The ISME Journal (2009) 3, 1365–1373

www.nature.com/ismej

ORIGINAL ARTICLE

Polymerase chain reaction primers miss half of rRNA microbial diversity

SunHee Hong[1,4], John Bunge[2], Chesley Leslin[1], Sunok Jeon[3] and Slava S Epstein[1,4]
[1]*Department of Biology, Northeastern University, Boston, MA, USA;* [2]*Department of Statistical Science, Cornell University, Ithaca, NY, USA;* [3]*Department of Environmental Science, Kangwon National University, Kangwon-Do, Korea and* [4]*Marine Science Center, Northeastern University, Nahant, MA, USA*

The rRNA approach is the principal tool to study microbial diversity, but it has important biases. These include polymerase chain reaction (PCR) primers bias, and relative inefficiency of DNA extraction techniques. Such sources of potential undersampling of microbial diversity are well known, but the scale of the undersampling has not been quantified. Using a marine tidal flat bacterial community as a model, we show that even with unlimited sampling and sequencing effort, a single combination of PCR primers/DNA extraction technique enables theoretical recovery of only half of the richness recoverable with three such combinations. This shows that different combinations of PCR primers/DNA extraction techniques recover in principle different species, as well as higher taxa. The majority of earlier estimates of microbial richness seem to be underestimates. The combined use of multiple PCR primer sets, multiple DNA extraction techniques, and deep community sequencing will minimize the biases and recover substantially more species than prior studies, but we caution that even this—yet to be used—approach may still leave an unknown number of species and higher taxa undetected.
The ISME Journal (2009) **3,** 1365–1373; doi:10.1038/ismej.2009.89; published online 20 August 2009
Subject Category: microbial ecology and functional diversity of natural habitats
Keywords: estimates of microbial richness; microbial diversity; PCR primers; PCR bias

Introduction

The rRNA approach (Olsen *et al.*, 1986) remains one of the most important tools to assess microbial diversity. This necessitates a thorough study of its potential biases. Several are widely recognized. Most importantly, polymerase chain reaction (PCR) primers seem to discriminate for and against certain sequences (Suzuki and Giovannoni, 1996; Polz and Cavanaugh 1998). Additionally, existing DNA extraction protocols are not 100% efficient (Stach *et al.*, 2001). It is well known that this skews the frequency distribution of rRNA gene species among the PCR products, in clone libraries, and in eventual sequence inventories (Tiedje, 1994; Suzuki and Giovannoni, 1996; Polz and Cavanaugh 1998; Acinas *et al.*, 2004; Caron *et al.*, 2004; Kurata *et al.*, 2004; Frey *et al.*, 2006; Sipos *et al.*, 2007). What is not known is whether an increase in sequencing efforts, such as that afforded by the 454 Life Sciences sequencing technology (Sogin *et al.*, 2006; Huber *et al.*, 2007) would theoretically ensure a recovery of all rare rRNA species. In other words, is it possible that certain rRNA sequences are not recoverable in principle using any (or all) PCR primer sets available? If so, the totality of microbial diversity would be inaccessible even with an unlimited sequencing effort, with the degree of undersampling likely unknown. Presently, this possibility remains hypothetical, partly because researchers have typically favored, in any given study, the use of a single PCR primer set aiming to sequence the resulting clone library as fully as possible, and as a result have rarely varied PCR primers in a single study. In addition, it is not clear whether statistical tools for the analysis of recovered vs unseen diversity were appropriate for the task (Hong *et al.*, 2006; Jeon *et al.*, 2006). Here, we address the possibility that each specific combination of PCR primers/DNA extraction technique recovers a specific fraction of target diversity, and cannot provide access to the rest of the diversity even with an increase in sequencing effort.

Correspondence: S Epstein, Department of Biology, Northeastern University, 360 Huntington Avenue, Boston, MA, USA.
E-mail: s.epstein@neu.edu
Received 26 February 2009; revised 24 June 2009; accepted 24 June 2009; published online 20 August 2009

Materials and methods

Sampling
The samples were collected from an intertidal sand flat in Massachusetts Bay, near the Marine Science Center of Northeastern University, Nahant, MA, USA. An undisturbed core of sediment 13 cm in

diameter and 15 cm deep was collected, thoroughly mixed, and 5-g subsamples stored at −80 °C until DNA extraction. One of the samples, used to establish clone library I (see below) was fully processed as part of an earlier study (Hong *et al.*, 2006).

DNA extraction and purification

Nucleic acids were extracted from three subsamples using two methods of genomic DNA extraction (Table 1). The DNA for clone libraries I and III were extracted by modification of Zhou *et al.*, 1996. Briefly, each 5-g sample was mixed with 13.5 ml of DNA extraction buffer (100 mM Tris–HCl (pH 8.0), 100 mM sodium EDTA (pH 8.0), 100 mM sodium phosphate (pH 8.0), 1.5 M NaCl, 1% w/v CTAB) and 100 μl of proteinase K (10 mg ml^{-1}) in Falcon tubes for 30 min at 37 °C. After incubation, 1.5 ml of 20% w/v sodium dodecyl sulfate was added, and heated in a 65 °C water bath for 2 h. The DNA was purified twice by extraction with an equal volume of chloroform–isoamyl alcohol (24:1) and was precipitated with 0.7 volume of isopropanol. The DNA for clone library II was extracted from a 0.5-g subsample using a Fast DNA SPIN kit (Bio 101, La Jolla, CA, USA) according to the manufacturer's instructions. The DNA was purified with the resin-based Wizard DNA cleanup system (Promega, Madison, WI, USA).

Cloning and sequencing

16S ribosomal DNA templates were amplified using two primer sets. The first primer set, 27F (5′-AGA GTT TGA TCC TGG CTC AG-3′) and 1492R (5′-GGT TAC CTT GTT ACG ACT T-3′ (Lane, 1991) was used in the clone libraries I and II. The other primer set 8F (5′-AGA GTT TGA TCC TGG-3′) and 1542R (5′-AAA GGA GGT GAT CCA-3′) (Buchholz-Cleven *et al.*, 1997) was applied for clone library III. Each PCR mixture (50 ml) contained 10 ng of DNA as a template, 10 pmol of each primer, 10 mmol of dNTP mixture, 2.5 U of *Taq* DNA polymerase (Promega), and the PCR buffer supplied with the enzyme. PCR were performed using the following conditions: an initial denaturing step at 95 °C for 5 min, followed by 30 cycles consisting of 95 °C for 1 min, 50 °C for 1 min, and 72 °C for 1 min, and a final elongation step at 72 °C for 10 min. PCR products were purified with Qiaquick PCR Purification Kit (Qiagen, Valencia, CA, USA) and cloned using TOPO TA cloning kit (Invitrogen, Carlsbad, CA, USA) according to the manufacturer's protocol. To check the correct length insert, colonies were screened by PCRs with M13F (5-GTA AAA CGA GGC CAG-3) and M13R (5-CAG GAA ACA GCT ATG AC-3) primers. The clones were sequenced at Seqwight DNA Technology Service (Houston, TX, USA). The usable sequence length varied between 500 and 800 nt; the number of clones sequenced is summarized in Table 1. We note that each of the three libraries I, II, and III were pooled clones from 2–4 smaller sublibraries, each made using the library-specific primer set and single DNA extract but different PCR products. For example, library II represents pulling together clones from libraries IIa, IIb, and IIc (184, 239, and 282 clones, respectively).

Phylogenetic analyses

The rRNA sequences were manually edited using Bioedit (version 5.0.7) and aligned by the CLUSTAL X (Thompson *et al.*, 1994). Potential chimeric sequences were detected using the Chimera Check Program available at the Ribosomal Database Project (RDP) II (Maidak *et al.*, 2001). Suspect sequences were eliminated from the database (7.0%, 4.7%, and 7.3% of the clones from libraries I, II, and III, respectively). The remaining sequences were grouped into OTUs based on 70%, 80%, 90%, 95%, 96%, 97%, 98%, and 99% rRNA gene sequence similarity levels. This grouping was achieved by first making all possible pair-wise sequence alignments by using CLUSTALW at default settings and calculating % sequence identities, followed by clustering the sequences into OTUs by using the unweighted pair group method with arithmetic mean (UPGMA) as implemented in the

Table 1 Number of OTUs registered in this study

DNA extraction approach	*Primer set*	*Clones sequenced*	*OTUs defined as % 16S rRNA gene sequence identity*							
			99	*98*	*97*	*96*	*95*	*90*	*80*	*70*
Clone library I										
Zhou *et al.* (1996)	27F 1492R (Lane, 1991)	556	387	315	285	249	233	144	47	7
Clone library II										
Bio 101 kit	27F 1492R (Lane, 1991)	705	373	320	272	222	202	115	19	1
Clone library III										
Zhou *et al.* (1996)	8F, 1542R (Buchholz-Cleven *et al.*, 1997)	473	338	301	268	245	222	140	27	2

OC clustering program (http:// www.compbio.dundee.ac.uk/Software/OC/oc.html). The OTU grouping was checked manually to verify that all OTUs were assembled at the cut-off level desired. The number of OTUs and their frequencies at each cut-off value became the subject of statistical analyses.

Statistical analysis

In this study we focused on discriminating between the three experimental protocols determined by the three combinations of DNA extraction techniques and PCR primers, vs the pooled data. Broadly speaking, then, we had four 'treatments' (three experimental protocols plus the pooled data), and we wished to compare the total recoverable richness for each of the treatments. However, we also had eight sequence similarity levels for OTU definition (70%, 80%, 90%, 95%, 96%, 97%, 98%, and 99% 16S rRNA gene sequence identity), so to obtain maximum precision we compared the four treatments at all eight similarity levels simultaneously, using a statistical linear regression model.

To apply the regression model we first had to estimate total richness for each dataset separately. We had $4 \times 8 = 32$ datasets, each consisting of the numbers of OTUs occurring exactly once, twice, three times, and so on, for a given treatment at a given % similarity level (this is known as 'frequency count' data). We estimated the total OTU richness (observed + unobserved) based on each dataset, so that our analyses are based on estimates of total richness and are in principle independent of the particular sample sizes (numbers of available sequences) used here. (For small sample sizes some bias may be present in the estimates of total richness; see below for a sensitivity analysis of this issue.) There are two main families of methods for richness estimation: parametric and nonparametric. We regard the former as more reliable for high-diversity data as encountered here (for a detailed comparison see (Hong *et al.*, 2006)), but we used both methods in parallel throughout. For a given sample, the parametric method essentially fits a parametric curve to the observed frequency-count data, and projects this curve to obtain the number of unobserved OTUs (and associated statistics such as standard errors, goodness-of-fit assessments, etc.). For example, Figure 1 shows the results for the pooled data at 97% similarity: the projected number of unobserved OTUs is 3402 and the estimated total richness is 4372 (s.e. 433). The selected model in this case is based on a mixture of two exponential abundance distributions, and the model fit is excellent. (The corresponding nonparametric richness estimate, based on Chao's ACE1 (Chao, 2005) is 4664 (s.e. 964)).

We obtained total richness estimates for each of the 32 datasets. Given an estimate of the total richness at every % sequence similarity level for each of the four treatments, we then fit a statistical linear regression (straight line) model for the total number of taxa vs % sequence similarity, within each treatment, and compared the resulting straight lines across treatments. In the final summary analysis we compared the total recoverable richness for each of the experimental protocols with the pooled data. As our final results are based on the linear regression model, we discuss that model and its implications in 'Results' below.

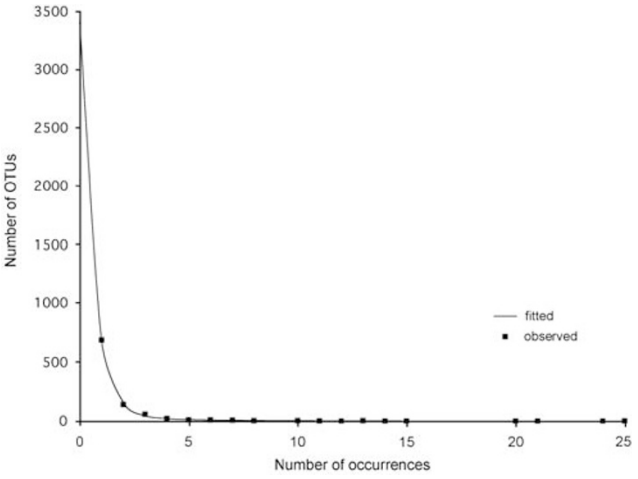

Figure 1 Fit of mixture-of-two-exponentials OTU abundance model (curve) to observed frequency-count data (points), with projection to zero frequency (unobserved OTUs), for pooled data at 97% gene sequence similarity level.

Results

Microbial inventory

Subsampling a single, thoroughly mixed marine sediment sample, we constructed three 16S rRNA gene clone libraries I, II, and III using three combinations of PCR primers and DNA extraction techniques (Table 1). Library I was reported earlier and used as a model to develop statistical approaches to estimate microbial richness (Hong *et al.*, 2006). Here, sequences from this library were pooled with original sequences from libraries II and III to create a unified list of the sample's OTUs (Table 1). Note that each library consists of 2–4 smaller libraries created from independent PCR reactions conducted using same PCR primers and same DNA extract. These smaller libraries are thus replicates of each method used. For example (as noted in Materials and methods), library II represents pulling together clones from libraries IIa, IIb, and IIc (184, 239, and 282 clones, respectively).

The combined inventory represented a diverse bacterial collection of species from 22 bacterial phyla, typical of rich marine sediment communities (Table 2). The inventory was dominated by representatives of Gammaproteobacteria, Bacteroidetes, and Deltaproteobacteria. Planctomycetes, Verrucomicrobia, Acidobacteria, Actinobacteria, Chloroflexi, and Alpha- and Epsilonproteobacteria were

Table 2 Number of rRNA gene sequences obtained in recent large scale surveys and sorted by bacterial phyla

Name of primer set	*27F, 1492R*					*27F or 515F, 1391R*[a]		*8F, 1542R*	*16S 3F, 16S 4R*	*27F, 1378R*	*27F, 907R*	*515F 1492R*	*515F, 1391R*
Reference	*This study; clone library I*	*This study; clone library II*	[b]	Gonzalez *et al.*(2000); Buchan *et al.*(2003); Mou *et al.* (2005)	Mesbah *et al.* (2007)	Walker and Pace (2007)	Ley *et al.* (2006)	*This study; clone library III*	[c]	*Hartmann and Widmer (2006)*	[d]	*Nemergut et al. (2007)*	*Papineau et al. (2005)*
Proteo-bacteria													
Alpha	**31**	**41**	212	52	140	83	284	47	7	295	3	26	+
Beta	1	1	303	40	–			12	27	75	13	109	–
Gamma	139	165	101	626	78			60	21	53	1	18	+
Delta	65	84	77	2	26			30	8	150	103	3	+
Epsilon	11	4	–	4	–			69	3		–	–	–
Unclear	26	2	98	21	–			**2**	2	–	–	–	–
Bacteroidetes	**133**	**205**	124	56	126	43	282	**2**	14	50	1	30	+
Chlorobi	**11**	**4**	–	–	–	1	–	–	–	–	–	–	–
Fibrobacteres	**2**	–	5	–	–	–	–	–	–	–	–	–	–
Gemmatimonadetes	**2**	–	48	–	–	4	–	–	–	34	–	11	+
Planctomycetes	**14**	**32**	124	–	–	2	68	**10**	–	11	2	8	+
Verrucomicrobia	**28**	**35**	25	–	8	6	44	**15**	3	76	–	10	+
Acidobacteria	**18**	**20**	470	–	–	13		**99**	3	224	10	37	+
Cyanobacteria	**26**	**3**	3	1	17	190	38	**9**	–	8	1	107	–
Spirochaetes	**2**	–	–	–	20	–	67	–	9	–	–	–	–
Fusobacterium	**1**	–	–	–	–	–	–	**1**	–	–	–	–	–
Firmicutes	**3**	–	476	3	159	17	–	**1**	12	78	5	–	+
Actinobacteria	**11**	**65**	64	53	7	168	–	–	23	700	4	5	+
Chloroflexi	**11**	**10**	4	–	8	22	422	**32**	2	29	14	2	+
Aquificae	**1**	–	–	–	–	–	–	–	–	–	–	–	–
Nitrospirae	–	–	6	–	–	–	–	–	4	14	–	–	–
Deinococcus-Thermus	–	–	–	–	–	9	–	–	–	–	1	6	–
Candidate division OD1	**4**	–	–	–	–	–	43	–	–	–	–	–	–
Candidate division OP1	**1**	–	–	–	–	-	–	–	–	–	–	–	+
Candidate division OP10	–	–	–	–	–	5	–	–	–	–	–	2	–
Candidate division OP3	**1**	–	–	–	–	–	–	–	–	–	–	–	–
Candidate division OP8	–	**5**	–	–	–	–	–	**3**	–	–	–	–	–
Candidate division SR1	**2**	–	–	–	–	–	–	–	–	–	–	–	–
Candidate division TM6	**1**	–	–	–	–	–	–	–	–	–	–	–	–
Candidate division KSB1	–	–	–	–	–	–	35	–	–	–	–	–	–
Candidate division WS3	**3**	–	–	–	–	–	–	**14**	–	–	–	–	–
Candidate division WS6	–	–	–	–	–	–	54	–	–	–	–	–	–
Candidate division WYO	–	–	–	–	–	–	–	–	–	–	–	10	–
Unclear affiliation	**8**	**29**	53	58	11	–	247	**67**	165	107	14	–	–
Total clone number	**556**	**705**	2193	916	600	563	1584	**473**	303	1904	172	384	NA

[a]Two primer sets used; 27F, 515R, and 27F, 1391R.
[b]Unpublished; sequence accession numbers EF072901–EF075329 at NCBI.
[c]Unpublished; sequence accession numbers EF434255–EF445544 at NCBI.
[d]Unpublished; sequence accession numbers AB293247–AB293418 at NCBI.
Original data are in bold.

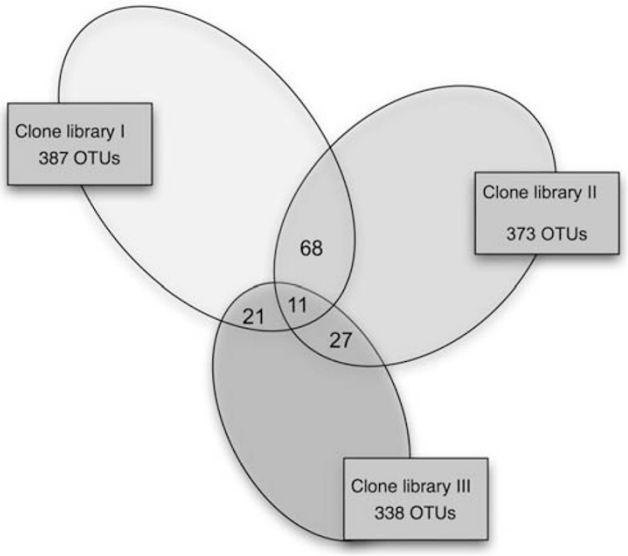

Figure 2 Overlap between OTU lists obtained from clone libraries I, II, and III; OTUs are defined as clusters of sequences sharing over 99% identity.

moderately abundant, comprising collectively 35% of all clones. Representatives of Sprirochaetes, Fusobacteria, Firmicutes, Chlorobi, Fibrobacteres, Gemmatimonadetes, Aquificae, and seven candidate phyla were rare, with no phylum being represented by more than 6% of all clones.

Several bacterial phyla, such as Alpha-, Gamma-, and Deltaproteobacteria, Planctomycetes, Verrucomicrobia, and Chloroflexi were represented equally well in the three clone libraries, whereas Bacteroidetes, Acidobacteria, Actinobacteria, and Epsilonproteobacteria exhibited a sharply asymmetrical distribution (Table 2). Differences between the libraries were more pronounced at lower levels of taxonomic identity. Species, defined as clusters of 16S rRNA gene sequences sharing over 97% of sequence identity, were predominantly unique to the respective library. As a result, the overlap between any 2 species lists did not exceed 17% (Figure 2), and only 11 species were shared between all three libraries.

Statistical analysis

Our goal was to compare the total taxonomic richness recoverable given unlimited sampling effort, under each of four 'treatments': the three experimental protocols plus the pooled data. This means that for each treatment we must estimate the total number of recoverable taxa, seen + unseen. However, estimates of the total number of taxa, or richness, are often accompanied by high standard errors, rendering comparisons across treatments statistically challenging. To overcome this difficulty we compared estimates of total recoverable richness across the four treatments, not just at one level of OTU definition, but at all eight measured levels

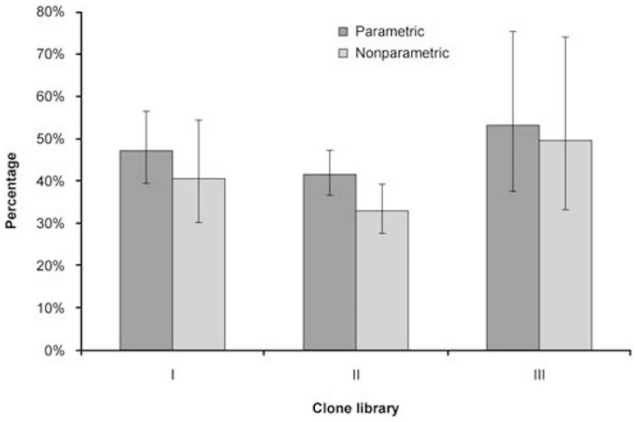

Figure 3 Estimated total OTU richness (points), and fitted parallel linear regressions (lines) for datasets I, II, III, and pooled, as a function of % sequence similarity. Arrow indicates estimated OTU richness based on pooled data at 97% gene sequence similarity level.

simultaneously (70%, 80%, 90%, 95%, 96%, 97%, 98%, and 99%). This is based on a result from Bunge *et al*. (2009), which states that the number of taxa increases exponentially as a function of % sequence identity, or equivalently that the logarithm of the number of taxa increases linearly as a function of % sequence identity; that is,

$$\log(\text{total } \textit{richness}) \approx \text{constant} + (\text{slope } \textit{coeffient}) * (\%\,\text{similarity } \textit{cutoff}) \quad (1)$$

For each of the four treatments we have eight data points ($x =$ % similarity cut-off, $y =$ estimated total OTU richness), and we fit a linear regression model (straight line function) with one line for each of the treatments, with parallel (equal) slopes but different intercepts (elevations). (Formal statistical hypothesis tests found no difference between the slopes.) The results are shown in Figure 3. Overall the fit is excellent, with $R^2 = 98.1\%$, 94.2%, 98.8%, and 95.9% for the pooled, I, II, and III datasets, respectively.

We note that the points within a given treatment are actually repeated measures, because they result from re-clustering the same collection of sequences at different % similarity levels. This implies that the points are correlated, and the correlation between successive points can be seen in the curved behavior of the points at the upper end of the % similarity scale, for some of the samples. The appropriate statistical model for this is regression with first-order autoregressive (1) errors, and we tested this model as well, but the data points here are too sparse and irregularly placed to obtain a precise fit, and the autoregressive (1) results were nearly identical to the standard regression model, so we do not report them here. We refer the reader to Bunge *et al*. (2009) for details on the full model.

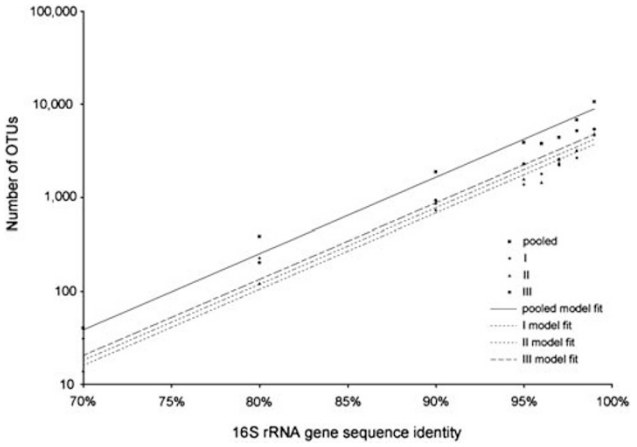

Figure 4 Estimated richness recoverable using experimental protocols I, II, and III, expressed as % of richness recoverable by pooling data from all three protocols (with associated 95% confidence bounds), based on parametric and nonparametric estimates of total richness.

The differences between the four treatments are evident in the displacements of the lines in Figure 3, and are quantified by the differences in the intercepts (elevations) of the lines. We estimated these differences using regression analysis, converted back from the log-scale to the original scale, and calculated confidence intervals using the Bonferroni correction for multiple comparisons. Finally, we used the results of the above analyses to compare each of the experimental protocols to the pooled data. The confidence intervals displayed in Figure 4 represent plausible ranges for the total OTU richness in principle recoverable by each of the three protocols, expressed as a percentage of the total richness recoverable by pooling all three protocols. Thus, for example, we estimate that protocol I can recover between 39.5% and 56.5% of the total richness recoverable by the pooled approach. Similar results were obtained by using the nonparametric estimates of total richness, shown in Figure 4. These results are somewhat lower, which may be because of the typical downward bias of nonparametric estimates in highly diverse situations; as noted above we regard the parametric results as more reliable in this case. Note that the most optimistic assessment of a given protocol relative to pooled (by considering the upper confidence bound), is that protocol III might be able to recover 75% of the diversity recoverable by the pooled approach, whereas the other two are considerably lower.

We note that though the parametric estimators of total richness are unbiased as the sample size increases to infinity, they are biased for small samples, to a degree that depends on the population structure and other factors. It is therefore reasonable to ask whether the differences between the three samples and the pooled data detected by our study could be the result of small-sample-size bias (or some other artifacts). To examine this we carried out two subsidiary studies. First, we subdivided original sample II, with 705 sequences, into three subsamples, IIa, IIb, and IIc, with 184, 239, and 282 sequences respectively. As sample II was homogeneously mixed, each of the subsamples should in principle yield comparable and roughly accurate estimates of total richness (though with larger standard errors than the complete sample II), except for small-sample-size bias or other artifacts. However, we found that these subsample sizes were generally too small to distinguish possible biases. For example, at the 97% similarity level, the complete sample II yielded a parametric total richness estimate of 2336 (s.e. 604), whereas subsamples IIa, IIb, and IIc yielded estimates 2207 (s.e. 1800), 3633 (s.e. 21680), and 2234 (s.e. 2550), respectively. At other % similarity levels some small-sample-size bias may have been present, but generally speaking the behavior of the estimates at these sample sizes was too erratic to be sure. We therefore carried out a second, simulation study, attempting to replicate homogenous mixing of the original pooled sample followed by homogeneous subsampling down to the sizes of the original three samples I, II, and III. Specifically, for the 97% similarity level, we postulated that the population had the total richness estimated from the original pooled sample, 4372 OTUs, and that the population followed the abundance structure of the parametric model fitted to the original pooled sample (a mixture of two exponential distributions). On the basis of this model, we simulated samples of the same sizes as the original samples I, II, and III, namely 556, 705, and 473 sequences, and estimated the total richness from each sample. Again, in principle, except for the small-sample-size bias (or other artifacts), the simulated samples should produce comparable and roughly accurate estimates of total richness (though with expectedly larger standard errors). We generated 10 simulation replicates from the above scenario, finding that across the 10 replicates, the estimates from the three simulated samples averaged 5595 ± 3291 (s.d.), 5212 ± 1483 (s.d.), and 4446 ± 2416 (s.d.) species. Thus, the results were as expected: for sample sizes 556, 705, and 473, as opposed to 184, 239, and 282, the estimates of total richness were roughly accurate but with higher standard errors than for the pooled sample (1734 sequences). We therefore conclude that, though the sizes of the subsamples IIa, IIb, and IIc were small enough so that total richness estimates could be affected to some extent by small-sample-size bias or other artifacts, the sizes of the original samples I, II, and III were large enough so that small-sample-size bias does not have a significant function, and the effects observed in the study can be attributed to the difference in the experimental treatments not the sample sizes.

Discussion

Microbial richness of environmental samples remains an elusive parameter, principally for two reasons. First, the totality of this richness seems to be very large (Tiedje, 1994; Curtis *et al.*, 2002; Hong *et al.*, 2006; Sogin *et al.*, 2006), whereas the samples of this diversity (i.e. clone libraries) researchers use to reconstruct this richness are typically small. Second, these samples (rRNA gene sequence inventories) (Muyzer and Smalla, 1998) are skewed. The first limitation may well be satisfactorily addressed by the progress in DNA sequencing technology, which allows an unprecedented increase in the size of microbial inventories (Sogin *et al.*, 2006; Huber *et al.*, 2007; Roesch *et al.*, 2007). It seems that the second limitation will determine how representative of the target communities these inventories will be. This is largely dependent on how incomplete the DNA extraction from the sample is, and on biases in the target gene amplification. Earlier studies have addressed both factors, showed their overall importance, and identified the principle source of amplification bias, such as PCR primer discrimination against certain templates and/or inhibition of the more abundant templates' amplification by self-annealing (Suzuki and Giovannoni, 1996; Polz and cavanaugh, 1998; Webster *et al.*, 2003; Acinas *et al.*, 2004; Kurata *et al.*, 2004; Frey *et al.*, 2006; Sipos *et al.*, 2007). This produces a qualitative picture of the overall importance of the PCR biases. The next logical question is does primer discrimination against certain species, coupled with biases of the DNA extraction method used to obtain the template, merely decrease the probability of their detection, or does it practically eliminate the possibility of such detection? In the first scenario, complete inventories are still possible given a sufficiently large sequencing effort; in the second, even an unlimited sequencing effort will not detect all the species in the community. This study attempts to discriminate between these two possibilities. (We note that metagenomics approaches are free from these biases (Venter *et al.*, 2004; Roesch *et al.*, 2007). However, the scale of such studies currently does not allow a complete recovery of all ribotypes present in environmental samples. Therefore, the potential of such approaches to recover the complete—and unbiased—picture of microbial community richness cannot be presently exploited. It follows that, in the foreseeable future, the rRNA approach, despite its biases, is likely to remain an important tool to recover and study microbial diversity in the environment.)

Our experimental approach is simple: using a single sample, obtain DNA using different extraction techniques, amplify the 16S rRNA gene using different primer sets, clone, and sequence the amplicons from several libraries, construct species (OTU) lists, and statistically predict the total microbial richness (observed + unobserved) as it appears from the individual libraries, as well as from the pooled data. Our null hypothesis is that the latter equals the former, indicating that the pooled data are no more than a result of an increase in sequencing efforts.

We note that this hypothesis must hold true for data obtained by replicating the same method (multiple libraries made using the same experimental protocols and applied to the same DNA extract). We checked this in two ways, first by considering three individual clone libraries (IIa–c) that collectively comprised library II (and are actual empirical replicates), and second by simulation. Modeling the species distribution in libraries IIa–c proved challenging: the resulting total richness estimates had very high standard errors in many cases, rendering comparisons imprecise and the hypothesis difficult to test. This is likely so because the sample sizes of libraries IIa–c are too small to reliably model the diversity represented by the complete library II. Nonetheless, for OTUs sharing 97% sequence identity, the libraries IIa–c all predicted roughly the same total richness as the complete library II (see Statistical Analysis, above). Second, to minimize the potential small-sample-size bias of libraries IIa–c, we proceeded by simulation. We randomly generated three homogeneous samples ('simulated clone libraries') equal in size to the empirically obtained libraries I, II, and III produced from the hypothesized underlying total population. We then compared the predictions of total richness based on the simulated samples to the hypothesized total diversity. These predictions were indeed comparable to one another and roughly accurate, but accompanied by larger standard errors because of the sample sizes. From this we conclude that any significant difference between treatments (libraries I, II, and III) would be real, and not because of small-sample-size bias. We then proceeded with evaluation of such differences.

We assessed the performance of individual primer sets/DNA extraction techniques by considering the number of recovered and recoverable OTUs at the species level, defined here as clusters of rRNA gene sequences sharing over 97% identity. The total richness of library I has been estimated in our earlier work to be 2400 species (Hong *et al.*, 2006). Libraries II and III, obtained with other combinations of PCR primers and DNA extraction techniques, seem to be equally rich (Figure 3). This suggests that, no matter what specific method is used to produce a clone library, sequencing this library to saturation will result in the same number of species, between 2300 and 2500. However, when we estimate the number of species in the pooled libraries, the predicted richness is almost twice as big (Figure 3). This can mean only one thing: a specific method recovers specific species, and the given sequencing effort using one method is not equivalent to using two methods with half the

sequencing effort each. Put differently, each specific combination of PCR primers/DNA extraction techniques can recover—even with unlimited sequencing—a specific subset of microbial species from the sample, which differs from other subsets recoverable with other methods. Each method used here seems to recover about 50% of the species recoverable from the combination of all the three methods, and this indicates that the null hypothesis should be rejected. We note that though the multiple PCR primer/multiple DNA extraction technique approach seems to recover a qualitatively richer set of species, this set is unlikely to be complete either, and it remains unknown what fraction this set constitutes in the true sample diversity.

Unexpectedly, we observe the same tendency at other, higher taxonomic levels. Bacterial OTUs based on 90% and 80% 16S rRNA gene sequence identities seem to be undersampled by using a single PCR primer/DNA extraction technique by just about as much as do species (Figure 3). It is widely understood that sequence-based OTUs cannot be easily translated into the units of traditional taxonomy, and no particular level of 16S rRNA gene identity stands for a specific classical taxon. Nonetheless, 16S gene sequence divergence of about 90% and 80% are believed to approximately correspond to differences between microbial family/classes and phyla, respectively (Hugenholtz *et al.*, 1998; Sait *et al.*, 2002; Schloss and Handelsman, 2004). If so, a typical rRNA survey seems to fail to detect at least half the classes and phyla present in the sample(s) studied. This is a significant bias, and we searched for evidence pointing in this direction in the past rRNA surveys deposited into the GenBank (Table 2; we considered 10 larger studies reporting over 150 sequences per investigation (including this study)). This examination showed a surprising fact: with a single exception of the 27F/1492R primer set (Lane, 1991) used by many researchers, the majority of the prior studies used study-specific PCR primer sets. The use of such primers sets has not been replicated by other researchers in other larger-scale surveys, and this makes it difficult to detect such sets' potential class- and phylum-level biases. Nonetheless, it is striking to see that several known and candidate phyla, .for example Chlorobi, Fibrobacteres, KSB1, WS6, OD1, OP1, have been detected repeatedly with one (27F/1492R) primer set, but not with any other. Similarly, Nitrosparae, OP8, OP10, WS3, and WYO have all been detected in studies using just 1–2 specific primer sets, whereas using other sets did not result in their detection. This apparent selectivity is consistent with our finding of a substantial discrimination caused by the DNA extraction techniques/PCR primers used here (Figures 1–4). We note that species composition and frequencies are different between different environments, and the degree of the discovered discrimination may vary depending on the environment (marine sediments, soils, etc.).

Conclusions

We showed that rRNA environmental gene surveys typically using one specific technique to extract community genomic DNA, and one specific PCR primer set to amplify its 16S rRNA genes, miss a significant amount—around 50%—of microbial diversity, from species to phyla. Simultaneous use of three such combinations, in combination with deep sequencing (for example using 454 sequencing technology), could in principle double the recoverable diversity, but what fraction of the true sample diversity this would represent remains uncertain.

Acknowledgements

We thank Dr V Ilyin from Northeastern University for help in clustering the sequencing data. This study was supported by NSF Grants OCE-0221267, MCB-0348341, and DEB-0816840 to SSE, and DEB-0816638 to JB. We thank Linda Woodard for supervising the statistical computations. This research was conducted using the resources of the Cornell University Center for Advanced Computing, which receives funding from Cornell University, New York State, the National Science Foundation, and other leading public agencies, foundations, and corporations.

References

Acinas SG, Klepac-Ceraj V, Hunt DE, Pharino C, Ceraj I, Distel DL *et al.* (2004). Fine-scale phylogenetic architecture of a complex bacterial community. *Nature* **430**: 551–554.

Bunge J, Epstein SS, Woodard L. (2009). Modeling species richness as a function of DNA sequence similarity. Department of Statistical Science, Cornell University: Ithaca, NY. Technical Report no. 2009-1. 18pp.

Buchholz-Cleven B, Rattunde B, Straub K. (1997). Screening for genetic diversity of isolates of anaerobic Fe(II)-oxidizing bacteria using DGGE and whole-cell hybridization. *Syst Appl Microbiol* **20**: 301–309.

Caron DA, Countway PD, Brown MV. (2004). The growing contributions of molecular biology and immunology to protistan ecology: molecular signatures as ecological tools. *J Eukaryot Microbiol* **51**: 38–48.

Chao A. (2005). Species richness estimation. In: Balakrishnan C, Read B, Vidakovic B (eds). *Encyclopedia of Statistical Sciences*. Wiley: New York. pp 7907–7916.

Curtis TP, Sloan WT, Scannell JW. (2002). Estimating prokaryotic diversity and its limits. *Proc Natl Acad Sci USA* **99**: 10494–10499.

Frey JC, Angert ER, Pell AN. (2006). Assessment of biases associated with profiling simple, model communities using terminal-restriction fragment length polymorphism-based analyses. *J Microbiol Methods* **67**: 9–19.

Gonzalez JM, Simo R, Massana R, Covert JS, Casamayor EO, Pedros-Alio C *et al.* (2000). Bacterial community structure associated with a dimethylsulfoniopropionate-producing North Atlantic algal bloom. *Appl Environ Microbiol* **66**: 4237–4246.

Hartmann M, Widmer F. (2006). Community structure analyses are more sensitive to differences in soil bacterial communities than anonymous diversity indices. *Appl Environ Microbiol* **72**: 7804–7812.

Hong SH, Bunge J, Jeon SO, Epstein SS. (2006). Predicting microbial species richness. *Proc Natl Acad Sci USA* **103**: 117–122.

Huber JA, Mark Welch DB, Morrison HG, Huse SM, Neal PR, Butterfield DA *et al.* (2007). Microbial population structures in the deep marine biosphere. *Science* **318**: 97–100.

Hugenholtz P, Goebel BM, Pace NR. (1998). Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *J Bacteriol* **180**: 4765–4774.

Jeon SO, Bunge J, Stoeck T, Barger KJ, Hong SH, Epstein SS. (2006). Synthetic statistical approach reveals a high degree of richness of microbial eukaryotes in an anoxic water column. *Appl Environ Microbiol* **72**: 6578–6583.

Kurata S, Kanagawa T, Magariyama Y, Takatsu K, Yamada K, Yokomaku T *et al.* (2004). Reevaluation and reduction of a PCR bias caused by reannealing of templates. *Appl Environ Microbiol* **70**: 7545–7549.

Lane DJ. (1991). 16S/23S rRNA sequencing. In: Stackebrandt E, Goodfellow M (eds). *Nucleic Acid Techniques in Bacterial Systematics*. John Wiley and Sons: Chichester. pp 115–175.

Ley RE, Harris JK, Wilcox J, Spear JR, Miller SR, Bebout BM *et al.* (2006). Unexpected diversity and complexity of the Guerrero Negro hypersaline microbial mat. *Appl Environ Microbiol* **72**: 3685–3695.

Maidak BL, Cole JR, Lilburn TG, Parker CT, Saxman PR, Farris RJ *et al.* (2001). The RDP-II (Ribosomal Database Project). *Nucleic Acids Res* **29**: 173–174.

Mesbah NM, Abou-El-Ela SH, Wiegel J. (2007). Novel and unexpected prokaryotic diversity in water and sediments of the alkaline, hypersaline lakes of the Wadi An Natrun, Egypt. *Microbiol Ecol* **54**: 598–617.

Mou X, Moran MA, Stepanauskas R, Gonzalez JM, Hodson RE. (2005). Flow-cytometric cell sorting and subsequent molecular analyses for culture-independent identification of bacterioplankton involved in dimethylsulfoniopropionate transformations. *Appl Environ Microbiol* **71**: 1405–1416.

Muyzer G, Smalla K. (1998). Application of denaturing gradient gel electrophoresis (DGGE) and temperature gradient gel electrophoresis (TGGE) in microbial ecology. *Antonie Van Leeuwenhoek* **73**: 127–141.

Nemergut D, Anderson S, Cleveland C, Martin A, Miller A, Seimon A *et al.* (2007). Microbial community succession in an unvegetated, recently deglaciated soil. *Microbiol Ecol* **53**: 110–122.

Olsen GJ, Lane DJ, Giovannoni SJ, Pace NR, Stahl DA. (1986). Microbial ecology and evolution: a ribosomal RNA approach. *Annu Rev Microbiol* **40**: 337–365.

Polz MF, Cavanaugh CM. (1998). Bias in template-to-product ratios in multitemplate PCR. *Appl Environ Microbiol* **64**: 3724–3730.

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

Sait M, Hugenholtz P, Janssen PH. (2002). Cultivation of globally distributed soil bacteria from phylogenetic lineages previously only detected in cultivation-independent surveys. *Environ Microbiol* **4**: 654–666.

Schloss PD, Handelsman J. (2004). Status of the microbial census. *Microbiol Mol Biol Rev* **68**: 686–691.

Sipos R, Szekely AJ, Palatinszky M, Revesz S, Marialigeti K, Nikolausz M. (2007). Effect of primer mismatch, annealing temperature and PCR cycle number on 16S rRNA gene-targetting bacterial community analysis. *FEMS Microbiol Ecol* **60**: 341–350.

Sogin ML, Morrison HG, Huber JA, Mark Welch D, Huse SM, Neal PR *et al.* (2006). Microbial diversity in the deep sea and the underexplored 'rare biosphere'. *Proc Natl Acad Sci USA* **103**: 12115–12120.

Stach JE, Bathe S, Clapp JP, Burns RG. (2001). PCR-SSCP comparison of 16S rDNA sequence diversity in soil DNA obtained using different isolation and purification methods. *FEMS Microbiol Ecol* **36**: 139–151.

Suzuki MT, Giovannoni SJ. (1996). Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. *Appl Environ Microbiol* **62**: 625–630.

Thompson JD, Higgens DG, Gibson TJ. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties, and weight matrix choice. *Nucleic Acids Res* **33**: 4673–4680.

Tiedje JM. (1994). Microbial diversity: of value to whom? *ASM News* **60**: 524–525.

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.

Walker JJ, Pace NR. (2007). Phylogenetic composition of rocky mountain endolithic microbial ecosystems. *Appl Environ Microbiol* **73**: 3497–3504.

Webster G, Newberry CJ, Fry JC, Weightman AJ. (2003). Assessment of bacterial community structure in the deep sub-seafloor biosphere by 16S rDNA-based techniques: a cautionary tale. *J Microbiol Methods* **55**: 155–164.

Zhou J, Bruns MA, Tiedje JM. (1996). DNA recovery from soils of diverse composition. *Appl Environ Microbiol* **62**: 316–322.