

## Understanding Bias in Microbial Community Analysis Techniques due to *rrn* Operon Copy Number Heterogeneity

BioTechniques 34: \_\_-\_\_ (April 2003)

**Laurel D. Crosby and Craig S. Criddle**  
Stanford University, Stanford, CA, USA

### ABSTRACT

*Molecular tools based on rRNA (rrn) genes are valuable techniques for the study of microbial communities. However, the presence of operon copy number heterogeneity represents a source of systematic error in community analysis. To understand the types and magnitude of such bias, four commonly used rrn-based techniques were used to perform an in silico analysis of a hypothetical community comprised organisms from the Comprehensive Microbial Resource database. Community profiles were generated, and diversity indices were calculated for length heterogeneity PCR, automated ribosomal intergenic spacer analysis, denaturing gradient gel electrophoresis, and terminal RFLP (using RsaI, MspI, and HhaI). The results demonstrate that all techniques present a quantitative bias toward organisms with higher copy numbers. In addition, techniques may underestimate diversity by grouping similar ribotypes or overestimate diversity by allowing multiple signals for one organism. The results of this study suggest a degree of caution should be used when interpreting rrn-based community analysis techniques.*

### INTRODUCTION

Microbial ecology addresses a variety of issues that range from the function of a single population to the myriad interactions of complex communities. Researchers in this field have added the challenge of scope because microbial diversity and community dynamics must be inferred using indirect methods. Unfortunately, the tools for discriminating and measuring microbial populations are far from ideal. One of the challenges is that microbial communities are exceedingly diverse, with estimates suggesting between 4000 and 10 000 different microbial genomes per gram of soil or sediment (1,2). These estimates are based on DNA-DNA re-annealing curves, yet traditional isolation techniques have recovered only a small fraction of this estimated diversity (3–5). Limitations to culturing include the inability to predict the proper culture medium to select unknown organisms, and the propensity of fast-growing organisms to outgrow and overshadow the more relevant organisms that grow more slowly. The development of molecular approaches to community analysis have circumvented the need for cultivation because phylogenetically informative DNA sequences can be directly screened from the environment.

The most widely used techniques for organism identification and community analysis include those based on 16S rRNA (*rrn*) genes because of the quality of phylogenetic information, rapid and straightforward procedures, and large databases of sequence infor-

mation. Despite the advantages of ribosomal DNA sequence analysis for studies of bacterial isolates, limitations exist for using rRNA genes to analyze mixed communities (6–9). Problems arise as a result of organisms that have variable numbers of copies of the *rrn* operon (10,11) and sequence heterogeneity between operons (12). While intracistronic heterogeneity has been cited as a source of “noise” for determining the phylogenetic rank of an isolate (12), the influence of *rrn* operon copy number and sequence heterogeneity on community analysis techniques is much more serious. The problem is that microbial communities, in almost all instances, are mixtures of unknown organisms with unknown numbers of copies of the *rrn* operon.

Techniques for the rapid evaluation of community diversity have several features in common. First, total genomic DNA is extracted from an environmental sample, and sequences of 16S genes or intergenic spacer regions are copied and amplified above the background of the genome using PCR. These copies of DNA are then subjected to various discrimination methods, including electrophoretic separation of fragments based on length, melting temperature, or restriction fragment lengths. The number of different electrophoretic bands or peaks in the analysis serves as a proxy for diversity, as the different ribosome types (ribotypes) are considered unique to a group of organisms. [Note that the term “ribotype” is used here in terms of a class of RNA, as opposed to the ribotype theory of the origin of life (13).] For all techniques,

the signal intensity of a particular peak reflects the number of copies of the DNA fragment that contribute to that peak. Unfortunately, the presence of variable numbers of operons for organisms in diverse communities leads to a mixture of overlapping signals, multiple signals for single populations, and distorted estimates of abundance between organisms. The result is a complicated portrait of community diversity that is difficult to interpret. To acknowledge the potential biases in these techniques, authors prudently advise readers to “interpret these data with caution.”

For researchers to develop sound experimental designs, accurate hypotheses, and meaningful conclusions regarding community structure and function, sources of systematic error in community analysis techniques must be identified and quantified. Biases related to genomic DNA extraction and PCR amplification are well documented (14–21). The goal of this paper is to illustrate how *rrn* operon copy number heterogeneity influences the interpretation of four commonly used 16S rDNA-based community analysis techniques. A hypothetical community was constructed using gene sequences retrieved from the Comprehensive Microbial Resource (CMR) database, which is a collection of completely sequenced and annotated genomes compiled by The Institute of Genomic Research (TIGR) (Rockville, MD, USA). DNA sequences encoding the 16S rRNA gene and adjacent spacer regions were used for an *in silico* comparison of four major community analysis tech-

niques: length heterogeneity PCR (LH-PCR) (22), automated ribosomal intergenic spacer analysis (ARISA) (8), denaturing gradient gel electrophoresis (DGGE) (23), and terminal RFLP analysis (T-RFLP; with restriction enzymes *RsaI*, *MspI*, and *HhaI*) (9). Sequences were analyzed according to the priming sequences and discrimination methods for each technique. Diversity indices were calculated based on the observed distribution of fragment sizes (or melting temperatures for DGGE) and then compared with the ideal or “true” diversity indices for the hypothetical community. The results demonstrate that *rrn* operon copy number heterogeneity strongly influences the interpretation of 16S rDNA-based community analysis techniques.

## MATERIALS AND METHODS

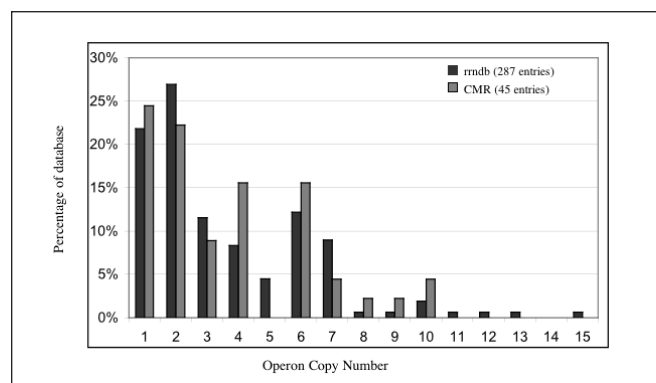
The CMR database, maintained by TIGR ([www.tigr.org](http://www.tigr.org)), was the source of the genome sequences analyzed in this report. The CMR database was selected over GenBank® because it contains all of the *rrn* operons for each organism. The organisms evaluated had *rrn* operon copy numbers ranging from 1 to 10, with 47% of organisms having either one or two operons and 71% having up to four (Figure 1). This distribution of operon frequency in the CMR database approximates the operon frequency for organisms in the Ribosomal RNA Operon Database (rrndb) Release 2.3 reported by Klappenbach (24). The CMR database provided coordinates for the sequence location of the *rrn* genes,

which allowed for intergenic spacer regions to be retrieved intact with the 16S rRNA genes. Sequences were retrieved using the segment retrieval function on the CMR Web site, where coordinates for the 5′ end of the 16S gene and 5000 bases downstream were used to retrieve the segment. For

operons with the typical configuration of rRNA genes (16S, 23S, and 5S), this reflected the entire 16S rRNA gene, the intergenic spacer region between the 16S and 23S rRNA genes, and a portion of the 23S rRNA gene. The analysis included all microbial species for which operons were reported but was limited to those organisms in which PCR primers matched potential targets by at least 65% for at least two techniques. With this constraint, members of the domain *Archaea* were excluded from the analysis. In addition, the hypothetical community was limited to a single strain of a given species in the event that multiple strains were reported. This reduced the magnitude of the *rrn* operon copy number bias that could be attributed to the characteristics of any particular species.

The sequences for forward and reverse primers, as previously published for each method, were used to search for corresponding sites within the DNA sequence. When these sites were found, subsequences were extracted that represented the fragment between the 5′ end of the forward primer and the 3′ end of the reverse primer. The lengths of these fragments were recorded for the LH-PCR and ARISA techniques, while fragments for DGGE and T-RFLP underwent further manipulation. Sequence fragments for DGGE were exported to the Winmelt software package (MedProbe AS, Oslo, Norway) to estimate the  $T_m$  of the lowest melting domain. For T-RFLP, sequences for a given restriction enzyme were used to identify the fragment length between the 5′ end of the forward primer and the first instance of the restriction enzyme cutting site. For all techniques, the fragment lengths (or melting temperatures, as in the case of DGGE) were plotted as histograms to simulate the electropherogram type of output that is common to the automated techniques. DGGE differs in this regard but was similarly plotted to facilitate comparisons between the techniques.

To reduce bias unrelated to copy number, all members of the hypothetical community were assumed to be equally abundant, with an equal ratio of genomes. This eliminates the complication of variable cell densities or growth rates and allows for meaningful com-



**Figure 1. Distribution of operon copy number frequency between the CMR database and rrndb (Release 2.3).** At this time, the CMR comprises 45 different species entries, while the rrndb contains 287 entries.

# Research Report

parison of diversity indices. A final assumption was that there was no PCR amplification bias as a result of primer annealing efficiency. The 65% similarity cutoff between the primer and potential targets represents a PCR amplification reaction with low stringency. In reality, the type of systematic error attributed to primer bias is a serious complication of PCR-based community analysis techniques (17,21) and only exacerbates the errors contributed by *rrn* operon copy number heterogeneity.

Diversity indices were calculated for each technique, based on the observed fragment distribution, where each unique fragment type represented a particular ribotype. The Shannon-Weaver index was used as a diversity index and was calculated as follows:

$$H = -\sum(p_i)(\log p_i) \quad [\text{Eq. 1}]$$

where the summation is over all unique fragments  $i$ , and  $p_i$  is the proportion of an individual "peak height" (i.e., number of same-sized fragments) relative to the sum of all peak heights (i.e., total number of fragments). Richness is the number of unique fragment sizes (or melting temperatures) identified by each technique. No minimum signal intensity threshold was used to determine peak richness, although cutoffs are commonly applied to the interpretation of real community analysis data. Evenness, or the equitability of the observed ribotypes, was calculated from the Shannon-Weaver diversity function, where:

$$E = H/\ln(\text{richness}). \quad [\text{Eq. 2}]$$

At the same time, true values for richness, evenness, and the Shannon-Weaver index were calculated based on the known species composition of the hypothetical community, assuming that each species would contribute only one ribotype for each technique. For this hypothetical community, 41 ribotypes were present at equal abundance. The observed values for the diversity indices were then compared with the true values to gain insight into the type and magnitude of bias as a result of operon copy number.

## RESULTS

Table 1 presents the amplification product lengths, melting temperatures, and restriction fragment lengths for

each organism. The table is organized such that the discriminating power of the different techniques can be evaluated for any given organism, while generalizations about each technique can be made by perusing the columns. Entries highlighted in bold represent fragments that have two or more members in common, while the number of fragments of each length are presented in parentheses. The histograms in Figure 2 represent the distribution of ribotypes for each particular technique. The scales for frequency distribution were kept relevant to each technique, while the gridlines for the vertical axis were set to 10 U. This allows for a visual comparison of the scales between the different techniques.

LH-PCR gave hypothetical amplification products with lengths ranging 314–371 bp, with an average product length of 346 bp and standard deviation of 13 bp. The technique produced a total of 26 unique product lengths for the hypothetical community, where 14 (54%) represented unique peaks, and 12 (46%) peaks contained fragments from two or more organisms. Of those peaks with multiple contributors, the number of organisms within the peak ranged from two to eight. Fragments of 348 bp were highest in frequency, with eight organisms contributing 40 copies of the LH-PCR fragment to this peak. In addition, there were 11 organisms that contributed 52 copies of the fragment within a size range of 5 bp (352–356 bp). The incidence of interoperon heterogeneity (heterogeneity within the same organism) was relatively low, with only five organisms with more than one fragment length. Of the organisms with heterogeneous copies of length heterogeneity fragments, four had fragments that differed by only one base pair. *Bacillus subtilis* was the exception, with four unique fragment lengths of 352, 353, 354, and 355 bp.

For ARISA, the hypothetical amplification products ranged 308–1576 bp, with an average of 751 bp (SD 239). The product lengths were typically unique, with six instances in which a maximum of two to three organisms contributed to a peak. Three organisms did not contribute an amplification product because of the orientation of the 16S and 23S rRNA genes within the operon or be-

cause they lacked a sufficiently similar priming site (less than the 65% sequence similarity criterion). Those organisms that did not contribute a product tended to have a low operon copy number, falling in the range from one to two copies. Of the organisms with multiple operons, only *Deinococcus radiodurans* gave an instance of a missed product for one of its operons. Despite the presence of organisms with no hypothetical product, ARISA yielded several peaks that exceeded the true richness of the community. The hypothetical community of 41 organisms gave a total of 68 peaks. For organisms with multiple *rrn* operons, the number of unique amplification products was almost equal to the number of operons (Table 1). For example, *Staphylococcus aureus* gave five unique product lengths for six operons, and *B. halodurans* gave six distinctly different product lengths for each of its six operons. These length differences were more than 1–2 bp, as would be indicative of a minor insertion or deletion event. In many instances, the length differences were tens or hundreds of bases apart, likely corresponding to the presence or absence of various tRNA sequences in the intergenic spacer region (25). This result demonstrates a combination of two systematic errors for the ARISA technique: (i) the underestimation of community diversity through missing or overlapping sequences and (ii) the overestimation of diversity due to heterogeneous amplification product lengths for a single organism.

The profile for DGGE showed a range of melting temperatures from 70.4°C to 79.4°C, with an average temperature of 74.0°C and standard deviation of 1.7°C. The analysis gave a total of 32 different melting temperatures, with nine temperatures representing amplification products from multiple organisms. Of the peaks with multiple contributors, seven contained only two members and the other two contained four members. The incidence of melting temperature heterogeneity for a single organism was low, with only five organisms that gave multiple signals. For those with multiple temperatures, the differences were frequently limited to a tenth of a degree.

For the T-RFLP analysis, three en-

Table 1. Hypothetical Fragments Retrieved for LH-PCR, ARISA, DGGE, and T-RFLP

Organism	LH-PCR	ARISA	DGGE	T-RFLP <i>RsaI</i>	T-RFLP <i>HhaI</i>
<i>Agrobacterium tumefaciens</i>	<b>314 (4)</b>	1494 (1), 1575 (1), 1576 (2)	<b>75.2 (4)</b>	824 (4)	339 (4)
<i>Aquifex aeolicus</i>	371 (2)	607 (2)	79.4 (2)	503 (2)	22 (2)
<i>B. halodurans</i>	<b>354 (6)</b>	965 (1), 1010 (1), 1091(1), 1135 (1), 1254 (1), 1281(1)	<b>75.5 (1), 75.6 (5)</b>	485 (5), 656 (1)	<b>240 (6)</b>
<i>B. subtilis</i>	<b>352 (1), 353 (1), 354 (7), 355 (1)</b>	448 (1), 449 (4), 452 (3), 629 (2)	74.8 (7), <b>75.2 (3)</b>	<b>454 (1)</b> , 455 (1), 456 (5), 457 (1), <b>475 (2)</b>	238 (1), <b>240 (8)</b> , 241 (1)
<i>Borrelia burgdorferi</i>	N	N	67.6 (1)	28 (1)	437 (1)
<i>Brucella mellitensis</i>	<b>314 (3)</b>	1048 (3)	75.8 (3)	<b>106 (3)</b>	<b>61 (3)</b>
<i>Campylobacter jejuni</i>	<b>346 (3)</b>	1074 (3)	<b>74.2 (3)</b>	<b>453 (3)</b>	98 (3)
<i>Caulobacter crescentus</i>	316 (2)	969 (2)	<b>74.6 (2)</b>	422 (2)	332 (2)
<i>Chlamydia pneumoniae</i>	<b>356 (1)</b>	513 (1)	<b>71.6 (1)</b>	<b>106 (1)</b>	734 (1)
<i>C. trachomatis</i>	357 (2)	531 (2)	74.5 (2)	488 (2)	735 (2)
<i>Clorobium tepidum</i>	<b>342 (2)</b>	737 (2)	76.2 (2)	465 (2)	91 (2)
<i>C. perfringens</i>	347 (9), <b>348 (1)</b>	466 (1), 468 (4), 469 (1), 700 (2), 702 (2)	<b>75.2 (9)</b> , 75.3 (1)	<b>453 (9), 454 (1)</b>	233 (10)
<i>D. radiodurans</i>	329 (3)	N (1), 308 (1), 1022 (1)	<b>75.7 (3)</b>	448 (3)	82 (3)
<i>Enterococcus faecalis</i>	366 (4)	508 (2), 609 (1), 610 (1)	73.8 (4)	903 (4)	218 (4)
<i>E. coli</i>	<b>348 (7)</b>	<b>636 (1)</b> , 637 (1), <b>713 (1)</b> , 719 (2), 722 (1), 728 (1)	<b>74.6 (7)</b>	<b>427 (7)</b>	<b>373 (7)</b>
<i>Hemophilus influenzae</i>	<b>348 (6)</b>	758 (3), 1003 (3)	71.0 (6)	463 (6)	364 (6)
<i>Helicobacter pylori</i>	333 (1), 334 (1)	N	71.3 (2)	846 (2)	99 (2)
<i>Listeria innocua</i>	<b>356 (6)</b>	<b>529 (4), 779 (2)</b>	<b>72.6 (6)</b>	<b>435 (6)</b>	<b>186 (6)</b>
<i>Listeria monocytogenes</i>	<b>356 (6)</b>	528 (4), <b>779 (2)</b>	<b>72.6 (6)</b>	<b>435 (6)</b>	<b>186 (6)</b>
<i>Mesorhizobium loti</i>	<b>314 (2)</b>	1197 (2)	<b>75.2 (2)</b>	682 (2)	<b>61 (2)</b>
<i>Mycobacterium leprae</i>	<b>356 (1)</b>	<b>569 (1)</b>	76.7 (1)	307 (1)	193 (1)
<i>Mycobacterium tuberculosis</i>	344 (1)	559 (1)	77.6 (1)	638 (1)	201 (1)
<i>Mycoplasma genitalium</i>	<b>343 (1)</b>	482 (1)	70.4 (1)	<b>475 (1)</b>	226 (1)
<i>Mycoplasma pulmonis</i>	<b>346 (1)</b>	<b>569 (1)</b>	72.5 (1)	<b>477 (1)</b>	841 (1)
<i>Neisseria meningitidis</i>	<b>348 (4)</b>	946 (4)	74.4 (4)	126 (4)	<b>213 (4)</b>
<i>Nostoc sp.</i>	315 (4)	<b>569 (1)</b> , 796 (3)	<b>72.2 (4)</b>	424 (4)	228 (4)
<i>Porphyromonas gingivalis</i>	<b>353 (4)</b>	1037 (4)	72.0 (4)	318 (4)	102 (4)
<i>Pseudomonas aeruginosa</i>	<b>342 (4)</b>	753 (4)	72.9 (4)	644 (4)	155 (4)
<i>Ralstonia solanacearum</i>	<b>346 (4)</b>	784 (4)	74.7 (4)	<b>477 (4)</b>	572 (4)
<i>Rickettsia conorii</i>	330 (1)	N	73.0 (1)	132 (1)	1060 (1)
<i>Salmonella enterica</i>	<b>348 (6)</b> , 349 (1)	<b>636 (4)</b> , 797 (3)	<b>75.7 (7)</b>	<b>427 (6)</b> , 428 (1)	<b>373 (6)</b> , 374 (1)
<i>S. aureus</i>	<b>355 (6)</b>	586 (1), 620 (1), 648 (1), 757 (1), 830 (2)	<b>72.2 (6)</b>	486 (6)	238 (6)

# Research Report

**Table 1. Hypothetical Fragments Retrieved for LH-PCR, ARISA, DGGE, and T-RFLP continued**

Organism	LH-PCR	ARISA	DGGE	T-RFLP <i>RsaI</i>	T-RFLP <i>HhaI</i>
<i>S. pneumoniae</i>	<b>352 (4)</b>	<b>529 (4)</b>	<b>74.2 (4)</b>	889 (4)	579 (4)
<i>S. pyogenes</i>	<b>354 (6)</b>	704 (6)	73.6 (1), 73.7 (5)	629 (5), 630 (1)	581 (5), 582 (1)
<i>Synechocystis sp.</i>	317 (2)	<b>746 (2)</b>	73.5 (2)	425 (2)	1048 (2)
<i>Thermotoga maritima</i>	351 (1)	525 (1)	77.1 (1)	86 (1)	1113 (1)
<i>Treponema pallidum</i>	<b>352 (1), 353 (1)</b>	578 (1), 588 (1)	<b>75.6 (2)</b>	639 (1), 640 (1)	850 (1), 851 (1)
<i>Ureaplasma urealyticum</i>	<b>345 (2)</b>	573 (2)	NA	283 (2)	370 (2)
<i>Vibrio cholerae</i>	<b>348 (8)</b>	707 (1), <b>713 (2)</b> , 792 (2), 793 (1), 968 (1), 994 (1)	<b>71.6 (5)</b> , 71.7 (1), <b>72.2 (2)</b>	<b>427 (8)</b>	<b>213 (8)</b>
<i>Xylella fastidiosa</i>	<b>348 (2)</b>	<b>746 (2)</b>	<b>72.2 (2)</b>	479 (2)	373 (2)
<i>Yersinia pestis</i>	<b>348 (6)</b>	<b>746 (3)</b> , 806 (3)	<b>75.5 (6)</b>	884 (6)	373 (6)

Values represent amplified PCR fragment lengths for LH-PCR and ARISA, melting temperatures of the lowest melting domains for DGGE fragments, and restriction fragment lengths for T-RFLP (*RsaI* and *HhaI* only, *MspI* not included). The number of copies of each fragment are noted in parentheses, and the values in **bold** represent fragments that overlap in size with one or more organisms within the technique.

zymes were used to generate terminal restriction fragments: *RsaI*, *MspI*, and *HhaI*. Each restriction enzyme was used to generate an independent T-RFLP profile of the hypothetical community. The range of fragment lengths for *RsaI* was 28–903 bp; for *MspI*, the range was 24–566 bp; and *HhaI* fragments ranged 22–1113 bp. The average fragment lengths for *RsaI*, *MspI*, and *HhaI* were 495 (SD 188), 300 (SD 184), and 313 (SD 210) bp, respectively. For brevity, only *RsaI* and *HhaI* fragments were included in Table 1. T-RFLP analysis with all three enzymes showed examples of interoperon heterogeneity within a single organism and overlapping fragment lengths from multiple organisms. Most examples of interoperon heterogeneity occurred as a result of fragment lengths that differed by a single base pair. *B. halodurans* was the only example of an organism that possessed operons that had distinctly different restriction sites, which resulted in two widely divergent fragment lengths. For the *RsaI* enzyme, one of the six operons of *B. halodurans* differed by 171 bases; whereas for *MspI*, two operons out of six differed by 389 bases. This is an interesting result because this pattern emerged from two different C→T transitions, disrupting cut sites for two different restriction enzymes. Instances of overlapping signals were also observed for the T-RFLP

analysis, with *RsaI* giving seven peaks with multiple contributors, three peaks for *MspI*, and five for *HhaI*.

Richness was estimated as the number of different ribotypes presented by each technique. The true value for species richness for the hypothetical community was 41, while the observed richness values for the four techniques ranged from 26 (LH-PCR) to 68 (ARISA) (Table 2). DGGE gave a richness estimate of 32 members, while T-RFLP gave values of 42, 40, and 38 for *RsaI*, *MspI*, and *HhaI*, respectively. In this hypothetical community, the true measure for evenness was equal to 1.0, as all species were equally abundant. ARISA ranked highest with a value of 0.951, followed by DGGE (0.900), T-RFLP (0.904 for *RsaI*, 0.885 for *MspI*, and 0.881 for *HhaI*), and finally, LH-PCR (0.814). The Shannon-Weaver diversity index was used to account for the abundance and evenness of ribotypes generated by each technique. A comparison of values for this index showed that most techniques underestimated diversity, with the exception of ARISA (due to biases noted earlier). All other values fell below the true value of 3.714. The diversity indices for each of the techniques, in rank order, are as follows: ARISA (4.012), T-RFLP *RsaI* (3.378), T-RFLP *HhaI* (3.206), T-RFLP *MspI* (3.263), DGGE (3.120), and LH-PCR (2.653).

## DISCUSSION

Clearly, the *rrn* operon copy number has an effect on community analysis techniques based on 16S rRNA genes. Some of these techniques tend to combine signals into a single peak, while others tend to generate multiple signals for a single organism. However, for all of these techniques, the fact that an organism has multiple copies of an operon leads to a quantitative bias for that organism. The magnitude of this bias depends on several factors, including the range of fragment sizes generated by the primer sets, the region of the *rrn* operon amplified, and the discriminating power of capillary and gel electrophoresis.

LH-PCR suffers most from overlapping signals because it requires the discrimination of small base pair differences. As an example, the original reference for LH-PCR illustrates how amplified products from soils form a contiguous distribution (22). This technique has been previously cited as a tool for quick assessments of changes, but the meaning of any such change is difficult to assess. For example, the loss of a high copy number organism would result in a more pronounced response than the loss of a low copy number organism. Thus, more attention would be drawn to drastic changes in large peaks, rather than subtle changes that may be

**Table 2. Diversity Indices for the Hypothetical Community**

Method	Shannon-Weaver Index	Richness	Evenness
LH-PCR	2.653	26	0.814
ARISA	4.012	68	0.951
DGGE	3.120	32	0.900
T-RFLP ( <i>Rsal</i> )	3.378	42	0.904
T-RFLP ( <i>MspI</i> )	3.263	40	0.885
T-RFLP ( <i>HhaI</i> )	3.206	38	0.881
Ideal (species level)	3.714	41	1.000

Calculations for the ideal values were based on an a community with all populations at equal abundance.

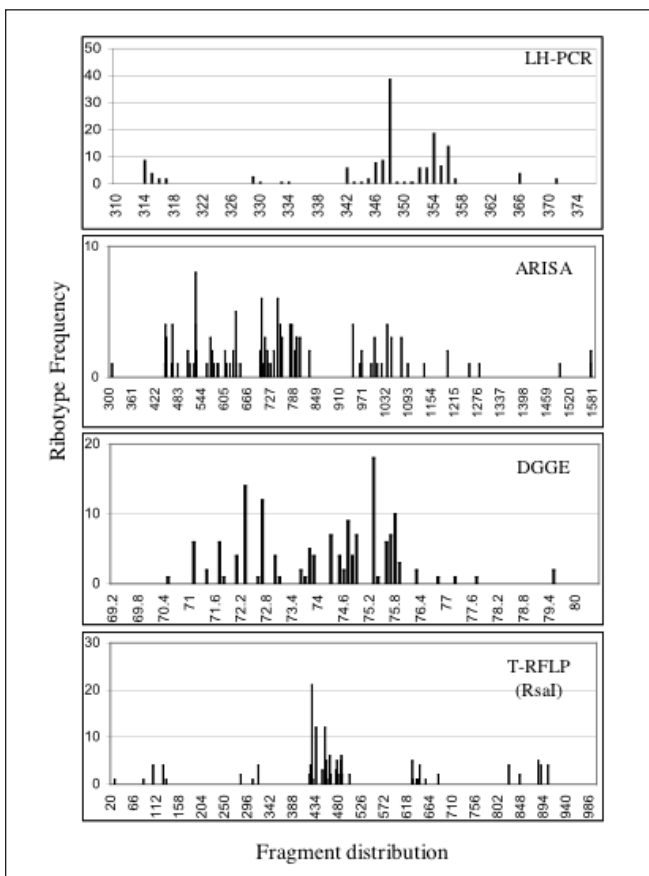
equally or more relevant.

ARISA has also been cited as a valuable tool due to its simplicity and rapidity. Use of the intergenic spacer region has the advantage that ISR fragment banding patterns confer a finer degree of phylogenetic resolution for microbial isolates compared to fragment analysis of 16S rRNA genes. Un-

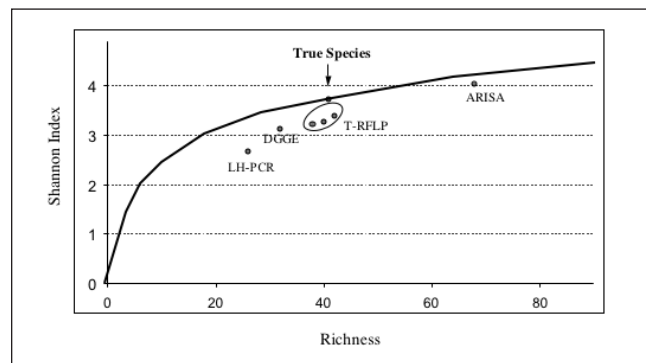
fortunately, heterogeneity in the lengths of intergenic spacer regions is a serious complication for studies of mixed communities. The magnitude of the problem is strongly influenced by the number of organisms with high copy number, as opposed to the number of organisms with fewer copies. Of the 22 organisms that gave a singular response, the average copy number was 2.45 (SD 1.37), compared to the copy number average of the whole population, which was 3.67 (SD 2.49). For comparison, the 16 organisms that gave multiple signals had an average of 5.88 copies of the operon (SD 2.25). Another observation is that the organ-

isms that did not give a signal (because of sequence variation relative to the “universal” PCR priming sites) also tended to have a low copy number. Again, this suggests that organisms of potential relevance to the function of the community may be overlooked. Techniques such as DGGE and T-RFLP also demonstrate copy number bias, but to a lesser extent than LH-PCR or ARISA. For DGGE, heterogeneity between operons was typically limited to a single base change between the DGGE fragments, which corresponded to a temperature difference of 0.1°C. The technique also displayed examples of overlapping sequences as a result of two or more organisms sharing the same ribotype. One point to consider for the hypothetical DGGE analysis is that melting temperatures were estimates rather than discreet values. Thus, in this hypothetical scenario, the bias due to overlapping fragments may be greater or less than the bias observed experimentally. Another consideration for the hypothetical analysis of the DGGE technique is that DNA fragments are normally separated and visualized by gel electrophoresis, rather than automated capillary electrophoresis. Thus, the degree of resolution of a gel may differ from the electropherogram type of output that is common to the other techniques.

T-RFLP analysis with each of the three restriction enzymes showed heterogeneity between operons but was typically limited to 1–2 bp. This corresponds to minor insertion or deletion events (indels) occurring between the operons. The disruption of a restriction



**Figure 2. Frequency distribution of fragments generated by LH-PCR, ARISA, DGGE, and T-RFLP (*Rsal*).** Scales for the x-axis are relative to each technique, while the gridlines for the y-axis were set to units of 10.



**Figure 3. Plot of Shannon-Weaver index versus richness for communities of equally abundant populations.** The relative positions of LH-PCR, ARISA, DGGE, and T-RFLP indicate how far the techniques deviate from the true value for these indices.

# Research Report

---

enzyme cutting site was also observed in the hypothetical community, although the incidence of this type of mutation was much less frequent. Regarding the discriminating power of T-RFLP, the wider range of possible fragment lengths led to a finer level of separation than for LH-PCR or DGGE. However, the T-RFLP profile also included several small peaks that abutted larger peaks, which, in practice, may not be finely differentiated. Another consideration for T-RFLP profiles and the other techniques is that interpretation often involves a cutoff for peaks that fall below a given intensity threshold. For this hypothetical analysis, no threshold was set, although several peaks were present at very low frequency relative to the larger peaks. Including such a cutoff would have had the effect of omitting signals from some of the heterogeneous operons of a single organism and from low copy number organisms. This would have influenced the calculation of the diversity indices and resulted in a lower estimate of diversity. According to the Shannon-Weaver diversity index calculations, T-RFLP comes close to approximating the actual diversity of the hypothetical community. However, it should be emphasized that, in this case and for all techniques described here, the bias of overlapping fragments directly offsets the bias of multiple signals by a single organism. Thus, two compensating errors do not necessarily yield a correct answer.

A potential limitation of this study is the fact that the CMR database currently emphasizes medically relevant organisms. These organisms may be systematically different in their copy number compared to environmental isolates, although comparison with the less medically oriented rrndb suggests that the copy number distributions are quite similar. The current size of the CMR database and the restrictions for inclusion in the hypothetical community also limit the scope of this analysis. However, the behavior of the Shannon-Weaver diversity index with respect to richness (for communities of uniform abundance and therefore an evenness value of 1.0) lends some insight into the value of this small subset of organisms (Figure 3). At low diversity, the

addition of a single population has a large impact on the diversity index value because the proportion of the new population is relatively large compared to the total number of populations. As the community grows, the impact of each additional population becomes smaller and smaller. Given the observation that the diversity index changes most abruptly for communities of less than 20 organisms, it would appear that a hypothetical community of 41 organisms is sufficient to make meaningful conclusions regarding the analysis techniques described here. (Note that the various diversity indices used in this analysis may not be appropriate for “real” 16S rDNA-based fragment analyses due to biases inherent in PCR amplification, fragment discrimination, and operon copy number issues.) Additions to the database and updated hypothetical analyses will determine whether these trends remain consistent.

It is instructive to observe the histograms generated by each technique for the hypothetical community. All organisms were equally represented on the basis of population densities, but the histograms show a wide variation in the ribotype abundance and diversity. This illustrates how peak amplitude can be deceptive in community analysis. Changes in the height of a particular peak can be caused by the growth or loss of a single population, while subtle changes in smaller peaks are overlooked or discounted. Understanding the *rrn* distribution for organisms in a particular environment would improve the application and interpretation of molecular analyses. Recent studies (11,26) suggest that variation in the copy number of the rRNA genes is related to the ecological strategy of an organism. That is, organisms with multiple copies of the *rrn* operon are able to mobilize quickly in response to rich growth conditions. Organisms with fewer *rrn* operons are more limited in their rate of ribosome synthesis and mobilize less quickly to an influx of nutrient into an environment. How does the dynamic nutrient profile of an environment shape the composition and, in turn, function of the microbial community? It may be that organisms of low *rrn* operon copy number comprise a significant portion of the microbial di-

versity, while high copy number organisms flourish during nutrient perturbations. These “fast responders” with high copy number are the same kinds of organisms that promptly appear on culture plates under traditional culturing methods, overshadowing the organisms that grow more slowly. Note that molecular-based tools were developed in part to avoid the bias of culture-based techniques, while the results of this study suggest that 16S rDNA-based molecular techniques may overemphasize the same organisms.

Another point to consider is the term ribotype, which is often meant to convey the sequence similarity of the 16S rRNA genes between two organisms. Ribotyping is used to describe the unique banding patterns of the rRNA gene using various methods of discrimination (restriction fragments, length heterogeneity, etc.) Two organisms are said to have common ribotypes when they give identical signals for a given technique. This hypothetical analysis demonstrates that while restriction fragment sites and gene fragment lengths may be in common for one technique, minute differences in the DNA sequence may yield divergent responses for another technique. Thus, it should be made clear that the concept of ribotype as a measure for diversity is entirely technique dependent.

This study provides an initial exploration of *rrn* operon copy number bias, based on the content of the databases available to date. Further investigation may lead to the refinement of existing methods and/or the development of correction factors for improved estimates of community diversity. In the meantime, method development should be directed toward technologies that are based on single copy genes and/or new discrimination methods. Until these new techniques are readily available and broadly applicable, researchers should continue to interpret *rrn*-based techniques with caution.

## ACKNOWLEDGMENTS

Support for this work was provided by a National Institutes of Health Training Grant in Biotechnology (no. T32GM008412) and a NASA Graduate

Student Researchers Program Fellowship (no. NGT-10-52619) to L.D.C. and by project no. DE-FG03-00ER63046-A001 from the U.S. Department of Energy NABIR program.

## REFERENCES

1. **Torsvik, V., F.L. Daae, R.A. Sandaa, and L. Ovreas.** 1998. Novel techniques for analyzing microbial diversity in natural and perturbed environments. *J. Biotechnol.* *64*:53-62.
2. **Torsvik, V., J. Goksoyr, and F.L. Daae.** 1990. High diversity in DNA of soil bacteria. *Appl. Environ. Microbiol.* *56*:782-787.
3. **Amann, R.I., W. Ludwig, and K.H. Schleifer.** 1995. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiol. Rev.* *59*:143-169.
4. **Giovannoni, S.J., T.B. Britschgi, C.L. Meyer, and K.G. Field.** 1990. Genetic diversity in Sargasso Sea bacterioplankton. *Nature* *345*:60-63.
5. **Ward, D.M., R. Weller, and M.M. Bateson.** 1990. 16S rRNA sequences reveal numerous uncultured microorganisms in a natural community. *Nature* *345*:63-65.
6. **Dahllöf, I., H. Baillie, and S. Kjelleberg.** 2000. rpoB-based microbial community analysis avoids limitations inherent in 16S rRNA gene intraspecies heterogeneity. *Appl. Environ. Microbiol.* *66*:3376-3380.
7. **Dahllöf, I.** 2002. Molecular community analysis of microbial diversity. *Curr. Opin. Biotechnol.* *13*:213-217.
8. **Fisher, M.M. and E.W. Triplett.** 1999. Automated approach for ribosomal intergenic spacer analysis of microbial diversity and its application to freshwater bacterial communities. *Appl. Environ. Microbiol.* *65*:4630-4636.
9. **Liu, W.T., T.L. Marsh, H. Cheng, and L.J. Forney.** 1997. Characterization of microbial diversity by determining terminal restriction fragment length polymorphisms of genes encoding 16S rRNA. *Appl. Environ. Microbiol.* *63*:4516-4522.
10. **Farrelly, V., F.A. Rainey, and E. Stackebrandt.** 1995. Effect of genome size and rrrn gene copy number on PCR amplification of 16S rRNA genes from a mixture of bacterial species. *Appl. Environ. Microbiol.* *61*:2798-2801.
11. **Klappenbach, J.A., J.M. Dunbar, and T. Schmidt.** 2000. rRNA operon copy number reflects ecological strategies of bacteria. *Appl. Environ. Microbiol.* *66*:1328-1333.
12. **Stackebrandt, E.** 2002. Defining taxonomic ranks. In *The Prokaryotes: an Evolving Electronic Resource for the Microbiological Community*. (Online reference.) Title No. 10125.
13. **Barbieri, M.** 1981. The ribotype theory on the origin of life. *J. Theor. Biol.* *91*:545-601.
14. **Frostegard, A., S. Courtois, V. Ramisse, S. Clerc, D. Bernillon, F. Le Gall, P. Jeannin, X. Nesme, et al.** 1999. Quantification of bias related to the extraction of DNA directly from soils. *Appl. Environ. Microbiol.* *65*:5409-5420.
15. **Martin-Laurent, F., L. Philippot, S. Hallet, R. Chaussod, J.C. Germon, G. Soulas, and G. Catroux.** 2001. DNA Extraction from soils: old bias for new microbial diversity analysis methods. *Appl. Environ. Microbiol.* *67*:2354-2359.
16. **Miller, D.N., J.E. Bryant, E.L. Madsen, and W.C. Ghiorse.** 1999. Evaluation and optimization of DNA extraction and purification procedures for soil and sediment samples. *Appl. Environ. Microbiol.* *65*:4715-4724.
17. **Polz, M.F. and C.M. Cavanaugh.** 1998. Bias in template-to-product ratios in multi-template PCR. *Appl. Environ. Microbiol.* *64*:3724-3730.
18. **Steffan, R.J., J. Goksoyr, A.K. Bej, and R.M. Atlas.** 1988. Recovery of DNA from soils and sediments. *Appl. Environ. Microbiol.* *54*:2908-2915.
19. **Wilson, I.G.** 1997. Inhibition and facilitation of nucleic acid amplification. *Appl. Environ. Microbiol.* *63*:3741-3751.
20. **Suzuki, M., M.S. Rappe, and S.J. Giovannoni.** 1998. Kinetic bias in estimates of coastal picoplankton community structure obtained by measurements of small-subunit rRNA gene PCR amplicon length heterogeneity. *Appl. Environ. Microbiol.* *64*:4522-4529.
21. **Suzuki, M.T. and S.J. Giovannoni.** 1996. Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. *Appl. Environ. Microbiol.* *62*:625-630.
22. **Ritchie, N.J., M.E. Schutter, R.P. Dick, and D.D. Myrold.** 2000. Use of length heterogeneity PCR and fatty acid methyl ester profiles to characterize microbial communities in soil. *Appl. Environ. Microbiol.* *66*:1668-1675.
23. **Muyzer, G., E.C. de Waal, and A.G. Uitterlinden.** 1993. Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA. *Appl. Environ. Microbiol.* *59*:695-700.
24. **Klappenbach, J.A., P.R. Saxman, J.R. Cole, and T.M. Schmidt.** 2001. rrndb: the ribosomal RNA operon copy number database. *Nucleic Acids Res.* *29*:181-184.
25. **Gurtler, V. and V.A. Stanisich.** 1996. New approaches to typing and identification of bacteria using the 16S-23S rDNA spacer region. *Microbiology* *142*:3-16.
26. **Fogel, G.B., C.R. Collins, J. Li, and C.F. Brunk.** 1999. Prokaryotic genome size and SSU rDNA copy number: estimation of microbial relative abundance from a mixed population. *Microb. Ecol.* *38*:93-113.

Received 7 November 2002; accepted 14 January 2003.

### Address correspondence to:

Dr. Craig S. Criddle  
Department of Civil and Environmental  
Engineering  
Terman Engineering Center, Rm B-9  
Stanford University  
Stanford, CA 94305, USA  
e-mail: criddle@stanford.edu