

Estimating prokaryotic diversity: When are 16S rDNA libraries large enough?

Paul F. Kemp* and Josephine Y. Aller

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

Abstract

As a necessary step in the study of prokaryotic diversity using 16S rDNA libraries, authors should evaluate how well their libraries represent diversity in the source environment. Phylotype–richness estimates can be used to judge whether a library represents diversity sufficiently for its intended purpose. We have argued that richness estimates are most useful if libraries are first shown to be large enough to yield stable estimates. In this article, we (1) evaluate two potentially suitable, non-parametric richness estimators (S_{ACE} and S_{Chao1}), tested against model libraries and libraries drawn from natural prokaryotic communities; (2) evaluate whether stable richness estimates are also unbiased; and (3) examine characteristics of prokaryotic libraries that influence the usefulness of richness estimators. Richness estimates consistently reached a stable asymptotic value for libraries that sampled diversity exhaustively. Stable estimates appear to be unbiased or minimally biased estimates of phylotype richness. The S_{ACE} estimator was often undefined, sometimes overestimated phylotype richness at intermediate sampling efforts, and sometimes stabilized at a larger library size than the S_{Chao1} estimator. The S_{Chao1} estimator appears well suited for estimating phylotype richness from prokaryotic 16S rDNA libraries. Libraries judged too small to yield a stable richness estimate typically had a highly uneven frequency distribution of phylotypes, with a preponderance of phylotypes that occurred only once in the library. Libraries considered large enough typically had a more even frequency distribution of phylotypes. A software tool is provided to aid others in assessing whether their libraries are large enough to yield stable phylotype–richness estimates.

The most commonly used approach to assay bacterial diversity in aquatic systems involves 16S rDNA library construction by polymerase chain reaction (PCR) amplification, cloning, sequencing, and subsequent phylogenetic analysis. Although very few 16S rDNA libraries sample diversity exhaustively, most published studies do not include an evaluation of how well the 16S rDNA library represents bacterial diversity in the source environment (Kemp and Aller 2003). Some limited questions may be addressed without such an evaluation; for example, the mere presence of a particular phylotype in a library may be informative. However, many fundamental questions cannot be addressed without first demonstrating that the library has captured a sufficiently large proportion of the diversity that it is intended to represent. For example, the absence of a particular phylotype from a library has little meaning unless it is known

that the library is large relative to the diversity in the source environment. Comparisons among libraries are particularly problematic. Even as simple a comparison as the number of bacterial phylotypes present in libraries from two sampling locations is of questionable value unless the libraries represent diversity equally well.

Species–richness estimators can be used to estimate the total number of phylotypes present in the source environment and provide a context for evaluating whether a library has captured a large enough fraction of diversity for its intended purpose. Richness estimators underestimate phylotype richness in small libraries (e.g., Hughes et al 2001; Hill et al. 2003; Stach et al. 2003; and others), and consequently have limited usefulness until the library size is large enough to yield a stable estimate. We proposed an assessment procedure in which 16S rDNA libraries are subsampled and phylotype richness is calculated for different subsample sizes to determine whether the library is large enough to yield a stable richness estimate (Kemp and Aller 2003). Using the abundance-based richness estimators S_{Chao1} (Chao 1984, 1987) and S_{ACE} (Chao et al. 1993), we identified 56 of 194 bacterial libraries that yielded stable richness estimates and compared the estimated phylotype richness across environments for these libraries only. This pro-

*E-mail: paul.kemp@stonybrook.edu

Acknowledgments

This project was supported in part by NSF OCE grants 9818574 to R.C. Aller and J.Y. Aller and 9907983 to P.F. Kemp, J.Y. Aller, and H.S. Dhadwal, and by a fellowship from the USIA Fulbright Senior Scholar Program to J.Y. Aller. This is contribution No. 1271 from the Marine Science Research Center, Stony Brook University.

cedure gave us greater confidence that the richness estimates and the comparisons based on them were valid.

The assessment procedure we applied was rudimentary, and critically important questions remain unanswered. Stable phylotype–richness estimates are not necessarily unbiased estimates, and as Hughes et al. (2001) commented “To test for bias, one needs to know the true richness to compare against the sample estimates. As yet, this comparison is impossible for microbes, because no communities have been exhaustively sampled.” Furthermore, the relationship of richness estimators to library size has been examined in several recent studies and by subsampling 16S rDNA libraries, and results have been mixed.

Hughes et al. (2001) observed that in several large (128 to 284 clones) prokaryotic libraries, richness estimates based on S_{Chao1} first increased and then usually stabilized with increasing subsample size and were independent of sample size thereafter. However, S_{Chao1} estimates did not stabilize for bacteria in a high-productivity aquatic mesocosm, and S_{ACE} did not yield stable estimates for any library. Hill et al. (2003) observed that S_{Chao1} estimates stabilized with one but not with a second large bacterial library, whereas S_{ACE} estimates did not stabilize with either. We obtained stable estimates from either S_{Chao1} or S_{ACE} in 56 bacterial libraries, but only a few libraries yielded stable estimates from both estimators (Kemp and Aller 2003). Stach et al. (2003) compared phylotype richness of actinobacteria at three depth horizons in marine sediment. Values of S_{ACE} did not stabilize with subsample size in two of three libraries. The relationship of S_{ACE} to subsample size was inconsistent and varied across depths, and it was evident that further sampling could alter even the rank order of estimated phylotype richness at the three depths.

These few published tests have not resolved which estimator might be most suitable or whether any richness estimator performs consistently with prokaryotic libraries. The true phylotype richness was unknown in all of these cases. Consequently, the failure to obtain a stable richness estimate from some libraries could indicate either failure of the richness estimator to perform as intended or simply that those libraries were too small to represent diversity adequately. Furthermore, bias could not be evaluated.

Foggo et al. (2003) addressed the question of bias by applying stringent criteria to select data sets that yield highly stable estimates of species richness, in three marine and estuarine faunal assemblages. Using the asymptotic richness estimate as an approximation of the true species richness, they subsampled the complete data sets and evaluated the relationship of bias to subsample size. At small subsample sizes, S_{Chao1} and five other estimators all underestimated species richness. At intermediate subsample sizes, S_{Chao1} slightly overestimated species richness for two relatively even assemblages of marine macrofauna, but slightly underestimated the species richness of a depauperate and highly patchy assemblage of estuarine oligochaetes. They judged S_{Chao1} the best compromise estimator, particularly at intermediate sampling efforts.

In summary, these studies are encouraging in that they demonstrate that richness estimators often yield stable estimates with prokaryotic libraries, and S_{Chao1} estimates appear to be minimally biased in tests with faunal assemblages. However, the significance and interpretation of unstable estimates is unclear, the results for different estimators disagree, and the direction and extent of bias have not been tested with prokaryotic libraries.

Before richness estimators can be used routinely to assess how well 16S rDNA libraries represent diversity, these uncertainties must be addressed. In this paper, we (1) evaluate two potentially suitable phylotype richness estimators tested with a variety of model libraries and libraries drawn from natural prokaryotic communities; (2) evaluate whether stable estimates are also unbiased; and (3) examine characteristics of aquatic prokaryotic libraries that influence the usefulness of richness estimators. Although our primary purpose is to evaluate the use of richness estimators to assess libraries from aquatic communities, we include examples from nonaquatic environments to illustrate that our results can be applied more generally. Finally, we provide a software tool to enable others to apply the same analysis to their own library and assess whether it is large enough to yield stable phylotype–richness estimates.

Materials and procedures

Definition—The definition of “phylotype” varies depending on the rules used to determine whether two PCR products are effectively identical or not. We use “phylotype” to mean 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. No specific relationship to traditional taxonomic nomenclature is implied.

Phylotype richness estimators—16S rDNA libraries are almost always unreplicated samples of diversity and require the use of abundance-based rather than incidence-based richness estimators. We examined the performance and suitability of two abundance-based estimators. S_{Chao1} (Chao 1984, 1987) is a non-parametric estimator based on mark-release-recapture techniques. It is calculated as

$$S_{\text{Chao1}} = S_{\text{obs}} + \frac{F_1^2}{2(F_2 + 1)} - \frac{F_1 F_2}{2(F_2 + 1)^2} \quad (1)$$

where S_{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. It is particularly appropriate for data sets in which most phylotypes are relatively rare (Chao 1987).

S_{ACE} (Chao et al. 1993) is a coverage-based estimator defined as

$$S_{\text{ACE}} = S_{\text{abund}} + \frac{S_{\text{rare}}}{C_{\text{ACE}}} + \frac{F_1}{C_{\text{ACE}}} \gamma_{\text{ACE}}^2 \quad (2)$$

where F_1 is the number of phylotypes occurring only once in the library, S_{rare} is the number of phylotypes occurring 10 or fewer times, and S_{abund} is the number occurring more than 10 times. γ_{ACE}^2 is the coefficient of variation of the F_i 's. C_{ACE} is a sample coverage estimate defined as the proportion of indi-

viduals in relatively rare phylotypes (<10 clones) that occur more than once in a library (Chao et al. 1993). The S_{ACE} estimator is particularly appropriate for data sets in which some phylotypes occur more frequently (Chao et al. 1993). S_{Chao1} and S_{ACE} estimators are highly correlated when most phylotypes are present only once or twice (Kemp and Aller 2003), as is often true of prokaryotic libraries.

Source data—Source data were constructed from theoretical distributions and obtained from published reports. Constructed data sets consisted of model libraries representing different underlying phylotype abundance distributions. In each of the model libraries, the true phylotype richness is known and the accuracy of phylotype richness estimators can be evaluated. We constructed a set of four model libraries based on different distribution models, including the geometric series, lognormal, broken stick, and uniform models.

The geometric series, lognormal, and broken stick models have a long history in classical ecology. The geometric series model (May 1975) stems from the assumption that each species, in rank order, preempts resources that are no longer available to any other species. It describes species-poor communities in which few phylotypes dominate and all others are rare. The lognormal distribution (Preston 1962) appears to fit many natural communities, especially in complex and species-rich environments (May 1975). It is particularly interesting because of a recent publication by Curtis et al. (2002) in which it provided the basis for a calculation of the global diversity of prokaryotes. The broken stick model (MacArthur 1957) arises from the assumption that resources in a community are allocated randomly among species. It is representative of the most even community compositions likely to occur in nature (May 1975). We included one additional model library in which all phylotypes are equal in abundance, as the extreme case of an even distribution.

We also used examples of real-world prokaryotic communities for which phylotype richness is known. Libraries in which all or nearly all phylotypes appear more than once are likely to have captured most of the phylotypes present in the source environment, and can be used as a proxy for ones in which diversity has truly been sampled exhaustively. Very few such libraries are available. We evaluate two aquatic archaeal libraries generated from Antarctic hypersaline lake sediment (Bowman et al. 2000a) and from a marine hydrothermal vent (Takai and Sako 1999). Two aquatic bacterial libraries were generated from Antarctic marine-salinity lake sediment (Bowman et al. 2000b) and from a lithotrophic biofilm (Bond et al. 2000). A third, very unusual aquatic bacterial library was obtained from a geothermal well where only two phylotypes were present, one much more abundant than the other (Marteinsson et al. 2001). In addition to the libraries from aquatic systems, we included two bacterial libraries from non-aquatic environments, one from chicken cecum (Gong et al. 2002) and one from contaminated soils (Nogales et al. 2001). For each of these libraries, the number of phylotypes actually

observed in the library should be a minimally biased estimate of the number present in the source community.

Finally, we examined seven libraries that did not exhaustively sample diversity in their source communities, as indicated by their including a number of phylotypes that occurred only once. These include both smaller and larger libraries, taken from both diverse and relatively simple source communities; they also differed in the degree to which one or a few phylotypes dominated in their abundance in the library.

Distribution parameters for model libraries were chosen to generate realistic phylotype abundance distributions similar to those found in the selected examples of natural communities. For the lognormal, broken stick, and uniform models, 1200 clones were allocated to 50 phylotypes, with the abundance of each phylotype rounded to the nearest whole number of clones. In the highly uneven geometric series model, the relative abundances of phylotypes differ by orders of magnitude, and only a few phylotypes can be represented with any manageable total number of clones. We employed a geometric series model distribution with 1200 clones allocated to 15 phylotypes.

Procedures—The assessment procedure is similar to rarefaction analysis (Heck et al. 1975). For each library, 1000 data subsets were derived by random sampling with replacement. These derived data subsets range in size n_i from $i = 1\%$ to 100% of the total size N of the library, in increments of 1% of N rounded to the nearest integer value. For each subsample size n_i , 10 replicate data subsets were drawn with replacement from the model library; at size $n_i = N$, all 10 data subsets were identical to the complete library. The values of S_{Chao1} and S_{ACE} were calculated for each derived data subset and plotted against subsample size n_i to determine whether an asymptote was reached. If the estimated phylotype richness reached an asymptote, we inferred that the library was large enough to yield a stable estimate of phylotype richness. We were satisfied with visual inspection of the plots, but any reasonable rule could be used to define what constitutes an asymptote.

Constructing 1000 derived data subsets from each of 18 libraries would be a laborious task. We prepared a simple program that employed a web form (HTML) interface and a Perl form processor (~150 lines of code). The data input to the web form consisted of the number of phylotypes that appeared in the library at frequencies ranging from 1 to 150 times (the upper limit of 150 is sufficient to include nearly all libraries published to date). The output consisted of 1000 lines of comma-separated data, each representing a derived data subset, suitable for import into a prepared spreadsheet. Formulas built into the spreadsheet calculated values of S_{Chao1} and S_{ACE} for each derived data subset and plotted them against subsample size.

Assessment

The S_{Chao1} and S_{ACE} richness estimators quickly reached an asymptotic maximum for all four model libraries (Fig. 1). In seven libraries from natural communities, each was considered to have sampled phylotype richness thoroughly, both estima-

Table 1. Comparison of the estimated number of phylotypes in the source community (predicted S_{Chao1} and S_{ACE} , averages of last five estimates) to the actual number (S)

Library	Size	S	S_{Chao1}	S_{Chao1}/S	S_{ACE}	S_{ACE}/S	%s*
Model libraries							
Geometric series	1200	15	14.8	0.99	15.8	1.05	—
Lognormal	1200	50	50.0	1.00	50.0	1.00	—
Broken stick	1200	50	51.7	1.03	51.1	1.02	—
Uniformly abundant	1200	50	50.0	1.00	50.0	1.00	—
Natural community libraries							
Archaea							
Hypersaline lake sediment†	76	8	8.0	1.00	8.0	1.00	98
Marine hydrothermal vent‡	83	12	12.4	1.03	12.3	1.03	98
Bacteria							
Saline lake sediment§	56	14	14.4	1.03	14.9	1.06	98
Lithotrophic biofilm	93	6	6.3	1.05	6.5	1.08	98
Geothermal well¶	73	2	2.0	1.00	2.0	1.00	98
Chicken cecum#	116	15	16.0	1.07	15.9	1.06	95
PCB-contaminated soil**	182	40	40.2	1.01	40.5	1.01	97
Mean estimated/actual phylotypes	1.02	1.03					

*%s is the percent similarity used by the original authors to define unique sequences

†Bowman et al. 2000a

‡Takai and Sako 1999

§Bowman et al. 2000b

||Bond et al. 2000

¶Marteinsson et al. 2001

#Gong et al. 2002

**Nogales et al. 2001

tors again reached a stable asymptote in all cases (Figs. 2 and 3). These range from a library in which only two phylotypes were present, one much more abundant than the other (Fig. 3A-B), to a large library containing many phylotypes (Fig. 3E-3F). In all but one case, richness estimates were identical to or slightly greater than the actual number of phylotypes present (Table 1). The sole exception was for a geometric series model library, where S_{Chao1} slightly underestimated phylotype richness.

Of the seven libraries that did not sample diversity exhaustively (Figs. 4 and 5), three were large enough that the value of S_{Chao1} reached a stable asymptote (Fig. 4). From the results with model libraries and libraries representing exhaustively sampled natural communities, we infer that these asymptotic values are unbiased or minimally biased estimates of phylotype richness. After concluding that all three estimates are unbiased, it is possible to make rigorous statistical comparisons of estimated phylotype richness among these three libraries. A solution for the variance of S_{Chao1} is available (Chao 1987) and provides the capacity for statistical comparisons. In these examples, we would judge the phylotype richness of bacteria in the geothermal spring example (Fig. 4E-4F, 65.3 phylotypes) to be significantly greater than that in coastal bacterioplankton (Fig. 4A-4B, 43.7 phylotypes) or Gulf of Papua sediment (Fig. 4C-4D, 44.0 phylotypes).

We see substantial disagreement between the S_{Chao1} and S_{ACE} estimators in two of these libraries. In the coastal bacterio-

plankton library (Fig. 4A-4B), values of S_{ACE} increased beyond the asymptotic maximum of S_{Chao1} , then decreased. At full library size, the two estimators were approximately equal. In the geothermal spring library (Fig. 4E-4F), the final values of S_{ACE} were somewhat higher than for S_{Chao1} . S_{ACE} was often undefined for subsample data sets in which no phylotypes occurred from one to 10 times (see Figs. 1 to 4).

Fig. 5 shows four examples of libraries from aquatic systems (if we consider a bioreactor to be an aquatic system) in which stable estimates of phylotype richness were not obtained for either S_{Chao1} or S_{ACE} . In two cases, values of S_{ACE} were generally higher than values of S_{Chao1} (Fig. 5A-5D). We infer that (1) these libraries were not large enough to obtain stable richness estimates, (2) richness estimates derived from these libraries are biased to an unknown extent, (3) we cannot determine how well these libraries represent their respective source environments, and therefore (4) comparisons among these examples would not be valid. For example, roughly the same number of phylotypes was recovered in the Antarctic sediment library (Fig. 5E-5F) as in the Arctic sediment library (Fig. 5G-5H). However, the apparent similarity in phylotype richness may be an artifact of sample size; more work might reveal a difference in the total number of phylotypes in each environment.

This raises an obvious question: how much larger would these libraries have to be to obtain unbiased estimates of phylotype richness? Richness estimators can indicate whether a

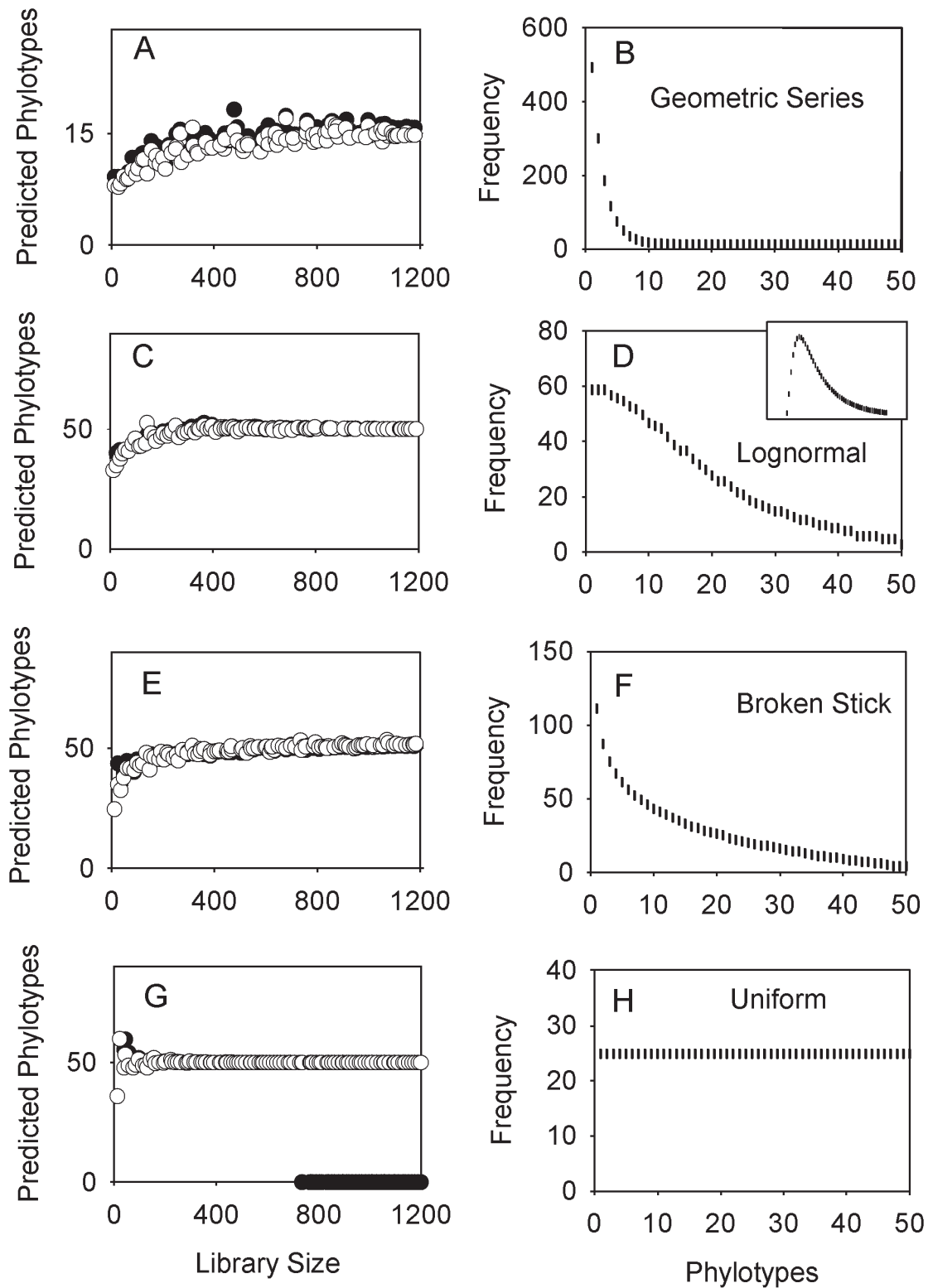


Fig. 1. Left panels: Predicted number of phylotypes based on S_{ACE} (filled symbols) and S_{Chao1} (open symbols) versus size of subsamples of four model libraries. Each point is the mean of 10 replicate subsamples of the library. Right panels: Each library represents a different underlying phylotype frequency distribution. Phylotype–richness estimates based on either estimator reach an asymptotic maximum at or near the correct number of phylotypes. Undefined values of S_{ACE} are plotted as zero phylotypes. Models used are A-B, geometric series; C-D, lognormal; E-F, broken stick; and G-H, uniform.

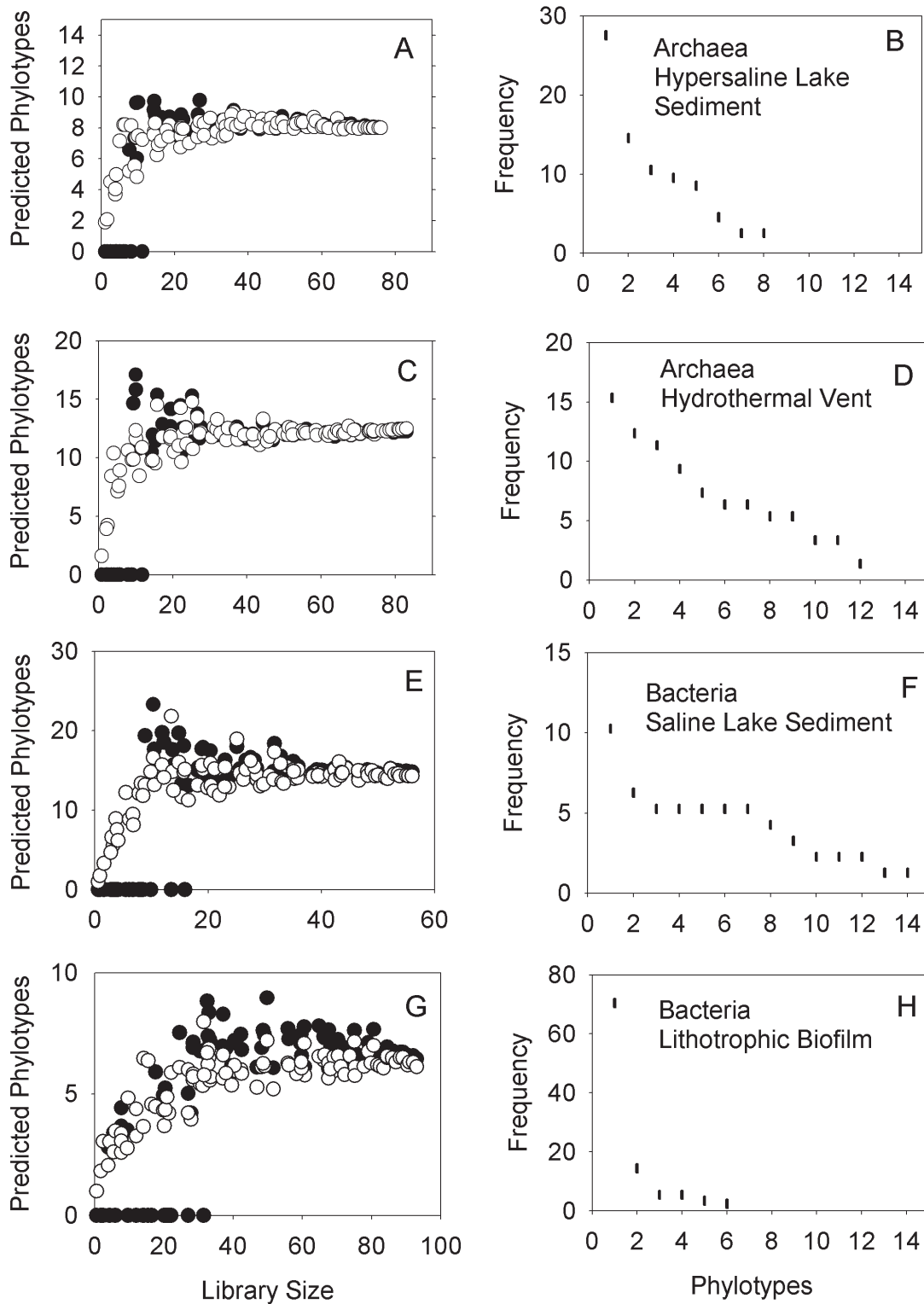


Fig. 2. Left panels: Predicted number of phylotypes based on S_{ACE} (filled symbols) and S_{Chao1} (open symbols) versus size of subsamples of four libraries derived from natural aquatic prokaryotic communities. Each point is the mean of 10 replicate subsamples of the library. All four were considered to have exhaustively sampled diversity in their source communities. Right panels: The corresponding phylotype frequency distributions for each library. Phylotype–richness estimates based on either estimator reach an asymptotic maximum at or near the number of phylotypes actually observed in these libraries. Undefined values of S_{ACE} are plotted as zero phylotypes. A-B: Bowman et al. 2000a. C-D: Takai and Sako 1999. E-F: Bowman et al. 2000b. G-H: Bond et al. 2000.

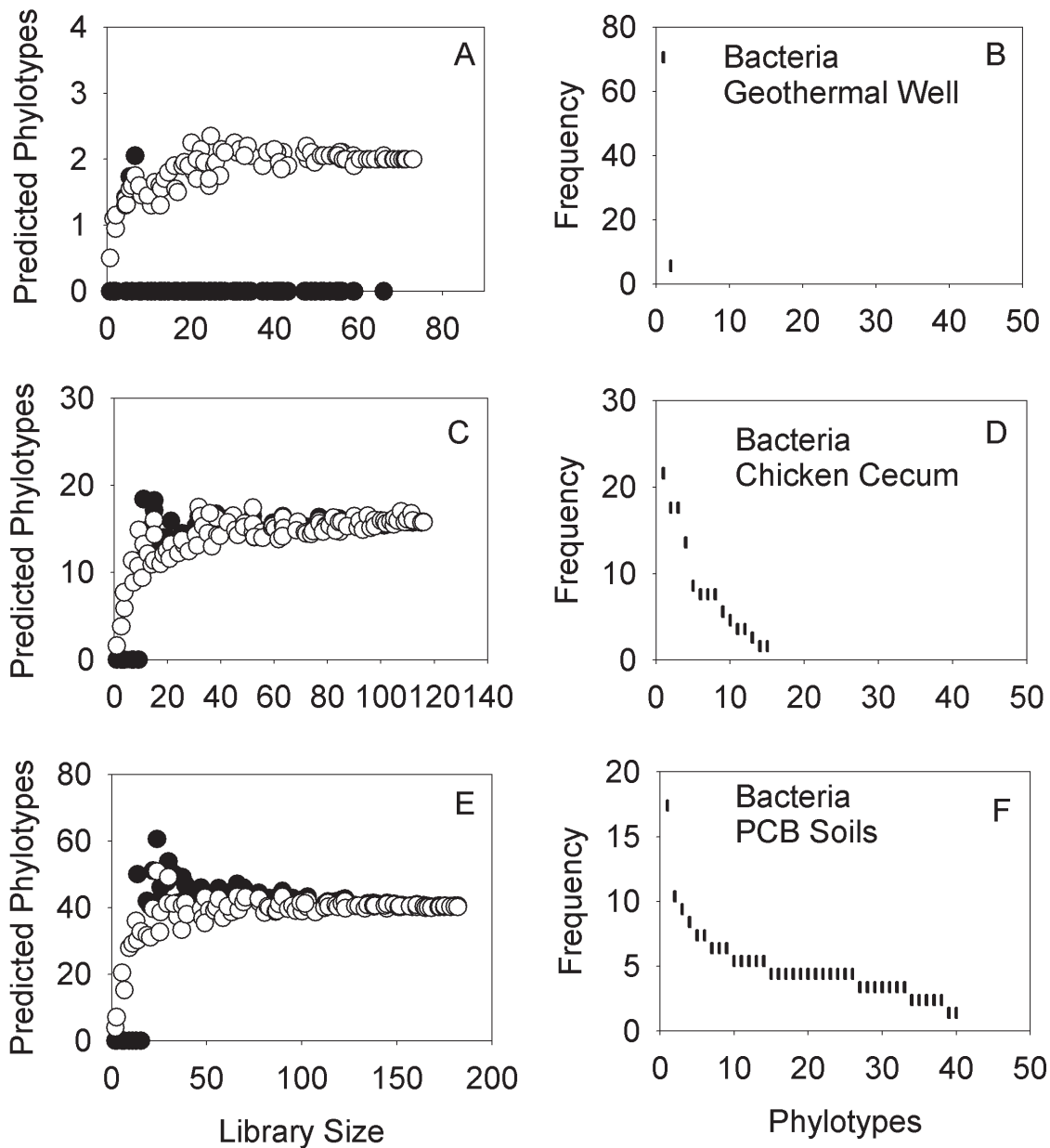


Fig. 3. Left panels: Predicted number of phylotypes based on S_{ACE} (filled symbols) and S_{Chao1} (open symbols) versus size of subsamples of three libraries derived from natural prokaryotic communities. Each point is the mean of 10 replicate subsamples of the library. All three were considered to have exhaustively sampled diversity in their source communities. Right panels: The corresponding phylotype frequency distribution for each library. Phylotype–richness estimates reach an asymptotic maximum for all three libraries, indicating that these libraries were large enough to yield stable and unbiased estimates of phylotype richness. A-B: Marteinsson et al. 2001. C-D: Gong et al. 2002. E-F: Nogaes et al. 2001.

library is large enough, but cannot predict how much larger an inadequate library should have been. However, it is possible to make some general comments about the minimum necessary library size by examining its relation to the distribution of phylotypes in libraries. We selected a wide variety of libraries that yielded a stable phylotype–richness estimate. These included 20 archaeal libraries and 26 bacterial libraries

from both aquatic and other environments. We estimated the minimum subsample size required to yield a phylotype–richness estimate within 5% of the asymptotic value of S_{Chao1} . The minimum library size was expressed as a multiple of the number of phylotypes present in the source community and plotted against the value of Simpson’s evenness index for the corresponding library (Fig. 6).

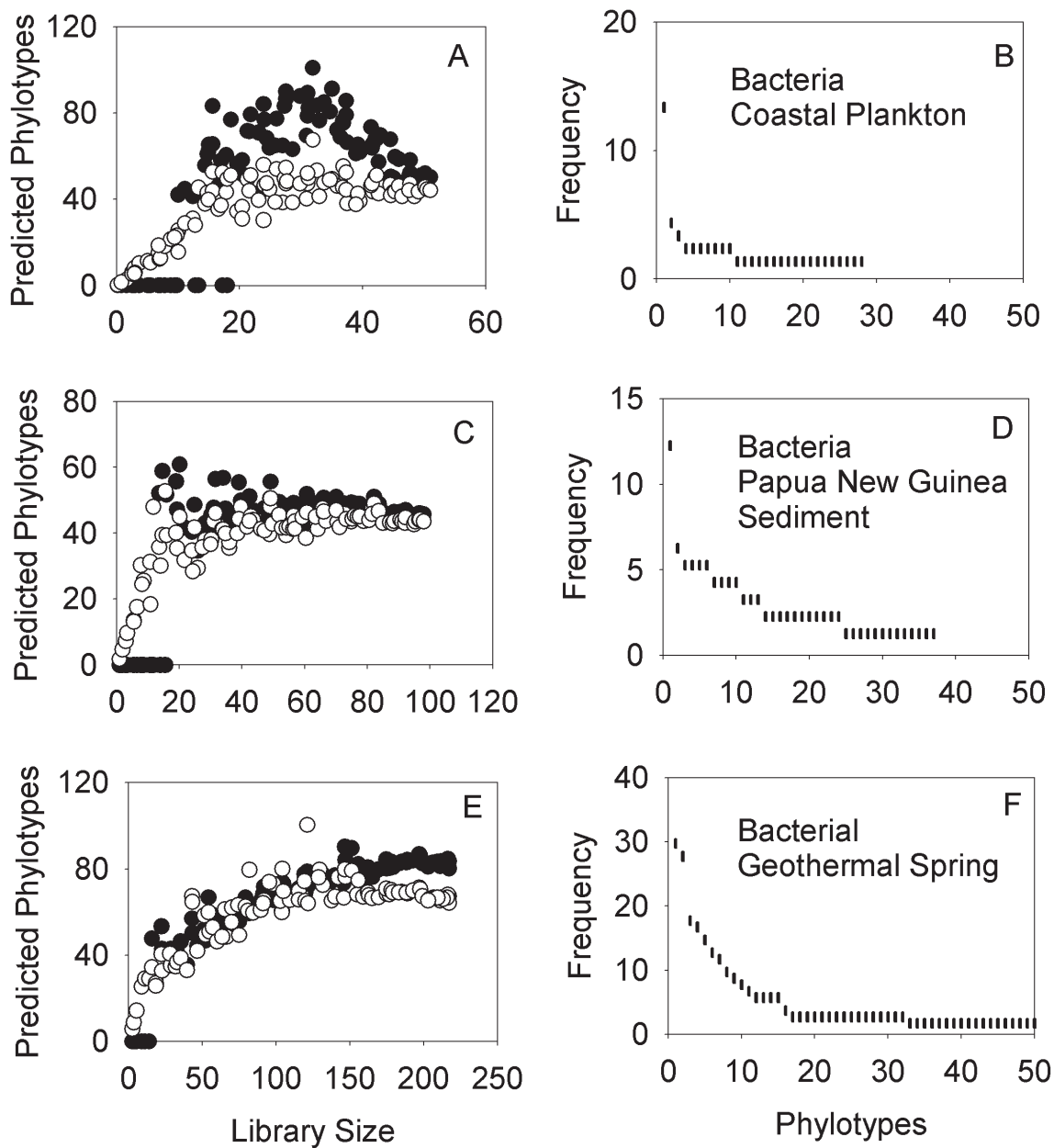


Fig. 4. Left panels: Predicted number of phylotypes based on S_{ACE} (filled symbols) and S_{Chao1} (open symbols) versus size of subsamples of three libraries derived from aquatic prokaryotic communities. Each point is the mean of 10 replicate subsamples of the library. None of these libraries were considered to have exhaustively sampled diversity. Right panels: The corresponding phylotype frequency distribution for each library. Phylotype–richness estimates reach an asymptotic maximum for all three libraries, indicating that these libraries were large enough to yield stable and unbiased estimates of phylotype richness. A-B: Rappe et al. 1997. C-D: Todorov et al. 2000. E-F: Hugenholtz et al. 1998.

Most of these “large enough” libraries have a relatively even distribution of phylotypes, comparable to the broken stick model library. The minimum relative library size decreases with increasing evenness ($r = 0.83$; Fig. 6) and was generally <4 for very even libraries, meaning that one could estimate phylotype richness S with a library of $<4S$ clones. In contrast, we would expect that highly uneven communities such as the ones depicted in Fig. 5 would require libraries that

are an order of magnitude larger than the number of phylotypes present (e.g., upper left quadrant of Fig. 6), before one could successfully estimate the number of phylotypes present.

Discussion

The majority of published 16S rDNA libraries for aquatic prokaryotic communities lack a context for judging whether they represent diversity in the environment sufficiently well to

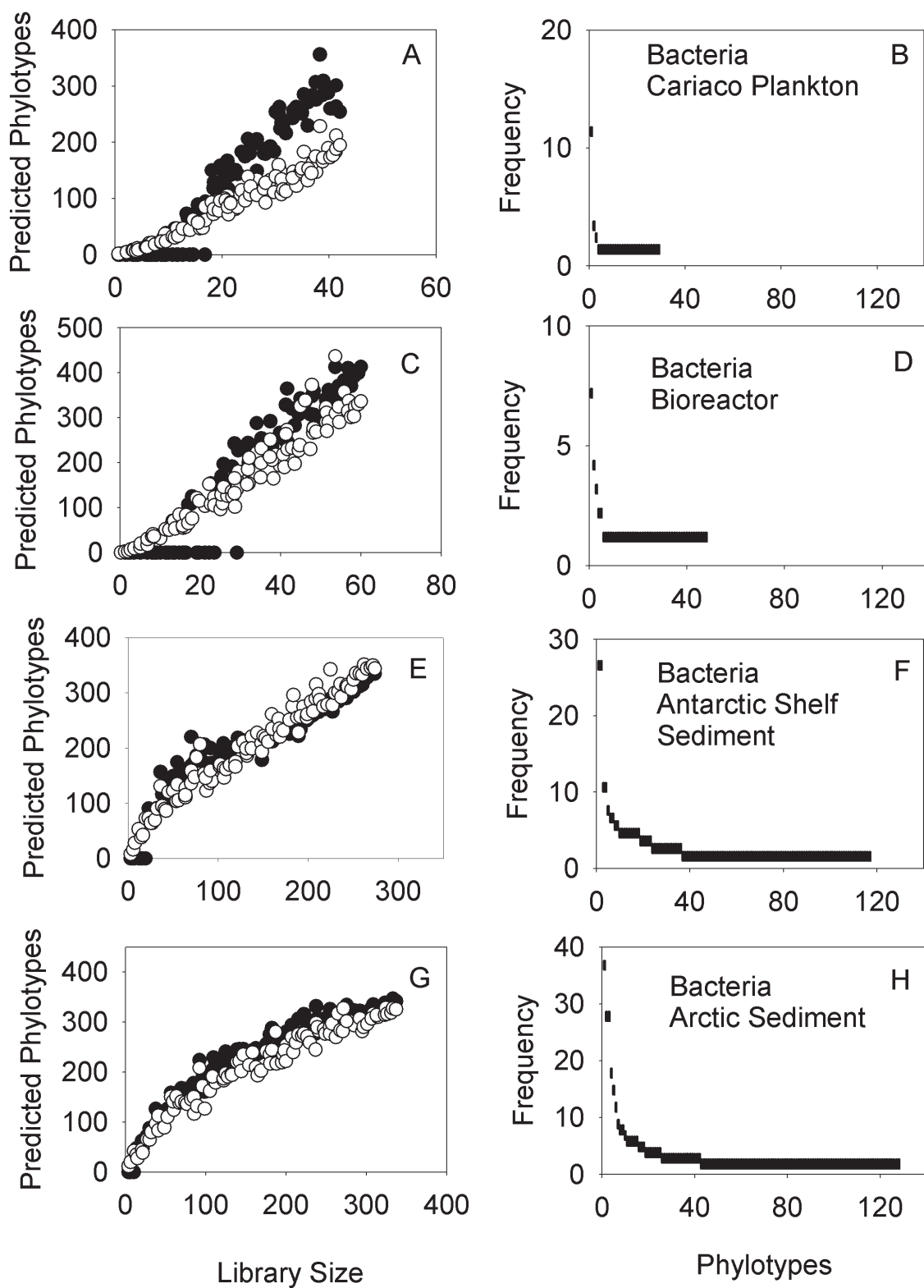


Fig. 5. Left panels: Predicted number of phylotypes based on S_{ACE} (filled symbols) and S_{Chao1} (open symbols) versus size of subsamples of four libraries derived from natural prokaryotic communities. Each point is the mean of 10 replicate subsamples of the library. Right panels: The corresponding phylotype frequency distribution for each library. For all four libraries, values of both estimators did not reach an asymptotic maximum, indicating that these libraries were not large enough to yield stable and unbiased estimates; i.e., phylotype richness was underestimated. A-B: Madrid et al. 2001. C-D: Sekiguchi et al. 1998. E-F: Bowman and McCuaig 2003. G-H: Ravensschlag et al. 1999.

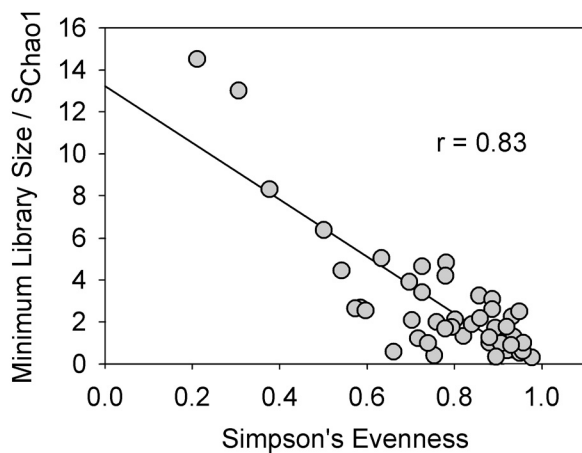


Fig. 6. Library size at which S_{Chao1} stabilized, expressed as multiples of the S_{Chao1} estimate of phylotype richness versus Simpson's evenness index.

address the questions being asked. Richness estimators offer a convenient and formulaically simple method to estimate total phylotype richness in the source community and can provide the context to make that judgment. The work presented here shows that the S_{Chao1} estimator performs reliably for a wide variety of model and real phylotype abundance distributions, ranging from completely uniform to high dominance and from very low to high diversity. Stable S_{Chao1} values are also unbiased or at worst minimally biased estimates of phylotype richness. S_{Chao1} has been favored by others as well (Foggo et al. 2003; Hill et al. 2003; Colwell and Coddington 1994).

The S_{ACE} estimator generally agreed with the S_{Chao1} estimator, but in some cases stabilized at a larger subsample size than the S_{Chao1} estimator, and in some cases, yielded richness estimates that were higher than estimates derived with S_{Chao1} . The S_{ACE} estimator was also undefined under some conditions, and in some cases, grossly overestimated phylotype richness at intermediate subsample sizes (e.g., Fig. 4A). It does not appear to be as well suited for estimating prokaryotic phylotype richness from 16S rDNA libraries.

A likely criticism of phylotype–richness estimates is that they are based on library composition, and libraries may not truly represent the relative abundances of phylotypes in nature. Both sampling artifacts and PCR amplification artifacts may increase or decrease the representation of phylotypes in a library (e.g., Reysenbach et al. 1992; Suzuki and Giovannoni 1996; Suzuki et al. 1998) leading to inaccurate estimates of phylotype richness. For obvious reasons, investigators constructing libraries do their best to ensure that libraries are not subject to such artifacts. Whether they are entirely successful or not, 16S rDNA libraries are the tool most frequently used to explore and develop a greater understanding of patterns in prokaryotic diversity. The procedure we have described and evaluated here allows investigators to avoid the pitfall of basing their conclusions upon an undersampled and inadequately represented diversity.

Martin (2002) has commented on two other problems associated with assessments of prokaryotic diversity based on 16S rDNA libraries. He noted, as have many others, that a given library will contain greater or fewer phylotypes depending on the percent similarity used to distinguish among sequences. Phylotype–richness estimates may be affected by the criteria used to separate sequences into phylotypes, e.g., if a researcher employs a 99% similarity cutoff instead of 97% similarity, or the converse. For libraries in which many sequences are considered identical at 97% but different at 99%, the effect of using a higher similarity cutoff is to generate more phylotypes, each represented by fewer clones. For both S_{ACE} and S_{Chao1} , phylotype–richness estimates increase with an increasing presence of rare phylotypes, and libraries with many rare phylotypes (e.g., as in Fig. 5) generally do not yield stable estimates of phylotype richness. However, we have observed that most published libraries are based on a very narrow range of percent similarities (97% to 99%), and we find no significant relationship evident between percent similarity and the number of phylotypes reported in libraries (Kemp and Aller 2003). It is possible that, in some libraries, changing the similarity cutoff level does not change the perceived number of phylotypes present, or at least not enough to materially affect richness estimates.

Martin (2002) also commented that although two libraries may have no phylotypes in common, the phylotypes in one library may have closely related counterparts in the other, and the libraries may be more similar than the absence of identical phylotypes would suggest. We agree with Martin that analyses of phylogenetic relatedness can provide different information and are valuable alternatives to comparisons of phylotype richness. However, we would argue that a comparison of two libraries based on phylogenetic relatedness will be just as flawed as a comparison based on phylotype richness if the two libraries represent different proportions of the diversity present in their respective source environments. In either case, one must first establish that both libraries represent diversity in the source environment sufficiently well for their intended purpose, whether that purpose is a comparison of phylotype richness or a comparison of phylogenetic relatedness.

Comments and recommendations

We are intrigued by finding that the libraries judged too small to yield stable estimates of phylotype richness ranged greatly in size but were usually characterized by a highly uneven phylotype abundance distribution (Fig. 5). The distribution of phylotypes generally resembled the geometric series model distribution but contained even more rare phylotypes (i.e., one clone each) than would be expected in a geometric series. In contrast, most of the “large enough” libraries tended to have a more even phylotype abundance distribution, comparable with broken stick or lognormal distributions (Figs. 2 to 4). Assuming that they are not entirely artificial, the preponderance of rare phylotypes in most libraries (including 138 of

194 libraries in our 2003 review) leads us to contemplate what processes might allow a very large number of phylotypes to persist at very low frequencies in a majority of aquatic prokaryotic communities.

A procedure comparable to the one used here could be applied while constructing a library in stages: calculating S_{Chao1} and characterizing additional clones until the estimated phylotype richness appears to reach an asymptote. At that point, it would be possible to assess whether the library adequately represents the diversity in the source environment for the problem under consideration. We strongly recommend this approach to avoid spending either too little or too much effort on library building.

Although the calculations required for this assessment procedure are not difficult, they are laborious. We constructed a web interface, form processor, and spreadsheet to enter library information, subsample the library, calculate values of S_{Chao1} and S_{ACE} , and plot richness estimates against subsample size. The web interface, form processor, spreadsheet, and instructions for their use are available through a web appendix located at <http://www.aslo.org/lomethods/free/2004/0114a.html>.

References

- Bond, P. L., S. P. Smriga, and J. F. Banfield. 2000. Phylogeny of microorganisms populating a thick, subaerial, predominantly lithotrophic biofilm at an extreme acid mine drainage site. *Appl. Environ. Microbiol.* 66:3842-3849.
- , S. A. McCammon, S. M. Rea, and T. A. McMeekin. 2000a. The microbial composition of three limnologically disparate hypersaline Antarctic lakes. *FEMS Microbiol. Lett.* 183:81-88.
- , S. M. Rea, S. A. McCammon, and T. A. McMeekin. 2000b. Diversity and community structure within anoxic sediment from marine salinity meromictic lakes and a coastal meromictic marine basin, Vestfold Hills, Eastern Antarctica. *Environ. Microbiol.* 35:227-237.
- , and R. D. McCuaig. 2003. Biodiversity, community structural shifts, and biogeography of prokaryotes within Antarctic continental shelf sediment. *Appl. Environ. Microbiol.* 69:2463-2483.
- Chao, A. 1984. Non-parametric estimation of the number of classes in a population. *Scand. J. Statistics* 11:265-270.
- . 1987. Estimating the population size for capture-recapture data with unequal catchability. *Biometrics* 43:783-791.
- , M. -C. Ma, and M. C. K. Yang. 1993. Stopping rules and estimation for recapture debugging with unequal failure rates. *Biometrika* 80:193-201.
- Colwell, R. K., and J. A. Coddington. 1994. Estimating terrestrial biodiversity through extrapolation. *Phil. Trans. R. Soc. Lond. B* 345:101-118.
- Curtis, T. P., W. T. Sloan, and J. W. Scannell. 2002. Estimating prokaryotic diversity and its limits. *Proc. Natl. Acad. Sci. U.S.A.* 99:10494-10499.
- Foggo, A., M. J. Attrill, M. T. Frost, and A. A. Rowden. 2003. Estimating marine species richness: an evaluation of six extrapolative techniques. *Mar. Ecol. Prog. Ser.* 248:15-26.
- Gong, J., R. J. Forster, H. Yu, J. R. Chambers, P. M. Sabour, R. Wheatcroft, and S. Chen. 2002. Diversity and phylogenetic analysis of bacteria in the mucosa of chicken ceca and comparison with bacteria in the cecal lumen. *FEMS Microbiol. Lett.* 208:1-7.
- Heck Jr., K. L., G. van Belle, and D. Simberloff. 1975. Explicit calculation of the rarefaction diversity measurement and the determination of sufficient sample size. *Ecology* 56:1459-1461.
- Hill, T. C. J., K. A. Walsh, J. A. Harris, and B. F. Moffet. 2003. Using ecological diversity measures with bacterial communities. *FEMS Microbiol. Ecol.* 43:1-11.
- Hugenholtz, P., B. M. Goebel, and N. R. Pace. 1998. Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *J. Bacteriol.* 180:4765-4774.
- Hughes, J. B., J. J. Hellmann, T. H. Ricketts, and B. J. M. Bohannan. 2001. Counting the uncountable: Statistical approaches to estimating microbial diversity. *Appl. Environ. Microbiol.* 67:4399-4400.
- Kemp, P. F., and J. Y. Aller. 2003. Bacterial diversity in aquatic and other environments: what 16S rDNA libraries can tell us. *FEMS Microbiol. Ecol.* 1600:1-17.
- MacArthur, R. H. 1957. On the relative abundance of bird species. *Proc. Nat. Acad. Sci. U.S.A.* 43:293-295.
- Madrid, V. M., G. T. Taylor, M. L. Scranton, and A. Y. Chistoserdov. 2001. Phylogenetic diversity of bacterial and archaeal communities in the anoxic zone of the Cariaco Basin. *Appl. Environ. Microbiol.* 67:1663-1674.
- Marteinsson, V. T., S. Hauksdóttir, C. F. V. Hobel, H. Kristmannsdóttir, G. L. Hreggvidsson, and J. K. Kreistjansson. 2001. Phylogenetic diversity analysis of subterranean hot springs in Iceland. *Appl. Environ. Microbiol.* 67:4242-4248.
- Martin, A. P. 2002. Phylogenetic approaches for describing and comparing the diversity of microbial communities. *Appl. Environ. Microbiol.* 68:3673-3682.
- May, R. M. 1975. Patterns of species abundance and diversity, p 81-120. *In* M. I. Cody and J. M. Diamond [eds.], *Ecology and evolution of communities*. Harvard Univ. Press.
- Nogales, B., E. R. B. Moore, E. Llobet-Brossa, R. Rossello-Mora, R. Amann, and K. N. Timmis. 2001. Combined use of 16S ribosomal DNA and 16S rRNA to study the bacterial community of polychlorinated biphenyl-polluted soil. *Appl. Environ. Microbiol.* 67:1874-1884.
- Preston, F. W. 1962. The canonical distribution of commonness and rarity. *Ecology* 43:185-215.
- Rappe, M. S., P. F. Kemp, S. J. Giovannoni. 1997. Phylogenetic diversity of marine coastal picoplankton 16S rRNA genes cloned from the continental shelf off Cape Hatteras, N.C. *Limnol. Oceanogr.* 42:811-826.
- Ravenschlag, K., K. Sahn, J. Pernthaler, and R. Amann. 1999. High bacterial diversity in permanently cold marine sediments. *Appl. Environ. Microbiol.* 6:3982-3989.

- Reysenbach, A. -L., L. J. Giver, G. S. Wickham, and N. R. Pace. 1992. Differential amplification of rTNA genes by polymerase chain reaction. *Appl. Environ. Microbiol.* 58:3417-3418.
- Sekiguchi, Y., Y. Kamagata, K. Syutsubo, A. Ohashi, H. Harada, and K. Nakamura. 1998. Phylogenetic diversity of mesophilic and thermophilic granular sludges determined by 16S rRNA gene analysis. *Microbiology* 144:2655-2665.
- Stach, J. E. M., L. A. Maldonado, D. G. Masson, A. C. Ward, M. Goodfellow, and A. T. Bull. 2003. Statistical approaches for estimating actinobacterial diversity in marine sediments. *Appl. Environ. Microbiol.* 69:6189-6200.
- Suzuki, M. T., and S. J. Giovannoni. 1996. Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. *Appl. Environ. Microbiol.* 62:625-630.
- , M., M. S. Rappe, and S. J. Giovannoni. 1998. Kinetic bias in estimates of coastal picoplankton community structure obtained by measurements of small-subunit rRNA gene PCR amplicon length heterogeneity. *Appl. Environ. Microbiol.* 64:4522-4529.
- Takai, K., and Y. Sako. 1999. A molecular view of archaeal diversity in marine and terrestrial hot water environments. *FEMS Microbiol. Ecol.* 28:177-188.
- Todorov, J. R., A. Y. Chistoserdov, and J. Y. Aller. 2000. Molecular analysis of microbial communities in mobile deltaic muds of Southeastern Papua New Guinea. *FEMS Microbiol. Ecol.* 33:147-155.

Submitted 2 January 2004

Revised 2 February 2004

Accepted 27 February 2004