacteria	0.062	0.273	0.792	0.069	0.265	0.141	0.028	0.388	0.784
Gram- bacteria	−0.030	0.568	0.851	0.046	0.364	0.340	−0.021	0.565	0.831
α-Proteobacteria	−0.045	0.615	0.855	0.005	0.486	0.402	−0.074	0.714	0.790
β-Proteobacteria	0.006	0.429	0.701	0.028	0.368	0.268	0.061	0.288	0.692
γ-Proteobacteria	−0.024	0.559	0.912	0.061	0.315	0.332	−0.086	0.759	0.831
δ-Proteobacteria	−0.062	0.699	0.792	0.100	0.186	0.233	0.100	0.175	0.521

r_M and P_1 represent the partial Mantel test statistic and the P value. P_2 values are based on F test on canonical axes of partial CCA. *, $P < 0.05$.