Estimation of microbial cover distributions at Mammoth Hot Springs using a multiple clone library resampling method

Héctor García Martín[1,2] and Nigel Goldenfeld[1,3*]

[1]Department of Physics and [3]Institute for Genomic Biology,
University of Illinois at Urbana-Champaign,
1110 West Green Street,Urbana, Illinois 61801

[2]Current address: DOE Joint Genome Institute,
Walnut Creek, California 94598, USA

*To whom correspondence should be addressed,
Phone : 217-333-8027, Fax: 217-333-9819, Email: nigel@uiuc.edu

September 4, 2005

Abstract

We propose the use of cover as a quick, low-resolution proxy for the abundance of microbial species, which is free from PCR bias. We showcase this concept in a computation that uses clone library information from travertine-forming hot springs in Yellowstone National Park to provide estimates of relative covers at different locations within the spring system. Samples were used from two media: the water column and the travertine substrate. The cover distribution is found to approximate a power law for samples within the water column. Significant commonality of species with the highest cover is observed in the water column for all locations, but not for species present in the substrate at different locations or between media at the same location. (122 words)

Keywords: Cover, Abundance, Bootstrap, Resampling, 16S rRNA, Clone library.

1 Introduction

Until recently, the study of microbial ecology was narrowly constrained by the difficulty of identifying microbes outside of cultures. Modern molecular methods, based upon the sequencing of small subunit rRNA genes (Olsen *et al.*, 1986), (Pace *et al.*, 1986) permit the classification and comparison of microbes directly from an environmental sample. A key step in many, but not all, molecular methods is the creation of a clone library containing representatives of the environmental 16S rRNA. Sequencing samples of the clone library has enabled estimates of diversity to be obtained in a variety of environments ranging from geothermal hot springs to the oral cavity. Clone libraries are, by now, numerous and relatively straightforward to assemble. Sequencing, whilst expensive, is becoming cheaper and high-throughput methods are available that enable huge datasets to be created from environmental samples.

However, diversity is not an adequate characterization of the dynamics, metabolism and community structure of an ecosystem. For this purpose, some measure of abundance is desirable; even though the most abundant organisms are not necessarily those which dominate the ecosystem dynamics, any quantitative understanding of geobiochemical cycles requires information about abundance. A variety of methods are available to measure abundance: Quantitative PCR, Most Probable Number PCR, competition PCR and dot-blot hybridization among others (Muyzer *et al.*, 1993), (Ding and

Cantor, 2004), (Amann *et al.*, 1995), (Head *et al.*, 1998), (Zoetendal *et al.*, 2004), each of them with their own advantages and disadvantages. These techniques are valuable probes of the environment, but are extremely local, providing information on scales that are often very much smaller than those characteristic of environmental spatio-temporal dynamics. Clone libraries are generally created from much larger, system wide samples, and so could provide, in principle, a more global, but still spatially-resolved measure of abundance. Unfortunately, attempts to estimate abundance using clone libraries are hampered by inherent biases in PCR amplification and cloning (Wintzingerode *et al.*, 1997).

The purpose of this paper is to propose a statistical method for estimating a coarse grained (or low resolution) measure of abundance based on the concept of cover, and using clone libraries alone. Our approach is fast, cheap, capable of high-throughput and only requires the use of a computer. Most importantly, we will show that our method is not significantly affected by extraction and PCR bias. Furthermore, being based upon clone libraries, it gives a large-scale, system-wide estimate of cover. We believe that our technique can provide a rapid and convenient first assay of an ecosystem, providing relative cover of the microbial population; such an assay would be expected to be followed up by local probes, using, for example, one of the techniques mentioned above.

We illustrate this method with data from a travertine-forming hot spring in Yellowstone National Park, where earlier studies (Fouke *et al.*, 2003) report

the nominal presence of 221 Operational Taxonomical Units (OTUs). Our technique yields information on which are the most abundant (in terms of cover) OTUs, and the ones with the greatest potential impact to drive the ecosystem metabolism. In this way, our analysis focuses attention on the 10 or 15 OTUs with highest cover, distinguishing them from the several hundred OTUs detected in previous work (Fouke *et al.*, 2003). The putative metabolic characteristics of these organisms can provide a clue as to their environmental role, and the likely dominant biogeochemical pathways that are active in the system.

2 Relative cover estimation

We define the relative abundance r_i as the fraction of total individuals in the system belonging to OTU i : $r_i = n_i/n$. Here n_i is the number of individuals belonging to OTU i and n is the total number of individuals. The index i takes on values from 1 to S, S being the total number of OTUs observed in the system. Samples are assumed to have been collected at each facies (see section 5.1 and to have been processed through the standard procedure of DNA extraction, 16S rRNA gene PCR amplification, cloning and clone screening to create a clone library as explained in (Fouke *et al.*, 2003).

Ideally, one would count every individual in the system and assign it to an OTU to calculate r_i. This is unfeasible first and most obviously because of the impossibility of sampling the whole system, and secondly because each

sample does not give information on relative abundance. The reason for the latter is that clone library abundances are not representative of abundance in the real system, because of biases present in PCR amplification. A small preference in primer binding for a certain OTU type is exponentially amplified and will distort abundances greatly. Other biases are introduced by the DNA extraction, ligation and transformation but they lack the exponential growth inherent to PCR DNA amplification. We will therefore use only information on the presence or absence of each OTU in each sample.

Having surrendered the (biased) abundance information for OTUs that is reflected in the clone library abundance, we need to find an alternative way to estimate abundance. The idea that we propose here is that if one has collected many environmental samples from the same location, and generated a clone library, the samples will show variations in which OTUs are present. These variations reflect in a non-trivial way the spatial abundance distribution of the organisms, and our task now is to extract this in the least biased way.

To this end, we use the collected data to obtain estimates of coarse-grained abundances or covers as explained in figure 2. Cover, sometimes known as occurrence or range, is a concept from macroscopic ecology, and is strongly linked to abundance (Kunin, 1998), (He and Gaston, 2000), (Kunin, 2000) although not equivalent. Referring to figure 2, assume that the square represents one of the facies in the system, properly divided into smaller subcells of size l. This length, which we call the correlation length l, is defined to be small enough so that sampling within the boundaries of a subcell would

	OTU number						
Sample	1	2	3	4	5	6	7
A	•	•		•	•		
B	•	•	•		•		
C	•	•		•			•
D	•					•	•
$\hat{C}_i$	(4	3	1	2	2	1	2)/4
$\hat{\rho}_i$	26.7%	20%	6.7%	13.3%	13.3%	6.7%	13.3%

Figure 1: Example of how to calculate the estimates of the cover $\hat{C}_i$ and relative cover $\hat{\rho}_i$ for a given number of samples according to equations 4 and 5. Only information of presence or absence of a given OTU is used.

always yield the same result. One can define the *cover* of OTU i to be the fraction of subcells in which OTU i is present over the total number of subcells:

$$C_i^t = \frac{X_i^t}{X} \tag{1}$$

and the time-averaged cover is:

$$C_i = \sum_{t=1}^{T} \frac{C_i^t}{T} \tag{2}$$

The relative cover is defined by simply normalizing the cover:

$$\rho_i = \frac{C_i}{\sum_i C_i} \tag{3}$$

Sampling the whole facies to find the true cover C_i is out of reach. Random sampling from each facies yields estimates (denoted by a caret) that

should converge quickly as the number of samples increases:

$$\hat{C}_i = \frac{N_i}{N} \tag{4}$$

$$\hat{\rho}_i = \frac{\hat{C}_i}{\sum_i \hat{C}_i} \tag{5}$$

where N_i is the number of samples in which OTU i is present and N is the total number of samples (see fig. 1). This last equation then provides a quick estimate of relative covers and hints at which OTUs are more likely to influence the microbial ecosystem.

Our method relies critically on the variability of detected OTUs from sample to sample. Why does this variation arise? In general, it is due to two main effects: (i) spatial and temporal variation, and (ii) detection errors.

As explained in section 5, samples were taken in different spatial location within the same facies and at different times of the year and day. Microbial species show preferred ranges of temperature and pH ranges, and have been shown to partition fairly tightly to given facies (Fouke *et al.*, 2003). It is therefore not surprising that spatial and temporal variations of pH, temperature and other facies characteristics within a given facies give rise to distinctive patterns in the location of OTUs.

Detection errors arise because the processes of extraction, amplification, ligation, transformation and sequencing have an intrinsic variability in their success rate, that can be dependent on the skill and expertise of the experimenter. For example, the large scale of the survey described below(more

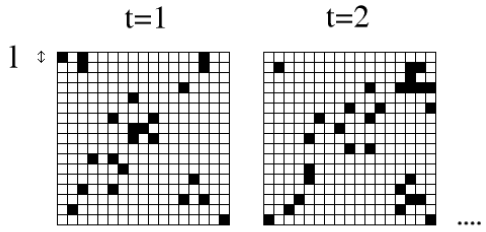

Figure 2: Cover, a concept borrowed from macroscopic ecology (Kunin, 1998),(He and Gaston, 2000),(Kunin, 2000), is a coarse-grained or low-resolution measure of abundance in the sense that each subcell will contribute if the species is present inside it, independent of its abundance in the subcell. The cover is the number of subcells in which a given OTU is present divided by the total number of subcells (equation 1), for each of the possible times t ($t = 1..T$). For example, for $t = 1$ the cover is $C_i^1 = 23/289$, and for $t = 2$, $C_i^2 = 29/289$. Each of the squares is a diagram representing one of the facies in the system.

than 14000 clones were screened) implied assignment of these tasks to persons of different levels of expertise (Fouke *et al.*, 2003). With the amount of available data it is hard to tell apart how much variance is due to spatial and temporal variation and how much is due to detection errors. High throughput, standardized "pipelines" for 16S rRNA analysis reduce these detection errors to a minimum, and allow for the large number of samples necessary for this method. So, to proceed, we will assume here that detection error has been minimized by careful and reproducible laboratory practice, and that there are spatial or temporal trends in the possible causes of detection error. Thus, we take into account explicitly only the variability arising from intrinsic spatial and temporal dynamics.

3 The library resampling method

Equation 5 offers an estimate for the relative cover but not its variance, critical to ascertaining the variability of ρ_i across the facies and in time. In this section, we use a computer-intensive library resampling method to estimate variability.

The library resampling method is an application of the original bootstrap method introduced by Efron in 1979 to assess the accuracy of statistical estimates and provide bias corrections (Efron, 1979). A familiar example of the bootstrap principle is its application in estimating confidence limits on phylogenies (Felsenstein, 1985), but the original bootstrap is a data resampling statistical method of much wider applicability. It is the broader method that we use here. A basic exposition of the data resampling bootstrap method can be found in (Shao and Tu, 1995), (Efron and Tibshirani, 1993) and (Chernick, 1999). Here we will limit ourselves to explaining its use for the case at hand, but in a self-contained way.

How can the data resampling bootstrap method be used to obtain a variance for the relative cover estimate in 5? Traditionally, one would divide the N samples in M groups of N/M samples and obtained the estimates of $\hat{\rho}_i^s$ ($s = 1..N/M$) for each of these groups as per equation 5. The variance would be obtained as usual: $var(\hat{\rho}_i) = \sum_s (\hat{\rho}_i^s - \hat{\rho}_i^{av})^2$, where $\hat{\rho}_i^{av}$ denotes the average of $\hat{\rho}_i^s$. For large enough N this would converge to the desired variance. Nonetheless, this procedure wastes samples for each estimate $\hat{\rho}_i^s$ ($s = 1..M$)

and leads to a poor estimation. For example, for $N = 8$ (as for data in section 5) having 4 groups would lead to a meager 2 samples per group.

The data resampling bootstrap explores the variance by forming groups of samples, whose content is randomly sampled from the original samples, but which have the same amount of samples per group as the total initial number of samples. This is achieved by choosing these groups through sampling with replacement as explained in figure 3: R bootstrap groups are created and a relative cover estimate $\hat{\rho}_i^s$ is calculated for each. The data resampling bootstrap theorem states that, for large enough R, the behaviour of the $\hat{\rho}_i^s$ around $\hat{\rho}_i$ mimics the behaviour of $\hat{\rho}_i$ around ρ_i (see appendix A in supplementary material and (Chernick, 1999)). One can therefore obtain an improved estimate and its variance by treating the bootstrap groups as independent measurements:

$$\rho_i^{BS} = \frac{1}{R}\sum_{s=1}^{R}\hat{\rho}_i^s \tag{6}$$

$$var(\hat{\rho}_i^{BS}) = \frac{1}{R}\sum_{s=1}^{R}(\hat{\rho}_i^s - \hat{\rho}_i^{BS})^2 \tag{7}$$

The data resampling bootstrap principle as stated above is not always applicable (e.g. extremal statistics (Chernick, 1999)) and the convergence to the right distribution must be proven for each estimator (Shao and Tu, 1995). Equations 6 and 7 refer to functions of sample means for which the probability distribution of the bootstrap resampling has been proved to converge to the probability distribution of the estimates in the limit of large N (see

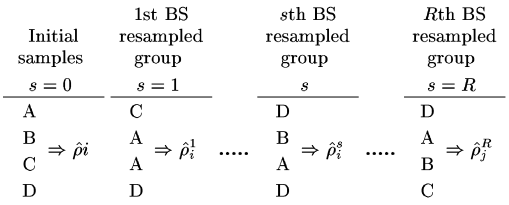

Figure 3: The data resampling bootstrap method applied to estimate the bias and variance of $\hat{\rho}_i$. R groups of four samples are generated by sampling with replacement from the original samples. This means that the samples in each group are chosen randomly among the original samples and each time a sample is selected for the group it is returned to the original set, so it can be chosen again. Therefore each group is not just a permutation of the initial samples. For each group, $\hat{\rho}_j^s$ is generated using equation 5 and the estimate of the bias and the variance are given by equations 6 and 7.

appendix A in supplementary material).

4 Model calculation to illustrate the use of the resampling method

In order to give a worked example of the use of the resampling method, we present in this section a model calculation on artificial data, and show to what extent the resampling method is capable of making faithful estimates from finite data sets. The artificial data has been constructed so that it mimics some aspects of the field data we will eventually analyze in the following section. To begin the discussion, we first explain how the artificial data were constructed from a model distribution, and the extent to which these artificial data have realistic properties. We would like our artificial data to be semi-

realistic, so that the success of the resampling algorithm on the artificial data has some relevance to the application of the resampling algorithm on field data. We conclude this section by exploring how well resampling converges with increasing sample size.

The theorem in appendix A (supplementary material) proves the consistency of the data resampling bootstrap estimator in the asymptotic limit, which is a necessary condition for the validity of the bootstrap method. The real interest of the bootstrap lies in its fixed sample properties. The variance must eventually converge to the true variance for $N \to \infty$, but for a finite N it also provides a measure of the variability of relative covers by omitting the use of certain samples, therefore yielding an account of its reliability.

There is no general theory for fixed sample properties of the bootstrap. Its performance is usually examined through empirical simulations (Shao and Tu, 1995), (Efron and Tibshirani, 1993). In this section we will assess the performance of the bootstrap by generating a series of samples from a known model cover distribution ρ_i and checking how close the bootstrap estimation using N of these samples is to the original, known distribution.

For this demonstration calculation, we assume that each of the S OTUs present in the system are present in each sample with probability $\rho(i) \propto i^{-0.65}$ ($i = 1..S$). As we will see, this model distribution actually mimics the cover distribution of microbes that, in section 5, we will obtain in the water column of the pond facies of the Yellowstone National Park data. The total number of OTUs S is chosen here to be $S = 200$ since the number of OTUs

in, for example, the water filters of the Pond is 43 and previous results (García Martín, 2004) indicate that $20\% - 25\%$ of the total diversity has been sampled.

This model distribution, and the parameters given above, represent the simplest possible ones that seem to capture certain realistic aspects of the field data used below. Note that here we are assuming that there are no correlations between samples. In the model distribution, the range of maximum and minimum number of OTUs found in one sample is close to that of the field data set used below: 3 to 26 OTUs for the model and 3 to 25 for the data. Nonetheless, the variances in the number of OTUs found in one sample are very different (3.28 for the model and 6.79 for the data), and the number of OTUs detected after a fixed number of samples is appreciably higher in the model (70.5 ± 6.2 vs 43). Thus, the model used for the demonstration calculation in this section captures only some of the features of the field data; the fact that the variance is so different from the field data suggests that the field data contain correlations not contained within the model distribution.

Now, we present results from the resampling calculation. Figures 4 and 5 show the results of the data resampling bootstrap estimates $\hat{\rho}_i(N)$, for sample numbers $N = 10, 100$ and $R = 10000$ as compared to the original relative cover ρ_i. The results are satisfactory, with the target cover within the variance of the estimate. As expected, estimates improve with increasing N. For low N the estimates overshoot slightly since not all S OTUs have been detected and therefore the detected OTUs are given a higher relative cover

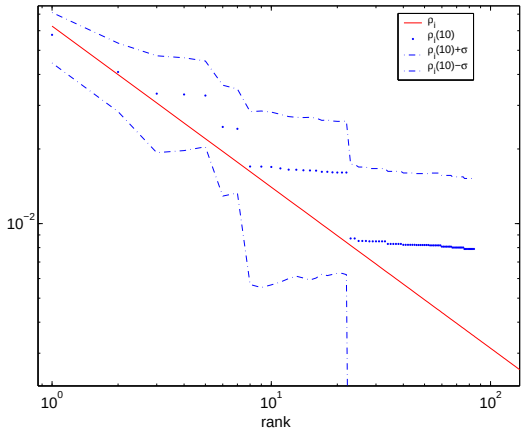

Figure 4: Data resampling bootstrap estimate for $N = 10$ samples. In spite of this low number of samples it is possible to get a hint of the underlying distribution. The estimate overshoots because for such low amount of samples not all OTUs have been detected and therefore the detected OTUs attain a higher relative cover so the sum of the relative covers adds up to unity.

than the real one.

R is in practice chosen large enough so that further increases don't change the estimate appreciably.

5 Analysis of Yellowstone National Park field data

5.1 Study site

We now turn to an application of the resampling method on field data from microbial communities at Yellowstone National Park, collected and published previously (Fouke *et al.*, 2003), (Bonheyo *et al.*, 2005) as part of a large

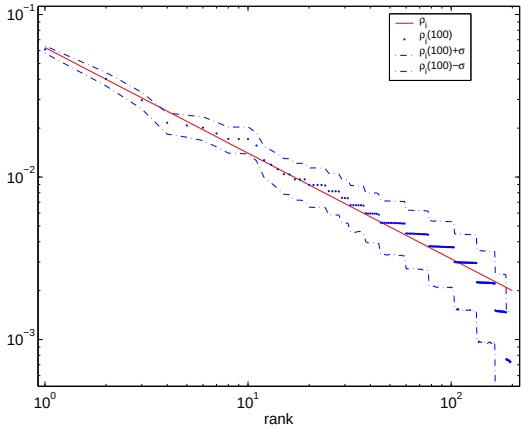

Figure 5: The data resampling bootstrap estimate improves for $N = 100$ as expected and approximates satisfactorily the target distribution.

biocomplexity study at the University of Illinois at Urbana-Champaign. Our purpose here is to illustrate how we have analyzed the microbial communities, and not to present a detailed description of the ecological context or our conclusions regarding the role of microbes in biomineralization.

For the dataset presented here, up to 50 samples were taken during an interval of 4 years at Spring AT-1, located on Angel Terrace, in the upper terrace region of the Mammoth Hot Springs complex at Yellowstone National Park. AT-1 is typical of the travertine-depositing springs at this site, and has been fully-characterized by (Fouke *et al.*, 2000): hot waters erupt from the vent and flow downhill cooling down, quickly degasing CO_2, increasing in pH and precipitating travertine at extremely high rates ($\sim$ 1.5 m per year) in a characteristic terraced architecture. The fast deposition rates produce a hostile environment for present microbial life, which must somehow avoid entrapment in the travertine substrate (Bonheyo *et al.*, 2005).

Samples were taken from all the five facies: vent, apron and channel, pond, proximal slope and distal slope. A facies is a subenvironment of sedimentary deposition within a system with specific physical, chemical, geological and biological characteristics. Biological subenvironments correlate tightly with facies. A broader explanation of the facies model can be found in (Fouke *et al.*, 2000). The samples were collected from two different media: filtered water from the flowing water column, and the surface of the deposited travertine substrate, with depths up to 2 cm. deep (see (Fouke *et al.*, 2003) for details of facies definitions and more specific information on the site).

Bacteria were identified through 16S rRNA gene identification as explained in (Fouke *et al.*, 2003). For each sample, clones were screened for unique sequences through RFLP. Three different sets of OTU definitions were used, based on sequence differences of 0.5%,1% and 3%, with the intention of determining to what extent, if any, our conclusions were affected by the OTU definition (Bonheyo *et al.*, 2005).

5.2 Data resampling bootstrap estimates

The procedure for obtaining the cover ρ_k^{BS} is the same as explained above with a total of $R = 10000$ bootstrap samples being used. The results are given in the form of rank abundance plots in figure 7 and figures 6 and 9 in the supplementary material. Rank tables for all facies and mediums are shown in figure 6 and figures 5 and 8 in the supplementary material, along with the number of samples for each case. A rank table with phylotype

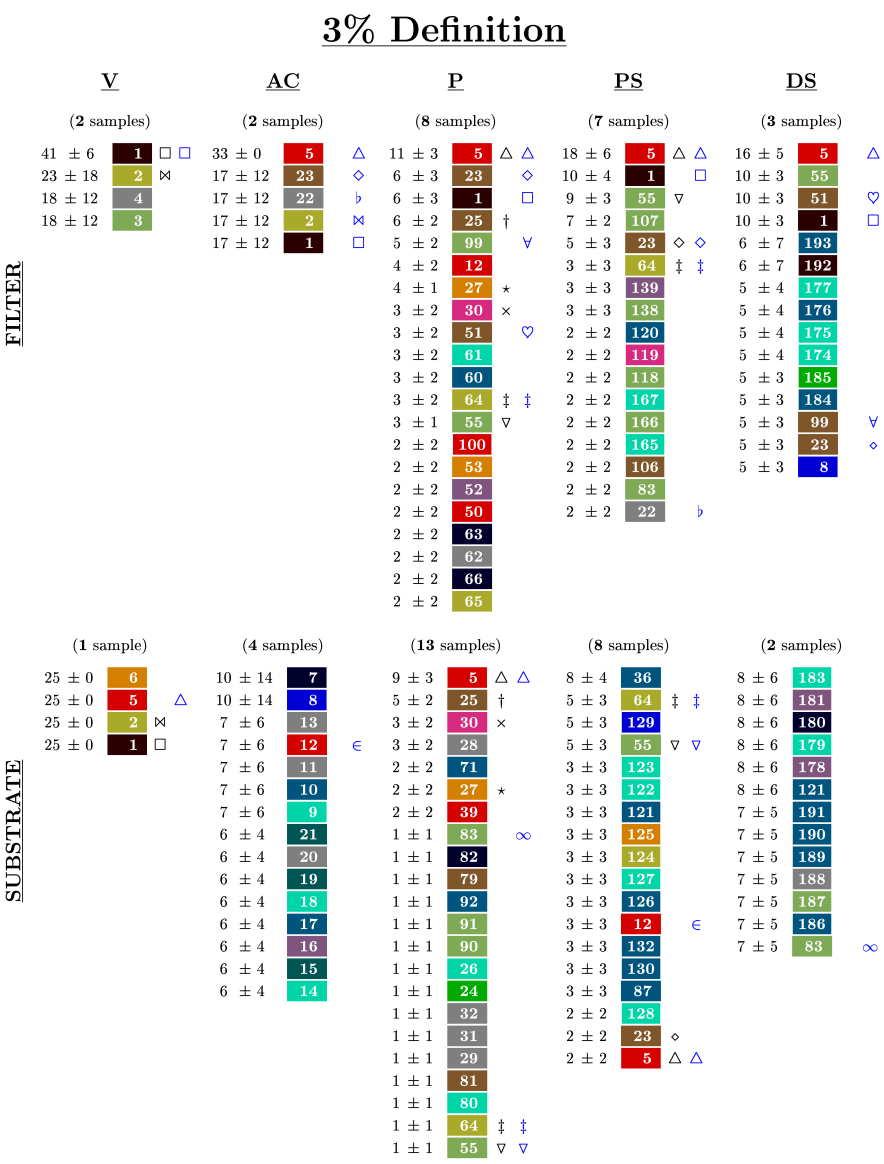

Figure 6: OTUs with highest coverage for the 3% difference definition for different facies and media. V stands for vent, AC for apron channel, P for pond, PS for proximal slope and DS for distal slope (Fouke *et al.*, 2003). Figures are relative covers with their variances. Numbers are identification OTU numbers given in figure 3 in the supplementary material. Black symbols mark OTUs that are present in another medium in the same facies. Blue symbols mark OTUs that are present in another facies in the same medium. Colors indicate phylotypes according to the code in figure 2 in the supplementary material. For reasons of space only the OTUs with highest covers are shown.

3% Definition

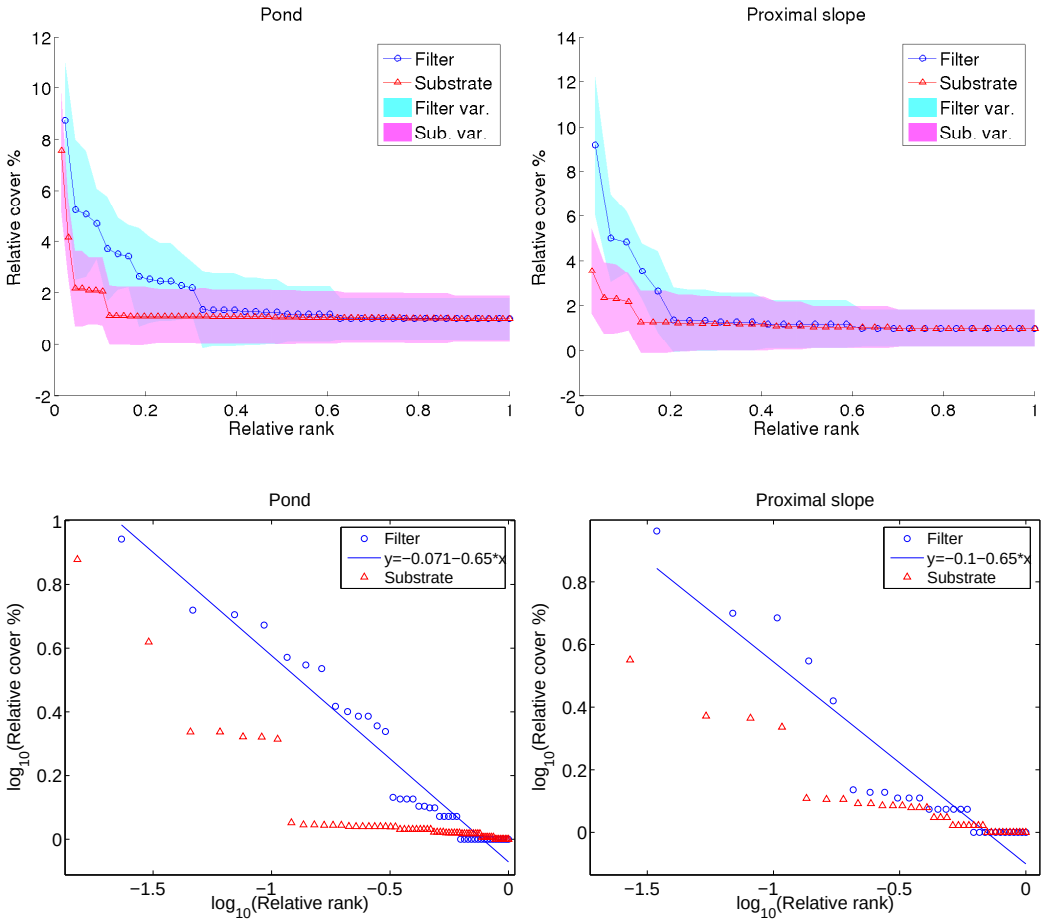

Figure 7: Plots of relative cover versus relative rank for the 3% definition in normal (above) and log-log axis (below). Relative covers are covers divided by the lowest cover. An OTU with rank i has the ith highest cover. Relative rank is rank divided by the total number of OTUs. Only the filter samples from the pond and proximal slope facies seem to be well-described by a power law, within the limits imposed by the small amount of samples used (i.e. steps in the lower right end). Substrate sample plots from both facies curve upwards.

Phylotype	Clone abundance	Cover estimate
Beta proteobacteria	22%	$10 \pm 4\%$
Cyanobacteria	16%	$18 \pm 3\%$
Aquificales	15%	$4 \pm 2\%$
Alpha proteobacteria	11%	$17 \pm 4\%$
Unknown division	9%	$14 \pm 4\%$
Green sulfur bacteria	9%	$5 \pm 2\%$
BCF group	7%	$9 \pm 3\%$
Delta proteobacteria	3%	$3 \pm 1\%$
Candidate div. OP-11	2%	$5 \pm 2\%$
Green non-sulfur bacteria	2%	$5 \pm 3\%$
Thermus/Deinococcus group	1%	$2 \pm 2\%$
Gamma proteobacteria	1%	$1 \pm 1\%$
Firmicutes	negligible	$3 \pm 2\%$
Eukaryota, Chloroplasts	negligible	$1 \pm 1\%$

Figure 8: Comparisons of covers and clone relative abundances from figure 6 in Bonheyo *et al.* (2005) for the proximal slope facies.

relative covers is available in figure 2 in the supplementary material.

Figure 8 presents a comparison of covers obtained through the resampling method and nominal clone abundances (figure 6 in (Bonheyo *et al.*, 2005)) for the proximal slope facies. The results are not wholly different: nominal clone library abundances are not completely misleading, although it is evident that library creation biases seems to have overepresented certain phylogenetic groups. Aquificales, for example, seem to have been overrepresented by a factor of more than three and beta-proteobacteria by a factor of two.

As can be seen, only the Pond and Proximal Slope facies have enough number of samples for the resulting covers to be statistically meaningful. Therefore only cover distributions for these mediums and facies are presented.

In the case of the ranked tables, nonetheless, even for 3 or 4 samples the results offer a qualitative idea of relative covers: the fact that the reported OTUs, and not others, are present is suggestive of a higher cover, although it cannot be quantified as would be the case with a larger sample size.

In the case of the Pond and Proximal Slope the covers seem to fit a power law for the water samples, in contrast with the substrate, where they do not. In the latter case, the rank-abundance curve is steeper, with the most dominant organisms having relatively more cover than in the former case.

It can also be noticed that among the highest ranking OTUs, there is a certain degree of commonality in the case of the water samples, but not in the substrate. This is in contrast with the reported biodiversity pattern, which is different for each facies in both facies and mediums (Fouke *et al.*, 2003), (Bonheyo *et al.*, 2005). We conclude that difference reflects the fact that the fluid motion provides a downstream flush of cells that is absent in the substrate. Remarkably, this is only noticeable for organisms with highest covers; less abundant organisms are niche dependent. Also, of the top ranking OTUs in the water very few appear in the top ranks of the sediment. If encrustment in substrate or adherence to the surface biofilm were a random process, it would be expected that the bacteria with highest cover in water would also have the highest cover in the substrate. Since this is not the case, it can be concluded that encrustment or surface biofilm adherence is not random: some species are more able to avoid it (or provoke

it) than others.

Figure 1 in the supplementary material presents the putative metabolic characteristics of the OTUs with highest cover, deduced from close relatives (in terms of 16S rRNA similarity). Although crude, lacking any other genomic information this is the only way to obtain a glimpse of the most abundant metabolisms. In agreement with Spear (Spear *et al.*, 2005), hydrogen metabolism seems to be a common feature in this spring.

Finally, we comment briefly on the highest cover organism identified by our analysis. OTU 5 (using the 3% OTU definition) seems to have highest cover in all facies in the pond and the water samples from the apron and channel and proximal slope. This OTU is an unknown beta proteobacterium and corresponds to OTU 8 in the 1% definition, and splits up into several different OTUs under the 0.5% definition. This seems to indicate that using too fine a distinction between sequences in the definition of OTUs is not ecologically useful. Consistent with this, high variances for cover estimations are noticed in the case of the 0.5% definition, suggesting that this may be too narrow a distinction for OTU definitions. Another possible explanation is, of course, that the sample size is too small.

6 Conclusion

We have presented a computational method that uses clone library information to provide a large-scale estimate of relative coarse grained abundance or

cover. Even though the role of an organism in an environment is not necessarily proportional to its abundance, this estimate can be used to generate hypotheses as to which bacterial OTUs have the potential for significantly influencing the ecosystem; thus our technique supplies possible candidates for later quantitative work involving (e.g.) hybridization probes.

The resampling method has been used with data from travertine-forming hot springs at Yellowstone National Park to provide estimations of relative covers for different facies and mediums. OTUs with highest cover are prime candidates to influence the degasing of CO_2 which, in turn, produces calcium carbonate precipitation and ultimately gives rise to the formation of the travertine terraces. The data for covers seems to fit well a power law for the water samples.

We report substantial commonality of species with highest cover in the water medium, but not in the substrate or between media in the same facies. This fact can be attributed to the water downflush of bacteria. In any case, commonality would be expected to be limited to bacteria with highest cover, since there is very little commonality of OTUs between facies (Fouke *et al.*, 2003). Lack of commonality between water and substrate samples indicates that substrate encrustment and surface biofilm adherence is not random, with some OTUs being able to avoid or provoke it.

The use of 3 different sets of OTU definitions permits us to explore the issue of the proper definition of OTUs/species. We conclude that differentiating OTUs by 0.5% may be excessive and advocate the 1% difference

definition.

7 Acknowledgements

We acknowledge helpful discussions with Bruce Fouke, Alison Murray and Philip Hugenholtz. This research was supported by the National Science Foundation through grant NSF-EAR-0221743.

References

Alexander, B., Andersen, J. H., Cox, R. P., and Imhoff, J. F. (2002). Phylogeny of green sulfur bacteria on the basis of gene sequences of 16S rRNA and of the Fenna-Matthews-Olson protein. *Arch. Microbiol.*, **178**(2), 131–40.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–402.

Amann, R. I., Ludwig, W., and Schleifer, K.-H. (1995). Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *FEMS Microbiology reviews*, **59**(1), 143–69.

Bonheyo, G., Frias-Lopez, J., García Martín, H., Veysey, J., Goldenfeld,

N., and Fouke, B. (2005). Partitioning of Mineralogical, Geochemical and Microbial Systems in Travertine Terraces at Yellowstone Hot Springs. Submitted to *Environmental Microbiology*.

Chelius, M. K. and Triplett, E. W. (2000). Dyadobacter fermentans gn. nov., sp. nov., a novel Gram-negative bacterium isolated from surface-sterilized Zea mays stems. *Int. J. Syst. Evol. Microbiol.*, **50**(2), 751–58.

Chernick, M. R. (1999). *Bootstrap methods, a practitioner guide*. Wiley, New York.

Ding, C. and Cantor, C. R. (2004). Quantitative analysis of nucleic acids - the last few years of progress. *Journal of Biochemistry and Molecular Biology*, **37**(1), 1.

Eder, W. and Huber, R. (2002). New isolates and physiological properties of the aquificales and description of thermocrinis albus sp. nov. *Extremophiles*, **6**(4), 309–18.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, **7**, 1–26.

Efron, B. and Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Chapman & Hall/CRC, Boca Raton Florida.

Felsenstein, J. (1985). Confidence limits on phylogenetics: An approach using the bootstrap. *Evolution*, **39**, 783–91.

Fouke, B., Bonheyo, G., Sanzenbacher, B., and Frias-Lopez, J. (2003). Partitioning of bacterial communities between travertine depositional facies at Mammoth Hot Springs, Yellowstone National Park, USA. *Canadian Journal of Earth Sciences*, **40**(11), 1531–48.

Fouke, B., Farmer, J., Des Marais, D., Pratt, L., Sturchio, N., Burns, P., and Discipulo, M. (2000). Depositional facies and aqueous-solid geochemistry of travertine-depositing hot springs (Angel Terrace, Mammoth Hot Springs, Yellowstone National Park, USA). *Journal of Sedimentary Research*, **70**(11), 565–85.

García Martín, H. (2004). *Statistical analysis of highly correlated systems in biology and physics*. Ph.D. thesis, University of Illinois at Urbana-Champaign, Department of Physics.

Genbank (1982). http://www.ncbi.nlm.nih.gov/Genbank/index.html.

Hayashi, N. R., Ishida, T., Yokota, A., Kodama, T., and Igarashi, Y. (1999). Hydrogenophilus thermoluteolus gen. nov., sp. nov., a termophilic, facultatively chemolithoautotrophic, hydrogen-oxidizing bacterium. *Int. J. Syst. Bacteriol.*, **49**(2), 783–6.

He, F. and Gaston, K. J. (2000). Estimating species abundance from occurrence. *The American Naturalist*, **156**(5), 553–59.

Head, I., Saunders, J., and Pickup, R. (1998). Microbial evolution, diversity

and ecology: a decade of ribosomal RNA analysis of uncultivated microorganisms. *Microbial Ecology*, **35**, 1–21.

Heising, S., Richter, L., Ludwig, W., and Schink, B. (1999). Chlorobium ferrooxidans sp. nov., a phototrophic green sulfur bacterium that oxidizes ferrous iron in coculture with a "Geospirillum" sp. strain. *Arch. Microbiol.*, **172**(2), 116–24.

Hiraishi, A., Yonemitsu, Y., Matsushita, M., Shin, Y., Kuraishi, H., and Kawahara, K. (2002). Characterization of Porphyrobacter sanguineus sp. nov., an aerobic bacteriochlorophyll-containing bacterium capable of degrading biphenyl and dibenzofuran. *Arch. Microbiol.*, **178**(1), 45–52.

Kunin, W. E. (1998). Extrapolating species abundance across spatial scales. *Science*, **281**, 1513–15.

Kunin, W. E. (2000). Scaling down: on the challenge of estimating abundance from occurrence patterns. *The American Naturalist*, **156**(5), 560–66.

Lonergan, D., Lenter, H., Coates, J., Phillips, J., Schmidt, T., and Lovley, D. (1996). Phylogenetic analysis of dissimilatory Fe(III)-reducing bacteria. *Journal of Bacteriology*, **178**(8), 2402–8.

Muyzer, G., de Waal, E. C., and Uitterlinden, G. A. (1993). Profiling of complex populations by denaturating gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA. *Applied and Enviromental Microbiology*, **59**, 695–700.

Olsen, G. J., Lane, D. J., Giovannoni, S. J., and Pace, N. (1986). MIcrobial ecology and evolution: a ribosomal RNA approach. *Annual Review of Microbiology*, **40**, 337–65.

Pace, N. R., Stahl, D. A., Lane, D. J., and Olsen, G. J. (1986). The analysis of natural microbial populations by ribosomal RNA sequences. *Advances in Microbial Ecology*, **9**, 1–55.

Shao, J. and Tu, D. (1995). *The Jackknife and Bootstrap*. Springer, New York.

Spear, J., Walker, J., McCollom, T., and Pace, N. (2005). Hydrogen and bioenergetics in the Yellowstone geothermal ecosystem. *Proc. Natl. Acad. Sci. USA*, **102**(11), 2555–60.

Stohr, R., Waberski, A., Liesack, W., Voelker, H., Wehmeyer, U., and Thomm, M. (2001). Hydrogenophilus hirschii sp. nov., a novel thermophilic hydrogen-oxidizing beta-proteobacterium isolated from Yellowstone National Park. *Int. J. Syst. Evol. Microbiol.*, **51**(2), 481–8.

Wintzingerode, F. v., Göbel, U. B., and Stackebrandt, E. (1997). Determination of microbial diversity in environmental samples: pitfalls of PCR-based rRNA analysis. *FEMS Microbiology Reviews*, **21**(3), 213–29.

Zoetendal, E. G., Collier, C. T., Koike, S., Mackie, R. I., and Gaskins, H. R. (2004). Molecular Ecological Analysis of the Gastrointestinal Microbiota: A Review. *Journal of Nutrition*, **134**(2), 465–72.

Supplementary material

Appendix A

This appendix states the result that proves the applicability of the bootstrap procedure for the estimator $\hat{\rho}_i$.

Let's define X_j^i such that $X_j^i = 1$ if species i is present in sample j and $X_j^i = 0$ otherwise. In this case, $\hat{\rho}_i(N) = \sum_j X_j^i / \sum_{i,j} X_j^i$ (as per equation 5). Then, if X_j^i are random with finite second moments, it can be proved that the distribution of $\sqrt{N}(\hat{\rho}_i^s(N) - \hat{\rho}_i(N))$ will converge in the asymptotic limit $(N \to \infty)$ to the same distribution as $\sqrt{N}(\hat{\rho}_i(N) - \rho_i)$, namely a gaussian with variance σ_i (which depends on the average of X_j^i). More explicitly (see (Shao and Tu, 1995), example 3.3):

$$\mathrm{P}\{\sqrt{N}\,(\hat{\rho}_i^s(N) - \hat{\rho}_i(N)) < x \;\; \forall i, s\} \to_{a.s.} \Phi(x/\sigma_i) \tag{8}$$

where $\Phi(x)$ is the standard normal distribution:

$$\Phi(x) = 1/\sqrt{2\pi} \int_{-\infty}^{x} e^{-x^2/2} \tag{9}$$

Notation: $P\{A\}$ denotes probability that clause A is true and $\to_{a.s.}$ denotes almost surely convergence or convergence with probability 1 : $P\{X_n \to X\} = 1 \Rightarrow X_n \to_{a.s.} X$.

OTU # (3%)	Putative Metabolism	Closest BLAST match
25	H and S oxidation	AJ320224 (88%) (Eder and Huber, 2002) AJ320219 (88%) Eder and Huber (2002)
7	Fe(III) reduction, H oxidation, S reduction	AF335183 (88%) (Lonergan *et al.*, 1996)
8	Anoxygenic photosynthesis, Fe(II) oxidation	AJ290834 (91%) (Alexander *et al.*, 2002) Y18253 (92%) (Heising *et al.*, 1999)
36	Heterotrophic	AB062105 (98%) (Hiraishi *et al.*, 2002)
183	Heterotrophic	AF137381 (91%) (Chelius and Triplett, 2000)
181	?	No close cultivated rep.
64	?	No close cultivated rep.
51	H and S oxidation	AJ320224 (88%) (Eder and Huber, 2002) AJ320219 (88%) (Eder and Huber, 2002)
22	?	No close cultivated rep.
5	H oxidation	AB009829 (94%) (Hayashi *et al.*, 1999) AJ131694 (93%) (Stohr *et al.*, 2001)
23	H and S oxidation	AJ320224 (88%) (Eder and Huber, 2002) AJ320219 (88%) (Lonergan *et al.*, 1996)
1	?	No close cultivated rep.
55	?	No close cultivated rep.

Figure 1: Putative metabolic characteristics of the OTUs with highest cover. Each OTU was compared with Genbank (Genbank, 1982) data through BLAST (Altschul *et al.*, 1997) and assumed to have a similar metabolism to the closest matches. The third column gives the Genbank accesion numbers for best matches with known metabolism along with the percentage similarity in the 16s rRNA gene and references for each accession number. Although crude, this method gives a rough idea of the possible environmental role of each OTU.

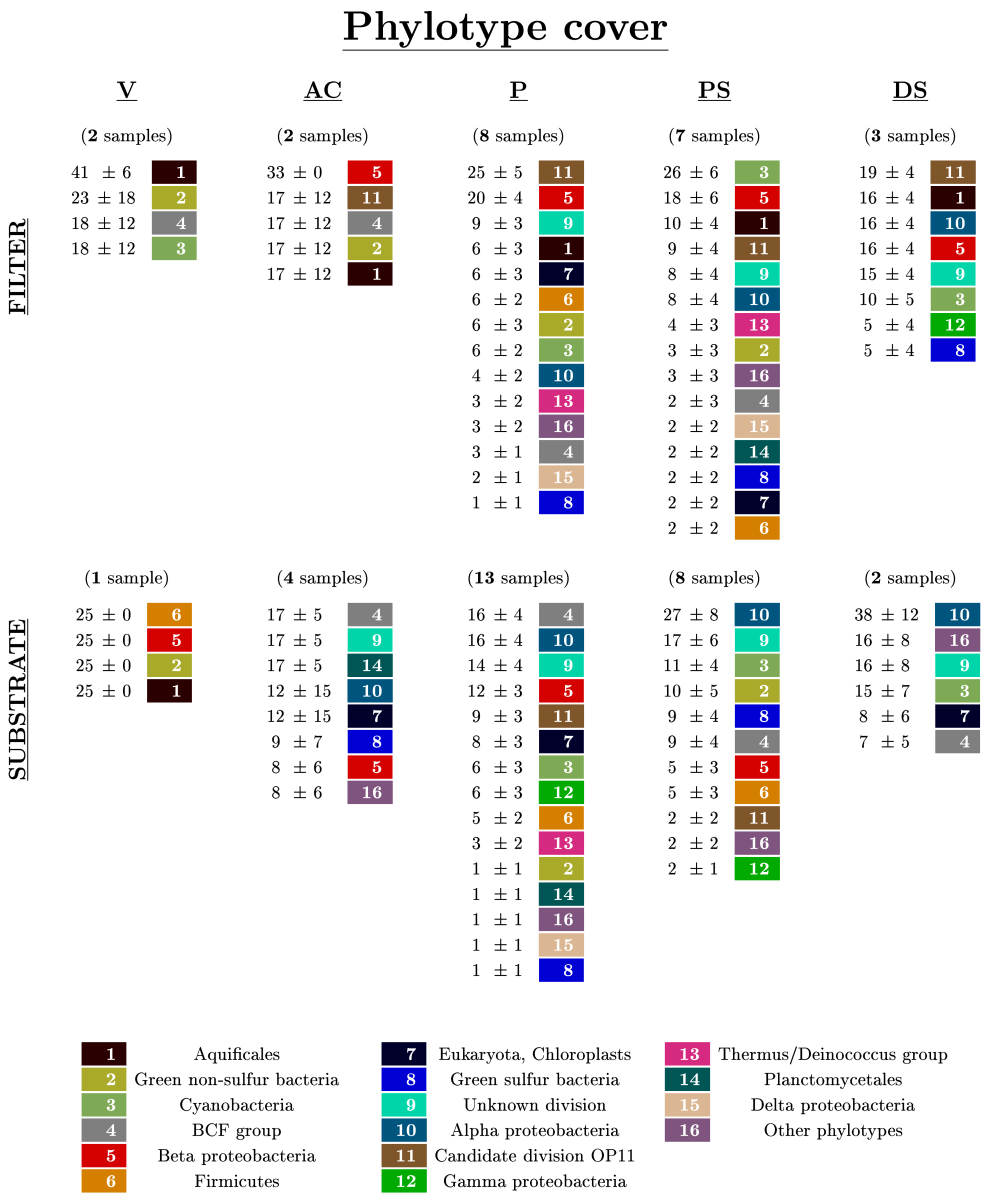

Figure 2: Phylotype relative covers and variances for each facies and medium. Each present phylotype is identified by a color throughout the whole paper. Numbers change for each grouping (phylotipes and 3%,1%,0.5% differences). The phylotype relative abudances are the sum of relative covers of OTUs belonging to a given phylotype. The variances are the square root of the sum of squared variances.

3% Definition

1	AF445736	Aquificales
2	AF445699	Green non-sulfur bacteria
3	AF446242	Cyanobacteria
4	AF446244	BCF group
5	AF445689	Beta proteobacteria
6	AF446260	Firmicutes
7	AF446272	Eukaryota, Chloroplasts
8	AF446276	Green sulfur bacteria
9	AF446331	Unknown division
10	AF446311	Alpha proteobacteria
11	AF446285	BCF group
12	AF445699	Beta proteobacteria
13	AF446280	BCF group
14	AF446328	Unknown division
15	AF446248	Planctomycetales
16	AF446291	Verrucomicrobium group
17	AF446303	Alpha proteobacteria
18	AF446345	Unknown division
19	AY293514	Planctomycetales
20	AF446287	BCF group
21	AY293515	Planctomycetales
22	AF445670	BCF group
23	AY293416	Candidate division OP11
24	AF445726	Gamma proteobacteria
25	AF445687	Candidate division OP11
26	AY293505	Unknown division
27	AF445720	Firmicutes
28	AY293507	BCF group
29	AY293499	BCF group
30	AF445644	Thermus/Deinococcus group
31	AF446288	BCF group
32	AF445721	BCF group
36	AF445711	Alpha proteobacteria
39	AF445679	Beta proteobacteria
50	AF446316	Beta proteobacteria
51	AY293536	Candidate division OP11
52	AF446333	Fibrobacteria/Acidobacteria
53	AF446262	Firmicutes
55	AF445722	Cyanobacteria
60	AY293544	Alpha proteobacteria
61	AY293543	Unknown division
62	AY293561	BCF group
63	AY293545	Eukaryota, Chloroplasts
64	AF445692	Green non-sulfur bacteria
65	AF446250	Green non-sulfur bacteria
66	AF446270	Eukaryota, Chloroplasts
71	AF446297	Alpha proteobacteria
79	AY293526	Candidate division OP11
80	AF446340	Unknown division
81	AY293525	Candidate division OP11
82	AY293557	Eukaryota, Chloroplasts
83	AF445678	Cyanobacteria
87	AY293518	Alpha proteobacteria
90	AF446265	Cyanobacteria
91	AY293477	Cyanobacteria
92	AF445724	Alpha proteobacteria
99	AY293414	Candidate division OP11
100	AF446243	Cyanobacteria
106	AF445690	Candidate division OP11
107	AF445667	Cyanobacteria
118	AF446266	Cyanobacteria
119	AF446258	Thermus/Deinococcus group
120	AY293481	Alpha proteobacteria
121	AY293404	Alpha proteobacteria
122	AF445709	Unknown division
123	AF446335	Unknown division
124	AY293483	Green non-sulfur bacteria
125	AF446259	Firmicutes
126	AF446302	Alpha proteobacteria
127	AF446332	Unknown division
128	AF446323	Unknown division
129	AF445706	Green sulfur bacteria
130	AF445712	Alpha proteobacteria
132	AF445713	Alpha proteobacteria
138	AY293417	Cyanobacteria
139	AF445743	Epsilon proteobacteria
165	AF445666	Unknown division
166	AF445677	Cyanobacteria
167	AF445738	Unknown division
174	AF446346	Unknown division
175	AY293510	Unknown division
176	AF446305	Alpha proteobacteria
177	AF445728	Unknown division
178	AF446274	Eukaryota, Mitochondria
179	AF446327	Unknown division
180	AF445714	Eukaryota, Chloroplasts
181	AF446273	Eukaryota, Mitochondria
183	AF445715	Unknown division
184	AF446310	Alpha proteobacteria
185	AY293509	Gamma proteobacteria
186	AY293495	Alpha proteobacteria
187	AF445719	Cyanobacteria
188	AF446284	BCF group
189	AF446306	Alpha proteobacteria
190	AF446312	Alpha proteobacteria
191	AF445716	Alpha proteobacteria
192	AF446241	Aquificales
193	AF446245	Alpha proteobacteria

Figure 3: OTU numbers with their corresponding defining sequence and division for 3% difference definition.

1% Definition

1	AF445736	Aquificales
2	AY293408	Aquificales
3	AF445659	Green non-sulfur bacteria
4	AF446242	Cyanobacteria
5	AF446244	BCF group
6	AF445658	Aquificales
7	AY293406	Aquificales
8	AF445689	Beta proteobacteria
9	AY293422	Aquificales
10	AY293423	Aquificales
11	AY293425	Aquificales
12	AF446260	Firmicutes
13	AY293427	Aquificales
14	AY293428	Aquificales
15	AY293429	Aquificales
16	AF446272	Eukaryota, Chloroplasts
17	AF446276	Green sulfur bacteria
18	AF446331	Unknown division
19	AF446311	Alpha proteobacteria
20	AF446285	BCF group
21	AY293516	Beta proteobacteria
22	AF446280	BCF group
23	AF446328	Unknown division
24	AF446248	Planctomycetales
25	AF446291	Verrucomicrobium group
26	AF446303	Alpha proteobacteria
27	AF446345	Unknown division
28	AY293514	Planctomycetales
29	AF446287	BCF group
30	AY293515	Planctomycetales
31	AY293411	Aquificales
32	AF445670	BCF group
33	AY293461	Candidate division OP11
35	AF445687	Candidate division OP11
37	AF445720	Firmicutes
38	AY293507	BCF group
48	AF445711	Alpha proteobacteria
52	AF445679	Beta proteobacteria
64	AF446316	Beta proteobacteria
65	AY293536	Candidate division OP11
66	AF446333	Fibrobacteria/Acidobacteria
67	AF446262	Firmicutes
69	AF445722	Cyanobacteria
72	AF445699	Beta proteobacteria
75	AY293540	Candidate division OP11
77	AY293543	Unknown division
80	AY293416	Candidate division OP11
83	AY293531	Unknown division
100	AY293526	Candidate division OP11
101	AF446340	Unknown division
102	AY293525	Candidate division OP11
103	AY293557	Eukaryota, Chloroplasts
104	AF445678	Cyanobacteria
108	AY293518	Alpha proteobacteria
112	AF446265	Cyanobacteria
113	AY293477	Cyanobacteria
114	AF445724	Alpha proteobacteria
121	AY293414	Candidate division OP11
122	AF446243	Cyanobacteria
132	AY293464	Aquificales
135	AF445667	Cyanobacteria
136	AF445692	Green non-sulfur bacteria
150	AY293473	Aquificales
151	AY293479	Beta proteobacteria
152	AY293480	Candidate division OP11
153	AF446266	Cyanobacteria
154	AF446258	Thermus/Deinococcus group
155	AY293481	Alpha proteobacteria
156	AY293404	Alpha proteobacteria
157	AF445709	Unknown division
158	AF446335	Unknown division
159	AY293483	Green non-sulfur bacteria
160	AF446259	Firmicutes
161	AF446302	Alpha proteobacteria
162	AY293491	Alpha proteobacteria
163	AY293492	Alpha proteobacteria
164	AY293493	Beta proteobacteria
165	AF446332	Unknown division
168	AF445706	Green sulfur bacteria
169	AF445712	Alpha proteobacteria
171	AF445713	Alpha proteobacteria
178	AY293417	Cyanobacteria
179	AF445743	Epsilon proteobacteria
210	AF445738	Unknown division
222	AF446274	Eukaryota, Mitochondria
223	AF446327	Unknown division
224	AF445714	Eukaryota, Chloroplasts
225	AF446273	Eukaryota, Mitochondria
226	AF446298	Alpha proteobacteria
227	AF445715	Unknown division
230	AY293495	Alpha proteobacteria
231	AF445719	Cyanobacteria
232	AF446284	BCF group
233	AF446306	Alpha proteobacteria
234	AF446312	Alpha proteobacteria
235	AF445716	Alpha proteobacteria

Figure 4: OTU numbers with their corresponding defining sequence and division for 1% difference definition.

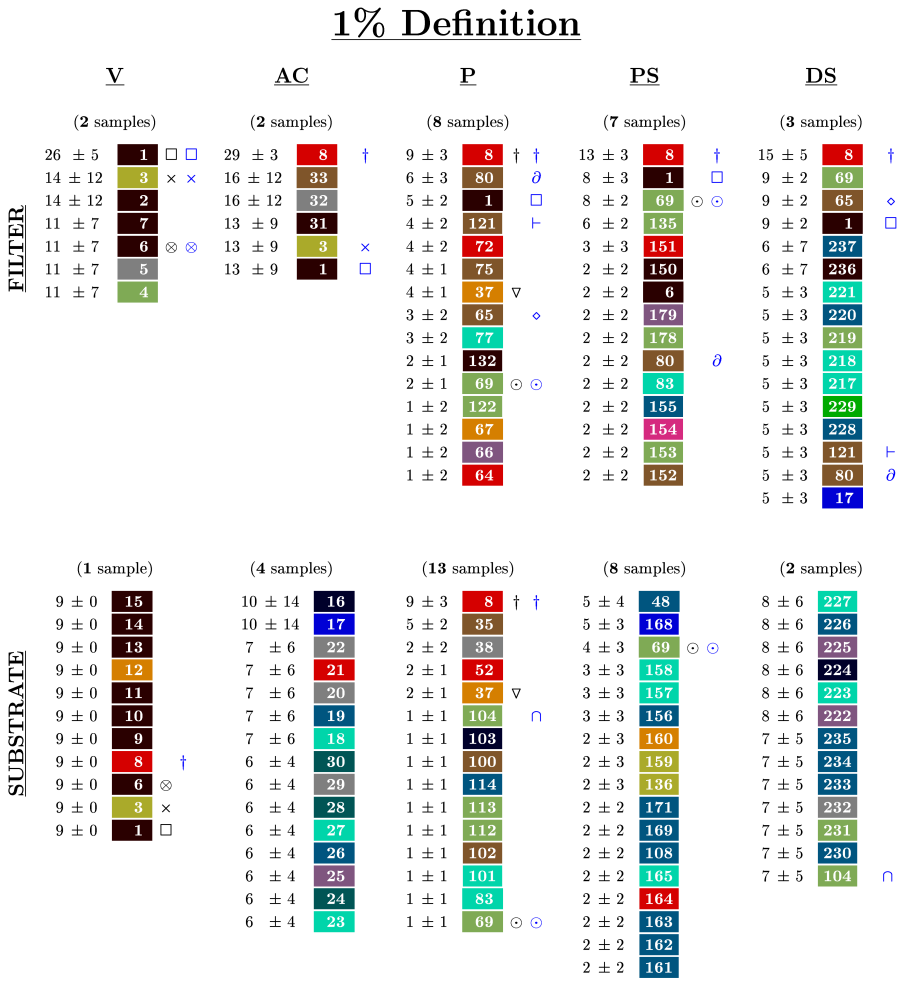

Figure 5: OTU covers for the 1% difference definition. For reasons of space only the OTUs with highest cover are shown.

1% Definition

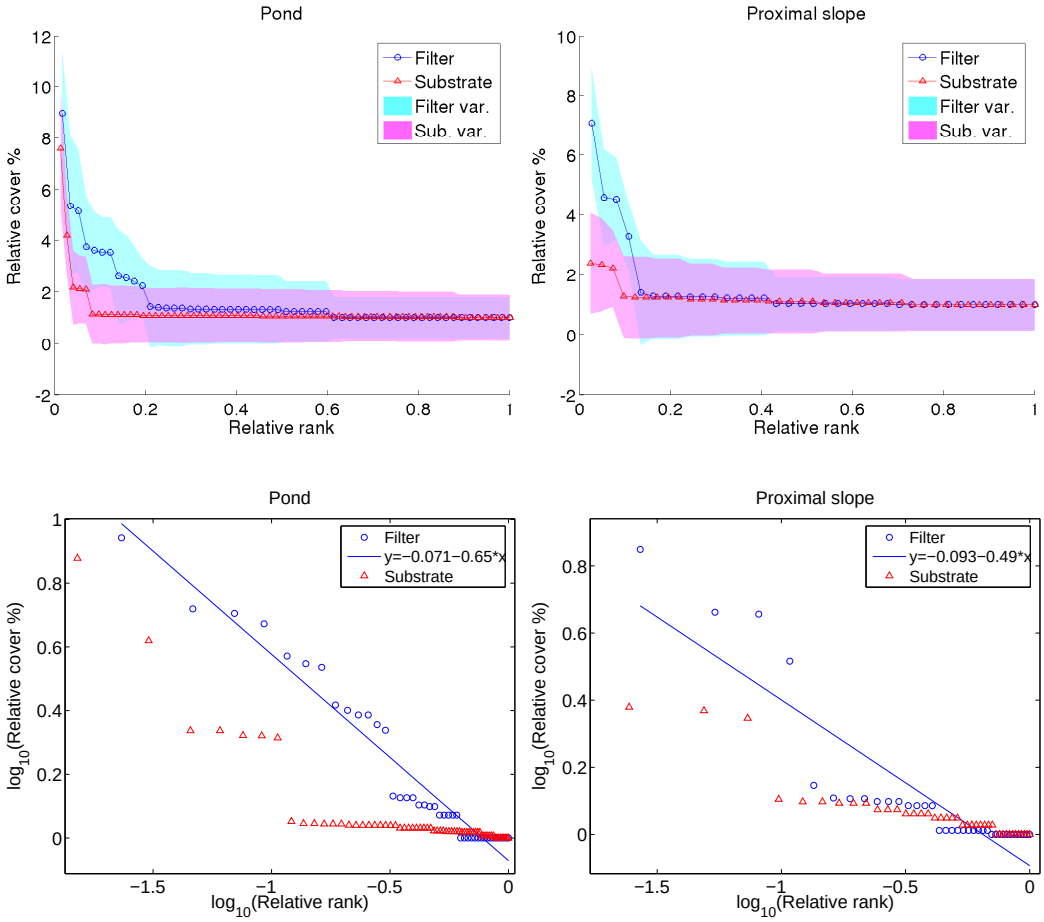

Figure 6: Plots of relative covers versus relative rank for the 1% difference definition in normal (above) and log-log axis (below).

0.5% Definition

1	AF445697	Aquificales
2	AY293408	Aquificales
3	AY293412	Aquificales
4	AY293409	Aquificales
5	AF445736	Aquificales
6	AF445659	Green non-sulfur bacteria
7	AF446242	Cyanobacteria
8	AF446244	BCF group
9	AY293407	Aquificales
10	AY293406	Aquificales
11	AY293420	Beta proteobacteria
12	AY293421	Aquificales
13	AY293422	Aquificales
14	AY293475	Aquificales
15	AY293424	Aquificales
16	AY293425	Aquificales
17	AF446260	Firmicutes
18	AY293426	Aquificales
19	AY293427	Aquificales
20	AY293474	Aquificales
21	AY293428	Aquificales
22	AY293429	Aquificales
23	AF445658	Aquificales
24	AF446272	Eukaryota, Chloroplasts
25	AF446276	Green sulfur bacteria
26	AF446331	Unknown division
27	AF446311	Alpha proteobacteria
28	AF446285	BCF group
29	AY293516	Beta proteobacteria
30	AF446280	BCF group
31	AF446328	Unknown division
32	AF446248	Planctomycetales
33	AF446291	Verrucomicrobium group
34	AF446303	Alpha proteobacteria
35	AF446345	Unknown division
36	AY293514	Planctomycetales
37	AF446287	BCF group
38	AY293515	Planctomycetales
39	AF445689	Beta proteobacteria
49	AF445687	Candidate division OP11
53	AY293507	BCF group
69	AF445679	Beta proteobacteria
70	AY293502	Thermus/Deinococcus group
71	AF446261	Firmicutes
86	AY293536	Candidate division OP11
92	AF445745	Firmicutes
94	AY293466	Beta proteobacteria
99	AY293543	Unknown division
102	AY293416	Candidate division OP11
105	AY293546	Candidate division OP11
106	AF445741	Beta proteobacteria
131	AY293526	Candidate division OP11
134	AF445722	Cyanobacteria
143	AY293557	Eukaryota, Chloroplasts
144	AF445678	Cyanobacteria
149	AY293518	Alpha proteobacteria
154	AF446265	Cyanobacteria
155	AY293477	Cyanobacteria
156	AF445724	Alpha proteobacteria
163	AY293414	Candidate division OP11
164	AF446243	Cyanobacteria
165	AY293413	Candidate division OP11
169	AY293415	Beta proteobacteria
181	AY293464	Aquificales
182	AF445739	Aquificales
190	AF445692	Green non-sulfur bacteria
212	AY293479	Beta proteobacteria
213	AY293480	Candidate division OP11
214	AF446266	Cyanobacteria
215	AF446258	Thermus/Deinococcus group
216	AY293481	Alpha proteobacteria
218	AY293404	Alpha proteobacteria
219	AF445709	Unknown division
220	AF446308	Alpha proteobacteria
221	AF446335	Unknown division
222	AY293483	Green non-sulfur bacteria
223	AY293482	Green non-sulfur bacteria
224	AF446259	Firmicutes
236	AF445706	Green sulfur bacteria
237	AF445712	Alpha proteobacteria
239	AF445713	Alpha proteobacteria
303	AY293513	Beta proteobacteria
304	AY293512	Beta proteobacteria
305	AF446346	Unknown division
306	AY293510	Unknown division
307	AY293511	Cyanobacteria
308	AY293532	Aquificales
309	AF446305	Alpha proteobacteria
310	AF445728	Unknown division
311	AY293494	Eukaryota, Mitochondria
312	AF446327	Unknown division
313	AF445714	Eukaryota, Chloroplasts
314	AF446273	Eukaryota, Mitochondria
315	AF446274	Eukaryota, Mitochondria
316	AF446298	Alpha proteobacteria
317	AF446275	Eukaryota, Mitochondria
318	AF445715	Unknown division
324	AY293495	Alpha proteobacteria
325	AF445719	Cyanobacteria
326	AF446284	BCF group
327	AF446306	Alpha proteobacteria
328	AF446312	Alpha proteobacteria
329	AF445716	Alpha proteobacteria
330	AF446241	Aquificales
331	AF446245	Alpha proteobacteria

Figure 7: OTU numbers with their corresponding defining sequence and division for 0.5% difference definition.

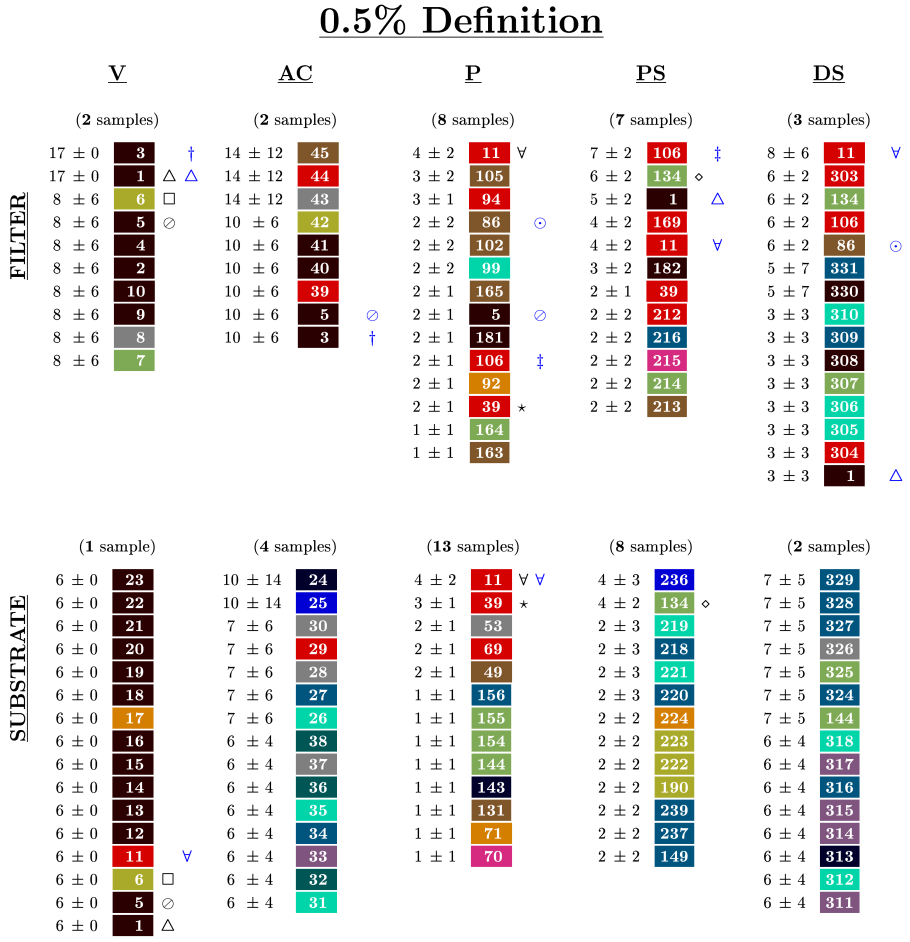

Figure 8: OTU covers for the 0.5% difference definition. For reasons of space only the OTUs with highest covers are shown.

0.5% Definition

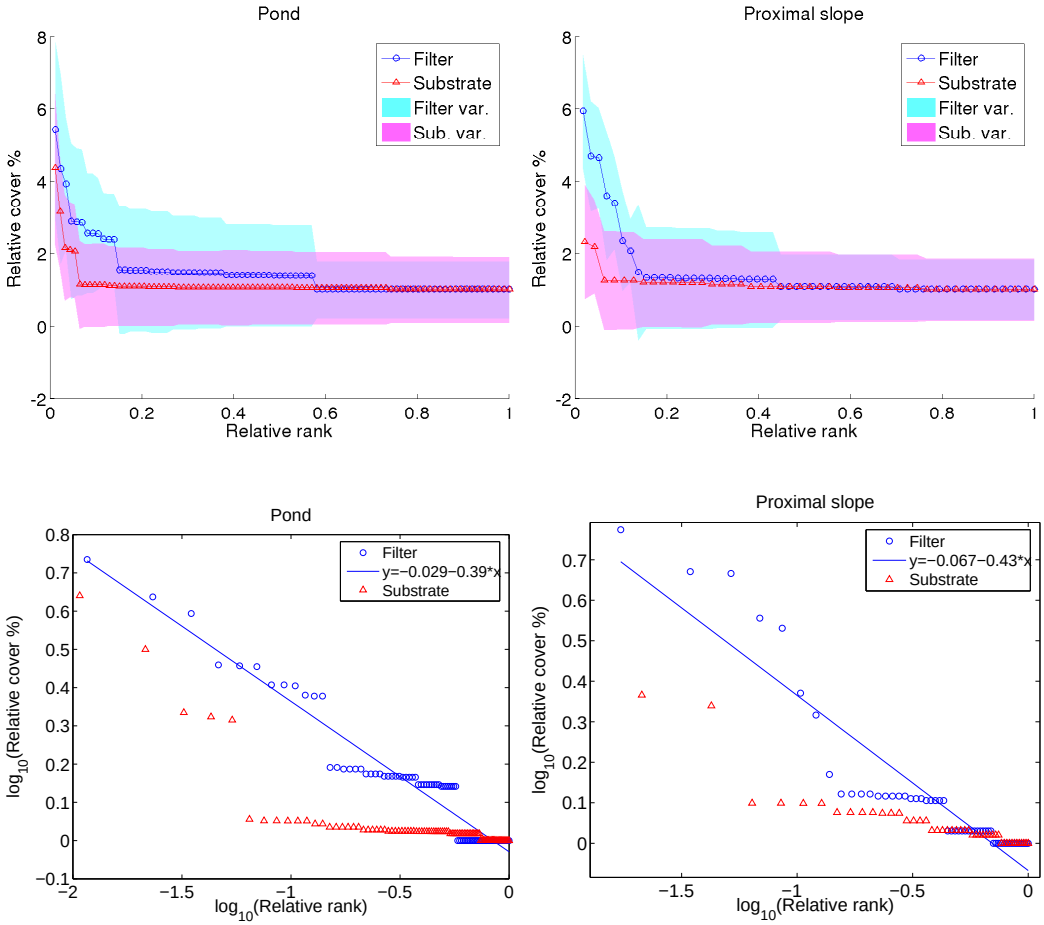

Figure 9: Plots of relative covers versus relative rank for the 0.5% difference definition in normal (above) and log-log axis (below).