New approaches to analyzing microbial biodiversity data

Brendan JM Bohannan* and Jennifer Hughes†

Modern molecular techniques have revealed an extraordinary diversity of microorganisms, most of which are as yet uncharacterized. This poses a major challenge to microbial ecologists: how can one compare the microbial diversity of different environments when the vast majority of microbial taxa are usually unknown? Three statistical approaches developed by ecologists and evolutionary biologists — parametric estimation, nonparametric estimation and community phylogenetics — are proving to be promising tools to meet this challenge. The combination of these tools with molecular biology techniques allow the rigorous estimation and comparison of microbial diversity in different environments.

Addresses
*Department of Biological Sciences, Stanford University, Stanford, CA 94305-5020, USA
†Department of Ecology and Evolutionary Biology, Brown University, Providence, RI, USA
Correspondence: Brendan JM Bohannan
e-mail: bohannan@stanford.edu

Current Opinion in Microbiology 2003, **6**:282–287

This review comes from a themed issue on
Techniques
Edited by Jo Handelsman and Kornelia Smalla

1369-5274/03/$ – see front matter

DOI 10.1016/S1369-5274(03)00055-9

Abbreviation
OTU operational taxonomic unit

Introduction

Microbial ecologists, like all ecologists, are often interested in the factors that regulate community diversity across temporal and spatial scales, the impact of human activity on this diversity and the consequences of this diversity for ecosystem processes. In addition, efficient exploitation of microorganisms (as sources of novel pharmaceuticals, for example) requires knowledge of the distribution of microbial diversity. Investigating these patterns requires that diversity be compared among different environments. Because of their extremely high abundance and diversity [1], however, this task has proven to be very difficult for microbes.

Microbiologists have recently rediscovered that ecologists and evolutionary biologists studying the diversity of macroorganisms have developed a range of approaches to document and analyze environmental diversity patterns, many of which may be applicable to microorganisms [2,3,4•]. Although most of this literature concentrates on comparing species richness (i.e. the number of distinct species), most if not all of these statistics are applicable to taxonomic levels other than species. This is fortunate for microbial ecologists, as defining prokaryotic species is difficult [5,6]. Out of necessity, microbial diversity studies usually examine the diversity of operational taxonomic units (OTUs). Although some OTU definitions try to capture a species-like unit, one can ask valid questions about biodiversity at any level, as long as the OTU definition is clear and consistent.

Community diversity is more than richness, however. It also includes evenness, or the relative abundance of OTUs [2]. A further aspect of community diversity is the genetic relatedness of the OTUs present. Indeed, because microbial ecologists often have phylogenetic data about their communities, they are in a unique position to compare evolutionary diversity (i.e. genetic distinctness among communities).

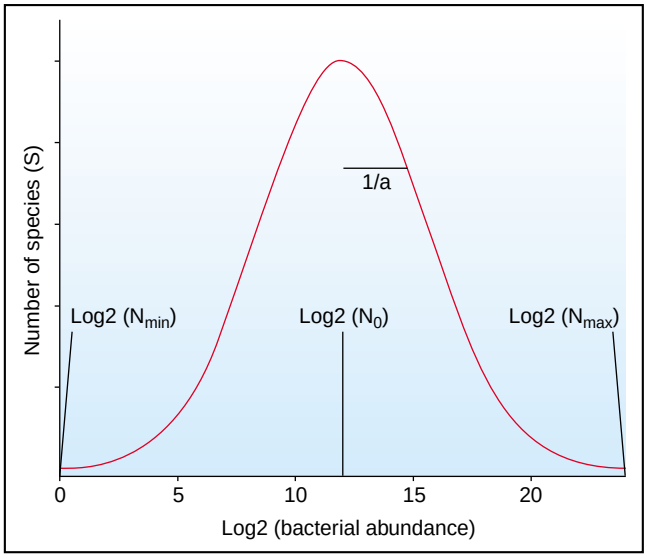
Several approaches to compare diversity have been applied recently to molecular studies of microbial diversity. In this review, we discuss three of these new (or, more accurately, newly rediscovered) approaches: parametric estimation, nonparametric estimation and community phylogenetics. The first two approaches are used to compare OTU richness among environments. The third approach compares evolutionary diversity among environments. Each approach has its own unique strengths and limitations.

Parametric methods: using species-abundance models to estimate diversity

Parametric approaches estimate the number of unobserved OTUs in a community by fitting sample data to models of relative OTU abundance (traditionally species abundance). These models include the lognormal ([7], Figure 1]) and Poisson lognormal [8], among others [9]. The advantage of this approach is that, given a few simplifying assumptions, one can use the model to estimate diversity from relatively small samples of individuals from a given environment. Thus, this approach could be ideal for estimating the diversity of hyperdiverse organisms such as microbes.

However, there are several impediments to using parametric approaches to estimate microbial diversity. The primary impediment is that there are not large datasets of microbial diversity data to support the use of any of the many competing abundance models. In the absence of empirical data, only theoretical arguments can be made for the appropriateness of some models over others. For

Figure 1



The lognormal species-abundance curve. This distribution assumes that a few taxa contain many individuals, a few taxa contain a few individuals and that most taxa contain moderate numbers of individuals. N_{max} is the number of individuals in the most abundant species, N_{min} is the number of individuals in the least abundant species, and N_0 is the modal species abundance. The total diversity (S_T) is the area under the species-abundance curve. The width of the curve is inversely proportional to the spread parameter *a* ($a = 2\ln 2\sigma^2$, where σ^2 is the standard deviation of the distribution) (Copyright © 2002 National Academy of Sciences, USA. Used with permission.)

example, Tom Curtis, John Dunbar and their respective colleagues [10••,11•] have recently argued that the growth dynamics and large population sizes, characteristic of many microorganisms, could lead to lognormal distributions of diversity.

A second problem with using parametric approaches is that even if compelling arguments can be made in favor of a particular model, the models still require large datasets to evaluate the distribution parameters, unless simplifying (and untested) assumptions are made. For instance, Curtis *et al.* [10••] developed two methods for parameterizing the lognormal curve that do not rely on large datasets, although they do require some simplifying assumptions. These approaches only rely on estimates of the total number of individuals in a sample (N_T) and the number of individuals of the most abundant OTU (N_{max}; Figure 1). An attractive feature of this method is that in theory these two parameters can be measured directly; the former can be measured relatively simply by microscopy and the latter with quantitative fluorescent *in situ* hybridization (FISH). Alternatively, one could estimate these values from a clone library and extrapolate to the larger sample.

The validity of these methods rest, of course, on the appropriateness of the lognormal distribution as a model of microbial diversity. The theoretical arguments in favor of this model are compelling, but it should be noted that attempts to determine empirically if bacterial diversity is indeed lognormally distributed have failed (e.g. [11•]), most likely because of the small proportion of diversity sampled. There is also some evidence that other models, such as a uniform distribution [12] or Fisher's negative binomial [13], may also be appropriate for microorganisms.

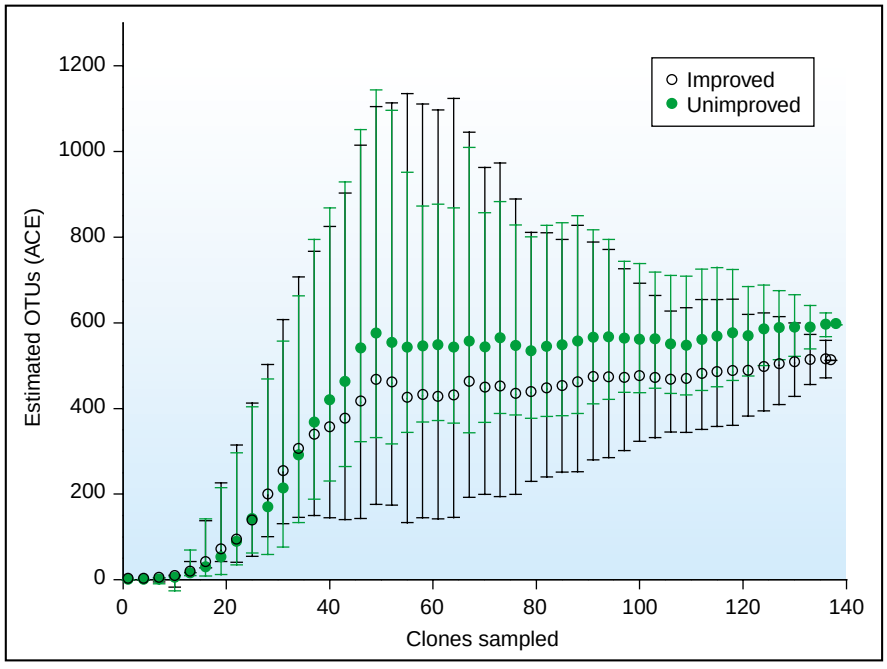
Nonparametric methods: using detection probabilities to estimate diversity

In contrast to parametric approaches, nonparametric approaches estimate OTU richness from small sample sizes without assuming a particular OTU abundance model [14••]. Many of these estimates are adapted from mark-release-recapture (MRR) statistics for estimating the size of animal populations [15,16]. Such approaches consider the proportion of OTUs that have been observed before ('recaptured') to those that are observed only once. The probability of detecting an OTU more than once will be higher in samples from less diverse communities. By contrast, samples from more diverse communities are predicted to contain fewer recaptures.

For instance, the Chao1 estimator uses the number of singletons (OTUs represented by only one individual in a sample) and doubletons (OTUs represented by two individuals in a sample) to estimate the diversity of a given environment. This estimator is particularly useful because a closed-form solution for the variance of the Chao1 estimator has been derived [17]. This variance is an estimate of the precision of Chao1; that is, it estimates the variance of diversity estimates that one expects if many different samples were drawn from the same community. The Chao1 variance can be used to calculate confidence intervals about the Chao1 estimate, and thus can be used to determine whether a difference in diversity between two samples (and the environments from which they are taken) is statistically significant. Hughes *et al.* [14••] used this approach to compare the microbial diversities of the human mouth and gut, two grassland soils under different agricultural management (Figure 2), and several aquatic mesocosms differing in nitrogen input. The Chao1 method has also been used for estimating diversity from environmental genomics data [18].

One disadvantage of nonparametric approaches is that they rely on estimates of the relative abundance of OTUs. Many studies have revealed that sampling biases can accompany genetic surveys of microbial diversity. For example, the abundance of PCR-amplified genes might not reflect the relative abundance of template DNA because of differences in primer binding and elongation efficiency [19–21]. Certainly, steps should be taken to reduce these biases when possible. Yet as long as the biases are similar across samples, robust comparisons using nonparametric estimators can still be made [14••].

Figure 2



Chao1 estimates of bacterial OTU richness in improved (black open circles) and unimproved (green solid circles) soil as a function of sample size. Error bars are 95% confidence intervals and were calculated with the variance formula derived by Chao (see text). The black lines are error bars for the improved sample. The green lines are error bars for the unimproved sample. (Copyright © 2001, the American Society for Microbiology. Used with permission.)

Another disadvantage of nonparametric estimators is that they only provide a lower bound of OTU diversity. The nonparametric approach relies solely on the information from the OTUs observed and, unlike parametric approaches, does not assume a distribution of relative OTU abundances. As a result, these methods do not account for very rare classes of OTUs. Thus for bacterial communities, nonparametric estimators will tend to underestimate OTU diversity.

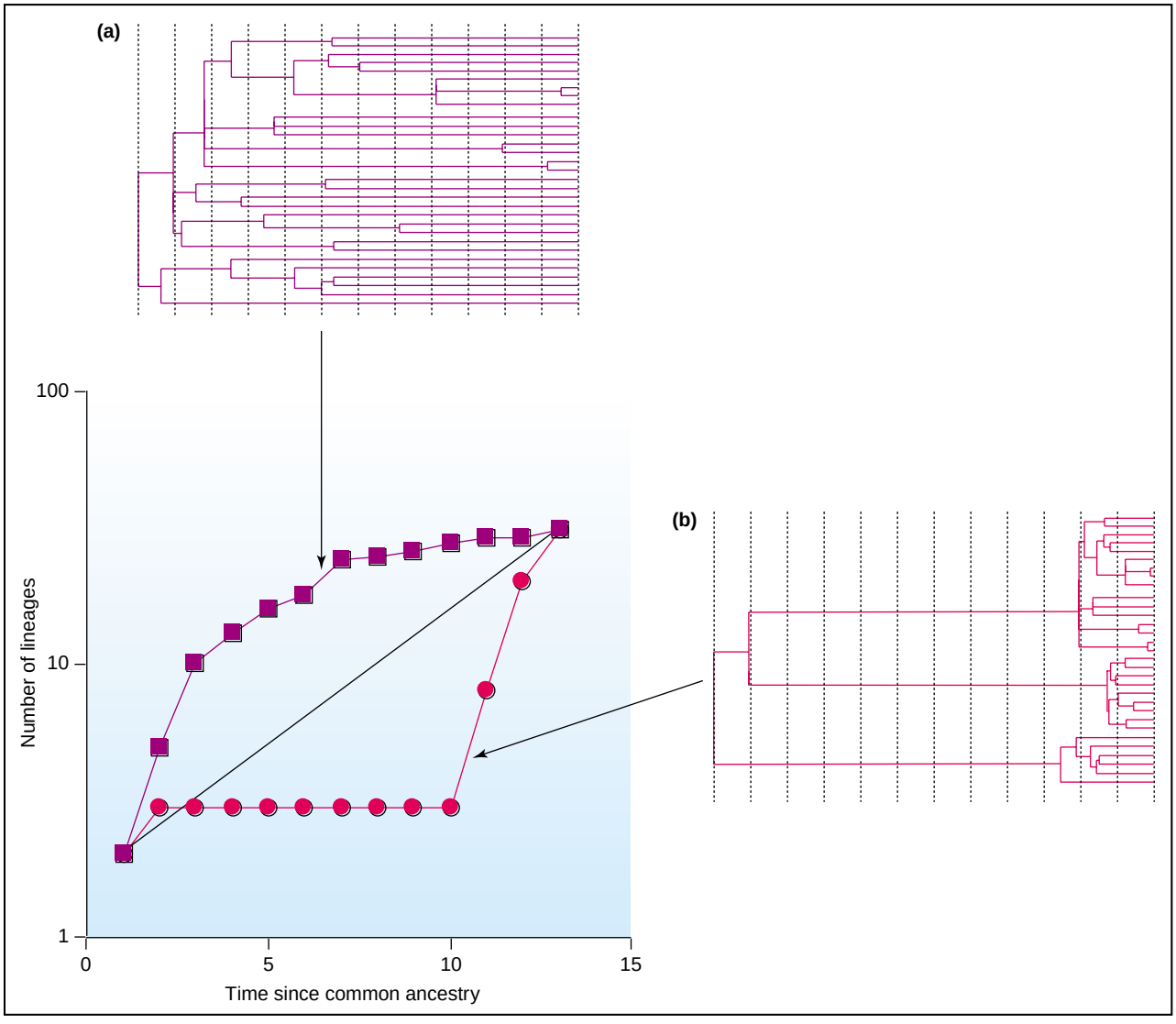
Using trees to see the forest: comparing phylogenetic diversity

Most microbial diversity data published recently are molecular in nature; that is, they consist primarily of sequence data from a given target gene (most commonly a ribosomal gene) obtained from an environmental sample. To be analysed by most of the approaches described above, this genetic diversity must first be grouped into taxonomic groups (OTUs). Although this grouping allows use of some potentially powerful analytical tools, it also has drawbacks. First, there is not a common criterion for grouping microbial sequence data into taxons, making comparisons of results based on OTUs difficult across studies. Second, by grouping sequences into OTUs, potentially valuable information concerning relatedness is often lost.

By contrast, phylogenetic approaches incorporate this information to compare genetic diversity among communities. For example, population biologists have developed methods for comparing the topology (shape) of phylogenetic trees for different communities. Martin [22••] used one of these techniques to compare microbial communities from different soil samples. He compared these communities by comparing lineage-per-time plots, graphs of the number of lineages present on a phylogenetic tree as a function of time. Constant rates of birth and extinction of lineages are predicted to yield exponential lineage-per-time plots. Concave departures from an exponential relationship indicate an overabundance of highly divergent lineages (i.e. a very genetically diverse community); convex departures indicate an overabundance of closely related lineages (i.e. a less genetically diverse community) (Figure 3). A more genetically diverse community would be predicted to be more phenotypically (i.e. functionally) diverse, if phenotypic variation is positively correlated with genetic variation (as has been shown for plants, animals and some microbes) [22••].

Differences in tree topology can further be used to make inferences regarding the processes important in the assembly of communities [4•]. For example, communities structured primarily by competitive exclusion are predicted to be less closely related than expected by chance, whereas communities structured primarily through habitat filtering are predicted to be more closely related than expected by chance. Webb [23] has developed two indices for comparisons of tree topology that allow such comparisons to be made: the net relatedness index (NRI)

Figure 3



Lineage-per-time plots constructed by counting the numbers of lineages present at different time intervals. Time intervals are equal lengths and were defined arbitrarily. **(a)** In this tree, there is an excess of highly divergent lineages, yielding a concave lineage-per-time plot. **(b)** In this tree, there is an excess of closely related lineages, yielding a convex lineage-per-time plot. Constant rates of birth and extinction of lineages yield exponential lineage-per-time plots, the signature of which is indicated by the solid straight line.

and the nearest taxa index (NTI). These two indices measure different aspects of 'clumpedness' or clustering of taxa in a given sample. These indices have been successfully applied to microbial diversity data (MC Horner-Devine and BMJ Bohannan, unpublished data).

The major disadvantage of the phylogenetic approaches described above is that they assume that patterns of genetic diversity observed in samples directly reflect patterns in the environment sampled. This may or may not be true. Thus, these approaches may tell you something about the sample analyzed (a clone library, for example) but not necessarily the environment sampled. This is in contrast to the parametric and nonparametric approaches discussed above, which attempt to extrapolate from sample to environment.

Application of approaches to microbial diversity data

How might these three classes of approaches contribute to our understanding of microbial diversity? To answer this, we consider one dataset in detail. McCaig *et al.* [24] sequenced two 16S rDNA clone libraries, one from an unfertilized pasture and the other from a fertilized pasture. They partially sequenced a total of 281 clones, grouped these sequences into OTUs (using a criterion of 97%

Table 1

Analyses of the diversity and composition of fertilized and unfertilized grassland soils.

Analytical approach	Diversity (number of OTUs)	Significant effect of fertilization on diversity?	Significant effect of fertilization on composition?	References
Descriptive indices	113–117	No	No[‡]	[24]
Parametric extrapolation	6300	N/A[*]	N/A	[10••]
Nonparametric extrapolation	467–590	No	N/A	[14••]
Nonparametric methods & phylogenetics	N/A	N/A	No	[25•]
Genetic diversity & phylogenetics	N/A	No[†]	Yes[§]	[22••]

[*]Parametric extrapolation can be used to statistically compare diversity; however, it has not yet been done for this data set. [†]Diversity was measured as total genetic diversity. [‡]Similarity indices were not significantly different; however, some differences in composition were evident upon inspection of the data. [§]Lineage-per-time plots were not significantly different; however, tests of covariation between unique sequences and phylogeny were significant.

sequence similarity), and then described the data using a variety of traditional ecological diversity and similarity indices. On the basis of this analysis, they concluded that the libraries were not substantially different in either diversity (i.e. OTU richness and evenness) or composition (i.e. the identity of the taxons present), although some differences in composition (most notably among the α-*Proteobacteria*) were noted upon inspection of the data (Table 1). The new techniques described above can augment this analysis in three ways. First, they can be used to estimate the richness of the environments from which these samples were taken, allowing one to determine how representative a sample is of the environment. For example, Curtis *et al.* [10••] used a parametric approach to estimate the total diversity in this grassland soil by fitting a lognormal model to the data of McCaig *et al.* [24]. They estimated that the diversity of this soil was approximately 6300 OTUs, indicating that the 113–117 OTUs observed by McCaig *et al.* represented less than 2% of the total soil diversity (Table 1). By contrast, Hughes *et al.* [14••] used the nonparametric Chao1 estimator to estimate a total soil diversity ranging from 467 OTUs (in the fertilized soil) to 590 OTUs (in the unfertilized soil), suggesting that McCaig *et al.* sampled approximately 20% of the total diversity (Table 1). This discrepancy could be due to the underestimation characteristic of nonparametric estimators, or to an overestimation by the parametric estimator caused by departure of the diversity data from lognormality. Regardless, the dataset of McCaig *et al.* represents a severe undersampling of the total diversity.

The second contribution that these new approaches can make is to add statistical robustness to comparisons of microbial communities. For example, Hughes *et al.* [14••] demonstrated that the decrease in diversity observed in the fertilized soil (relative to the unfertilized) was not statistically significant (Table 1, Figure 2). Furthermore, they demonstrated that if this decrease was indeed real (and the lack of significance was due to a lack of statistical power) then a sampling of at least 250 clones from each soil would be necessary to have sufficient power to detect this difference. Third, these techniques can increase the resolution of community comparisons. For example, Martin [22••] demonstrated using phylogenetic methods that although total genetic diversity of the fertilized and unfertilized soils examined by McCaig *et al.* did not differ significantly, the compositions of these libraries were indeed significantly different. He was also able, using phylogenetic methods, to identify the specific clade in which the majority of these compositional differences occurred, increasing the resolution of the comparison.

Conclusions

The approaches described above offer windows to different aspects of microbial diversity. However, each approach has particular strengths and limitations, as well as different requirements for input data. These factors, as well as the research question being addressed, must be considered when choosing an analytical approach. The most robust alternative is to use a combination of approaches to analyze microbial data, relying, for example, on parametric approaches to make absolute estimates of environmental diversity, nonparametric approaches to compare the diversity of different environments, and phylogenetic approaches to provide genetic comparisons of environmental communities. Combined with culture-independent molecular techniques for surveying the diversity of microorganisms, these statistical approaches make possible the detailed and rigorous description of the distribution of microbial life.

Acknowledgements

We are grateful to Jo Handelsman for the invitation to contribute this article, and to MC Horner-Devine for helpful comments on a previous draft of this manuscript. This work was supported by the National Science Foundation (DEB-0108556).

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Pace NR: **A molecular view of microbial diversity and the biosphere**. *Science* 1997, **276**:734-740.

2. Magurran AE: *Ecological Diversity and its Measurement*. Princeton, New Jersey: Princeton University Press; 1988.

3. Colwell RK, Coddington JA: **Estimating terrestrial biodiversity through extrapolation**. *Phil Trans Royal Soc Lond B* 1994, **345**:101-118.

4. • Webb CO, Ackerly DD, McPeek MA, Donoghue MJ: **Phylogenies and community ecology**. *Annual Review of Ecology and Systematics* 2002, **33**:475-505.
A review of the emerging field of 'community phylogenetics' — the use of phylogenetic analyses to make inferences about community ecology. The review does not mention microorganisms, but the approaches discussed have potential applications in microbial ecology.

5. Stackebrandt E: **Summary of Workshop- — Part 2 microbial species.** In *Microbial Diversity in Time and Space*, Edited by R Colwell. New York: Plenum Press; 1996.

6. Staley JT: **Biodiversity: are microbial species threatened?** *Curr Opin Biotechnol* 1997, **8**:340-345.

7. Preston FW: **The commonness and rarity, of species**. *Ecology* 1948, **29**:254-283.

8. Bulmer MG: **On fitting the Poisson lognormal distribution to species abundance data**. *Biometrics* 1974, **30**:101-110.

9. Hubbell SP: *The Unified Neutral Theory of Biodiversity and Biogeography*. Princeton, New Jersey: Princeton University Press; 2001.

10. •• Curtis TP, Sloan WT, Scannell JW: **Estimating prokaryotic diversity and its limits**. *Proc Natl Acad Sci USA* 2002, **99**:10494-10499.
An elegant application of the lognormal taxon abundance model to microbial diversity data. The authors offer a unique approach for parameterizing the model, tailored to microbial diversity data.

11. • Dunbar J, Barns SM, Ticknor LO, Kuske CR: **Empirical and theoretical bacterial diversity in four Arizona soils**. *Appl Environ Microbiol* 2002, **68**:3035-3045.
An application of the lognormal taxon abundance model to estimate the coverage (the proportion of total diversity sampled) of microbial clone libraries.

12. Zhou J, Xia B, Treves DS, Wu LY, Marsh TL, O'Neill RV, Palumbo AV, Tiedje JM: **Spatial and resource factors influencing high microbial diversity in soil**. *Appl Environ Microbiol* 2002, **68**:326-334.

13. Kroes I, Lepp PW, Relman DA: **Bacterial diversity within the human subgingival crevice**. *Proc Natl Acad Sci USA* 1999, **96**:14547-14552.

14. •• Hughes JB, Hellmann JJ, Ricketts TH, Bohannan BJ: **Counting the uncountable: statistical approaches to estimating microbial diversity**. *Appl Environ Microbiol* 2001, **67**:4399-4406.
A review of statistical approaches to analysing microbial operational taxonomic unit data with an emphasis on nonparametric techniques.

15. Seber GAF: *The Estimation of Animal Abundance and Related Parameters*. London: Griffin; 1973.

16. Krebs CJ: *Ecological Methodology*. New York: Harper and Row; 1989.

17. Chao A: **Estimating the population size for capture-recapture data with unequal catchability**. *Biometrics* 1987, **43**:783-791.

18. Breitbart M, Salamon P, Andresen B, Mahaffy JM, Segall AM, Mead D, Azam F, Rohwer F: **Genomic analysis of uncultured marine viral communities**. *Proc Natl Acad Sci USA* 2002, **99**:14250-14255.

19. Reysenbach A-L, Giver LJ, Wickham GS, Pace NR: **Differential amplification of rRNA genes by polymerase chain reaction**. *Appl Environ Microbiol* 1992, **58**:3417-3418.

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

21. Speksnijder A, Kowalchuk GA, De Jong S, Kline E, Stephen JR, Laanbroek HJ: **Microvariation artefacts introduced by PCR and cloning of closely related 16S rRNA gene sequences**. *Appl Environ Microbiol* 2001, **67**:469-472.

22. •• Martin AP: **Phylogenetic approaches for describing and comparing the diversity of microbial communities**. *Appl Environ Microbiol* 2002, **68**:3673-3682.
An excellent review of phylogenetic techniques (originally developed by population geneticists) that might be useful for the estimation and comparison of microbial diversity. The author applies several of these techniques to published datasets to illustrate their utility.

23. Webb CO: **Exploring the phylogenetic structure of ecological communities: an example for rain forest trees.** *Am Nat* 2000, 145-155.

24. McCaig AE, Glover A, Prosser JI: **Molecular analysis of bacterial community structure and diversity in unimproved and improved upland grass pastures**. *Appl Environ Microbiol* 1999, **65**:1721-1730.

25. • Singleton DR, Furlong MA, Rathbun SL, Whitman WB: **Quantitative comparisons of 16S rRNA gene sequence libraries from environmental samples**. *Appl Environ Microbiol* 2001, **67**:4374-4376.
The authors combine a nonparametric coverage index with a measure of phylogenetic distance to create a similarity index for microbial clone libraries.