

Estimating Bacterial Diversity from Environmental DNA: A Maximum Likelihood Approach

Frederick Cohan¹, Danny Krizanc², and Yun Lu²

¹ Department of Biology,
Wesleyan University, Middletown, CT, 06459
fcohan@wesleyan.edu

² Department of Mathematics and Computer Science,
Wesleyan University, Middletown, CT, 06459
dkrizanc@wesleyan.edu, ylu@wesleyan.edu

Abstract. The ability to measure bacterial diversity is a prerequisite for the systematic study of bacterial biogeography and ecology. In this paper we describe a method of estimating diversity from an environmental sample of DNA and apply it to data taken from samples from the Sargasso Sea. Our approach combines the coverage depth method of Venter *et al.* [2] and the contig spectrum approach of Angly *et al.* [4], but uses maximum likelihood to recover the diversity rather than using hand-fit models as in [2]. We assume four species abundance distributions, then maximize the likelihood of fitting the coverage depth at different positions of the consensus sequence provided in the Sargasso Sea sample. The resulting estimates match well with those obtained using less mathematically rigorous approaches.

1 Introduction

The extent of prokaryote diversity has been hotly debated and rightly so. But measuring prokaryote diversity is not a trivial task [1]. There are two general approaches to estimate microbial diversity that have been applied in the past: nonparametric methods, which use detection probabilities to estimate diversity; and parametric methods, which use species abundance models to estimate diversity. Each approach has its particular strengths and limitations as well as different requirements for the input data [5].

Nonparametric methods use detection probabilities to estimate diversity. In contrast to parametric approaches, these approaches estimate OTU (operational taxonomic units) richness (the number of species in a community) from small sample sizes without assuming a particular OTU abundance model [14]. Such approaches consider the proportion of OTUs that have been observed before to those observed only once. The probability of detecting an OTU more than once will be higher in samples from less diverse communities. 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 ([19], [20] and [21]). Another disadvantage is that they only provide a lower bound of OTU diversity. These methods do not account for very rare classes of OTUs, thus for bacterial communities, nonparametric estimators will tend to underestimate OTU diversity.

Parametric methods use species abundance models to estimate diversity. These models include the lognormal [7] and Poisson lognormal [8] among others [9]. The advantage of this method is that one can use the model to estimate diversity using relatively small samples of individuals from a given environment. However, there are several impediments to using parametric approaches to estimate microbial diversity. One limit is that there are no large data sets of microbial diversity data to support the use of any of the many competing abundance models. Another limit is that even if a compelling argument can be made in favor of a particular model, the models still require large data sets to evaluate the distribution parameters unless simplifying assumptions are made.

In this paper, we use a parametric method that combines the coverage depth method of Venter *et al.* [2] and the contig spectrum approach of Angly *et al.* [4], but uses maximum likelihood to recover the diversity rather than using hand-fit models as in the case of [2]. We use our method to estimate the number of bacterial species represented in a sample of DNA drawn from water from the Sargasso Sea. We assume four possible abundance distributions for our data, then try to recover the true diversity by maximizing the likelihood of fitting the coverage depth at every position of the consensus sequence of an environmental DNA sample.

1.1 Venter's Coverage Depth Method

A method based upon coverage depth was introduced by Venter *et al.* [2]. The method is applied to the results of whole metagenome shotgun sequencing of an environmental sample of DNA. For a single genome analysis, assembly coverage depth should be approximated by a Poisson distribution; for multiple genome analysis, assembly coverage depth should be approximated by a mixture of Poisson distributions. The empirical distribution of coverage depth at every position in the full set of assemblies was computed, then compared with hand-constructed mixtures of Poisson distributions. An excellent fit can be obtained; and to the extent that a limited range of mixtures give acceptable fits, this model may be used to estimate the diversity of the bacteria represented in the DNA extracted. However, there are obvious challenges to genome assembly in the environmental context. Additionally, the method is based on hand fitting the observed depth of coverage to a theoretical model of assembly progress for a sample corresponding to a mixture of organisms at different abundances, and is therefore ad hoc and likely less reliable.

1.2 PHACCS Methods

A method based upon the contig spectrum was introduced by Angly *et al.* [4] for estimating diversity in viral communities. They call it PHACCS (PHAge

Communities from Contig Spectrum). PHACCS uses a modified version of the Lander-Waterman model to predict a contig spectrum from assumed population parameters [4]. Since there are several genotypes in this modified model, an assumption about their underlying distribution within the community in terms of abundance has to be made. A number of distributions have been suggested including the power law, logarithmic, exponential, broken stick, niche preemption and lognormal distributions. However, predictions by PHACCS are dependent on the quality of the contig spectrum input. As a general rule, the higher the contig degree is, the better the estimations are, since the model fitting is done over a larger number of points. Additionally, the present implementation of the Lander-Waterman model assumes that all DNA fragments and all the genotypes have the same size. For these reasons, PHACCS estimates as well as ours should be only considered approximations.

2 A Maximum Likelihood Approach

We introduce our method, which essentially combines the coverage depth method and the contig spectrum method assuming a variety of abundance distributions. We calculate the likelihood of fitting the coverage depth at every position of the sequence, then recover the diversity of the sample by maximizing the log-likelihood. We use the following notation:

- S : Size of all reads sequenced from the sample
- N : Size of all the genomes in the sample
- T : Total number of genomes in the sample
- g : Average length of a genome in the sample
- k : Number of abundance levels
- a_i : the i -th abundance level ($i = 1, 2, 3, \dots$)
- c_i : Average coverage given by the species with abundance a_i
- p_i : the proportion of the species with abundance a_i to all the species in the sample
- m_i : Number of different species with abundance a_i
- K_j : Coverage depth observed at position j of the assembly consensus sequence ($j = 1, 2, \dots, J$)
- n_l : Number of positions on the assembly consensus sequence with coverage depth K_l ($l = 0, 1, \dots, L$)
- J : Total number of positions on the assembly consensus sequence
- L : Total number of different coverage depth levels observed
- M : Total number of different genomes in the sample

Let x denote the fraction $\frac{S}{N}$. We have the following relations by definition:

$$J = \sum_{i=1}^k m_i * g = g * \sum_{i=1}^k m_i \quad (1)$$

and

$$T = \frac{N}{g}. \quad (2)$$

The relation among T , a_i and m_i can be expressed by the equation

$$T = \sum_{i=1}^k a_i * m_i. \quad (3)$$

The diversity M is

$$M = \sum_{i=1}^k m_i. \quad (4)$$

Since the choice of abundance distribution is very important in estimating the diversity, we will deal with the problem separately by assuming four models of the possible abundance distributions for our sample: discrete distribution with a fixed number of abundance levels, power law distribution, broken stick distribution and log-normal distribution.

2.1 Model A: Discrete Distribution

Assuming discrete distribution of abundance with a fixed number of abundance levels, sequencing of an environmental sample with more than one species should result in the sequence coverage depth reflecting a mixture of Poisson distributions. Let K_j be the coverage depth at the j -position of the assembly sequence and $P(K_j)$ be the probability of coverage depth K_j on the consensus sequence. Then the expected number of positions on the consensus sequence with coverage depth K_j is $L * P(K_j) = M * g * P(K_j)$. Let $P(K_j, a_i)$ be the probability of coverage depth K_j on the consensus sequence given by the species with abundance a_i . Then the expected number of positions on the consensus sequence with coverage depth K_j is given by $\sum_{i=1}^k m_i * g * P(K_j, a_i)$. Hence we have

$$M * g * P(K_j) = \sum_{i=1}^k m_i * g * P(K_j, a_i). \quad (5)$$

By the definition of p_i , we have

$$p_i = \frac{m_i}{\sum_{i=1}^k m_i} = \frac{m_i}{M} \quad (6)$$

and

$$\sum_{i=1}^k p_i = 1. \quad (7)$$

Dividing both sides of (5) by $M * g$, we have

$$P(K_j) = \sum_{i=1}^k \left(\frac{m_i}{M}\right) * P(K_j, a_i) = \sum_{i=1}^k p_i * P(K_j, a_i). \quad (8)$$

Assuming that the coverage depth K_j given by the species with abundance a_i satisfies the Poisson distribution, we have

$$P(K_j, a_i) = \frac{e^{-y} y^{K_j}}{K_j!} \tag{9}$$

where y is the average coverage given by the species with abundance a_i .

We assume that $y = c_i$ where c_i is the total size of fragments from species with abundance a_i divided by the genome length of the species. Then c_i is given by

$$c_i = \frac{(\frac{a_i * g}{N}) * S}{g} = (\frac{S}{N}) * a_i = x * a_i. \tag{10}$$

Then (8) can be rewritten as

$$P(K_j) = \sum_{i=1}^k p_i * (\frac{e^{-x*a_i} (x * a_i)^{K_j}}{K_j!}). \tag{11}$$

Hence the likelihood of fitting the coverage depth at every position of the consensus sequence is given by

$$P = \prod_{j=1}^J P(K_j). \tag{12}$$

Suppose there are n_l positions with the same coverage K_l for $0 \leq l \leq L$. Then we can express (12) as

$$P = \prod_{l=0}^L P(K_l)^{n_l} = \prod_{l=0}^L [\sum_{i=1}^k p_i * (\frac{e^{-x*a_i} (x * a_i)^{K_l}}{K_l!})]^{n_l} \tag{13}$$

where the a_i 's and p_i 's are the parameters.

Now we can maximize the likelihood P with respect to the parameters a_i and p_i where $a_i > 0$, $0 < p_i < 1$ for $i = 1, 2, \dots, k - 1$ and $p_k = 1 - \sum_{i=1}^{k-1} p_i$. Hence m_i (the number of species with abundance a_i) can be solved using the equations (3) and (6)

$$m_i = \frac{(p_i * T)}{\sum_{i=1}^k (p_i * a_i)}. \tag{14}$$

Thus the diversity M is

$$M = \sum_{i=1}^k m_i = \sum_{i=1}^k \frac{(p_i * T)}{\sum_{i=1}^k (p_i * a_i)}. \tag{15}$$

Since P is extremely small, for practical reasons, we maximize the logarithm of the likelihood instead of the likelihood. The formula for $Log[P]$ is

$$Log[P] = \sum_{l=0}^L [n_l * Log(\sum_{i=1}^k p_i * \frac{e^{-x*a_i} (x * a_i)^{K_l}}{K_l!})] \tag{16}$$

where the a_i 's and p_i 's are the parameters.

2.2 Model B: Power Law Distribution

This model assumes the power law distribution of abundance. In this case, the diversity M itself is a parameter of the likelihood, so we obtain the diversity directly by maximizing the likelihood (or log-likelihood).

Let a_i denote the abundance level of the species i , then the formula for the power law distribution is

$$a_i = a * i^{-b} \tag{17}$$

for $1 \leq i \leq M$.

The parameter a represents the abundance of the most abundant genotype; b is a parameter related to the evenness (the relative abundance of individuals within a species) and M is the number of different genotypes in the community, which is the diversity we want to estimate.

A similar calculation to the above yields

$$\text{Log}[P] = \sum_{l=0}^L [n_l * \text{Log} \sum_{i=1}^M (\frac{1}{M} * \frac{e^{-x*a*i^{-b}} (x * a * i^{-b})^{K_j}}{K_j!})] \tag{18}$$

where the a, b and M are the parameters.

2.3 Model C: Broken Stick Distribution

This model assumes the broken stick distribution of abundance. Let a_i denote the abundance level of the species i , then the formula for the broken stick distribution is

$$a_i = \frac{T}{M} \sum_{q=i}^M \frac{1}{q} \tag{19}$$

for $1 \leq i \leq M$ where T is the total number of genomes in the sample.

A similar calculation to the above yields

$$\text{Log}[P] = \sum_{l=0}^L [n_l * \text{Log} \sum_{i=1}^M (\frac{1}{M} * \frac{e^{-x*\frac{T}{M} \sum_{q=i}^M \frac{1}{q}} (x * \frac{T}{M} \sum_{q=i}^M \frac{1}{q})^{K_j}}{K_j!})] \tag{20}$$

where M is the only parameter.

2.4 Model D: Log-Normal Distribution

The model assumes the Log-normal distribution of abundance. Let a_i denote the abundance level of the species i , then the formula for the lognormal distribution is

$$a_i = \frac{e^{m_i \sigma}}{\sum_{j=1}^M e^{m_j \sigma}} \tag{21}$$

where $m_i = \frac{M}{\sqrt{2\pi}} (e^{-t_i^2/2} - e^{-t_i+1^2/2})$.

Here $t_1 = -\infty$, $t_{i+1} = \sqrt{2} \operatorname{erf}^{-1}[\frac{2}{M} + \operatorname{erf}(\frac{t_i}{\sqrt{2}})]$ and $t_{M+1} = +\infty$ for $1 \leq i \leq M$ where erf is the error function ($\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$) and erf^{-1} its inverse. A similar calculation to the above yields

$$\operatorname{Log}[P] = \sum_{l=0}^L [n_l * \operatorname{Log}(\sum_{i=1}^M (\frac{1}{M} * \frac{e^{-x * \frac{e^{m_i \sigma}}{\sum_{j=1}^M e^{m_j \sigma}}}} (x * \frac{e^{m_i \sigma}}{\sum_{j=1}^M e^{m_j \sigma}})^{K_j}) K_j!))] \quad (22)$$

where σ, M are the parameters.

3 Sargasso Sea Data

We tested our models on data obtained from Sargasso Sea water samples. The cell counts imply approximately one billion cells per liter, while the relative abundance of the most common organism ranged from 3% to 12% of the total. The total sequences from samples were pooled and assembled to provide a single master assembly. The empirical distribution of coverage depth at every position in the full set of assemblies was computed.

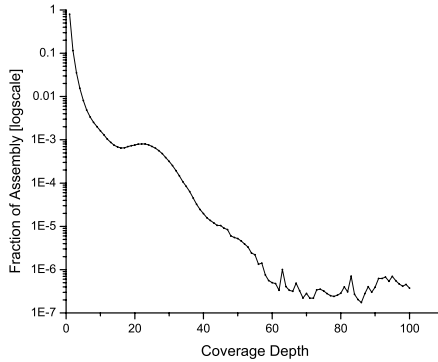


Fig. 1. Fraction of consensus sequence f_l with coverage depth K_l

The total size of the reads in Sargasso Sea water sample is

$$S = 1.66 * 10^6 * 818 = 1.36 * 10^9. \quad (23)$$

and the total number of base pairs N in the samples is

$$N = 2 * 10^{11} * (170 + 340 + 250 + 170) * 2 * 10^6 / 200 = 1.86 * 10^{18} \quad (24)$$

with average genome size $g = 2 * 10^6 \text{bp/genome}$.

Hence the total number of genomes T is

$$T = \frac{N}{(2 * 10^6)} = 9.3 * 10^{11}. \quad (25)$$

Let x be the fraction of S to N , then

$$x = \frac{S}{N} = 7.31 * 10^{-10}. \quad (26)$$

Fig. 1 shows the fraction of the consensus sequence f_l as the y -axis and the coverage depth K_l as the x -axis. (Data provided by A. Halpern of the Venter Institute.)

4 Our Results

We assume the four species abundance distributions: models A through D, then maximize the likelihood of fitting the coverage depth at different positions of the consensus sequence provided in the Sargasso Sea Sample. **Table 1** lists the corresponding results.

Table 1. Estimates assuming different distribution model for Sargasso Sea data

Model	Abundance Levels	M	$MaxLog[P]$
Model A	one	909.4	$-1.2994 * 10^9$
	two	924.0	$-0.9984 * 10^9$
	three	997.3	$-0.9983 * 10^9$
	four	951.0	$-0.9797 * 10^9$
Model B		3504.0	$-0.9888 * 10^9$
Model C		871.0	$-1.0996 * 10^9$
Model D		917.0	$-1.0287 * 10^9$

Model A

We assume that the abundance distribution is a discrete distribution with a fixed number of abundance levels. We estimate the diversity by assuming different discrete abundance levels. If we assume two abundance levels, then by calculating the log-likelihood with $100 \leq M \leq 20,000$, we generate **Fig. 2** with the log-likelihood $Log[P]$ as the y -axis and the diversity M as the x -axis, after approximately $M = 924.0$, the value of $Log[P]$ is decreasing.

Model B

We assume that the abundance distribution is the power law distribution. In general, we don't know the abundance of the most abundant genotype, but we can still estimate the diversity by making a one parameter as well as b and M . By calculating the log-likelihood with $100 \leq M \leq 20,000$, we generate **Fig. 3** with the log-likelihood $Log[P]$ as the y -axis and the diversity M as the x -axis. After approximately $M = 3504$, the value of $Log[P]$ decreases slowly.

Model C

We assume that the abundance distribution is the broken stick distribution. By calculating the log-likelihood with $100 \leq M \leq 20,000$, we generate **Fig. 4** with

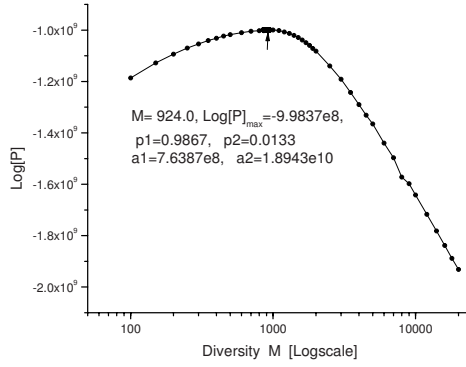


Fig. 2. $\text{Log}[P]$ at different diversity M

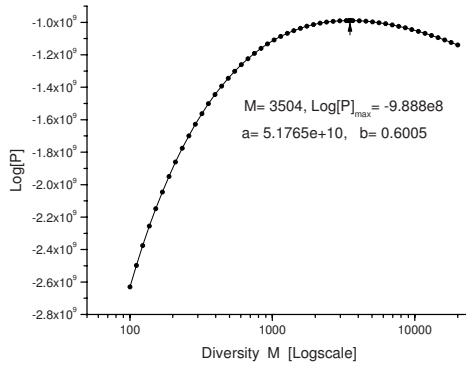


Fig. 3. $\text{Log}[P]$ at different diversity with $a = 5.0469 * 10^{10}$ and $b = 0.6001$

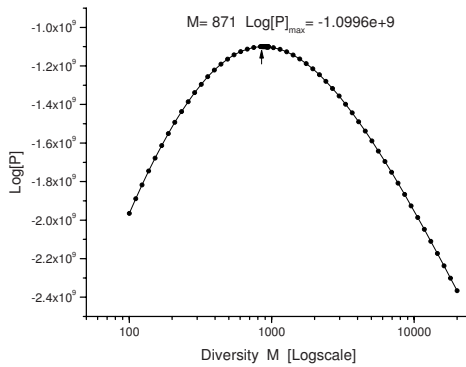


Fig. 4. $\text{Log}[P]$ at different diversity M

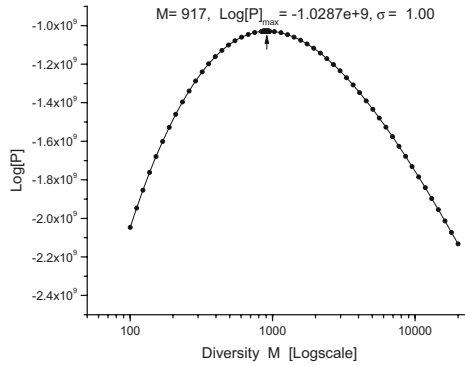


Fig. 5. $\text{Log}[P]$ at different diversity M

the log-likelihood $\text{Log}[P]$ as the y -axis and the diversity M as the x -axis. $\text{Log}[P]$ approaches its maximum when $M = 871$.

Model D

We assume the abundance distribution is the lognormal distribution. By calculating the log-likelihood assuming $\sigma = 1.0$, we generate the **Fig. 5** with the log-likelihood $\text{Log}[P]$ as the y -axis and the diversity M as the x -axis. $\text{Log}[P]$ approaches its maximum when $M = 917$.

5 Discussion

Our method is based on the method of coverage depth described in [2] and the method of contig spectrum in [4]. But unlike [2], we apply a mathematical tool of maximum likelihood estimation to maximize the likelihood of fitting the coverage depth at every position of the assembly sequence.

We assume four different abundance distributions: the discrete distribution with a fixed number of abundance levels, the power law distribution, broken stick distribution and lognormal distribution. We gave the general formulas for different cases, and developed the corresponding programs for the tests on the Sargasso Sea data. The results on the Sargasso Sea data are within the range of the estimated diversity in [2] (1000 – 47733). We estimated the diversity to be approximately 900 when assuming three distributions: discrete abundance distribution, the broken stick distribution and the lognormal distribution; the diversity is estimated at approximately 3500 if the abundance distribution is the power law distribution.

We note that as in [4] our estimates are sensitive to the quality of the data provided and especially to the assembly parameters used to determine the coverage. While three out of the four distributions suggest a value close to the low end of the range estimated in [2], it has been suggested that the power law distribution is the best fit to observed diversity levels in phage data [4]. For this reason, we believe the 900 figure is best interpreted as a lower bound on the number

of species present in the sample and the power law estimate is the most accurate. We plan on experimenting with more data sets in order to further evaluate this discrepancy. We also intend to work with other suggested distributions such as the logarithmic distribution, exponential distribution, and niche preemption distribution.

Acknowledgments. The authors would like to acknowledge Karin Remington and Aaron Halpern from the Venter Institute for responding to our emails and providing us with the Sargasso Sea data.

References

1. T.P. Curtis, W.T.Sloan: Exploring Microbial Diversity - A Vast Below, *Science*, 2005, 309: 1331–1333.
2. J.C. Venter *et al.*: Environmental Genome Shotgun Sequencing of the Sargasso Sea, *Science*, 2004, 304: 66–74.
3. Supporting Online Material:
URL: www.sciencemag.org/cgi/content/full/1093857/DC1.
4. F. Angly *et al.*: PHACCS, an Online Tool for Estimating the Structure and Diversity of Uncultured Viral Communities Using Metagenomic Information, *BMC Bioinformatics*, 2005, 6: 41. URL: www.biomedcentral.com/147-2105/6/41
5. B.J.M. Bohannan, J. Hughes: New Approaches to Analyzing Microbial Biodiversity Data, *Current Opinion in Microbiology*, 2003, 6: 282–287.
6. G. Myers: Whole-Genome DNA Seqencing, *Computing in Science and Engineering*, May-June 1999: 33–43.
7. F.W. Preston: The Commonness and Rarity of Species, *Ecology*, 1948, 29: 254–283.
8. M.G. Bulmer: On Fitting the Poisson Lognormal Distribution to Species Abundance Data, *Biometrics*, 1974, 30: 101–110.
9. S. Hubbell: *The Unified Neutral Theory of Biodiversity and Biogeography*, Princeton, New Jersey: Princeton University Press; 2001
10. T.P. Curtis, W.T. Sloan, J.W. Scannell: Estimating Prokaryotic Diversity and Its Limits, *Proc Natl Acad Sci USA*, 2002, 99: 10494–10499.
11. J. Dunbar, S. Barns, L. Ticknor, C. Kuske: Empirical and Theoretical Bacterial Diversity in Four Arizona Soils, *Appl Environ Microbiol*, 2002, 68: 3035–3045.
12. J. Zhou *et al.*: Spatial and Resource Factors Influencing High Microbial Diversity in Soil, *Appl Environ Microbiol*, 2002, 68: 326–334.
13. I. Kroes, P.W. Lepp, D. Relman: Bacterial Diversity Within the Human Subgingival Crevice, *Proc Natl Acad Sci USA*, 1999, 96: 14547–14552.
14. J.B. Hughes *et al.*: Counting the Uncountable: Statistical Approaches to Estimating Microbial Diversity, *Appl Environ Microbiol*, 2001, 67: 4399–4406.
15. G. Seber: *The Estimation of Animal Abundance and Related Parameters*, London: Griffin; 1973.
16. C. Krebs: *Ecological Methodology*, New York: Harper and Row; 1989.
17. A. Chao: Estimating the Population Size for Capture-recapture Data with Unequal Catchability, *Biometrics*, 1987, 43: 783–791.
18. M. Breitbart *et al.*: Genomic Analysis of Uncultured Marine Viral Communities, *Proc Natl Acad Sci USA*, 2002, 99: 14250–14255.
19. A. Reysenbach *et al.*: Differential Amplification of rRNA Genes by Polymerase Chain Reaction, *Appl Environ Microbiol*, 1992, 58: 3417–3418.

20. M. Suzuki, S. Giovannoni: Bias caused by Template Annealing in the Amplification of Mixtures of 16S rRNA Genes by PCR, *Appl Environ Microbiol*, 1996, 62: 625–630.
21. A. Speksnijder *et al.*: Microvariation Artefacts Introduced by PCR and Cloning of Closely Related 16S rRNA Gene Sequences, *Appl Environ Microbiol*, 2001, 67: 469–472.
22. G. Jasons, M. Wolinsky, J. Dunbar: Computational Improvements Reveal Great Bacterial Diversity and High Metal Toxicity in Soil, *Science*, 2005, 309: 1387–1390.
23. P.G. Falkowski, C. de Vargas: Shotgun Sequencing in the Sea: a Blast from the Past? *Science*, 2004, 304: 58–60.
24. J.M. Travis, D.R. Larsen: Measures of Diversity, *Natural Resource biometrics*, 1995.