Journal of Microbiology (2014) Vol. 52, No. 7, pp. 574–580
Copyright ⓒ 2014, The Microbiological Society of Korea

DOI 10.1007/s12275-014-3636-9

Assessment of Microbial Diversity Bias Associated with Soil Heterogeneity and Sequencing Resolution in Pyrosequencing Analyses[§]

Sokhee P. Jung and Hojeong Kang*

School of Civil and Environmental Engineering, Yonsei University, Seoul 120-749, Republic of Korea

(Received Dec 10, 2013 / Revised Mar 26, 2014 /Accepted Apr 10, 2014)

It is important to estimate the true microbial diversities accurately for a comparative microbial diversity analysis among various ecological settings in ecological models. Despite drastically increasing amounts of 16S rRNA gene targeting pyrosequencing data, sampling and data interpretation for comparative analysis have not yet been standardized. For more accurate bacterial diversity analyses, the influences of soil heterogeneity and sequence resolution on bacterial diversity estimates were investigated using pyrosequencing data of oak and pine forest soils with focus on the bacterial 16SrRNA gene. Soil bacterial community sets were phylogenetically clustered into two separate groups by forest type. Rarefaction curves showed that bacterial communities sequenced from the DNA mixtures and the DNAs of the soil mixtures had midsize richness compared with other samples. Richness and diversity estimates were highly variable depending on the sequence read numbers. Bacterial richness estimates (ACE, Chao 1 and Jack) of the forest soils had positive linear relationships with the sequence read number. Bacterial diversity estimates (NPShannon, Shannon and the inverse Simpson) of the forest soils were also positively correlated with the sequence read number. One-way ANOVA shows that sequence resolution significantly affected the α-diversity indices (P<0.05), but the soil heterogeneity did not (P>0.05). For an unbiased evaluation, richness and diversity estimates should be calculated and compared from subsets of the same size.

***Keywords*:** pyrosequencing, diversity index, community analysis, 16S rRNA gene, oak forest soil, pine forest soil

Introduction

Microbial diversity is a fundamental measurement of a microbial community in ecology, and it underlies many ecological models for the establishment of ecological conservation strategies. Microbial diversity takes into account richness and evenness, and diversity indices are quantitative estimates representing how many species there are and how evenly they are distributed in a sample (McCaig *et al.*, 1999; Nübel *et al.*, 1999; Colwell, 2009). For a comparative analysis of microbial diversities in a variety of ecological settings, it is important to estimate the true microbial diversities accurately.

Microbial richness estimations targeting whole microbial genomes reveal great numbers of microbial genomes in soils, but they vary depending on the analytical methods. In a DNA hybridization study, the estimated number of bacterial genomes was ~10^5/g of soil (Torsvik *et al.*, 1990). A computational approach associated with the reassociation kinetics found ~10^7 microbial genomes per gram of soil, which exceeds the previous estimate by two orders of magnitude (Gans *et al.*, 2005). However, a metagenomic approach predicted far lower numbers (Daniel, 2005). In a metagenomic approach, one gram of soil was found to harbor about 2,000 to 18,000 bacterial genomes, and estimations varied depending on the soil texture (Daniel, 2005).

Bacterial diversity analyses based on pyrosequencing targeting 16S rRNA show far less variable results than genome-targeting methods. ~1,000 to ~5,000 OTUs (operational taxonomic units) were found in 0.5 to 1.0 g of soils (Roesch *et al.*, 2007; Acosta-Martínez *et al.*, 2008; Kwon *et al.*, 2010; Will *et al.*, 2010; Nacke *et al.*, 2011; Deng *et al.*, 2012; Lee *et al.*, 2013). Bacterial diversity estimates are highly variable among studies because different analytical approaches are employed. In some forest soil studies, much larger numbers of sequence reads than previous studies were analyzed to get diversity estimates more closely to the true diversity (Roesch *et al.*, 2007; Nacke *et al.*, 2011). In other studies, diversity estimates of samples were compared without any normalization of sequence read numbers (Dunbar *et al.*, 1999, 2000; Roesch *et al.*, 2007; Acosta-Martínez *et al.*, 2008; Kwon *et al.*, 2010; Will *et al.*, 2010; Nacke *et al.*, 2011; Deng *et al.*, 2012; Lee *et al.*, 2013). Averages of replicate diversity estimates were compared to each other, or replicate diversity estimates were comparatively evaluated without any averaging (Hur *et al.*, 2011; Nacke *et al.*, 2011; Deng *et al.*, 2012). Replicate soil samples were combined, and DNA was extracted from the mixture for pyrosequencing analyses (Lee *et al.*, 2013).

Because all these approaches have not been evaluated, the influences of soil heterogeneity and sequence resolution on diversity indices are investigated for more accurate comparative analyses of bacterial diversities in this study. Soil bacterial communities were selected for this study, because soil bacterial communities are very diverse compared to other bacterial communities in artificial systems (Lee *et al.*,

*For correspondence. (H. Kang) E-mail: hj_kang@yonsei.ac.kr; (S. Jung) sokheejung@gmail.com; Tel.: +82-2-2123-5803; Fax: +82-2-364-5300

[§] Supplemental material for this article may be found at http://www.springerlink.com/content/120956.

2010; Jung and Regan, 2011; Jung *et al.*, 2012). Triplicate soil sets from two forests were taken. The DNA of each soil replicate, their DNA mixtures, and the DNA of a soil replicate mixture were investigated, and diversity estimates were evaluated in each bacterial community set with a different sequence read number. In order to overcome bias induced by random selection, five subsets per sample were generated, and their averages were evaluated among different samples. It is found that richness estimates and diversity estimates are highly variable depending on the sequence read numbers.

Materials and Methods

Sampling and DNA extraction

Soil samples were taken from an oak forest and pine forest in the Hong-Reung experimental forest of the Korea Forest Research Institute (South Korea). Triplicate soil samples for each tree type were collected from the top 5 cm from the surface of three sampling points located one meter away from the tree trunk and 120° apart from each other (OA, OB, and OC for the oak tree and PA, PB, and PC for the pine tree). Extracted DNA samples from each tree were mixed (OD and PD), and DNA was extracted from soil mixtures (OS and PS). Total five DNA samples were tested in each forest. DNA was extracted from 0.5 g of soil sample in a 40-μl elution solution using FastDNA SPIN Kit for Soil (MP BIO, Cat. No. 6560-200). The integrity of genomic DNA was confirmed using gel electrophoresis (Supplementary data Fig. S1). DNA concentrations were measured using an Epoch Microplate Spectrophotometer (BioTek® Instruments, Inc.), and they ranged from 273 to 750 ng/μl. DNA samples were diluted to equilibrate concentrations and purified using an UltraClean DNA purification kit (Mo-bio, Cat No. 12100-300) before PCR amplification.

PCR and pyrosequencing

Purified DNA samples were amplified by targeting V1 - V3 regions of the bacterial 16S rRNA gene (~450 bp based on *E. coli* genome) using the primer set of forward primer V1-9F (5′-CCTATCCCCTGTGTGCCTTGGCAGTC-TCAG-AC-GAGTTTGATCMTGGCTCAG-3′) and reverse primer V3-541R (5′-CCATCTCATCCCTGCGTGTCTCCGAC-TCAG-*X*-AC-WTTACCGCGGCTGCTGG-3′). The first two primer sections are the adaptor and key, AC is a linker, and underlined sequences are gene-specific primers. *X* in the reverse primer is a barcode primer. PCR amplification was performed in a 50-μl volume containing 1.25 U Taq DNA Polymerase, 5 μl of 10× PCR reaction buffer, 0.2 mM dNTP mix, 0.4 μM of each primer, and 1 μ of template DNA (Roche Cat. No. 04-728-882-001) with the following thermalcycler program: initial denaturation at 95°C for 5 min; 30 cycles of denaturation at 95°C for 30 sec, annealing at 55°C for 30 sec, and extension at 72°C for 60 sec; and a final extension at 72°C for 7 min in a PTC-200 DNA Engine (MJ Research, USA). The size and contamination of PCR amplicons were confirmed by gel electrophoresis. The quality of PCR products was confirmed by gel electrophoresis. PCR products were purified using a QIAquick PCR Purification Kit (QIAGEN, Cat. No. 28106), and several reactions were pooled in a 1.5-ml tube. Bands shorter than 300 bp were removed using a QIAquick Gel Extraction Kit (QIAGEN, Cat. No. 28706) in subsequent gel electrophoresis. 1 μg of PCR product was subjected to pyrosequencing. The Pyrosequencing was performed with 454 GS FLX Titanium (454 Life Science, Rosche) in Chunlab, Inc. (Korea) according to the manufacturer's instructions.

Pyrosequencing data analysis

DNA sequences were separated by unique barcodes. After sequencing, barcodes, linkers, and gene-specific primers were removed from original sequencing reads. The resultant sequences were filtered to select sequences above 300 bp containing 0 to 1 ambiguous base calls (Ns). Nonspecific sequences (expectation value of $>e^{-5}$) in a BLASTN search and chimeric sequences were removed. For the taxonomic assignment of each pyrosequencing read, the EzTaxon-e database (http://www.eztaxon-e.org) was used (Chun *et al.*, 2007). Operational taxonomic units (OTUs) were defined at the 3% divergence threshold using the average neighbor clustering algorithm. CD-HIT was used for the massive clustering of metagenomic sequences (Fu *et al.*, 2012), and corresponding graphical representations were generated using CLcommunity 3.0 (Chunlab Inc.). Abundance-based coverage estimator (ACE), Chao 1 estimator (Chao), interpolated Jackknife richness estimator (Jack), non-parametric Shannon

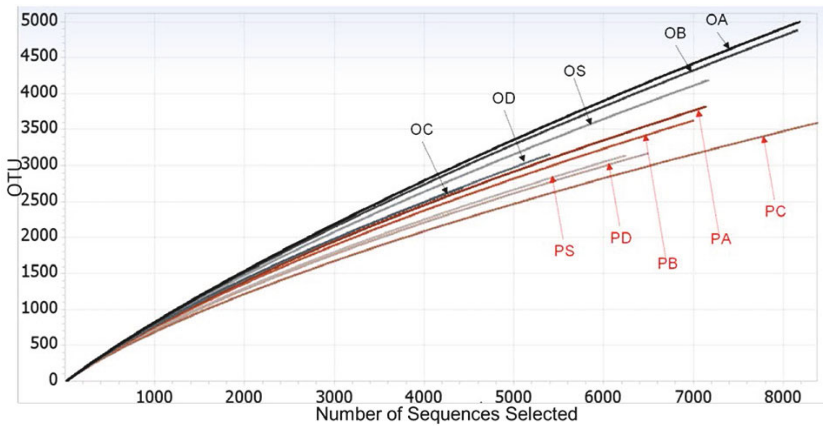

Fig. 1. **Rarefaction curves of pyrosequenced bacterial communities of forest soils. OA, OB and OC are triplicate samples in the oak tree, and PA, PB and PC are triplicate samples for the pine tree. Extracted DNAs from each tree were mixed (OD and PD), and DNA was extracted from soil mixtures (OS and PS).**

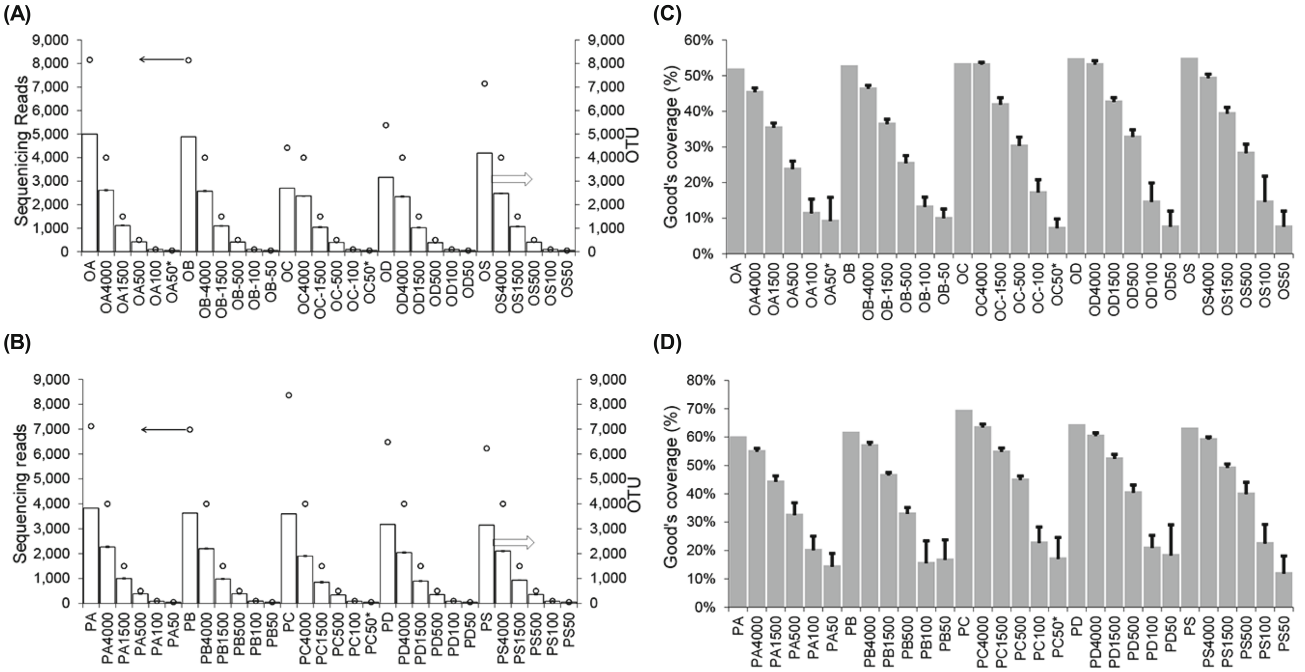

Fig. 2. **Sequencing reads, OTUs and good's coverage values of pyrosequenced bacterial communities of forest soils and their subsets.** Bars and vertical capped bars indicate averages and standard deviations, respectively (n=5). OA, OB, and OC are triplicate samples in the oak tree, and PA, PB, and PC are triplicate samples for the pine tree. Extracted DNAs from each tree were mixed (OD and PD), and DNA was extracted from soil mixtures (OS and PS).

diversity index (NpShannon), Shannon index of diversity (Shannon), Simpson index of diversity (Simpson), and Good's coverage were calculated using Mothur 1.28.0 (Schloss *et al.*, 2009). Analysis of variance (ANOVA) was performed using SPSS 18 (IBM, NY). Fast Unifrac, a variant of the UniFrac algorithm, was used to calculate the distance matrix for β-diversity analysis with sequence normalization and by treating unclassified OTUs as different (Hamady *et al.*, 2010). The multidimensional Fast UniFrac distance matrix was converted into two vectors using principal coordinate analysis (PCoA).

Results and Discussion

The ten soil bacterial community groups were phylogenetically clustered using UPGMA (Unweighted pair group method with arithmetic mean) and PCoA (Supplementary data Fig. S1). In both analyses, the bacterial community groups were phylogenetically clustered into two separate groups according to the forest type (oak and pine). Over the entire range of the sequence reads in rarefaction analysis (Gotelli and Colwell, 2001), the oak soils had higher bacterial richness than the pine soils (Fig. 1), possibly due to the difference in

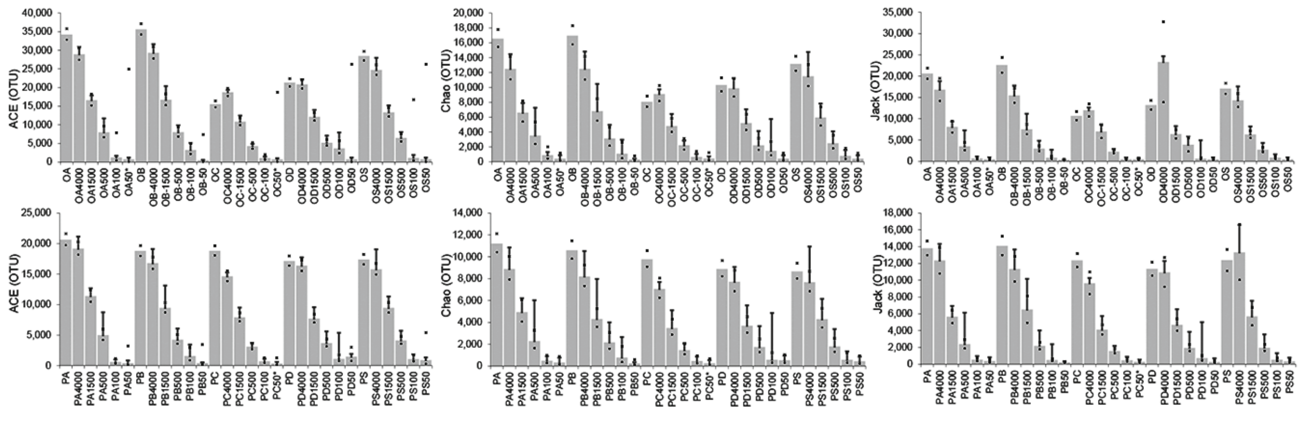

Fig. 3. **Diversity index values of ACE, Chao, and Jack of pyrosequenced bacterial communities of the oak forest soils and the pine forest soils.** Bars and vertical capped bars indicate averages and standard deviations, respectively (n=5). Averages of higher confidence interval (HCI) and lower confidence interval (LCI) at 95% confidence are indicated as × and +, respectively (n=5). OA, OB, and OC are triplicate samples in the oak tree, and PA, PB and PC are triplicate samples for the pine tree. Extracted DNAs from each tree were mixed (OD and PD), and DNA was extracted from soil mixtures (OS and PS).

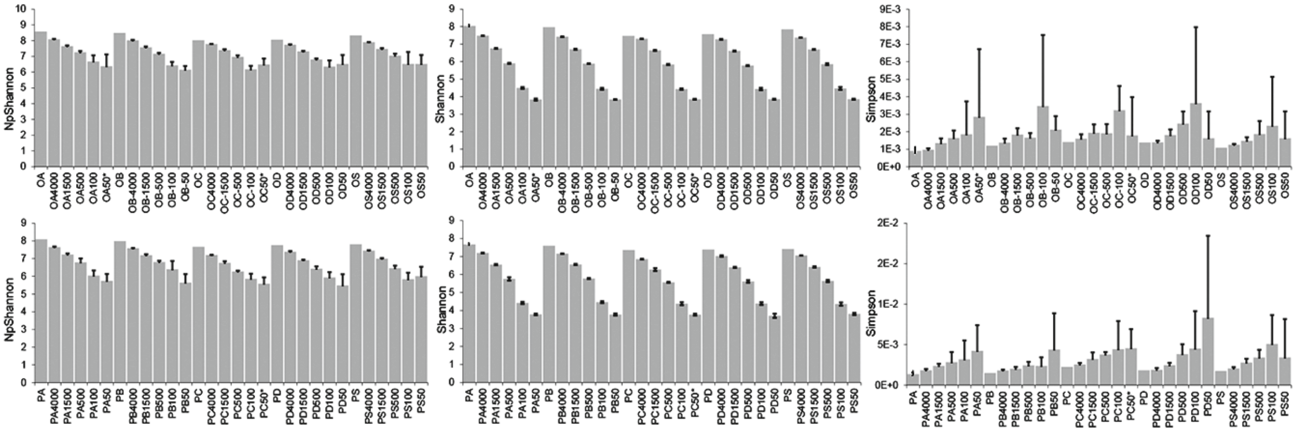

Fig. 4. Diversity index values of NpShannon, Shannon and Simpson of pyrosequenced bacterial communities of the oak forest soils and the pine forest soils. Bars and vertical capped bars indicate averages and standard deviations, respectively (n=5). Average values of higher confidence interval (HCI) and lower confidence interval (LCI) at 95% confidence are indicated as × and +, respectively (n=5). OA, OB and OC are triplicate samples in the oak tree, and PA, PB and PC are triplicate samples for the pine tree. Extracted DNAs from each tree were mixed (OD and PD), and DNA was extracted from soil mixtures (OS and PS).

Table 1. Regression equations of various diversity index values on varying sequence read in the oak- and pine-soil bacterial communities

	OA	OB	OC	OD	OS	OT
OTU	y=0.6473x + 61.80	y=0.6364x + 61.37	y=0.5838x + 71.07	y=0.5764x + 68.01	y=0.6110x + 65.53	y=0.6110x + 65.56
R^2	0.9973	0.9973	0.9956	0.99761	0.9968	0.9934
ACE	y=6.9938x + 2515	y=6.9125x + 3060	y=4.5331x + 1566	y=4.77x + 2658	y=5.9545x + 1967	y=5.8328x + 2353
R^2	0.925	0.9219	0.9956	0.8832	0.9334	0.8668
CHAO1	y=2.9507x + 1174	y=2.9727x + 1125	y=2.1604x + 803	y=2.2843x + 1042	y=2.762x + 851.9	y=2.626x + 999.2
R^2	0.947	0.9589	0.9703	0.9428	0.9963	0.9244
Jack	y=4.1019x + 879	y=3.7798x + 783	y=2.9571x + 808	y=5.6858x + 43.57	y=3.4589x + 686.8	y=3.99670x + 622.7
R^2	0.8957	0.9588	0.884	0.582	0.986	0.7071
	PA	PB	PC	PD	PS	PT
OTU	y=0.5593x + 69.80	y=0.5407x + 75.79	y=0.4667x + 71.24	y=0.5018x + 67.66	y=0.5174x + 67.97	y=0.5172x + 70.49
R^2	0.9953	0.9946	0.9947	0.9956	0.9955	0.9887
ACE	y=4.6909x + 1555	y=3.9881x + 1605	y=3.5785x + 951.8	y=3.8154x + 1392	y=3.7436x + 1681	y=3.9633x + 1437
R^2	0.9327	0.9329	0.9507	0.9604	0.9275	0.9206
CHAO1	y=2.1287x + 777.5	y=1.9276x + 780.0	y=1.6977x + 467.8	y=1.8081x + 639.32	y=1.8205x + 725.2	y=1.8765x + 677.9
R^2	0.9571	0.9605	0.9708	0.9821	0.9597	0.9499
Jack	y=3.0076x + 574.5	y=2.8068x + 706.6	y=2.3558x + 325.1	y=2.6508x + 459.5	y=3.2872x + 334.9	y=2.8217x + 480.1
R^2	0.9685	0.9187	0.9764	0.9473	0.8911	0.9139
	OA	OB	OC	OD	OS	OT
NpShannon	y=0.3814ln(x)+4.870	y=0.4248ln(x)+4.494	y=0.3411ln(x)+4.895	y=0.2992ln(x)+5.133	y=0.325ln(x)+5.121	y=0.3543ln(x)+4.908
R^2	0.7708	0.9653	0.8311	0.7049	0.647	0.7626
Shannon	y=0.8337ln(x)+0.6391	y=0.8232ln(x)+0.6674	y=0.7965ln(x)+0.7832	y=0.7872ln(x)+0.8158	y=0.8089ln(x)+0.7442	y=0.8099ln(x)+0.7299
R^2	0.9976	0.9976	0.997	0.9979	0.9971	0.9962
Simpson^{-1}	y = 0.0874x+660	y=0.0698+459	y=0.0300x+500	y=0.0477x+505	y=0.0105x+715	y=0.0491x+568
R^2	0.1693	0.4385	0.0537	0.0799	0.0013	0.0583
Coverage	y=0.0841ln(x)-0.2563	y=0.0836ln(x)-0.2408	y=0.1008ln(x)-0.3082	y=0.1040ln(x)-0.3255	y=0.0941ln(x)-0.2886	y=0.0933ln(x)-0.2839
R^2	0.9395	0.9774	0.9821	0.9753	0.7568	0.9435
	PA	PB	PC	PD	PS	PT
NpShannon	y=0.4363ln(x)+4.045	y=0.3981ln(x)+4.310	y=0.3573ln(x)+4.169	y=0.4118ln(x)+3.9324	y=0.3577ln(x)+4.382	y=0.3922ln(x)+4.168
R^2	0.9171	0.8145	0.898	0.8411	0.785	0.8095
Shannon	y=0.7806ln(x)+0.8051	y=0.772ln(x)+0.8641	y=0.7034ln(x)+1.101	y=0.753ln(x)+0.8571	y=0.747ln(x)+0.923	y=0.751ln(x)+0.910
R^2	0.9956	0.994	0.9944	0.9941	0.9967	0.9906
Simpson^{-1}	y = 0.0534x+336	y=0.0448+386	y=0.0378x+246	y=0.0797x+232	y=0.0467x+295	y=0.0525x+299
R^2	0.2957	0.2618	0.5635	0.5992	0.3126	0.3151
Coverage	y=0.0915ln(x)-0.2181	y=0.0983ln(x)-0.2544	y=0.1095ln(x)-0.2536	y=0.1018ln(x)-0.2279	y=0.1052ln(x)-0.2684	y=0.1013ln(x)-0.2445
R^2	0.9566	0.9145	0.9552	0.9237	0.9496	0.9121

their soil textures. Bacterial communities from the DNA mixtures and the DNA samples of the soil mixtures (OD, PD, OS, PS) had medium richness compared with other samples, implying that bacterial richness was equalized by mixing soil or DNA samples. In the tested range of sequence reads, all rarefaction curves reached a plateau (Fig. 1). In a previous study, rarefaction curves were also not saturated in soil bacterial communities with >25,000 reads (Nacke *et al.*, 2011). It may be hard for high-throughput pyrosequencing to capture all the diversity of a soil bacterial community.

α-Diversity analyses showed that observed OTUs and Good's coverage were highly dependent on sequence read numbers (Fig. 2). OTU and Good's coverage should reach certain thresholds, but they increase as sequence reads increase (Fig. 2). Richness estimates and diversity estimates were also variable depending on sequence read numbers (Figs. 3 and 4).

Table 2. Bacterial diversity analysis by targeting 16S rRNA gene

Sample	Method	*n*	Target	Enzyme	No. of Sequence	OTUs or Phylotype	ACE	Chao	Shannon	Ref.
Sandy loam between trees	Culture[a]	1	8F-1492R	*Rsa*I+ *Bst*UI	37	7±0.5	-	-	2.405	Dunbar *et al.* (1999)
Sandy loam rhizosphere		1	8F-1492R		37	14	-	-	3.254	
Cinder between trees		1	8F-1492R		37	7±0.6	-	-	1.541	
Cinder rhizosphere		1	8F-1492R		37	15±1.6	-	-	3.337	
Sandy loam between trees	Cloning[a]	1	8F-1492R	-	190	150±1.3	-	-	7.067	Dunbar *et al.* (1999)
Sandy loam rhizosphere		1	8F-1492R	-	190	147±4.3	-	-	7.092	
Cinder between trees		1	8F-1492R	-	190	127±2.3	-	-	6.612	
Cinder rhizosphere		1	8F-1492R	-	190	150	-	-	7.018	
Sandy loam between trees	TRFLP[b]	4	8F-1492R	*Hae*III, *Hha*I, *Msp*I, *Rsa*I	-	22.0±8.0	-	-	4.08±0.51	Dunbar *et al.* (2000)
Sandy loam rhizosphere		4	8F-1492R		-	20.3±5.9	-	-	3.92±0.41	
Cinder between trees		4	8F-1492R		-	18.5±5.0	-	-	3.80±0.28	
Cinder rhizosphere		4	8F-1492R		-	19.8±10.2	-	-	3.81±0.63	
Maize field, Brazil	Pyro	1	V9	-	26140	2369	4888	5021	-	Roesch *et al.* (2007)
Sugarcane field, Florida		1	V9	-	28328	2700	5820	5666	-	
Campus, Illinois		1	V9	-	31818	2692	5890	6040	-	
Boreal forest, Ontario		1	V9	-	53533	5543	13329	20244	-	
Unmanaged beech forest	Pyro	3	V2-V3	-	27642±1374	1734±1254	3794±406	3824±397	5.78±0.19	Nacke *et al.* (2011)
Fertilized grassland		3	V2-V3	-	21808±3153	1134±1301	2828±1333	2887±1277	5.75±0.25	
Fertilized pasture		3	V2-V3	-	26363±3249	1498±1648	2650±550	2720±453	5.70±0.10	
Beech forest		3	V2-V3	-	26954±2989	1134±1669	3600±926	3639±921	5.63±0.22	
Spruce forest		3	V2-V3	-	27881±5584	1509±1584	2089±581	2195±560	5.43±0.60	
Unfertilized pasture		3	V2-V3	-	29555±2653	1302±1482	2082±1157	2226±1074	5.20±0.38	
Heavymetalsites[c]	Pyro	5	V1-V3	-	1300	221	-	354	-	Hur *et al.* (2011)
Heavymetalsites-WT[c]		5	V1-V3	-	1300	421	-	836	-	
Heavymetalsites-GM[c]		5	V1-V3	-	1300	547	-	1188	-	
Soils under ambient CO2	Pyro	12	V4-V5	-	2501±387	847±109	-	-	6.04±0.13	Deng *et al.* (2012)
Soils under elevated CO2		12	V4-V5	-	2424±519	801±122	-	-	5.98±0.16	
Tundra soil, organic 0-2 cm	TRFLP	11	27F-927R	*Hha*I	-	45.7±4.2	-	-	3.47±0.12	Lee *et al.* (2013)
Tundra soil, mineral 5 cm	TRFLP	13	27F-927R		-	43.0±8.4	-	-	3.35±0.21	
Tundra soil, organic 0-2 cm[d]	Pyro	1	V1-V3	-	832	465	2027	1140	5.76	
Tundra soil, mineral 5 cm[d]	Pyro	1	V1-V3	-	2190	1085	4895	2893	6.48	
Oak forest soils	Pyro	5	V1-V3	-	6650±1683	3987±1028	27052±8552	13022±3866	7.78±0.25	This study
Pine forest soils		5	V1-V3	-	7034±827	3474±301	18558±1401	9828±1091	7.49±0.140	
Oak forest soils		25	V1-V3	-	4000	2468±111	24417±4798	11049±1539	7.37±0.08	
Pine forest soils		25	V1-V3	-	4000	2099±130	16473±1762	7888±756	7.06±0.12	
Oak forest soils		25	V1-V3	-	1500	1065±35	13818±3153	5857±972	6.68±0.06	
Pine forest soils		25	V1-V3	-	1500	929±56	9141±1855	4111±626	6.44±0.12	
Oak forest soils		25	V1-V3	-	500	404±13	6378±2559	2724±866	5.85±0.06	
Pine forest soils		25	V1-V3	-	500	363±22	4039±1114	1882±387	5.67±0.10	
Oak forest soils		25	V1-V3	-	100	91±3	2117±2301	1001±735	4.46±0.06	
Pine forest soils		25	V1-V3	-	100	88±4	1053±854	596±239	3.85±0.03	
Oak forest soils		25	V1-V3	-	50	48±1	707±355	424±139	3.77±0.07	
Pine forest soils		25	V1-V3	-	50	45±2	734±1039	376±278	6.44±0.12	

[a] Diversity indices derived from RFLP profiles using *Rsa*I-*Bst*UI

[b] Averages of diversity indices from TRFP profiles using *Hae*III, *Hha*I, *Msp*I, and *Rsa*I

[c] Heavy metal-contaminated sites planted with no poplar, wild type poplar (WT), and genetically-modified poplar (GM)

[d] DNA from a soil mixture.

Regression analysis was performed to investigate the numerical relationships between sequence read numbers and each α-diversity indices. Bacterial richness estimates (ACE, Chao 1, and Jack) of the forest soils had positive linear relationships with the sequence read number (Table 1), where the oak soils had stronger correlations than the pine soils. The inverse Simpson index, a transformed form of the Simpson index, is a commonly used index in application because it is equal to the true diversity of order 2 (Hill, 1973; Jost, 2006). Bacterial diversity estimates (NPShannon, Shannon and the inverse Simpson) of the forest soils were also positively correlated with the sequence read number (Table 1). One-way ANOVA shows that sequence resolution significantly affected the α-diversity indices ($P<0.05$), but soil heterogeneity did not affect the α-diversity indices ($P>0.05$), showing that sampling strategies does not make significant difference in α-diversity analyses on complex soil bacterial communities.

In previous richness and diversity analyses on bacterial communities (Table 2), bacterial richness and diversity estimates were calculated and compared without any normalization of sequence reads. Because bacterial richness and diversity estimates are highly dependent on sequence read numbers, it is necessary to construct multiple random subsets of the same size and to compare the average estimates of the subsets with the values of interest when it comes to comparing richness and diversity estimates from different sources (Jung *et al.*, 2014a, 2014b).

Richness and diversity estimates have also been calculated from various community characterization methods such as bacterial cultivation (Dunbar *et al.*, 1999), cloning and sequencing (Dunbar *et al.*, 1999), and T-RFLP (Dunbar *et al.*, 2000; Lee *et al.*, 2013) (Table 2). For those diversity estimates from different microbial ecological methods, great care is required for comparative diversity analysis. In a previous study on four forest soils, two bacterial community diversities from DNA sequencing of cultivated cells and direct DNA sequencing were compared for data calibration between two methods (Dunbar *et al.*, 1999) (Table 2). Though both methods showed the same site-specific characteristics, significant discrepancies were observed in α-diversity estimates and phylotypes. T-RFLP was applied to the same samples, and it showed that T-RFLP did not provide reliable diversity measures (Dunbar *et al.*, 2000). Diversity estimates were highly dependent on the restriction enzyme and were not consistent in each sample. However, average estimates of multiple T-RFLP sets showed a concurrent site-specificity as the two methods above, indicating that T-RFLP can be an effective comparative diversity analysis tool among low-resolution methods (Dunbar *et al.*, 2000). However, diversity estimates were very different between T-RFLP and pyrosequencing analysis, as revealed by Tundra soil study (Lee *et al.*, 2013), suggesting that comparative diversity analysis between a low-resolution T-RFLP and a high-resolution pyrosequencing should be avoided.

DNA concentrations of the pine forest soil (537 ± 172 ng/μl) were more variable than those of the oak forest soils (565 ± 71 ng/μl), with a total average DNA concentration of 551 ± 125 ng/μl (Supplementary data Table S2). Absorbance ratios (OD260/OD280) in the pine forest soils (1.78 ± 0.06) were generally lower than those of the oak forest soils (1.91 ± 0.01). 109,411 sequences (≥300 bp) were recovered from pyrosequencing. 63 ± 2% in total sequences were classified into the bacteria domain. One-way ANOVA shows that DNA quality did not affect sequence read numbers. The average numbers of bacteria sequences and observed OTU were 6,842 ± 1,267 and 3,731 ± 763 per sample, respectively, where the observed OTU number was ~55% of the bacterial sequence number (Supplementary data Table S2). Previous soil studies utilizing the 454 pyrosequencing platform produced 4–5 times more sequence reads with fewer OTUs than those in our study, where different hyper-variable regions of shorter lengths were targeted. In one forest soil study with sequence lengths of ~100 bp, the average number of bacterial sequences was 34,955 ± 3,326, and the observed OTU number was 3,326 ± 1,486 (9.5%) (Roesch *et al.*, 2007). In another soil study with a minimum read length of >200 bp, the average number of bacterial sequences was 26,700 ± 3,778, and the observed OTU number was 1,415 ± 283 (5.3%) (Nacke *et al.*, 2011). Despite the smaller size of sequence data sets than these previous two studies, our data detected more bacterial diversity. This suggests that the pyrosequencing platform in our study generated high-quality data by capturing more diversity with less sequencing effort.

Acknowledgements

The authors thank In-Seon Son (ChunLab, Inc.) for her assistance in bioinformatics. This research was supported by the National Research Foundation of Korea (# 2013056833).

References

Acosta-Martínez, V., Dowd, S., Sun, Y., and Allen, V. 2008. Tag-encoded pyrosequencing analysis of bacterial diversity in a single soil type as affected by management and land use. *Soil Biol. Biochem.* **40**, 2762–2770.

Chun, J., Lee, J.H., Jung, Y., Kim, M., Kim, S., Kim, B.K., and Lim, Y.W. 2007. EzTaxon: a web-based tool for the identification of prokaryotes based on 16S ribosomal RNA gene sequences. *Int. J. Syst. Evol. Microbiol.* **57**, 2259–2261.

Colwell, R.K. 2009. Biodiversity: concepts, patterns, and measurement. The Princeton Guide to Ecology, pp. 257–263. *In* Levin, S.A., Princeton University Press, Princeton, NJ, USA.

Daniel, R. 2005. The metagenomics of soil. *Nat. Rev. Microbiol.* **3**, 470–478.

Deng, Y., He, Z., Xu, M., Qin, Y., Van Nostrand, J.D., Wu, L., Roe, B.A., Wiley, G., Hobbie, S.E., Reich, P.B., and Zhou, J. 2012. Elevated carbon dioxide alters the structure of soil microbial communities. *Appl. Environ. Microbiol.* **78**, 2991–2995.

Dunbar, J., Takala, S., Barns, S.M., Davis, J.A., and Kuske, C.R. 1999. Levels of bacterial community diversity in four arid soils compared by cultivation and 16S rRNA gene cloning. *Appl. Environ. Microbiol.* **65**, 1662–1669.

Dunbar, J., Ticknor, L.O., and Kuske, C.R. 2000. Assessment of microbial diversity in four southwestern united states soils by 16S rRNA gene terminal restriction fragment analysis. *Appl. Environ. Microbiol.* **66**, 2943–2950.

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152.

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

Gotelli, N.J. and Colwell, R.K. 2001. Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecol. Lett.* **4**, 379–391.

Hamady, M., Lozupone, C., and Knight, R. 2010. Fast UniFrac: facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and PhyloChip data. *ISME J.* **4**, 17–27.

Hill, M.O. 1973. Diversity and evenness: A unifying notation and its consequences. *Ecology* **54**, 427–432.

Hur, M., Kim, Y., Song, H.R., Kim, J.M., Choi, Y.I., and Yi, H. 2011. Effect of genetically modified poplars on soil microbial communities during the phytoremediation of waste mine tailings. *Appl. Environ. Microbiol.* **77**, 7611–7619.

Jost, L. 2006. Entropy and diversity. *Oikos* **113**, 363–375.

Jung, S., Ahn, Y.H., Oh, S.E., Lee, J., Cho, K.T., Kim, Y., Kim, M.W., Shim, J., and Kang, M. 2012. Impedance and thermodynamic analysis of bioanode, abiotic anode, and riboflavin-amended anode in microbial fuel cells. *Bull. Korean Chem. Soc.* **33**, 3349–3354.

Jung, S. and Regan, J.M. 2011. Influence of external resistance on electrogenesis, methanogenesis, and anode prokaryotic communities in microbial fuel cells. *Appl. Environ. Microbiol.* **77**, 564–571.

Jung, S.P., Cheong, Y., Yim, G., Ji, S., and Kang, H. 2014a. Performance and bacterial communities of successive alkalinity-producing systems (SAPSs) in passive treatment processes treating mine drainages differing in acidity and metal levels. *Environ. Sci. Pollut. Res.* **21**, 3722–3732.

Jung, S.P., Yoon, M.H., Lee, S.M., Oh, S.E., Kang, H., and Yang, J.K. 2014b. Power generation and anode bacterial community compositions of sediment fuel cells differing in anode materials and carbon sources. *Int. J. Electrochem. Sci.* **9**, 315–326.

Kwon, S., Kim, T.S., Yu, G.H., Jung, J.H., and Park, H.D. 2010. Bacterial community composition and diversity of a full-scale integrated fixed-film activated sludge system as investigated by pyrosequencing. *J. Microbiol. Biotechnol.* **20**, 1717–1723.

Lee, S.H., Jang, I., Chae, N., Choi, T., and Kang, H. 2013. Organic layer serves as a hotspot of microbial activity and abundance in arctic tundra soils. *Microb. Ecol.* **65**, 405–414.

Lee, T., Doan, T., Yoo, K., Choi, S., Kim, C., and Park, J. 2010. Discovery of commonly existing anode biofilm microbes in two different wastewater treatment MFCs using FLX titanium pyrosequencing. *Appl. Microbiol. Biotechnol.* **87**, 2335–2343.

McCaig, A.E., Glover, L.A., and Prosser, J.I. 1999. Molecular analysis of bacterial community structure and diversity in unimproved and improved upland grass pastures. *Appl. Environ. Microbiol.* **65**, 1721–1730.

Nübel, U., Garcia-Pichel, F., Kühl, M., and Muyzer, G. 1999. Quantifying microbial diversity: Morphotypes, 16S rRNA genes, and carotenoids of oxygenic phototrophs in microbial mats. *Appl. Environ. Microbiol.* **65**, 422–430.

Nacke, H., Thürmer, A., Wollherr, A., Will, C., Hodac, L., Herold, N., Schöning, I., Schrumpf, M., and Daniel, R. 2011. Pyrosequencing-based assessment of bacterial community structure along different management types in german forest and grassland soils. *PLoS One* **6**, e17000.

Roesch, L.F.W., Fulthorpe, R.R., Riva, A., Casella, G., Hadwin, A.K.M., Kent, A.D., Daroub, S.H., Camargo, F.A.O., Farmerie, W.G., and Triplett, E.W. 2007. Pyrosequencing enumerates and contrasts soil microbial diversity. *ISM J.* **1**, 283–290.

Schloss, P.D., Westcott, S.L., Ryabin, T., Hall, J.R., Hartmann, M., Hollister, E.B., Lesniewski, R.A., Oakley, B.B., Parks, D.H., Robinson, C.J., Sahl, J.W., Stres, B., Thallinger, G.G., Van Horn, D.J., and Weber, C.F. 2009. Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* **75**, 7537–7541.

Torsvik, V., Goksøyr, J., and Daae, F.L. 1990. High diversity in DNA of soil bacteria. *Appl. Environ. Microbiol.* **56**, 782–787.

Will, C., Thürmer, A., Wollherr, A., Nacke, H., Herold, N., Schrumpf, M., Gutknecht, J., Wubet, T., Buscot, F., and Daniel, R. 2010. Horizon-specific bacterial community composition of german grassland soils, as revealed by pyrosequencing-based analysis of 16S rRNA genes. *Appl. Environ. Microbiol.* **76**, 6751–6759.