The ISME Journal (2012) 6, 1137–1147

www.nature.com/ismej

ORIGINAL ARTICLE

454 Pyrosequencing reveals bacterial diversity of activated sludge from 14 sewage treatment plants

Tong Zhang, Ming-Fei Shao and Lin Ye
Environmental Biotechnology Laboratory, Department of Civil Engineering, The University of Hong Kong, Hong Kong

Activated sludge (AS) contains highly complex microbial communities. In this study, PCR-based 454 pyrosequencing was applied to investigate the bacterial communities of AS samples from 14 sewage treatment plants of Asia (mainland China, Hong Kong, and Singapore), and North America (Canada and the United States). A total of 259 K effective sequences of 16S rRNA gene V4 region were obtained from these AS samples. These sequences revealed huge amount of operational taxonomic units (OTUs) in AS, that is, 1183–3567 OTUs in a sludge sample, at 3% cutoff level and sequencing depth of 16 489 sequences. Clear geographical differences among the AS samples from Asia and North America were revealed by (1) cluster analyses based on abundances of OTUs or the genus/family/order assigned by Ribosomal Database Project (RDP) and (2) the principal coordinate analyses based on OTUs abundances, RDP taxa abundances and UniFrac of OTUs and their distances. In addition to certain unique bacterial populations in each AS sample, some genera were dominant, and core populations shared by multiple samples, including two commonly reported genera of *Zoogloea* and *Dechloromonas*, three genera not frequently reported (i.e., *Prosthecobacter*, *Caldilinea* and *Tricoccus*) and three genera not well described so far (i.e., Gp4 and Gp6 in *Acidobacteria* and Subdivision3 genera incertae sedis of *Verrucomicrobia*). Pyrosequencing analyses of multiple AS samples in this study also revealed the minority populations that are hard to be explored by traditional molecular methods and showed that a large proportion of sequences could not be assigned to taxonomic affiliations even at the phylum/class levels.
The ISME Journal (2012) **6,** 1137–1147; doi:10.1038/ismej.2011.188; published online 15 December 2011
Subject Category: microbial population and community ecology
Keywords: pyrosequencing; activated sludge; bacteria; core population; cluster analysis; PCoA

Introduction

Similar to soil and sediment, activated sludge (AS) is a highly complex system of eukaryotes (protozoa and fungi), bacteria, archaea and viruses, in which bacteria are dominant. Despite its importance in biological wastewater treatment, the microbial influence on the formation, development and function of AS remains largely unstudied (Wagner and Loy, 2002). This is partly attributable to the lack of robust techniques needed to explore the complex microbial communities.

The low sequencing depth of the traditional PCR-cloning approach when compared with the vast genetic diversity present in AS systems hindered a comprehensive characterization of the microbial community structure. In this aspect, the current community analyses typically represent a mere snapshot of the dominant members, with little information on taxa with medium to low abundances. High-throughput sequencing is a promising method, as it provides enough sequencing depth to cover the complex microbial communities (Shendure and Ji, 2008). So far, it has been applied to analyze microbial communities in marine water (Qian *et al.*, 2010), soil (Roesch *et al.*, 2007; Lauber *et al.*, 2009), human hand surface (Fierer *et al.*, 2008), human distal intestine (Claesson *et al.*, 2009), etc. But few studies have been conducted using this method to investigate AS, though this approach has been used to study the microbial community in the raw sewage (McLellan *et al.*, 2010) and the residual biosolids after digestion of AS (Bibby *et al.*, 2010).

Using computational ecology tools such as the UniFrac β-diversity metric (Lozupone and Knight, 2005) and principal coordinates analysis (PCoA), researchers can compare differences in microbial communities between ecosystems and along environmental gradients.

With the aid of high-throughput sequencing technology and the well-established β-diversity analytical tool, an attempt could be made to answer some fundamental questions related to AS microbial communities. In this study, AS samples were collected from 14 sewage treatment plants (STPs) in mainland China, Hong Kong, Singapore, the United States and Canada. Pyrosequencing using the

Correspondence: T Zhang, Environmental Biotechnology Laboratory, Department of Civil Engineering, The University of Hong Kong, Pokfulam Road, Hong Kong.
E-mail: zhangt@hku.hk
Received 22 May 2011; revised 9 November 2011; accepted 9 November 2011; published online 15 December 2011

16S rRNA gene as the biomarker was conducted to examine the bacterial diversity of these AS samples, to evaluate the similarity/difference of different samples, to compare the unique dominant bacterial populations, and to identify the core populations shared by different samples. To the best of our knowledge, this study is the first application of PCR-based 454 pyrosequencing to characterize and compare multiple AS samples.

Materials and methods

STPs and AS

As shown in Table 1, AS samples were taken from the aeration tanks of 14 full-scale STPs that treat municipal wastewater in Asia (mainland China, Hong Kong and Singapore) and North America (Canada and the United States). These STPs treat mainly municipal wastewater, except that CN-QD-TD, CN-GZ-DT and CN-NJ-SJ have a significant industrial wastewater component. The three STPs from North America apply the conventional AS (CAS) or oxidation ditch processes. Those from Asia use A/O (anoxic/aerobic) or A/A/O (anaerobic/anoxic/aerobic) processes, two of them are membrane bioreactors, and one of them treats saline (1.1% salinity) sewage. When the samples were taken, the water temperature ranged from 10 to 31 °C. All sludge samples were briefly settled on site to be concentrated and finally fixed in 50% (v/v) ethanol aqueous solution. The fixed samples were immediately transported to the laboratory for further treatment.

DNA extraction, PCR amplification and pyrosequencing

Of these samples, the same sludge sample from Sha-Tin (Hong Kong) was divided into two aliquots, that is, CN-HK-ST1 and CN-HK-ST2. Then the two subsamples were treated as independent samples from DNA extraction to pyrosequencing, to evaluate the reproducibility of the methods applied in this study. Samples of 10 ml were centrifuged at 4000 r.p.m. for 10 min at 4 °C. Pellet (200 mg) of each sample was collected for DNA extraction in duplicate with the FastDNA SPIN Kit for Soil (MP Biomedicals, Solon, OH, USA), which was found to be the most suitable (having the lowest contamination) for the samples in this study, compared with the ZR Soil Microbe DNA Kit, the SoilMaster DNA Extraction Kit, the PowerSoil DNA Isolation Kit and the UltraClean Soil DNA Isolation Kit. The duplicate DNA extracts were then merged together for the following PCR amplification.

The forward primer 563F (5′-AYTGGGYD TAAAGNG-3′) at the 5′-end (*E. coli* positions 563–578) of the V4 region (239 nucleotides) and a cocktail of four equally mixed reverse primers, that is, R1 (5′-TACCRGGGTHTCTAATCC-3′), R2 (5′-TAC CAGAGTATCTAATTC-3′), R3 (5′-CTACDSRGGTMTC TAATC-3′) and R4 (5′-TACNVGGGTATCTAATC-3′), at the 3′-end of the V4 region (*E. coli* positions 785–802) were selected for PCR, because they may capture 95% of all bacterial 16S rRNA gene sequences in databases (Murphy *et al.*, 2010). They have also been used in other microbial diversity studies employing the PCR-based pyrosequencing method (Claesson *et al.*, 2009; Jesus *et al.*, 2010; Murphy *et al.*, 2010). The 5′-fused primer includes a 10 nucleotide 'barcode' inserted between the Roche 454 life Science (Branford, CT, USA) adapter primer A and the 563F primer. The barcode is permuted for each sample and allows the identification of individual samples in a mixture in a single pyrosequencing run (Sogin *et al.*, 2006). A 100 μl reaction system was set up for each PCR amplification, using MightyAmp polymerase (TaKaRa, Otsu, Japan). The amplification was conducted in an i-Cycler (BioRad, Hercules, CA, USA) under the following conditions: initial denaturation at 98 °C for 2 min, and 28 cycles at 98 °C for 15 s, 56 °C for 20 s and 68 °C for 30 s, and a final extension at 68 °C for 10 min. Amplicon libraries were prepared by a cocktail of three independent PCR products for each sample to minimize the impact of potential early-round PCR errors (Sogin *et al.*, 2006). PCR amplicons were purified with a quick-spin Kit (iNtRON, Seoul, Korea), and concentrations were measured using spectrometry (NanoDrop-1000, Thermo Scientific, Wilmington, DE, USA). Amplicons from different sludge samples were then mixed to achieve equal mass concentrations in the final mixture, which was sent out for pyrosequencing on the Roche 454 FLX Titanium platform (Roche) at the Genome Research Center of the University of Hong Kong. The raw reads have been deposited into the NCBI short-reads archive database (Accession Number: SRA026842.2).

Post-run analysis

All the raw reads were treated with the Pyrosequencing Pipeline Initial Process (Cole *et al.*, 2009) of the Ribosomal Database Project (RDP), (1) to sort those exactly matching the specific barcodes into different samples, (2) to trim off the adapters, barcodes and primers using the default parameters, and (3) to remove sequences containing ambiguous 'N' or shorter than 150 bps (Claesson *et al.*, 2009). The reads selected above were defined as 'raw reads' for each AS sample.

Sequences were denoised using the 'pre.cluster' command in Mothur platform to remove sequences that are likely due to pyrosequencing errors (Huse *et al.*, 2010; Roeselers *et al.*, 2011). PCR chimeras were filtered out using Chimera Slayer (Haas *et al.*, 2011). The reads flagged as chimeras was extracted out and submitted to RDP. Those being assigned to any known genus with 90% confidence were merged with the non-chimera reads, to form the collection of 'effective reads' for each AS sample.

Although bacteria-specific primers were applied, very small portions of unexpected archaeal sequences might be obtained (Qian *et al.*, 2010). To remove these

Table 1 Characteristics of 14 STPs

Group	*ID*	*STP*		*Percentage (%) of municipal wastewater*	*Process*	*Flow rate (10^3 m^3 d^{-1})*	*Average (or range) ($mg\,l^{-1}$)*			*Latitude*	*Temperature (°C)*		*Sampling time (mm/year)*
		Code	*Full name (city and country)*				*COD*	*TN*	*TP*		*Range*	*At sampling*	
I	01	CN-NJ-SJ	Suo-Jin-Cun (Nanjing, PRC)	85	A/O	15	330	57	5.8	32.1	16–29	22	11/2009
	02	CN-QD-TD	Tuan-Dao (Qingdao, PRC)	70	A/A/O	100	900	120	10	36.1	12–20	17	12/2009
	03	CN-HR-UN	Ha-Er-Bin (Haerbin, PRC)	Predominant	A/O	325	420	63	NA	45.7	9–24	10	12/2009
	04	CN-SH-MH	Min-Hang (Shanghai, PRC)	70	A/O	44	267 (140–576)	40 (34–53)	2.9 (2-6)	31.2	12–28	16	11/2009
	05	CN-BJ-BX	Bei-Xiao-He (Beijing, PRC)	95	A/A/O+MBR	100	462	51	9.3	39.9	16–25	16	12/2009
	06	CN-WH-LW	Long-Wang-Zui (Wuhan, PRC)	Predominant	A/A/O	161	277	23	4.9	30.5	16–29	17	12/2009
	07	CN-GZ-DT	Da-Tan-Sha (Guangzhou, PRC)	60	A/A/O	150	595	55	5.2	23.1	20–30	26	04/2010
	08	SG-SG-UP	Ulu Pandan (Singapore)	Predominant	CAS+MBR	23	265	45	9.2	1.3	23–32	27	05/2010
II	09	US-CO-CO	Columbia Regional (Columbia, USA)	Predominant	CAS	76	700 (500–1100)	32 (21–42)	NA	38.9	11–23	17	04/2010
	10	US-GR-PG	Potato Creek (Griffin, USA)	Predominant	OD	8	402	30	4	33.2	15–20	18	05/2010
	11	CA-GP-GP	Guelph (Guelph, Canada)	Predominant	CAS	55	270	45	7	43.5	13–23	13	01/2010
III	12	CN-HK-ST1	Sha-Tin (Hong Kong, PRC)[a]	95	A/O	216	(226–491)	(26–40)	(3.1–5.3)	22.3	22–32	27	11/2009
	13	CN-HK-ST2	Sha-Tin (Hong Kong, PRC)[a]	95	A/O	216	(226–491)	(26–40)	(3.1–5.3)	22.3	22–32	27	11/2009
IV	14	CN-HK-SH	Shek-Wu-Hui (Hong Kong, PRC)	90	A/O	216	(206–529)	(28–41)	(3.0–5.6)	22.3	21–31	30	08/2010
V	15	CN-HK-SL	Stanley (Hong Kong, PRC)	100	A/O	8	(196–481)	(25–42)	(3.2–5.5)	22.3	24–32	31	08/2010

Abbreviations: A/O, anoxic/aerobic; A/A/O, anaerobic/anoxic/aerobic; CAS, conventional activated sludge; MBR, membrane bioreactor; NA, not available; OD: oxidation ditch; STPs, sewage treatment plants.
[a]Saline (1.1% salinity) sewage due to seawater toilet flushing practice in that area.

cross-talking sequences, the effective sequences of each AS sample were submitted to the RDP Classifier (Wang *et al.*, 2007) to identify the archaeal and bacterial sequences, and filter out the archaeal sequences using a self-written Python script. The average length of all bacterial sequences without the primers was 207 bp.

After the above filtration, the minimum number of selected bacterial sequences in the 15 AS samples was 16 489. To fairly compare the 15 AS samples at the same sequencing depth, normalization of the sequence number was conducted by extracting the first 16 489 sequences in each sample for all the following analyses.

Taxonomic classification of the bacterial sequences of each AS sample was carried out individually, using the RDP Classifier. A bootstrap cutoff of 50% suggested by the RDP was applied to assign the sequences to different taxonomy levels.

The normalized sequence set of each AS sample was also individually aligned by Infernal (Nawrocki and Eddy, 2007) using the bacteria-alignment model in Align module of the RDP. By applying the Complete Linkage Clustering, sequences in each set were assigned to phylotype clusters at two cutoff levels of 3% and 6%. On the basis of these clusters, rarefaction curves, ACE and Chao1 richness were calculated using the relevant RDP modules, including Rarefaction and Chao1 Estimator. Good's coverage was calculated as $G = 1 - n/N$, where n is the number of singleton phylotypes and N is the total number of sequences in the sample.

The cluster analysis (CA) was conducted to group the bacterial communities of different AS samples on the basis of (1) taxonomy results obtained using the RDP Classifier (excluding those unclassified sequences), and (2) operational taxonomic units (OTUs) generated using RDP Complete Linkage Clustering from the merged pool of sequences of all the AS samples. The names of bacterial sequences from each sample were encoded specifically using another self-written Python script to identify their sources in the merged sequence pool. With the encoded sequence name, a matrix of these OTUs and their abundances in the 15 AS samples was compiled using the third self-written Python script. The number of OTUs in each AS sample was counted from this matrix and summarized in Supplementary Table S1. Then, the CA was conducted using the unweighted-pair group mean averages, based on the Bray–Curtis distance calculated from the matrix using PAST software (McLellan *et al.*, 2010; Qian *et al.*, 2010). Similarly, CA was also conducted using a matrix of the RDP taxa (at each level from genus to class) and their abundances in each sample.

PCoA was conducted based on (1) RDP Classifier results, (2) OTUs described above and (3) weighted UniFrac (Hamady *et al.*, 2010). The first two are phylogenetically independent methods and were conducted using PAST, somewhat similar to the above CA. For UniFrac, which is a phylogeny-dependent method, the Greengenes coreset tree was selected as the reference tree. A sample mapping file showing the frequency of each reference taxon in different AS samples was generated through local BLAST following the protocol in UniFrac's tutorial (http://bmf2.colorado.edu/fastunifrac/tutorial.psp) and then uploaded to http://bmf2.colorado.edu/fastunifrac/ to conduct weighted UniFrac PCoA according to the instructions.

Results and discussion

Diversity of microbial communities

As shown in Supplementary Table S1, after filtering the low quality reads using the RDP Initial Process in Pyrosequencing Pipeline (PP) and trimming the adapters, barcodes and primers, there were 20 276 ~ 28 260 effective reads for the 15 AS samples. After denoising, filtering out chimeras, and removing the archaeal sequences, the library size of each sample was normalized to 16 489 sequences, which were the smallest among the 15 samples, to conduct the downstream analyses for different samples at the same sequencing depth.

The numbers of OTUs, Chao 1 and ACE at two cutoff levels of 3% and 6% are summarized in Supplementary Table S1. On the basis of the OTU number, the AS sample from Shek-Wu-Hui (Hong Kong) had the richest diversity, followed closely by those from Sha-Tin (Hong Kong) and Ulu Pandan (Singapore), whereas the three samples from North America displayed considerably less richness, especially that from Columbia Regional STP (USA). The patterns of Chao 1 and ACE values were very similar to the OTU numbers. All three indices, that is, OTU number, Chao 1 and ACE, demonstrated that the richness values varied by 2–3 times among these AS samples. Plots of OTU number versus sequence number, that is, the rarefaction curves, are shown in Supplementary Figure S2.

As a common hypothesis, diversity may affect the performance of the AS process. However, an accurate estimation of OTUs in an AS sample, based on DNA sequencing, has not been conducted before. In this study, based on the 247 335 effective bacterial sequences, there could be a total of 13 951 (3% cutoff) and 8493 OTUs (6%) in the 15 samples. The OTU numbers ranged from 1183 to 3567 in different samples, similar to the numbers of bacterial OTUs (3% cutoff) in soils from Brazil, Florida (USA) and Illinois (USA), but much less than that in the soil of Canada (Roesch *et al.*, 2007), at the same sequencing depth. It should be noted that the richness values were certainly affected by sequencing noise.

The 16 489 selected effective bacterial sequences in each samples were assigned to different taxa levels (from genus to phylum) using the RDP Classifier at 50% threshold. Although it has been reported that the V4 region used in this study

displayed the highest number of correctly classified sequences, followed by the V3 and V6 regions (Claesson *et al.*, 2009), quite a large portion of effective bacterial sequences in this study could not be assigned to any taxa of different level at the 50% threshold, indicating the extent of novel sequences captured by this study. Supplementary Figure S1 shows that the unclassified sequence portions in the total community increased from the domain level to the genus level, and were significantly different among the samples, especially at the family and genus levels. For example, 43% and 57% of sequences in the CN-QD-TD sample could not be assigned to any taxa at families and genera levels, respectively, whereas the US-CO-CO and CA-GP-GP samples only contained 20–21% (family) and 32–34% (genus) unclassified taxa, respectively.

As shown in Figure 1, *Proteobacteria* was the most abundant phylum in all samples, accounting for 36–65% of total effective bacterial sequences. This is similar to the analytical results of bacterial communities in soil (Roesch *et al.*, 2007) and sewage (McLellan *et al.*, 2010), in which *Proteobacteria* was also the most dominant community. The other dominant phyla were *Firmicutes* (1.4–14.6%, averaging at 8.1%), *Bacteroidetes* (2.7–15.6%, averaging at 7.0%) and *Actinobacteria* (1.3–14.0%, averaging at 6.5%). Similar to a few previous studies on AS using microarray (Xia *et al.*, 2010) and cloning (Snaidr *et al.*, 1997), these four groups were dominant (56–86%) in bacterial communities of the 15 AS samples in this study, followed by a few other major (average abundance >1%) phyla, including *Verrucomicrobia* (4.2%), *Chloroflexi* (3.4%), *Acidobacteria* (3.0%) and *Planctomycetes* (2.4%). A few phyla, including TM7, *Thermotogae*, OD1, *Spirochaetes*, WS3, *Nitrospira* and *Synergistetes*, were only the major (abundance >1%) phyla in one of the 15 samples. The abundances of other phyla were <1% in all samples.

Within *Proteobacteria*, ε-*Proteobacteria* only occurred at very low levels (0.01–0.51%, averaging 0.08%). Except for the four Hong Kong sludge samples, which had the α-subdivision as the most dominant class within *Proteobacteria*, in all other 11 samples, the β-subdivision was the most dominant *Proteobacteria*, followed by α-, γ- and δ-subdivisions. This is different from results of a study using microarray (Xia *et al.*, 2010), which showed that α-subdivision was the most abundant of the *Proteobacteria*. However, the findings are similar to the results of another study about soil bacteria using pyrosequencing (Roesch *et al.*, 2007), which demonstrated that in most soils, the β-subdivision was the most abundant one within the *Proteobacteria*.

It is interesting that in the AS sample from the unique saline (due to sea water toilet flushing practice) STP at Sha-Tin of Hong Kong, the α-subdivision was the dominant class within the *Proteobacteria*, accounting for 18.5–25.7% of total bacterial effective sequences, whereas the α-*Proteobacteria* only averaged 10.7% in the other 13 AS samples. This might be explained by the dominance (up to 25–50%) of the α-*Proteobacteria* in the marine microbial community of the surface/subeuphotic layers (Venter *et al.*, 2004; Qin *et al.*, 2010). In addition, the δ-*Proteobacteria* was the most dominant class in the other two STPs (12.9% at Stanley and 10.2% at Shek-Wu-Hui) of Hong Kong, but their most dominant orders in the δ-*Proteobacteria* were different, that is, *Myxococcales* (11%) and *Desulfobacterales* (5.8%), respectively.

In addition to the four classes of the *Proteobacteria*, other dominant (>1%) shared (occurring in >7

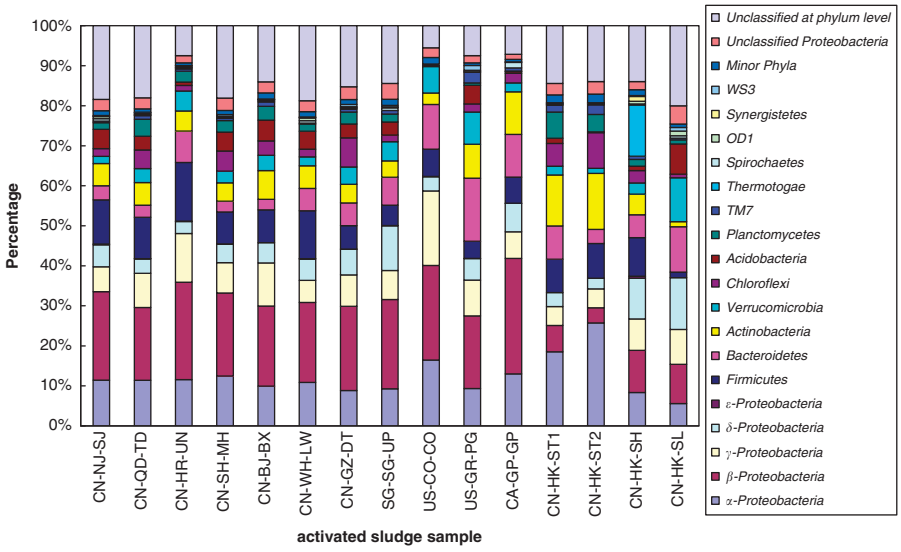

Figure 1 Abundances of different phyla and classes in *Proteobacteria* in the 15 AS samples. The abundance is presented in terms of percentage in total effective bacterial sequences in a sample, classified using RDP Classifier at a confidence threshold of 50%. Taxa represented occurred at >1% abundance in at least one sample. Minor phyla refer to the taxa with their maximum abundance <1% in any sample.

samples) classes included *Actinobacteria*, *Sphingobacteria*, *Clostridia*, *Bacilli*, *Planctomycetacia*, *Verrucomicrobiae*, *Anaerolineae*, *Verrucomicrobia* Subdivision3, *Flavobacteria*, *Caldilineae*, *Acidobacteria*_Gp4 and *Acidobacteria*_Gp6 (Supplementary Table S2).

Similarity analysis of the 15 sludge samples

The similarity of the 15 sludge samples was evaluated using two independent methods: CA and PCoA.

Cluster analysis. As shown in Figure 2a, CA, based on abundances of orders, revealed that bacterial communities in the 15 samples could be clustered into five groups: (1) Group I contains all AS samples from mainland China and that from Singapore; (2) Group II contains the three samples from North America; (3) Group III is sludge from Sha-Tin (Hong Kong), which treats saline sewage; (4) Group IV is sludge from Shek-Wu-Hui (Hong Kong), which treats sewage containing slaughterhouse wastewater; (5) Group V is sludge from Stanley (Hong Kong), which is located inside a cave.

This grouping pattern was similar at the family and the genus levels (Supplementary Figure S4b S4c), but slightly changed at the class levels (Supplementary Figure S4a), and eventually disappeared at the phylum level. Using 0.6 as a benchmark (McLellan *et al.*, 2010), the bacterial compositions (at the family/order levels) of the different sludge samples are quite similar in Group I, even though the STPs in mainland China and Singapore are separated by over thousands km. Using the OTUs abundances, similar grouping patterns were observed at both 3% (Figure 2b) and 6% cutoffs (Supplementary Figure S4d).

Principal coordinates analysis. PCoA was also conducted to evaluate similarities of different AS samples using three different approaches, that is, RDP Classifier taxa, OTUs and UniFrac. For the first two approaches, taxa or OTUs are regarded as equally related, whereas UniFrac incorporates the degree of divergence in the phylogenetic tree of OTUs into PCoA (Hamady *et al.*, 2010; Qian *et al.*, 2010). The PCoA analysis results are shown in Figure 3 (UniFrac at 3% cutoff) and Supplementary Figure S5a (order) and Supplementary Figure S5b (OTUs at 3% cutoff). PCoA at the family, genus and 6% cutoff levels were also conducted. Although there are slight variances among these PCoA results, the same general trend was observed, that is, the sludge samples from mainland China forming clusters different from the sludge from North America.

We hypothesized that there would be significant differences among the sludge from different geographical areas. As demonstrated by the CA and PCoA, the sludge samples were certainly similar to those from the same geographical area, possibly due to unique sewage compositions, as well as different plant design and operation in each area (such as A/O or A/A/O process applied in Chinese STP for biological nutrient removal).

Shared and distinct orders/families

At the order level, the 27 abundant ($>1\%$ in any AS sample) orders accounted for 45–80% of the classified sequences, as shown in Supplementary Figure S3. Among the 27 orders, *Rhodocyclales*, *Burkholderiales*, *Rhizobiales*, *Myxococcales*, *Clostridiales*, *Sphingobacteriales*, *Actinomycetales*, *Rhodobacteriales*, *Xanthomonadales*, *Planctomycetales*, *Pseudomonadales* and *Verrucomicrobiales* were the orders commonly shared by all sludges. However, two orders were significantly ($P<0.1$) more abundant in

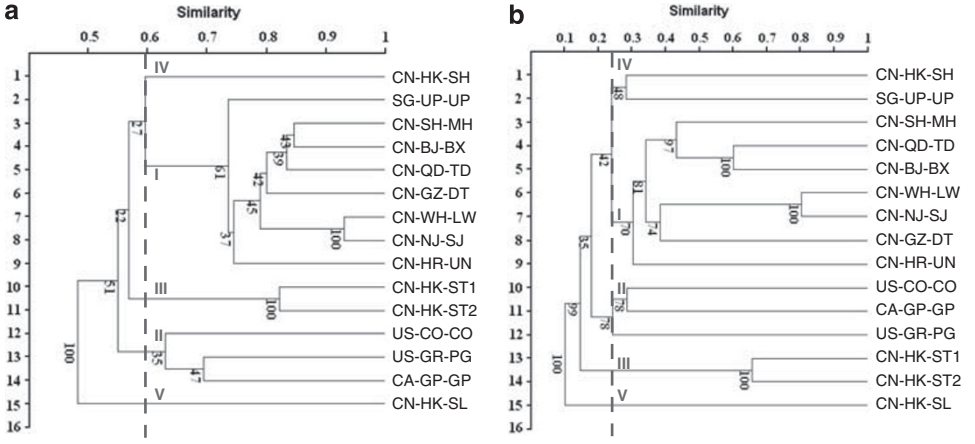

Figure 2 CA based on Bray–Curtis distances of 15 AS samples. (**a**) At order level; (**b**) at 3% cutoff-OTU level. The dot lines show the similarity cutoff levels to cluster the 15 AS samples into five groups: Group I contains all AS from mainland China and that from Singapore; Group II contains the three samples from North America; Group III is sludge from Sha-Tin (Hong Kong), which treats saline sewage; Group IV is sludge from Shek-Wu-Hui (Hong Kong), which treats sewage containing slaughterhouse wastewater; Group V is sludge from Stanley (Hong Kong), which is located inside a cave. For 3% cutoff, 0.25 was selected, whereas 0.6 were selected at the order level, as the lower taxonomy level shows more differences.

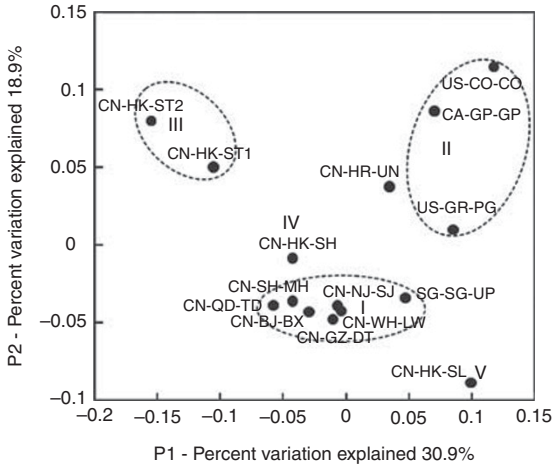

Figure 3 Principal coordinate analysis of 15 AS samples by weighted UniFrac. For UniFrac, Greengenes coreset tree was selected as the reference tree, a sample mapping file showing the frequency of each reference taxon in different AS samples was generated through local BLAST and was uploaded to http://bmf2.colorado.edu/fastunifrac/ to conduct weighted UniFrac PCoA following the instructions (Hamady *et al.*, 2010). PCoA at order level (Supplementary Figure S5a) and 3% cutoff-OTU level (Supplementary Figure S5b) were also conducted using PAST in a way similar to the above CA. For these three PCoA results, similar group patterns were obtained as CA shown in Figure 2. The color reproduction of this figure is available at the *ISME Journal* online.

North American sludge, that is, *Burkholderiales* and *Sphingobacteriales*, whereas three (i.e., *Rhizobiales*, *Planctomycetales* and *Desulfobacterales*) were more abundant in mainland China's sludge. CA after removing these distinct five orders shows mixed clusters of sludge from the two geographical areas.

At the family level, the top 10 families in each sample are summarized in Supplementary Table S3. The CA based on these data showed distinct geographical clusters for AS samples. *Nocardioidaceae*, *Streptococcaceae*, *Cystobacteriaceae*, *Polyangiaceae*, *Rhodocyclaceae* and *Verrucomicrobiaceae* were the families commonly shared by all sludge, whereas *Flavobacteriaceae*, *Comamonadaceae* and *Sphingomonadaceae* were significantly more abundant in North American sludge. *Anaerolineaceae*, *Planctomycetaceae*, *Bradyrhizobiaceae*, *Desulfobacteraceae*, *Hyphomicrobiaceae* and *Rhodospirillaceae* were more abundant in mainland China's sludge.

Core and distinct genera

Supplementary Figure S5 shows the genus profiles for the 15 AS samples. Each panel represents one of the 739 genera, which were sorted alphabetically by phyla and then genus (see Supplementary Table S4 for their names and abundances in 15 AS samples). Examination of these genera captured the similarities and differences in the genus profiles of the 15 samples.

Comparative analysis revealed a core microbiota across the 15 AS samples. As shown in Table 2, among the 744 assigned genera, 70 (accounting for 63.7% of the classified sequences) were shared by all 15 samples. A total of 235 genera were commonly shared by more than 10 AS samples, accounting for 92.7% of all classified sequences. There were 271 rare genera that only appeared in one or two samples, accounting for only 0.9% of total classified sequences, a very minor part of the bacterial communities in AS.

Table 2 Percentages of the shared genera and their corresponding sequences

Number of sample[a]	*Number of shared genera*	*Percentage in classified genera*[b]	*Percentage in classified sequences*[c]
15	70	9.5	63.7
14	121	16.4	81.2
13	158	21.4	86.5
12	184	24.9	89.3
11	208	28.1	91.0
10	235	31.8	92.7
9	253	34.2	93.6
8	278	37.6	94.0
7	305	41.3	94.5
6	339	45.9	97.3
5	373	50.5	97.7
4	404	54.7	98.4
3	468	63.3	99.1
2	556	75.2	99.7
1	739	100.0	100.0

Abbreviation: AS, activated sludge.
Among the 744 assigned genera, 70 (accounting for 63.7% of the classified sequences) were shared by all 15 samples. A total of 235 genera were commonly shared by more than 10 AS samples, accounting for 92.7% of all classified sequences. There were 273 rare genera that only appeared in one or two samples, accounting for only 0.9% of total sequences.
[a]Number of AS samples, which share the genera in the second column.
[b]Percentage of the number of shared genera in the number of total classified genera.
[c]Percentage of sequences of the shared genera in the total classified sequences.

The top 10 abundant genera in each sample were selected (a total of 58 genera for all 15 samples) and compared with their abundances in other samples, as shown in Figure 4 and Supplementary Table S5. Eight genera were abundant (>1%) in at least six samples, including two genera extensively reported in AS, that is, *Zoogloea* and *Dechloromonas*, three genera rarely reported before, that is, *Caldilinea*, *Tricoccus* and *Prosthecobacter*, plus three not well-described genera, that is, Gp4, Gp6 and Subdivision3_genera_incertae_sedis. The species in the genus *Zoogloea*, such as *Zoogloea ramigera*, are known to form characteristic cell aggregates embedded in extracellular gelatinous matrices, often called zoogloeal matrices (Dugan *et al.*, 1992), and are the main agent for the flocculation of AS (Rossello-Mora *et al.*, 1995). They existed in extremely high abundances in two AS samples, that is, CN-HR-UN (8.3%) and CA-GP-GP (11.1%), in high abundances (1.38–11.1%) in most of the other samples, but in low abundances in all four samples from Hong Kong (0.05–1.50%). *Dechloromonas* is a

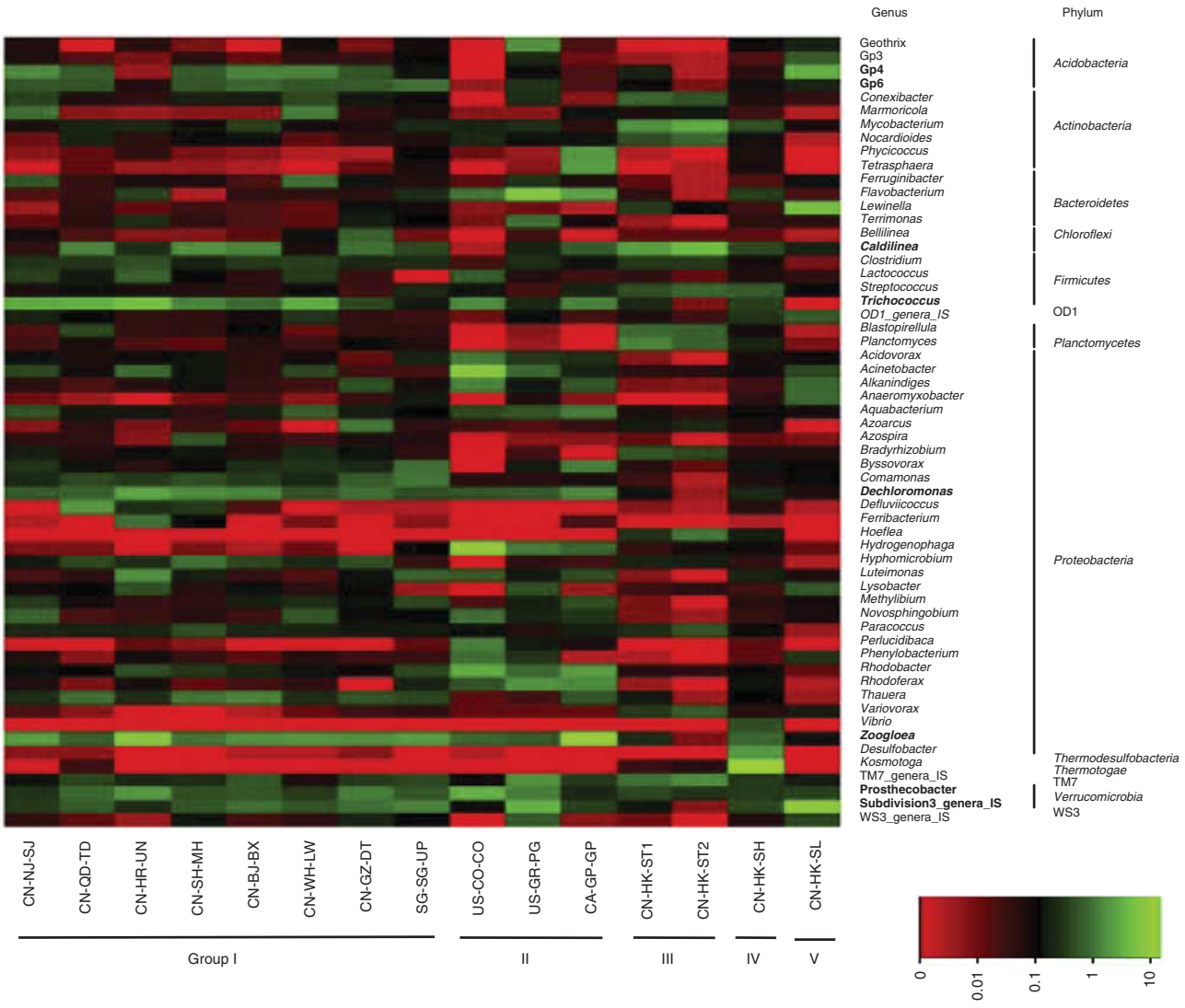

Figure 4 Heat map of top 10 genera in each sample. The top 10 abundant genera in each sample were selected (a total of 58 genera for all 15 samples) and compared with their abundances (percentages) in other samples. The color intensity (log scale) in each panel shows the percentage of a genus in a sample, referring to color key at the right bottom. Those in bold font are the core genera in different AS samples.

genus capable of reducing perchlorate (Achenbach *et al.*, 2001) and also frequently reported as phosphate accumulating organisms in enhanced biological phosphorus removal reactors (Liu *et al.*, 2005). *Caldilinea* has some filamentous species and has a role in stabilizing flocs of AS in a wide range of STPs (Yoon *et al.*, 2010). The information about the existence and roles of other genera in AS are limited.

Just like at the family level, the dominant genera showed some geographical characteristics. For example, *Flavobacterium* were only dominant in the three samples from North America, with abundance levels ranging from 1.83 to 7.44%, whereas its abundances were $<1\%$ in all the samples from mainland China and Singapore. In addition, all three North American samples had high levels of the photosysthetic bacteria *Rhodobacter* (1.43–3.72%), whereas the samples from mainland China and Singapore contained relatively less (0.32–0.99%). As shown in Figure 5, a few other genera more dominant in North America included *Acinetobacter*, *Rhodoferax, Acidovorax*, *Aquabacterium* and *Alkanindiges*, whereas Gp4, Gp6, *Clostridium*, *Hyphomicrobium*, *Comamonas* and *Pirellula* were more abundant in mainland China's sludge.

The distribution of some abundant genera also depended on temperature. For instance, the species in *Trichococcus* genus have been identified as psychrotolerant mesophiles and are able to grow at low temperatures. The sequences assigned to *Trichococcus* were found at high abundances (1.55–5.53%) in AS samples from colder areas, such as that from Ha-Er-Bin (10 °C at sampling), and at low abundances (0–0.96%) in AS samples from sub-tropical/tropical areas, including Hong Kong, Guangzhou, Potato Creek (Georgia, USA) and Singapore.

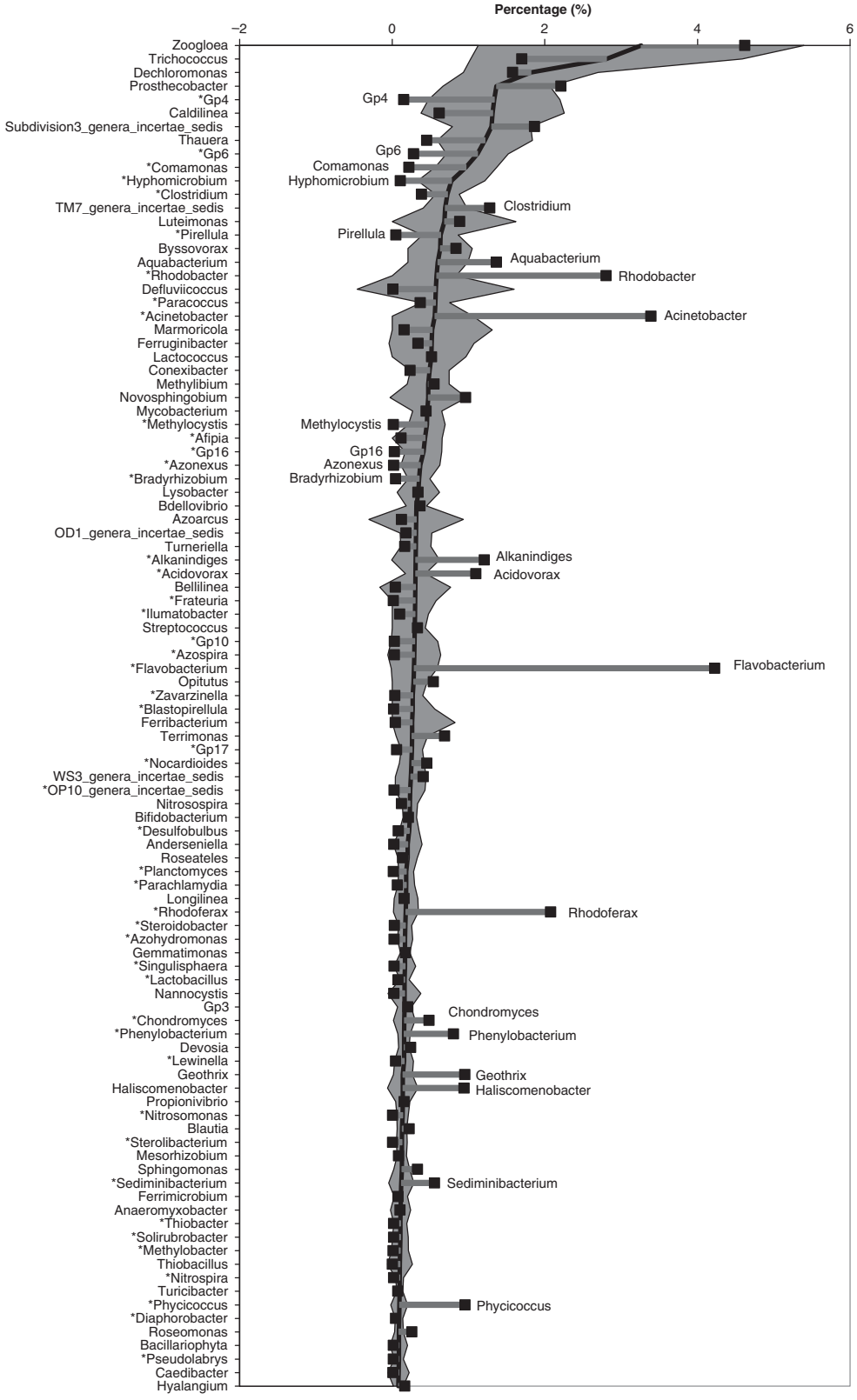

Figure 5 Average abundances (percentages) of the top 100 genera in eight AS samples of Group I (mainland China and Singapore), and their average percentages in three AS samples of Group II (North America). The blue line shows the average percentages in Group I and the shadow area shows the variation ranges (average ± s.d.). The up-bars and down-bars show the average percentages of the corresponding genera in Group II. *Shows the genera, which had significantly different average abundances in Groups I and II. The color reproduction of this figure is available at the *ISME Journal* online.

The sludge samples also contained some other unique populations, possibly due to salinity, exposure to sunlight, industrial wastewater contribution and so on. Two of the Hong Kong samples (ST1 and ST2), which had been exposed to seawater, contained unique genera such as *Hoeflea*, of which all three known species are related to marine environments (Palacios *et al.*, 2006). The CN-HK-SL sample, possibly due to its unique location inside a cavern in Hong Kong, had a very low abundance of the photosysthetic bacteria *Rhodobacter* (0.07%) when compared with the very high range in three North American samples (1.43–3.72%) and the moderately high range in other samples (0.19–0.99%). In addition, the sum of *Rhodobacteriales*, *Rhodocyclales* and *Rhodospirillales* was only 4.0% in CN-HK-SL, far less than the average value of 13.5% (7.5–24.2%) in all other samples. The CN-HK-SH sludge from Shek-Wu-Hui in Hong Kong had fecal bacteria, including *Enterobacteriales*, *Synergistales* and *Campylobacterales*, at levels which were 14, 14 and 10 times higher than their averages in all other samples, possibly due to the contribution of wastewater from a large-scale local slaughtering plant. The unique populations in different AS samples could be due to many possible causes, including geographical isolation, concentration and types of organic substrates in the sewage, which are significantly affected by the diets of eastern and western peoples, operation mode (A/A/O or A/O vs conventional AS), dissolved oxygen concentration, temperature, salinity, season, type and percentage of industrial wastewater, etc. Although it is out of the scope of the present study, the factors shaping microbial community deserve more comprehensive and systematic studies in the future, using the methodology demonstrated here.

Compare with previous studies on AS

This is the first systematic analysis of the bacterial communities of multiple AS samples conducted by examining >16 489 16S rRNA gene fragments per sample. Our analysis demonstrates that the 15 samples had some core genera even though they were collected from geographically separated STPs operating under different conditions and used to treat different types of sewage. At the same time, deep sequencing using the PCR-based 454 pyrosequencing technique also revealed the unique and distinct taxa in different samples.

Previously, most studies on the diversity of AS microbial communities have heavily relied on clone library analysis, which sequenced far fewer 16S rRNA gene fragments (Snaidr *et al.*, 1997) or microarray (Xia *et al.*, 2010). Our findings are generally consistent with these previous studies. For instance, *Proteobacteria* were found to dominate in all AS samples, followed by *Bacteroidetes*, *Firmicutes* and *Actinobacteria*. This is also similar to the analytical results of bacterial communities in soil (Roesch *et al.*, 2007) and sewage influent (McLellan *et al.*, 2010), in which *Proteobacteria* was the most dominant phylum. However, our findings also differ from the previous studies, as extremely diverse microbial communities were found in AS samples, which were impossible to be detected in earlier studies because of their limited methodologies. Although it is still difficult to draw firm conclusion on the roles that these species might have in sludge floc formation and pollution control, the sequences obtained in this study provide us a glimpse of minor microbial taxa (groups) that have been overlooked due to methodological limitations. More attention should be paid to them in further work.

The sludge samples varied greatly due to differences in sewage composition, organic loading, pH, temperature, dissolved oxygen and sludge retention time applied at the aeration tank. More comparative studies based on well-designed sampling plans are needed to determine the main factors shaping microbial communities.

The technical limitations of pyrosequencing may affect our results. First, the primers applied may generate 16S rRNA gene fragments with limited efficiency (Hong *et al.*, 2009; Wang and Qian, 2009), although the primer selection bias is the same for all samples and would not significantly affect the overall comparison of these sludges. Second, biases are also likely associated with the DNA extraction, although the optimal extraction kit has been used in this study after comparing five kits. Additionally, the length of the 16S rRNA gene amplicons obtained in this study was short, about 207 bps (excluding primer sequences), although this length and region may have taxonomic classification accuracy of ~85% (Nossa *et al.*, 2010). Finally, the rarefaction curves (Supplementary Figure S2) suggest that some samples, such as that from Shek-Wu-Hui in Hong Kong, were still under-sampled even with 16 489 effective sequences.

Acknowledgements

We would like to thank the Hong Kong General Research Fund (HKU7197/08E) for financially supporting this study. Dr Ming-Fei Shao thanks HKU for the postdoctoral fellowship and Mr Lin Ye thanks HKU for the postgraduate studentship. We would like to thank Professor Gao DW, Professor Deng BL, Dr Huang QG, Dr Zhu HG, Dr Liang DW, Dr Duan JZ, Dr Zhang M, Dr Zhang XX, Mr Yu K and Miss Yang Y, for the sludge sampling.

References

Achenbach LA, Michaelidou U, Bruce RA, Fryman J, Coates JD. (2001). *Dechloromonas agitata* gen. nov., sp. nov. and *Dechlorosoma suillum* gen. nov., sp. nov., two novel environmentally dominant (per)chlorate-reducing bacteria and their phylogenetic position. *Int J Syst Evol Microbiol* **51**: 527–533.

Bibby K, Viau E, Peccia J. (2010). Pyrosequencing of the 16S rRNA gene to reveal bacterial pathogen diversity in biosolids. *Water Res* **44**: 4252–4260.

Claesson M, O'Sullivan O, Wang Q, Nikkila J, Marchesi J, Smidt H *et al.* (2009). Comparative analysis of pyrosequencing and a phylogenetic microarray for exploring microbial community structures in the human distal intestine. *PloS One* **4**: e6669.

Cole JR, Wang Q, Cardenas E, Fish J, Chai B, Farris RJ *et al.* (2009). The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res* **37**: D141–D145.

Dugan PR, Stoner DL, Pickrum HM. (1992). The genus Zoogloea. In: *The prokaryotes*. Balows A, Trüper HG, Dworkin M, Harder W, Schleifer K-H (eds). Springer-Verlag: New York, NY, pp 3952–3964.

Fierer N, Hamady M, Lauber CL, Knight R. (2008). The influence of sex, handedness, and washing on the diversity of hand surface bacteria. *Proc Natl Acad Sci USA* **105**: 17994.

Haas BJ, Gevers D, Earl AM, Feldgarden M, Ward DV, Giannoukos G *et al.* (2011). Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Res* **21**: 494–504.

Hamady M, Lozupone C, 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.

Hong S, Bunge J, Leslin C, Jeon S, Epstein S. (2009). Polymerase chain reaction primers miss half of rRNA microbial diversity. *ISME J* **3**: 1365–1373.

Huse SM, Welch DM, Morrison HG, Sogin ML. (2010). Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environ Microbiol* **12**: 1889–1898.

Jesus EC, Susilawati E, Smith S, Wang Q, Chai B, Farris R *et al.* (2010). Bacterial communities in the rhizosphere of biofuel crops grown on marginal lands as evaluated by 16S rRNA gene pyrosequences. *BioEnergy Res* **3**: 20–27.

Lauber CL, Hamady M, Knight R, Fierer N. (2009). Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale. *Appl Environ Microbiol* **75**: 5111.

Liu Y, Zhang T, Fang HHP. (2005). Microbial community analysis and performance of a phosphate-removing activated sludge. *Bioresour Technol* **96**: 1205–1214.

Lozupone CA, Knight R. (2005). Unifrac: A new phylogenetic method for comparing microbial communities. *Appl Envrion Microbiol* **71**: 8228–8235.

McLellan S, Huse S, Mueller Spitz S, Andreishcheva E, Sogin M. (2010). Diversity and population structure of sewage derived microorganisms in wastewater treatment plant influent. *Environ Microbiol* **12**: 378–392.

Murphy E, Cotter P, Healy S, Marques T, O'Sullivan O, Fouhy F *et al.* (2010). Composition and energy harvesting capacity of the gut microbiota: relationship to diet, obesity and time in mouse models. *Gut* **59**: 1635–1642.

Nawrocki EP, Eddy SR. (2007). Query-dependent banding (QDB) for faster RNA similarity searches. *PLoS Comput Biol* **3**: e56.

Nossa CW, Oberdorf WE, Yang L, Aas JA, Paster BJ, Desantis TZ *et al.* (2010). Design of 16S rRNA gene primers for 454 pyrosequencing of the human foregut microbiome. *World J Gastroenterol* **16**: 4135–4144.

Palacios L, Arahal D, Reguera B, Marin I. (2006). *Hoeflea alexandrii* sp. nov., isolated from the toxic dinoflagellate *Alexandrium minutum* AL1V. *Int J Syst Evol Microbiol* **56**: 1991.

Qian P, Wang Y, Lee O, Lau S, Yang J, Lafi F *et al.* (2010). Vertical stratification of microbial communities in the Red Sea revealed by 16S rDNA pyrosequencing. *ISME J* **5**: 507–518.

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.

Roesch L, Fulthorpe R, Riva A, Casella G, Hadwin A, Kent A *et al.* (2007). Pyrosequencing enumerates and contrasts soil microbial diversity. *ISME J* **1**: 283–290.

Roeselers G, Mittge EK, Stephens WZ, Parichy DM, Cavanaugh CM, Guillemin K *et al.* (2011). Evidence for a core gut microbiota in the zebrafish. *ISME J* **5**: 1595–1608.

Rossello-Mora R, Wagner M, Amann R, Schleifer K. (1995). The abundance of Zoogloea ramigera in sewage treatment plants. *Appl Environ Microbiol* **61**: 702.

Shendure J, Ji H. (2008). Next-generation DNA sequencing. *Nat Biotechnol* **26**: 1135–1145.

Snaidr J, Amann R, Huber I, Ludwig W, Schleifer KH. (1997). Phylogenetic analysis and *in situ* identification of bacteria in activated sludge. *Appl Environ Microbiol* **63**: 2884–2896.

Sogin M, Morrison H, Huber J, Welch D, Huse S, Neal P *et al.* (2006). Microbial diversity in the deep sea and the underexplored 'rare biosphere'. *Proc Nat Acad Sci* **103**: 12115.

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA *et al.* (2004). Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**: 66–74.

Wagner M, Loy A. (2002). Bacterial community composition and function in sewage treatment systems. *Curr Opin Biotechnol* **13**: 218–227.

Wang Q, Garrity G, Tiedje J, Cole J. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* **73**: 5261–5267.

Wang Y, Qian P. (2009). Conservative fragments in bacterial 16S rRNA genes and primer design for 16S ribosomal DNA amplicons in metagenomic studies. *PloS One* **4**: e7401.

Xia S, Duan L, Song Y, Li J, Piceno Y, Andersen G *et al.* (2010). Bacterial community structure in geographically distributed biological wastewater treatment reactors. *Environ Sci Technol* **44**: 1043–1045.

Yoon DN, Park SJ, Kim SJ, Jeon CO, Chae JC, Rhee SK. (2010). Isolation, characterization, and abundance of filamentous members of Caldilineae in activated sludge. *J Microbiol* **48**: 275–283.

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