doi:10.1111/j.1462-2920.2006.01003.x

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

Héctor García Martín[1†] and Nigel Goldenfeld[1,2*]
[1]*Department of Physics and* [2]*Institute for Genomic Biology, University of Illinois at Urbana-Champaign, 1110 West Green Street, Urbana, IL 61801, USA.*

Summary

We propose the use of cover as a quick, low-resolution proxy for the abundance of microbial species, which reduces polymerase chain reaction 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.

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, while expensive, is becoming cheaper and high-throughput methods are available that enable huge data sets 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 biogeochemical cycles requires information about abundance. A variety of methods are available to measure abundance: Quantitative polymerase chain reaction (PCR), Most Probable Number PCR, competition PCR and dot-blot hybridization among others (Muyzer *et al*., 1993; Amann *et al*., 1995; Head *et al*., 1998; Ding and Cantor, 2004; 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 article 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, when used with clone libraries of large enough size. 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).

Accepted 17 January, 2006. *For correspondence. E-mail nigel@uiuc.edu; Tel. (+1) 217 333 8027; Fax (+1) 217 333 9819. †Present address: DOE Joint Genome Institute, Walnut Creek, CA 94598, USA.

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.

Relative cover estimation through library resampling

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 *Study site*) 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. For this procedure to eliminate the aforementioned biases it is necessary for the number of clones sequenced per library in each sample (sample size) to be sufficient to allow the detection of all relevant OTUs. Appendix S2 in *Supplementary material* offers an estimation of the required sample size depending on the expected biases and the minimum relative abundance needed for an OTU to be considered relevant.

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.

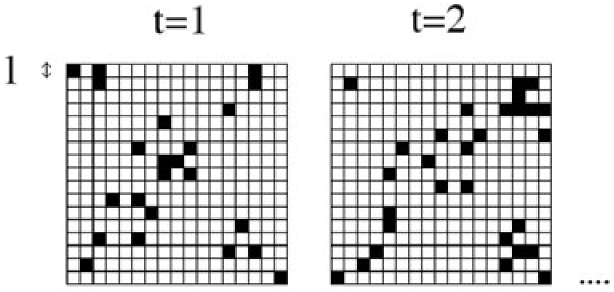


Fig. 1. Cover, a concept borrowed from macroscopic ecology (Kunin, 1998; 2000; He and Gaston, 2000), is a coarse-grained or low-resolution measure of abundance in the sense that each subcell will contribute if the species is present inside it, independent of its abundance in the subcell. The cover is the number of subcells in which a given OTU is present divided by the total number of subcells (Eq. 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.

To this end, we use the collected data to obtain estimates of coarse-grained abundances or covers as explained in Fig. 1. Cover, sometimes known as occurrence or range, is a concept from macroscopic ecology, and is strongly linked to abundance (Kunin, 1998; 2000; He and Gaston, 2000) although not equivalent. Referring to Fig. 1, 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 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. S1 in *Supplementary material*). 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 *Analysis of Yellowstone National Park field data*, 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 than 14 000 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.

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 (Efron and Tibshirani, 1993; Shao and Tu, 1995; 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 Eq. 5. The variance would be obtained as usual: $var(\hat{\rho}_i) = \Sigma_s(\hat{\rho}_i^s - \hat{\rho}_i^{av})^2$, 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 *Analysis of Yellowstone National Park field data*) having four groups would lead to a meager two 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 Fig. 2: 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 S1 in *Supplementary mate-*

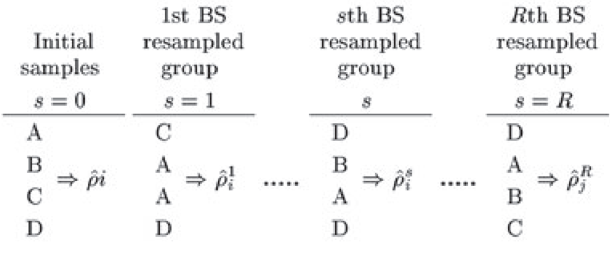


Fig. 2. The data resampling bootstrap method applied to estimate the bias and variance of $\hat{\rho}_i$. R groups of four samples are generated by sampling with replacement from the original samples. This means that the samples in each group are chosen randomly among the original samples and each time 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 Eq. 5 and the estimate of the bias and the variance are given by Eqs 6 and 7.

rial 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 Appendix S1 in *Supplementary material*).

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. While the theorem in Appendix S1 (*Supplementary material*) proves the consistency of the data resampling bootstrap estimator in the asymptotic limit, the real interest of the bootstrap lies in its fixed sample properties. The artificial data have been constructed so that it mimics some aspects of the field data we will eventually analyse 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.

For this demonstration calculation, we assume that each of the S OTUs present in the system are present in each sample with probability $\rho(i) \propto i^{-d}$, where different distributions with $d = 2$, 1.5, 1, 0.65, 0.3 are considered ($i = 1.S$). As we will see, this model distribution for $d = 0.65$ actually mimics the cover distribution of microbes that, in *Analysis of Yellowstone National Park field data*, 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 $S = 200$ because 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.

The OTUs present in each sample are generated though a Montecarlo algorithm, as explained in *Supplementary material*.

Figure 3 shows the results of the data resampling bootstrap estimates $\hat{\rho}_i(N)$ for $d = 0.65$, sample numbers $N = 10$, 100 and $R = 10\,000$ as compared with 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 because not all S OTUs have been detected and therefore the detected OTUs are given a higher relative cover than the real one. R is in practice chosen large enough so that further increases don't change the estimate appreciably.

The choice of different values of d allows us to explore the robustness of the method and the required number of samples for accurate cover estimation as a function of the steepness of the rank-cover distribution. As can be observed in Fig. S12 in *Supplementary material* similar results are found for the different exponents with inaccuracies showing for low ρ_i OTUs, the number of which is different for each exponent. As a rule of thumb, to obtain an adequate estimation of relative cover ρ_i the necessary number of samples can be calculated to be (see *Supplementary material*):

$$N >> \frac{1}{\rho_i \sum_j C_j} \tag{8}$$

where C_j is the cover as defined before. In practice, for the studied distributions, $N > 1.2/(\rho\Sigma_j C_j)$ seems sufficient to obtain an adequate estimate of ρ_i.

Analysis of Yellowstone National Park field data

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; G. Bonheyo *et al*., submitted) as part of a large biocomplexity study at the University of Illinois at Urbana-Champaign. Our purpose here is to illustrate how we have analysed 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 data set 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*.,

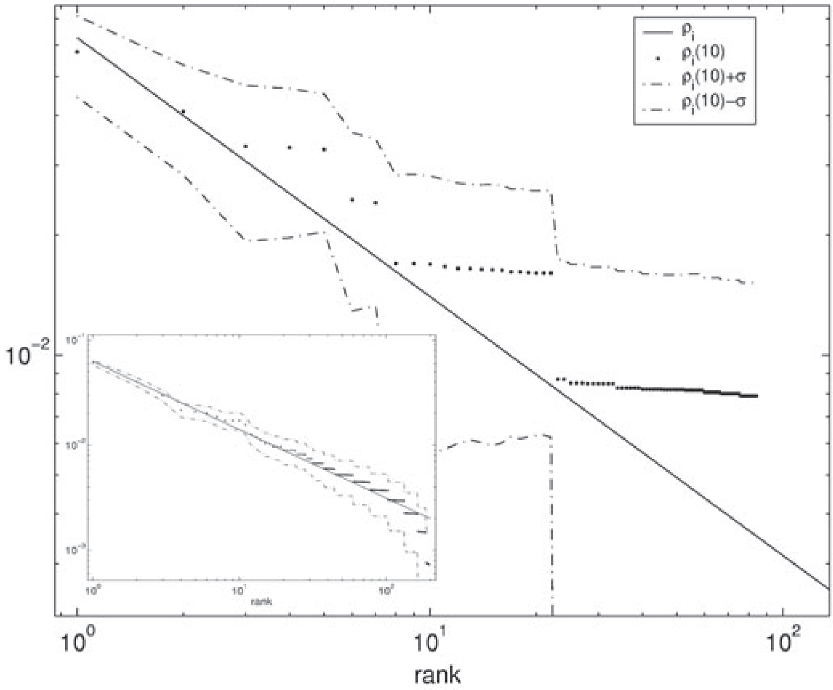


Fig. 3. Data resampling bootstrap estimate for $N = 10$ samples. In spite of this low number of samples it is possible to get a hint of the underlying distribution. The estimate overshoots because for such low amount of samples not all OTUs have been detected and therefore the detected OTUs attain a higher relative cover so the sum of the relative covers adds up to unity. The inset shows how the bootstrap estimate improves for $N = 100$ as expected and approximates satisfactorily the target distribution.

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 (~1.5 m year^{-1}) 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 (G. Bonheyo *et al*., submitted).

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 restriction fragment length polymorphism (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 (G. Bonheyo *et al*., submitted).

Data resampling bootstrap estimates

The procedure for obtaining the cover ρ_k^{BS} is the same as explained above with a total of $R = 10\,000$ bootstrap samples being used. The results are given in the form of rank–abundance plots in Fig. 4 and Figs S8 and S11 in *Supplementary material*. Rank tables for all facies and mediums are shown in Fig. 5 and Figs S7 and S10 in *Supplementary material*, along with the number of samples for each case. A rank table with phylotype relative covers is available in Fig. S2 in *Supplementary material*.

Table 1 presents a comparison of covers obtained

Table 1. Comparisons of covers and clone relative abundances from fig. 6 in Bonheyo and colleagues (submitted) for the proximal slope facies.

Phylotype	Clone abundance (%)	Cover estimate (%)
Beta-proteobacteria	22	10 ± 4
Cyanobacteria	16	18 ± 3
Aquificales	15	4 ± 2
Alpha proteobacteria	11	17 ± 4
Unknown division	9	14 ± 4
Green sulfur bacteria	9	5 ± 2
BCF group	7	9 ± 3
Delta proteobacteria	3	3 ± 1
Candidate division OP-11	2	5 ± 2
Green non-sulfur bacteria	2	5 ± 3
Thermus/deinococcus group	1	2 ± 2
Gamma proteobacteria	1	1 ± 1
Firmicutes	Negligible	3 ± 2
Eukaryota, chloroplasts	Negligible	1 ± 1

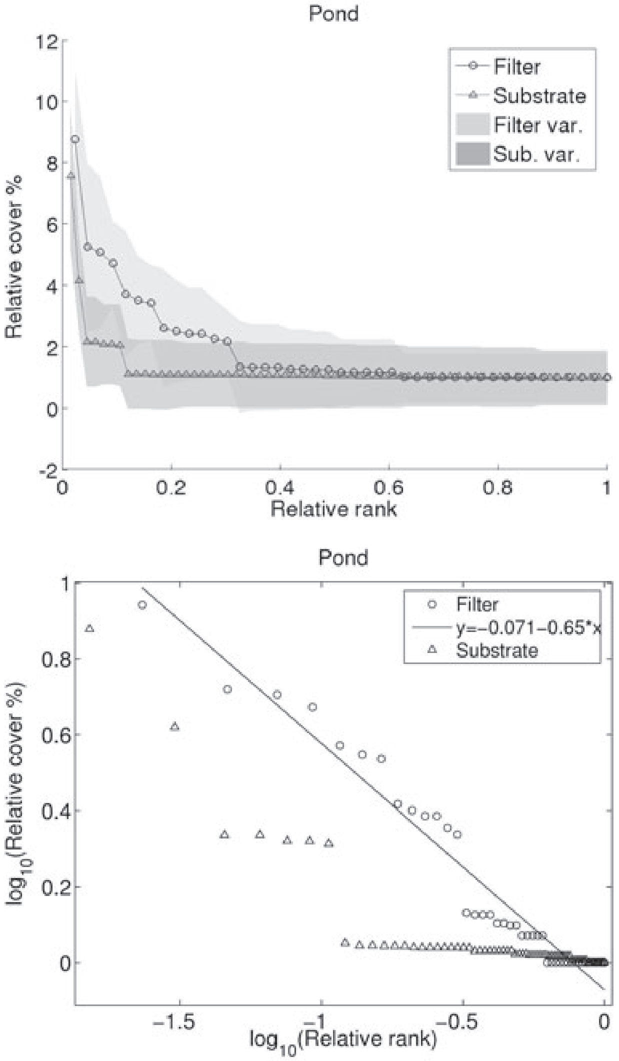


Fig. 4. Plots of relative cover versus relative rank for the 3% definition in normal and log–log axis. Relative covers are covers divided by the lowest cover. An OTU with rank *i* has the *i*th highest cover. Relative rank is rank divided by the total number of OTUs. The filter samples from the pond 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). The substrate sample plot curves upwards.

through the resampling method and nominal clone abundances (fig. 6 in G. Bonheyo *et al*., submitted) 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 over-represented by a factor of more than three and beta-proteobacteria by a factor of two.

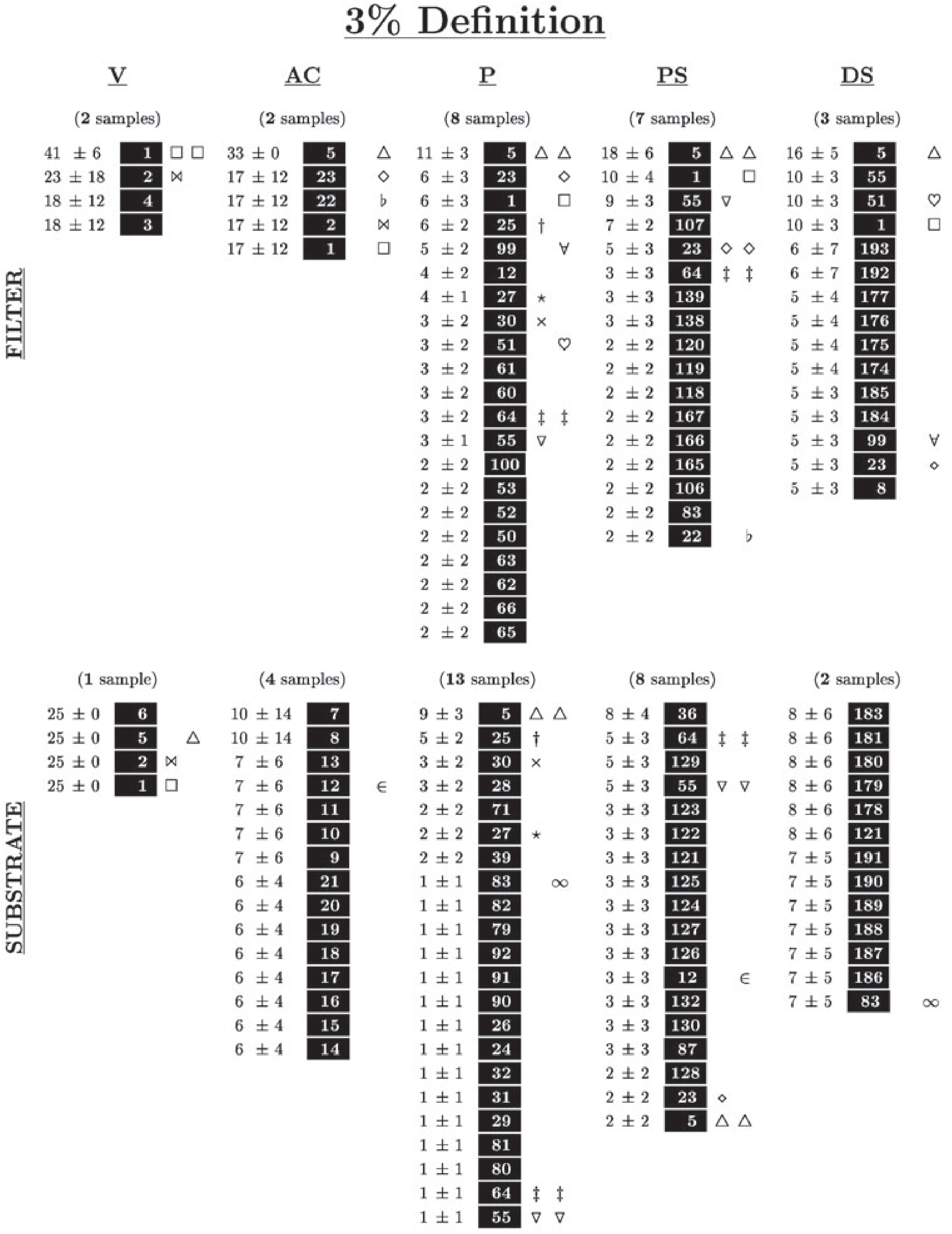
As can be seen, only the Pond and Proximal Slope facies have enough number of samples for the resulting covers to be statistically meaningful. Therefore, only cover distributions for these mediums and facies are presented. In the case of the ranked tables, nonetheless, even for three or four samples 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; G. Bonheyo *et al*., submitted). 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. As 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.

Table 2 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. Operational taxonomical unit 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.



3% Definition

FILTER

V				AC				P				PS				DS			
(2 samples)				(2 samples)				(8 samples)				(7 samples)				(3 samples)			
41 ± 6	1	□	□	33 ± 0	5		△	11 ± 3	5	△	△	18 ± 6	5	△	△	16 ± 5	5		△
23 ± 18	2	⋈		17 ± 12	23		◇	6 ± 3	23		◇	10 ± 4	1		□	10 ± 3	55		
18 ± 12	4			17 ± 12	22		♭	6 ± 3	1		□	9 ± 3	55	∇		10 ± 3	51		♡
18 ± 12	3			17 ± 12	2		⋈	6 ± 2	25	†		7 ± 2	107			10 ± 3	1		□
				17 ± 12	1		□	5 ± 2	99		∀	5 ± 3	23	◇	◇	6 ± 7	193		
								4 ± 2	12			3 ± 3	64	‡	‡	6 ± 7	192		
								4 ± 1	27	⋆		3 ± 3	139			5 ± 4	177		
								3 ± 2	30	×		3 ± 3	138			5 ± 4	176		
								3 ± 2	51		♡	2 ± 2	120			5 ± 4	175		
								3 ± 2	61			2 ± 2	119			5 ± 4	174		
								3 ± 2	60			2 ± 2	118			5 ± 3	185		
								3 ± 2	64	‡	‡	2 ± 2	167			5 ± 3	184		
								3 ± 1	55	∇		2 ± 2	166			5 ± 3	99		∀
								2 ± 2	100			2 ± 2	165			5 ± 3	23		◇
								2 ± 2	53			2 ± 2	106			5 ± 3	8		
								2 ± 2	52			2 ± 2	83						
								2 ± 2	50			2 ± 2	22		♭				
								2 ± 2	63										
								2 ± 2	62										
								2 ± 2	66										
								2 ± 2	65										

SUBSTRATE

(1 sample)				(4 samples)				(13 samples)				(8 samples)				(2 samples)			
25 ± 0	6			10 ± 14	7			9 ± 3	5	△	△	8 ± 4	36			8 ± 6	183		
25 ± 0	5		△	10 ± 14	8			5 ± 2	25	†		5 ± 3	64	‡	‡	8 ± 6	181		
25 ± 0	2	⋈		7 ± 6	13			3 ± 2	30	×		5 ± 3	129			8 ± 6	180		
25 ± 0	1	□		7 ± 6	12		∈	3 ± 2	28			5 ± 3	55	∇	∇	8 ± 6	179		
				7 ± 6	11			2 ± 2	71			3 ± 3	123			8 ± 6	178		
				7 ± 6	10			2 ± 2	27	⋆		3 ± 3	122			8 ± 6	121		
				7 ± 6	9			2 ± 2	39			3 ± 3	121			7 ± 5	191		
				6 ± 4	21			1 ± 1	83		∞	3 ± 3	125			7 ± 5	190		
				6 ± 4	20			1 ± 1	82			3 ± 3	124			7 ± 5	189		
				6 ± 4	19			1 ± 1	79			3 ± 3	127			7 ± 5	188		
				6 ± 4	18			1 ± 1	92			3 ± 3	126			7 ± 5	187		
				6 ± 4	17			1 ± 1	91			3 ± 3	12		∈	7 ± 5	186		
				6 ± 4	16			1 ± 1	90			3 ± 3	132			7 ± 5	83		∞
				6 ± 4	15			1 ± 1	26			3 ± 3	130						
				6 ± 4	14			1 ± 1	24			3 ± 3	87						
								1 ± 1	32			2 ± 2	128						
								1 ± 1	31			2 ± 2	23	◇					
								1 ± 1	29			2 ± 2	5	△	△				
								1 ± 1	81										
								1 ± 1	80										
								1 ± 1	64	‡	‡								
								1 ± 1	55	∇	∇								

Fig. 5. Operational taxonomical units with highest coverage for the 3% difference definition for different facies and media. V, vent; AC, apron channel; P, pond; PS, proximal slope; DS, distal slope (Fouke *et al*., 2003). Figures are relative covers with their variances. Numbers are identification OTU numbers given in Fig. S3 in *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. Colours indicate phylotypes according to the code in Fig. S2 in *Supplementary material*. For reasons of space only the OTUs with highest covers are shown.

Table 2. Putative metabolic characteristics of the OTUs with highest cover.

OUT no. (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.

Each OTU was compared with GenBank (http://www.ncbi.nlm.nih.gov/Genbank/index.html) data through BLAST (Altschul *et al.*, 1997) and assumed to have a similar metabolism to the closest matches. The third column gives the GenBank accession numbers for best matches with known metabolism along with the percentage similarity in the 16 s rRNA gene and references for each accession number. Although crude, this method gives a rough idea of the possible environmental role of each OTU.

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 (for example) 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. Operational taxonomical units 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 seem 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, because 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 three 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.

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:** 131–140.

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:** 3389–3402.

Amann, R.I., Ludwig, W., and Schleifer, K.-H. (1995) Phylogenetic identification and *in situ* detection of individual microbial cells without cultivation. *FEMS Microbiol Rev* **59:** 143–169.

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:** 751–758.

Chernick, M.R. (1999) *Bootstrap Methods, A Practitioner Guide*. New York, NY, USA: Wiley.

Ding, C., and Cantor, C.R. (2004) Quantitative analysis of nucleic acids – the last few years of progress. *J Biochem Mol Biol* **37:** 1.

Eder, W., and Huber, R. (2002) New isolates and physiological properties of the aquificales and description of *Thermocrinis albus* sp. nov. *Extremophiles* **6:** 309–318.

Efron, B. (1979) Bootstrap methods: another look at the jackknife. *Ann Stat* **7:** 1–26.

Efron, B., and Tibshirani, R.J. (1993) *An Introduction to the Bootstrap*. Boca Raton, FL, USA: Chapman & Hall/CRC.

Felsenstein, J. (1985) Confidence limits on phylogenetics: an approach using the bootstrap. *Evolution* **39:** 783–791.

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). *J Sediment Res* **70:** 565–585.

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. *Can J Earth Sci* **40:** 1531–1548.

García Martín, H. (2004) *Statistical Analysis of Highly Correlated Systems in Biology and Physics*. PhD Thesis, University of Illinois at Urbana-Champaign, Department of Physics.

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:** 783–786.

He, F., and Gaston, K.J. (2000) Estimating species abundance from occurrence. *Am Nat* **156:** 553–559.

Head, I., Saunders, J., and Pickup, R. (1998) Microbial evolution, diversity and ecology: a decade of ribosomal RNA analysis of uncultivated microorganisms. *Microbial Ecol* **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:** 116–124.

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:** 45–52.

Kunin, W.E. (1998) Extrapolating species abundance across spatial scales. *Science* **281:** 1513–1515.

Kunin, W.E. (2000) Scaling down: on the challenge of estimating abundance from occurrence patterns. *Am Nat* **156:** 560–566.

Lonergan, D., Lenter, H., Coates, J., Phillips, J., Schmidt, T., and Lovley, D. (1996) Phylogenetic analysis of dissimilatory Fe(III)-reducing bacteria. *J Bacteriol* **178:** 2402–2408.

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. *Appl Environ Microbiol* **59:** 695–700.

Olsen, G.J., Lane, D.J., Giovannoni, S.J., and Pace, N. (1986) Microbial ecology and evolution: a ribosomal RNA approach. *Annu Rev Microbiol* **40:** 337–365.

Pace, N.R., Stahl, D.A., Lane, D.J., and Olsen, G.J. (1986) The analysis of natural microbial populations by ribosomal RNA sequences. *Adv Microb Ecol* **9:** 1–55.

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

Spear, J., Walker, J., McCollom, T., and Pace, N. (2005) Hydrogen and bioenergetics in the Yellowstone geothermal ecosystem. *Proc Natl Acad Sci USA* **102:** 2555–2560.

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:** 481–488.

Wintzingerode, F.V., Göbel, U.B., and Stackebrandt, E. (1997) Determination of microbial diversity in environmental samples: pitfalls of PCR-based rRNA analysis. *FEMS Microbiol Rev* **21:** 213–229.

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. *J Nutr* **134:** 465–472.

Supplementary material

The following supplementary material is available for this article online:

Appendix S1. Bootstrap applicability
Appendix S2. Clone library size and cover estimates
Appendix S3. Montecarlo generation of probability distributions
Appendix S4. Finite size effects on relative cover estimation

Fig. S1. 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 Eqs 4 and 5. Only information of presence or absence of a given OTU is used.
Fig. S2. Phylotype relative covers and variances for each facies and medium. Each present phylotype is identified by a colour throughout the whole paper. Numbers change for each grouping (phylotypes 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.
Fig. S3. OTU numbers with their corresponding defining sequence and division for 3% difference definition.
Fig. S4. Colour version of Fig. 5.
Fig. S5. 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 *i*th 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.

Fig. S6. OTU numbers with their corresponding defining sequence and division for 1% difference definition.

Fig. S7. OTU covers for the 1% difference definition. For reasons of space only the OTUs with highest cover are shown.

Fig. S8. Plots of relative covers versus relative rank for the 1% difference definition in normal (above) and log–log axis (below).

Fig. S9. OTU numbers with their corresponding defining sequence and division for 0.5% difference definition.

Fig. S10. OTU covers for the 0.5% difference definition. For reasons of space only the OTUs with highest covers are shown.

Fig. S11. Plots of relative covers versus relative rank for the 0.5% difference definition in normal (above) and log–log axis (below).

Fig. S12. Data resampling bootstrap estimates for $d = 0.3$, 1, 2 (top to bottom). Results are similar as for the $d = 0.65$ case.

This material is available as part of the online article from http://www.blackwell-synergy.com