Bacterial diversity in aquatic and other environments: what 16S rDNA libraries can tell us

Paul F. Kemp *, Josephine Y. Aller

Marine Sciences Research Center, Stony Brook University, Stony Brook, NY 11794-5000, USA

Received 13 June 2003; received in revised form 19 September 2003; accepted 28 September 2003


Abstract

We evaluate the substantial amount of information accumulated on bacterial diversity in a variety of environments and address several fundamental questions, focusing on aquatic systems but including other environments to provide a broader context. Bacterial diversity data were extracted from 225 16S rDNA libraries described in published reports, representing a variety of aquatic and non-aquatic environments. Libraries were predominantly composed of rare phylotypes that appeared only once or twice in the library, and the number of phylotypes observed was correlated with library size (implying that few libraries are exhaustive samples of diversity in the source community). Coverage, the estimated proportion of phylotypes in the environment represented in the library, ranged widely but on average was remarkably high and not correlated with library size. Phylotype richness was calculated by methods based on the frequency of occurrence of different phylotypes in 194 libraries that provided appropriate data. For 90% of aquatic-system libraries, and for 79% of non-aquatic libraries, the estimated phylotype richness was < 200 phylotypes. Nearly all of the larger estimates were in aquatic sediments, digestive systems and soils. However, the approaches used to estimate phylotype richness may yield underestimates when libraries are too small. A procedure is described to provide an objective means of determining when a library is large enough to provide a stable and unbiased estimate of phylotype richness. A total of 56 libraries, including 44 from aquatic systems, were considered ‘large enough’ to yield stable estimates suitable for comparing richness among environments. Few significant differences in phylotype richness were observed among aquatic environments. For one of two richness estimators, the average phylotype richness was significantly lower in hyperthermal environments than in sediment and bacterioplankton, but no other significant differences among aquatic environments were observed. In general, and with demonstrated exceptions, published studies have captured a large fraction of bacterial diversity in aquatic systems. In most cases, the estimated bacterial diversity is lower than we would have expected, although many estimates should be considered minimum values. We suggest that on local scales, aquatic bacterial diversity is much less than any predictions of their global diversity, and remains a tractable subject for study. The global-scale diversity of aquatic Bacteria, on the other hand, may be beyond present capabilities for effective study.

© 2003 Federation of European Microbiological Societies. Published by Elsevier B.V. All rights reserved.

Keywords: Bacteria; Diversity; Species richness; Aquatic

1. Introduction

Aquatic Bacteria are notoriously difficult to obtain in culture, and consequently the biogeochemical roles of aquatic Bacteria have been studied using black-box techniques such as ATP content, epifluorescence direct counts, DNA and protein synthesis rates, frequency of dividing cells, and a host of other methods that are inherently blind to variations in community composition. This blind spot in the aquatic microbial ecologist’s perception is universally recognized, and probably has been universally frustrating.

The advent of molecular methods, especially ribosomal RNA-based techniques, led to an explosion of microbial biodiversity papers starting in the late 1980s (see review by Morris et al. [1]). Much of what is now known of the diversity of aquatic Bacteria is based on distinguishing among different organisms, as represented by their extracted and polymerase chain reaction (PCR)-amplified nucleic acids, without actually culturing them or having any direct knowledge of their morphology, physiology or ecology. Molecular techniques have revealed a surprisingly high prokaryotic diversity. Some studies have reported an...
unsuspected abundance or even dominance of groups that were previously unknown or thought to be relatively rare [2,3]. These surprises have spurred continued interest in exploring prokaryotic diversity and its relationship to ecosystem processes.

In this paper, we assess the substantial amount of information accumulated on bacterial diversity in a variety of environments. We address several fundamental questions, focusing on aquatic systems but including other environments to provide a broader context.

1. Is aquatic bacterial diversity immeasurably high, or is it a tractable subject for study?
2. Have published reports adequately represented aquatic bacterial diversity?
3. How diverse are bacterial assemblages in aquatic environments?

We address these questions by a meta-analysis (sensu Glass [4]) of published studies of bacterial diversity. In a subsequent paper we will consider archael diversity, for which we have located 48 libraries from nine different environments.

2. Materials and methods

2.1. Terminology

Terminology varied considerably among the studies from which our data are drawn. All studies considered here employed PCR amplification of the 16S ribosomal RNA gene (16S rDNA) derived from environmental sources. Because most studies involved cloning of the amplified DNA, we will use 'clones' to denote the identified products of PCR amplification; 'library' to denote a collection of such identified PCR products derived from a given environmental source; 'library size' to denote the number of PCR products that were actually characterized; and 'phytotype' to denote a group of PCR products judged by the original authors to be essentially identical, regardless of the method or criteria they used to assess phylogenetic similarity. We make no assumptions regarding the relationship of 'phytotype' to traditional concepts of genera and species. Finally, we will refer to libraries as being samples of diversity in the environment. However, we recognize that while 16S rDNA libraries are ideally representative of the diversity present in the source environment, technically they are only samples of the PCR-amplified material from which clones are made.

2.2. Source data

Source data were obtained from published reports. Most of these studies (197 of 225 libraries in total employed sequencing to compare among PCR products and determine how many different PCR products were obtained. In many cases, sequencing was coupled to screening techniques (e.g. restriction fragment length polymorphism, denaturing gradient gel electrophoresis, amplified rDNA restriction analysis) to pre-sort PCR products prior to sequencing. A few studies employed techniques other than sequencing (10 amplified rDNA restriction analysis, 18 restriction fragment length polymorphism) to assess phylogenetic similarity. These alternative techniques are also dependent on differences in base sequence, although the sequence is not actually determined.

We applied a consistent set of standards when selecting data to be included in our analysis.

1. PCR amplification must have employed bacterial or universal primers. We excluded studies that used amplification primers targeting groups below the domain level.
2. We also excluded studies that involved characterization of cultured cells, which may not be representative of organisms in the source material. The sole exception to this rule was for bioreactors initiated with natural assemblages.
3. We excluded studies reporting very small libraries of fewer than five clones.
4. Some of the original studies obtained full-length 16S rDNA, while others obtained shorter fragments of several hundred bases. In all cases when shorter fragments were obtained, PCR primers were designed to include one or more variable regions of 16S rDNA. Borneman et al. [5] have shown that partial sequences containing variable regions produce much the same phylogenetic tree structures as full-length 16S rDNA sequences. Therefore, we did not consider PCR fragment length in our selection of data.
5. Some reports pooled several libraries derived from separate samples of a study site. In such cases, we used the pooled data. When both pooled data and data for individual libraries were presented, we used the data from individual libraries.
6. Each report included at minimum two essential pieces of information, either given directly or readily derived from the published data. These are the total number of clones actually characterized (=library size), and the number of different phytotypes obtained.
7. A subset of these studies provided more complete data regarding the frequency distribution of phytotypes. These were used to calculate the probable total number of phytotypes in the environment using one parametric and two non-parametric estimators.

A total of 225 bacterial libraries met criteria 1–6 [5–125]. Of these, sufficient detail was provided for 194 libraries to allow us to estimate total phytotype richness (criterion 7) by at least one method. A number of studies, including some reporting impressively large libraries, did not meet all of our requirements and were excluded from consideration. Although our primary interest is in Bacteria in aquatic systems (160 libraries in total; 138 with sufficient information provided to estimate total phytotype richness
by at least one method), we have included bacterial data from other environments to put the aquatic data in a larger context. ‘Aquatic’ refers here to both freshwater and marine environments, including free-living and particle-associated plankton, biofilms, hyperthernal springs and vents, sediment, gas hydrates, and groundwater. Non-aquatic environments included soils and paleosols, bioreactors, human and other animal digestive systems, and compost.

2.3. Estimating the number of unseen phylotypes

In 1953, Fisher et al. [126] described a procedure to estimate the number of unseen species present in a species assemblage from which samples are drawn. Since then, a variety of parametric and non-parametric methods for evaluating the number of unseen species have been proposed. Several reviews and comparisons of these methods are available (e.g.[127–130]) and provide a solid basis for understanding their limitations and strengths. We have employed several of the many alternative approaches, as described briefly below. Our intent is to illustrate that different approaches yield different results, but not to exhaustively compare and evaluate methods. Our selection includes methods that use different subsets of the original phylotype frequency data to calculate coverage and phylotype richness. Eqs. 2–6 are as provided in the EstimateS software package for species richness calculations (Version 5, R.K. Colwell, http://viceroy.eeb.uconn.edu/estimates).

We have rephrased the following descriptions to use the terms defined above (phylotype, library, library size) rather than alternative terms often used in the diversity literature (species, sample, sample size).

Because reports provided different levels of detail regarding library composition, some studies could not be used to calculate some estimators, i.e. n differs among the various estimators.

2.3.1. Observed number of different phylotypes

The most basic estimate of phylotype richness we considered was the observed number of different phylotypes in a library, which provides a minimum estimate of the total number of phylotypes present in the source assemblage. Most of the studies we considered used sequence similarity as the criterion for distinguishing among phylotypes, and most employed a similarity cutoff of 97% or greater (81 of 91 studies reporting a cutoff value used 97% or greater; seven others used 95%; three used 90%).

2.3.2. Accumulation curves and rarefaction analysis

Accumulation curves, also known as collector curves, are commonly used to assess whether diversity has been sampled exhaustively. They are constructed by plotting the cumulative number of phylotypes identified against the number of clones characterized (= level of effort). The order of addition of clones is randomized because it affects the shape of the curve. The number of different phylotypes identified increases with level of effort until diversity is exhaustively sampled and no new phylotypes can be recovered from the sampled environment. Strongly asymptotic plots provide evidence that few additional phylotypes will be recovered per increment of additional work beyond the existing library size. Weakly curvilinear plots indicate that phylotype richness in the source environment is underestimated [127]. Accumulation curves are sometimes extrapolated to estimate the total number of phylotypes that may be present; however, the estimated total number of phylotypes is highly dependent on the extrapolation method used [131]. Colwell and Coddington [127], and Gotelli and Colwell [132] discuss the requirements and limitations of such extrapolations.

Rarefaction analysis also generates a plot of phylotype richness against level of effort, by subsampling an observed frequency distribution in order to predict the total number and relative abundance of phylotypes that would be expected per level of effort. The result is a smoothed accumulation curve that may be accompanied by confidence intervals representing the statistical uncertainty in the estimated phylotype richness at any given level of effort. Rarefaction analysis is most often used to compare multiple sites at an equal level of effort (e.g.[8,19,74]) or to evaluate changes in community diversity (e.g. during incubations [90]). Rarefaction analysis is not appropriately used to extrapolate to total phylotype richness [133], but a strongly curvilinear rarefaction curve is evidence that diversity was exhaustively sampled.

We illustrate the relationship between level of effort and the corresponding estimates of total phylotype richness with several examples of rarefaction curves (calculated as in [134]) from different environments.

2.3.3. Coverage

Good’s coverage is a non-parametric estimator of the proportion of phylotypes in a library of infinite size that would be represented in a smaller library. As defined by Good [135], coverage C is calculated as:

\[ C = 1 - \frac{n_1}{N} \]  

where \( n_1 \) is the number of phylotypes appearing only once in a library, and \( N \) is the library size.

Chao et al. [136] and Lee and Chao ([137]; see also [128]) presented a non-parametric, abundance-based coverage estimator of phylotype richness in which sample coverage \( C_{ACE} \) is estimated as the proportion of individuals in relatively rare phylotypes (\( \leq 10 \) clones) that occur more than once in a library:

\[ C_{ACE} = 1 - \frac{F_1}{N_{rare}} \]

where \( F_1 \) is the number of phylotypes occurring only once in the library, and \( N_{rare} \) is the total number of individuals
in phylotypes occurring 10 or fewer times. Keating and Quinn (personal communication), in an unpublished comparison of various coverage estimators, found that this estimator performed best when sampling moderately even communities. As shown later, we found that phylotypes tended to be evenly represented in these libraries, because most phylotypes are relatively rare.

2.3.4. \textit{S}_\text{Chao1} estimator

Chao [138, 139] described a non-parametric estimator based on mark-release-recapture techniques that yields an unbiased estimate of the probable total number of phylotypes present in the source assemblage. Hill et al. [140] and Hughes et al. [130] summarized various parametric and non-parametric approaches to estimating total phylotype richness from the frequency of phylotypes in a clone library. They recommended \textit{S}_\text{Chao1} as a particularly promising richness estimator.

The \textit{S}_\text{Chao1} estimator is derived from the number of phylotypes appearing either one or two times in a given library, and therefore is particularly appropriate for data sets in which most phylotypes are relatively rare in the library. It is calculated as:

\[ S_{\text{Chao1}} = S_{\text{obs}} + \frac{F_1^2}{2(F_2 + 1)} - \frac{F_1F_2}{2(F_2 + 1)^2} \tag{3} \]

where \( S_{\text{obs}} \) is the number of phylotypes observed in the library, and \( F_1 \) and \( F_2 \) are the number of phylotypes occurring either one or two times. \( S_{\text{Chao1}} \) is correlated with sample size when sample size is small [127], and therefore tends to underestimate phylotype richness for small sample sizes.

Chao [139] also described the calculation of variance for \( S_{\text{Chao1}} \):

\[ \text{var}(S_{\text{Chao1}}) = F_2 \left[ \frac{G^4}{4} + G^3 + \frac{G^2}{2} \right] \text{ where } G = \frac{F_1}{F_2} \tag{4} \]

From this, the confidence limits about estimates of \( S_{\text{Chao1}} \) may be calculated.

2.3.5. \textit{S}_\text{ACE} estimator

\( S_{\text{ACE}} \) also provides an estimate of the total number of phylotypes in a source environment [136]:

\[ S_{\text{ACE}} = S_{\text{abund}} + \frac{S_{\text{rare}}}{C_{\text{ACE}}} + \frac{F_1}{C_{\text{ACE}}} \gamma_{\text{ACE}} \tag{5} \]

\( F_1 \) is the number of phylotypes occurring only once in the library, \( S_{\text{rare}} \) is the number of phylotypes occurring 10 or fewer times, and \( S_{\text{abund}} \) is the number occurring more than 10 times. \( \gamma_{\text{ACE}} \) is the coefficient of variation of the \( F_i \)s and is given by:

\[ \gamma_{\text{ACE}} = \max \left[ \frac{\sum_{i=1}^{10} i(1-i)F_1}{C_{\text{ACE}}(N_{\text{rare}})(N_{\text{rare}}-1)} , 1, 0 \right] \tag{6} \]

The \( S_{\text{ACE}} \) estimator is derived from a larger subset of the complete frequency distribution than the \( S_{\text{Chao1}} \) estimator, and therefore is particularly appropriate for data sets in which some phylotypes occur more frequently.

2.3.6. Boneh et al. estimator

Boneh et al. [140] described a parametric procedure for estimating the total number of phylotypes in a sampled environment, derived from the original Fisher et al. [126] method in which all frequency distribution data are used. We selected this procedure for its use of all data, and in contrast to the non-parametric methods described above. The Boneh et al. procedure is derived from a continuous time model of independent Poisson processes in which:

\[ \psi(t) = \sum_{k=1}^{\max t} N_k e^{-k - \sum_{k=1}^{\max t} N_k e^{-(1+t)}} \tag{7} \]

where \( \psi(t) \) is the number of new phylotypes expected from time 0 to \( t \), when the original observations were from time \(-1\) to 0, \( k \) is the number of times a phylotype is seen, and \( N_k \) is the number of phylotypes seen \( k \) times [140]. Re-phrased in terms of library size, \( t \) is equal to multiples of the size of the original library. If the original library is of size 10, setting \( t = 4 \) would provide an estimate of the expected number of new phylotypes for a library of size \( 1(10) + 4(10) = 50 \). If \( t \) is large, the Boneh et al. equation provides an estimate of the total number of new phylotypes that could be expected.

We will designate the Boneh et al. estimator \( S_B \) in all following discussion.

3. Results

3.1. Library size

The average size of 220 bacterial libraries was 81 ± 10 (mean ± 95% confidence limits), ranging in size from 5 to 417 clones. Five libraries, all from non-aquatic systems, were much larger (968–4270 clones from pig digestive systems [60]; 1386 clones from human periodontitis [78]). The
average size of 160 libraries from aquatic systems was 73 ± 10 (mean ± 95% confidence limits; range 5–353 clones).

3.2. Library composition

The number of different phylotypes observed was highly variable but strongly correlated with library size, over a wide range of library sizes (Fig. 1; \( r^2 = 0.62, n = 225, P < 0.001 \)). In aquatic environments, the average return per effort invested was approximately 70 different phylotypes per 100 clones analyzed (i.e. regression slope = 0.70). Rare phylotypes that appeared only once or twice contributed a large proportion (79%) of the phylotypes observed in most libraries (Fig. 2; linear regression slope = 0.79, \( r^2 = 0.88, n = 201 \)). Libraries from a variety of human and other animal digestive systems were a notable exception: many of these libraries had a much lower proportion of rare phylotypes than would be expected by comparison to libraries from other environments (see Fig. 2, Digestive Systems). The similarity cutoff for all of these exceptional studies was 98–99%.

3.3. Coverage

Two indices of coverage were used to estimate the proportion of phylotypes in the source assemblage that were represented in sample libraries: Good’s \( C \) [135] and Chao et al.’s \( C_{ACE} \) [136]. Coverage ranged widely but was not apparently related to library size for either index (Fig. 3A,B). The mean values of both indices were similar in aquatic versus other environments (Table 1). However, \( C_{ACE} \) estimates of coverage were often substantially lower than the corresponding estimate based on Good’s \( C \) (Fig. 3A,B).
and on average, estimates based on $C_{ACE}$ were significantly lower than estimates based on $C$ (Table 1).

### 3.4. Phylotype richness

Phylotype richness was calculated using three different approaches that estimate the number of phylotypes in the source environment from which samples were drawn. With two exceptions, the highest phylotype richness calculated for a library from an aquatic system was 795 phylotypes using $S_{ACE}$, 968 using $S_{Chao1}$, and 168 using $S_B$. The two exceptions were 4950 phylotypes ($S_{ACE}$, sulfide-rich mud [109]), and 2716 phylotypes ($S_{ACE}$, marine snow [89]). For 91% of aquatic-system libraries, estimates of phylotype richness were $\leq 200$ phylotypes using $S_{ACE}$, $\leq 115$ phylotypes using $S_{Chao1}$, and $\leq 59$ phylotypes using $S_B$ (Fig. 4).

### 3.5. Methods performance and comparisons

Among the three methods used to estimate phylotype richness, the non-parametric $S_{Chao1}$ and $S_{ACE}$ methods agreed closely. They differed primarily in that $S_{Chao1}$ yielded lower estimates for several libraries from hyper-thermal environments (Fig. 5A). The parametric method proposed by Boneh et al. [140] resulted in systematically lower estimates of phylotype richness than either non-parametric method, especially at higher values of $S_{Chao1}$ and $S_{ACE}$ (Fig. 5B,C). Indeed, nearly all values of $S_B$ were less than the corresponding values of $S_{Chao1}$ and $S_{ACE}$.

Confidence intervals were calculated for $S_{Chao1}$ from Eq. 4. Confidence limits decreased with increasing library size (Fig. 6; $r^2 = 0.56$, $n = 160$, $P < 0.001$). Rarefaction curves representative of a variety of aquatic and non-aquatic environments were generated and compared to phylotype richness estimates derived from $S_{Chao1}$, $S_{ACE}$ and $S_B$ (Fig. 7). Rarefaction curves ranged from nearly linear for temperate grassland soil [39] and French Guiana sediment [66], to strongly curvilinear for the Caribbean Basin [67].

### 4. Discussion

#### 4.1. Biases in the source data

All of the studies considered are based on PCR amplification of portions of 16S rDNA, and most are based on sequencing PCR products. Diversity estimates based on the types and relative proportions of sequences found in 16S rDNA libraries are subject to biases and methodological artifacts that may increase or decrease the frequency
with which a particular phylotype is found in a library [84,141–143]. Analysis of the effects of methodological problems is well beyond the scope of this paper. While such artifacts introduce noise into our analyses, they do not preclude worthwhile analysis of the wealth of information represented in 16S rDNA libraries. We note that these libraries have withstood the test of peer review and have been accepted as valid contributions to the exploration of bacterial diversity. They are the primary source of information we have available to assess the phylogenetic

Fig. 5. Comparisons among phylotype richness estimators. A: $S_{\text{Chao1}}$ and $S_{\text{ACE}}$ were highly correlated, and differed primarily in that $S_{\text{ACE}}$ estimates were much higher than $S_{\text{Chao1}}$ in hyperthermal environments. $S_B$ estimates were frequently lower than the corresponding estimates for either (B) $S_{\text{ACE}}$ or (C) $S_{\text{Chao1}}$.

Fig. 6. Confidence intervals for $S_{\text{Chao1}}$ expressed as a percent of the corresponding mean value of $S_{\text{Chao1}}$, versus library size. Confidence intervals decrease very rapidly (note the log-log plot) with increasing library size. At the mean size of libraries in this analysis, estimated confidence intervals about $S_{\text{Chao1}}$ range from 1 to 200%.
richness and complexity of bacterial communities. Analysis of this information provides insight into the likely complexity of bacterial communities in nature, and guidance and inspiration for future studies.

Aside from such artifacts, the percent similarity used as the cutoff for discriminating among phylotypes varies over a narrow range. However, the percent similarity cutoff used was not significantly correlated with the number of phylotypes observed, or with coverage estimates obtained \((n = 135)\). We did not consider this potential issue further.

### 4.2. Diversity measures vary

Some caution is needed when using diversity estimators to assess phylotype richness or the degree to which one has sampled it, or to compare among different environments. The conclusions reached may depend on which diversity measure is used. For example, the \(S_0\) estimator yielded lower estimates of phylotype richness than either \(S_{\text{Chao1}}\) or \(S_{\text{ACE}}\) (Fig. 5B,C), in most libraries. The \(S_{\text{ACE}}\) estimator yielded higher estimates in many libraries from hyperthermal environments than either \(S_{\text{Chao1}}\) or \(S_0\) (Fig. 5A,B), but otherwise was very strongly correlated with \(S_{\text{Chao1}}\).

These similarities and differences are to be expected. The parametric \(S_0\) estimator appears poorly suited for libraries in which most phylotypes appear only once or twice. The two non-parametric estimators differ primarily in that the \(S_{\text{ACE}}\) estimator takes into account a larger subset of the frequency distribution data (phytolotypes appearing 10 or fewer times in a library) than the \(S_{\text{Chao1}}\) estimator (phytolotypes appearing once or twice). However, rare phylotypes that appeared only once or twice dominated most of these libraries. Thus, the \(S_{\text{ACE}}\) and \(S_{\text{Chao1}}\) estimators usually were functionally equivalent, and consequently highly correlated. They differed for hyperthermal environments because many phylotypes appeared more often than once or twice in these libraries, i.e. in these libraries, the \(S_{\text{ACE}}\) estimator included some 'less rare' phylotypes that were ignored by the \(S_{\text{Chao1}}\) estimator.

We illustrate the differences among phylotype richness estimators in several examples. Libraries dominated by a few abundant phylotypes yielded strongly curvilinear rarefaction plots, and estimates of phylotype richness by any method were generally low (e.g. Fig. 7, Cariaco Basin). Conversely, libraries primarily composed of unique phylotypes yielded nearly linear rarefaction plots, and estimates of phylotype richness by any method were generally high (e.g. Fig. 7, French Guiana sediment). In intermediate cases (e.g. Fig. 7, hyperthermal spring), phylotype richness estimates by different procedures varied substantially for the same library, reflecting differences in the response of richness estimators to the frequency distribution of phylotypes.

Like richness estimates, coverage estimates often agreed but sometimes differed. The classic coverage index proposed by Good \([135]\) frequently yields higher estimates than Chao et al.'s \([136]\) \(C_{\text{ACE}}\) (Fig. 3C). Coverage estimates are frequently used as proof that sufficient work has been done to capture most of the diversity in a sampled environment. We urge caution if only one coverage estimator is used: alternative methods of estimating coverage may be less optimistic.

### 4.3. Reliability of richness estimates

Larger libraries tend to yield more statistically robust estimates of phylotype richness than smaller libraries. Fig. 6 shows the strong relationship of confidence limits for \(S_{\text{Chao1}}\) to library size. At the mean size of aquatic-system libraries (73 clones), the observed range in confidence limits was quite large (±1–200%). It would be difficult to predict whether a new library of average size would provide a reliable estimate of total phylotype richness. However, confidence limits were invariably less than ±50% for all libraries greater than ~300 clones in size. We conclude that reasonably robust (i.e. ±50%) estimates of phylotype richness may be obtained from moderately large libraries, and in many cases reliable estimates can be generated from average-sized libraries.

### 4.4. When is a library large enough?

As library size increases, both accumulation curves and phylotype richness estimators converge on a stable estimate of the true phylotype richness \([127]\). Bowman and McCuaig \([144]\) commented that "most published data sets are too small to provide accurate values [of species richness]; i.e. most obtain only a nonasymptotic species richness curve." Gotelli and Colwell \([132]\) commented that species richness, measured by the number of species actually observed, can only be compared when accumulation curves have reached an asymptotic maximum, meaning that all species have been recovered at least once. The
critical point is that while too small data sets can be used to estimate phylotype richness, the estimates will be biased to an unknown extent. For accumulation curves, the bias is always negative. For phylotype richness estimators, the bias may be either positive or negative, depending on the estimator used [127]. Both $S_{\text{Chao1}}$ and $S_{\text{ACE}}$ yield negatively biased estimates at small sample sizes [130].

Generalizing Gotelli and Colwell’s [132] comments, no valid statistical comparisons of the estimated total phylotype richness in different environments can be made, unless it is known that the estimates are unbiased. For example, the estimated phylotype richness of soils appears higher than that of aquatic sediments (Table 1). However, although soils would appear to be more diverse, it may only be that they have been better sampled than sediments; the estimates for aquatic sediments may greatly underestimate the true phylotype richness.

Few authors of the papers we considered commented on whether they regarded their libraries as adequately large, and most of those did so only by noting either that coverage was high, or that a curvilinear accumulation or rarefaction curve was obtained. Curvilinearity is an inconsistently interpreted, highly subjective measure of adequacy.

Phylotype richness estimators such as $S_{\text{ACE}}$ and $S_{\text{Chao1}}$ provide an efficient and less subjective alternative. The utility of phylotype richness estimators is that they approach this stable maximum before all phylotypes have been recovered, i.e. well before the corresponding accumulation curve would have reached an asymptotic maximum. Colwell and Coddington [127] found that the $S_{\text{Chao1}}$ estimator reached an asymptotic maximum most rapidly of several estimators, and yielded the least (negatively) biased estimate of richness at small sample sizes.

We propose using the following procedure to provide a more objective means of determining when a library is large enough to yield stable, unbiased estimates of phylotype richness. Investigators should explore libraries in progressive stages. At each stage a new subset of clones is characterized. The investigator calculates phylotype richness estimates for the subset, and determines whether estimates of phylotype richness have stabilized. If they have not, another subset of the library is analyzed, until stable estimates are obtained. The investigator then can evaluate the proportion of diversity in the source community that is represented in the library, and determine whether this proportion is sufficient for the purpose at hand.

We illustrate this approach with examples taken from six published libraries representing a variety of environments (coral reefs, 42; chicken ceca, 47; estuarine water, 97; human guts, 101; arctic sediment, 90; hyperthermal spring, 98). We constructed pseudo-libraries of varying size by randomly subsampling clones from the published libraries, and calculating phylotype richness estimates for each subset of the data. This is conceptually identical to procedures used to generate rarefaction curves.

Fig. 8 shows the results for $S_{\text{ACE}}$ (A) and $S_{\text{Chao1}}$ (B). A large library size did not guarantee a stable estimate of phylotype richness, nor was a small library necessarily too small. For example, the value of $S_{\text{ACE}}$ did not reach an asymptotic maximum for the largest library examined (hyperthermal spring microbial mat), but did approach an asymptotic maximum for one of the smaller libraries (chicken ceca). In the chicken ceca data set, the rarefaction curve for the full library was barely curvilinear (not shown). This illustrates that richness estimators stabilize at smaller library sizes than accumulation curves, and can provide an estimate of phylotype diversity with less effort.

Based on this progressive sampling approach, either three ($S_{\text{ACE}}$) or four ($S_{\text{Chao1}}$) of the six libraries would be considered ‘large enough’ to provide an unbiased estimate of phylotype richness. One library from a hyperthermal spring differed between the two estimators. This library would not be considered large enough based on $S_{\text{ACE}}$, but would be considered large enough based on $S_{\text{Chao1}}$. Again, the difference between the two estimators is not surprising; remember that the two estimators are highly correlated except in hyperthermal environments (Fig. 5A).
We applied this approach to 194 libraries for which frequency distribution data were available. We judged 56 libraries to be large enough to yield stable and unbiased richness estimates (Table 2), meaning that a stable estimate was obtained with either $S_{ACE}$ or $S_{Chao1}$, or with both. These libraries ranged widely in size (23–347 clones), spanning nearly the entire range of the libraries we examined, excepting only a few large libraries from digestive systems (Fig. 9). They include libraries containing both many and few phylotypes. For all environments combined, the 56 libraries contained $37\pm 3\%$ or $63\pm 6\%$ of the estimated total number of phylotypes in the environment (Table 2). Comparable results were obtained for all libraries combined, including those judged ‘not large enough’ to yield stable richness estimates ($S_{ACE}: 35\pm 3\%; S_{Chao1}: 58\pm 4\%$). In short, phylotype richness was surprisingly well represented.

<table>
<thead>
<tr>
<th></th>
<th>$n$</th>
<th>$C$</th>
<th>$C_{ACE}$</th>
<th>$S_{ACE}$</th>
<th>$S_{Chao1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Aquatic systems</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bacterioplankton</td>
<td>9</td>
<td>76 ± 11</td>
<td>73 ± 11</td>
<td>66 ± 36</td>
<td>54 ± 31</td>
</tr>
<tr>
<td>Biofilms</td>
<td>5</td>
<td>70 ± 17</td>
<td>64 ± 16</td>
<td>55 ± 33</td>
<td>39 ± 35</td>
</tr>
<tr>
<td>Gas hydrates</td>
<td>3</td>
<td>85 ± 8</td>
<td>79 ± 10</td>
<td>33 ± 4</td>
<td>33 ± 35</td>
</tr>
<tr>
<td>Groundwater</td>
<td>4</td>
<td>76 ± 16</td>
<td>67 ± 21</td>
<td>114 ± 61</td>
<td>81 ± 52</td>
</tr>
<tr>
<td>Hyperthermal</td>
<td>8/9</td>
<td>94 ± 3</td>
<td>74 ± 22</td>
<td>96 ± 31*</td>
<td>20 ± 11</td>
</tr>
<tr>
<td>Sediment</td>
<td>11</td>
<td>73 ± 10</td>
<td>65 ± 9</td>
<td>227 ± 165</td>
<td>190 ± 169</td>
</tr>
<tr>
<td>Suspended particles</td>
<td>3</td>
<td>67 ± 15</td>
<td>58 ± 14</td>
<td>65 ± 64</td>
<td>52 ± 63</td>
</tr>
<tr>
<td><strong>Other systems</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Digestive systems</td>
<td>4</td>
<td>53 ± 23</td>
<td>51 ± 19</td>
<td>175 ± 77</td>
<td>127 ± 35</td>
</tr>
<tr>
<td>Compost</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Soils</td>
<td>4</td>
<td>65 ± 27</td>
<td>61 ± 23</td>
<td>90 ± 67</td>
<td>88 ± 102</td>
</tr>
<tr>
<td>Bioreactors</td>
<td>4</td>
<td>83 ± 12</td>
<td>78 ± 12</td>
<td>81 ± 38</td>
<td>49 ± 31</td>
</tr>
<tr>
<td>All aquatic</td>
<td>43/44</td>
<td>78 ± 5</td>
<td>69 ± 6</td>
<td>114 ± 47</td>
<td>80 ± 46</td>
</tr>
<tr>
<td>Observed/Predicted (%)</td>
<td></td>
<td>38 ± 6</td>
<td>65 ± 7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>All other systems</td>
<td>12</td>
<td>67 ± 13</td>
<td>63 ± 12</td>
<td>115 ± 41</td>
<td>88 ± 39</td>
</tr>
<tr>
<td>Observed/Predicted (%)</td>
<td></td>
<td>34 ± 5</td>
<td>53 ± 13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>All systems</td>
<td>55/56</td>
<td>76 ± 5</td>
<td>68 ± 5</td>
<td>114 ± 38</td>
<td>82 ± 37</td>
</tr>
<tr>
<td>Observed/Predicted (%)</td>
<td></td>
<td>37 ± 5</td>
<td>63 ± 6</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Richness estimates should be considered minimally biased. Other abbreviations as in text. Numbers are mean±95% confidence intervals, with $n$ in parentheses. Note that $n$ for particular estimators varies from the total $n$ available per environment because data for some libraries were not sufficient to calculate all estimators. *$n=8$ for $S_{ACE}$ in hyperthermal environments.

Hughes et al. [130] and Hill et al. [145] observed that in a limited number of tests with bacterial data sets, the $S_{Chao1}$ estimator tended to stabilize with increasing library size, whereas $S_{ACE}$ did not. We found that both the $S_{ACE}$ and $S_{Chao1}$ estimators stabilized, but often not for the same libraries. While $S_{ACE}$ stabilized for 30 libraries, $S_{Chao1}$ did so for 34 libraries, and both estimators stabilized in only eight libraries. The corresponding results for aquatic environments were 23 libraries for $S_{ACE}$, 27 libraries for $S_{Chao1}$, and six libraries for both estimators. This illustrates again that richness estimators are not all equivalent.

Our suggested approach needs further exploration, for example to examine which of many alternative richness estimators reaches an asymptotic maximum most rapidly, is closest to the true phylotype richness in a given environment, and performs most consistently among communities differing in absolute diversity and relative dominance. It is possible that multiple estimators must be employed (as we did here), if no one richness estimator will work equally well with all libraries. We argue that phylotype richness estimates must be regarded as potentially biased unless a comparable approach is used to establish that a stable estimate of phylotype richness has been obtained.

### 4.5. Differences among aquatic environments

Some of the fundamental questions often asked are ‘How many different Bacteria are there?’, ‘How diverse are bacterial communities?’ and ‘What controls bacterial diversity?’ Such questions are often addressed by comparing the number of phylotypes recovered in a library to...
observations for other sites and environments. As noted above, comparisons of the diversity of phylotypes actually recovered in libraries will be valid only if the libraries all appear to have exhaustively sampled the diversity present in their source communities [132,145]. However, valid comparisons of diversity between libraries can still be made, provided that unbiased estimates of total phylotype richness can be obtained. We outlined a simple procedure (Section 4.4) for ascertaining whether a library is large enough to obtain a stable and unbiased estimate of phylotype richness, and applied this procedure to all libraries for which sufficient data were available. The resulting subset of 56 ‘large enough’ libraries represents those in which we have the most confidence that phylotype richness estimates are valid. We consider that a comparison among environments is most likely to be valid if it is limited to this subset.

For aquatic environments, only three to nine libraries per environment were large enough for a valid comparison (43 in total for $S_{ACE}$, 44 in total for $S_{Chao1}$), limiting our ability to resolve real differences in diversity among these environments. Thus, we are not surprised to find only a few significant differences among aquatic environments. Such differences were observed using $S_{Chao1}$ but not using $S_{ACE}$ (log-transformed data; $S_{Chao1} F=2.54$, df = 6,37, $P < 0.05$; $S_{ACE} F=1.90$, df = 6,36, $P > 0.05$). Multiple comparison procedures (Newman–Keuls test [146]) indicated that the average phylotype richness was significantly lower in hyperthermal environments than in sediments and bacterioplankton, but not less than in groundwater, suspended particles, biofilms and gas hydrates. The often low bacterial diversity of hyperthermal environments is widely recognized, and has been emphasized by studies in which only one or a few bacterial taxa inhabited a particular hot spring [92]. We note that this apparently low diversity might sometimes result from sampling a restricted local environment (e.g. filamentous mats [91]; microbial mats [98]).

Phylotype richness also varied greatly within the environmental categories we used. For example, the estimated phylotype richness in sediments ranged from the most diverse of all sampled environments, to sediments in which only a few different phylotypes occur (Fig. 10). Conversely, the mean phylotype richness of suspended particles was very nearly the lowest of all aquatic environments, yet one study reported an exceptionally high bacterial diversity in marine snow [89].

These striking variations within environments may represent unexplored, important differences in environmental conditions among the locations we have grouped together. For example, we did not separate environments by redox state, but it may influence patterns of diversity. This is illustrated by Madrid et al.’s [67] study of anoxic basin waters, bacterial diversity was low in samples from 310 and 320 m near the redox boundary, but higher in samples from 500 and 1310 m, below the redox boundary. Much more work is needed to determine which environmental variables are most important for determining patterns of diversity.

4.6. A high proportion of rare phylotypes in libraries – what does it mean?

With few exceptions, bacterial libraries of all sizes are predominantly composed of rare phylotypes that appeared only once or twice (Fig. 2). Among the studies we considered, four of five exceptionally large libraries were also exceptional in that they contained proportionally fewer rare phylotypes than most other libraries (Fig. 2, the four highest observed phylotypes). It is likely that these few libraries were sufficiently large (1262–4270 clones) that many phylotypes were sampled repeatedly. Indeed, coverage estimates for these exceptionally large libraries were very high (94–98%, Good’s C). Most of the other libraries we considered probably represent small samples of the available phylotypes, in which the likelihood is that any given phylotype would only be sampled once.

If a library is an unbiased sample of the source material, then a preponderance of rare phylotypes also suggests that most phylotypes occur in roughly uniform abundance in the source material. Given the possibility of various PCR
artifacts, gene copy number and other biases, making the leap from the composition of libraries to the composition of bacterial communities in nature is certainly problematic. We regard the widespread occurrence of rare phylotypes in libraries as interesting evidence that in nature, bacterial communities might be predominantly composed of uniformly abundant phylotypes.

4.7. How diverse are Bacteria?

If PCR artifacts and other potential biases have influenced the composition of published libraries, it would appear to be in the direction of generating more nearly uniform representation of phylotypes (e.g. [142,143]), i.e. most phylotypes appeared only once or twice in most libraries. A bias toward uniformly rare phylotypes leads to an overestimate of the true phylotype richness, and an underestimate in the apparent coverage of libraries by any of the estimators we used. Despite this, our phylotype richness estimates were remarkably low, and coverage estimates were remarkably high in aquatic environments.

For aquatic environments, most (90%) libraries yielded estimates of \( \leq 200 \) phylotypes in their source environments. The largest number of phylotypes actually observed in any of the studies we considered was 375 in pig digestive systems [60]. The largest number actually observed in an aquatic environment was 255 in Antarctic sediment [144]. Coverage estimates for aquatic systems range widely, but suggest that on average libraries captured more than half of the estimated total number of bacterial phylotypes in their source environments (mean \( C_{\text{ACE}} = 0.57 \); mean \( C = 0.71 \)). The libraries considered large enough to generate stable estimates of phylotype richness, on average, contained 37–63% of the estimated total number of phylotypes in their source communities. These results lead us to suggest that libraries often capture a large fraction of the diversity present in the source community.

We remain open to the possibility that bacterial diversity may be very high in some systems. Some indirect and direct evidence supports this. For example, in one study all clones in a library were unique [15], making it impossible to estimate total phylotype richness, but suggesting a very high bacterial diversity. Two exceptional libraries yielded very high phylotype richness estimates of 2716 phylotypes [90] and 4950 phylotypes [109]. Bowman and McCuaig [144] recently used the \( S_{\text{Chao1}} \) estimator to evaluate phylotype richness at three depths within Antarctic sediment. Their moderately large libraries contained 138–255 phylotypes, and they estimated that the source communities contained 442–1128 phylotypes.

We conclude that in general (and undoubtedly with some exceptions), the level of effort expended in published studies of aquatic systems has not been sufficient to exhaustively sample bacterial diversity, but is often sufficient to capture a large fraction of it. In many cases, bacterial diversity estimated from these libraries is lower than we would have expected.

This is surprising, given hints that Bacteria are extremely diverse in nature. Curtis et al. [147] presented arguments that the probable diversity of Bacteria in all marine planktonic systems combined may approach \( 2 \times 10^6 \) different phylotypes in total, much higher than we derived from any of the case studies we considered. Hughes et al. [130] observed that “Ultimately, microbes – like tropical insects – are too diverse to count exhaustively.” Our analyses suggest that in fact, studies of bacterial diversity have often captured a great deal of the phylotypes present, and that the study of bacterial diversity is not an intractable problem. How do we reconcile our estimates with the widespread perception that aquatic bacterial diversity is extremely high?

Curtis et al. based their analyses on an assumption of log-normal distribution of species richness, an assumption supported by observations in soils [38]. However, it is not known whether this assumption is generally valid, and Ward [148] has pointed out that published species richness estimates range so broadly that they “provide no support for the assumption of a log-normal – or any other – distribution of abundance.” We based our analyses on non-parametric estimates of phylotype richness, using estimators reported to perform well for relatively small samples of relatively uniformly distributed populations. The difference in our underlying assumptions may contribute to differences in our conclusions.

The predictions of Curtis et al. are actually similar to our calculations, when considered on comparable spatial scales. Although we did not impose a specific standard for spatial scale in selecting data, the data we used for our calculations were all from samples collected on local scales, e.g. grams of sediment, or milliliters to liters of water. At local scales, Curtis et al. calculated that there may be about 9000 phylotypes in a single sample from an anaerobic digester, 70 in a sample of activated sludge, 163 in 1 ml of Sargasso Sea water, 6000–60000 in 1 g of soil, and fewer than 8000 phylotypes in an entire lake. Our estimates for aquatic systems overlap with this broad range. The study by Bowman and McCuaig [144] illustrates the importance of spatial scale to diversity estimates. They estimated that Antarctic sediment at a single site contained 442, 1128, and 541 phylotypes at 0 cm, 2 cm and 21 cm depth. The libraries from these depths contained a total of 496 phylotypes, but only 76 of them overlapped between depths. Pooling samples separated by such small distances greatly increased the perceived total diversity of the site.

Fuhrman and Campbell [149] noted that the diversity of Bacteria and Archaea in the ocean, in the absence of ecological or physiological correlates, makes Hutchinson’s [150] ‘paradox of the plankton’ question (‘How is it possible to maintain so many phytoplankton species in a homogeneous environment with only a few potentially limit-
ing resources?) even more difficult to answer. Any answers must also consider both the low diversity and the uniform abundance (i.e. equally rare) of phylotypes as represented in rDNA libraries. We suggest that on local scales, aquatic bacterial diversity is much less than any predictions of their global diversity, and remains a tractable subject for study. The global-scale diversity of aquatic Bacteria, on the other hand, may be beyond present capabilities for effective study.

Acknowledgements

We thank J. Bowman, T. Leser, and V. Madrid for supplying us with unpublished data. We are indebted to K. Keating and J. Quinn for sharing their extensive analysis of 17 non-parametric procedures for estimating species richness. R. Cerrato kindly supplied programs for generating rarefaction curves. Four anonymous reviewers provided very thoughtful and constructive comments. This project was supported in part by NSF OCE Grants 9818574 to R.C. Aller and J.Y.A., and 9907983 to P.F.K., J.Y.A. and H.S. Dhadwal, and by a fellowship from the USAI Fulbright Senior Scholar Program to J.Y.A. This is contribution No. 1211 from the Marine Science Research Center, Stony Brook University.

References


