

## MINI-REVIEW

# The tragedy of the uncommon: understanding limitations in the analysis of microbial diversity

Stephen J Bent<sup>1,2</sup> and Larry J Forney<sup>1,2</sup>

<sup>1</sup>Initiative for Bioinformatics and Evolutionary Studies, University of Idaho, Moscow, ID, USA  
and <sup>2</sup>Department of Biological Sciences, University of Idaho, Moscow, ID, USA

**Molecular microbial community analysis methods have revolutionized our understanding of the diversity and distribution of bacteria, archaea and microbial eukaryotes. The information obtained has adequately demonstrated that the analysis of microbial model systems can provide important insights into ecosystem function and stability. However, the terminology and metrics used in macroecology must be applied cautiously because the methods available to characterize microbial diversity are inherently limited in their ability to detect the many numerically minor constituents of microbial communities. In this review, we focus on the use of indices to quantify the diversity found in microbial communities, and on the methods used to generate the data from which those indices are calculated. Useful conclusions regarding diversity can only be deduced if the properties of the various methods used are well understood. The commonly used diversity metrics differ in the weight they give to organisms that differ in abundance, so understanding the properties of these metrics is essential. In this review, we illustrate important methodological and metric-dependent differences using simulated communities. We conclude that the assessment of richness in complex communities is futile without extensive sampling, and that some diversity indices can be estimated with reasonable accuracy through the analysis of clone libraries, but not from community fingerprint data.**

*The ISME Journal* (2008) 2, 689–695; doi:10.1038/ismej.2008.44; published online 8 May 2008

**Subject Category:** microbial population and community ecology

**Keywords:** diversity; communities; diversity indices

## Background

The ability of researchers to quantify diversity and test many important hypotheses regarding patterns and processes in microbial communities hinges on their ability to characterize the diversity and distribution of microbes in a wide range of habitats. An accurate assessment of the composition of these communities permits us to characterize spatial and temporal patterns of diversity, as well as responses to changing environmental conditions, perturbations and treatments. A necessary first step, however, is to reach a consensus regarding what inferences can and cannot be made regarding microbial community structure given the inherent limitations imposed by the methods now used in studies of microbial community ecology and by the extraordinary diversity found in most habitats (Dahllof, 2002; Ward, 2002).

Contemporary studies on the diversity of prokaryotes often employ methods based on the analysis of nucleic acid sequences, especially those of 16S rRNA genes. These have gained favor because they allow investigators to detect and quantify phylotypes that are difficult to culture and thereby obtain a more comprehensive assessment of diversity than was previously possible. This has led to an improved understanding of the extraordinary richness of prokaryotic biodiversity (Woese, 1987; DeLong and Pace, 2001; Wellington *et al.*, 2003; Oremland *et al.*, 2005). Indeed, the extent of prokaryotic diversity in most habitats is almost incomprehensible, with complex habitats containing an estimated  $10^4$ – $10^6$  species in a single gram (Dykhuizen, 1998; Torsvik *et al.*, 1998; Ovreas *et al.*, 2003; Gans *et al.*, 2005) and the Earth's biosphere containing more than  $10^{30}$  individuals (Whitman *et al.*, 1998) and an untold number of species.

The challenges faced in efforts to characterize this extraordinary diversity are compounded by the fact that the observed and inferred rank-abundance distributions for most communities show a long tail of numerically minor species or phylotypes

---

Correspondence: Larry J Forney, Department of Biological Sciences, Life Sciences South, Rm. 252, University of Idaho, Moscow, ID 83843, USA.  
E-mail: lforney@uidaho.edu

(Preston, 1948; MacArthur, 1960; May, 1975; Tokeshi, 1993; Curtis and Sloan, 2004). In other words, most communities are dominated by a small number of species whereas the vast majority of populations are quite uncommon. This characteristic of prokaryotic communities has sparked debate over the most appropriate mathematical distribution for modeling community composition (Hughes, 1986; Wilson, 1991; Tokeshi, 1993), and exposed the limitations of current methods (Dunbar *et al.*, 1999; Curtis *et al.*, 2002; Zhou, 2003; Hewson and Fuhrman, 2004; Osborne *et al.*, 2006), which are by and large unable to detect the many uncommon members of these communities. Perhaps most worrisome is the tendency of many investigators to simply ignore the uncommon, and draw conclusions regarding microbial community diversity based solely on the number and rank abundance of numerically common organisms. This approach, where investigators tacitly acknowledge the existence of uncommon organisms (but do not consider them further), at best constitutes an innocent oversimplification that can still allow valid inferences to be drawn, but at worst it leads to the misinterpretation of data and faulty conclusions.

The cultivation-independent molecular methods now commonly used to characterize microbial diversity can be grouped into two basic categories: (a) methods based on the phylogenetic analysis of cloned nucleic acid sequences, and (b) a family of methods collectively, and colloquially, known as 'community fingerprinting'. The data produced by sequencing and fingerprinting methods differ due to the reliance of the latter on proxy information (for example, restriction sites or %G + C content) rather than full sequence data (Abdo *et al.*, 2006), and both kinds of methods have problems and biases that have been previously noted (Reysenbach *et al.*, 1992; Farrelly *et al.*, 1995; Suzuki and Giovannoni, 1996; Hansen *et al.*, 1998; Frostegard *et al.*, 1999; Maarit Niemi *et al.*, 2001; Qiu *et al.*, 2001; Baker *et al.*, 2003; Crosby and Criddle, 2003). As clone library and fingerprinting methods can generally be performed using the same DNA extraction procedure and comparable primers, the two categories of methods can effectively be used to analyze the same pool of 16S rRNA amplicons. The effect of these DNA extraction and PCR biases on different methods is therefore similar, and can be discounted to some extent for purposes of assessing differences between community analysis methods.

Here, we focus specifically on differences between methods to assess the richness and rank-abundance of phylotypes as measured by several diversity indices and methods explicitly used to calculate similarity measures are not discussed. Although one could debate the merits of diversity indices, the reality is that they are commonly used summary statistics. As we continue to gain in understanding of the extant microbial variety and distribution, microbial ecologists will continue to

need to express the observed patterns using summary statistics. If diversity indices are to be used, they should be used with a full understanding of how the method used can affect the index value and the subsequent interpretation of the data.

## Quantifying diversity

Diversity is a general ecological concept that has various shades of meaning and many metrics, both of which are often used loosely. Calculation of a diversity index involves distilling information contained in community analysis data into a single numerical value that reflects the number and relative abundance of phylotypes in a single community. The utility of diversity metrics rests in the fact that they capture information about biodiversity by summarizing species richness and evenness into a single real number. Researchers must, therefore, classify the observed diversity into 'kinds' before calculating many of the commonly used metrics of diversity. This classification step is particularly problematic in the microbial world, as asexual reproduction and horizontal gene transfer across species boundaries leads to ill-defined species within which consistent and meaningful boundaries are difficult to draw.

Most investigators nowadays rely on phylogenetic approaches for the classification of microbial diversity (Staley, 2006) as bacterial species are currently defined on the basis of a rather odd phenetic-genotypic species concept where multiple characters are used to group related organisms (Stackebrandt *et al.*, 2002), and the organisms must first be cultured. In practice, DNA sequence polymorphisms, often in the small subunit rRNA gene, are used to classify diversity in terms of phylotypes or operational taxonomic units (OTUs) that are defined in an *ad hoc* manner. By doing so, investigators can classify organisms into discrete categories, which enables them to quantify prokaryotic diversity using conventional diversity or similarity indices. Because most studies on microbial diversity do not measure species diversity *per se*, we eschew the use of this term, favoring phylotypes or OTUs instead.

The three most widely used diversity indices are richness, the Simpson index (Simpson, 1949) and the Shannon–Weaver index (Shannon and Weaver, 1949). Any of these three indices can be used to compare multiple communities to each other, but the values for different indices cannot be compared to each other in a simple, intuitive way. The cause of this incomparability is rooted in the intrinsically different meanings of each index. Richness is simply the number of phylotypes present, whereas the Simpson index reflects the probability that any two organisms sampled will be the same phylotype. The Shannon–Weaver index is an information theory measure of the entropy, or nonredundancy, of a system such that a community in which every

organism is different would have minimal redundancy and therefore maximum entropy.

Measures of phylotype richness are independent of whether phylotypes are rare or common in a community. As none of the existing molecular microbial ecology methods capture more than a small proportion of the total richness in most microbial communities, richness must be estimated. The methods used to do this include nonparametric estimators such as Chao1 and ACE (Hughes *et al.*, 2001), extrapolation of accumulation curves (Soberon and Llorente, 1993) and parametric estimation based on model fitting (Curtis *et al.*, 2006). Nonparametric estimators and extrapolation of accumulation curves rely on counting individuals sampled from a community, and, therefore, their application is largely limited to data from the analysis of clone libraries. On the other hand, parametric methods of data analysis that use observed relative abundance data to choose a model distribution can also be used with microbial community fingerprint data. The choice of model distributions used to estimate the underlying community structure can radically affect the resulting richness estimate, but other information can be used to inform this choice (Gans *et al.*, 2005). All of these methods suffer from uncertainty that often ranges several orders of magnitude, thus, greatly reducing the reliability of richness estimates (Hong *et al.*, 2006). Other diversity indices, such as the Shannon and Simpson indices, can be estimated more accurately because rare phylotypes generally have a smaller relative numerical impact.

When two or more community fingerprints or clone libraries are compared, it is tempting to conclude that ones with more OTUs are more diverse, but this is not necessarily true. Changes in the rank-abundance can alter the number of detectable phylotypes without changing the actual phylotype richness in the underlying community. Estimates based on postulated rank-abundance distributions can mitigate this problem (Dunbar *et al.*, 2002; Narang and Dunbar, 2004).

## A conceptual framework

Hill (Hill, 1973) has proposed a conceptual framework that provides a useful way to describe and quantify biological diversity. He defines different 'orders' ( $q$ ) of diversity ( $D$ ) that summarize information about the number and relative abundances of species or phylotypes. Hill states that  ${}^qD$  can be regarded as the 'effective number of species,' or phylotypes, present in a sample for a given order  $q$ , and that diversity indices represented by different values of  $q$  are distinguished by the weighting applied to phylotypes that differ in abundance. This family of diversity indices has the property that for all values of  $q$ , they are equal to phylotype richness when all phylotypes are equally abundant. The most generally useful diversity indices are of orders  $q = 0$ ,

1 and 2 (Jost, 2006). Richness, or number of species or phylotypes, corresponds to a diversity index of order  $q = 0$ . For calculation of richness, all phylotypes are weighted evenly, as the relative abundance is not considered, yielding  ${}^0D = S$ , where  $S$  is the number of phylotypes in the community. The exponentially transformed Shannon–Weaver index corresponds to  $q = 1$ , with phylotypes weighted proportionally to their relative abundance. The formula for this index is  ${}^1D = \exp(-\sum(p_i \ln p_i))$ , where  $p_i$  is the proportional abundance of the  $i$ th phylotype. Finally, the reciprocal Simpson index calculated with replacement (due to the large population size) represents  $q = 2$ , with phylotypes weighted by the square of their relative abundance, yielding  ${}^2D = 1/(\sum p_i^2)$ . Another index in this family,  ${}^\infty D = 1/p_{i(\max)}$ , expresses the reciprocal of the proportional abundance of the most abundant species ( $p_{i(\max)}$ ), which is known as the Berger–Parker index (Berger and Parker, 1970). This has recently found use as a parameter in a richness estimator based on a log-normal, model (Curtis *et al.*, 2002; Loisel *et al.*, 2006). This set of diversity indices provides a consistent theoretical framework for assessing the behavior of the index values with different data sets. Each of these indices reflects different properties of the community and, hence, the choice of index must be based on the questions being asked in a particular study.

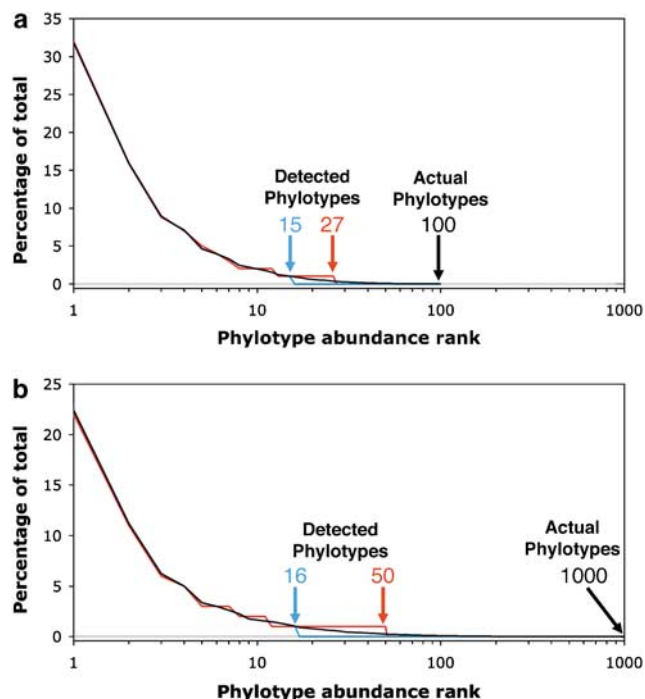
## Sampling vs screening communities

The cultivation-independent methods commonly used to quantify diversity or compare communities differ from each other in a fundamentally important way, as suggested in recent studies (Hartmann and Widmer, 2006). When using methods based on the phylogenetic analysis of cloned nucleic acid sequences, individual DNA molecules are *sampled* from a PCR product pool, cloned and then sequenced. By sampling a community through analysis of a clone library one can obtain information about some of the organisms found in the tail of a rank-abundance distribution. In contrast, community fingerprinting methods determine the absolute quantity of different amplicons using some analytical method. In T-RFLP (terminal restriction fragment length polymorphism analysis) of 16S rRNA genes (Liu *et al.*, 1997), which we will use as our example, the sizes and the fluorescence intensities of labeled DNA fragments are quantified by capillary gel electrophoresis. If the quantity of a given DNA fragment is below a chosen threshold value, it is indistinguishable from noise and discarded (Abdo *et al.*, 2006). This amounts to *screening* samples to determine the presence or absence of phylotypes. Numerically rare phylotypes are generally not detected by community fingerprinting methods. The distinction between sampling and screening communities becomes important whenever there are

many organisms representing diverse phylotypes that transpose individually below the detection limit of an assay, but collectively above it.

To illustrate the implications of differences between methods that sample the diversity in a community and those that screen diversity, we constructed two computer simulated log-normally distributed communities. The phylotypes constituting these communities were simply 'kinds' of organismal variety (or OTUs), defined in a way that permits them to be distinguished equally well using clone library and fingerprint methods. By doing so we simulated a scenario in which community structure was primarily driven by large genetic differences rather than microheterogeneity. This effectively constitutes a best-case scenario for fingerprint analysis. The hypothetical communities contained either 100 or 1000 log-normally distributed phylotypes that span 18  $\log_2$  octaves ( $N_T = 10^8$ ,  $\sigma = 15.95$ ,  $S_0 = 10$ ,  $N_0 = 62100$ ) and 27  $\log_2$  octaves ( $N_T = 10^8$ ,  $\sigma = 15.95$ ,  $S_0 = 100$ ,  $N_0 = 2737$ ), respectively, with the phylotypes evenly spaced within each octave. Analyses of clone libraries were simulated by multiplying the relative proportion of each species in the community by the number of clones analyzed and rounding the result to the nearest integer. The difference between the sum of these integers and the size of the clone library was made up by adding the required number of single clones from among the species that were previously not sampled. Microbial community fingerprints were simulated by converting the relative proportion of each species in the community into a T-RFLP peak height value, with all peak heights below the threshold discarded from the analysis.

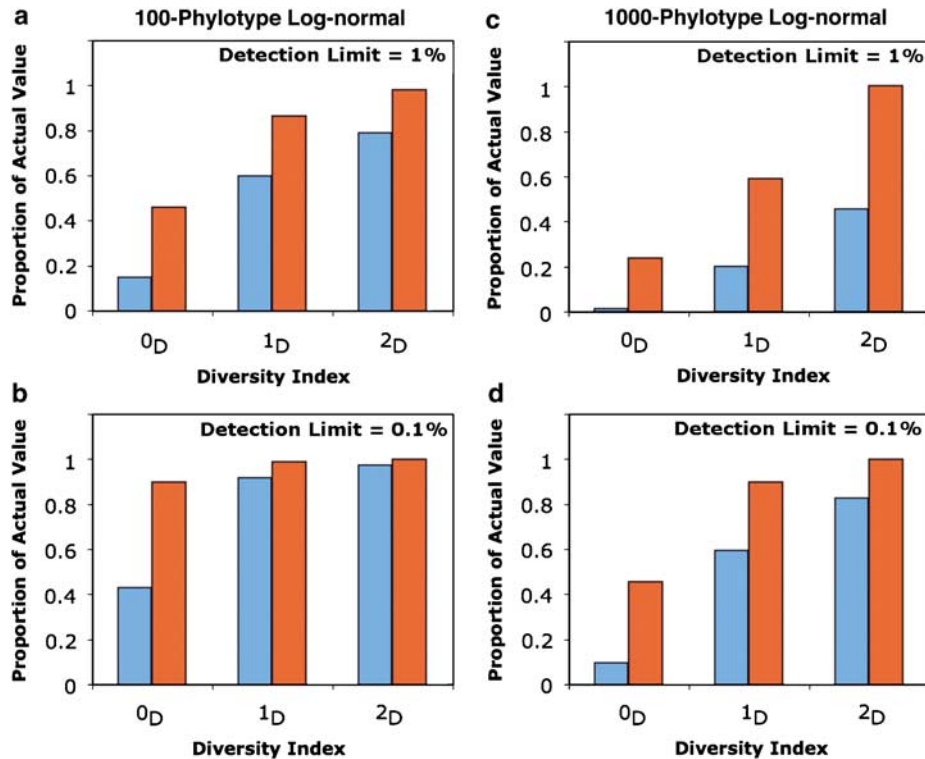
We simulated the expected values for the  ${}^0D$ ,  ${}^1D$  and  ${}^2D$  diversity indices obtained from sampling diversity through the analysis of clone libraries that differ in size, and screening diversity using community fingerprints. In the latter we imposed different detection thresholds. When a 1% detection limit was used the community fingerprints revealed 15 and 16 phylotypes in the 100- and 1000-phylotype communities, respectively. In contrast, simulations of clone library analyses detected 27 and 50 phylotypes, respectively (Figure 1). Thus, for both communities a greater number of phylotypes were detected through the analysis of clone libraries, which reflects the power of sampling communities as opposed to screening diversity on the basis of community fingerprints. Likewise, the simulated clone library analyses consistently yielded more accurate values for  ${}^0D$ ,  ${}^1D$  and  ${}^2D$  diversity indices than did community fingerprints (Figure 2). Of course as the detection limit of a diversity screening assay is lowered, the ability to detect minor phylotypes increases (Figure 2). As the detection limit was lowered from 1 to 0.1%, the accuracy of the inferred values of diversity indices substantially increased. The inverse Simpson ( ${}^2D$ ) index was found to be most robust and less affected



**Figure 1** Comparison of the actual (black lines) and observed rank abundances of phylotypes detected by analysis of microbial community fingerprints (blue lines) and clone libraries of 16S rRNA genes (red lines). The numbers of phylotypes observed in microbial community fingerprints (blue) and clone libraries (red) are shown. The simulations assumed the detection threshold for community fingerprint analysis was 1% of the total fluorescence, and that 100 clones were sampled from each library. The hypothetical communities contained either (a) 100 or (b) 1000 log-normally distributed phylotypes.

by assay sensitivity or the absolute level of diversity in a community. However, accurate estimates of richness ( ${}^0D$ ) in communities with high diversity required greater sensitivity than current fingerprint and clone library methods typically provide. The use of nonparametric estimators of diversity, such as Chao1, produced the most accurate estimate of  ${}^0D$  from clone library data. This implies that estimates of richness in microbial communities are unreliable unless highly intensive sampling is employed.

As diversity indices are often used to compare communities and assess relative diversity, we calculated the true ratios of diversity indices and compared them to those based on data from simulated analyses of the communities described above. The true ratio of the  ${}^0D$  value of the community with 100 phylotypes was 0.10 times that of the community with 1000 phylotypes ( $100/1000 = 0.1$ ), whereas the ratio of the  ${}^1D$  indices was 0.30 ( $13.8/46.6 = 0.30$ ), and that of the  ${}^2D$  indices was 0.51 ( $6.8/13.3 = 0.51$ ). Neither analytical method yielded accurate estimates of  ${}^0D$  or  ${}^1D$  ratios (Table 1). Likewise, the ratio of  ${}^2D$  estimates based on community fingerprint data was also far from the true value. The only instance in which the ratio of  ${}^2D$  indices closely approximated the true value was when data from the analysis of clone libraries were



**Figure 2** The ratio of observed to actual values for the diversity indices  ${}^0D$ ,  ${}^1D$  and  ${}^2D$  when the diversity is screened by community fingerprints (blue) or sampled from a clone library of 16S rRNA genes (red). For the  ${}^0D$  index, we used the number of observed phylotypes for the community fingerprint values and the Chao1 estimator for the clone library values. The data are from simulated communities with 100 (a and b), and 1000 (c and d) phylotypes. The upper graphs (a and c) are based on a 1% detection limit, corresponding to a library of 100 clones or a fingerprint detection threshold of 1% of the total fluorescence. The lower graphs (b and d) are based on a 0.1% detection limit, corresponding to a library of 1000 clones or a fingerprint detection threshold of 0.1% of the total fluorescence.

**Table 1** A comparison of analytical methods used to estimate the true ratio of diversity indices based on the simulated analysis of communities using clone libraries and community fingerprints

Diversity index	True ratio <sup>b</sup>	Ratio from simulations <sup>a</sup>	
		Clone library	Community fingerprints
${}^0D$	0.10	0.52 <sup>c</sup>	0.94
${}^1D$	0.30	0.43	0.87
${}^2D$	0.51	0.50	0.89

<sup>a</sup>Estimated ratio of diversity indices based on the simulated analysis of diversity using clone libraries or community fingerprints with a detection threshold of 1%.

<sup>b</sup>True ratio of diversity indices for a community with 100 phylotypes divided by that for a community with 1000 phylotypes as described elsewhere in the text.

<sup>c</sup>Actual ratio of phylotypes detected. Using the Chao1 nonparametric estimator for both clone libraries, this ratio is 0.19.

used. This analysis suggests that efforts to compare communities using diversity indices estimated from community fingerprinting or the analysis of clone libraries may lead to misleading conclusions, and this is largely because the ratios calculated are ultimately subject to the same limitations as estimates of the indices themselves.

## Summary

The tragedy of the uncommon is that they are often ignored. Although numerically dominant organisms are likely to be responsible for the majority of metabolic activity and energy flux in a system (Tilman, 1982), it is well known that uncommon organisms serve as a reservoir of genetic and functional diversity (Yachi and Loreau, 1999; Nandi *et al.*, 2004), often play key roles in ecosystems (Phillips *et al.*, 2000; Louda and Rand, 2002), and can become numerically important if environmental conditions change. Ideally, the presence and abundance of the uncommon but important organisms would be reflected in the values of diversity indices. But due to the distorted lenses through which we observe microbial communities they usually are not, and so diversity indices need to be applied judiciously in studies on microbial community ecology and biodiversity.

The simulations of community analysis performed here illustrate that different methods of examining community structure can produce radically different metrics of diversity, even when many of the well-documented biases of molecular methods are excluded from consideration. One way to increase the accuracy of diversity metrics is to

choose metrics such as the  $^2D$  reciprocal Simpson's index, which is comparatively insensitive to numerically minor constituents (Lande *et al.*, 2000). However, this insensitivity comes with a trade-off, in that the calculated diversity measures are more sensitive to errors and biases that affect the apparent abundance of numerically dominant members of communities. These problems can be ameliorated by advances in sequencing technology, novel modeling approaches and new diversity metrics. These advances allow for more intensive sampling of communities, new ways of assessing the composition of those communities, and better use of the data (Curtis *et al.*, 2002). For now, the use of multiple methods in concert, such as fingerprinting of a large set of samples followed by cluster analysis and then clone library analysis of a subset (Zhou *et al.*, 2007), can provide an optimal balance between the resources required and information gained.

## Acknowledgements

We thank Dr Eva Top, Jacob Pierson and Margaret-Mary McEwen for their critical reviews of this paper and helpful suggestions. This work was supported by an NIH Center of Biomedical Research Excellence grant (PT20 RR 16448) from the National Center for Research Resources to LJJ and SJB was supported by a fellowship from the Subsurface Science Research Initiative of the Inland Northwest Research Alliance.

## References

- Abdo Z, Schütte UME, Bent SJ, Williams CJ, Forney LJ, Joyce P. (2006). Statistical methods for characterizing diversity of microbial communities by analysis of terminal restriction fragment length polymorphisms of 16S rRNA genes. *Environ Microbiol* **8**: 929–938.
- Baker GC, Smith JJ, Cowan DA. (2003). Review and re-analysis of domain-specific 16S primers. *J Microbiol Methods* **55**: 541–555.
- Berger W, Parker FL. (1970). Diversity of planktonic Foraminifera in deep sea sediments. *Science* **168**: 1345–1347.
- Crosby LD, Criddle CS. (2003). Understanding bias in microbial community analysis techniques due to rrn operon copy number heterogeneity. *Biotechniques* **34**: 790–794, 796, 798 passim.
- Curtis TP, Head IM, Lunn M, Woodcock S, Schloss PD, Sloan WT. (2006). What is the extent of prokaryotic diversity? *Philos Trans R Soc Lond B Biol Sci* **361**: 2023–2037.
- Curtis TP, Sloan WT, Scannell JW. (2002). Estimating prokaryotic diversity and its limits. *Proc Natl Acad Sci USA* **99**: 10494–10499.
- Curtis TP, Sloan WT. (2004). Prokaryotic diversity and its limits: microbial community structure in nature and implications for microbial ecology. *Curr Opin Microbiol* **7**: 221–226.
- Dahllöf I. (2002). Molecular community analysis of microbial diversity. *Curr Opin Biotechnol* **13**: 213–217.
- DeLong EF, Pace NR. (2001). Environmental diversity of bacteria and archaea. *Syst Biol* **50**: 470–478.
- Dunbar J, Barns SM, Ticknor LO, Kuske CR. (2002). Empirical and theoretical bacterial diversity in four Arizona soils. *Appl Environ Microbiol* **68**: 3035–3045.
- Dunbar J, Takala S, Barns SM, Davis JA, Kuske CR. (1999). Levels of bacterial community diversity in four arid soils compared by cultivation and 16S rRNA gene cloning. *Appl Environ Microbiol* **65**: 1662–1669.
- Dykhuizen DE. (1998). Santa Rosalia revisited: why are there so many species of bacteria? *Antonie Van Leeuwenhoek* **73**: 25–33.
- Farrelly V, Rainey FA, Stackebrandt E. (1995). Effect of genome size and rrn gene copy number on PCR amplification of 16S rRNA genes from a mixture of bacterial species. *Appl Environ Microbiol* **61**: 2798–2801.
- Frostegard A, Courtois S, Ramišse V, Clerc S, Bernillon D, Le Gall F *et al.* (1999). Quantification of bias related to the extraction of DNA directly from soils. *Appl Environ Microbiol* **65**: 5409–5420.
- Gans J, Wolinsky M, Dunbar J. (2005). Computational improvements reveal great bacterial diversity and high metal toxicity in soil. *Science* **309**: 1387–1390.
- Hansen MC, Tolker-Nielsen T, Givskov M, Molin S. (1998). Biased 16S rDNA PCR amplification caused by interference from DNA flanking the template region. *FEMS Microbiol Ecol* **26**: 141–149.
- 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.
- Hewson I, Fuhrman J. (2004). Richness and diversity of bacterioplankton species along an estuarine gradient in Moreton Bay, Australia. *Appl Environ Microbiol* **70**: 3425–3433.
- Hill MO. (1973). Diversity and evenness: a unifying notation and its consequences. *Ecology* **54**: 427–432.
- Hong SH, Bunge J, Jeon SO, Epstein SS. (2006). Predicting microbial species richness. *Proc Natl Acad Sci USA* **103**: 117–122.
- Hughes JB, Hellmann JJ, Ricketts TH, Bohannan BJ. (2001). Counting the uncountable: statistical approaches to estimating microbial diversity. *Appl Environ Microbiol* **67**: 4399–4406.
- Hughes RG. (1986). Theories and models of species abundance. *Am Nat* **128**: 897–899.
- Jost L. (2006). Entropy and diversity. *Oikos* **113**: 363–375.
- Lande R, DeVries PJ, Walla TR. (2000). When species accumulation curves intersect: implications for ranking diversity using small samples. *Oikos* **89**: 601–605.
- Liu WT, Marsh TL, Cheng H, Forney LJ. (1997). Characterization of microbial diversity by determining terminal restriction fragment length polymorphisms of genes encoding 16S rRNA. *Appl Environ Microbiol* **63**: 4516–4522.
- Loisel P, Harmand J, Zemb O, Latrille E, Lobry C, Delgenes J-P *et al.* (2006). Denaturing gradient electrophoresis (DGE) and single-strand conformation polymorphism (SSCP) molecular fingerprintings revisited by simulation and used as a tool to measure microbial diversity. *Environ Microbiol* **8**: 720–731.
- Louda SM, Rand TA. (2002). Native Thistles: expendable or integral to ecosystem resistance to invasion?. In: Kareiva P and Levin SA (eds). *The Importance of Species: Perspectives on Expendability and Triage*. Princeton University Press: Princeton NJ. pp 5–15.

- Maarit Niemi R, Heiskanen I, Wallenius K, Lindstrom K. (2001). Extraction and purification of DNA in rhizosphere soil samples for PCR-DGGE analysis of bacterial consortia. *J Microbiol Methods* **45**: 155–165.
- MacArthur RH. (1960). On the relative abundance of species. *Am Nat* **94**: 25–36.
- May RM. (1975). Patterns of species abundance and diversity. In: Diamond JM, Cody ML (eds). *Ecology and Evolution of Communities*. Belknap Press: Cambridge, MA. pp 81–120.
- Nandi S, Maurer JJ, Hofacre C, Summers AO. (2004). Gram-positive bacteria are a major reservoir of Class 1 antibiotic resistance integrons in poultry litter. *Proc Natl Acad Sci USA* **101**: 7118–7122.
- Narang R, Dunbar J. (2004). Modeling bacterial species abundance from small community surveys. *Microb Ecol* **47**: 396–406.
- Oremland RS, Capone DG, Stolz JF, Fuhrman J. (2005). Whither or wither geomicrobiology in the era of ‘community metagenomics’. *Nat Rev Microbiol* **3**: 572–578.
- Osborne CA, Rees GN, Bernstein Y, Janssen PH. (2006). New threshold and confidence estimates for terminal restriction fragment length polymorphism analysis of complex bacterial communities. *Appl Environ Microbiol* **72**: 1270–1278.
- Ovreas L, Daae FL, Torsvik V, Rodriguez-Valera F. (2003). Characterization of microbial diversity in hypersaline environments by melting profiles and reassociation kinetics in combination with terminal restriction fragment length polymorphism (T-RFLP). *Microb Ecol* **46**: 291–301.
- Phillips CJ, Harris D, Dollhopf SL, Gross KL, Prosser JI, Paul EA. (2000). Effects of agronomic treatments on structure and function of ammonia-oxidizing communities. *Appl Environ Microbiol* **66**: 5410–5418.
- Preston FW. (1948). The commonness, and rarity, of species. *Ecology* **29**: 254–283.
- Qiu X, Wu L, Huang H, McDonel PE, Palumbo AV, Tiedje JM *et al.* (2001). Evaluation of PCR-generated chimeras, mutations, and heteroduplexes with 16S rRNA gene-based cloning. *Appl Environ Microbiol* **67**: 880–887.
- Reysenbach A, Giver LH, Wickham GS, Pace NR. (1992). Differential amplification of rRNA genes by polymerase chain reaction. *Appl Environ Microbiol* **58**: 3417–3418.
- Shannon CE, Weaver W. (1949). *The Mathematical Theory of Communication*. University of Illinois Press: Urbana, IL.
- Simpson EH. (1949). Measurement of diversity. *Nature* **163**: 688.
- Soberon J, Llorente J. (1993). The use of species accumulation functions for the prediction of species richness. *Conserv Biol* **7**: 480–488.
- Stackebrandt E, Frederiksen W, Garrity GM, Grimont PA, Kampf P, Maiden MC *et al.* (2002). Report of the *ad hoc* committee for the re-evaluation of the species definition in bacteriology. *Int J Syst Evol Microbiol* **52**: 1043–1047.
- Staley JT. (2006). The bacterial species dilemma and the genomic-phylogenetic species concept. *Philos Trans R Soc Lond B Biol Sci* **361**: 1899–1909.
- 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.
- Tilman D. (1982). *Resource Competition and Community Structure*. Princeton University Press: Princeton, NJ.
- Tokeshi M. (1993). Species abundance patterns and community structure. In: Begon M, Fitter AH (eds). *Advances in Ecological Research*. Academic Press: London, UK. pp 112–186.
- Torsvik V, Daae FL, Sandaa RA, Ovreas L. (1998). Novel techniques for analysing microbial diversity in natural and perturbed environments. *J Biotechnol* **64**: 53–62.
- Ward BB. (2002). How many species of prokaryotes are there? *Proc Natl Acad Sci USA* **99**: 10234–10236.
- Wellington EM, Berry A, Krsek M. (2003). Resolving functional diversity in relation to microbial community structure in soil: exploiting genomics and stable isotope probing. *Curr Opin Microbiol* **6**: 295–301.
- Whitman WB, Coleman DC, Wiebe WJ. (1998). Prokaryotes: the unseen majority. *Proc Natl Acad Sci USA* **95**: 6578–6583.
- Wilson JB. (1991). Methods for fitting dominance/diversity curves. *J Veg Sci* **2**: 35–46.
- Woese CR. (1987). Bacterial evolution. *Microbiol Rev* **51**: 221–271.
- Yachi S, Loreau M. (1999). Biodiversity and ecosystem productivity in a fluctuating environment: the insurance hypothesis. *Proc Natl Acad Sci USA* **96**: 1463–1468.
- Zhou J. (2003). Microarrays for bacterial detection and microbial community analysis. *Curr Opin Microbiol* **6**: 288–294.
- Zhou X, Brown CJ, Abdo Z, Davis CC, Hansmann MA, Joyce P *et al.* (2007). Differences in the composition of vaginal microbial communities found in healthy Caucasian and black women. *ISME J* **1**: 121–133.