Chapter 16

Extensive Phylogenetic Analysis of a Soil Bacterial Community Illustrates Extreme Taxon Evenness and the Effects of Amplicon Length, Degree of Coverage, and DNA Fractionation on Classification and Ecological Parameters

SERGIO E. MORALES, THEODORE F. COSART, JESSE V. JOHNSON, AND WILLIAM E. HOLBEN

16.1 INTRODUCTION

Intense interest in surveying bacterial species richness and diversity has moved microbial ecology heavily toward molecular-based approaches due to the inability to cultivate the vast majority of bacteria from natural environments [Rappe and Giovannoni, 2003]. Amongst the primary methods used to study bacterial diversity developed in the last two decades [Tiedje et al., 1999; Nocker et al., 2007], the majority have relied on PCR amplification followed by cloning and have targeted the 16S rRNA gene based on its utility for phylogenetic placement and comparisons. However, this approach has also posed some problems due to the high level of conservation within the 16S gene [Hanage et al., 2006], as well as other inherent limitations caused by the utilized approaches (e.g., PCR and cloning biases).

PCR amplification has been linked to biases caused by preferential amplification of certain templates in the mixture based on their primary sequence, percent G + C content (hereafter GC content), or other factors, resulting in so-called PCR bias [Polz and Cavanaugh, 1998]. A review of the pitfalls of PCR-mediated 16S rRNA gene analysis is presented by Stackebrandt [Chapter 12, Vol. 1]. Other factors, including template concentration, have been shown to skew detection toward the most abundant members of the community [Polz and Cavanaugh, 1998], a major concern due to the typical distribution of bacterial species in many environments, such as soil, where the tendency is toward having a few numerically predominant populations and a veritable slew of rare or unique ones [Sogin et al., 2006]. Since current PCR-based approaches predominantly recover numerically dominant species, this bias in detection might invalidate hypothesis testing on complex communities where limited sampling is performed [Brose et al., 2003].

Once PCR-amplified, most studies undertake a cloning-dependent approach prior to sequencing. Many
of the problems associated with cloning dependent strategies have been addressed or acknowledged [Rappé and Giovannoni, 2003; Holben et al., 2004; Schloss and Handelsman, 2004, 2005; Curtis, 2006], but others still linger. Until the recent development of cloning-independent sequencing, PCR amplification followed by random (shotgun) cloning of amplicon mixtures resulted in the detection of numerically dominant sequences, especially where relative abundance can vary over orders of magnitude as it often does in complex communities. This has been mostly dealt with by increasing the number of sequences obtained and analyzed (e.g., via high-throughput sequencing), but the need to develop methods to detect and characterize members of the so-called “rare biosphere” [Sogin et al., 2006] is still present. This is of special concern where limited sampling is performed on highly complex microbial communities exhibiting mostly even distribution of populations with only a few showing any degree of dominance, as typically observed for soils [Janssen, 2006]. Access to the bacterial diversity present in the “underrepresented minority” populations [Rappé and Giovannoni, 2003; Sogin et al., 2006] is needed, given their overwhelming contribution to the total diversity in many ecosystems. Furthermore, the role of numerically rare populations in important ecosystem functions (e.g., nitrification, certain infectious diseases) may be of special interest [Holben et al., 2004].

One approach to increase recovery of organisms in low abundance in the community involves prefractionation of total bacterial community genomic DNA based on its GC content (hereafter GC fractionation) prior to subsequent molecular manipulations of total community DNA [Holben and Harris, 1995]. The technique reduces total community complexity by binning total community DNA into fractions of similar GC content by physical separation into multiple fractions. Once fractionated, the total diversity in each bin or fraction is reduced compared to the unfractonated total community allowing for deeper sampling of minority populations [Holben et al., 2004].

Lastly, the final step in any current sequencing project targeting the 16S rRNA gene typically requires the processing and analysis of large numbers of sequences that must be catalogued into phylogenetically relevant groups. The creation of programs, pipelines, websites, and servers dedicated to such tasks has greatly aided the advancement of research in this area. Some of the most commonly used tools include the RDP Classifier and Sequence Match functions [Cole et al., 2009], Greengenes [DeSantis et al., 2006], and ARB [Ludwig et al., 2004]. Despite the tools currently available, the huge surge in available gene 16S sequences has flooded databases with massive numbers of unclassified or poorly annotated sequences that can lead to errors in lineage placements if included in phylogenetic analyses.

This review discusses the bacterial diversity uncovered within a single soil sample generated from GC fractionated DNA that was used to develop an extensive (5000 clones) partial 16S rRNA gene library using Sanger diode sequencing. It also compares the fractionated DNA library results to a library generated from the same soil community DNA sample using nonfractionated DNA. We go on to compare analyses of these sequences using multiple phylogenetic alignment tools and different amplicon and library sizes to gain a better understanding of the effects that each parameter has on the final interpretation of the composition and diversity of the total bacterial community in this soil sample.

16.2 MATERIALS AND METHODS

For a full outline of the materials and methods, please refer to the original publication [Morales et al., 2009]. What follows is an abbreviated version.

16.2.1 Study Site and Sample Collection

Samples were collected from the KBS LTER Row-Crop Agriculture site in mid-Michigan (for an overview of that project, see http://lter.kbs.msu.edu/). The current study examined the bacterial community in the replicate plots of Treatment 1 at the main experimental site, which is representative of canonical agricultural practice in the upper Midwest. The treatment consisted of conventional wheat, corn, and soybean annual rotations receiving standard levels of chemical inputs, with chisel plowing. Soil was classified as a fine-loamy, mixed, mesic Typic Hapludalf. For this bacterial population survey, five randomly positioned, 0- to 20-cm soil cores were taken from each of six treatment replicates in July 2004, at the height of the growing season, and composited to provide a single sample for that replicate. Each replicate treatment sample was sieved through 2-mm mesh and mixed thoroughly, providing six replicate samples for Treatment 1. All soil samples for this study were stored on dry ice or at −70 °C immediately after soil processing (i.e., sieving and mixing) prior to bacterial community DNA extraction.

16.2.2 DNA Manipulations

Total microbial community DNA was extracted and purified by direct lysis [Holben, 1997] and equal amounts of DNA (10 μg) from each replicate sample were pooled to provide a representative sample for this treatment regimen. The combined sample was subsequently fractionated based on percent GC content of the DNA of the
component populations of the community [Holben and Harris, 1995]. Fifteen separate fractions were obtained representing percent GC contents ranging from 20% to 80% (the full range observed in the domain Bacteria), and the amount and percent GC content of the DNA at each position in the gradient were determined as described elsewhere [Holben and Harris, 1995]. Each individual fraction was then PCR amplified independently for creation of the 16S rRNA gene clone library using primers 536f (5′-CAGCMGCCGCGGTAATWC-3′) and 907r (5′-CCGTCAATTCMTTTRAGTTT-3′). Complete reaction conditions are described fully in Morales and Holben [2009]. PCR products were cloned using the plasmid vector pT7Blue-3 and the Perfectly Blunt cloning kit (Novagen, Inc., Madison, WI) according to the manufacturer’s instructions. Plasmid DNA was sequenced by using the universal primer T7 and standard dideoxy sequencing conditions.

16.2.3 Phylogenetic Placement and Tree Creation Based on Clone Libraries

All 16S rRNA gene sequences were manually trimmed of vector and primer sequence, and they were screened for chimaeras using Pintail [Ashelford et al., 2005] prior to alignment and analysis. Multiple Fasta files were created and independently aligned in ARB [Ludwig et al., 2004] using the Fast Aligner and at least three reference sequences for each clone from the 16S rRNA gene database PT server containing 51,024 reference sequences (http://www.arb-home.de/downloads.html). Sequences from the current study were integrated into the annotated tree based on parsimony.

16.2.4 Assignment to Similarity-Based OTUs and Species Richness Estimators

ARB-generated 16S sequence alignments were used to create Jukes–Cantor corrected distance matrices that were used as input for the DOTUR program [Schloss and Handelsman, 2005], which calculated Simpson’s and Shannon–Weaver diversity indices, Chao1 richness estimates, and OTU bins using default settings. Comparison of GC-fractionated to nonfractionated data was performed by creating a master sequence library containing both fractionated and nonfractionated sequence libraries. Approximately 500 (487 and 490, respectively) sequences were compared for fractionated and nonfractionated libraries by comparing 33 sequences obtained from each of the 15 GC-based fractions of the total community to a library of 490 sequences randomly cloned from nonfractionated total community DNA from the same sample. The sequences obtained were aligned in ARB and then processed with the DOTUR program. DOTUR data files were then used as input for the SONS program [Schloss and Handelsman, 2006b], which was used to compare OTU representation within each library.

16.2.5 Identification of Phylum-Specific Taxonomic Bins and OTU Composition

Distance score cutoff values for individual phyla were calculated using the DAM (DOTUR-ARB Matching) program [Morales et al., 2008], available at http://dbs. umt.edu/research_labs/holbenlab/links.php. This allowed comparison between ARB-generated group lists and DOTUR list files created from the total dataset of 4889 sequences.

Sequence representation within specific OTUs was identified by means of the DOTMAN (for “DOTUR manipulation” program, available at http://dbs.umt.edu/research_labs/holbenlab/links.php) [Morales et al., 2008]. DOTMAN queries selected OTUs (based on DOTUR bins) against a sequence database, generating FASTA files from a user-given file.

16.2.6 Sample Size Simulations

To explore the effects of sampling size on ecological parameters (Chao1 richness estimation, Shannon–Weaver indices, and dominance), we used EcoSim700 null model software for ecology (version 7.0) to analyze data created from the first 500, 2000, 3390, and 5000 sequences contained in our library. Input files were created from OTUs that clustered with 97% similarity and were subsequently used as the data matrix for running the program.

16.2.7 Nucleotide Sequence Accession Numbers

All sequences used in this chapter have been deposited in the GenBank database (accession no. EU352912 to EU357802).

16.3 RESULTS AND DISCUSSION

16.3.1 Effect of Sample Size on Observed and Estimated Richness

Environmental rRNA gene clone libraries vary considerably in size and until very recent have typically contained 500 sequences or less [Narang and Dunbar, 2004; Ashelford et al., 2006]. This has changed with the recent advent of high-throughput sequencing
technologies, with early studies using such approaches generating tens or hundreds of thousands of sequences [Sogin et al., 2006; Roesch et al., 2007]. More recently, studies have foregone depth of sequencing for the increased comparative capabilities [Lauber et al., 2009] afforded by barcoded sequencing [Hamady et al., 2008]. This trend has moved studies looking at 16S rDNA gene diversity back to numbers akin to those generated from traditional Sanger sequencing and cloning dependent strategies, with the depth of sequencing ranging from 1000 to 2000 sequences.

Increased depth of sequencing likely provides a more accurate view of the dominant phylotypes in an ecosystem [Schloss and Handelsman, 2006a] but still misses a large proportion of the diversity in the system, a problematic option given our lack of knowledge regarding the role of this “underrepresented majority” of low-abundance populations. The study described herein (and in Morales et al. [2009]) provides an in-depth analyses demonstrating the effects of sequence sample size on ecological parameters, such as richness estimation, dominance, diversity indices, and evenness. The effect of sample size was tested by creating datasets from the first 500, 2000, 3390, and 5000 sequenced clones from a GC-fractionated library with subsequent removal of anomalous and nonbacterial sequences producing sets of 487, 1962, 3322, and 4889 sequences, respectively. These datasets were analyzed based on “bins” created as a function of 16S sequence similarity at multiple levels of similarity (Table 16.1 and Fig. 16.1), but discussion in this report is focused primarily on the widely utilized 97% sequence similarity level.

A 5.1-fold increase in the number of OTUs and a 3.5-fold increase in the richness estimation were observed (at 97% sequence similarity) between the smallest and the largest dataset (Table 16.1 and Fig. 16.1). Analysis of the full dataset (4889 sequences) indicated a total of 1714 OTUs at 97% sequence similarity (Table 16.1), with a predicted total diversity of 3555 OTUs. A similar pattern was recently reported by Roesch et al. [2007] when a single soil sample was analyzed using pyrosequencing and library sizes ranging from 10,000 to 53,632 sequences were compared. Despite an apparent asymptote seemingly being reached with each library size, additional sequencing always resulted in higher OTU detection and subsequent higher estimated richness.

Shannon diversity index values increased approximately 1.2-fold across this same span of sequence library sizes, with the Simpson’s reciprocal index (1/D) exhibiting a 1.7-fold increase. At 97% sequence similarity, the Shannon–Weaver score for the 4889 sequence library was calculated to be 6.75, much greater than the values of 4.35 and 4.68 previously estimated for an Amazon and a Scottish soil, respectively [Schloss and Handelsman, 2005]. Our value is also higher than what has been reported for other ecosystems, including surface ocean waters [Biers et al., 2009], the human gut [Andersson et al., 2008], and glacial ice [Simon et al., 2009]. Our value is also higher than what has been reported for other ecosystems, including surface ocean waters [Biers et al., 2009], the human gut [Andersson et al., 2008], and glacial ice [Simon et al., 2009].

Our study also supported previous findings indicating a “rare biosphere” [Sogin et al., 2006] and the increasingly common perception that numerous taxa present in comparably low overall abundance comprise the bulk of the soil bacterial community [Schloss and Handelsman, 2006a]. Although individually underrepresented in our library (determined <3 times), collectively these rarely identified sequences accounted for 82% of all identified
Table 16.1: Effect of Sample Size on Similarity-Based OTUs, Shannon–Weaver Diversity Index, Evenness, and Richness Estimation

| Sequence Sample Size, % Similarity Level | Number of OTUs | Shannon–Weaver Index | Evenness | Unique Estimate<sup>a</sup> | OTU1 | OTU2 | OTU3 | OTU4 | OTU5 | OTU6 | OTU7 | OTU8 | OTU9 | OTU10 |
|------------------------------------------|----------------|----------------------|----------|-----------------------------|------|------|------|------|------|------|------|------|------|------|------|
| First500<sup>b</sup>                     | 100            | 461                  | 6.10     | 0.995                       | 5.851| 6    | 3    | 3    | 2    | 2    | 2    | 2    | 2    | 2    |
|                                           | 97             | 335                  | 5.62     | 0.966                       | 1.020| 12   | 8    | 7    | 7    | 7    | 5    | 5    | 5    | 5    |
|                                           | 70             | 30                   | 2.94     | 0.866                       | 0.32 | 59   | 45   | 43   | 41   | 40   | 31   | 27   | 24   | 24   |
|                                           | 55             | 6                    | 0.78     | 0.434                       | 6    | 379  | 51   | 43   | 6    | 5    | 3    |      |      |      |
|                                           | 47             | 2                    | 0.13     | 0.188                       | 2    | 473  | 14   |      |      |      |      |      |      |      |
| First2000<sup>c</sup>                    | 100            | 1,682                | 7.33     | 0.986                       | 12.163| 22   | 9    | 8    | 8    | 7    | 6    | 5    | 5    | 5    |
|                                           | 97             | 928                  | 6.39     | 0.935                       | 2.126| 44   | 33   | 24   | 24   | 22   | 19   | 18   | 16   | 12   |
|                                           | 70             | 67                   | 3.34     | 0.794                       | 0.80 | 174  | 156  | 139  | 134  | 132  | 110  | 100  | 97   | 96   | 91   |
|                                           | 55             | 14                   | 1.69     | 0.641                       | 0.20 | 608  | 597  | 297  | 224  | 91   | 82   | 29   | 12   | 10   |      |
|                                           | 38             | 6                    | 0.04     | 0.025                       | 0.12 | 1,950| 8    | 1    | 1    | 1    |      |      |      |      |      |
| First3390<sup>d</sup>                    | 100            | 2,680                | 7.73     | 0.980                       | 15.015| 40   | 17   | 15   | 15   | 11   | 10   | 8    | 8    | 8    | 7    |
|                                           | 97             | 1,319                | 6.59     | 0.918                       | 2.991| 76   | 54   | 54   | 41   | 38   | 31   | 27   | 27   | 27   | 26   |
|                                           | 70             | 84                   | 3.48     | 0.785                       | 0.88 | 309  | 243  | 237  | 210  | 182  | 165  | 159  | 150  | 147  |      |
|                                           | 55             | 14                   | 1.62     | 0.615                       | 0.14 | 1,284| 1,064| 387  | 205  | 85   | 85   | 60   | 54   | 46   | 28   |
|                                           | 38             | 2                    | 0.02     | 0.032                       | 0.02 | 3,311| 11   |      |      |      |      |      |      |      |      |
| First5000<sup>e</sup>                    | 100            | 3,789                | 8.04     | 0.976                       | 20.790| 54   | 25   | 22   | 17   | 15   | 14   | 13   | 12   | 12   | 10   |
|                                           | 97             | 1,714                | 6.75     | 0.906                       | 3.555| 99   | 81   | 81   | 63   | 62   | 61   | 46   | 39   | 38   | 38   |
|                                           | 70             | 102                  | 3.60     | 0.778                       | 0.119| 474  | 345  | 297  | 256  | 248  | 236  | 233  | 215  | 211  | 210  |
|                                           | 55             | 18                   | 1.98     | 0.685                       | 0.20 | 1,402| 1,026| 729  | 504  | 453  | 231  | 221  | 175  | 58   | 37   |
|                                           | 38             | 5                    | 0.03     | 0.017                       | 0.05 | 4,873| 13   | 2    | 2    | 1    |      |      |      |      |      |

<sup>a</sup> Based on full biased corrected Chao1 richness estimates.
<sup>b</sup> Based on 487 starting sequences.
<sup>c</sup> Based on 1962 starting sequences.
<sup>d</sup> Based on 3322 starting sequences.
<sup>e</sup> Based on 4887 starting sequences and 2 archeal sequences used as references.
OTUs in this soil system, resulting in an extremely high evenness estimate of 0.906 (Table 16.1).

16.3.2 Community Composition
Proper assignment of obtained sequences into a phylogenetic context is crucial to understanding the diversity and ultimately the role of individual taxa in any ecosystem. Many tools are available for this task, but comparative assessment of the coherence of taxonomic placements achieved via multiple methods is lacking in most studies due to the challenge of analyzing increasingly large and complex datasets. Here, we compare various commonly used methods of taxonomic placement for 16S rRNA gene sequence data. Sequences were first analyzed using the Classifier (version 1.0; taxonomical hierarchy release 6.0) and the SeqMatch tools [Cole et al., 2007] of the RDP II. Individual sequences were considered classified only if both programs showed agreement at the phylum level. Unclassified sequences based on this criterion were subsequently assigned a potential placement based on Classifier. Using this method, 3233 (66%) of the sequences in our library were classified into 17 recognized phyla (Table 16.2). The entire set of 4889 sequences was also classified using ARB [Ludwig et al., 2004] by placement into an ARB-generated phylogenetic tree of 51,024 classified sequences. With this approach, a 33% increase in placement to known phyla was obtained, with 4854 (99%) of the sequences assigned to 25 recognized phylogenetic groups. It is worth noting that the classification of certain groups was comparable using both methods (Table 16.2); but in groups with low sequence representation within databases (e.g., refer to Chlorobi, Acidobacteria, Fibrobacteres, and candidate divisions of Table 16.2), the ARB-based approach allowed for more-consistent assignment of bacteria at the phylum level. We also note that since the analysis described herein was performed, a new release of the Classifier tool (version 2.0; taxonomical hierarchy release 7.8) has been made available [Wang et al., 2007]. Reanalysis of our dataset with this new release prior to this writing produced taxonomic placements nearly identical to those obtained with ARB for classified sequences. However, we note that the new version of Classifier was still unable to classify 1013 (21%) of the sequences in our library compared to the 1% unclassified by ARB.

We also compared classification and other results from two additional methods, Clustal W alignment following Thompson et al. [1994] and classification using MUSCLE software [Edgar, 2004], which produced datasets with similar numbers of OTUs, Chao1 richness estimates, and other diversity parameters. However, phylogenetic trees generated from the results of these approaches did not produce clustering coherent with phylogenetic assignments obtained using RDP tools (not shown). In contrast, phylogenetic trees generated using ARB alignments were reproducible and provided consistent phylogenetic placement to that obtained with the RDP toolset.

16.3.3 Community Structure and Taxonomic Distribution
Based on an ARB tree-based classification approach, all of the sequences in the study were assigned into 25 recognized phyla, with just 35 of the 4889 sequences not assignable to any known phylum or group (Table 16.2). The most predominant phylum in this Midwestern agricultural soil was the Proteobacteria, which comprised 35% of the sequence library, followed by the Acidobacteria with 26% of the total. Six other phyla, including Actinobacteria, Bacteroidetes, Thermomicrobia, Gemmatimonadetes, Planctomycetes, and Verrucomicrobia, averaged 7% representation. The remaining 17 phyla were represented by phylotype numbers totaling <2% of the total library.

The presence of a “rare biosphere” was first presumed by Sogin et al. [2006], but since then the presence of numerically underrepresented members of the bacterial community has become a common observation in complex environments [Bent and Forney, 2008; Elshahed et al., 2008; Fuhrman, 2009]. For the soil sample analyzed in our study, 1405 OTUs (at 97% sequence similarity), comprising 82% of all identified OTUs, were represented three or fewer times in this 4889-sequence library. When the data were reanalyzed to include all OTUs represented 19 or fewer times (one-half the value of the tenth most predominant OTU), 99% of all OTUs were included in this category. This represents 83% of all sequences in the full library, indicating that although individually they are underrepresented, they should not be ignored as they collectively represent the vast majority of the diversity in this ecosystem. As considered in the previous section, depth of sequencing is crucial to understanding the importance of varying microbial diversity in ecosystems, and the reality is that for comprehensive sampling of the hidden diversity in the rare biosphere, extensive sampling is needed [Bent and Forney, 2008]. Smaller library sizes are sometimes rationalized by the argument that major biogeochemical processes are not significantly affected by organisms present in small numbers [Pedros-Alio, 2006; Fuhrman, 2009]. However, in order to fully understand the role that these minority populations play in an ecosystem, how they are maintained and how they affect diversity, richness and evenness these rare organisms must be sampled and characterized.

Since numerically dominant phylotypes are expected to control major biogeochemical processes, we explored...
### Results and Discussion

#### Table 16.2  Taxonomic Classification Based on Multiple Methods

<table>
<thead>
<tr>
<th>Taxon</th>
<th>Total Number of Sequences</th>
<th>Classifier(^*)</th>
<th>Unclassified(^*)</th>
<th>% Sequence Similarity(^*)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Phylum</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acidobacteria</td>
<td>955</td>
<td>88</td>
<td>41</td>
<td>38</td>
</tr>
<tr>
<td>Actinobacteria</td>
<td>491</td>
<td>452</td>
<td>39</td>
<td>38</td>
</tr>
<tr>
<td>Bacteroidetes</td>
<td>453</td>
<td>450</td>
<td>17</td>
<td>59</td>
</tr>
<tr>
<td>CD OD1</td>
<td>68</td>
<td>0</td>
<td>0</td>
<td>49</td>
</tr>
<tr>
<td>CD OPI0</td>
<td>43</td>
<td>18</td>
<td>0</td>
<td>38</td>
</tr>
<tr>
<td>CD OPI1</td>
<td>14</td>
<td>6</td>
<td>6</td>
<td>27</td>
</tr>
<tr>
<td>CD OPI3</td>
<td>12</td>
<td>0</td>
<td>0</td>
<td>53</td>
</tr>
<tr>
<td>CD TM6</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>83</td>
</tr>
<tr>
<td>CD TM7</td>
<td>17</td>
<td>11</td>
<td>2</td>
<td>56</td>
</tr>
<tr>
<td>CD WS1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>99</td>
</tr>
<tr>
<td>CD WS3</td>
<td>34</td>
<td>17</td>
<td>5</td>
<td>38</td>
</tr>
<tr>
<td>Chlamydiae</td>
<td>2</td>
<td>2</td>
<td>10</td>
<td>88</td>
</tr>
<tr>
<td>Chlorothe</td>
<td>22</td>
<td>0</td>
<td>1</td>
<td>70</td>
</tr>
<tr>
<td>Chloroflexi</td>
<td>27</td>
<td>32</td>
<td>18</td>
<td>55</td>
</tr>
<tr>
<td>Cyanobacteria</td>
<td>12</td>
<td>6</td>
<td>57</td>
<td>38</td>
</tr>
<tr>
<td>Deferribacteria</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>—</td>
</tr>
<tr>
<td>Deinococcus-Thermus</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>100</td>
</tr>
<tr>
<td>Deltaproteobacteria</td>
<td>0</td>
<td>0</td>
<td>21</td>
<td>—</td>
</tr>
<tr>
<td>Firmicutes</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>60</td>
</tr>
<tr>
<td>Gemmatimonadetes</td>
<td>251</td>
<td>195</td>
<td>33</td>
<td>38</td>
</tr>
<tr>
<td>Lentisphaerae</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>—</td>
</tr>
<tr>
<td>Nitrospirae</td>
<td>59</td>
<td>47</td>
<td>0</td>
<td>38</td>
</tr>
<tr>
<td>Planctomycetes</td>
<td>243</td>
<td>174</td>
<td>22</td>
<td>38</td>
</tr>
<tr>
<td>Proteobacteria</td>
<td>1690</td>
<td>1631</td>
<td>506</td>
<td>21</td>
</tr>
<tr>
<td>Spirochaetes</td>
<td>3</td>
<td>1</td>
<td>5</td>
<td>38</td>
</tr>
<tr>
<td>Thermodesulfovibrio</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>—</td>
</tr>
<tr>
<td>Thermotoga</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>—</td>
</tr>
<tr>
<td>Verrucomicrobia</td>
<td>103</td>
<td>89</td>
<td>50</td>
<td>43</td>
</tr>
<tr>
<td>Unclassified</td>
<td>35</td>
<td>1654</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td><strong>Class</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alphaproteobacteria</td>
<td>374</td>
<td>368</td>
<td>13</td>
<td>21</td>
</tr>
<tr>
<td>Betaproteobacteria</td>
<td>485</td>
<td>475</td>
<td>0</td>
<td>43</td>
</tr>
<tr>
<td>Deltaproteobacteria</td>
<td>348</td>
<td>278</td>
<td>22</td>
<td>38</td>
</tr>
<tr>
<td>Epsilonproteobacteria</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>99</td>
</tr>
<tr>
<td>Gammaproteobacteria</td>
<td>478</td>
<td>468</td>
<td>7</td>
<td>38</td>
</tr>
<tr>
<td>Unclassified</td>
<td>4</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

\(^*\)Classification based on ARB-generated tree.
\(^*\)Sequences were considered classified if assigned to the same phylum using both Seqmatch and Classifier tools from RDP.
\(^*\)Unclassified sequences were assigned to likely phylum based on Classifier results.
\(^*\)Based on ARB phylum level classification.
16.3.4 Effect of Sequence Length on Community Analysis

Metagenomics and high-throughput sequencing have recently come of age for exploring microbial assemblages. One shortcoming of the new sequencing platforms is the reduction in sequence lengths obtained. This has been shown to reduce the capability of searches to identify similar targets [Wommack et al., 2008]. We explored the effect of sequence length on community indices by comparing differently sized regions of the same set of sequences. A 400-bp fragment, spanning between *E. coli* positions 518 and 927 and encompassing two hypervariable regions (V4 and V5), was amplified for generation of the library in this study. In order to compare the outcome of ecological parameter estimates made using these ~400-bp sequence fragments to estimates based on longer sequences, 1184 full-length 16S gene sequences from the ARB database were selected that covered all of the phyla detected in our study and had the greatest similarity to sequences within our own library, thus serving as proxies for those sequences. These reference sequences were analyzed as both full-length and truncated sequences (generated by trimming to match the 536–907 region in our library, excluding primers) to create distance matrices at 97% sequence similarity that were used as input for the DOTUR program.

The use of full-length sequences produced a modest increase in some values. Observed richness increased from 911 to 1031 OTUs identified (Table 16.3), while the Shannon–Weaver index increased from 6.65 to 6.86, from truncated to full-length sequences, respectively. A more substantial effect on (Chao1) richness estimation was observed, with an almost two-fold increase in estimated numbers of OTUs when full-length sequences were used for the analysis, presumably reflecting greater phylogenetic resolution afforded by the additional sequence information. As previously noted, the use of smaller fragment sizes has been linked to reduced ability to assign sequences to specific lineages. We tested the effect of sequence length on taxonomic placement by comparing results from the full-length and truncated sequence sets and found that comparable results were obtained for both. Indeed, in 1166 of the 1184 cases (98.5%), congruent taxonomic assignments were obtained with the truncated sequences; while only in 16 cases (1.5%) did the additional sequence information result in a different taxonomic assignment (Table 16.3). Phylum-level classification based on ARB-based tree generation was highly reproducible independent of fragment length (Table 16.4). Collectively, this suggests that the 400-bp V4–V5 region examined for our survey, which is readily obtained from a single dideoxy sequence reaction, is sufficient to provide reliable phylogenetic placement at phylum and higher-order levels. The effect of sequence length on finer-level placement (genus and species) was not examined, being outside the context of the current study.

16.3.5 Assignment of Cutoff Values for Phylogenetic Clusters

16S rRNA gene-based studies have been abundant in microbial ecology in the last decade; perhaps as a result, several practices and rules of thumb have become accepted almost to the point of being dogmatic, while revisions of these based on current knowledge have been lagging. Once such rule of thumb is the assumption that sequences sharing 60–80% identity likely belong to the same phylum [Schloss and Handelsman, 2005, 2006a]. Using these values as a guideline, we empirically
16.3 Results and Discussion

Table 16.3 Effect of Fragment Length on Similarity-Based OTU Number, Shannon–Weaver Diversity Index, and Richness Estimation

<table>
<thead>
<tr>
<th>Sequence Length, % Similarity Level</th>
<th>Number of Unique OTUs</th>
<th>Shannon–Weaver Index</th>
<th>Richness Estimate(^a)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full Length(^b)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>1183</td>
<td>7.08</td>
<td>350,169</td>
</tr>
<tr>
<td>97</td>
<td>1031</td>
<td>6.86</td>
<td>7,452</td>
</tr>
<tr>
<td>70</td>
<td>54</td>
<td>3.01</td>
<td>67</td>
</tr>
<tr>
<td>55</td>
<td>6</td>
<td>0.67</td>
<td>6</td>
</tr>
<tr>
<td>46</td>
<td>2</td>
<td>0.01</td>
<td>2</td>
</tr>
<tr>
<td>Truncated(^c)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>1166</td>
<td>7.05</td>
<td>61,646</td>
</tr>
<tr>
<td>97</td>
<td>911</td>
<td>6.65</td>
<td>4,175</td>
</tr>
<tr>
<td>70</td>
<td>80</td>
<td>3.41</td>
<td>93</td>
</tr>
<tr>
<td>55</td>
<td>15</td>
<td>1.68</td>
<td>18</td>
</tr>
<tr>
<td>46</td>
<td>5</td>
<td>0.47</td>
<td>6</td>
</tr>
</tbody>
</table>

\(^a\)Based on full biased corrected Chao1 richness estimates.
\(^b\)Based on 1184 full-length sequences.
\(^c\)Based on 1184 truncated sequences. Truncations were created by deleting the upstream base-pair region from the E. coli consensus position 536 and downstream of consensus position 906.

Table 16.4 Effect of Sequence Length on Taxonomic Placement and Distance Based on ARB Alignment

<table>
<thead>
<tr>
<th>Taxon(^a)</th>
<th>Full-Length</th>
<th>Truncated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phylum</td>
<td>% Sequence Similarity</td>
<td>Number of Sequences</td>
</tr>
<tr>
<td>Acidobacteria</td>
<td>58</td>
<td>176</td>
</tr>
<tr>
<td>Bacteroidetes</td>
<td>52</td>
<td>106</td>
</tr>
<tr>
<td>CD OD1</td>
<td>57</td>
<td>20</td>
</tr>
<tr>
<td>Gemmatimonadetes</td>
<td>75</td>
<td>42</td>
</tr>
<tr>
<td>Planctomycetes</td>
<td>46</td>
<td>88</td>
</tr>
<tr>
<td>Class</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Betaproteobacteria</td>
<td>64</td>
<td>141</td>
</tr>
</tbody>
</table>

\(^a\)CD, candidate division.

Based on our data, we can reject the hypothesis that the outcome could have been altered by sample size (number of sequences within a given group).
A lack of correlation was found between sample size and the percent sequence similarity required for correctly clustering the suite of classified sequences into their respective phyla, suggesting that there are true differences in the degree of 16S sequence variation between different phyla (Table 16.4). This suggests that different evolutionary strategies are pursued by different phyla, as reflected by differences in the degree of sequence divergence in the 16S rRNA gene that are generated or tolerated. Differences in genomic sequences are common, especially when concerning ribosomal sequences that can have varying operon copy numbers even amongst the same species [Crosby and Criddle, 2003; Binnemans et al., 2006]. Combine this with other factors like codon usage bias, which has been linked to differences in growth capabilities and levels of gene expression [Sharp and Li, 1987; Willenbrock et al., 2006], and it becomes clear that utilizing a uniform cutoff value to delineate phylogenetically related groups based on a single gene is an overly simplistic approach.

### 16.3.6 Comparison of Fractionated Versus Nonfractionated DNA Libraries

It has become clear from recent studies that despite extensive efforts to fully sequence all bacteria in an environmental sample, using library-based or metagenomic approaches, the probability of success is small. Given that the majority of diversity is hidden in underrepresented groups, specialized or targeted approaches need to be applied to effectively sample the bacterial diversity within an ecosystem. Previous work by Holben and co-workers [Holben and Harris, 1995; Holben et al., 2004] has successfully used GC fractionation as a way to enhance recovery of targets when using DNA-based methods. The technique physically separates a complex bacterial community genomic DNA sample based on the GC content of its component populations allowing for targeted sampling of DNA within a given GC range. We examined how this technique affects recovery of low-abundance sequences in the complex mixture of bacterial community DNA from our Midwestern agricultural soil sample by comparing a GC-fractionated library to an unfractionated one. Based on OTU representation, no substantive differences in phylum or genus level community composition were detected (see supplemental material [Morales et al., 2009]). However, when the set of aligned sequences were analyzed using the DOTUR and EcoSim programs, the species detection (also known as rarefaction) curve of OTUs detected at 97% sequence similarity indicated a higher rate of recovery of new phylotypes for the GC-fractionated library (Table 16.5). Also see supplemental material [Morales et al., 2009]). In addition, the values for the Shannon–Weaver diversity index, evenness, and Chao1 richness estimation were all higher for GC-fractionated DNA compared to randomly sampled total bacterial community DNA (Table 16.5). The %G + C content of chromosomal DNA in the prokaryotic domain is (with some inherent variability) a characteristic indicator of groups of related bacteria at genus to phylum level. Genomic GC content values range from ~26% to over 70% and the complex mixture of total bacterial community DNA in an environmental sample can be fractionated as a function of the chromosomal GC content of the component populations using CsCl equilibrium density gradient centrifugation as described previously [Holben and Harris, 1995; Holben et al., 2004]. This separation is based on differential density imposed by the AT-dependent DNA binding dye bis-benzimidazole. Following ultracentrifugation, fractionation of the gradient through an inline UV detector can provide a profile of the bacterial community in terms of DNA abundance versus GC content. Furthermore, the %G + C content represented by each gradient fraction can be determined by linear regression analysis of data obtained from control gradients containing standard DNA samples of known %G + C composition. This procedure is described and discussed in detail in Chapter 25 of Volume I. Perhaps the main benefit of this approach is that the complex mixture of total community DNA in a sample can be physically separated into discrete fractions (again as a function of GC content), thereby reducing the overall complexity of the complete mixture into multiple components prior to downstream analyses and manipulations such as PCR amplification. By thorough

<table>
<thead>
<tr>
<th>GC-fractionated</th>
<th>Number of Unique OTUs</th>
<th>Shannon–Weaver Index</th>
<th>Evenness</th>
<th>Richness Estimate</th>
<th>% of Shared OTUs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>335</td>
<td>5.62</td>
<td>0.966</td>
<td>1020</td>
<td>64</td>
</tr>
<tr>
<td>Not GC-Fractionated</td>
<td>301</td>
<td>5.45</td>
<td>0.954</td>
<td>780</td>
<td>74</td>
</tr>
</tbody>
</table>

*Based on 487 starting sequences.
*OTUs identified in both libraries.
sampling of individual GC fractions, it is easier to cover all the diversity present in a sample, very deep coverage is essential for detection of minority populations, which collectively are the major reservoir of bacterial diversity in most communities. Although some reports suggest that major biogeochemical transformations are controlled by numerically predominant organisms, we do not at this time possess sufficient information to disregard the role of the rare biosphere in important ecosystem processes. A good example for soil environments is the nitrifying bacteria whose numerical abundance in agricultural soils typically ranges from \(10^4\) to \(10^9\) bacteria/g dry weight in a background community typically numbering \(>10^4\) bacteria/g ([Kowalchuk and Stephen, 2001]). It is our opinion that additional, deeper, and more robust sampling is essential in order to obtain an accurate assessment of community diversity parameters. This is especially true where even near-saturation of sampling curves has not previously been feasible due to large numbers of taxa exhibiting high degrees of evenness, or where theoretical estimates based on relatively small sample sizes (<1000) do not appear to be accurate (e.g., asymptotic behavior is not yet apparent in a sampling curve). The importance of at least approaching sampling saturation is supported by a recent publication indicating that surveys missing or ignoring a small subset (e.g., <10% of species) result in minimal loss of information, but that more extensive gaps in data substantially increase information loss ([Vellend et al., 2007]).

Secondly, although only indirectly examined herein, primer selection is critical to future studies of microbial diversity via sequence analysis. Using short sequences has effects on various diversity parameters, and this effect can be exacerbated due to selection of inferior primers. Youssef et al. [2009] conducted a thorough comparison of different primer locations within the 16S rRNA gene and of PCR product sizes to examine effects on species richness estimates. They found substantial differences as a result of primer selection and also supported our findings on the effect of sequence length. Additionally, it has been demonstrated that libraries built using a single primer pair or DNA extraction method are not sufficiently robust to detect the total diversity within a sample ([Feinstein et al., 2009; Hong et al., 2009]). This means that even after extremely deep sequencing efforts, one would still be missing some of the diversity present using currently favored approaches.

Thirdly, the use of currently available computational tools for determining alignments, phylogenetic comparisons, and classification must be done with care. Theoretically, these approaches should support each other if properly performed with valid data, but in many cases they do not. This lack of coherence is highly detrimental when comparison of results from different studies is desired. Freely accessible, high-capacity tools like those provided by the RDP II website are a great service, especially when they can produce results congruent to those obtained using other, more labor-intensive methods as demonstrated here.

Lastly, it is likely a matter of time before rapidly evolving technologies allow complete access to all of the diversity-based microbial diversity in a sample; but until then, we can accomplish much by employing synergistic techniques. This could help minimize reliance on “brute force” and randomly based approaches when studying the depths of microbial diversity. We present one option, GC fractionation; but other approaches, such as community-level subtractive hybridization, are available that can facilitate access to the diversity present in a given community with fewer sequences required.

16.4 CONCLUSIONS

Several key concepts are outlined in this study that are salient to future metagenomic and high-throughput sequencing studies of microbial diversity. First, if we are to accurately portray the total diversity present in a sample, very deep coverage is essential for detection of minority species ([Holben et al., 2004]) and more-even detection of total diversity, thereby reducing the required survey size needed to approach complete coverage of the entire bacterial community ([Morales et al., 2009]).

INTERNET RESOURCES

The KBS LTER Site (http://lter.kbs.msu.edu/)
ARB (http://www.arb-home.de/)
DOTUR (http://schloss.micro.umn.edu/software/)
ECOSIM (http://www.garyentsminger.com/ecosim/index.htm)
DAM and DOTMAN programs (http://holben-lab.dbs.umt.edu/links.php)
RDP Classifier and SeqMatch (http://rdp.cme.msu.edu/index.jsp)

Acknowledgments

Funding for this project was provided by the United States Department of Agriculture, National Research Initiative (USDA-CSREES grant no. 2004–03501). Soil samples for this project were graciously provided by the KBS-LTER. We also gratefully acknowledge the Ribosomal Database Project (RDP-II) for aid in the chimera check
process, as well as Linda Schimmelpenning and Tara West-
lec for technical assistance in performing GC fractionation,
cloning, and sequence processing. We thank Jared Rapp
for aiding in computer-based data analysis.

REFERENCES

ANDERSSON AF, LINDBERG M, IVASON J, NOLLA AT, RUKOVA E, YKVIN P,
ENGSTRAND L. 2008. Comparative analysis of human gut microbiota

At least 1 in 20 16S rRNA sequence records currently held in
public repositories is estimated to contain substantial anomalies.

New screening software shows that most recent large 16S rRNA
72:3574–3578.

BENT SJ, FORNEY LJ. 2008. The tragedy of the common: Understanding
limitations in the analysis of microbial diversity. ISME J.
2:689.

BIER EJ, SUN S, HOWARD EC. 2009. Prokaryotic genomes and diversity
in surface ocean waters: Interrogating the global ocean sampling

Ten years of bacterial genome sequencing: Comparative-

Sensitivity to sample coverage and intensity to spatial pat-

COLE JR, CHAI B, FARKAS R, USAING W, Q KI, KOSMAN-AI, MIRSHAHRI
DAM et al. 2007. The ribosomal database project (RDP-
II): Introducing myRDP space and quality controlled public data.
Nucleic Acids Res. 35:D169–D172.

COLE JR, WANG Q, CAI Z, LUGAES Q, CHUZHANOVA NA, BIRCH-HYATT
EJ, BENDS J, BÁC A, NETTLE TN, WENZEL M, KRAMER KA, HAHN
GH, et al. 2006. Greengenes, a chimera-checked 16S rRNA
gene database and workbench compatible with ARB. Appl.

CROOKS LD, CHU Z. 2003. Understanding bias in microbial com-
munity analysis techniques due to rRNA turnover. Nucleic
Acids Res. 31:1178–1183.

CURTIS T. 2006. Microbial ecologists: It’s time to ‘go large’. Nature
440:860.

DÉKERTIS TZ, HUGGINSOLZ P, LARSEN N, RUSAN M, BUDGE IR, KALLER
K, et al. 2006. Gist-organizer, a chimera-checked 16S rRNA
gene database and workbench compatible with ARB. Appl.

DEININGER EC. 2004. MUSCLE: Multiple sequence alignment with high

DUNBAR J, LESLIE C, WU T, SHIER J, WANG S, CHEN D, LIU B, YUAN
of rare members of the soil microbiome. Appl. Environ.
Microbiol. 74:5302–5405.

EINSTEIN LM, SUI WE, BLACKWOOD CB. 2009. Assessment of bias
associated with incomplete extraction of microbial DNA from soil.

FURMAN JA. 2009. Microbial community structure and its functional

GARRITY GM, BILL JA, LIEBENS TG. 2004. Taxonomic outline of the
Bergey’s Manual Trust.

HÄMÄVY M, WALKER J, HARRIS J, GOLD N, KNIGHT R. 2008. Error-
correcting barcoded primers allow hundreds of samples to be pyro-

HANGGE WP, FRAZIER C, SWAETZ BG. 2006. Sequences, sequence clas-

HOLLÉN WE. 1997. Isolation and purification of bacterial com-
munity DNA from environmental samples. In Manual of Environmental

HOLLÉN WE, HARRIS D. 1995. DNA-based monitoring of total bac-
terial community structure in environmental samples. Mol. Ecol.

HOLLÉN WE, FERGUSON KJ, ROLLINSON A, APINALIET J, B. 2004. GC
fractionation enhances microbial community diversity assessment
and detection of minority populations of bacteria by denaturing gradient

HONG SH, BUNGE J, LEVIN C, ION S, ENVIN SJ. 2009. Polymerase
chain reaction primers miss half of 16S rDNA microbial diversity. ISME
J. 3:1365–1373.

JANSSEN PH. 2006. Identifying the dominant soil bacterial taxa in
libraries of 16S rDNA and 16S rRNA genes. Appl. Environ.

KOWALCHUK GA, STEINHELF JR. 2001. Ammonia-oxidizing bacteria:

LÆRCH CL, HAMMÈN M, KNOCHER K, POLKER K. 2009. Pyrosequencing-
based assessment of soil ph as a predictor of soil bacterial community

LEUWAJ J, STRUNK D, WEBSTRA R, BESLER L, MEYER H, KUSAY Y,
Acids Res. 32:1633–1637.

MORSELS SE, HOLLÈN WE. 2009. Empirical testing of 16S rRNA
gene PCR primer pairs reveals variance in target specificity and effi-

MORSELS SE, COHRT T, JOHNSON JV, HOLLÈN WE. 2008. Supplement-
mental programs for enhanced recovery of data from the DOTUR

MORSELS SE, COHRT TP, JOHNSON JV, HOLLÈN WE. 2009. Extensive
phylogenetic analysis of a soil bacterial community illustrates extreme
taxon evenness and the effects of amplicon length, depth of coverage
and DNA fractionation on classification and ecological parameters.


NIEDERBERGER TD, McDONALD IR, HACKER AL, SOO RM, BARRETT
JE, WALL DH, CART SC. 2008. Microbial community composition
10:1713.

NOCKER A, BURK M, CUSCEPT AK. 2007. Genotypic microbial commu-

PIERON-ALOISI, C. 2004. Mane microbe diversity: can it be deter-


RAVENS MS, GIVOGRIVISI SJ. 2003. The uncultured microbial majority.

RUSHTON C, BOTTIE H. 2009. Diversity of total, nitrogen-fixing and den-

ROCHÉ LF, PULITOREE RR, REÍN A, CASTELA G, HAWKIN AKM,
KENT AD, et al. 2007. Phylogenetic enumerates and contrasts soil
microbial diversity. ISME J. 1:283–290.

SANJUAN H, RESHANTRA B, DEGENS A, TIMOFIEVSHE J, VERGIL TM,
NIEMI X, et al. 2006. Potential of a 16S rDNA-based taxonomic
microarray for analyzing the zebrafish effects of maize on Agrobac-
72:4302–4312.


Queries in Chapter 16

Q1. Confirm chap. no.
Q2. We have shortened the running head, since it exceeds the textwidth. Please confirm
Q3. Where is this?
Q4. Confirm chap. no.
Q5. Supply page range
Q6. Supply city & publisher
Q7. cite refs
Q8. Cite refs